# PREDICTION OF PROTEIN FUNCTIONS THROUGH THE PROCESSING OF PROTEIN-PROTEIN INTERACTION DATA NETWORK

**By**

**Eng. Khaled El-Sayed Ahmed Mostafa**

**A Thesis submitted to the**

**Faculty of Engineering, Cairo University**

**In partial fulfillment of the**

**Requirements for the Degree of**

**DOCTOR OF PHILOSOPHY**

**IN**

**SYSTEMS AND BIOMEDICAL ENGINEERING**

**FACULTY OF ENGINEERING CAIRO UNIVERSITY**
**GIZA EGYPT**
**2011**

# PREDICTION OF PROTEIN FUNCTIONS THROUGH THE PROCESSING OF PROTEIN-PROTEIN INTERACTION DATA NETWORK

**By**

**Eng. Khaled El-Sayed Ahmed Mostafa**

**A Thesis submitted to the**

**Faculty of Engineering, Cairo University**

**In partial fulfillment of the**

**Requirements for the Degree of**

**DOCTOR OF PHILOSOPHY**

**IN**

**SYSTEMS AND BIOMEDICAL ENGINEERING**
**Under the supervision of**

**Prof.Dr. Abou-Bakr M. Yousef**
**System and Biomedical Engineering department**
**Cairo university, Egypt**

**Prof.Dr. Yasser Mostafa Kadah**
**System and Biomedical Engineering department**
**Cairo university, Egypt**

**Associate Prof. Nahed Hussien Salouma**
**Laser Institute Cairo university, Egypt**

**FACULTY OF ENGINEERING CAIRO UNIVERSITY**
**GIZA EGYPT**
**2011**

# Contents

Acknowledge
List of Figures
List of Tables
List of Abbreviations

Abstract ……..1

1- Introduction
1.1 Thesis Objective ……..……...3
1.2 Problem Definition and Motivation…….....4
1.3 Thesis Objectives ……....6
1.4 Thesis Organization …….....7
2- Biological Background
2.1 Introduction to Proteomics ……..……..8
2.1.1 Historical development……...9
2.1.2 The need for Proteomics…….....9
2.1.3 Proteomics Impact on heath life…….……...10
2.1.4 Proteomics Research Areas……...11
2.2 Organisms and cells …… 11
2.2.1 Prokaryotic cells …….....11
2.2.2 Eukaryotic cells …….……....12
2.3 Molecules of life……..13
2.3.1 Small molecules …….....13
2.3.2 DNA……14
2.3.3 RNA……15
2.3.4 Genes and Protein synthesis…….....17
2.4 Biological Database ……20
2.5 Summary ……22

3- Protein-Protein Interactions
3.1 Protein-Protein Interaction network …….....23
3.2 Challenging of PPI ……..25
3.3 PPI experimental methods ……....…….....26
3.3.1 Yeast two hybrid …….…….....27
3.3.2 Mass spectrometry……28
3.3.3 Microarray.……29
3.4 TAP method of complex purification ……30
3.3.5 Gene Co-expression ……31
3.3.6 Synthetic lethality method ……...32
3.4 Mathematical and graphical models for PPIs ……....33
3.4.1 Presentation of PPI network ……33
3.4.2 PPI network concepts……34
3.5 PPI prediction …….. ……..39
3.5.1 PPI prediction using known structure……...40
3.5.2 Prediction of PPI in absence of protein structure ……40
3.6 PPI databases ……..41
3.7 Summary ……43

4- Review of literature

4.1 Overview……...……....45
4.2 Neighbor counting ……...…….45
4.3 Chi-square method ……...48
4.4 Markov Random Field ……50
4.5 Prodistin ……....52
4.6 Samanta……..……..53
4.7 Support Vector Machine ……..53
4.8 Functional flow……54
5.9 Leave one out ……...……54
4.10 Summary……..……....54

5- A New Weighted Protein-Protein Interaction for Improvement of Yeast Protein Functions Prediction

5.1 Yeast Saccharomyces Cerevisiae ……....56
5.1.1 Yeast history ……...57
5.1.2 Why Yeast ……..58
5.1.3 Yeast features ……59
5.2 Challenges of the study ……61
5.2.1 Yeast protein naming ……..61
5.2.2 Protein clusters and interactions ……....62
5.2.3 Protein-protein interactions ……63
5.3 Protein function prediction using a new weighting algorithm for PPI ……..64
5.3.1 Protein degree and level ……....66
5.3.2 PPI weighting strategy ……68
5.3.2.1 Experimental verification method ……....68
5.3.2.2 Interaction Generality 1……....75
5.3.2.3 Interaction Generality 2……77
5.3.3 Protein function prediction by weights integration ……79
5.4 Summary… ……....82

6- Protein function correlation based on overlapping proteins and cluster interaction

6.1 Protein function correlation ……85
6.2 Protein cluster interaction ……88
6.3 Function categories and overlapping proteins ……91
6.4 Overlapping and interaction integration ……..93
6.5 Function relations ……96
6.5.1 Direct relations ……..96
6.5.2 Indirect relations ……...97
6.5.3 Protein function integration ……100
6.6 Summary… ……101

7- Yeast Protein Function Motif Extraction Based on Sequence Alignment

7.1 Protein sequence overview ……...103
7.2 Motif extraction ……104
7.2.1 Protein sequence collection ……104
7.2.2 Multiple sequence alignement ……..106
7.3 Summary.. ……112

8- Summary and future work ……113

References ……..117

# ACKNOWLEDGMENT

First of all, I would like to thank ALLAH who support me and supplied me with power and faithful to comprehensive with this very difficult topic.

Second, this work could not come to light without Prof. Yasser Kadah guidance. I feel very privileged to have worked with Prof. Yasser, a brilliant scientist; I learned much from our many insightful discussions in scientific field or in usual life. Really I hope to be as Prof. Yasser.

My gratitude and deep appreciation is to Prof. Nahed, for her generous advice and motivation toward finishing this topic.

Also the support of my family specially my father Mr. Sayed, my mother, Dr. Farouk, wife's mother , wife (Eng. Fayroz), my brother (Dr. Waleed), my sisters, and my lovely and honey kids (Abdullah and Haneen). In addition to Prof. Abdullah in the System and Biomedical Engineering Department; Cairo University.

Also, I would like to thank Prof. M. Salah Abbas the Dean of faculty of Engineering in Modern University for Technology and Information (MTI). In addition to all staff and friends in my college.
I present this work to MTI and Cairo University, and I wish to contribute some thing valuable to these communities.

## List of Figures


| | | |
|---|---|---|
| Fig.2.1 | Different fields affect on proteomics | 9 |
| Fig.2.2 | Eukaryote versus Prokaryote | 12 |
| Fig.2.3 | A single stranded DNA polynucleotide | 14 |
| Fig.2.4 | Double stranded DNA and its nucleotides | 15 |
| Fig.2.5 | Double helix DNA and its hydrogen bonds | 15 |
| Fig.2.6 | Secondary structure for E. coli RNA | 16 |
| Fig.2.7 | mRNA processing | 16 |
| Fig.2.8 | tRNA secondary structure | 17 |
| Fig.2.9 | 64 amino acids codons | 17 |
| Fig.2.10 | Basic structure of amino Acid | 18 |
| Fig.2.11 | The different types of AAs | 18 |
| Fig.2.12 | Central dogma of molecular biology | 19 |
| Fig.2.13 | The different structure of proteins | 20 |
| Fig.2.14 | Different formats for DNA and protein sequences | 21 |
| Fig.3.1 | Protein-protein interaction network | 24 |
| Fig.3.2 | Example of Y2H interaction | 28 |
| Fig.3.3 | Example of microarray | 30 |
| Fig.3.4 | Example of Tap-Affinity experimental method | 30 |
| Fig.3.5 | Example of gene co-expression | 32 |
| Fig.3.6 | Example of synthetic lethality experiment | 32 |
| Fig.3.7 | Basic structure of protein network | 33 |
| Fig.3.8 | The number of red nodes is the degree of node A | 34 |
| Fig.3.9 | Random reconnection procedure of a regular ring graph. | 36 |

| | | |
|---|---|---|
| Fig.3.10 | A scale free distribution | 37 |
| Fig.3.11 | Examples of modular networks composed of two modules | 38 |
| Fig.3.12 | The architecture of the hierarchical network model | 39 |
| Fig.3.13 | X-ray crystallography device | 40 |
| Fig.4.1 | Annotated and un-annotated proteins and their physical interaction | 46 |
| Fig.4.2 | Sample of proteome network | 48 |
| Fig.5.1 | Yeast as simple model organism | 57 |
| Fig.5.2 | Yeast function annotation | 59 |
| Fig.5.3 | Yeast function categories | 61 |
| Fig.5.4 | Yeast proteins interactions | 64 |
| Fig.5.5 | (a) small routing connection system (b) large routing system | 64 |
| Fig.5.6 | A graph of connected proteins indicates the leafs (yellow nodes) | 64 |
| Fig.5.7 | Degree of the Protein (black node A has degree equal 6) | 66 |
| Fig.5.8 | Real part of yeast proteome using Inter View program | 66 |
| Fig.5.9 | Experimental methods related to the interaction pairs | 70 |
| Fig.5.10 | Experimental method type and its number of interaction pairs | 70 |
| Fig.5.11 | Experimental methods and protein function prediction | 73 |
| Fig.5.12 | The sensitivity and specificity of protein function prediction in cell location for w0, w1, w2 with k=2-5 as number of interactions | 72 |
| Fig.5.13 | The sensitivity and specificity of protein function prediction in Biochemical for w0, w1, w2 with k=2-5 as number of interactions | 74 |
| Fig.5.14 | The sensitivity and specificity of protein function prediction in cellular role for w0, w1, w2 with k=2-5 as number of interactions | 75 |
| Fig.5.15 | IG1 value for protein interaction network | 75 |
| Fig.5.16 | The relation between the number of interactions and IG1 | 76 |

| | | |
|---|---|---|
| Fig.5.17 | The most common five sub-graphs of network | 78 |
| Fig.5.18 | The sensitivity and specificity of the Biochemical function category for number of interactions k=2:5 | 81 |
| Fig.5.19 | The sensitivity and specificity of the cell location function category for number of interactions k=2:5 | 81 |
| Fig.5.20 | The sensitivity and specificity of the cellular role function category for number of interactions k=2:5 | 82 |
| Fig.6.1 | proteins have the same functions; correlation between functions | 86 |
| Fig.6.2 | the relation between the proteins and functions | 87 |
| Fig.6.3 | two interacted clusters | 88 |
| Fig.6.4 | relations between the Biochemical sub-function category_2 towards the sub-function category_1 | 92 |
| Fig.6.5 | directed relation between the two sub-function categories 11, 1 and its weight equal (100%) | 96 |
| Fig.6.6 | conditional relationship between the sub-function categories | 97 |
| Fig.6.7 | anti correlation between the two sub-functions category B, C given sub-function category A | 98 |
| Fig.6.8 | Conditional relation between two functions | 99 |
| Fig.6.9 | a complete combination between sub-function category 4 and the rest functions of the cellular role category in Yeast | 100 |
| Fig.7.1 | GCG sample format | 106 |
| Fig.7.2 | Staden format sequence | 106 |
| Fig.7.3 | Gene-bank sequence format | 107 |
| Fig.7.4 | FASTA format for protein BNI5 of Yeast | 107 |
| Fig.7.5 | Specific motif (consensus) extracted from the first 10 proteins locations | 108 |
| Fig.7.6 | The relation between the conservation and alignment position | 109 |

Fig.7.7 spectrum alignment graphs (consensus versus the position) for Biochemical sub-function categories 110

## List of Tables


| | | |
|---|---|---|
| Table 2.1 | Comparison between Prokaryote and Eukaryote | 11 |
| Table 4.1 | Basic and estimated functions for Yeast proteome using NCM | 47 |
| Table 4.2 | Basic and estimated functions for Yeast proteome using Chi-square method | 49 |
| Table 4.3 | Edges number for the three states of the interaction network | 51 |
| Table 5.1 | Yeast sub-function categories, function name, and the number of proteins for each function | 60 |
| Table 5.2 | Yeast proteins and their different names | 62 |
| Table 5.3 | Numbers of Annotated and un-annotated proteins for All Proteins Based on Three Functional categories | 63 |
| Table 5.4 | Sample of proteins and their interactions | 67 |
| Table 5.5 | Yeast interaction pairs, number of identification methods | 69 |
| Table 5.6 | The reliability scale of some experimental methods regarding to the GRID datasets | 71 |
| Table 5.7 | comparison between the two used methods for determining the reliability for protein interactions | 72 |
| Table 5.8 | the reliability score of IG1 of protein interactions | 77 |
| Table 5.9 | IG2 values for yeast protein interactions | 78 |
| Table 5.10 | Weights for yeast protein interactions. w1- w5 are experimental methods, IG1, IG2, average, and weights PCA | 80 |
| Table 6.1 | Annotated and un-annotated proteins for All Proteins Based on Three Functional categories | 86 |

| | | |
|---|---|---|
| Table 6.2 | Yeast sub-function categories, function name and the number of proteins for each function | 86 |
| Table 6.3 | relations between yeast protein functions based on the number of interactions | 90 |
| Table 6.4 | overlapping number of proteins over Yeast biochemical function categories | 92 |
| Table 6.5 | direct relations over Yeast Biochemical sub-function categories when the threshold is greater than 0.85 | 92 |
| Table 6.6 | yeast biochemical functions and estimated numbers of proteins | 95 |
| Table 6.7 | an integrated algorithm relating to the overlapping number of proteins and number of interactions | 95 |
| Table 7.1 | Sample of protein names and their sequences from DIP database… | 105 |
| Table 7.2 | Different data sources for protein names and their codes | 105 |
| Table 7.3 | Yeast protein functions and its extracted motifs | 111 |

# ABSTRACT

Protein function prediction is one of the most important and hot tasks in the field of proteomics, since it leads to understanding cell activities. Protein functions may be predicted from protein sequences, gene expression, protein domains, protein localizations, protein structure, and protein-protein interactions (PPI) as recent computational techniques.

Although protein function prediction through PPI networks is a powerful modality, it lacks the following points: 1) the reliability of the protein interactions to be considered in the prediction process where each interaction can be identified by one or more experimental method. And each experimental method has its score of stability and reliability. 2) The relations between the known protein functions and correlation which affect the prediction process. and 3) the features that identify these functions. Most of the previous computational techniques do not consider these points; that is why it decreases the confidence of the prediction process.

In this thesis, some algorithms are provided with new ideas to overcome the above-mentioned drawbacks. Regarding the reliability, an integrated algorithm is proposed. It includes the experimental identification method; that includes the number of experimental methods furthermore their reliabilities, local topology which indicates the number of surroundings for the studied proteins, and global topology which illustrates the most common graphs for the proteins through the network. In addition, a new weighting algorithm has been calculated using all the previous data. This new technique explores the collected data to create reliable interactions and enhance the prediction process.

Moreover, a novel technique is introduced to express the relations between protein functions, including number of interactions between the protein clusters and overlapping number of proteins that have the same functions. This technique indicates the correlation, anti-correlation, and independency between some protein functions which affects the protein function prediction.

Motif extraction is also performed using different techniques as multiple sequence alignment (MSA) in order to take advantage of the features that identify protein functions. This consensus (the most common positions of amino acids for proteins in multiple sequences alignment) is considered as the signature of that function and is used to identify it.

The proposed techniques are applied to Yeast data "Saccharomyces Cerevisiae" the simple eukaryote species which has complete genome and sequences. Yeast has a round 6500 proteins which can be classified into three main function categories. Each one of those function categories (biochemical – cell location – cellular role) contains many sub-functions.

The obtained results are validated via valuable methods and the results revealed great enhancement in protein function prediction process. The sensitivity and specificity of the results are more reliable than the previous techniques.

# Chapter 1

# Introduction

## 1.1 Thesis Overview

One of the most important challenges of the post-genomic era is determining protein functions. Due to this reason, Automated Function Prediction is currently one of the most active research fields. Furthermore, the availability of entire genome sequences and high-throughput capabilities that helps determining gene co-expression patterns has shifted the research focus from studying single proteins or small complexes to the entire proteome. Since, discovering the new functions of un-annotated proteins has led to understanding the cell function; a lot of methods have been implemented to satisfy this object.

Here, integrated techniques that can be applied to the protein interaction networks are presented to predict more reliable protein functions. These techniques use a new weighting method for determining protein reliability; moreover use a novel algorithm to discover the correlation between protein functions. Feature selection techniques are applied to focus on the subset of relevant variables. **S**everal computational approaches will be used for finding specific genetic signatures characteristic of each function. We subsequently validated the robustness of those signatures with a set of test sequences. The results obtained from the proposed algorithms will be analyzed, validated and compared with the previous work.

## 1.2Problem Definition and Motivation

In the past, Biologists tried to determine protein functions from the structure of the protein and similar proteins. Possible roles of similarity between the protein and its homologies - from other organisms - were suggested and investigated to predict protein functions.

Due to the different groups of homologous, these methods were found to be exhaustive and uncertain. That is why other techniques have been used to predict the protein functions by analyzing gene expression patterns [1, 2], phylogenetic profiles [3, 4, 5], protein sequences [6, 7], protein domains [8, 9]. These technologies suffered from high error rates because of their inherent limitations. Another technique which depends on integrated multi sources was used [10, 11].

The computational approach, which has been adopted to solve these problems, uses information gained from physical and genetic interaction maps to predict protein functions.

Recently, researchers introduced different techniques to determine the probability of protein function prediction using the information extracted from Protein-Protein Interactions (PPI). Although these trials are promising, they lack the solving major problems such as network topology and strength of interaction.

Network topology represents the interaction between proteins and how they are connected. This means that, a lot of information can be extracted from these networks regarding the strength of interaction and its contribution to new function prediction i.e. weighted contribution. A PPI network can be described as a complex system of proteins linked by interactions. The computational analysis of PPI networks begins with the representation of the PPI network structure. On the other hand, the simplest representation takes the form of a network graph consisting of nodes and edges [12]. Proteins are represented as nodes in the graph and two proteins that interact physically are represented as adjacent nodes connected by an edge [13]. Based on this graphical representation, various computational approaches, such as data mining, machine learning, and statistical approaches can be performed to reveal the PPI networks at different levels.

In general, the computational analysis of PPI networks faces some major problems. First, the unreliability of protein interactions, which comes from large-scale

experiments, that yields numerous false positions as Yeast two hybrids (Y2H). Second, protein may have more than one function and may be considered in one or more functional groups which lead to overlapped function clusters. Third, proteins with different functions may interact this means that PPI has connections between proteins in different functional groups which expand the topological complexity of the PPI networks.

*Neighbor counting* is a method proposed by Schwikowski et al. in [14] to infer the functions of an un-annotated protein from the PPI. This method finds the neighbor proteins and gets their assigned functions and the frequencies of occurrence of these functions. Then, these functions are arranged in descending order according to their frequencies. The first *k* functions are considered and assigned to the un-annotated protein. The authors in [15] used this technique with *k* equals to 3. Although this method exploits the information from the neighbors, it has some drawbacks: 1) it considers the interactions to be of equal weights which is not logic, 2) it does not take into consideration the nature of the function and whether it is dominant or not and 3) it does not provide a confidence level for assigning a function to the protein. The problem of confidence level was addressed in [16] where the authors used Chi-square statistics to calculate significant value based on the probability of the presence of different functions. Although chi-square method provides more deep analysis, it produces lower sensitivity and specification compared to the neighbor counting method. Deng et al., in [17] considered different situations of the presence of certain function for a protein of interest and described them as: 1) number of all proteins sharing this function, 2) number of protein pairs (interacted) and having the function, 3) number of protein pairs with one of them has the function and the other does not and 4) number of protein pairs does not have this function. A weighted sum of these numbers is calculated according to random Markov field algorithm at a time and assigned different weights, so Markov random field method [17, 18] introduces the overcome of all the above problems by considering the entire interaction network.

For it considers the frequency of proteins having the function of interest, as well as all the neighbors with less weight placed on, far away neighbors close ones, it can calculate the probability that an un-annotated protein has a function of interest. This method presents good results compared to the previous two methods.

## 1.3 Thesis Objective

Since the protein function prediction is one of the most important tasks in proteomics, the target of the thesis is to predict the un-annotated functions for proteins.

In this thesis, new integrated methods are provided to overcome the drawbacks of PPIs and to improve the accuracy of prediction. These drawbacks are: 1) - the reliability of the protein interactions which is considered in the prediction process, 2) - the relations between protein functions and 3) - the features that identify these functions.

For the reliability, an advanced integrated algorithm will be proposed. It includes the experimental identification method; the methods that identify the interactions in lab, local topology; the topology that indicates the position of the protein and its direct neighbors, and global topology, the topology that identify the position of the protein through the network. Moreover, new weighted algorithms have been calculated including Average weighted summer, Principal Component Analysis (PCA) and exploring the similarity between the proteins interactions and the connected routers in certain autonomous number of network explored. By applying the same idea of network linked list protocols as OSPF (Open Shortest Path First) information of surrounding routers will be clear according to the principals of the cost and level (hop count) [19, 20]. On the contrary, in the relation between proteins, a novel technique will be introduced to express these relations including number of interactions between the protein clusters, overlapping number of proteins that have the same functions, and integrated algorithm to collect the previous two techniques. **S**everal computational approaches will be used for finding specific genetic signatures characteristic of each function. We subsequently validated the robustness of those signatures with a set of test sequences.

The techniques applied to Yeast data “Saccharomyces Cerevisiae” and the generated results were validated via valuable methods to reveal great enhancement in protein function prediction process.

## 1.4 Thesis Organization

The remainder of this thesis is divided into 7 chapters.

**Chapter 2** presents the basic biological concepts to understand the rest of the thesis and presents a description of the data source as Yeast protein interactions used in this thesis.

**Chapter 3** describes the definition of PPI, challenges, methods of identification, mathematical and graphical models of protein interactions, PPI prediction methods, and PPI databases.

**Chapter 4** reviews the current literature pertaining to protein function prediction methods through PPI.

**Chapter 5** presents the contribution procedure; weighting the protein interactions and predicting their functions using the neighbor counting method. This chapter introduces the interaction weights by number and reliability of experimental methods furthermore the network topology either local or global then using neighbor counting method to get the functions.

**Chapter 6** describes an approach of estimating the correlations between the protein functions through the cluster interactions and overlapping number of proteins.

**Chapter 7** explains the genomic signatures and motif extraction methods for identifying protein functions.

**Chapter 8** discusses the results and the possible improvements in research as well as identifying the future work related to research areas.

# Chapter 2

# Biological background

This chapter is written as a suitable starting point for the readers who lack necessary biological background to read the rest of the thesis. In this chapter, we will discuss in brief, an introduction about Proteomics, the need for proteomics, and its impact on health life with current hot research areas (section 2.1). The second section is about the typical structure of a cell down to molecular level including major biological terminologies and genome organization in Eukaryotic and Prokaryotic cells (section 2.2). Later on, section 2.3 introduces a short description of the molecules of life as DNA, RNA, and protein as well as, its construction, functions, and structures. In the end of this chapter (section 2.4), the most important biological data sources used in this study is introduced.

## 2.1 Introduction to Proteomics

It is noticed that there is no absolute definition for proteomics. The most common definition is:"Proteomics is a modern science that is collected from Mathematics, Statistics, Biology and other fields to introduce the secrets of life" [21] as shown in Fig.2.1. The term Proteomics more properly refers to the identification of the cell functions. Another definition for proteomics is the theory to solve formal and practical problems imposed by or inspired from the management and analysis of biological data. Proteomics is concerned with developing new tools for analyzing of proteomic and molecular biological data including sequence analysis, proteomic algorithms, phylogenetic inference, and biological inspired computational models [22].

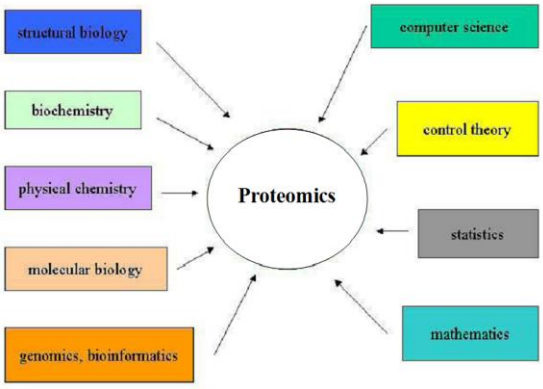


Fig.2.1 Different fields affect on proteomics

### 2.1.1 Historical development

Proteomics was coined in the early 1990s by Macquarie University PhD candidate, Marc Wilkins. "The protein complement of the genome". New technologies that allow researchers to visualize thousands of proteins at the same time revealing patterns that may have important clinical implications. Proteomics is the field of studying the proteome (Protein complement to a genome) [23]. Proteomics can be divided into three parts [24]: *Functional proteomics* **(**The identification of protein functions, activities or interactions at organism-wide scale). *Expressional proteomics* **(**The analysis of organism-wide changes in protein expression). *Structural proteomics* **(**The determination of protein structure by X-ray, NMR or computer-based methods). Proteomics uses the mathematical models algorithms to discover the functions of the cells and try to visualize the proteins response for the treatment. The concepts of the proteome and the field of proteomics are rapidly developing as new technology, and high-throughput techniques making the mapping of the entire human proteome seem like a dream – as the complete sequencing of genomes once was that has come true.

### 2.1.2 The need for Proteomics

In general, proteomics aims are in three directions; **P**roteomics tries to discover the functions of all the cells (the functions of all proteins inside the cell and determine each protein function). Second, **p**roteomics tries to analyze organisms and

explain the protein expression and its processes. Third, **p**roteomics attempts to determine the structure of the proteins of the cell (3D structure) because the structure is responsible for the interactions and the functions. Proteomics researchers try to explore the data stored in databases as PIR, DIP, SCOP and PDB for 3D macromolecular structures and use mathematical models and machine learning algorithms to reach the explanation of the protein function. The information stored in these databases is essentially useless until being analyzed. Thus, the purpose of proteomics extends much further to develop tools and resources that aid in the analysis of the data.

In this study, the estimation of protein functions through the protein interaction networks is performed. The used data is collected from different databases (database of protein interactions), and variable techniques have been applied.

### 2.1.3 Proteomics Impact on heath life

Proteomics is considered an empowering technology that helps the researchers in biotechnology taking a proactive role in defining and shaping the future of their fields and the world. A new technology of proteomics where researchers are to detect how cancer drugs work into the cells is created [25]. That can be performed by detecting proteins response for cancer and how proteins interact with. On the contrary, pharmaceutical industry has operated without bringing together the disciplines of biology, chemistry, and information technology [26]. That is why pharmaceutical industry appears to have been retarded, so other industries are implementing information technology to improve their operations. According to the genome project and the resultant data explosion, it is important to join these fields of science together to exploit the available data and thus expedite the drug discovery process. In the past, the drug discovery process used to take an average of 15 years to develop each new medicine before offering it to the market. Nearly 75% of drug candidates currently being tested by pharmaceutical companies fall short and never reach the market [27]. In an attempt to improve and reduce the cost of drug discovery, the pharmaceutical industry has recently turned to bioinformatics and proteomics which may reduce the cost to half.

### 2.1.4 Proteomics Research Areas

The main research areas for proteomics are: ***functional proteomics***; that discovers the functions of all the cells. It determines the functions for each protein and the integration between the interacted ones. Then ***structural proteomics***; it visualizes the folding shape structure of each protein (3D) to put it into relation with the function. Finally, the ***expressional proteomics***; it explains the protein expression. From the previous research areas, many techniques have been created to enhance specific actions.

## 2.2 Organisms and cells

All organisms consists of small cells, typically too small to be seen by naked eye, but big enough for an optical micro scope [28]. They are estimated about $6 * 10^{13}$ cells in human body of about 320 different types. For instance there are several types of skin cells, brain cells (neurons), and many others. The world of organisms could be divided into two types: prokaryotic and eukaryotic cells. The main differences between the prokaryotic and eukaryotic cells are introduced in Table 2.1 as follows:

Table 2.1 Comparison between Prokaryote and Eukaryote

| Feature | **Prokaryote** | **Eukaryote** |
|---|---|---|
| Size | Small about 0.5µm | Up to 40µm |
| Feature | Prokaryote | Eukaryote |
| Organelles | No organelles | Organelles |
| Genetic material | Circular DNA | Linear DNA and chromosome |

### 2.2.1 Prokaryotic Cells

Prokaryotic Cells which have a typical size of about 1 micron in diameter are smaller than eukaryotic cells as shown in Fig.2.2[1] and have simpler structure (e.g.,

they do not have any inner cellular membranes that are always present in eukaryotic cells) [29].

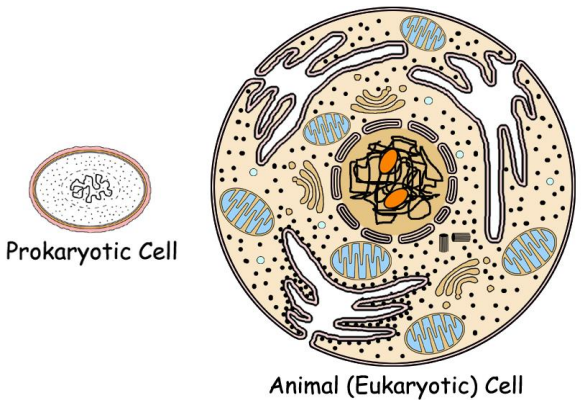


Fig.2.2 Eukaryote versus Prokaryote [Prokaryotic and Eukaryotic Cells]

Prokaryotic cells are single cellular organisms, but take into consideration that being a single cell does not mean that an organism is a prokaryote. Being smaller than eukaryotes does not mean that prokaryotes are of less important. For instance, it is quite likely that the number of bacteria living in the mouth and digestive tract of human are being larger than the number of eukaryotic cells in the same individual, and many of these bacteria are necessary for a human being to live a normal life (these numbers are rather difficult to estimate, rather a hypothesis). Some times prokaryotes are known as microbes.

### 2.2.2 Eukaryotic Cells

Eukaryotic cell has a nucleus; which is separated from the rest of the cell by a membrane. The nucleus contains chromosomes, which are the carrier of the genetic materials. There is internal membrane enclosed compartment within eukaryotic cells. Other organelles are found as: centiroles, lysosomes, golgi complexes, mitochondria which are specialized for particular biological processes.

The mitochondria are found in all eukaryotes and are specialized for energy production (respiration) chloroplasts are organelles found in plant cells which produce sugar using light. Light is the ultimate source of energy for all life on earth. The area of the cell outside the nucleus and the organelles is called the cytoplasm. Membranes

are complex structure and an effective barrier to the environment that regulates the flow of food, energy and information in and out the cell.

An essential feature of most (prokaryote and eukaryote) living cells is their ability to grow in an appropriate environment and to undergo cell division. The growth of a single cell and its subsequent division is called cell cycle. However, not all cells continually grow and divide, for example neurons only undergo an initial growth phase. Prokaryotes, particularly bacteria, are extremely successful in multiplying. It is likely that natural selection has favored single called organisms able to grow and divide quickly. Multi-cellular organisms typically begin life as a single cell, as a result of fusion of a male and female sex cell (gametes). The single cell has to grow; divide and differentiation need to be controlled. Cancerous cells grow without control and can go to form tumors. Such development of single cells into complex organisms is in itself an area of study called biology development.

## 2.3 Molecules of life

There are four basic types of molecules involved in life:

1) Small molecules.
2) DNA.
3) RNA.
4) Proteins.

DNA, RNA, Proteins are known collectively as biological macromolecules.

### 2.3.1 Small molecules

Small molecules are the building blocks of the macromolecules or they play independent roles, such as single transmission, or being a source of energy, or material for a cell. Some important examples besides water are sugars, fatty acid, amino acids and nucleotides. For instance, biological membrane is constructed from fatty acids, into which macromolecules are embedded. There are 20 different amino acids molecules, which are building blocks for proteins. They differ by R side chains which determine their properties. The order of these different amino acids within the

protein determines the three dimensional structure of the protein. There is a convention that each amino-acid is denoted by a letter in Latin alphabet, for instance Arginine is denoted by R, Histidine by H, Lysine by L and there are 20 such letters.

### 2.3.2 DNA

**Deoxyribonucleic** (DNA) is the main information carrier molecule in the cell. Also it is the basis for the building blocks encoding the information of life in a single or double stranded. A single stranded called a polynucleotide (as shown in Fig.2.3) is a chain of small molecules, called nucleotides. There are four different nucleotides grouped into two types, purines: adenosine (A) and guanine (G) and pyrimidines: cytosine (C) and thymine (T). They are usually referred to as bases (in fact bases are the only distinguishing element between different nucleotides), and denoted by their initial letters, A, C, G and T. However, the ends of the polynucleotide are different, meaning that each polynucleotide sequence will have directionality, the ends of the polynucleotide are marked either 3' or 5'. The general convention is to label the coding strand from 5' to 3' (left to right) [30].

5′ G–T–A–A–A–G–T–C–C–C–G–T–T–A–G–C 3′

Fig.2.3 a single stranded DNA "polynucleotide

DNA can be double stranded. When DNA is double-stranded, the second strand is referred to as the reverse complement strand. This name is derived from the fact that the directionality of this second strand runs in the opposite direction of the first, and the bases in the second strand are complementary to the bases in the first. Complementary bases are determined by which pairs of nucleotides can form bonds between them. In the case of DNA, A binds to T and C binds to G. For the polynucleotide given above, the double-stranded polynucleotide is as shown in Fig.2.4. Hydrogen bonding between functional groups on the bases is the cause of forming the double strands.

5′ G–T–A–A–A–G–T–C–C–C–G–T–T–A–G–C 3′

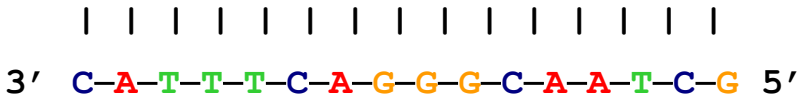


Fig.2.4 double stranded DNA and its nucleotides

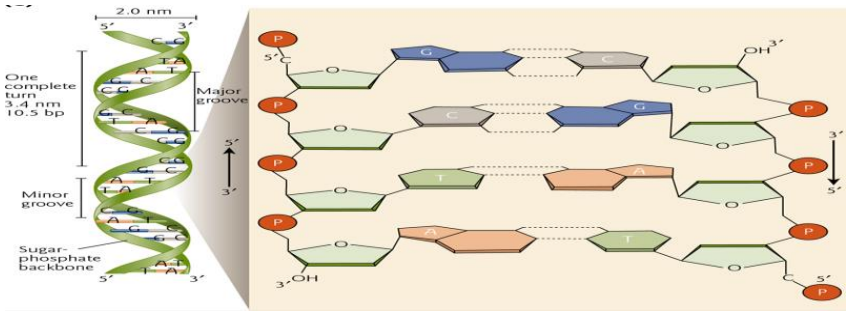


Fig.2.5 Double helix DNA and its hydrogen bonds [DNA structure]

The orientation of the bases is stacked allowing the rotation of the DNA helix [31]. Due to base stacking, DNA completes a turn every 10.5 bp forming a major and minor grove as shown in Fig.2.5[2].

### 2.3.3 RNA

**Ribonucleic Acid (RNA)** is similar to DNA in the fact that it is constructed from nucleotides. However, instead of thymine (T), an alternative base uracil (U) is found in RNA. It can also be a part of a hybrid helix where one strand is an RNA strand and the other is a DNA strand. RNA is generally found as a single stranded molecule that may form a secondary structure (Fig.2.6)[3] or tertiary structures due to the complementary bases between parts of the same strand. One of the most important roles of RNA is the protein synthesis. It is divided into three types: messenger RNA, transfer RNA, and ribosome RNA. Two of the major RNA molecules involved in protein synthesis are messenger RNA (mRNA), and transfer RNA (tRNA) [32].

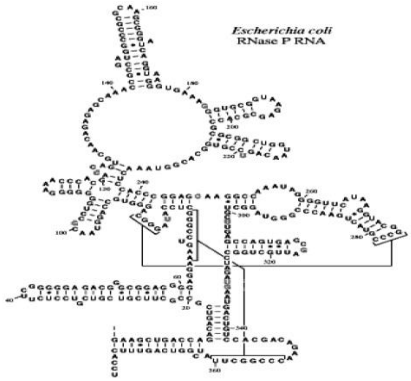


Fig.2.6 Secondary structure for E. coli RNA [Wiki/RNA]

**a) mRNA**

mRNA encodes the genetic information copied from the DNA molecules. The *transcription* is the process in which DNA is copied into RNA molecule. The resulting linear molecule is mRNA transcript as shown in Fig.2.7[3]. In eukaryotic cells, before mRNA is translated into a protein, it needs to be modified. The nature of most eukaryotic genes is that the genes are created in pieces, where coding regions, called *exons*, are interspersed with non-coding regions, called *introns*.

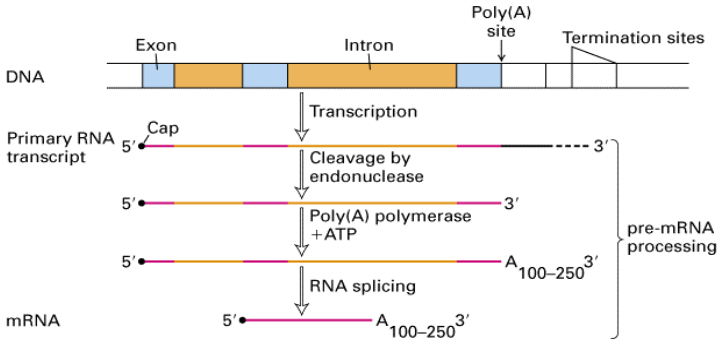


Fig.2.7 mRNA processing [Wiki/RNA]

**b) tRNA**

tRNA molecules develop a well-defined three-dimensional structure (Fig.2.8[3]) which is critical in the creation of proteins. Attached to each tRNA molecule is an amino acid (which will be discussed momentarily). The amino acid is determined by a three base sequence called an anti-codon sequence, which is complementary to the

sequence in the mRNA. *Translation* is the process of converting the ribosomes and tRNA into protein.

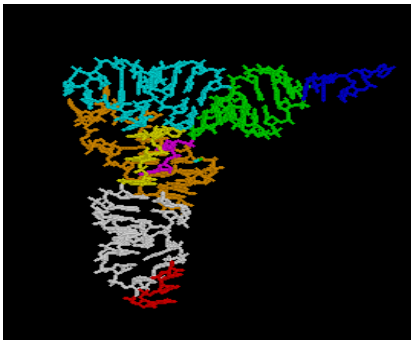

Fig.2.8 tRNA secondary structure [Wiki/RNA]

### c) rRNA

Finally, Ribosomal RNA is a part of the ribosome which is involved in translation.

## 2.3.4 Genes and Protein synthesis

### a) Genetic Code

Since, there are 4 possible bases (A, C, G, U) and 3 bases in the codon, there are 4 * 4 * 4 = ($4^3$) = 64 possible codon sequences (as shown in Fig.2.9)[4]. The codon AUG is used as a signal to initiate the translation process, while the codons UAA, UAG, and UGA are terminal codons signaling the end of translation. The amino acids are coded by the 61 codon sequences.

| First Position | Second Position of Codon: U | C | A | G | Third Position |
|---|---|---|---|---|---|
| U | UUU Phe [F] | UCU Ser [S] | UAU Tyr [Y] | UGU Cys [C] | U |
| | UUC Phe [F] | UCC Ser [S] | UAC Tyr [Y] | UGC Cys [C] | C |
| | UUA Leu [L] | UCA Ser [S] | UAA *STOP* | UGA *STOP* | A |
| | UUG Leu [L] | UCG Ser [S] | UAG *STOP* | UGG Trp [W] | G |
| C | CUU Leu [L] | CCU Pro [P] | CAU His [H] | CGU Arg [R] | U |
| | CUC Leu [L] | CCC Pro [P] | CAC His [H] | CGC Arg [R] | C |
| | CUA Leu [L] | CCA Pro [P] | CAA Gln [Q] | CGA Arg [R] | A |
| | CUG Leu [L] | CCG Pro [P] | CAG Gln [Q] | CGG Arg [R] | G |
| A | AUU Ile [I] | ACU Thr [T] | AAU Asn [N] | AGU Ser [S] | U |
| | AUC Ile [I] | ACC Thr [T] | AAC Asn [N] | AGC Ser [S] | C |
| | AUA Ile [I] | ACA Thr [T] | AAA Lys [K] | AGA Arg [R] | A |
| | AUG Met [M] | ACG Thr [T] | AAG Lys [K] | AGG Arg [R] | G |
| G | GUU Val [V] | GCU Ala [A] | GAU Asp [D] | GGU Gly [G] | U |
| | GUC Val [V] | GCC Ala [A] | GAC Asp [D] | GGC Gly [G] | C |
| | GUA Val [V] | GCA Ala [A] | GAA Glu [E] | GGA Gly [G] | A |
| | GUG Val [V] | GCG Ala [A] | GAG Glu [E] | GGG Gly [G] | G |

Fig.2.9 64 amino acids codons [tutorials/AAs]

### b) Amino Acid

Amino acid is a molecule that consists of amino group (NH2), carboxyl group (COOH), and R group (side chain) which determine the type of amino acid (Fig. 2.10). It is considered as the building block from which protein is made. There are 20 different amino acids that vary relating to their side chain groups (Fig.2.11)[5]. Amino acids are classified into different groups based on their solubility in water; *Hydrophilic* amino acids are water soluble, while *hydrophobic* are not.

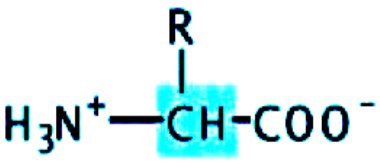


Fig.2.10 basic structure of Amino Acid

This property becomes important when a protein sequence is made. Amino acids are linked to one another via a single chemical bond called a *peptide bond* [33, 34].

As shown in the following Figure, there are 20 different amino acids. They are divided into: neutral non-polar (9 a.a), neutral polar (6 a.a), acidic (2 a.a), and basic (3 a.a).

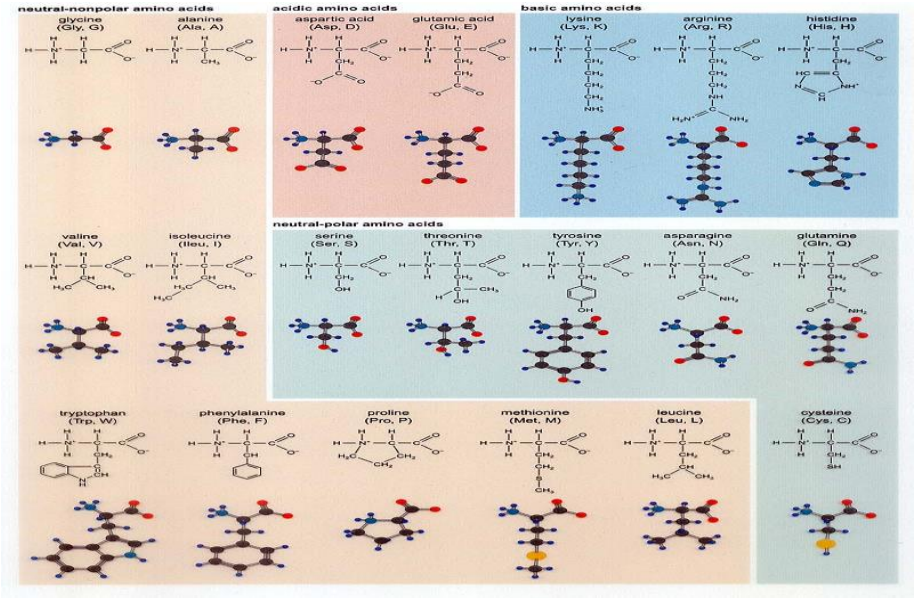


Fig.2.11 the different types of AAs [biochemistry/AA]

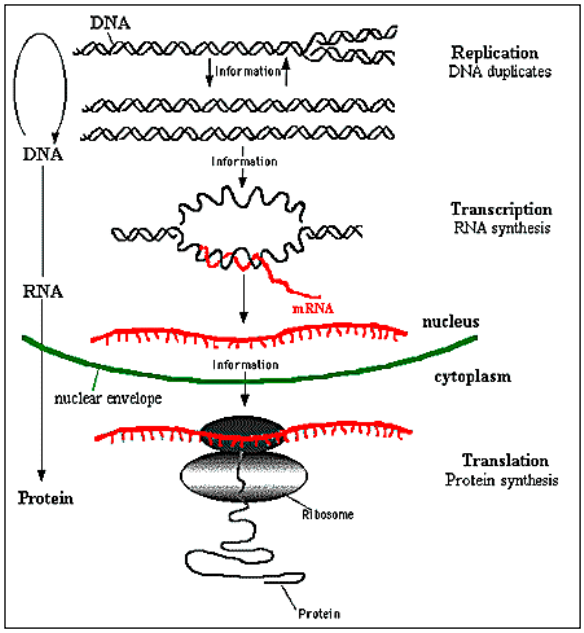


Fig.2.12 central dogma of molecular biology [articles/protein]

### c) Protein

Protein is a Greek word "Pro Teios" which means holding first place. Proteins are the fundamental components of all living cells. They perform variety of biological tasks as controlling physioco-chemical conditions inside the cell, and transmitting biological signals. They have high molecular weights reach to millions and they consist of sequences of amino acids. Although DNA is kept in nucleus, protein synthesis happens in cytoplasm. The central dogma of molecular biology is shown in Fig.2.12[6] which indicate the transcription and translation processes [35].

Since protein consists of groups of amino acids which are responsible for its function. Any mutation or exchange in this sequence will change the shape and cause the dieses. Each part of sequence has its shape which combines with other sequences leading to the folding shape (final structure of protein). Each protein has its own folding shape which may change by time or for making another function. Two or more proteins can be combined to specify certain function. The folding is an identification vector or signature for each protein. Proteins fold into one or more specific spatial conformations driven by a number of non covalent interactions such as:

- Hydrogen bonds
- Ionic interactions
- Van der Waals' forces
- Hydrophobic packing

To identify the protein function, 3D structure of protein should be determined which can be collected by using such techniques as X-ray crystallography, and NMR spectroscopy. There are four types of protein structures as shown in Fig.2.13[6]: *Primary structure*; sequence of the amino acids which is cross linked, *secondary structure*; highly regular sub-structures (*alpha helix* & *strands of beta sheet*), *Tertiary*

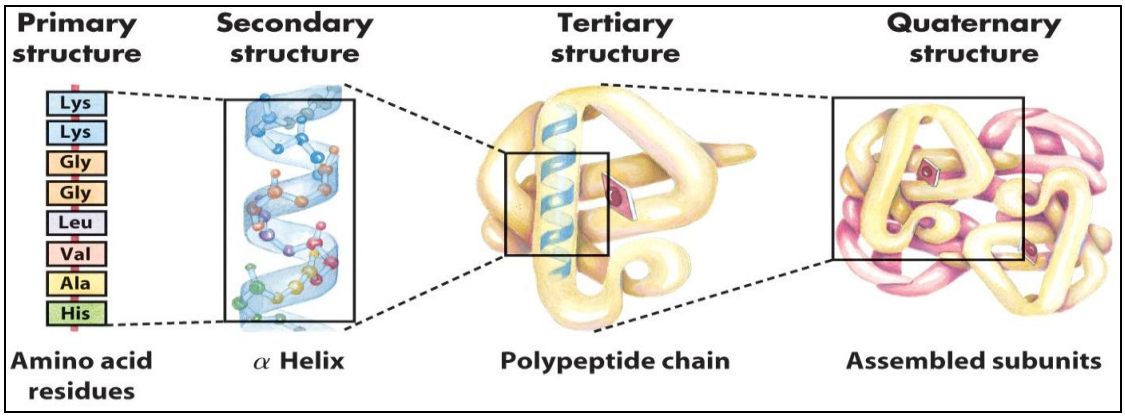


Fig.2.13 the different structure of proteins [articles/protein]

*structure*; the three-dimensional structure of a single protein molecule which is spatial arrangement of the secondary structures, and *Quaternary structure*; complex of several protein molecules or polypeptide chains, usually called protein subunits which function is a part of the larger assembly or protein complex.

## 2.4 Biological databases

Biological databases are libraries of life science information collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics.

Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures [36]. Fig.2.14 introduces different formats for DNA and proteins.

- **Primary sequences databases**

  All public DNA sequences are stored in the EMBL database (also known as EMBL –Bank), which is in fact a collaboration of three databases EMBL in Europe, Gen Bank in the USA and DDBJ in Japan (each database mirrors the others and they exchange data every 24 hours).


Fig.2.14 Different formats DNA and protein sequences

- **Protein sequence databases**

  1. UniProt: Universal Protein Resource (UniProt Consortium: EBI, Expasy, PIR)

  2. PIR Protein information Resource (Georgetown University Medical Center (GUMC).

  3. Swiss-Prot: Protein knowledgebase (Swiss Institute of Bioinformatics). Its format is very similar to EMBL format, except considerably more information about the physical and biochemical properties of the protein is provided.

- **Protein Structure Databases:**

  1. Protein Data Bank (PDB) (Research Collaborator for Structure Bioinformatics (RCSB))
  2. CATH Protein Structure Classification.
  3. SCOP Structural Classification of Proteins
  4. Swiss-MODEL Server and Repository for Protein Structure Models.
  5. Mod Base Database of Comparative Protein Structure Models (Sali Lab, UCSF).

- **Protein-Protein Interactions**

  1. Bio-GRID A General Repository for Interaction Database (Samuel Lunenfeld Research Institute).
  2. STRING: STRING is a database of known and predicted protein –protein interactions. (EMBL).
  3. DIP Database of Interacting Proteins.

## 2.5 Summary

In this chapter, an introduction to the proteomics was presented including Historical development, the need for Proteomics, proteomics impact on heath life, and proteomics research areas. Further more, a comparison between the different organisms (single and multi cell) was introduced. Also the different types of molecules of life were discussed especially the protein which was presented in details having its different types of structure. Finally the most common databases of proteins were described.

# Chapter 3

# Protein-Protein Interaction Networks

There is no doubt that the analysis of protein–protein interactions is one of the most important principals of proteomics that enable understanding the cellular organization, processes, and functions. Proteins seldom act as single isolated species; they often interact with each other in the same cellular processes, while functions of uncharacterized proteins (un-known) is predicted through comparison with the interactions of similar known proteins.

In this chapter, the details of protein-protein interactions are discussed. Section 1 introduces the definition of PPI followed by their challenges (section 2). In addition, large-scale experimental methods of protein–protein interactions (section 3) that use techniques as two-hybrid systems, mass spectrometry, and protein microarrays, have enriched the available protein interaction data and facilitated the construction of integrated protein–protein interaction networks . In section 4, graphical and mathematical models are introduced to illustrate the proteins and their connections. Prediction methods are proposed in section 5. Finally, description for the available databases and repositories of protein–protein interactions has introduced in section 6.

## 3.1 Protein-Protein Interaction Network

A Protein-Protein interaction network can be described as a complex system of proteins linked by interactions [13] as shown in Fig.3.1. Protein-protein interactions play an important role in the field of proteomics. They regulate a wide array of biological processes, including transcriptional activation/ repression immune, endocrine, pharmacological signaling, cell-to-cell interactions, and metabolic and developmental control [37-40].

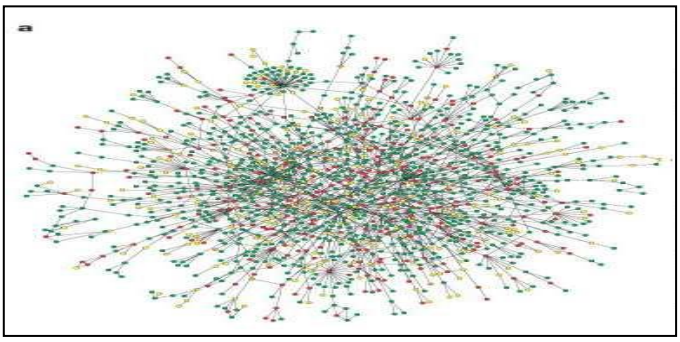

Fig.3.1 Example of protein-protein interaction network [cytoscape]

Most proteins within cells do not exist in isolation, but interact with other proteins either directly or indirectly, via mechanisms that ultimately involve some forms of binding. Most proteins in a cell function cooperatively in highly ordered networks, in order to let cellular metabolism run smoothly. They may interact with a cell's nucleic acids, in which the function may regulate gene transcription or control DNA replication. Another example is the binding of small non protein (ligand) as seen in the case of enzyme binding to prevent catalytic activity.

An enzyme's catalytic function may also be modified by binding a highly reactive metal ion to the active site, such as electron transfer reactions. Finally, proteins may interact with other proteins; these interactions may provide a variety of functions such as catalytic, structural, localization, cleavage, transferring, or inhibitory functions. Protein localization occurs to position of protein in a specific cellular location, this allows the cellular contents to remain in their highly ordered compartments, and prevents any mixing that may produce undesirable potentials [41]. PPIs play different roles in biology based on the composition, affinity, and lifetime of the complex association. The basics of protein folding, protein assembly and PPI are the non-covalent contacts between residue side chains [42]. These contacts facilitate a variety of interactions and associations within and between proteins. Based on their diverse structural and functional characteristics, PPIs are categorized in three categorical ways [43]. Regarding their interaction surface, they are homo - or hetero [44] oligomeric, while judged by their stability, they are obligate or non-obligate [45] and for as their persistence, they are transient or permanent [46]. A given PPI can fall into any combination of these categorical pairs. An interaction also requires

reclassification under certain conditions; for example, transient in vivo or become permanent under certain cellular conditions.

It is observed that in the same cellular processes, proteins often interact with each other regarding the analysis of annotated proteins [47]. Recently, PPI is one of the most reliable methods used to predict the protein functions through the known proteins. PPI is not used to predict the protein functions only but also to facilitate the modeling of pathways to indicate the molecular mechanisms of cellular processes. For understanding the biochemistry of the cell, the interactions inside its proteome should be characterized.

PPIs are much more wide spread than once suspected, and the degree of regulation in the cell that they confer is large. It is important to identify the different interactions, understand the extent to which they take place in the cell, and determine the consequences of the interactions to understand the degree of significance in the cell.

## 3.2 Challenges of PPI

In general, PPI networks building is challenging and it has some major difficulties:

1. The reliability of protein interactions
    - Large-scale experiments have yielded numerous false positives [48], high throughput yeast two-hybrid (Y2H) assays are ~50% reliable.
    - It is also likely that there are many false negatives in the PPI networks and are currently under study.

2. A protein has several different functions up to eight functions in Yeast. A protein is imbedded in more than one functional group. This means that overlapping clusters should be identified in the PPI networks. Also it is recommended not apply the conventional clustering method for their pair-wise disjoint clusters.

3. It is known that two proteins with different functions can interact with each other. Such observations, random connections between the proteins in different

functional groups expand the topological complexity of the PPI networks, showing difficulties when detecting unambiguous partitions.

4. Many biological data do not provide complete information because the nature and limitations of the experiments used to derive them.

5. Many useful biological databases contain overlapping or complementary information on the same proteins. The mapping between genes and names is many-to-many. Multiple names may refer to the same genes and multiple genes may also be referred to by the same name. Each of these databases may refer to the same protein using different names. For example, the yeast gene product GIP4, is identified by an EMBL accession number (U12980) in EMBL-Bank, a RefSeq accession number (NP_009371) in GenBank, an UniProt ID is (P39732) in UniProt, a systematic name is (YAL031C) in CYGD, and an SGD ID is (S000000029) in SGD.

For the above mentioned difficulties of PPIs, many studies attempt to characterize and understand the behaviors of network interactions [49, 50]. This study takes the features of PPIs as small-world properties [51], while others take the scale-free degree [52, 53] or hierarchical modularity [51].

## 3.3 PPI experimental methods

Consequently, an examination of protein–protein interactions (PPIs) is essential to understand the molecular mechanisms of underlying biological processes [54]. This section intends to provide an overview of the more common experimental methods currently used to generate PPI data. Although experimental methods are rather expensive and find out a small number of interactions which are specifically targeted, they are very important to determine the protein interactions. Recently, high-throughput approaches involve genome-wide detection of protein interactions. Other studies that use the yeast two-hybrid (Y2H), which is the most widely used method to study protein–protein interactions system [55- 57], mass spectrometry (MS) [58-63], and protein microarrays [64, 65] generates large amounts of interaction data.

### 3.3.1 Yeast two hybrid system

Y2H system is one of the most common approaches that detects the pairs of interacting proteins in vivo [66,67] which takes (a bottom up) genomic approach for detecting possible binary interactions between any two proteins encoded in the genome of interest.

It has been introduced in 1989 [68]. It is a molecular–genetic tool that facilitates the study of PPI. Since the interaction of the two proteins activates a reporter gene; a color reaction is seen on specific media. This indication tracks the interaction between two proteins, revealing "prey" proteins that interact with a known "bait" protein.

Two-hybrid procedures are typically carried out by screening a protein of interest against a random library of potential protein partners. Fig.3.2 depicts the Y2H process [57, 69].

In Fig.3.2(a), the fusion of the "bait" protein and the DNA-binding domain of the transcriptional activator does not turn on the reporter gene; no color change occurs; and the interaction cannot be tracked. Fig.3.2(b), the fusion of the "prey" protein and the activating region of the transcriptional activator is sufficient to switch on the reporter gene. In Fig.3.2(c), the "bait" and the "prey" associate, bringing the DNA-binding domain and activator region into sufficiently close proximity to switch on the reporter gene.

Although in vivo, the Y2H system enables both highly sensitive detection of PPIs and screening, indicates physical interactions, and good for pair wise and transient interactions [66], it has several limitations. The most common limitations are: 1) it cannot, by definition, detect interactions involving three or more proteins and those depending on posttranslational modifications (PTMs) except those applied to the budding yeast itself [66]. 2) since some proteins (e.g., membrane proteins) cannot be reconstructed in the nucleus, the Y2H system is not suitable for the detection of interactions involving these proteins [70, 71]. 3) The method does not guarantee that interaction.

### 3.3.2 Mass Spectrometry

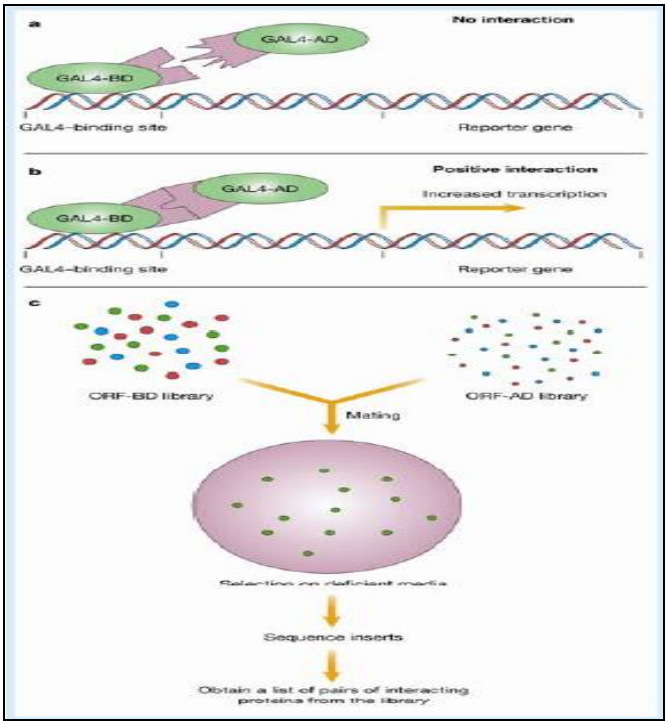


Fig.3.2 example of Y2H interaction [cmbi.bjmu.cn]

Quantitative MS is a method used to analyze the composition of a partially purified protein complex together with a control purification in which the complex of interest is not enriched. Mass spectrometry analysis has been performed in three steps:

1. Bait presentation
2. Affinity purification of the complex,
3. Analysis of the bound proteins [70].

MS analysis is applied on the PPI network in yeast [72, 73]. Mass-spectrometry-based proteomics which adopts a top-down proteomic approach by analyzing the composition of protein complexes, is applied not only to identify and quantify individual proteins [72-75] but also to identify and quantify protein analysis, including protein profiling [76], PTMs [77, 78], and in particular, identification of PPIs.

In general, mass spectrometric analysis is more physiological than the Y2H system. Actual molecular assemblies composed of all combinations of direct and cooperative interactions are analyzed in vivo, as opposed to the examination of reconstituted bimolecular interactions vivo or in vitro. MS detects more complex interactions and is not limited to binary interactions, permitting the isolation of large protein complexes and the detection of networks of interactions.

However, the technique is best applied on interactions of high abundance and stability, while two-hybrid approaches are able to reliably detect transient and weak interactions.

### 3.3.3 Microarray

Microarray-based analysis is a relatively high-throughput technology that allows the simultaneous analysis of thousands of parameters within a single experiment. The key advantage of the microarray format is the use of a nonporous solid surface, such as glass, that permits precise deposition of capturing molecules (probes) in a highly dense and ordered fashion. The early applications of microarrays and detection technologies were largely centered on DNA-based applications. Today, DNA microarray technology is a robust and reliable method for the analysis of gene function [79]. However, gene expression arrays provide no information on protein. PTMs (such as phosphorylation or glycosylation) that affect cell function. Toexamine expression at the protein level and acquire quantitative and qualitative information about proteins of interest, the protein microarray was developed. A protein microarray is a piece of glass on which various molecules of protein have been affixed at separate locations in an ordered manner, forming a microscopic array [80]. These are used to identify PPIs, the substrates of protein kinases, or the targets of biologically active small molecules. The experimental procedure for protein microarray analysis involves choosing solid supports, arraying proteins on the solid supports, and screening for PPIs. Experiments with the yeast proteome microarray reveal a number of PPIs that are not previously identified through Y2H or MS-based approaches. Global protein interaction studies are performed with a yeast proteome chip. Ge et al [66] describes a universal protein array that permits quantitative detection of protein interactions with a range of proteins, nucleic acids, and small molecules. Zhu et al [65] generate a yeast

proteome chip from recombinant protein probes of 5,800 open-reading frames, in contrast, mass spectrometric Protein-protein interactions network. As shown in Fig.3.3, example of micro array plate.

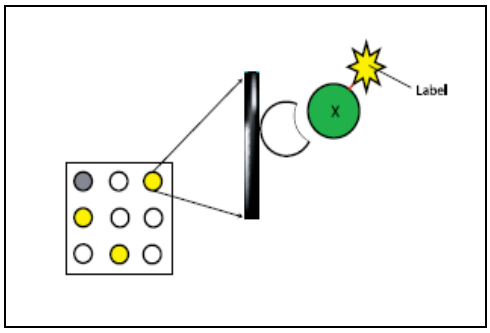


Fig.3.3 example of microarray [dnassequencing]

### 3.3.4 TAP method of complex purification

A TAP tag consists of two IgG binding domains of Staphylococcus protein A and a calmodulin binding peptide separated by the tobacco etch virus protease cleavage site [81, 82] as shown in Fig.3.4.

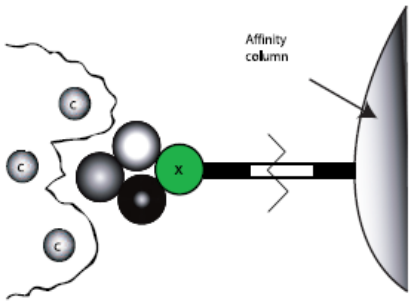


Fig.3.4 example of Tap-Affinity experimental method [embl.ed]

A target protein open reading frame (ORF) is fused with the DNA sequences encoding the TAP tag and is expressed in yeast, where it forms native complexes with other proteins. At the first step of the TAP purification, protein A binds tightly to an IgG matrix; and after washing out the contaminants, the protease cleaves the link between protein A and IgG matrix. The eluate of this first step is then incubated with calmodulin-coated beads in the presence of calcium. After washing, the target protein

complex is released. The components of each complex are screened by poly-crylamide gel electrophoresis, cleaved by proteases, and the fragments are identified by MS. Comparing Y2H and TAP–MS, it is noted that both methods generate a lot of false positives and miss a lot of known interactions. TAP–MS reports on higher order interactions beyond binary and, therefore, provides direct information on protein complexes.

Several large-scale studies of protein complexes are performed using TAP–MS and Y2H methods [60, 83-85]. For example, Krogan et al. showed that 7,123 protein interactions identified with high confidence in yeast can be clustered into 547 protein complexes [86].

### 3.3.5 Gene co-expression

Since the function of a protein complex depends on the functionality of all subunits, subunits present in stoichiometric amounts and gene expression levels of subunits in a complex are related.

Gene expression profiles are provided, for example, from cell cycle experiments and expression levels of a gene under different conditions.

Expression profile similarity is calculated as a correlation coefficient between relative expression levels of two genes/proteins or the normalized difference between their absolute expression levels or calculated using other methods [87-91] (Fig.3.5). The distribution of these quantities for target proteins then is compared with the distributions for random non interacting protein pairs. It is shown that the most obvious co-expression comes from permanent complexes such as ribosome and proteasome [87]. Several studies tackle the problem of gene co-expression and demonstrate that interacting proteins in yeast are more likely to have their genes co-expressed compared with non interacting proteins [87, 92–99].

Moreover, it is shown that expression levels of physically interacting proteins coevolve, and coevolution of gene expression is a better predictor of protein interactions than coevolution of amino acid sequences [100].

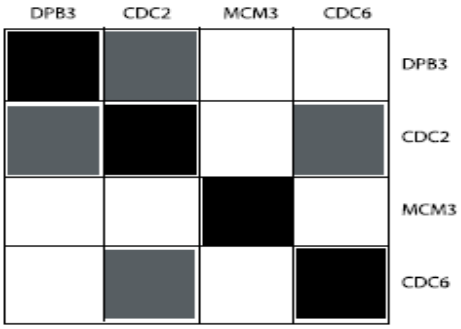


Fig.3.5 Example of gene co-expression [journal.ppat]

### 3.3.6 Synthetic lethality method

The synthetic lethality method is successfully used with the DNA microarray methodology. The synthetic lethality method produces mutations or deletions of two separate genes which are viable alone but cause lethality when combined together in a cell under certain conditions [100–109]. Since these mutations are lethal, they are not isolated directly and should be synthetically constructed. Synthetic interaction can point to the possible physical interaction between two gene products, their participation in a single pathway, or a similar function.

Synthetic lethality experiments are used in predicting the unknown function of the proteins (monitoring specific protein interactions). The most detailed information about protein interaction interfaces at the atomic level is provided by X-ray crystallography and NMR spectroscopy, but the number of solved protein complexes remains low [96]. At the same time, the real-time characterization of interacting proteins in vivo is achieved with various spectroscopic techniques requiring the attachment of a spectroscopic label to a target protein [98, 99]. As shown in Fig.3.9, different cases for the identification of the interaction are introduced.

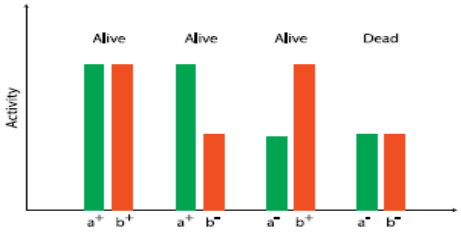


Fig.3.6 example of synthetic lethality experiment [journal.ppat]

## 3.4 Mathematical and graphical models for PPIs

This section introduces the concept of mathematical model relating to the PPI and basic properties and metrics applied to PPI networks. It also indicates the basic concepts and measurements in graphic representation employed to characterize various properties of PPI networks.

### 3.4.1 Presentation of PPI network

The computational methods of PPI network mechanisms begin with a representation of the interactions network structure. As mentioned earlier, the simplest representation takes the form of a mathematical graph consisting of nodes and edges [12, 13]. Proteins are represented as nodes as shown in Fig.3.7 and two proteins that interact physically are represented as adjacent nodes connected by an edge.

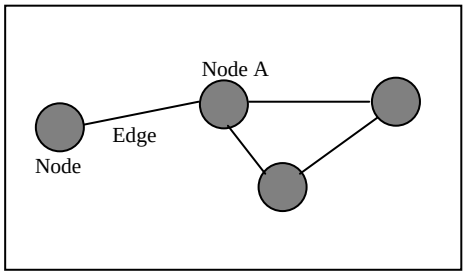


Fig.3.7 proteins as nodes and the interactions as edges

Proteins interact with each other to perform a specific cellular function or process. These interacting patterns form a PPI network that is represented by a graph $G = (V, E)$ with a set of nodes $V$ and a set of edges $E$.

$$V \times V = \{(v_i, v_j) \mid v_i \in V, v_j \in V, i \neq j\}. \quad (3.1)$$

An edge $(vi, vj) \in E$ connects two nodes $vi$ and $vj$. The vertex set and edge set of a graph are denoted by $V(G)$ and $E(G)$, respectively. Graphs can be directed in path ways or in enzymes network or undirected as in the functions relations. In directed graphs, each directed edge has its source and a destination vertex (target). However,

undirected graphs, the order of the incident vertices of an edge is immaterial. Also graphs can be weighted or un-weighted.

### 3.4.2 PPI network concepts

A number of fundamental concepts of these graphical representations are introduced to understand the used techniques.

#### a) Degree

The degree (or connectivity) of a node is the number of surrounding nodes which have direct connections in an undirected graph [110]. For example, in the undirected network graphed in Fig.3.8, node A has degree $k = 6$.

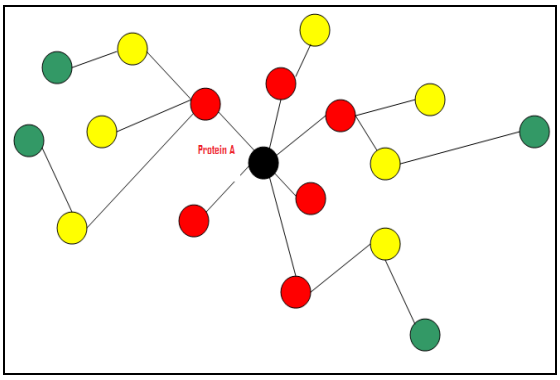


Fig.3.8 the number of red nodes is the degree of node A

Let *N(vi)* denote the neighbors of node *vi*; that is the set of nodes connected to *vi*. The degree *d(vi)*of *vi* is then equivalent to the number of neighbors of *vi*, or |*N(vi)*|. In directed graphs, the edges will be denoted by d+ or d- relating out or in respectively. The summation of corresponding edge weights is used in the weighted graphs.

#### b) Paths and walk

In this sub-section the difference between the path and walk is discussed indicating the different types of paths. Many relationships within a graph is showed by means of conceptual “walks” and “paths.” A walk is defined as a sequence of

nodes in which each node is linked to its sequencing node. While a path is a walk in which each node in the walk is distinct.

In the path that starts from *vi (source)*, passes through *vk*, and ends with *vj(target)* , path (*vi*, *vk*, *vj*). All paths starting with source node *vi* and end by target node *vj* are denoted by *P(vi*, *vj)*. The number of edges in the sequence of the path acts the length of the path. The minimal-length path connecting two nodes is the shortest path between them. *SP(vi*, *vj)* denotes the set of the distinct shortest paths between *vi* and *vj* . The distance between these nodes is the length of the shortest path between them and is denoted by dist*(vi*, *vj)*. A graph *G*' = *(V*', *E*'*)* is a sub-graph of the graph *G* = *(V*, *E)* if $V' \subseteq V$ and $E' \subseteq E$. A *vertex-induced* sub-graph is a vertex subset *V'* of a graph *G* together with any edges in edge subset *E*' whose end points are both in *V*'.

### c) PPI networks properties

The most famous characteristics of PPI networks are Small-world and scale free distribution.

- **Small-World Property**

PPI networks are highly dynamic and structurally complex. They are characterized by the inherent properties of complex systems [49,111,112]. PPI networks indicate the property of small-world networks which means that the average shortest-path length between any two nodes in a network is relatively small.

In small-word networks, all nodes can be reached quickly from any node via a few hops to its adjacent neighbors. It has found that the sub-networks in the middle of either a regular network or a random network are highly clustered and have short average path lengths between nodes [113].

The procedure for random reconnection of a regular graph is illustrated in Fig.3.9. The procedure starts with a regular ring graph with 20 nodes and four directly connected neighbors for each node. After selecting node and its connected edge, reconnection of that edge random with probability p to another node. By repeating this process, a disordered random graph is obtained for $p = 1$. For the value of $p$

between 0 and 1, the graph becomes a small-world network. Like a regular graph, it is highly clustered, but it has short path lengths like a random graph.

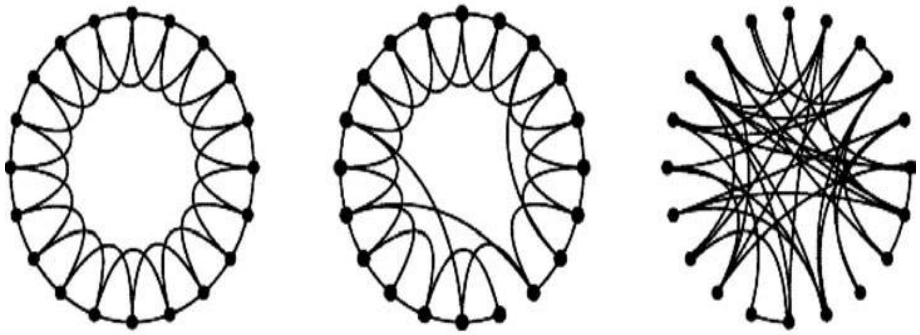

Fig.3.9 random reconnection procedure of a regular ring graph [Reprinted by permission from Macmillan Publishers Ltd:[113]]

The small-world network with high clustering coefficients and short path lengths is detected when *p* is around 0.01. One of the most famous networks for this phenomenon is Yeast PPI networks. The average shortest path length and average clustering coefficient for these networks extract from the DIP [114] and MIPS [115] databases. Although both networks are large and very sparse, with more than 5,500 nodes, the average value of the shortest path lengths between all possible node pairs is very small, at ~4.

- **Scale-Free distribution**

As another special property of PPI networks is their scale free distribution [110]. The degree distribution refers to the probability that a given node is of degree *k*, is approximated by a power law $P(k) \sim k^{-Y}$. A scale free network has a few high-degree hub nodes, while most nodes have only a few connections. The structure and dynamics of these networks are independent of the network size as measured by the number of nodes in the network. Growth and preferential attachment are the two important features of scale-free networks [52]. The growth property means that networks are continuously expanded by the addition of new nodes connecting to the presented nodes. As a preferential attachment, the new nodes are likely linked to high-

degree nodes. Since the topological structure is characterized by a few ultrahigh-degree nodes and abundant low-degree nodes, scale-free networks are robust to random attacks [116]. Scale-free networks do not possess an inherent modularity, so the average clustering coefficient is somewhat independent [110]. As shown in Fig.3.10 representation of a scale-free network. Relating to the study [117], the scale-free distributions in yeast PPI networks is examined and got $\gamma$ values in the power-law degree distributions as 1.77 and 1.64 in DIP and MIP respectively.

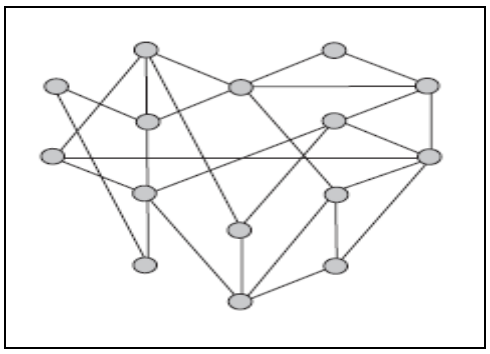

Fig.3.10 A scale free distribution [Reprinted by permission from Macmillan Publishers Ltd:[110]]

- **Modular network**

The discussed properties suggest two important topological issues in the analysis of PPI networks: modularity and the presence of hubs. A module in a PPI network is a region with dense internal connections and sparse external interconnections to other regions. Assuming that a PPI network is composed of a collection of modules, it categorizes nodes in the network as *modular nodes*, *peripheral nodes*, and *interconnecting* nodes. Modular nodes are the core of a module. They have a relatively high connectivity to members of the same module. Peripheral nodes are trivial nodes with a low degree of connectivity. They are linked to modular nodes or to the other peripheral nodes in the same module. The connected nodes between two modules are interconnecting nodes. The edge that connects two nodes in different modules is defined as a *bridge*. As shown in Fig.3.11, example of modular networks. (a) Five dark gray nodes represent *interconnecting nodes*. Light

gray and white nodes are *modular nodes* and *peripheral nodes*, respectively. Three thick edges are *bridges* connecting two modules. (b) A black node represents *bridging nodes*. Three dark gray nodes are *interconnecting nodes*, and three thick edges are *bridges* connecting from the *bridging node* to each module.

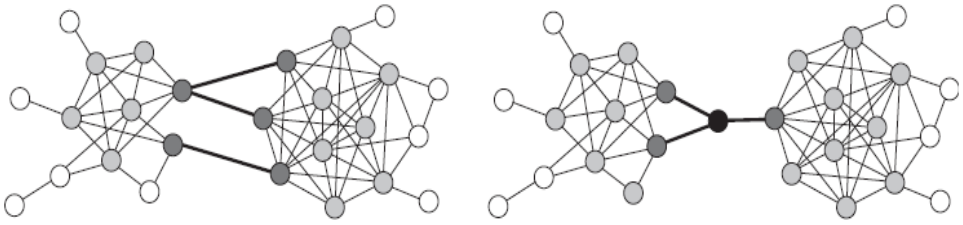

Fig.3.11 Examples of modular networks composed of two modules [Reprinted by permission from Macmillan Publishers Ltd:[110]]

The architecture of the hierarchical network model is characterized by scale-free topology with embedded modularity [118] as shown in Fig.3.12. In this model, a few hub nodes are emphasized as the determinants of survival during network perturbation and as the backbone of the hierarchical structure. This model suggests that low-degree nodes are connected to form a small module. A core node within the module interconnects not only with the cores of other small modules but also with a higher-degree node, which, in turn, becomes the core of a larger module consisting of a group of the small modules. By repeating these steps, a hierarchy of modules is structured through the hubs. The degree distribution of hierarchical networks is similar to that of scale-free networks, showing locally disordered effects within modules. However, unlike scale-free networks, the pattern of clustering coefficients in hierarchical networks has an inverse relationship to degree [110]. Therefore, low-degree nodes are clustered better than high-degree nodes, since low-degree nodes are interconnected within a module, whereas high-degree nodes are typically interconnected between modules. A schematic view of a hierarchical network, degree distribution, and the average clustering coefficients with respect to degree are illustrated in Fig.3.12. The modular and hierarchical network models can reasonably be applied to PPI networks because cellular functionality is typically envisioned as having a hierarchical structure. Extracting these structures from PPI networks may provide valuable information regarding cellular function.

Based on this graphic representation, various computational approaches, such as data mining, machine learning, and statistical approaches, can be designed to reveal the organization of PPI networks at different levels. An examination of the graphic form of the network can yield a variety of insights. For example, neighboring proteins in the graph are generally considered to share functions.

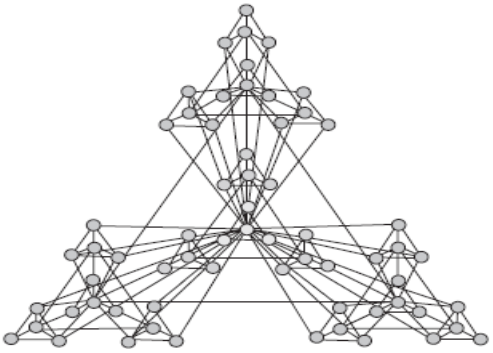

Fig.3.12 The architecture of the hierarchical network model [Reprinted by permission from Macmillan Publishers Ltd:[110]]

Thus, the functions of a protein may be predicted by looking at the proteins with which it interacts and the protein complexes to which it belongs. In addition, densely connected sub-graphs in the network are likely to form protein complexes that function as a unit in a certain biological process. An investigation of the topological features of the network (e.g., whether it is scale-free, a small network, or governed by the power law) can also enhance our understanding of the biological system [54].

## 3.5 PPI prediction

This section presents different ways for predicting the protein-protein interactions relating to the knowledge of the structure.

### 3.5.1 Protein–Protein Interaction Prediction Using Known Structures

The crystallization structure of proteins is the cause of using some prominent methods of protein interaction prediction. A protein's structure can be solved from the synthesized crystals using X-ray diffraction or neutron diffraction analysis using x-ray crystallography. Also the nuclear magnetic resonance (NMR spectra are generated by placing a sample in a magnetic field and applying radiofrequency pulses) is a technique that is generally used for proteins in solution and cannot be crystallized. Fig.3.13 shows the x-ray crystallography device.

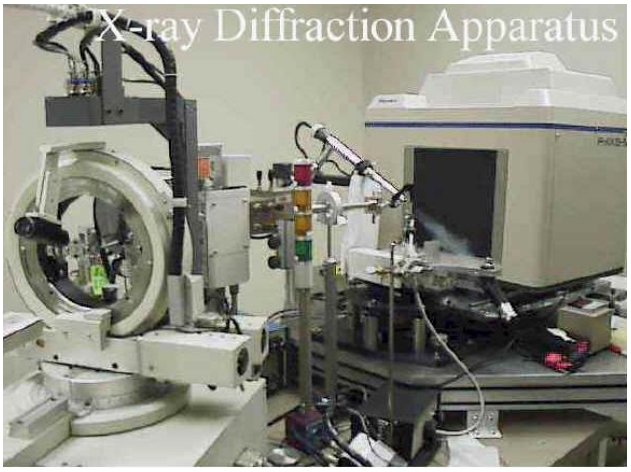


Fig.3.13 X-ray crystallography device [wiki/ protein structure]

### 3.5.2 Prediction of PPI in the Absence of Protein Structures

There are numerous methods also designed to predict interactions without structural data. These methods can be divided into computational methods and others non computational methods [119]. The final forms of the protein interfaces affects on the interaction process. It is possible however, that in some cases most of the protein surface contributes to a protein–protein interaction with one or multiple interaction partners (binding with distinct interfaces on the surface of the same protein. Thus, the problem with defining the rest of the protein surface is this surface may form part of other interfaces, which would invalidate a proper statistical comparison. There are six types of protein interfaces [120]. From an evolutionary perspective, protein–protein

interfaces have evolved over time to optimize the interface to suit their individual biological functions. This function may have required the evolution of specific binding strength. It is also important to distinguish between the different types of complexes, in terms of their type of physical interaction, when analyzing the intermolecular interfaces.

## 3.6 Protein–Protein Interactions Databases

Protein interaction databases are especially useful for generating a collection of known interactions. With the help of computational inference methods, accurate interaction predictions can be made. But computational prediction methods still need a lot of improvement. The Database of Interacting Proteins (DIP) has combined data from a variety of sources to create a single, consistent set of PPI. It contains experimentally determined protein interactions and includes a core subset of interactions that have passed a quality assessment [121].

Interaction data are obtained from the literature; PDB; and high-throughput methods such as Y2H, DNA and protein microarrays; and TAP–MS analysis of protein complexes. DIP has links to a couple of related databases including Live DIP which records information about the state of a biological interaction, such as covalently modified, conformational, or cellular location states [122].

For the yeast PPI data, the core PPIs have been selected from full data by a computational curative process based on the correlation of protein sequence and RNA expression profiles [88]. Another database related to DIP is Prolinks, which brings together four methods of linking proteins: phylogenetic profiles, Rosetta Stone, gene neighbors, and gene clusters [123].

Also there are number of open databases that provide comprehensive PPI data for several different organisms. There is little standardization among these databases, with each having a unique data structure, format, and mode of description. The data have been curated using various computational methods. The major open PPI databases will be briefly described as follows:

- *MIPS*: The Munich Information Center for Protein Sequences (MIPS) [124] is the repository of a significant body of protein information including sequence, structure, expression, and functional annotations. This database also includes PPI data for selected organisms, including *Homo sapiens*. The human PPI data have been manually created on the basis of literature review and include the experimental approach, a description, and the binding regions of interacting partners [125]. MIPS is often used as a standard of truth database for evaluating the quality of data and the accuracy of interaction prediction methods.

- *BIND*: The Biomolecular Interaction Network Database (BIND) [126], a component of BOND (the Biomolecular Object Network Databank), includes interactions, molecular complexes as a collection of two or more molecules that together form a functional unit, and pathways as a collection of two or more molecules that interact in a sequence.

- *BioGRID*: The General Repository for Interaction Database (BioGRID) [127] is a unified and continuously updated source of physical and generic interactions. It comprises more than 55,000 non redundant interactions for yeast, making it the largest database for this organism, and more than 130,000 non redundant interactions across a total of 22 different organisms.

- *MINT*: The Molecular Interaction Database (MINT) [128] uses expert curators to extract various experimental details from published literature; these are then stored in a structured format. Homo-MINT [129] is a separate database of human protein interactions that have been inferred from orthologs in model organisms.

- *IntAct*: IntAct [130] is a database and toolkit for modeling, storing, and analyzing molecular interaction data. In addition to PPI data, it also includes extensive information on DNA, RNA, and small-molecule interactions.

- *HPRD*: The Human Protein Reference Database (HPRD) [131] provides a comprehensive collection of human PPI with protein features such as protein functions, PTMs, enzyme–substrate relationships, and sub-cellular localization. The human PPI data have been obtained from various experimental methods including the Y2H systems.

## 3.7 Summary

In this chapter, the basic structure of protein-protein interactions was introduced that it consisted of nodes and edges. Later on the challenges of these data were introduced and how we could overcome these problems. The most common experimental methods used in determining the protein interactions were discussed as (Y2H, mass spectrometry, microarray). An introduction for the most common mathematical and graphical models was discussed. Finally the different databases and data sources were introduced.

# Chapter 4

# Review of Literature

With the completion of the Human Genome Project (HGP), new challenges lie ahead in deciphering the complex functional and interactive processes between proteins and multi component molecular machines that contribute to the majority of operations in cells, as well as the transcriptional regulatory mechanisms and pathways that control these cellular processes [132].

Getting large amount of biological data from high-throughput processes such as genomic and proteomic sequencing, gene expression profiling, immuno-precipitation, mass spectrometry and more recently, flow cytometry, it is now possible to study the characteristics and interactions of cellular components from a global perspective.

Meanwhile, the maturation of high-throughput techniques for various genome analysis makes available a large quantity and variety of genomic information. These information offer possible avenues to shed light on the functions of proteins which cannot be easily characterized by sequence homology alone by providing complementary information related to the functionality and behavior of proteins. The computational approach, which has been adopted to solve the problems, is to use information gained from physical and genetic interaction maps to predict protein functions. Recently, the researchers introduced different techniques to determine the probability of protein function prediction using the information extracted from PPI.

In this chapter, most of the conventional methods that use protein-protein interactions to predict protein functions relying on the basis that interacting proteins share functions are introduced. These methods are: neighbor counting, Chi-square, Markov random field, Prodistin, Samanta, Support vector machine, and functional flow.

## 4.1 Overview

A PPI network is described as a complex system of proteins linked by interactions. The computational analysis of PPI networks begins with the representation of the PPI network structure. The simplest representation takes the form of a network graph consisting of nodes and edges [12]. Proteins are represented as nodes in the graph and two proteins that interact physically are represented as adjacent nodes connected by an edge [13]. Based on this graphical representation, various computational approaches, such as data mining, machine learning, and statistical approaches are performed to reveal the PPI networks at different levels. In general, the computational analysis of PPI networks is challenge with some major problems.

## 4.2 Neighbor Counting Method

Neighbor counting is a method proposed by Schwikowski et al. in [14] to infer the functions of an un-annotated protein from the PPI. This method finds the neighbor proteins and gets their assigned functions and the frequencies of occurrence of these functions. Then, these functions are arranged in descending order according to their frequencies. The first $k$ functions are considered and assigned to the un-annotated protein. The authors in [15] used this technique with $k$ equals to 3. Although this method exploits the information from the neighbors, it has some drawbacks: 1) it considers the interactions to be of equal weights which is not logic, 2) it does not consider the nature of the function and whether it is dominant or not and 3) it does not provide a confidence level for assigning a function to the protein. Also the method couldn't predict functions of protein found in un-annotated proteins group. As shown in Fig.4.1, four annotated proteins (p1, p2, p3, p5) and their functions ((f8, f9), (f1, f7), (f1, f3), and (f1-f6)) respectively.

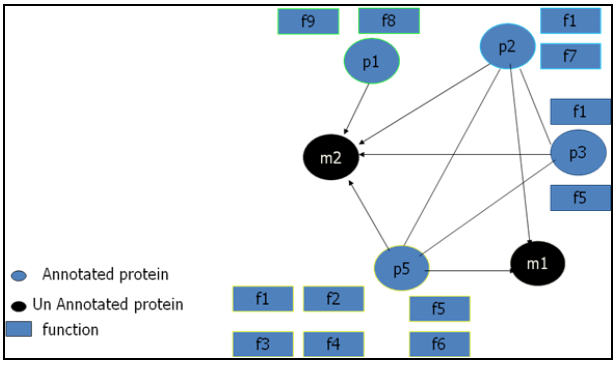


Fig.4.1 annotated and un-annotated proteins and their physical interactions

According to the procedure of the mentioned method (neighbor counting method):

- m1 (un-annotated protein) may have the most frequent functions as:
  - f1 has been found two times from (p2, p5).
  - the rest of functions have been seen only one.
- m2 (un-annotated protein) may have the most frequent functions as:
  - f1 has been found three times from (p2, p3, p5).
  - f5 has been found two times from (p3, p5).
  - The rest of functions have been seen only one.

It is noticed that, the method does not take the strength of the interactions into consideration where the interaction between p2 and m2 may be more reliable than interaction between p3 and m2, it leads to:

m2 may have function f7 instead of f5,

So the reliability of interactions should be considered in the proteome study. As well as, there is not confidence level for each estimated function. By implementing this method on the collected data of yeast proteome network, the results are collected as shown in Table 4.1 which indicates the basic, estimated, and overlapping functions for each protein.

Although some estimated functions seem correct as proteins ID (7, 19, 24, and 33) as shown in Table 4.1, there are some functions estimated as false positive as protein ID (1, 19, 22, 24, and 32) and others can not be predicted.

The leave one out (will be introduced later in the end of this chapter) is applied to this method and the sensitivity and specificity are calculated. The overlapping cells show

the overlapping numbers between the basic and estimated cells as shown in protein ID 7 which has function ID 25, and protein ID 19 which has functions IDs 8, 43.

Table 4.1 basic and estimated functions for Yeast proteome using neighbor counting method

| Protein ID | Basic functions | | | | | | | | Estimated functions | | | | Overlap |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *28* | 35 | 0 | 0 | 0 |
| 2 | 15 | 42 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 3 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 4 | 7 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 5 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 6 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 7 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *25* | 0 | 0 | 0 | 1 |
| 8 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 9 | 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 10 | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 11 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 12 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 13 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 14 | 15 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 15 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 16 | 10 | 13 | 27 | 28 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 17 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 18 | 3 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 19 | 8 | 43 | 0 | 0 | 0 | 0 | 0 | 0 | *8* | 43 | 9 | 17 | 2 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 21 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 22 | 7 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | *22* | 0 | 0 | 0 | 0 |
| 23 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 24 | 16 | 22 | 36 | 0 | 0 | 0 | 0 | 0 | *12* | 18 | 37 | 36 | 1 |
| 25 | 28 | 0 | | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | | 0 |
| 26 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 27 | 3 | 16 | 18 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 29 | 2 | 3 | 15 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 30 | 15 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 31 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *0* | 0 | 0 | 0 | 0 |
| 32 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | *37* | 36 | 0 | 0 | 0 |
| 33 | 6 | 8 | 10 | 17 | 24 | 28 | 43 | 0 | *6* | 8 | 43 | 17 | 4 |

Although this method is very simple and easy, it introduces the basic idea of estimating the functions exploring the data of surrounding proteins and produces roughly good results compared to other techniques.

## 4.3 Chi-square method

Chi-square method is to infer protein functions based on $X^2$- statistics and overcome the third problem of the neighboring count method (confidence level) that takes the fraction of each function among total database of proteins. It is developed by Hishigaki [16]. As shown in Fig.4.2, a sample of interacted proteins contains annotated proteins which they have functions (white proteins) and others do not have (black ones) are introduced.

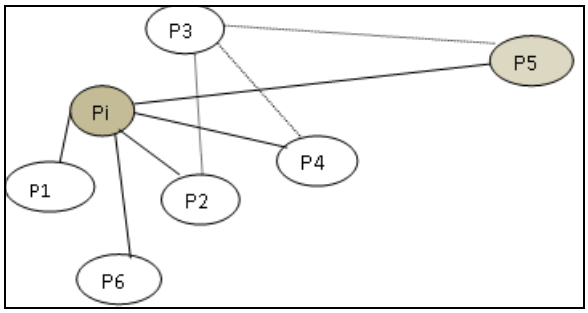


Fig.4.2 sample of proteome network

For a protein $p_i$, let

$n_i(j)$ = the number of proteins interact with $P_i$ and have function $F_j$.

$e_i(j)$ = #Nei(i) $x$ $\pi_j$ be the expected number of proteins in Nei(i) having function $F_j$, where, #Nei(i) is the number of proteins in Nei(i), and $\pi_j$ is the fraction of the function among proteins. $s_i(j)$ is defined as the score value.

$$s_i(j) = [n_i(j) - e_i(j)]^2 / e_i(j) \quad (4.1)$$

This method is one of the significance methods used to specify the confidence (significance) level and value for each estimated function. Although Chi-square method depends on the statistical measurements, their results do not reach for the required target as well as it introduces poor results compared to the neighbor counting method.

Although Protein function prediction approaches consider the frequency of proteins having the function of interest as well as all the neighbors with less weight placed on far away neighbors than close neighbors, the chi-square method does not take the distance between the proteins, it just overcomes the significance value. By implementing this method to Yeast proteome network data, the next results are collected as shown in Table 4.2. The estimated results are poor comparing to the

previous method. It is noticed that protein ID 19 has only one overlapping function (8), and protein ID 33 has two overlapping functions (6, 8), where, the neighbor counting method presents two functions in ID 19 and four functions in ID 33, in addition the other estimated and overlapping functions for different protein ID. Also as mentioned above the leave one out method will be applied to this method to calculate the sensitivity and specificity for chi-square method.

Table 4.2 basic and estimated functions for Yeast proteome using Chi-square method

| Protein_ID | Basic Functions | | | | | | | | Estimated Functions | | | | | Overlapping |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 15 | 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 7 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 15 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 10 | 13 | 27 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 3 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 8 | 43 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 1 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 7 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 0 | 0 |
| 23 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | 16 | 22 | 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 3 | 16 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | 2 | 3 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | 15 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | 6 | 8 | 10 | 17 | 24 | 28 | 43 | 0 | 6 | 8 | 11 | 0 | 0 | 2 |

## 4.4 Markov random field method

Markov random field method overcomes all the above problems by considering the entire interactions network. Because it considers the frequency of proteins having the function of interest as well as all the neighbors with less weight placed on far away neighbors than close neighbors. Also it can calculate the probability that an un-annotated protein has a function of interest [17].

The approach considers that for a given function $f_i$, it assigns 1 to proteins which are annotated and have that function, and assigns 0 to proteins which are annotated and do not have this function.

Let X = ($x_1$, ... $x_n$, $x_{n+1}$, ..., $x_{n+m}$) denotes the functional annotations of all proteins where $x_1$….. $x_n$ are un-annotated, and $x_{n+1}$ … $x_{n+m}$ are annotated. It gets the prior probability distribution of *X* based on the interaction network technique; *Gibbs distribution* [18]. Defining Gibbs distribution for protein-protein interaction network without the internal interactions is done as:

$$P(X) = \prod_{i=1}^{N} \pi^{Xi}(1-\pi)^{1-Xi} = \left(\frac{\pi}{1-\pi}\right)^{Ni}(1-\pi)^{N} \tag{4.2}$$

Where N is the total number of proteins (N = n + m) and $N_i$ is the number of proteins which are assigned with 1.

The approach needs to define the following items:

• $N_{11}$ the number of edges in which both the edges proteins got assignment 1.

$$N_{11} = \sum_{(i,j)\in S} x_i x_j \tag{4.3}$$

= # {(1←→1) pairs in S},

• $N_{10}$ the number of edges in which the edges proteins have different assignments.

$$N_{10} = \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \tag{4.4}$$

= #{(1←→0) pairs in S},

• $N_{00}$ the number of edges in which both the edges vertices got assignment 0.

$$N_{00} = \sum_{(i,j)\in S} (1-x_i)(1-x_j) \tag{4.5}$$

= #{(0←→0) pairs in S}.

Also the approach defines $H_1(X)$ and $H_2(X)$ as a model:

$$H_1(x) = \alpha N_1 . \quad (4.6)$$

$$H_2(x) = \beta N_{10} - \gamma N_{11} - N_{00} . \quad (4.7)$$

Where α, β and γ define the model parameter set Θ and are calculated.

The Gibbs distribution can be written as:

$$P(X \mid \theta) = (1/Z(\theta))e^{-H1(X)-H2(X)} . \quad (4.8)$$

Where $Z(\theta)$ is the function of the parameters.

The approach has used sampling method (Gibbs sampling) to get the unknown $x_i$. After estimating the parameters and getting the prior distribution, approach uses the Bayesian algorithm to get the posterior probability given the prior probability (count of given function $f_i$ by number of known proteins). By applying this approach to the yeast proteome, better results (high sensitivity and high specificity) are collected. As shown in Table 4.3, some of $x_{00}$, $x_{10}$, and $x_{11}$ values are collected. It is noticed, some values as function ID 4 has $x_{11}$ equal 0 which reflects the poor estimation for this function through the surrounding proteins and interactions using this technique.

Table 4.3 Edge number for the three states of the interaction network

| Function ID | $X_{00}$ | $X_{10}$ & $X_{01}$ | $X_{11}$ |
|---|---|---|---|
| 1 | 2495 | 52 | 12 |
| 2 | 2437 | 101 | 21 |
| 3 | 2372 | 164 | 23 |
| 4 | 2547 | 12 | 0 |
| 5 | 2153 | 297 | 109 |
| 6 | 2245 | 240 | 74 |
| 7 | 2212 | 299 | 48 |
| 8 | 2413 | 101 | 45 |
| 9 | 2388 | 156 | 15 |
| 10 | 2105 | 346 | 108 |
| 11 | 2485 | 55 | 19 |
| 12 | 2382 | 123 | 54 |

## 4.5 Prodistin

*PRODISTIN* [133] uses the Czekanowski-Dice distance between each pair of proteins as a distance metric and clusters the proteins using the BIONJ clustering algorithm [134]. The Czekanowski-Dice distance between every two proteins u and v is calculated. The approach uses a simple first-order model of the sampling variances and covariance of evolutionary distance estimates. This model leads to a simple expression of the minimum variance reduction, which is fully consistent with the agglomerative approach. These elements are combined to form BIONJ. And the minimum results will be selected. The formula of this technique is shown in equation 4.9.

$$D(u,v) = \frac{\left|Nu \Delta Nv\right|}{\left|Nu \text{ Y } Nv\right| \left|Nu \text{ I } Nv\right|} \quad . \tag{4.9}$$

Where: Nu refers to the set that contains u and its level-1 neighbors. Nu Δ Nv refers to the symmetric difference between two sets u and v. D (u, v) is the distance between u and v.

- D (u, v) < 1 if u and v are level-1 neighbors,
- D (u, v) will be evaluated to 0. If Nu = Nv, and
- D (u, v) will be evaluated to 1. if Nu ∩ Nv = ∅.

The largest connected components in a protein interaction network are only used. The BIONJ algorithm produces a hierarchical classification tree. A PRODISTIN functional class for a function is defined to be the largest possible sub-tree in the classification tree that: 1) it contains at least three proteins having the function; and 2) it has at least 50% of its annotated members having the function. Un-annotated proteins in the functional class are then predicted with the function.

## 4.6 Samanta

Samanta technique is like *PRODISTIN*. It applies clustering techniques to partition the proteome into functional modules, but using a different distance metric [135]. A p-value between two proteins is computed as follows:

$$p(N,u,v,m) = \frac{\binom{N}{m}}{\binom{N}{n1}} \frac{\binom{N-m}{n1-m}\binom{N-n1}{n2-m}}{\binom{N}{n2}} \tag{4.10}$$

Where: N refers to all proteins in the interaction network,
$m = |N_u \cap N_v|$, $n_1 = |N_u|$, and $n_2 = |N_v|$

The p-value is reflective of the likelihood of proteins u and v sharing m neighbors given that u has n1 neighbors and v has n2 neighbors. A similar measure known as the Hyper-geometric distance is also introduced in [136] for estimating interaction reliability. Using the p-value as a distance metric, proteins are clustered using a hierarchical clustering approach. Begin with each protein as a cluster. The two clusters with the smallest p-value are merged to form a cluster. The p-value between two clusters is computed by the geometric mean of the p-value of its components.

## 4.7 Support Vector Machines

Lanckriet et al. [137] has introduced an integrated Support Vector Machines classifier for function prediction, in which protein-protein interaction data was used to derive one of the kernels using pair-wise interaction similarity between proteins based on interaction data.

### 4.8 Functional flow

Nabieva et al. [138] has proposed a network-based algorithm that simulates functional flow between proteins. Proteins are initially assigned infinite potential for a function if a protein is annotated with that function and 0 potential otherwise. Functions are then simulated to flow from proteins with higher potential to their level-1 neighbors that have lower potential. The amount of flow is influenced by the reliability of the interactions between interaction partners, which is derived similarly as in our approach.

## 4.9 Leave one out

The relationship between quality of prediction and network's information is estimated by applying the leave-one-out method which is applied to proteins have at least one interaction. Calculation of sensitivity and specificity is required to determine the quality of the prediction. The sensitivity (SN) and specificity (SP) are defined as:

$$SN = \frac{\sum_{i}^{k} Ki}{\sum_{i}^{k} ni} \tag{4.11}$$

$$SP = \frac{\sum_{i}^{k} Ki}{\sum_{i}^{k} mi} \tag{4.12}$$

Where $n_i$ is the number of observed functions for protein $P_i$, $m_i$ is the number of predicted functions for protein *Pi*, and *ki* is the overlap between them.

## 4.10 Summary

In this chapter, the most common methods used to estimate the protein function prediction exploring data of protein-protein interactions network were defined and discussed (neighbor counting method, chi-square method, MRF, Prodistin, Samanta, support vector machine, functional flow). It was noticed that neighbor counting method was the seed of these methods which could get enhanced results by integrating its technique with weighted interactions. Markov random field

method was the best method (high sensitivity) that because it used weights for the interacted proteins. It considered the weights for seeds (single proteins), interacted proteins having the same function, and interacted proteins which one of them had the function and the other did not have it. Also support vector machine method got better results.

Chapter 5

# Improvement of Yeast Protein Functions Prediction Using Weighted Protein-Protein Interaction

Protein function prediction is among the most important tasks in the field of bioinformatics as it can lead to understanding cell activities. Since the most recent methods of protein functions prediction are performed using protein-protein interaction network data, so the reliability of these interactions is very critical point in prediction process. Protein-protein interactions are identified by high-throughput experimental methods as Y2H, mass spectrometry of co-immunoprecipitated protein complexes Co-IP, Gene co-expression, TAP purification Cross link, Co-purification, biochemical and other methods. It is noticed that a challenging technical problem is done by using the first two methods which lead to spurious interactions as self activation in Y2H and abundant with contaminants CO-IP. This problem leads to false positive interactions [98]. So getting a quantitative method for evaluating the path way through proteomics data is required. A number of experimental and computational approaches are implemented for large-scale mapping of PPIs to realize the potential of protein networks for systems analysis. One method utilizes multiple independent sets of training positives to reduce the potential bias in using single training set as association with publishing identifier, or foundation in two species or more, or have expression correlation more than 0.6 [139]. Another technique is getting the conserved patterns of protein interaction in multiple species [140]. Also there are many methods for determining the reliability of interactions [138, 141-145].

Usually, equally-weighted protein-protein interactions (PPI) are used to predict the protein functions. In this chapter, we provide a new weighting strategy for PPI to improve the prediction of protein functions. These weights are dependent on the local and global network topologies as well as the number and type of experimental verification methods. The proposed methods are applied to yeast proteome and integrated with neighbor counting method to predict the protein functions of unknown proteins. The results reveal great improvement in the sensitivity and specificity of prediction of two functional categories: cellular role and cell locations. The studied species is presented in the first section. Later on, the study challenges, suggested methodology, and the estimated results are indicated respectively. The target of this study is to improve the reliability of the interactions and increasing the confidence of the protein function prediction.

## 5.1 Yeast Saccharomyces Cerevisiae

### 5.1.1 Yeast history

Yeast (Saccharomyce Cerevisiaeas) is a model of organisms. It is a very simple eukaryote as shown in Fig.5.1. Yeast is used as a model for human Genome which has about 6400 protein-coding genes. In April 1996, the complete genome sequence of the brewers and bakers yeast saccharomyces Cerevisiae was sequenced. The project was launched by an initiative of A. Goffeau 1989 and the European Commission (EC) to sequence chromosome 111 in a pilot study. This was an important event, not just because it was the first complete eukaryotic genome sequence, but also because it was the first total sequence for an important model organism for which there is a large constituency of researches ready and able to exploit the sequence data. One third of their functions are un-annotated as shown in Fig.5.2.

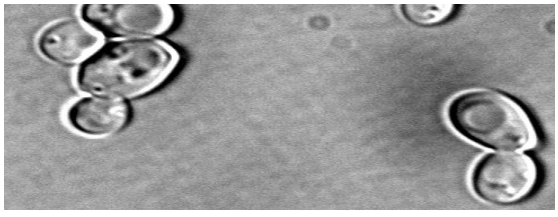

Fig.5.1 Yeast as simple model organism [wiki/Yeast]

### 5.1.2 Why Yeast

Herein, we can summarize why the current research focus on yeast as following:

1- Yeast has already provided biologists with a valuable resource for determining the function of individual human genes involved in medical problems, such as cancer, neurological disorders, and skeletal disorder. Over the next few years, scientists in the United States and Europe will piece together for the first time a comprehensive look at how all the genes in a eukaryotic cell function as an integrated system. ”The yeast genome is closer to the human genome than anything completely sequenced so far, “said Dr. Francis Collins, director of the National Center for Human Genome Research NCHGR, part of the National Institutes of health NIH .

2- Biologists have studied yeast, known by its scientific name saccharomyces Cerevisiae, for many decades because it offers valuable clues to understanding the work of more-advanced organisms. Humans and yeast, for example, share a number of similarities in their genetic make up. For example, many regions of yeast DNA contain stretches of DNA subunits, called bases that are very close or identical to those in human DNA. These similarities tell scientists that, genes in those regions play a critical role in cell function in both species, or they would have been lost during the 1 billion years of evolution that separate yeast and humans. About one-third of yeast genes are related to those in the human. Some of these critical processes include DNA coping and repair of damaged DNA, Protein synthesis and transport across membranes, and control of metabolic processes.

3- In cancer research, S. Cerevisiae has emerged as an important model for studying control of the eukaryotic cell cycle. Although yeast DNA shares many similarities with human DNA, finding yeast genes is easier because the yeast genome lacks the long stretches of filler DNA and repeated bases the human genome contains, which often cause scientists problems when examining a long DNA piece for the presence of genes. Yet, scientists know.

4- The difficulty of experiments on human body.

5- Some researchers made seminal discoveries concerning the control of the cell cycle. They have identified key molecules that regulate the cell cycle in all eukaryotic organisms, including yeast, plants, animals and human based on study of Yeast. Defects in cell cycle control may lead to the type of chromosome alteration seen in cancer cells.

6- Yeasts have recently been used to generate electricity in microbial fuel cells, and produce ethanol for the bio-fuel industry.

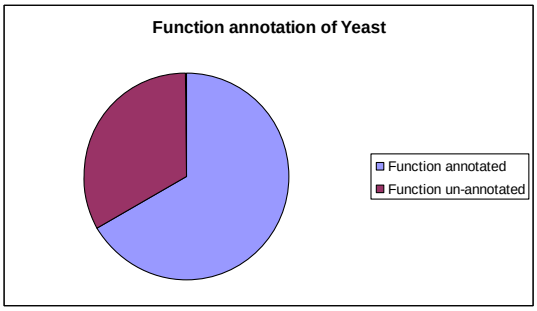


Fig.5.2 Yeast function annotation

### 5.1.3 Yeast Features

In this study, 6416 proteins are studied and divided into three functional categories. Fig.5.3: Cellular role functions including 43 sub-function categories, Cell location function including 29 sub-function categories, and Biochemical functions including 57 sub-function categories. Table 5.1 shows some examples of sub-function categories.

Table 5.1 Yeast sub-function categories, function name, and the number of proteins for each function.

| Function category | Function name | # proteins |
|---|---|---|
| Cellular role | Aging | 39 |
| Cellular role | Amino-acid metabolism | 218 |
| Cellular role | Carbohydrate metabolism | 254 |
| Cellular role | Cell adhesion | 4 |
| Cellular role | Cell cycle control | 213 |
| Cellular role | Cell polarity | 216 |
| Cellular role | Cell stress | 331 |
| Cellular role | Cell structure | 120 |
| Cellular role | Cell wall maintenance | 184 |
| Cellular role | Chromatin/chromosome structure | 274 |
| Cellular role | Cyto kinesis | 40 |
| Cellular role | DNA repair | 154 |
| Cell location | Bud neck | 61 |
| Cell location | Cell ends | 6 |
| Cell location | Cell wall | 70 |
| Cell location | Centrosome/spindle pole body | 72 |
| Cell location | Contractile ring | 3 |
| Cell location | Cytoplasmic | 755 |
| Cell location | Cytoskeletal | 107 |
| Cell location | Endoplasmic reticulum | 225 |
| Cell location | Endosome/Endosomal vesicles | 36 |
| Cell location | Extracellular excluding cell wall | 34 |
| Biochemical | ATPase | 247 |
| Biochemical | ATP-binding cassette | 31 |
| Biochemical | Activator | 46 |
| Biochemical | Active "transporter," primary | 93 |
| Biochemical | Active "transporter," secondary | 201 |
| Biochemical | Adhesin/agglutinin | 7 |
| Biochemical | Anchor Protein | 13 |
| Biochemical | Channel [passive transporter] | 15 |
| Biochemical | Chaperones | 90 |
| Biochemical | Complex assembly protein | 76 |

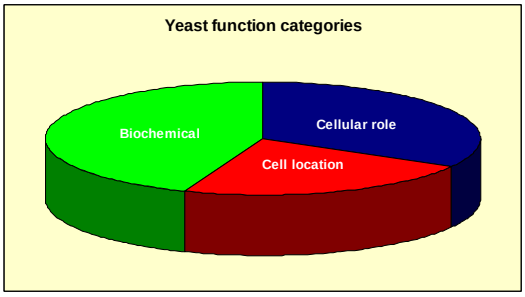


Fig.5.3 Yeast function categories

## 5.2 Challenges of the study

Protein function prediction through protein-protein interaction network is considered as one of the most difficult problems in proteomic research era. It has a lot of problems as follow:

### 5.2.1 Yeast protein Naming

The official Gene Name of S. Cerevisiae gene is referred as the standard name on SGD locus page, and generally becomes the standard name based on its publication in a peer-reviewed paper describing characterization of that gene. Any alternative Gene Name is referred to as an Alias. Gene Names in S.Cerevisiae are generally three letters followed by number. For example: CDC28-A Gene Name conferred on a nuclear ORF on the basis of genetic characterization. Different copies of duplicated genes may be indicated by an extension to the end of the Gene Name. This extension can be made by either adding a letter, e.g. ‘A’ or ‘B’ as in the case of the ribosomal protein genes, or by adding a hyphen and a number, e.g. ‘-1’, ‘-2’, as in the case of YRF1genes encoding the ‘Y’ helicase or the ribosomal RNA genes.

As one of the problems shown in this research area is the protein naming. Since the proteins are collected from different centers, universities, institutes, and countries, each protein may have more than one name. In Yeast proteome, protein may have up to eight names as shown in Table 5.2. Protein ID =1 has four names: two as gene names (AAC1 – ANC1) and others as accessions (YM9796.09 - YMR056C). Also protein ID = 17 has 8 different names (ABF2- CDRP1- HIM1- HM- P19 HM- mt

TFA- YM9916.11- YMR072W). Since different names lead to a conflict especially in PPI, the reliability of each protein and its name should be found. So herein, all possible names for each protein are created and collected in single field.

Table 5.2 Yeast proteins and their different names

| Protein ID | Name 1 | Name 2 | Name 3 | Name 4 | Name 5 | Name 6 | Name 7 | Name 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | AAC1 | ANC1 | YM9796.09 | YMR056C | | | | |
| 2 | AAC3 | ANC3 | YBR0753 | YBR085W | | | | |
| 3 | AAD3 | YCR107W | | | | | | |
| 4 | AAD4 | D0752 | YDL243C | | | | | |
| 5 | AAD6 | YFL056C | | | | | | |
| 6 | AAD10 | J2245 | YJR155W | | | | | |
| 7 | AAD14 | AKR9B1 | N0300 | YNL331C | | | | |
| 8 | AAD15 | O0205 | YOL165C | | | | | |
| 9 | AAH1 | N1208 | N1825 | YNL141W | | | | |
| 10 | AAP1 | H8179.24 | YHR047C | | | | | |
| 11 | AAR2 | YBL0611 | YBL06.06 | YBL074C | | | | |
| 12 | AAT1 | YKL461 | YKL106W | | | | | |
| 13 | AAT2 | AAT1 | ASP5 | L1746 | YLR027C | | | |
| 14 | ABC1 | COQ8 | G2920 | YGL119W | | | | |
| 15 | ABD1 | YBR1602 | YBR236C | | | | | |
| 16 | ABF1 | BAF1 | OBF1 | REB2 | GFI | YKL505 | YKL112W | |
| 17 | ABF2 | CDRP1 | HIM1 | HM | P19-HM | mtTFA | YM9916.11 | YMR072W |

## 5.2.2 Protein clusters and interaction

Proteins are divided into clusters according to special characteristic as common function, structure, pathway; eg. Protein may be found in more than one cluster. Assuming the function category as cluster, protein is considered in one to eight different clusters (function categories). For example protein name (ACT1, ABY1, END7, YFL039C) in cellular role category has 7 sub-functions (Cell polarity-cell structure- chromatin/chromosome structure- mating response- other metabolism-poll 2 transcription- vesicular transport). Although there are 2523 proteins in yeast have no annotated functions (one third of proteome number), there are 1236 proteins have two functions and 383 proteins have three functions. It is noticed that proteins in different cluster may have interactions which inverse/contradictory the basic idea of prediction “If two proteins interact, they are neighbors of each others”. For the un-

annotated proteins, the functions of their neighbors contain information about the function of the un-annotated protein. Since the considered interactions are physical interactions, the interactions over different functional clusters are not important in the study and will be neglected from the calculations regarding the estimated correlation over protein clusters/functions.

### 5.2.3 Protein-protein interactions

Since more than one third of yeast proteome are losing for their functions in every category (Table 5.3), it should get reliable protein interactions. So the used data is collected from MIPS (Munich Information Center for Protein Sequences (MIPS, http://mips.gsf.de)) which is the most reliable and robust data sources for protein interactions. MIPS contains 2559 physical interactions among 6416 proteins. These interactions have 120 ones as self interactions which lead to worth results especially if the protein is found in two different function categories (clusters).

Self interactions have been removed from the proposed technique. Moreover, both the interaction network and the functional annotations of the proteins are incomplete. It is noticed that, there is variety in the number of interactions for each protein which reaches for 29 interactions in some proteins. Fig.5.4 shows the variety of interactions over the number of proteins.

Table 5.3 Numbers of Annotated and un-annotated proteins for All Proteins Based on Three Functional categories

| Biochemical function | |
|---|---|
| Annotated | 3353 |
| Un-annotated | 3063 |
| Sub-cellular location | |
| Annotated | 3181 |
| Un-annotated | 3235 |
| Cellular role | |
| Annotated | 3894 |
| Un-annotated | 2522 |

As shown in Fig.5.4, more than 300 proteins have more than 5 interactions and about 1000 proteins have more than two interactions (two neighbors). These interactions reach for 29 interactions as Proteins ID =1556 "KRE28", and protein ID= 3258 "SNP1". Since the reliability of these interactions is very critical issue in the study, furthermore that all interactions take the same weight "same strength" which it does not indicate the strength of the interaction between the two studied proteins, the study introduces a new and good reliable measure for determining the strength of interactions. These measures produce a weight for each interaction relating to a lot of parameters will be discussed later.

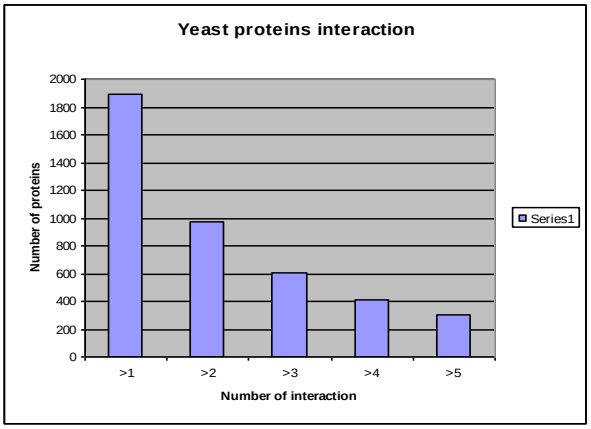


Fig.5.4 yeast proteins interactions

## 5.3 Protein function prediction using a new weighting algorithm for PPI

Usually, equally-weighted protein-protein interactions (PPI) are used to predict the protein functions. In this study, it is provided a new weighting strategy for PPI to improve the prediction of protein functions. These weights are dependent on the local and global network topologies as well as the number and type of experimental verification methods. The proposed methods are applied to the studied material "yeast proteome" and integrated with neighbor counting method to predict the protein functions of un-known proteins. The study introduces a novel algorithm by comparing the proteins in protein-protein interaction network to the connected routers in the same autonomous number of networking. Protein acts as router (node) and edge (interaction between two proteins) as connection between two routers as shown in

Fig.5.5, 5.6, where routers can have up to 100 interactions but 29 interactions as maximum in yeast proteome.

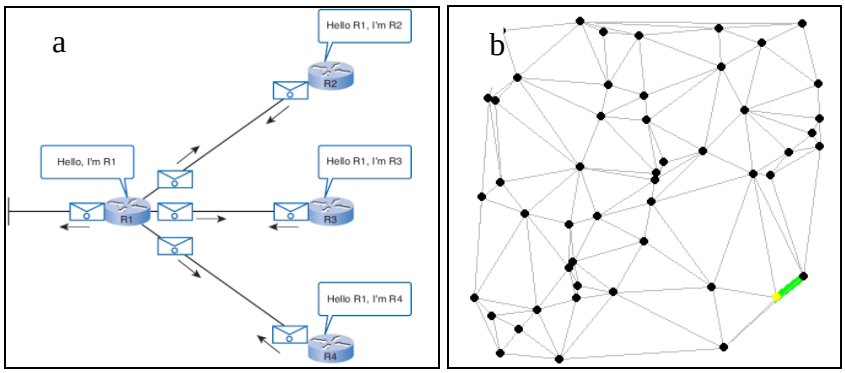


Fig.5.5 (a) small routing connection system (b) large routing system

The routing system is performed using many ways as (LAN, WAN, Serial) as the different ways of connections (different experimental methods for protein interactions in protein system). Initially, the router is not aware of any neighbor routers on the link. So linked state protocol is applied to the routing system, where a *link* is an interface on a router and the protocol is the control system of all connected routers. The protocol includes some information as: 1)-Interface's IP address/mask, 2)-The type of network; Ethernet (broadcast) or serial point-to-point link, 3)-The cost of the link, and 4)-Any neighbor routers on that link. In protein system the same protocol is happened 1)- the protein is identified by name, ID, sequence, and its functions (if known), 2)- the type of network; just interaction between two proteins or dense interactions (cluster), 3)- the weight of the interaction (our contribution), and 4)- all neighbors of the adjacent protein.

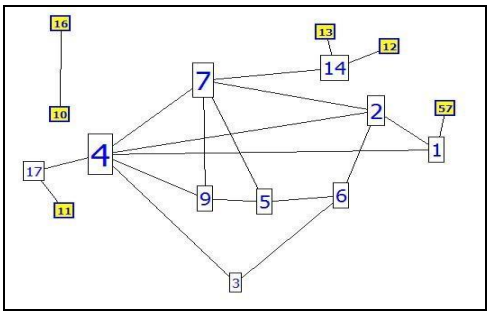


Fig.5.6 a part of connected proteins indicates the leafs (yellow nodes)

The protein interactions are calculated till reach the second level. The algorithm is performed on three steps 1)- determining the level and degree for each adjacent protein, 2)- calculating the weight (cost) for each interaction, and 3)- integrating these data to predict the function of the un-annotated proteins using neighborhood counting method.

### 5.3.1 Protein degree and level

There is a difference between the degree and the level of certain node. The degree of protein is defined as the total number of adjacent proteins (proteins directly connected) [54]. As shown in Fig.5.7, protein A has degree equal 6. But the protein level is the layer of nodes related to the studied one.

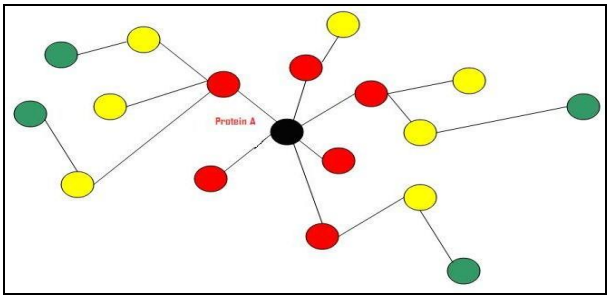


Fig.5.7 Degree of the Protein (black node A has degree equal 6)

The directed nodes are the first level and their neighbors are the second level and so on as shown in Fig.5.7. The red nodes are the first level of protein A (Black one) where yellow nodes are the second level proteins (nodes connected to protein's A neighbors) and the green nodes are the third level.

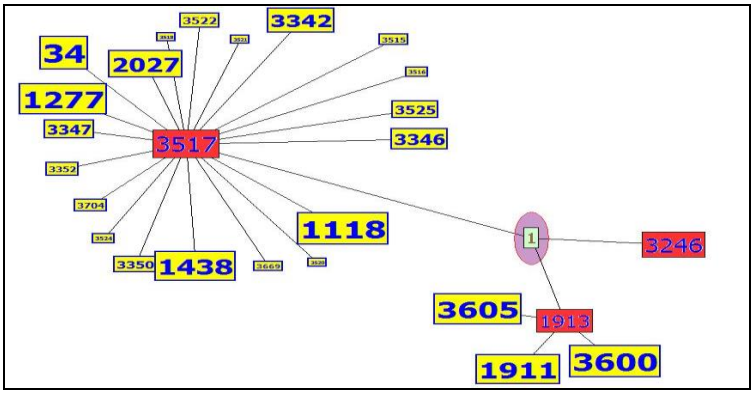


Fig.5.8 real part of yeast proteome using Inter View program

In router networks, the hop count principal (the number of routers which send packets through routing system) is performed to determine the router level. In our study, the second level is assumed to be enough to get the most important information about the function of protein. Also, some cases of interactions act as closed loops which have been dealt as first level [54, 146-148]. By applying the concept of node level to the proposed data, it can be noted as shown in Fig.5.8, proteins ID numbers (1913, 3246, and 3517) are the first level for the studied protein (1) and all the yellow nodes are second level. As shown in Table 5.4, protein IDs, number of interactions, and the IDs of neighbors are produced.

Table 5.4 sample of proteins and their interactions

| Protein ID | # interactions | p1 | p2 | p3 | p4 | p5 | p6 | p7 | P8 | p9 | p10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 1 | 3258 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | 23 | 19 | 33 | 33 | 84 | 304 | 333 | 370 | 407 | 568 | 1065 |
| 34 | 17 | 56 | 475 | 1118 | 1277 | 2027 | 3350 | 3352 | 3342 | 3346 | 3347 |
| 35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 36 | 5 | 36 | 36 | 2557 | 3092 | 4052 | 0 | 0 | 0 | 0 | 0 |
| 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 39 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 40 | 1 | 3802 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 41 | 3 | 1726 | 3275 | 386 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 46 | 1 | 3708 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47 | 1 | 4590 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 51 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

### 5.3.2 PPI weighting strategy

In this study, the Protein-Protein interactions are weighted. Three factors are considered to calculate these weights for all interactions. These factors are: experimental verification methods, Interaction Generality for local topology, and Interaction Generality for global topology.

#### 5.3.2.1 Experimental verification method

Protein-protein interactions (PPI) can be identified by a lot of methods as: affinity precipitation, affinity chromatography, Yeast to hybrid, purified complex, reconstituted complex, biochemical assay, chemical lethality and chemical rescue [149]. Since data sets of PPIs contain a lot of false positive interactions, a crucial step in analyzing proteomics data is separating the subset of credible interactions from the background noise. The technique of this method is divided into two parts: the first technique takes the count number of experimental methods only without determining their strength (reliability). The second technique takes the reliability of each method [150] into consideration furthermore the count number of methods. Then a comparison between these two techniques is introduced to indicate their effect on the protein function prediction process.

**A. Number of experimental methods**

According to the different chemical structure of amino acids, each protein interaction pair is identified by one method or more. By applying the previous mentioned experimental methods on Yeast protein interactions, it is found that interactions are identified by one or two up to ten different methods.

Table 5.5 shows some of Yeast interaction pairs, number of identification methods. And determine the state of each used experimental method for identification ((1, 2) for satisfy and (–) not found).

| Protein_1 | Protein_2 | # Identification Method | Y2H | Cross-link | affinity chromo | precipitation | assay | purification | in Vetro | Others |
|---|---|---|---|---|---|---|---|---|---|---|
| YKL161C | RLM1 | 1 | 1 | - | - | - | - | - | - | - |
| AAC1 | YHR005C-A | 1 | | 1 | - | - | - | - | - | - |
| AAD14 | AAD14 | 1 | 1 | - | - | - | - | - | - | - |
| AAD6 | YNL201C | 1 | 1 | - | - | - | - | - | - | - |
| ABP1 | ACT1 | 3 | 1 | - | 1 | - | - | 1 | - | - |
| ABP1 | RVS167 | 4 | 1 | - | - | 1 | - | - | - | 2 |
| ABP1 | SRV2 | 3 | - | - | - | - | - | - | 1 | 2 |
| YER045C | PSE1 | 1 | 1 | - | - | - | - | - | - | - |
| ACC1 | DMC1 | 1 | 1 | - | - | - | - | - | - | - |
| ACC1 | SNP1 | 1 | 1 | - | - | - | - | - | - | - |
| ACE2 | YNL157W | 1 | 1 | - | - | - | - | - | - | - |
| ACS2 | SNP1 | 1 | 1 | - | - | - | - | - | - | - |
| ACT1 | ACT1 | 4 | 1 | 1 | 1 | - | 1 | - | - | - |
| ACT1 | AIP1 | 1 | 1 | - | - | - | - | - | - | - |
| ACT1 | BEM1 | 2 | 1 | - | - | 1 | - | - | - | - |
| ACT1 | BNI1 | 1 | 1 | - | - | - | - | - | - | - |

Regarding the data taken from MIPS database, each interaction pair (protein names) has its number of experimental methods as shown in Table 5.5. For example, it is found that, protein interaction (ABP1 and ACT1) is identified by 3 methods (Yeast two hybrid, Affinity chromatography, and Purification) while interaction pair (ACT1 and BEM1) is identified by 2 methods (Y2H, and precipitation).

It is noticed that, cells have title ‘other’ mean that, the interaction can be happened relating to one of these reasons: (redundancy, gel retardation, gel filtration or identified in Pub Med). Also self interaction can be created as week point in the experimental method as shown in interactions (AAD14 and ACT1).

Some interactions are identified by nine or ten experimental methods as (ADA2 and GCN5) and (CMD1 and NUF1) respectively. Fig.5.9 shows that most of the Yeast protein interactions are identified by one experimental method (~ 1890 interactions). Also shows the maximum number of identifications reached for 10.

Herein, our technique considers the interactions that identified by just one experimental method as low confidence (with weight 0.5) and interactions identified by more than one as high confidence (with weight 1). As shown in Fig.5.10, each identified method and its number of interaction is indicated. It is noticed that, the Y2H

is the common method which identifies more than 30% of the Yeast protein interactions.

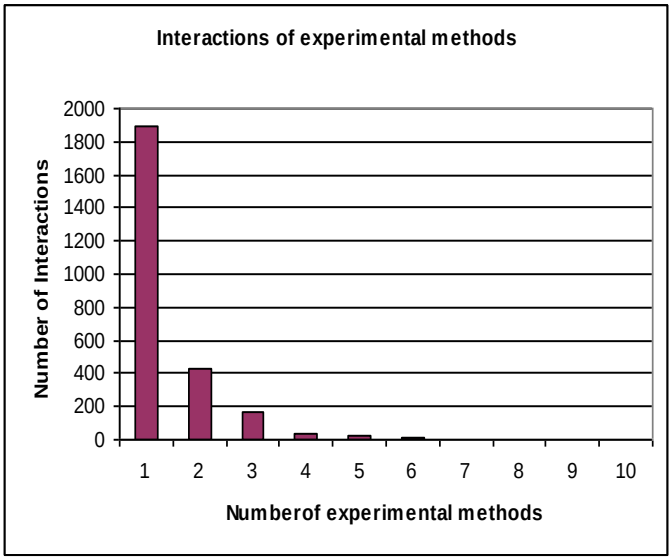


Fig.5.9 Experimental methods related to the interaction pairs that reach ten in some interactions. (More than one third of Yeast interactions have been identified by one method Y2H)

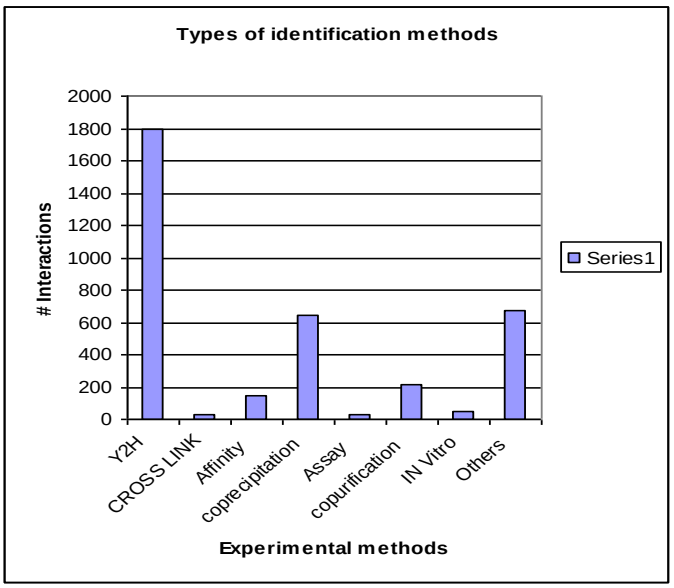


Fig.5.10 Experimental method type and its number of interaction pairs.

**B. Number of experimental methods and their reliability**

As shown in Fig.5.10, most of the interactions are identified by Y2H method which contains spurious interactions as self activations furthermore the false positive

interactions. So reliability confidence scale should be considered as well as the number of interaction. The reliability scale of experimental methods was taken from the GRID database as shown in Table 5.6.

Table 5.6 the reliability scale of some experimental methods regarding to the GRID datasets

| Experimental method | Reliability score |
|---|---|
| Chromatography | 0.82 |
| Precipitation | 0.46 |
| Two hybrid | 0.27 |
| Biochemical Assay | 0.67 |
| Purification | 0.89 |
| Other | 0.35 |

Suppose the reliability $Rel_i$ of each experimental method (i) is estimated as in Table 5.6. And assume that the experimental methods are independent. The reliability Rel (u, v) of the interaction of (u, v) is taken as the probability that at least one of the involved experimental methods is reliable [151]. The reliability score will be calculated from the formula shown in equation-5.1. It is noticed that, the term $n_{i,\,u,\,v}$ was assumed to equal one (assuming one trial only for each experimental method).

$$Rel\,(u,v) = 1 - \Pi_{i \epsilon E(u,v)}\,(1-Rel_i)^{ni} \qquad (5.1)$$

**C. Comparison between the reliability methods**

Regarding the number and the type of experimental methods, there are different values in confidence scores of the reliability. As shown in Table 5.7; the first interaction between (YKL161C and RLM1) is identified by one method (Y2H). It has confidence scores 0.5 and 0.27 respectively relating to the two suggested methods. It is noticed that, in spite of the first two rows have been identified by one experimental method, the score in the second method is different (the range reaches 0.4). Also for interactions are identified by more than one method, they have varieties in the scores according to the type of interactions.

Table 5.7 comparison between the two used methods for determining the reliability for protein interactions

| Protein 1 | Protein 2 | Experimental methods | Reliability Method 1 | Reliability Method 2 |
|---|---|---|---|---|
| YKL161C | RLM1 | Y2H | 0.5 | 0.27 |
| ACT1 | TPM2 | Link Assay | 0.5 | 0.67 |
| ABP1 | ACT1 | Y2H, Purification, Precipitation | 1 | 0.956 |
| ABP1 | RVS167 | Y2H,Precipitation, two of Others | 1 | 0.833 |

As shown in Fig.5.11, estimating the Protein functions from the surrounding proteins is created. The prediction process is performed relating to the number of interactions, number of surrounding functions (most frequent), reliability of interactions according to the type of identification methods. The types are (1 for Yeast two hybrid, 2 for affinity purifications, 3 for precipitation), and the maximum score is calculated. Protein (YGL245W) has function number 6 (Cytoplasm) in cell location function category. Also it has functions numbers 34 and 36 as cellular role function category (protein synthesis, RNA processing/modifications) respectively. The yellow cells are assumed un-known and their functions are estimated and compared by the real functions. For protein *ARC1*, it has two interactions; the first produces functions 6-34-36 by weights (0.5 (one experimental method) -0.89 (affinity purification)).

The second interaction presents 6-22-34 by weights (1- 0.94) two experimental methods and (affinity purification and precipitation) respectively. The estimated functions according to the neighbor counting method with reliability of the number of interactions are (6-34-22-36) with weights (1.5-1.5-1-0.5) respectively. The estimated functions according to the neighbor counting method with reliability of type of experimental methods are (6-34-22-36) by weights (1.83-1.83-0.94-0.89) respectively. On the other hand, in Fig.11b the protein *ACC1* has estimated functions as (20) by weight (1-0.54) respectively. The estimation in both cases will be roughly good if a threshold more than one is determined. The first case presents functions 6, 34 only but the second case does not produce any functions and can be said, this method does not provide good results. In another words, the reliability of interactions does not lead to positive prediction for the second case.

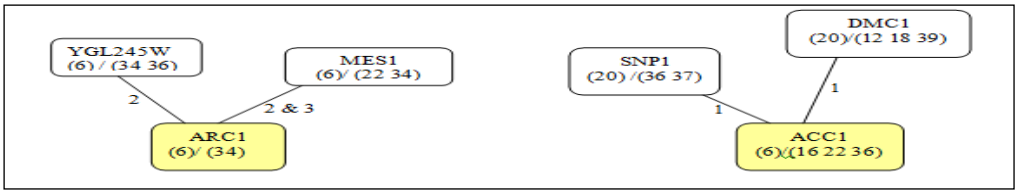


Fig.a Fig.b

Fig.5.11 protein function prediction with (a) different number of experimental methods and (b) different confidence scores of reliability.

By applying the neighbor counting method with equal weights, weights relating to the number of interactions, and weights for the strength of interaction on the three types of Yeast functions categories. The curves of sensitivity and specificity of the three protein functions categories are created as shown in Fig.5.12-14 respectively.

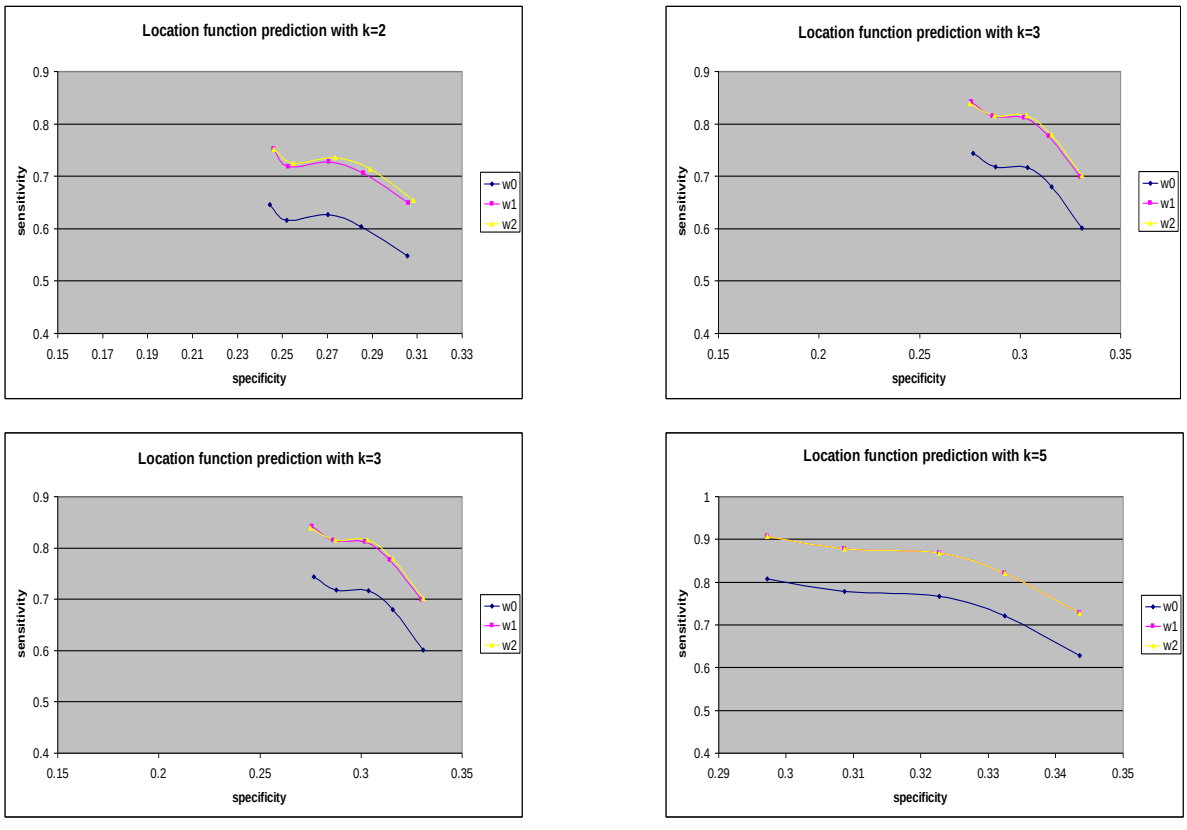


Fig.5.12; the sensitivity and specificity of protein function prediction in cell location for w0, w1, w2 (equal weight, number of experimental methods, and the reliability of interactions) with k=2-5 as number of interactions.

Since the reliability of protein interactions is very critical point in determining the PPI strength and protein function prediction, a qualitative comparison between different weights (unity, weights related to number of interactions, and weights related to the reliability) have been introduced. The reliability of interactions is affected on protein function prediction. The weights of reliability (strength and number) of experimental methods have introduced better sensitivity than equal weights (traditional neighbour counting method) in cell location and cellular role functions categories. For biochemical function category, the old method has introduced roughly good results compared to the new ones.

The sensitivity was enhanced by increasing the number of interactions for each predicted protein. The combination of these two scores will be the first measure in calculating the global interaction weight.

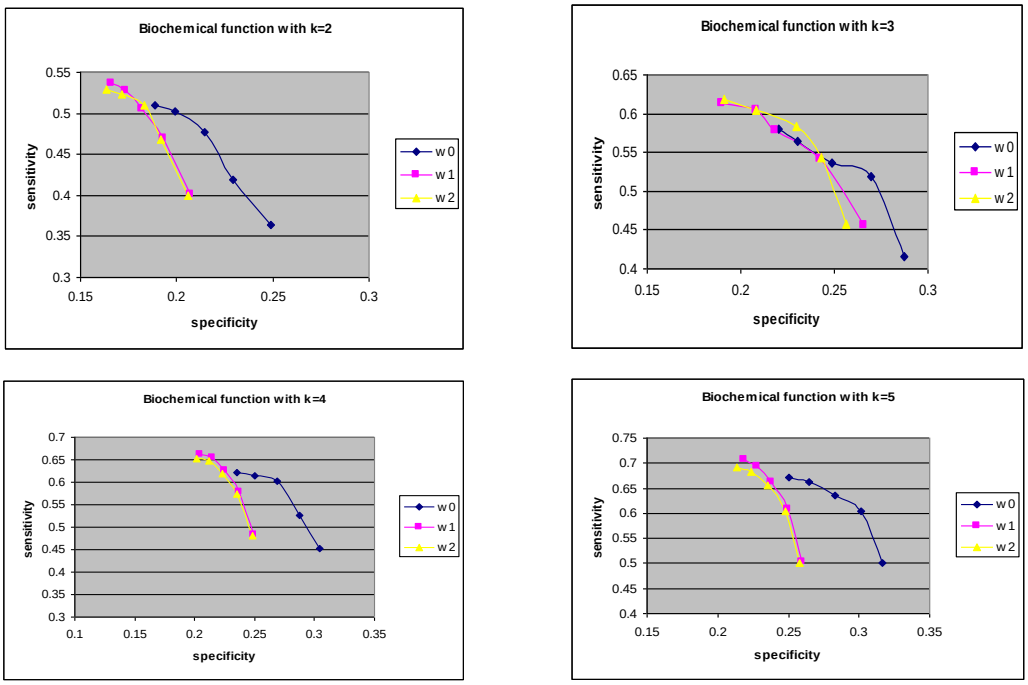


Fig.5.13; the sensitivity and specificity of protein function prediction in Biochemical for w0, w1, w2 (equal weight, number of experimental methods, and the reliability of interactions) with k=2-5 as number of interactions.

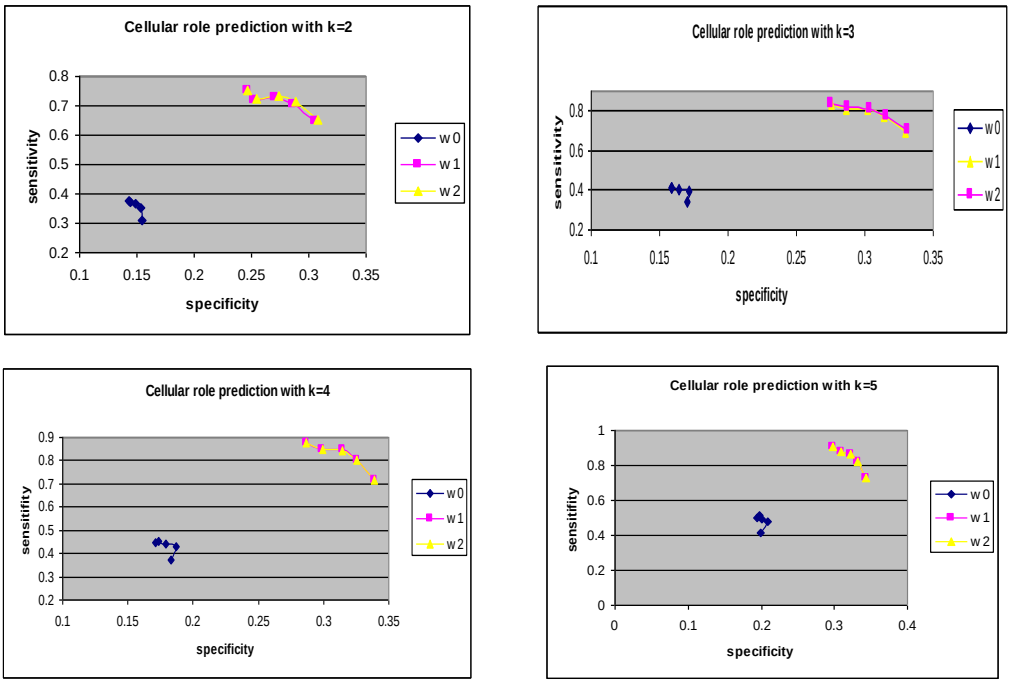


Fig.5.14 the sensitivity and specificity of protein function prediction in cellular role for w0, w1, w2 (equal weight, number of experimental methods, and the reliability of interactions) with k=2-5 as number of interactions.

### 5.3.2.2 Interaction Generality IG1 (local topology)

The second method for calculating the weights is using IG1 concept (Interaction Generality 1) [138, 144, 151]. A new method for assessing the reliability of protein–protein interactions (local topology) is obtained in biological experiments basically by getting the number of proteins involved in a given interaction (number of leafs (proteins) connecting to the two studied proteins incremented by one) as shown in Fig.5.15.

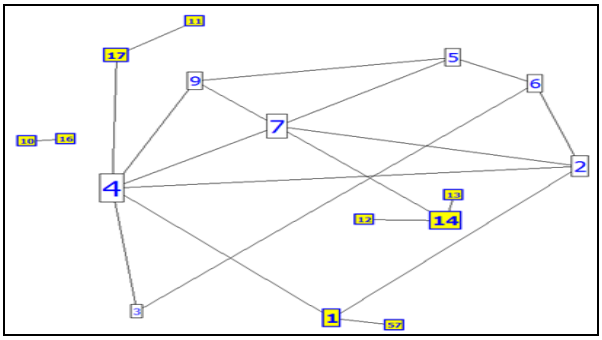


Fig.5.15 protein network data, the edge between proteins (4, 17) has IG1 value 2 where the edge between proteins (7, 14) has IG1 value equal 3.

IG1 assumes that complicated interaction networks are likely to be true positives. By implementing the IG1 on the collected data (yeast proteins interactions) it has been found that the range of IG1 is from 1 up to 21 as shown in Fig.5.16 that means there are some interactions that have many leafs proteins reach twenty. According to IG1 concept, increasing values leads to not correct interactions (false positive interaction). In the suggested algorithm, it assumed that interaction that has IG1 value less than 4 (as threshold) has high confidence (100%) where more than this value is low confidence. For example the interaction between proteins (YMR056C and YHRS01C) has IG1 value equal 3 (weight = 100%) where the interaction between proteins (YMR056C and YDR167W) has IG1 value equal 4 (weight = 50%). Also on the other hand, the interaction between proteins (YDL043C and YMR117C) has IG1 value equal 21 (weight = 50%).

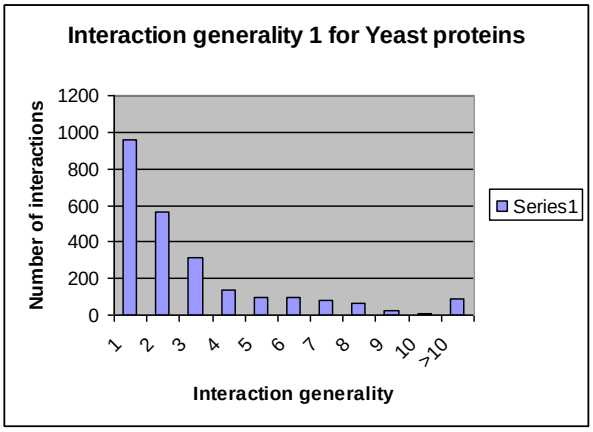


Fig.5.16 the relation between the number of interactions and IG1

Table 5.8 the reliability score of IG1 of protein interactions

| PID_1 | PID_2 | IG1 | Reliability score |
|---|---|---|---|
| 1 | 1913 | 3 | 1 |
| 1 | 3246 | 1 | 1 |
| 1 | 3517 | 4 | 0.5 |
| 7 | 7 | 0 | 0 |
| 19 | 33 | 7 | 0.5 |
| 19 | 2980 | 1 | 1 |
| 19 | 3384 | 1 | 1 |
| 22 | 2483 | 2 | 1 |
| 24 | 785 | 4 | 0.5 |
| 24 | 3258 | 14 | 0.5 |
| 25 | 5838 | 2 | 1 |
| 32 | 3258 | 13 | 0.5 |
| 33 | 33 | 0 | 0 |
| 33 | 84 | 7 | 0.5 |
| 33 | 304 | 8 | 0.5 |
| 33 | 333 | 8 | 0.5 |

As shown in Table 5.8, the reliability score of IG1 is indicated in the last column. It is noticed that, there are some cells showing zero (0) which correspond the self interactions as in cases of protein ID 7, and ID 33. According to the determined threshold (<4), the reliability score is high confidence (100%), else lead to low confidence. Although this method introduces a measure for the interaction reliability, it indicates the local topology only and does not introduce a measure for the global topology of the network. So this method will be integrated with other techniques as it will be introduced later.

### 5.3.2.3 Interaction Generality IG2 (global topology)

The third method is calculating the weights by using IG2 concept (Interaction Generality 2), [145, 151]. This algorithm explores the major five sub-graphs of network to get information about the global topology of the network. After collecting the five values for each interaction according to Fig.5.17, the principal component analysis (PCA) concept has been implemented regarding to Saito definition.

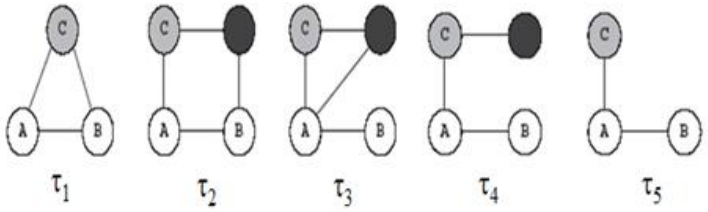


Fig.5.17 the most common five sub-graphs of network

By implementing PCA for the previous major five topologies of yeast proteins network, it has been found a range of varieties for IG2 values from -281 up to ~27 as shown in Table 5.9. The cells of t1-t5 are the values of the suggested five sub-graphs for yeast protein network. The average values of the collected data (t1-t5) are 1.415, 10.234, 24.826, 6.107, and 1.89 respectively. By determining threshold equal 19 as the margin of reliability to assume that IG2 values less than the threshold are more accurate (100%) than higher ones.

Table 5.9 IG2 values for yeast protein interactions

| P_name 1 | P_name 2 | PID_1 | PID_2 | t1 | t2 | t3 | t4 | t5 | IG2 |
|---|---|---|---|---|---|---|---|---|---|
| AAC1 | YHR005C-A | 1 | 1913 | 0 | 0 | 0 | 2 | 2 | 26.94071 |
| ANC1 | SNF5 | 1 | 3246 | 0 | 0 | 0 | 3 | 0 | 26.90991 |
| ANC1 | TAF25 | 1 | 3517 | 0 | 0 | 164 | 5 | 3 | -121.486 |
| ABP1 | ACT1 | 19 | 33 | 2 | 0 | 4 | 10 | 6 | 26.97287 |
| ABP1 | RVS167 | 19 | 2980 | 1 | 1 | 2 | 13 | 0 | 23.08996 |
| ABP1 | SRV2 | 19 | 3384 | 1 | 1 | 2 | 12 | 0 | 24.42532 |
| YER045C | PSE1 | 22 | 2483 | 0 | 0 | 2 | 3 | 1 | 24.44631 |
| ACC1 | DMC1 | 24 | 785 | 0 | 0 | 0 | 20 | 3 | 25.10544 |
| ACC1 | SNP1 | 24 | 3258 | 0 | 0 | 0 | 10 | 13 | 26.56783 |
| ACE2 | YNL157W | 25 | 5838 | 0 | 0 | 0 | 0 | 1 | 26.8268 |
| ACS2 | SNP1 | 32 | 3258 | 0 | 0 | 0 | 7 | 12 | 26.97778 |
| ACT1 | AIP1 | 33 | 84 | 0 | 0 | 8 | 10 | 6 | 26.88486 |
| ACT1 | BEM1 | 33 | 304 | 0 | 2 | 20 | 14 | 7 | 26.97287 |
| ACT1 | BNI1 | 33 | 333 | 2 | 2 | 4 | 14 | 7 | 19.55493 |
| ACT1 | BUD6 | 33 | 370 | 1 | 1 | 10 | 22 | 7 | 7.772698 |
| ACT1 | CAP2 | 33 | 407 | 0 | 0 | 8 | 10 | 6 | 22.16398 |
| ACT1 | COF1 | 33 | 568 | 0 | 0 | 8 | 10 | 5 | 17.03328 |
| ACT1 | FUS1 | 33 | 1065 | 0 | 0 | 8 | 10 | 6 | 19.55493 |
| ACT1 | GLK1 | 33 | 1164 | 0 | 0 | 8 | 7 | 9 | 19.55002 |
| ACT1 | IQG1 | 33 | 1470 | 0 | 2 | 8 | 13 | 6 | 19.55493 |
| ACT1 | LAS17 | 33 | 1583 | 0 | 0 | 8 | 9 | 7 | 19.63262 |
| ACT1 | MYO4 | 33 | 1983 | 0 | 0 | 8 | 10 | 5 | 18.64504 |

### 5.3.3 Protein function prediction by weights integration

Regarding the three previous methods of calculating the weights, number of high confidence interactions is collected compared to low confidence ones. After collecting the weights from the three previous methods (number of experimental methods, IG1 and IG2), new weights strategies can be created as average or PCA of these three values. Six different weights for each interaction are collected. As indicated in Table 5.10 interaction between proteins (AAC1 and YHR005C-A) has W0=1 which means equal weight for any interaction, W1=0.5 which means that it is identified by only one experimental method, W2=1 that means it has less than four leafs in IG1 (IG1<4), W3=0.5 that indicates that IG2 is more than 19, W4 is the average of the three weights which equal 0.66 ($1/3 \Sigma Wi$, i=1..3 ), and the last weight W5 (PCA of the three weights with threshold equal zero) is 0.5 that indicates that its value is more than zero. The previous example shows a weak interaction (edge) between the protein ID 1 (AAC1) and protein ID 1913 (YHR005C-A). Another example is high confidence (strong interaction) as shown in the second row, protein interaction (edge) between ANC1 and SNF5 where the weights are (1, 1, 1, 0.5, 0.83, 1) for W0-W5 respectively.

Relating to the main three measurements (number of experimental methods, IG1 and IG2), a lot of weights can be created as (applying AND/OR processes on the three weights or each weight has its coefficient according to its important role in determining the edge 0.35 - 0.2 - 0.4 respectively for W1, W2 and W3).

After collecting the levels and degree for each protein, and the six different weights, the neighbor counting method (frequencies of interaction partners having certain functions of interest) is implemented to predict the functions to all un-known proteins. Comparison between the new weights (W1-W5) and weight less technique W0 (edges with equal weights) algorithms is performed for proteins having interactions up to five. It is shown that for most selected new weights at specific specificity the sensitivity is higher than for weight less interactions. Fig.5.18-20 show the sensitivity and specificity for the three yeast protein function categories relating to the number of interactions k =2-5.

Table 5.10 new suggested weights for yeast protein interactions. w0- w5 are equal weights, experimental methods, IG1, IG2, average, and weights PCA.

| Protein A name | Protein B name | Protein A-ID | Protein B-ID | W0 | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|---|---|---|---|
| AAC1 | YHR005C-A | 1 | 1913 | 1 | 0.5 | 1 | 0.5 | 0.66 | 0.5 |
| ANC1 | SNF5 | 1 | 3246 | 1 | 1 | 1 | 0.5 | 0.83 | 1 |
| ANC1 | TAF25 | 1 | 3517 | 1 | 0.5 | 0.5 | 1 | 0.66 | 1 |
| ABP1 | ACT1 | 19 | 33 | 1 | 1 | 0.5 | 0.5 | 0.66 | 1 |
| ABP1 | RVS167 | 19 | 2980 | 1 | 1 | 1 | 0.5 | 0.83 | 1 |
| ABP1 | SRV2 | 19 | 3384 | 1 | 1 | 1 | 0.5 | 0.83 | 1 |
| YER045C | PSE1 | 22 | 2483 | 1 | 0.5 | 1 | 0.5 | 0.66 | 0.5 |
| ACC1 | DMC1 | 24 | 785 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| ACC1 | SNP1 | 24 | 3258 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| ACE2 | YNL157W | 25 | 5838 | 1 | 0.5 | 1 | 0.5 | 0.66 | 0.5 |
| ACS2 | SNP1 | 32 | 3258 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| ACT1 | AIP1 | 33 | 84 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 1 |
| ACT1 | BEM1 | 33 | 304 | 1 | 1 | 0.5 | 1 | 0.83 | 1 |
| ACT1 | BNI1 | 33 | 333 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| ACT1 | BUD6 | 33 | 370 | 1 | 0.5 | 0.5 | 1 | 0.66 | 1 |
| ACT1 | CAP2 | 33 | 407 | 1 | 1 | 0.5 | 0.5 | 0.66 | 1 |
| ACT1 | COF1 | 33 | 568 | 1 | 1 | 0.5 | 0.5 | 0.66 | 1 |
| ACT1 | FUS1 | 33 | 1065 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 1 |
| ACT1 | GLK1 | 33 | 1164 | 1 | 1 | 0.5 | 0.5 | 0.66 | 1 |
| ACT1 | IQG1 | 33 | 1470 | 1 | 1 | 0.5 | 1 | 0.83 | 1 |
| ACT1 | LAS17 | 33 | 1583 | 1 | 1 | 0.5 | 0.5 | 0.66 | 1 |
| ACT1 | MYO4 | 33 | 1983 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 1 |

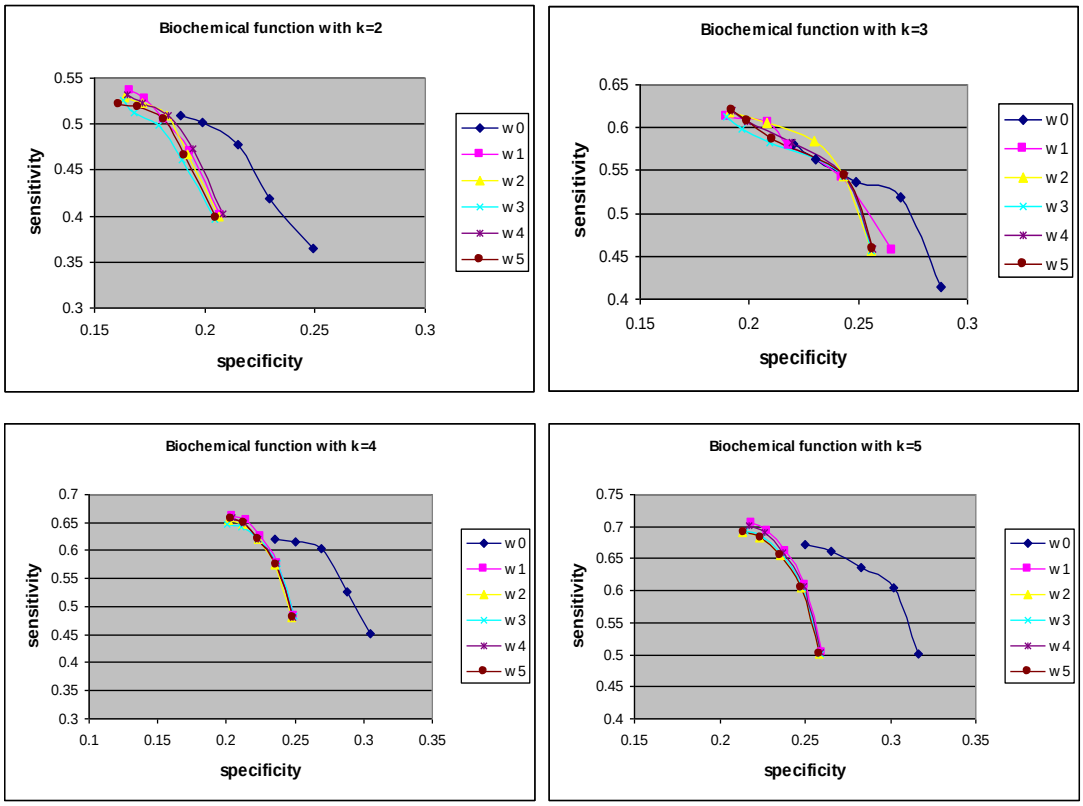


Fig.5.18 the sensitivity and specificity of the Biochemical function category for number of interactions k=2-5.

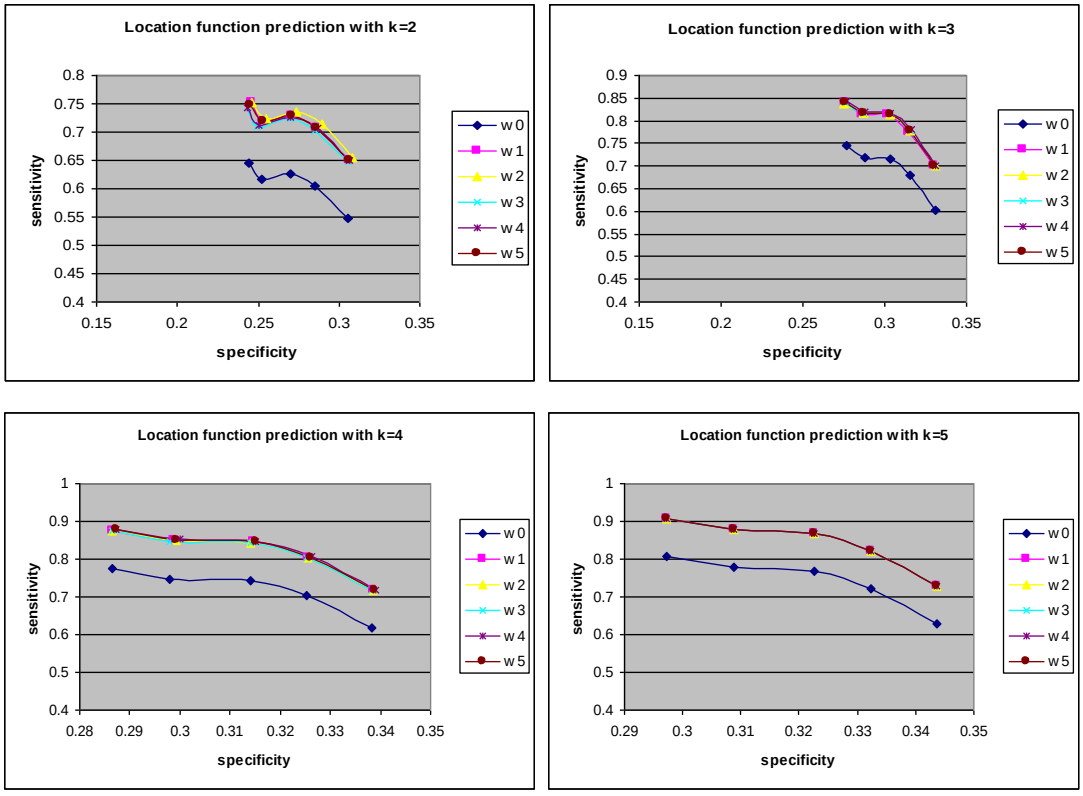


Fig.5.19 the sensitivity and specificity of the cell location function category for number of interactions k=2-5.

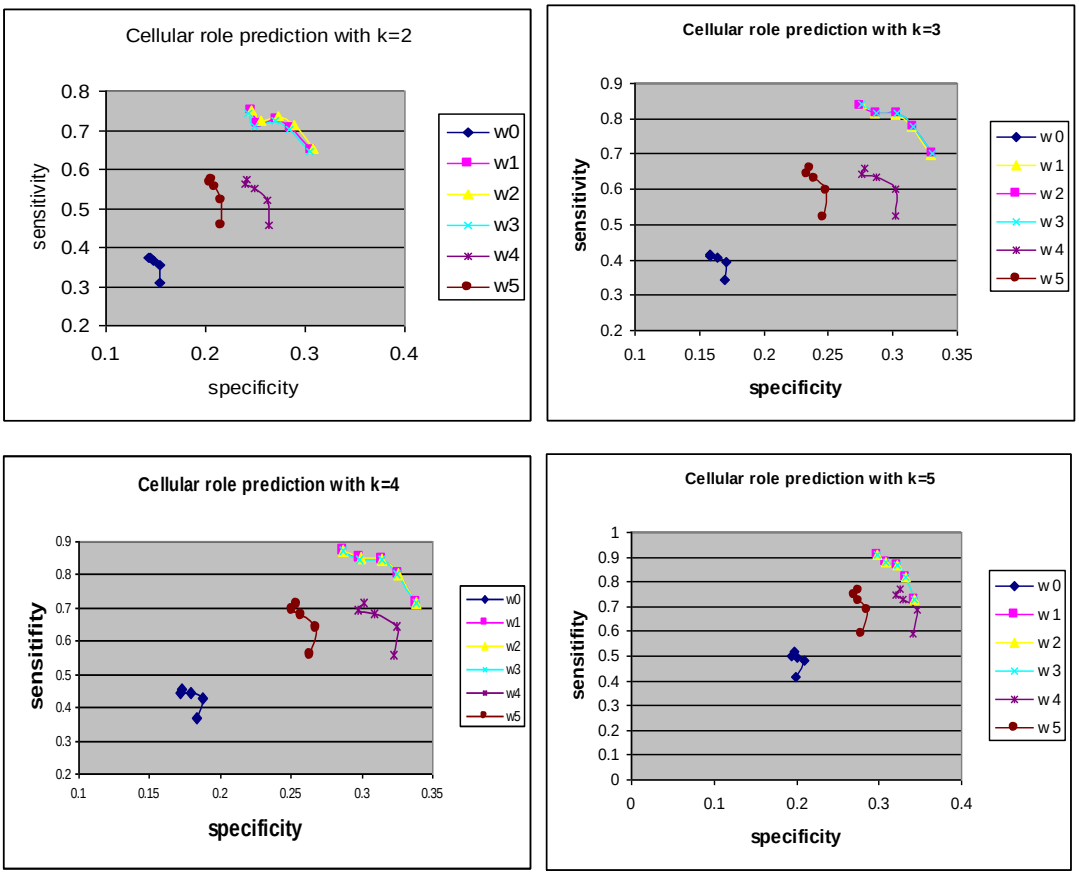


Fig.5.20 the sensitivity and specificity of the cellular role function category for number of interactions k=2-5.

## 5.4 Summary

In this chapter, a new weighting technique for PPI was introduced. This technique depended on the number of experimental identification methods furthermore the local and global topologies. As shown in the previous Figures, the results of the cellular role function category was enhanced and improved specially in the three suggested basic weights w1, w2, and w3. Although the average and PCA weights introduced better results than the weight less (unity weight) technique, they gave poor results comparing to the three weights (w1, w2, and w3).

Also it was noticed that, the sensitivity was improved with increasing number of interactions. In the second function category (cell location), all new weights were roughly overlapped and made shift in positive direction to improve the sensitivity and specificity. Also the larger number of interactions improved the results. But in the third functional category Fig.5.18 (Biochemical function), the weight less technique

was better than the suggested algorithms which meant that biochemical function category did not depend on the interactions and the protein neighbors and the computational methods failed in estimating the protein functions of un-known proteins from the surrounding neighbors. So we could conclude our work as:

1- The new weighting technique enhanced the sensitivity and specificity for two function categories (cell location and cellular role).
2- Increasing the number of interactions improved the sensitivity and specificity.
3- The technique did not reveal good results in biochemical function category which indicated, the estimation of this function category was very difficult using the neighbor data.

# Chapter 6

# Protein functions correlation based on overlapping proteins and cluster interactions

A lot of methods are developed to predict protein functions based on different information sources as protein structure, sequences, protein domain, protein-protein interactions, genetic interactions, and gene expression analysis. The accuracy of prediction can be enhanced by integrating multiple sources of information or collecting relations between the known functions. Recently, the researchers introduce different methods to determine the probability of protein function prediction using the information extracted from PPI. Although these techniques are promising, they lack the addressing of effective problems such as determining the relations between the protein functions "Usually, the relations between the protein functions do not be considered in protein function/interaction prediction".

In this chapter, a strategy for determining the relation between the protein functions is proposed. The technique is based on the overlapping number of proteins and interactions over protein clusters to determine the correlation between the sub-function categories as well as improve the protein function prediction process. Herein, we try to estimate values that represent the relation between each function and other functions within the same category depending on integrating the number of overlapping proteins [152], and cluster interactions [153]. The proposed method is applied to Yeast proteome (selected species) for the mentioned reasons in previous chapter (5.1). The revealed results are promising where the interactions are regarding the fact, the interacted proteins have common function (major function) (Brown et al. 2000; Eisen et al. 1998; Pavlidis et al. 2001).

## 6.1 Protein functions correlations

Because proteins are fundamental components for all living cells and each one Consists of sequences of Amino Acids (AAs) and performs a variety of biological tasks as Control physicochemical conditions inside the cell or transmit biological signals, so determining the functions for each protein is an important task. Since proteins work in complex system, they can bind to each other and interact. One of the most important problems in proteomics is protein complex isolation and mapping protein-protein interactions. The target of these processes is to understand the cell functions and to have basic idea about the relations between the proteins functions. Although estimating the protein functions correlations is very important, many researchers are interested in determining the individual protein functions not the relations between these functions. Protein functions may be predicted from sequences [6, 7], gene expression [1, 2], protein domains [8, 9], protein localizations [3, 4, 5], and protein-protein interactions [12- 18, 154].

In most cases, obtaining information about the relations between different functions is of great importance, since this would increase the certainty of protein function prediction.

As mentioned before, protein may have more than one function (up to 8 functions in Yeast Saccharomyces cerevisiae). Some of these functions may be correlated, anti correlated or independent. Protein may be seed (self dependent) or participate in certain function or in-complex (temporary or permanent).

If protein has certain function Fx_1 and it has another function as Fx_2 but it should not have function Fx_3, so it can be said that functions (Fx_1, Fx_2) have specific relations and functions (Fx_1, Fx_3) are anti correlated. The proposed technique is to estimate the relation between the protein functions based on the overlapping proteins and interactions over the protein clusters.

Table 6.1 Numbers of Annotated and un-annotated proteins for All Proteins Based on Three Functional categories

| Biochemical function | |
|---|---|
| Annotated | 3353 |
| Un-annotated | 3063 |
| Sub-cellular location | |
| Annotated | 3181 |
| Un-annotated | 3235 |
| Cellular role | |
| Annotated | 3894 |
| Un-annotated | 2522 |

Table 6.2 Yeast sub-function categories, function name and the number of proteins for each function.

| F_ID | Function category | Function name | # proteins |
|---|---|---|---|
| 1 | Cellular role | Aging | 39 |
| 2 | Cellular role | Amino-acid metabolism | 218 |
| 3 | Cellular role | Carbohydrate metabolism | 254 |
| 4 | Cellular role | Cell adhesion | 4 |
| 5 | Cellular role | Cell cycle control | 213 |
| 6 | Cellular role | Cell polarity | 216 |
| 7 | Cellular role | Cell stress | 331 |
| 8 | Cellular role | Cell structure | 120 |
| 9 | Cellular role | Cell wall maintenance | 184 |
| 10 | Cellular role | Chromatin/chromosome structure | 274 |
| 11 | Cellular role | Cyto kinesis | 40 |
| 12 | Cellular role | DNA repair | 154 |
| 1 | Cell location | Bud neck | 61 |
| 2 | Cell location | Cell ends | 6 |
| 3 | Cell location | Cell wall | 70 |
| 4 | Cell location | Centrosome/spindle pole body | 72 |
| 5 | Cell location | Contractile ring | 3 |
| 6 | Cell location | Cytoplasmic | 755 |
| 7 | Cell location | Cytoskeletal | 107 |
| 8 | Cell location | Endoplasmic reticulum | 225 |
| 9 | Cell location | Endosome/Endosomal vesicles | 36 |
| 10 | Cell location | Extracellular (excluding cell wall) | 34 |
| 1 | Biochemical | ATPase | 247 |
| 2 | Biochemical | ATP-binding cassette | 31 |
| 3 | Biochemical | Activator | 46 |
| 4 | Biochemical | Active "transporter," primary | 93 |
| 5 | Biochemical | Active "transporter," secondary | 201 |
| 6 | Biochemical | Adhesin/agglutinin | 7 |
| 7 | Biochemical | Anchor Protein | 13 |
| 8 | Biochemical | Channel [passive transporter] | 15 |
| 9 | Biochemical | Chaperones | 90 |
| 10 | Biochemical | Complex assembly protein | 76 |

The proposed method is applied to yeast proteome for the previous mentioned reasons. Regarding the yeast (Saccharomyces cerevisiae) studied species, it has three function categories; Cellular role functions (C.R) (contains 43 sub-function category), Cell location functions (C.L) (contains 29 sub-function category) and Bio-chemical functions (Bio-ch) (contains 57 sub-function category) as shown in Table 6.1 and Table 6.2. Yeast proteins are defined in the Yeast Proteome Databases. Each function category has certain number of proteins. And some of those proteins are involved in more than one sub-function category.

As shown in Table 6.1, each function category has its annotated and un-annotated numbers of proteins. For biochemical functions categories, there are 3353 annotated proteins and 3063 as un-annotated proteins. Roughly one third of the global numbers of all yeast proteins are un-annotated. In Table 6.2, each sub-function category has its ID and its number of proteins is shown. The basic idea of function correlation is coming from example as shown in Fig.6.1; group of proteins have the same functions which leads to there is a relation between those functions.

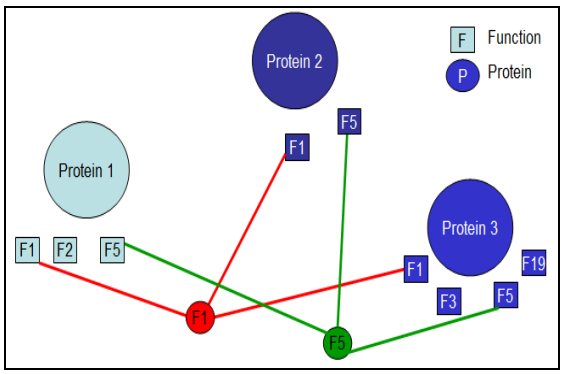


Fig.6.1 proteins have the same functions; correlation between functions

As shown in Fig.6.1, protein_1 has functions f1, f2, f5 and protein_2, protein_3 have functions f1, f5, and f1, f3, f5, f19 respectively. It is noticed that functions f1and f5 are common between the three proteins which leads to new idea “is there a correlation between those functions”. The answer depends on the natural of the functions. The previous methods try to estimate the relations regarding the interactions between the proteins [12]. Although this method try to estimate the relations through the interactions, the method does not introduce a clear view (relation probability) about

these relations and does not present the anti correlation property. Herein, the relations are calculated relating to three methods; cluster interactions, overlapping proteins, and integrating the two methods.

## 6.2 Protein cluster interaction

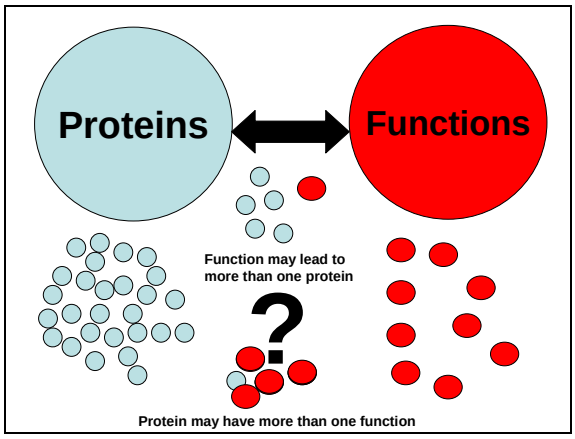


Fig.6.2 the relation between the proteins and functions

Proteins are acted as network. The simplest representation takes the form of a network graph consisting of nodes and edges. Proteins are represented as nodes in the graph and two proteins that interact physically are represented as adjacent nodes connected by an edge. Each group of proteins doing similar action called cluster (may have sequence similarity or similar function). So the network consists of groups of clusters. The clusters may be self assembled or have external interactions. The interactions may be real interactions (physical interactions between proteins in the two different groups or clusters) or from overlapping proteins (same proteins are found in the two clusters and have self interactions). Fig.6.2 shows the relations between the proteins and clusters; cluster has group of proteins and protein may be found into more than one cluster.

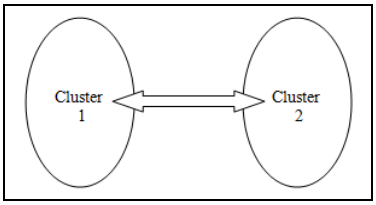


Fig.6.3 Two interacted clusters

As shown in Fig.6.3, two clusters can interact and these interactions are bidirectional. Herein, the relations between protein functions are estimated from these interactions. Table 6.3 shows the intersection numbers of cluster interactions in Yeast Biochemical function categories. It introduces the first 26 sub-functions (yellow colored), number of proteins (red cells) and interactions for each cluster respectively. Each crossed cell in table indicates the interaction number of the two indicating functions or clusters. For example proteins sub-function category_1 (cluster-1) interacts with proteins sub-function category_14 (cluster-14) by 49 interactions; cluster-1 contains 247 proteins, cluster-14 contains 283 proteins, and there are 49 interactions between those clusters as shown in Table 6.3. Although this method introduces a new concept for determining the relations between the protein functions, it can not indicate clear view for calculating these interactions. There are many drawbacks for this method as: 1)- the number of interactions is small comparing to the number of proteins in both clusters. 2)- some of these interactions may be considered as false positive interactions (from verification method). 3)- some interactions are considered as self interactions (17) in case cluster-1 and cluster-14 (proteins in both clusters). Also as shown in Table 6.3, some cells contain a few number of interactions as 3 interactions between clusters-1 and cluster-2 which can not be used as measure for these relations. On the other hand, there are some clusters have no interactions with any ones (self interactions) as clusters (24, 6). The reasons are, these clusters have group of proteins does not have the ability to interact with others or because the required function needs only one protein. Because the number of interactions between the crossed clusters is small, the threshold that determines the strength of correlation for cluster interactions is very difficult to be specified. The threshold can be estimated as certain number (specific) or as percentage of the number of proteins. In the proposed technique, the threshold is suggested as more than 10% of the number of proteins found in one of the two studied clusters. By applying this technique to yeast biochemical sub-function

categories, the green cells are considered as higher numbers of interactions between the two interacted clusters. As shown between cluster-1 and cluster-14, there are 49 interactions as ~20% of number of proteins.

Table 6.3 the relations between yeast protein functions based on the number of interactions

| | | 247 | 31 | 46 | 93 | 201 | 7 | 13 | 15 | 90 | 76 | 23 | 23 | 24 | 283 | 30 | 26 | 61 | 23 | 33 | 84 | 640 | 69 | 48 | 6 | 98 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| 247 | 1 | 33 | 3 | 8 | 9 | 1 | 0 | 0 | 0 | 11 | 6 | 11 | 1 | 2 | 49 | 1 | 0 | 5 | 3 | 4 | 7 | 53 | 2 | 2 | 0 | 3 | 3 |
| 31 | 2 | 3 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 5 | 1 | 0 | 0 | 0 | 0 |
| 46 | 3 | 8 | 1 | 6 | 0 | 1 | 0 | 1 | 0 | 1 | 4 | 2 | 1 | 0 | 8 | 7 | 1 | 3 | 1 | 1 | 3 | 11 | 11 | 0 | 0 | 0 | 0 |
| 93 | 4 | 9 | 2 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 2 | 0 | 0 | 0 | 0 |
| 201 | 5 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 1 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 0 |
| 7 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 7 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 15 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 90 | 9 | 11 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 26 | 1 | 2 | 4 | 0 | 0 | 0 | 0 | 2 | 5 | 19 | 0 | 15 | 4 | 2 | 0 | 0 | 0 |
| 76 | 10 | 6 | 0 | 4 | 6 | 0 | 0 | 0 | 0 | 1 | 7 | 0 | 1 | 0 | 9 | 1 | 1 | 1 | 0 | 0 | 1 | 7 | 3 | 0 | 0 | 0 | 0 |
| 23 | 11 | 11 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 5 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 |
| 23 | 12 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 4 | 1 | 0 | 0 | 0 |
| 24 | 13 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 283 | 14 | 49 | 1 | 8 | 1 | 3 | 0 | 0 | 1 | 0 | 9 | 3 | 1 | 3 | 60 | 0 | 0 | 2 | 0 | 0 | 5 | 36 | 9 | 1 | 0 | 3 | 3 |
| 30 | 15 | 1 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 16 | 2 | 3 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| 26 | 16 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1 | 3 | 0 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 1 |
| 61 | 17 | 5 | 2 | 3 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 2 | 3 | 3 | 5 | 17 | 2 | 2 | 15 | 4 | 2 | 0 | 0 | 0 |
| 23 | 18 | 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 3 | 5 | 1 | 21 | 4 | 0 | 0 | 0 | 0 |
| 33 | 19 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 5 | 6 | 0 | 5 | 2 | 1 | 0 | 0 | 0 |
| 84 | 20 | 7 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 5 | 0 | 0 | 2 | 1 | 0 | 2 | 12 | 0 | 1 | 0 | 0 | 0 |
| 640 | 21 | 53 | 5 | 11 | 9 | 2 | 0 | 1 | 0 | 15 | 7 | 10 | 0 | 1 | 36 | 4 | 3 | 15 | 21 | 5 | 12 | 79 | 9 | 8 | 0 | 2 | 0 |
| 69 | 22 | 2 | 1 | 11 | 2 | 2 | 0 | 0 | 0 | 4 | 3 | 0 | 4 | 0 | 9 | 0 | 1 | 4 | 4 | 2 | 0 | 9 | 6 | 1 | 0 | 0 | 1 |
| 48 | 23 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 8 | 1 | 1 | 0 | 1 | 0 |
| 6 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 98 | 25 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 |
| 90 | 26 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 5 |

Also it can be noted that, some values can satisfy the suggested threshold as number of intractions between clusters 1& 21, clusters 9, 19, and clusters 18 & 21 as 53 (21%),19 (58%) and 21(91%) respectively.

## 6.3 Function categories and overlapping proteins

The second technique for estimating the function correlation is the overlapping proteins. This technique is based on the number of proteins found in the two studied clusters (overlapped number of proteins). If there are some proteins in the two studied clusters, it can be said that there is a correlation between those clusters. By applying the proposed technique on the yeast function category (Biochemical), it has found a lot of direct relations between the sub- function categories as shown in Table 6.4 and Table 6.5. Method collects all sub-function categories on the two axes as shown in Table 6.4 and puts the number of overlapped proteins in each cross section cell (square) and compares this number (cell) with the smaller number of the two surrounding sub- categories (red cells). As shown the first top left cell indicates the sub-function category number one and contains 247 that mean the first sub-function category contains 247 proteins. The rest cells in the first row indicate the overlapping number of proteins between the first sub-function and residuals of the same sub-functions category according to the column number. Percentage between each cell number and the smaller number of the two surrounding sub-function categories is calculated. By determining threshold equal to 0.85, direct relationships between the two sub-function categories is estimated.

As illustrated in Table 6.5; the method can determine 9 direct relationships among 57 functions in biochemical sub-function categories. It is noticed that if the threshold value is decreased to 0.72, the direct relations between the sub-function categories will increase (22 values). The red cells (diagonal) show the number of proteins in each sub-function and the green cells show the overlapping cross section for high correlated functions. The direct relations between the functions mean correlation between those functions.

For example if protein has function **x**, it should have function **y** because there is high correlations between function x and function y. as shown in Table 6.5, there are 9 direct relations (green cells in Table 6.4).

Table 6.4 the overlapping number of proteins over the Yeast biochemical function categories

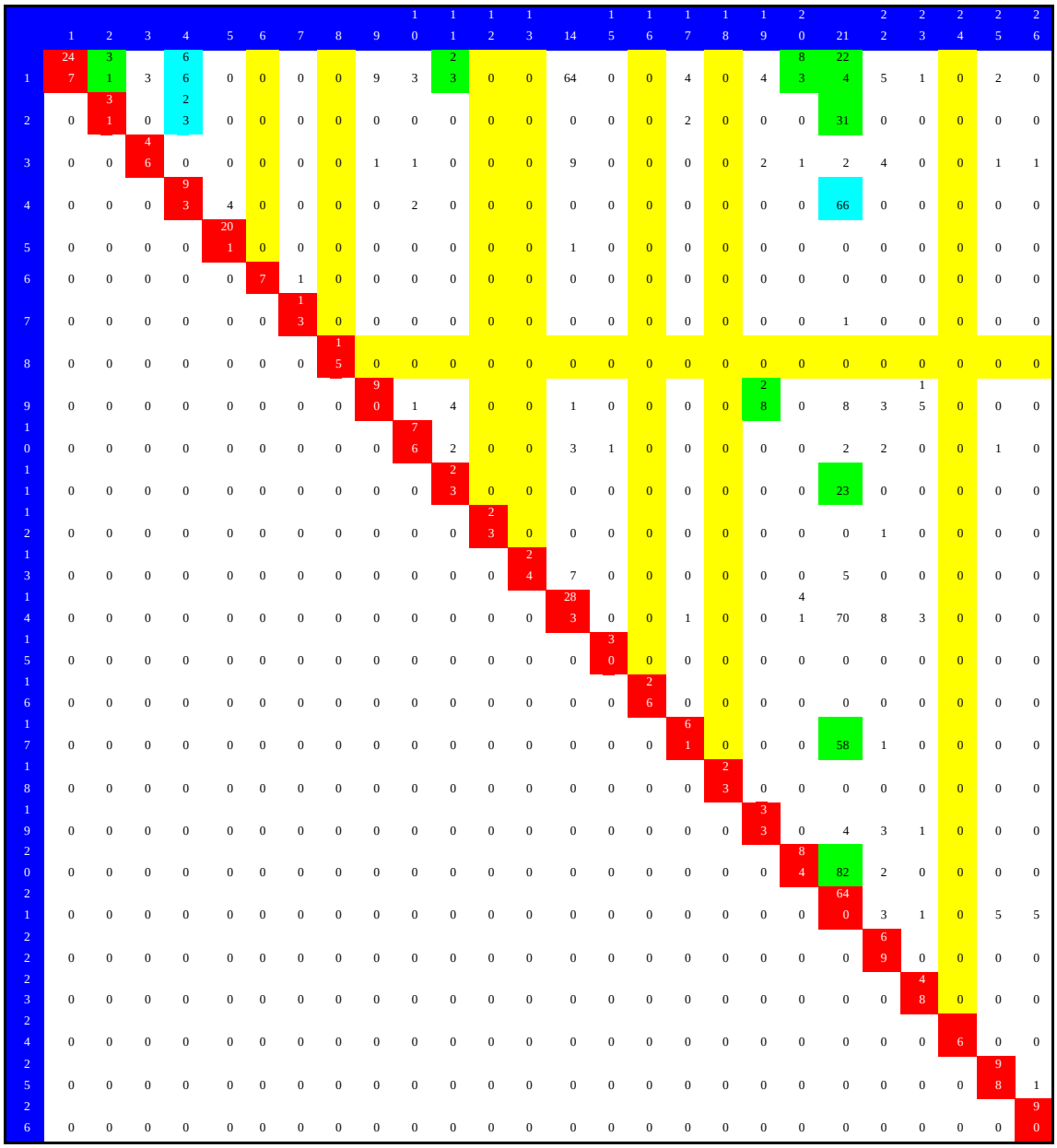

Table 6.5 the direct relations over Yeast Biochemical sub-function categories when the threshold is greater than 0.85.

| Fx_1 ID | Fx_2 ID | Fx_1 name | Fx_2 name | Score |
|---|---|---|---|---|
| 1 | 2 | ATPase | ATP-binding | 1 |
| 1 | 11 | ATPase | Conserved ATP | 1 |
| 1 | 20 | ATPase | Helicase | 0.99 |
| 1 | 21 | ATPase | Hydrolase | 0.91 |
| 2 | 21 | ATP-binding | Hydrolase | 1 |
| 9 | 19 | Chaperones | Heat shock protein | 0.85 |

| | | | | |
|---|---|---|---|---|
| 11 | 21 | Conserved ATP | Hydrolase | 1 |
| 17 | 21 | GTP-binding protein/GTPase | Hydrolase | 0.95 |
| 20 | 21 | Helicase | Hydrolase | 0.98 |

Because the correlation score between functions (1, and 2) is 1, proteins have sub-function category_2 (ATP-binding) will have by default sub-function category_1 (ATPase) as shown in Table 6.5. And each protein has sub-function category_11 (conserved ATP) will have sub-function category_1 (ATPase).

The scores between four relations are ones which means, all proteins have the first function they will have the second function. If the threshold decreases into 0.72, a lot of direct relations can be created (blue cells (22)) as relation between sub-function category_1 (ATPase) and sub-function category_4 (transporter) which has scored 0.72. These scores are collected and integrated with scores of cluster interactions. The direct relations between the studied functions are acted as shown in Fig.6.4.

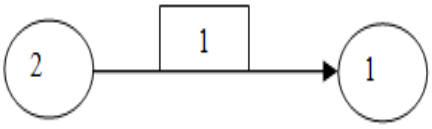


Fig.6.4 relations between the Biochemical sub-function category_2 towards the sub-function category_1

## 6.4 Overlapping and Interaction integration

In this section, the scores of overlapping and cluster interactions are integrated to determine the relation between the functions either positive (to participate in the same functions) or negative (anti correlations, if protein has one function, it should not have the other one) or independent (there is no relations between the studied functions).

Herein, if the score of overlapping proteins is more than the threshold (0.85), it will be positive otherwise will be negative. Also for the cluster interactions, if its score has

more than 10% of the numbers found in one of the two clusters, it will be positive and other wise will be negative.

The function relation technique is integrated with the traditional method of protein function prediction (neighbor counting method). Improved results are gained than previous. As known in neighborhood method, it finds the neighbor proteins and gets their assigned functions and the frequencies of occurrence of these functions. Then, these functions are arranged in descending order according to their frequencies. The first *k* functions are considered and assigned to the un-annotated protein. The authors in [18] used this technique with *k* equals to 3. By applying the proposed technique on the yeast function categories, the results are as shown in Table 6.6 and Table 6.7. The algorithm shows increasing number of true positive (TP) and decreasing the true negative (TN) and false positive (FP).

Table 6.6 shows each yeast Biochemical function category and its results. Function category_1 has 247 proteins, 47 of them identified as TP and the rest (200) identified as TN and there are 141 proteins identified as FP. On the other hand function category_2 has 2 proteins as (TP), 29 proteins as (TN) and 15 proteins as (FP). Also function category_11 has 6 proteins as (TP), 17 proteins as (TN) and 10 proteins as (FP). It can be noted that the integrated algorithm enhanced (increased) the numbers of TP and decreased the numbers of TN and FP. As shown in Table 6.7 the integration between function_1 and Function_2 (positive overlapping and positive interactions) shows the same numbers of function_2 (least one). And integration between function_1 and function_11 has 6 proteins as (TP, the same number of function_11 true positive) and decreases the number of FP (141 & 10 → 7). The integration process has been divided into for cases regarding to the states of overlapping and interactions. The collected cases are 1)- Positive overlapping & positive interactions (the score of overlapping more than the threshold (0.85) and the number of interactions are more than 10% of the minimum number of proteins in one category), 2)- positive overlapping and negative interactions, 3)- negative overlapping and positive interactions, and 4)- negative overlapping and negative interactions. It can be noted that in case of (positive & positive), enhanced results has been gained specially in increasing the TP and decreasing the TN and FP. Although the number of TP is small relating to one function of them, but it is very accurate and equal to the minimum number over the two functions. It is very clear that the numbers of TN and

FP are decreased as in cases functions (1-21) which they have FP equal to 141 and 395 respectively and now it is 74. The results will be poor when the two scores are negative which reflects the effect of overlapping and interactions. But when one of them is positive and the other is negative, it has variety in results. The negative of interaction score fixes the number of TN and the negative of overlapping score increases the TN. It can be indicated that, the overlapping numbers of proteins and the number of interactions has affected the protein function prediction process in positive way. And the relations between the protein functions have enhanced the degree of confidence.

Table 6.6 yeast biochemical functions, estimated numbers of proteins as true positive (TP), true negative (TN), and false positive (FP)

| Function category | TP | TN | FP |
|---|---|---|---|
| 1 | 47 | 200 | 141 |
| 2 | 2 | 29 | 15 |
| 4 | 9 | 84 | 18 |
| 9 | 25 | 65 | 29 |
| 11 | 6 | 17 | 10 |
| 14 | 67 | 216 | 194 |
| 18 | 4 | 19 | 36 |
| 19 | 7 | 26 | 16 |
| 20 | 4 | 80 | 45 |
| 21 | 91 | 549 | 395 |

Table 6.7 an integrated algorithm relating to the overlapping number of proteins and number of interactions according to the determined thresholds

| F:x-y | Overlapping | | Interactions | | TP | TN | FP |
|---|---|---|---|---|---|---|---|
| 1-2 | 31/31 | + | 3/31 | (~)+ | 2 | 29 | 15 |
| 1-11 | 23/23 | + | 11/23 | + | 6 | 17 | 7 |
| 2-21 | 31/31 | + | 5/31 | + | 2 | 29 | 13 |
| 11-21 | 23/23 | + | 10/23 | + | 6 | 17 | 7 |
| 1-20 | 83/84 | + | 7/84 | - | 4 | 80 | 26 |
| 1-21 | 224/247 | + | 53/247 | + | 29 | 218 | 74 |
| 20-21 | 82/84 | + | 12/84 | + | 4 | 80 | 26 |
| 1-4 | 66/99 | - | 9/93 | - | 4 | 89 | 14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4-21 | 66/93 | - | 9/93 | - | 4 | 89 | 13 |
| 1-14 | 64/247 | - | 49/247 | + | 12 | 235 | 35 |
| 18-21 | 0/23 | - | 21/23 | + | 0 | 23 | 12 |

In this chapter, an integrated technique is introduced to estimate the correlations or relations between yeast protein functions. The technique depends on the overlapping numbers of proteins as well as number of interactions over the protein clusters. By applying the proposed algorithm to the collected data, the results are improved; reducing the number of true negative and false positive furthermore increasing the true positive results.

The results are good when the two measures are positive. Although the number of interactions is important for enhancement the results but the overlapping number is more critical. In protein function prediction problem, the effect of the function correlations has been indicated and the results are better than the absolute method (neighbor counting method without function correlation).

## 6.5 Function Relations

It can be noted that the relations over the protein functions is divided into two classes: direct and indirect relations.

### 6.5.1 Direct relationships

By applying the same concept of overlapping number of proteins, the direct relations can be estimated through the threshold (over threshold). As illustrated in Table 6.4, method can determine 9 direct relations (more than the threshold (0.85)) between 57 functions in Bio-chemical sub-function categories and 4, 7 direct relations in cellular role and cell location with thresholds 0.72 and 0.79 respectively. It can be noted that the threshold value is a big number to express the correlation between the two sub-function categories. If the threshold decreases till 0.70 in biochemical function, there are more than 22 direct relations.

The technique indicates that there is a direct relationship between sub-function category 1 (ATPase) and the second function (ATP-binding cassette) with weight equal 100% towards sub-function category one. It means that if protein has function category 2 it should have function category one. In the fourth row of Table 6.5, the weight is equal to the 0.9 that means 90% of the proteins which have sub-function category 21 (Hydrolase) have sub-function category 1 (ATPase). This technique converts the undirected graph of physical interactions between the proteins (protein interaction network) into directed graph between the sub-function categories which have been taken into consideration to enhance the accuracy of protein function prediction. As shown in Fig.6.5, direct relation between sub-function 11 towards sub-function 1 by 100% which means that, any protein will have sub-function 11 should have sub-function 1.


Fig.6.5 directed relation between the two sub-function categories 11, 1 and its weight equal (100%).

Because the directed relations cannot give a wide screen for all relations between the sub-function categories, so the study has discussed the indirect relationships and anti-correlations between the proteins sub-functions category.

### 6.5.2 Indirect relationships

Because the few number of direct relations, the study puts some conditions to estimate the indirect relationships and uncorrelated functions. If there are three sub - functions categories A, B and C and each function contains number of proteins X1, X2 and X3 respectively and

If $A \cap B = n1$ proteins,

$A \cap C = n2$ proteins, and

$B \cap C = n3$ proteins

The next combinations can be collected:

a- n1 = 0 and / or n2 = 0

b- n1 = n2

1- (n1=n2=n3)

2- (n1=n2 and n3=0)

3- (n1=n2 ≠n3)

c- n1 ≠ n2

**a. [n1= 0 and/or n2=0]**

If the number of proteins in the cross section between two sub-function category is zero (no overlapped proteins are found) that leads to *uncertainty* case**.** We cannot say that there is anti correlation between these two sub-function categories which have been intersected by zero. So, it should calculate the indirect relationship between two or more function categories if they interact in the same number of proteins for the same function category.

**b. [n1=n2]**

**1- [n1= n2 = n3]**

The same proteins found in the three sub-function categories *(A∩B∩C=n1=n2=n3)* that leads to there is a *correlation* between B, C toward A and so on. If number m of proteins have functions B and C, they should have function A. as shown in Fig.6.6. If protein has the two sub-functions category it should have the third one or by the statistical view p (A\B, C) = 1 the probability of protein to have sub-function category A conditional sub-function categories B and C equal the unity as shown in Fig.6.6. Protein has function A (first function) and given (conditional) function B (second function), it should have the third one function C.

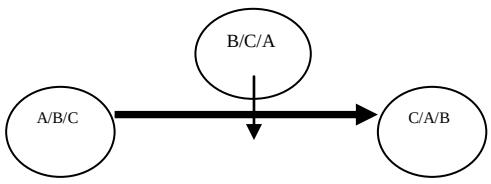


Fig.6.6 the conditional relationship between the sub-function categories.

**2- [n1= n2 & n3=0]**

If protein has function A and B it should not have function C as shown in Fig.6.7 or in the statistical view P (B\A, C) = P (C\A, B) = 0 the probability of protein to have the third function conditional the two functions is zero.

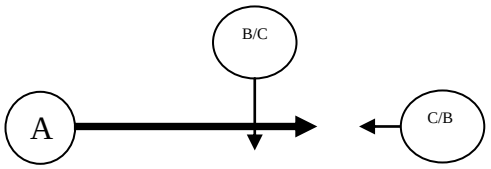


Fig.6.7 anti correlation between the two sub-functions category B, C given sub-function category A.

**3- [n1= n2 ≠ n3]**

If the protein is in two sub-function categories and is not in the third so it leads to *uncertainty* case.

**c. [n1< > n2]**

If the number of proteins is not the same in the two sub-function categories, it should have three combinations condition:

**1- [n1< n2] & [A ∩ B] = 0**

There is no intersection between the proteins of the two sub-function categories that leads to *uncertainty* case.

**2- [n1< n2] & [A ∩ B] < n1**

n1 is fraction of n2 (some proteins of the first sub-function category are found in the second sub-function category) that also leads to *uncertainty*.

**3- [n1< n2]**

All n1 are found in n2 that leads to correlation between the two studied functions. Protein has function category A as shown in Fig.6.8 has correlation towards C given function B.

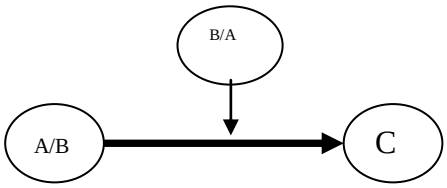


Fig.6.8 protein has function A/B and has conditional function B/A respectively it will have function C.

### 6.5.3 Protein functions integrations

The function relation technique has integrated with the traditional methods of protein function prediction. And improved results have been collected than previous. As known in neighborhood counting method, the function with high frequency has been taken as first one then the others (less frequency) without taking the relation between the functions into consideration. Now, the function has accepted by the highest frequency and bigger number of correlation or relations. For example sub-function category 1 (ATPase) contains 247 proteins (yeast protein function database) and has directed relation with sub-function category 2 (ATP-binding cassette) regarding to our technique with weight 100%. After applying the combination between neighbor counting method and the studied technique, it has been found that the results are as shown in Table 6.7, It can be noted that the numbers of the false positive and true negative in combination technique are less than in single mode (neighbor method) in addition the number of the true positive is roughly the same or less (the difference in the values according to the weight value). The studied technique has clear addition on the accuracy of the prediction.

By applying the Chi-square method to get the correlation between the two sub-function categories 1 and 2, it has been found that for each prediction, the score values of the two functions are the highest scores. An illustrated example; if protein has function 4 as a basic function category and has direct relationship with function category 17 by weight 0.75 and indirect relationship with function categories 6, 7, 28 respectively the technique can illustrate the relationships as shown in Fig.6.9.

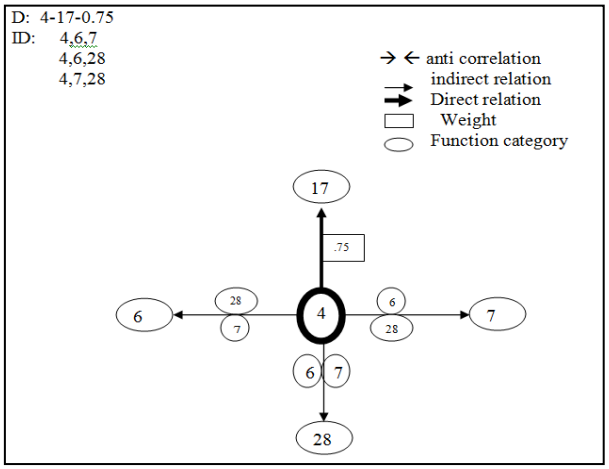


Fig.6.9 an example for complete combination between sub-function category 4 and the rest functions of the cellular role category in Yeast.

## 6.6 Summary

In this chapter, a novel technique was introduced to get the relations (correlation) between the functions in the same function category in yeast (Saccharomyces cerevisiae). In this technique, the overlapping numbers and cluster interactions were considered in determining these relations. By applying the technique on all the functions categories and mixing the results with any method of protein function predictions as neighborhood and Chi-square methods, enhanced results and increasing of the accuracy were achieved. The conversion of undirected interactions between the proteins into directed interactions between the clusters was very critical point for determining the function correlation. Also producing a global figure for the function relations is performed.

# Chapter 7

# Yeast Protein Function Motif (Signature) Extraction Based on Sequence Alignment

Protein function prediction is one of the most important problems in the field of proteomics since it leads to determining cell functions. Since proteome is divided into clusters, each cluster (group of proteins) should have common characteristics. One of these characteristics is to have the same functions. In this study we try to extract motifs for each sub-function category in yeast proteins. The technique is based on applying multiple sequence alignment (MSA) to all yeast protein function categories. The protein sequences are collected from different data sources as DIP, PIR, and SWISS PROT.

BIO- CLC program is used to apply the sequence alignment. Threshold is determined for every protein function category to indicate the most common frequent amino acids to be as feature for this category. After implementing the algorithm, sequence is verified with some proteins have the correct functions and the gained results are good. The technique is considered as verification method for protein function prediction.

## 7.1 Protein sequence Overview

Proteins are macromolecules response for doing many functions. They are the main building blocks and functional molecules of the cell. Every 3 bases of RNA (codon) correspond to one amino acid which is arranged to build the protein. Proteins are consisted of sequence of amino acids which are the basic units of structure. When the 20 amino acids (natural components) are sequenced in different numbers and different orders, infinite number of proteins can be created. If the length of amino acids is more than 40, it is called protein otherwise called multi peptide.

The sequence of amino acids is response for the folding shape of protein (3D structure) as well as its main functions. In particular, proteins transmit regulatory signals throughout the cell, catalyze a tremendous number of chemical reactions, and are critical for the stability of numerous cellular structures. As known, each group of proteins having specific characteristic is called cluster (group). One example for these clusters is the similarity in doing specific function. Many methods have been developed to predict the protein functions as analyzing gene expression patterns, phylogenetic profiles, protein sequences, protein domains, and protein interaction networks, and estimated correlation. Since most of the prediction methods depend on the protein sequences and the fact that if two proteins have similar sequences, they may have the same function, the protein sequences is our interest in this study.

Each protein function category (cluster) has group of proteins is defined and their protein sequences are collected. Many data sources as DIP, PIR, and SWISS PROT are used to get these sequences that because protein may have more than one name and its sequence is not found in one data source. Accurate multiple sequence alignment technique is performed to those sequences using Bio-CLC program.

So in this chapter, we introduce technique using multiple sequence alignment to extract certain motif for each sub-function category. The technique is applied to Yeast protein sequences (complete sequence genome) and extracted consensuses are collected and considered as feature (signature) for each sub-function category. By collecting these extracted motifs (features), protein function prediction process is

verified. This chapter is organized as follows; the proposed algorithm is explained in next section following by the results of the work together with their discussion.

## 7.2 Motif extraction

Function category motif extraction is one of the most important tasks in proteomics that because it is considered as a feature to identify this category. Protein sequence search in BLAST or NCBI is considered one of the multiple diverse sources in identification the proteins and determining their functions [155].

In this study, an integrated method is used among different data sources to get the annotated protein sequences. The sequences which have same sub-function category are collected and multiple sequence alignment is used to extract specific motif (consensus) for each sub-function category.

### 7.2.1 Protein sequence collection

Although BLAST and NCBI web sites are used to get protein data, it is very exhaustive process to gain the protein sequences manually. A group of databases as DIP, PIR, SWISS-PROT, and MIPS are integrated to collect these sequences. This integration is performed, since all annotated proteins are not found in one database. Although DIP (Database of Interacting Proteins) is the most famous data source used to get the sequences of yeast proteins, it misses for some proteins which collected from other databases.

As shown in Table 7.1, a sample of protein names and parts of their sequences is indicated. It can be noted that, the protein names are gene names which means the DIP databases deal with gene names not the international name (accession / standard name). So comparison between the protein names and data sources is performed to identify all data about the protein (it is considered as one of the challenges of yeast naming as mentioned in section 5.2). It can be shown in Table 7.2, some missed cells which loss the corresponding names for these proteins. As example; protein code DIP:

239N equal Q27272 code in SWISS PROT equal A49067 in PIR database. But DIP 772 did not have corresponding code in SWISS PROT.

Table 7.2 indicates the relations over different data sources for yeast proteins. Relating to the distinguished names of proteins (Gene name, Accession number, and ORF) and the different databases, the standard core for protein should be given (Accession number).

Table 7.1 Sample of protein names and their sequences from DIP database.

| Protein Name | Protein sequence |
|---|---|
| BNI1 | MLKNSGSKHSNSKESHSNSSSGIFQNLKRLANSNATNSNTGSPTYASQQQHSPVGNEVSTSPASSSS…… |
| BNI4 | MSDSISDSKSSELLNSTFYSSTSINTLDHARTFRNSLILKEISDQSLNSSIKPCESVLDRDVESSVLQ…… |
| BNI5 | MGLDQDKIKKRLSQIEIDINQMNQMIDENLQLVEPAEDEAVEDNVKDTGVVDAVKVAETALFSGND…. |
| BUD2 | MSSNNEPAQSRTSYFKLNEFLSNVKHYKNTFKGEIQWCNNLSLNDWKTHYLQITSTGALTHSIDELTA…. |
| BUD3 | MEKDLSSLYSEKKDKENDETLFNIKLSKSVVETTPLNGHSLFDDDKSLSDWTDNVFTQSVFYHGSDD… |
| BUD4 | MAQDIDKLARDEEKPVKLSSSPLKFTLKSTQPLLSYPESPIHRSSIEIETNYDDEDEEEEDAYTCLTQS…. |
| BUD5 | MRTAVPQLLEATACVSRECPLVKRSQDIKRARKRLLSDWYRLGADANMDAVLLVVNSAWRFLAVWR… |
| BUD6 | MKMAVDDPTYGTPKIKRTASSSSSIETTVTKLLMSTKHLLQVLTQWSKGTTSGRLVSDAYVQLGNDF… |

Table 7.2 Different data sources for protein names and their codes

| DIP interaction | DIP code | SP code | PIR code | GI code | DIP code | SP code | PIR code | GI code |
|---|---|---|---|---|---|---|---|---|
| DIP:193E | DIP:239N | SWP:Q27272 | PIR:A49067 | GI:1079142 | DIP:368N | SWP:P04637 | PIR:DNHU53 | GI:8400738 |
| DIP:196E | DIP:237N | SWP:P47825 | PIR:A48184 | GI:477148 | DIP:36N | SWP:P08047 | PIR:A29635 | GI:88887 |
| DIP:199E | DIP:772N | | PIR:S41672 | GI:1085161 | DIP:368N | SWP:P04637 | PIR:DNHU53 | GI:8400738 |
| DIP:207E | DIP:387N | SWP:P12428 | PIR:FYFFB | GI:72497 | DIP:388N | SWP:P10090 | PIR:FYFFW | GI:17136592 |
| DIP:229E | DIP:237N | SWP:P47825 | PIR:A48184 | GI:477148 | DIP:570N | SWP:P03254 | PIR:Q2AD2 | GI:74182 |
| DIP:271E | DIP:121N | SWP:P19538 | PIR:A38926 | GI:24638496 | DIP:754N | SWP:P41044 | PIR:S37695 | GI:17136674 |
| DIP:272E | DIP:492N | | PIR:JC4234 | GI:17137760 | DIP:121N | SWP:P19538 | PIR:A38926 | GI:24638496 |
| DIP:273E | DIP:45N | | PIR:A31225 | GI:24639671 | DIP:526N | | PIR:JU0092 | GI:17136554 |
| DIP:274E | DIP:54N | SWP:P13677 | PIR:A32392 | GI:17136716 | DIP:526N | | PIR:JU0092 | GI:17136554 |
| DIP:275E | DIP:769N | | PIR:S40691 | GI:2119474 | DIP:526N | | PIR:JU0092 | GI:17136554 |
| DIP:276E | DIP:537N | SWP:P07181 | PIR:MCFF | GI:17647231 | DIP:526N | | PIR:JU0092 | GI:17136554 |
| DIP:342E | DIP:187N | SWP:P10083 | PIR:A43731 | GI:17136654 | DIP:73N | SWP:P18491 | PIR:A34688 | GI:24654863 |
| DIP:344E | DIP:40N | SWP:P16371 | PIR:A30047 | GI:24650241 | DIP:637N | | PIR:S06956 | GI:85137 |
| DIP:345E | DIP:325N | SWP:P10084 | PIR:B43731 | GI:17136616 | DIP:356N | SWP:Q01068 | PIR:D46177 | GI:24650229 |

Yeast protein functions are divided into three function categories: Bio-chemical functions (contains 57 sub-function categories), Cellular role functions (contains 43 sub-function categories), and Cell location (contains 29 sub-function categories). The study collects all possible protein sequences related to every specific sub-function category in one place. The average of collected protein sequences is 41% of the total number of protein sequences. An example for the collected number of sequences, Biochemical ATPase sub-function category which has 247 proteins, 112 protein sequences are collected. But for Biochemical protein motor sub-category which has 17 proteins, 12 protein sequences are collected.

### 7.2.2 Multiple Sequence Alignment

Although there are many methods used in motif extraction as Deterministic algorithm (match or mismatch), Probabilistic algorithm, Combination between Deterministic and Probabilistic presentation and M-PST (mismatch probabilistic suffix tree) [58], the multiple sequence alignment has produced good results. Also it has been used in determining the interactions protein [60] and probabilistic approach [156].

In this study, CLC BIO package is used to perform MSA (multiple sequence alignment) for all collected protein sequences that have the same function. As shown in Fig.7.5, the alignment process after applying MSA to the FASTA format protein sequences is indicated.

It is noticed that, all collected proteins should be in the same sequence format to have the multiple sequence alignment. Some data sources have other formats as GCG, Staden, EMBL, Clustal, MSF, Gen-bank, RSF, and FASTA formats.

As shown in Fig.7.1:7.4, some examples of these formats are introduced. So the collected sequences from different data sources should be converted into FASTA format.

!!NA_SEQUENCE 1.0 test.seq Length: 5390 April 22, 1999 13:50 Type: N Check: 8167

.. 1 ttatataaaa aatgctgaaa acaggatcaa ggaggaagat ttaaatatag 51 atataatata

tgggaagaaa cataaaaacg aaataagaac agctaaatat

Fig.7.1 GCG sample format

The hallmarks of a GCG formatted sequence are: it begins with the line (all uppercase) *!!NA_MULTIPLE_ALIGNMENT* 1.0 for nucleic acid sequences or *!AA_MULTIPLE_ALIGNMENT* 1.0 for amino acid sequences. A description line which contains informative text describing what is in the file. A dividing line which contains the number of bases or residues in the sequence, when the file was created, and importantly, two dots (..) which act as a divider between the descriptive information and the following sequence information. The programs of the Staden suite of biological analysis software accept sequences in staden format as shown in Fig.7.2. A typical staden format file: Staden formatted sequence files contain the sequence (Ns, AAs) and nothing else

GGTACGTAGTAGCTGCTGCTACGTGCGCTAGCTAGTACGTCATTA
CGACGTAGATGCTAGCTGACTCGATGCAGTACGTAGTAGCTGCTG
CTACGTGCGCTAGCTAGTACGTCACGACGTAGATGCTAGCTGACT CGATGC

Fig.7.2 Staden format sequence

Another format as shown in Fig.7.3 is Gen-bank format which is considered as one of the most important used sequences in protein files.

```
     CDS             1..735
                     /function="putative transcriptional activator of the
                     carbapenem biosynthetic genes"
                     /note="LuxR homolog"
                     /codon_start=1
                     /transl_table=11
                     /product="CarR"
                     /protein_id="AAC38168.1"
                     /db_xref="GI:2873371"
                     /translation="MNKEISYFIERKLKAYGNVLFAYFMMDKSSLSNPVFISNYPQKC
                     IDTYIDNKLFINDPVIHYSLKRVTPFSWDDNDLAVLRSENEDVAMYLREHDITVGYTF
                     VLHDHDNNLAILTIANNDEKNDFEDFIKNRENDLQMLLVTTHEKAMKHKHFVKGKTAP
                     LDCLQSALITPRETEVLFLVSRGNTYKEVSRTLGISEATVKFHINNSVRKLNVINSRH
                     AISKALELNLFRAFTGSLMTRKLVAI"
ORIGIN
        1 atgaataaag agatcagtta ttttatagaa agaaagctaa aggcctatgg gaatgtttta
       61 ttcgcttact ttatgatgga taaatcttct ttatcaaatc ctgtttttat ttctaactac
      121 ccccaaaaat gtattgatac ctatattgat aataaacttt ttatcaatga tcctgttata
      181 cattactctt taaaaagagt aactccattt tcctgggatg ataacgatct cgctgtatta
      241 cggtccgaaa atgaagatgt tgccatgtat ctaagggagc atgacatcac tgtaggttac
      301 acatttgttc ttcacgacca tgataacaat ctggcgatcc tgactattgc taacaatgat
      361 gaaaaaaatg attttgagga ttttataaag aacagagaga atgatttaca aatgttgtta
      421 gtgactactc acgaaaaagc aatgaaacat aaacacttcg ttaaaggtaa aacggcgccc
      481 ttggattgct tgcaaagtgc attgattaca ccacgtgaaa cagaagtact tttcttggtc
      541 agtagaggga atacttataa agaggtgtcc agaacactgg gtatcagtga agcaacagtg
      601 aaattccata tcaataactc tgtcagaaaa cttaatgtca ttaattctcg ccatgccata
      661 agcaaagcac ttgagctcaa tctgtttcga gcctttacgg gatctctcat gaccagaaaa
      721 ttggttgcaa tatag
//
```

Fig.7.3 Gen-bank sequence file

Sequences in FASTA formatted files are preceded by a line starting with >. The first word on this line is the name of the sequence. The rest of the line is a description of the sequence. The remaining lines contain the sequence itself. An example of a FASTA file containing a single sequence is:

```
>1-BNI5

MGLDQDKIKKRLSQIEIDINQMNQMIDENLQLVEPAEDEAVEDNVKDTGVVDAVKVAETALFSGNDGAD
SNPGDSAQVEEHKTAQVHIPTENEANKSTDDPSQLSVTQPFIAKEQITHTAIAIGDSYNSFVANSAGNEKA
KDSCTENKEDGTVNIDQNRGEADVEIIENNDDEWEDEKSDVEEGRVDKGTEENSEIESFKSPMPQNNTLG
GENKLDAELVLDKFSSANKDLDIQPQTIVVGGDNEYNHESSRLADQTPHDDNSENCPNRSGGSTPLDSQT
KIFIPKKNSKEDGTNINHFNSDGDGQKKMANFETRRPTNPFRVISVSSNSNSRNGSRKSSLNKYDSPVSSPI
TSASELGSIAKLEKRHDYLSMKCIKLQKEIDYL…..
```

Fig.7.4 FASTA file format for protein BNI5 of Yeast

The protein sequences in FASTA format are collected and used in Bio-CLC program. The MSA is performed in accurate mode option. On the left hand side of Fig.7.5, the names of proteins are shown and by different colors the amino acids are arranged. Each 10 AAs are separated by gab to indicate the protein sequences. In the button, the

conservation and gap fraction are indicated. The conservation means the most frequent (strength) of amino acids in one place but the gap fraction means the difference between amino acids in this location. High conservation and low gap fraction are good indication for extracting the consensus. The first ten amino acids have high density relationships so the conservation level is increased and gab fraction is decreased as low percentage. So consensus is created as [SxNDSGx-P].

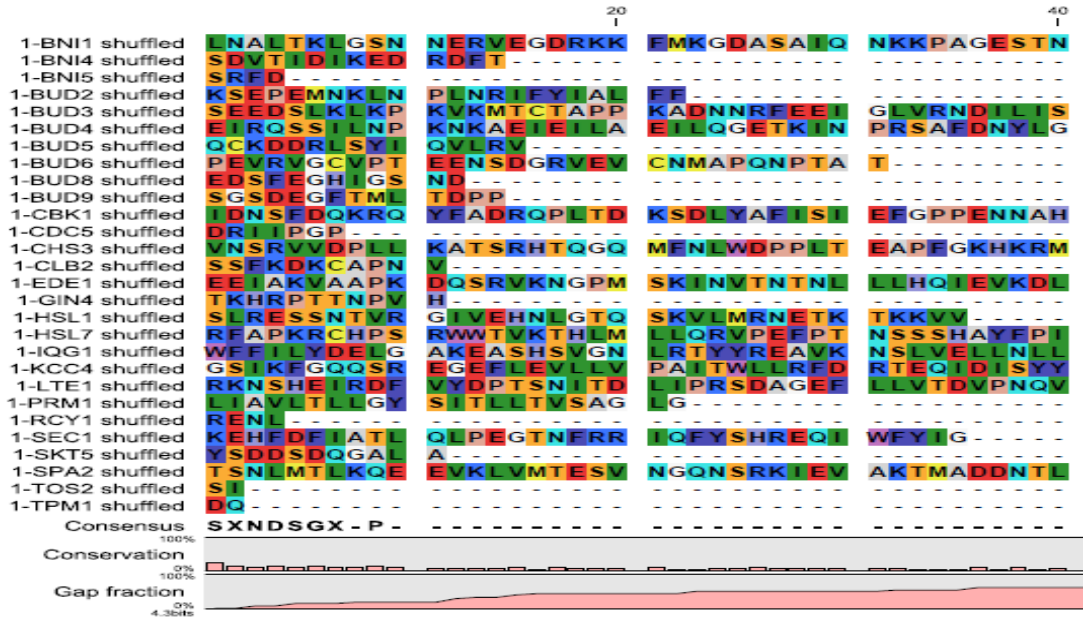


Fig.7.5 MSA over collected proteins. Specific motif (consensus) has been extracted from the first 10 proteins locations

The extracted amino acids (letters) are divided into three parts. Capital letters (SNDSGP) which means the first letters of most proper amino acids, (-) means gap (no amino acid in this location) and (x) means any amino acid can be found in this location. On the other hand, there are the unrelated sequences which have poor relations so they have high gap fraction.

Although the motif is clear for each specific area, the motif is difficult process to be extracted. Since the manual extraction is very exhaustive process, a detected threshold is created for consensus spectrum.

The threshold is detected relating to the maximum conservation percentage of the sequence alignment. The relation between the conservation percentage and alignment position is created as shown in Fig.7.6. The threshold is around 20% which means any alignment conservation more than threshold (~0.2) will be as motif location (measure for the function). Fig.7.6 shows 11 peaks more than the determined threshold where there is just one as maxima.

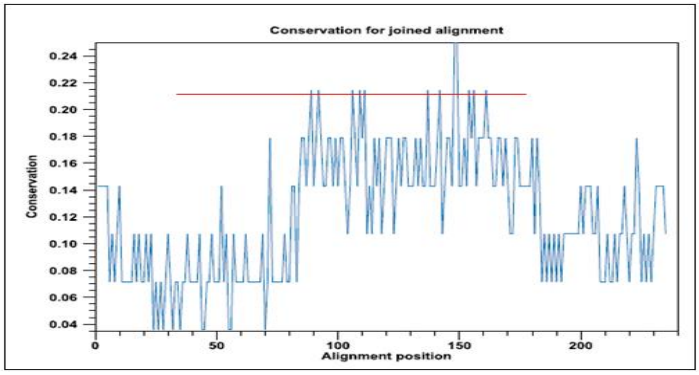


Fig.7.6 the relation between the conservation and alignment position

It is noticed that, the threshold value for determining the consensus is different for each sub-function category according to the sequences alignment strength as shown in Fig.7.7. This value is determined visually from the spectral graph or from the two dimensional data array between the sequence position and frequent percentage.

In this study, the multiple sequence alignment is applied to all protein sequences of Yeast (127 sub-functions). For each sub-function, protein sequences have been collected and sequence alignment is performed to extract specific motif. This motif is considered as feature (signature) for this function. As shown in Table 7.3, some motifs are collected for each sub-function category. These motifs are considered as identified features for each function. And it can be used to verify the predicted functions. If motif of function (**A)** for example is found in protein sequence and the mathematical methods estimate that protein to have this function (**A**), protein has high confidence to perform this function (high probability). Also it is noticed that, there are some sub-functions have no motifs. The reasons in this cases are as: 1)- the matching

of alignment is less than the suggested threshold, 2)- the sequences have other functions may change the main sequence of the protein, or 3)- the group of proteins (collected sequences) can not express the function. As shown in Table 7.3, function ID 2 has no clear motifs because in Fig.7.7a, the maximum amplitude does not reach for the required threshold. But the extracted motifs for protein functions 3, 8, and 5 as match as the peaks of the spectrum of the Fig.7.7. b, c, d respectively.

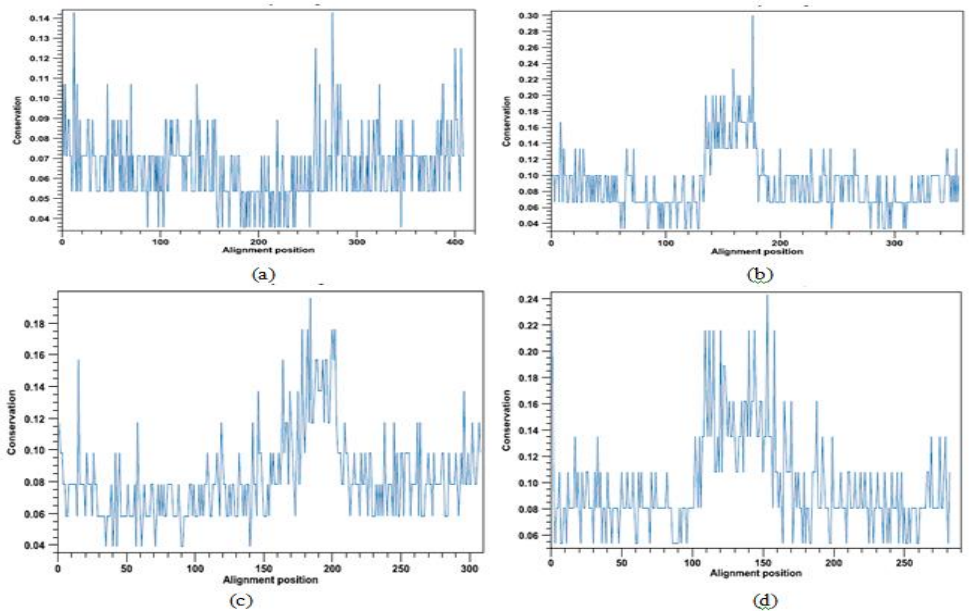


Fig.7.7 four different alignment graphs spectrum (consensus versus the position) for Bio-chemical sub-function categories

Table 7.3 Yeast protein functions and its extracted motifs indicating the start and end positions for biochemical sub-functions

| Function ID | Function Name | Starting position | End position | Consensus |
|---|---|---|---|---|
| 2 | Amino-acid metabolism | -- | -- | ---------- |
| 3 | Carbohydrate metabolism | 138 | 176 | TS-----ATELSR-R--T-A-AN---LEDL--------I |
| 8 | Cell structure | 182 | 203 | LDSRSSEXSEAALST---ESES |

| | | | | |
|---|---|---|---|---|
| 5 | Cell cycle control | 111 | 140 | SLSSNXTLNTXEXESS--SEELKXTTRSEQSRRSTSLKI---SSES-E |

## 7.3 Summary

In this chapter, we tried to extract motifs for each sub-function category in yeast proteins. The technique was based on applying multiple sequence alignment (MSA) to all yeast protein function categories sequences which were collected from different data sources as DIP, PIR, and SWISS PROT. CLC BIO program was used for applying sequence alignment. Motifs were extracted related to the most frequent positions of amino acids in the alignment. These motifs were considered as identified features for each function. And they could be used to verify the predicted functions. If motif of function for example was found in protein sequence and the mathematical methods estimated that protein to have this function, protein had high confidence to perform this function (high probability).

Although most of collected functions had specific motifs, there were few sub-function categories that had no motifs which reflected failure of this method in verifying the prediction process. So as advanced study, the collected protein sequences should have just one known functions (not more one) to avoid the confusion of two or more function motifs.

# Conclusion and Future work

Protein function prediction is one of the most important and hot tasks in the field of proteomics, since it leads to understanding cell activities. Protein functions may be predicted from protein sequences, gene expression, protein domains, protein localizations, protein structure, and protein-protein interactions (PPI) as recent computational techniques. Although protein function prediction through PPI networks is a powerful modality, it lacks the following points: 1) the reliability of the protein interactions to be considered in the prediction process where each interaction can be identified by one or more experimental method. And each experimental method has its score of stability and reliability, 2) the relations between the known functions which participate with the prediction process and 3) the features that identify these functions. Most of the previous computational techniques do not consider these points; that is why it decreases the confidence of the prediction process.

In this thesis, some algorithms were provided with new ideas to overcome the above-mentioned drawbacks. Regarding the reliability, an integrated algorithm was proposed. It included the experimental identification method; that contained the number of experimental methods furthermore their reliabilities, local topology which indicated the number of surroundings for the studied proteins, and global topology which illustrated the most common graphs for the proteins through the network. In addition, a new weighting algorithm was calculated using all the previous data. This new technique explored the collected data to create reliable interactions and enhance the prediction process.

Moreover, a novel technique was introduced to express the relations between protein functions, it included number of interactions between the protein clusters and overlapping number of proteins that had the same functions. This technique indicated the correlation,

anti-correlation, and independency between some protein functions which affected the protein function prediction.

Motif extraction was also performed using specific technique as multiple sequence alignment (MSA) in order to take advantage of the features that identified protein functions. This consensus (the most common positions of amino acids for proteins in multiple sequences alignment) was considered as the signature of that function and was used to identify it.

The proposed techniques were applied to Yeast data "Saccharomyces Cerevisiae" the simple eukaryote species which had complete genome and sequences. Yeast had a bout 6500 proteins classified into three main function categories. Each one of those function categories (biochemical – cell location – cellular role) contained many sub-functions.

The obtained results were validated via valuable methods and the results revealed great enhancement in protein function prediction process. The sensitivity and specificity of the results were more reliable than the previous techniques. The results of the cellular role function category were enhanced and improved specially in the three suggested basic weights w1 (experimental method), w2 (local topology), and w3 (global topology). The average and PCA weights introduced better results than weight less (unity weight) technique. Also it was noticed that, the sensitivity was improved with increasing number of interactions. In the second function category (cell location), all new weights were roughly overlapped and made shift in the positive direction to improve the sensitivity and specificity. Also the larger number of interactions improved the results. But the third functional category (Biochemical function), the weight less technique was better than the suggested algorithms which meant that biochemical function category did not depend on the interactions and the protein neighbors and the computational methods failed in estimating the protein functions of un-known proteins from the surrounding neighbors. On the second hand (protein function relations), many functions were estimated from these relations furthermore the anti-correlation concept which applied for all functions. Finally for the third technique (motif function extraction), many protein function

categories had one or more motifs (sequence of amino acids) which they had conservation value greater than the determined threshold. These motifs were considered as signatures for those functions.

So we concluded our work as:

1- The new weighting technique enhanced the sensitivity and specificity for two function categories (cell location and cellular role).
2- Increasing the number of interactions improved the sensitivity and specificity.
3- The technique did not reveal good results in biochemical function category which indicated, the estimation of this function category was very difficult using the neighbor data.
4- Some protein functions had high correlation reached for 100% which meant that any protein had one of the correlated functions it had the second one.
5- The correlation based on the overlapping number of protein was more accurate than based on the protein cluster interactions.
6- The integration between the cluster interactions and overlapping number of proteins gave higher accuracy than any one of them.
7- Motifs were extracted related to the most frequent positions of amino acids in the alignment.
8- These motifs were considered as identified features for each function. And it used to verify the predicted functions. If motif of function for example was found in protein sequence and the mathematical methods estimated that protein to have this function, protein had high confidence to perform this function (high probability).
9- There were few sub-function categories had no motifs which reflected the failure of this method in verifying the prediction process.

## FUTURE WORK

Herein, some methods and techniques can be performed to the data sources as:

1- The new weighting technique can be integrated with MRF method and other statistical techniques.
2- Calculating the reliability of the interactions by other techniques as protein domains.
3- Applying the MSA to proteins that contain just one function not more.
4- Collecting all these data and results in one package.
5- Applying all the previous techniques to human proteins.
6- Getting a relation between the function and the structure part that response for doing it.
7- Estimating the protein functions through other new techniques as fussy logic or neural network.

[1] M. Zhao, and K. Aihara, "Gene function prediction using labeled and unlabeled data," *BMC Bioinformatics*, vol. 9, pp. 57-71, 2008.

[2] H. Zhao, Wu, B., "DNA-Protein Binding and gene expression patterns" *Lecture Notes-Monograph Series, Statistics and Science: A Festschrift for Terry Speed*, vol. 40, pp. 259-274, 2003.

[3] M. Morin "Phylogenetic Networks Simulation, Characterization, and Reconstruction" *Proc New Mexico*,2007.

[4] J. Sun and Z. Zhao, "Construction of phylogenetic profiles based on the genetic distance of hundreds of genomes," *Biochemical and Biophysics Resources Communication*, vol. 355, pp. 849-53, 2007.

[5] M. Pellegrini, E. Marcotte "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles". *Proc National Academy of Science USA*, vol.96, pp. 4285-8, 1999.

[6] E. D. Harrington, A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, and P. Bork, "Quantitative assessment of protein function prediction from metagenomics shotgun sequences," *Proc National Academy of Science U S A*, vol. 104, pp. 13913-8, 2007.

[7] R. V. Spriggs, Y. Murakami, and S. Jones, "Protein function annotation from sequence: prediction of residues interacting with RNA," *Bioinformatics*, vol. 25, pp. 1492-7, 2009.

[8] I. Friedberg, "Automated protein function prediction--the genomic challenge," *Brief Bioinformatics*, vol. 7, pp. 225-42, 2006.

[9] N. Nariai, E. D. Kolaczyk, and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," *PLoS One*, vol. 2, p. e337-344,2007.

[10] T. G. Rohit Gupta, Gaurav Pandey, Michael Steinbach, and Vipin Kumar, "Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements Using Association Analysis " *in bibl. SIAM International Data Mining Conference*, Minneapolis, 2007.

[11] Y. Liu, I. Kim, and H. Zhao, "Protein interaction predictions from diverse sources," *Drug Discovery Today*, vol. 13, pp. 409-16,2008.

[12] A. Wagner, "How the global structure of protein interaction networks evolves," *Proc Biological Science*, vol. 270, pp. e35-40, 2003.

[13] A. Zhang, *Protein Interaction Networks: Computational Analysis*, Cambridge University Press, 2009.

[14] B. Schwikowski, and S. Fields, "A network of protein-protein interactions in yeast," *Natural Biotechnology*, vol. 18, pp. 1257-61,2000.

[15] R. Sharan "Analysis of biological networks: Protein-protein interaction networks – functional Annotation". *lecture note*,2006.

[16] H. Hishigaki, K. Nakai, T. Ono, and T. Takagi, "Assessment of prediction accuracy of protein function from protein--protein interaction data," *Yeast*, vol. 18, pp. 523-31,2001.

[17] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *Computational Biology*, vol. 10, pp. 947-60, 2003.

[18] Z. Wei and H. Li, "A Markov random field model for network-based analysis of genomic data," *Bioinformatics*, vol. 23, pp. 1537-44, 2007.

[19] D. D. Clark and D. L. Tennenhouse, "Architectural considerations for a new generation of protocols," *SIGCOMM: Proceedings of the ACM symposium on communications, architectures and protocols*, vol 90, pp. 200-208,1990.

[20] Wendell Odom "*CCNA ICND2 Official Exam Certification Guide*" USA press, 2007.

[21] S. Overby, "*Drug companies speed*" CIO magazine, 2001.

[22] P. Larranga, B. Calvo, R. Santana, and V. Robles "machine learning in bioinformatics" *Brief bioinformatics*, vol. 7, no. 1, pp. 86-112, 2006.

[23] http://newscenter.cancer.gov

[24] S.C Rastogi. "*BIOINFORMATICS METHODS AND APPLICATIONS*". Textbook,2005.

[25] http://www.mesotheliomaweb.org/proteomics.htm

[26] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, " Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome wide data," *Proceedings of the National Academy of Sciences of the United States of America*, vol.101, no. 9, pp. 2981-2986, 2004.

[27] A. lesk, "introduction to bioinformatics," Oxford Press, *ch. Archieves on information retrieval*, pp. 206-209, 2008.

[28] R. Jeremy, "Bioinformatics," 2nd edition, ser. *Computational Biology Series. Springer Verlag GmbH*, vol. 10, 2009.

[29] http://www.cod.edu/people/faculty/fancher/prokeuk.htm

[30] http://ghr.nlm.nih.gov/handbook/basics/dna

[31] http://en.wikipedia.org/wiki/DNA

[32] http://en.wikipedia.org/wiki/RNA

[33] http://www.biomed.curtin.edu.au/biochem/tutorials/AAs/AA.html

[34] http://www.biology.arizona.edu/biochemistry/problem_sets/aa/aa.html

[35] http://www.getbig.com/articles/protein.htm

[36] I.Avila-Campillo, K.Drew, J. Lin, D. J. Reiss, and R. Bonneau, "Bio Net Builder: automatic integration of biological networks," *Bioinformatics*,vol.23, no.3, pp. 392-393,2007.

[37] U. Alon, "Biological networks: the tinkerer as an engineer," *Science*, vol. 301, pp. 1866-7, 2003.

[38] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47-52, 1999.

[39] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc Natural Academy of Science U S A*, vol. 93, pp. 13-20, 1996.

[40] E. V. Koonin, Y. I. Wolf, and G. P. Karev, "The structure of the protein universe and genome evolution," *Nature*, vol. 420, pp. 218-23, 2002.

[41] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani, "Computational methods for predicting protein-protein interactions," *Advanced Biochemical Engineering and Biotechnology*, vol. 110, pp. 247-67,2008.

[42] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces," *Molecular Biology*, vol. 325, pp. 377-87, 2003.

[43] I. M. Nooren and J. M. Thornton, "Diversity of protein-protein interactions," *European Molecular Biology Conference*, vol. 22, pp. 3486-92, 2003.

[44] S. Jones, A. Marin, and J. M. Thornton, "Protein domain interfaces: characterization and comparison with oligomeric protein interfaces," *Protein Engineering*, vol. 13, pp. 77-82, 2000.

[45] C. J. Tsai and R. Nussinov, "Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association," *Protein Science*, vol. 6, pp. 1426-37, 1997.

[46] Lo Conte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *Molecular Biology*, vol. 285, pp. 2177-98, 1999.

[47] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399-403, 2002.

[48] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein-protein interaction data?," *Molecular Biology*, vol. 327, pp. 919-23, 2003.

[49] R. Albert and A.L. Barabȗsi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47–97,2002.

[50] S.B. Seidman, "Network structure and minimum degree," *Social Networks*, vol. 5, pp. 269–87, 1983.

[51] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-2, 1998.

[52] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-12, 1999.

[53] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Natural Revision Genetics*, vol. 5, pp. 101-13, 2004.

[54] M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill, "Interactome: gateway into systems biology," *Human Molecular Genetics*, vol. 14 Spec No. 2, pp. 171-81, 2005.

[55] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg, "A protein interaction map of Drosophila melanogaster," *Science*, vol. 302, pp. 1727-36, 2003.

[56] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," *Proc National Academy of Science U S A*, vol. 97, pp. 1143-7, 2000.

[57] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, pp. 623-7, 2000.

[58] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, pp. 141-7, 2002.

[59] A. C. Gingras, R. Aebersold, and B. Raught, "Advances in protein complex analysis using mass spectrometry," *Physiology*, vol. 563, pp. 11-21, 2005.

[60] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, pp. 180-3, 2002.

[61] N. J. Krogan, W. T. Peng, G. Cagney, M. D. Robinson, R. Haw, G. Zhong, X. Guo, X. Zhang, V. Canadien, D. P. Richards, B. K. Beattie, A. Lalev, W. Zhang, A. P. Davierwala, S. Mnaimneh, A. Starostine, A. P. Tikuisis, J. Grigull, N. Datta, J. E. Bray, T. R. Hughes, A. Emili, and J. F. Greenblatt, "High-definition macromolecular composition of yeast RNA-processing complexes," *Molecular Cell*, vol. 13, pp. 225-39, 2004.

[62] L. M. Markillie, C. T. Lin, J. N. Adkins, D. L. Auberry, E. A. Hill, B. S. Hooker, P. A. Moore, R. J. Moore, L. Shi, H. S. Wiley, and V. Kery, "Simple protein complex purification and identification method for high-throughput mapping of protein interaction networks," *Proteome Resources*, vol. 4, pp. 268-74, 2005.

[63] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni, "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules," *Science*, vol. 295, pp. 321-4, 2002.

[64] H. Ge, "UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions," *Nucleic Acids Resources*, vol. 28, p. e3, 2000.

[65] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder, "Global analysis of protein activities using proteome chips," *Science*, vol. 293, pp. 2101-5, 2001.

[66] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc, National Academy of Science U S A*, vol. 98, pp. 4569-74, 2001.

[67] S. Fields and O. Song, "A novel genetic system to detect protein-protein interactions," *Nature*, vol. 340, pp. 245-6, 1989.

[68] T. Ito, K. Ota, H. Kubota, Y. Yamaguchi, T. Chiba, K. Sakuraba, and M. Yoshida, "Roles for the two-hybrid system in exploration of the yeast protein interactome," *Molecular Cell and Proteomics*, vol. 1, pp. 561-6, 2002.

[69] I. Abe, T. Seki, K. Umehara, T. Miyase, H. Noguchi, J. Sakakibara, and T. Ono, "Green tea polyphenols: novel and potent inhibitors of squalene epoxidase," *Biochemical and Biophysics Resources Communications*, vol. 268, pp. 767-71, 2000.

[70] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198-207,2003.

[71] S. Oliver, "Guilt-by-association goes global," *Nature*, vol. 403, pp. 601-3, 2000.

[72] T. P. Conrads, H. J. Issaq, and T. D. Veenstra, "New tools for quantitative phosphoproteome analysis," *Biochemical and Biophysics Resources Communication*, vol. 290, pp. 885-90,2002.

[73] B. Kuster, P. Mortensen, J. S. Andersen, and M. Mann, "Mass spectrometry allows direct identification of proteins in large genomes," *Proteomics*, vol. 1, pp. 641-50, 2001.

[74] J. Peng, J. E. Elias, C. C. Thoreen, L. J. Licklider, and S. P. Gygi, "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry

(LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome," *Proteome Resources*, vol. 2, pp. 43-50,2003.

[75] M. P. Washburn, D. Wolters, and J. R. Yates, 3rd, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology," *Natural Biotechnology*, vol. 19, pp. 242-7,2001.

[76] E. Lasonder, Y. Ishihama, J. S. Andersen, A. M. Vermunt, A. Pain, R. W. Sauerwein, W. M. Eling, N. Hall, A. P. Waters, H. G. Stunnenberg, and M. Mann, "Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry," *Nature*, vol. 419, pp. 537-42, 2002.

[77] M. Mann, S. E. Ong, M. Gronborg, H. Steen, O. N. Jensen, and A. Pandey, "Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome," *Trends Biotechnology*, vol. 20, pp. 261-8, 2002.

[78] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Natural Biotechnology*, vol. 21, pp. 255-61, 2003.

[79] D. H. Blohm and A. Guiseppi-Elie, "New developments in microarray technology," *Current Opinion Biotechnology*, vol. 12, pp. 41-7, 2001.

[80] G. MacBeath and S. L. Schreiber, "Printing proteins as microarrays for high-throughput function determination," *Science*, vol. 289, pp. 1760-3, 2000.

[81] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin, "A generic protein purification method for protein complex characterization and proteome exploration," *Natural Biotechnology*, vol. 17, pp. 1030-2,1999.

[82] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, and B. Seraphin, "The tandem affinity purification (TAP) method: a general procedure of protein complex purification," *Methods*, vol. 24, pp. 218-29, 2001.

[83] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal, "A map of the interactome network of the metazoan C. elegans," *Science*, vol. 303, pp. 540-3, 2004.

.

[84] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman,

C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga, "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, pp. 631-6,2006.

.

[85] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, pp. 141-7, 2002.

[86] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. St Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, pp. 637-43,2006.

[87] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating whole-genome expression data with protein-protein interactions," *Genome Resources*, vol. 12, pp. 37-46, 2002.

[88] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular Cell and Proteomics*, vol. 1, pp. 349-56, May 2002.

[89] P. Kemmeren, N. L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma, and F. C. Holstege, "Protein interaction verification and functional annotation by integrated analysis of genome-scale data," *Molecular Cell*, vol. 9, pp. 1133-43,2002.

[90] S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie, and G. S. Davidson, "A gene expression map for Caenorhabditis elegans," *Science*, vol. 293, pp. 2087-92,2001.

[91] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics*, vol. 18, pp. 1454-61,2002.

[92] O. G. Troyanskaya, "Putting microarrays in a context: integrated analysis of diverse biological data," *Brief Bioinformatics*, vol. 6, pp. 34-43,2005.

[93] N. Bhardwaj and H. Lu, "Correlation between gene expression profiles and protein-protein interactions within and across genomes," *Bioinformatics*, vol. 21, pp. 2730-8, 2005.

[94] S. Tornow and H. W. Mewes, "Functional modules by relating protein interaction networks and gene expression," *Nucleic Acids Resources*, vol. 31, pp. 6283-9,2003.

[95] S. A. Teichmann and M. M. Babu, "Conservation of gene co-regulation in prokaryotes and eukaryotes," *Trends Biotechnology*, vol. 20, pp. 407-10, 2002.

[96] H. Ge, Z. Liu, G. M. Church, and M. Vidal, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae," *Natural Genetics*, vol. 29, pp. 482-6,2001.

[97] A. Grigoriev, "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae," *Nucleic Acids Resources*, vol. 29, pp. 3513-9,2001.

[98] R. Mrowka, A. Patzak, and H. Herzel, "Is there a bias in proteome research?," *Genome Resources*, vol. 11, pp. 1971-3,2001.

[99] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, pp. 249-55,2003.

[100] H. B. Fraser, A. E. Hirsh, D. P. Wall, and M. B. Eisen, "Coevolution of gene expression among interacting proteins," *Proc National Academy of Science U S A*, vol. 101, pp. 9033-8,2004.

[101] S. L. Rutherford, "From genotype to phenotype: buffering mechanisms and the storage of genetic information," *Bioessays*, vol. 22, pp. 1095-105,2000.

[102] J. L. t. Hartman, B. Garvik, and L. Hartwell, "Principles for the buffering of genetic variation," *Science*, vol. 291, pp. 1001-4,2001.

[103] A. Bender and J. R. Pringle, "Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in Saccharomyces cerevisiae," *Molecular Cell and Biology*, vol. 11, pp. 1295-305,1991.

[104] S. L. Ooi, X. Pan, B. D. Peyser, P. Ye, P. B. Meluh, D. S. Yuan, R. A. Irizarry, J. S. Bader, F. A. Spencer, and J. D. Boeke, "Global synthetic-lethality analysis and yeast functional profiling," *Trends Genetics*, vol. 22, pp. 56-63,2006.

[105] J. A. Brown, G. Sherlock, C. L. Myers, N. M. Burrows, C. Deng, H. I. Wu, K. E. McCann, O. G. Troyanskaya, and J. M. Brown, "Global analysis of gene function in yeast by quantitative phenotypic profiling," *Molecular System and Biology*, vol. 2, p. 2006-11, 2006.

[106] A. H. Tong, M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone, "Systematic genetic analysis with ordered arrays of yeast deletion mutants," *Science*, vol. 294, pp. 2364-8,2001.

[107] P. Ye, B. D. Peyser, X. Pan, J. D. Boeke, F. A. Spencer, and J. S. Bader, "Gene function prediction from congruent synthetic lethal interactions in yeast," *Molecular System and Biology*, vol. 1, p. 2005-26,2005.

[108] J. Lippincott-Schwartz and G. H. Patterson, "Development and use of fluorescent protein markers in living cells," *Science*, vol. 300, pp. 87-91,2003.

[109] J. Piehler, "New methodologies for measuring protein interactions in vivo and in vitro," *Current Opinion Structure Biology*, vol. 15, pp. 4-14,2005.

[110] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Natural Revision Genetics*, vol. 5, pp. 101-13,2004.

[111] C. R. Myers, "Software systems as complex networks: structure, function, and evolvability of software collaboration graphs," *Physics Revision*, vol. 68, p. 046-116, 2003.

[112] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268-76, 2001.

[113] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-2,1998.

[114] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Resources*, vol. 32, pp. D449-51,2004.

[115] H. W. Mewes, D. Frishman, K. F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stumpflen, "MIPS: analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Resources*, vol. 34, pp. D169-72,2006.

[116] R. Albert, H. Jeong, and A. L. Barabasi, "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378-82,2000.

[117] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41-2,2001.

[118] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-5, 2002.

[119] J. Kim and K. Tan, "Discover protein complexes in protein-protein interaction networks using parametric local modularity," *BMC Bioinformatics*, vol. 11, p. 521,2000.

[120] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces," *Molecular Biology*, vol. 325, pp. 377-87,2003.

[121] E. Estrada, "Virtual identification of essential proteins within the protein interaction network of yeast," *Proteomics*, vol. 6, pp. 35-40,2006.

[122] E. Estrada and J. A. Rodriguez-Velazquez, "Sub-graph centrality in complex networks," *Physics Revision*, vol. 71, pp. 56-103,2005.

[123] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On Power-law Relationships of the Internet Topology," *Proc. SIGCOMM*, pp.251-62, 1999.

[124] H. W. Mewes, D. Frishman, K. F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stumpflen, "MIPS: analysis and annotation of proteins from whole genomes in 2005," *Nucleic Acids Resources*, vol. 34, pp. D169-72, 2006.

[125] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H. W. Mewes, A. Ruepp, and D.

Frishman, "The MIPS mammalian protein-protein interaction database," *Bioinformatics*, vol. 21, pp. 832-4,2005.

[126] C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M. J. Dumontier, M. R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J. P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J. J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B. F. Ouellette, and C. W. Hogue, "The Biomolecular Interaction Network Database and related tools 2005 update," *Nucleic Acids Resources*, vol. 33, pp. D418-24,2005.

[127] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Resources*, vol. 34, pp. D535-9,2006.

[128] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni, "MINT: the Molecular INTeraction database," *Nucleic Acids Resources*, vol. 35, pp. D572-4,2007.

[129] M. Persico, A. Ceol, C. Gavrila, R. Hoffmann, A. Florio, and G. Cesareni, "HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms," *BMC Bioinformatics*, vol. 6, p. S21-5,2005.

[130] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob, "IntAct--open source resource for molecular interaction data," *Nucleic Acids Resources*, vol. 35, pp. D561-5,2007.

[131] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, S. Menon, G. Hanumanthu, M. Gupta, S. Upendran, S. Gupta, M. Mahesh, B. Jacob, P. Mathew, P. Chatterjee, K. S. Arun, S. Sharma, K. N. Chandrika, N. Deshpande, K. Palvankar, R. Raghavnath, R. Krishnakanth, H. Karathia, B. Rekha, R. Nayak, G. Vishnupriya, H. G. Kumar, M. Nagini, G. S. Kumar, R. Jose, P. Deepthi, S. S. Mohan, T. K. Gandhi, H. C. Harsha, K. S. Deshpande, M. Sarker, T. S. Prasad,

and A. Pandey, "Human protein reference database--2006 update," *Nucleic Acids Resources*, vol. 34, pp. D411-4,2006.

[132] M. E. Frazier, G. M. Johnson, D. G. Thomassen, C. E. Oliver, and A. Patrinos, "Realizing the potential of the genome revolution: the genomes to life program," *Science*, vol. 300, pp. 290-3,2003.

[133] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, and B. Jacq, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network," *Genome and Biology*, vol. 5, p. R6, 2003.

[134] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data," *Molecular Biology Evolution*, vol. 14, pp. 685-95,1997.

[135] M. P. Samanta and S. Liang, "Predicting protein functions from redundancies in large-scale protein interaction networks," *Proc National Academy of Science U S A*, vol. 100, pp. 12579-83,2003.

[136] D. S. Goldberg and F. P. Roth, "Assessing experimentally derived interactions in a small world," *Proc National Academy of Science U S A*, vol. 100, pp. 4372-6,2003.

[137] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," *Proc Pac Symp Biocomput*, pp. 300-11, 2004.

[138] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, "Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps," *Bioinformatics*, vol. 21 Suppl 1, pp. i302-10,2005.

[139] J. Yu and R. L. Finley, Jr., "Combining multiple positive training sets to generate confidence scores for protein-protein interactions," *Bioinformatics*, vol. 25, pp. 105-11,2009.

[140] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker, "Conserved patterns of protein interaction in multiple species," *Proc National Academy of Science U S A*, vol. 102, pp. 1974-9,2005.

[141] R. Okada, K. Asai and M. Arita. "Flow model of the protein-protein interaction network for finding credible interactions." *proceedings of the 5th asia-pacific bioinformatics conference*, pp. 317-26,2007

[142] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant, "Gaining confidence in high-throughput protein interaction networks," *Natural Biotechnology*, vol. 22, pp. 78-85,2004.

[143] A. S. Aytuna, A. Gursoy, and O. Keskin, "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," *Bioinformatics*, vol. 21, pp. 2850-5,2005.

[144] M. Deng, F. Sun, and T. Chen, "Assessment of the reliability of protein-protein interactions and protein function prediction," *Proc Pac Symp Biocomput*, pp. 140-51,2003.

[145] R. Saito, H. Suzuki, and Y. Hayashizaki, "Interaction generality, a measurement to assess the reliability of a protein-protein interaction," *Nucleic Acids Resources*, vol. 30, pp. 1163-8,2002.

[146] P. Bonacich, " Power and centrality: a family of measures." *American Journal of Sociology*, vol. 92, pp. 1170–82,1987.

[147] L. Katz, "On the Matric Analysis of Sociometric Data," *Sociometry* vol. 10, pp. 233-41,1947.

[148] J. R. Seeley, "The net of reciprocal influence. Study II: The balance of power," *Can J Psychol*, vol. 5, pp. 68-76,1951.

[149] B. J. Breitkreutz, C. Stark, and M. Tyers, "Osprey: a network visualization system," *Genome Biology*, vol. 4, p. R22,2003.

[150] B. J. Breitkreutz, C. Stark, and M. Tyers, "The GRID: the General Repository for Interaction Datasets," *Genome Biology*, vol. 4, p. R23,2003.

[151] H. N. Chua, W. K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions," *Bioinformatics*, vol. 22, pp. 1623-30,2006.

[152] K. Sayed, N. Soloma, and Y. Kadah, "Estimation of the correlation between protein sub-function categories based on overlapping proteins," *Proc. 27th National Radio Science Conference*, Menouf, Egypt,2010

[153] K. Sayed, N. Solouma, and Y. Kadah, "Determining The Relations Between Protein Sub-Function Categories Based On Overlapping Proteins,"*Journal of Communication and Computer*, 2011 (In Press).

[154] K. Sayed, N. Solouma, and Y. Kadah, "Comparison between different methods for protein function prediction," *proc. 1st International Joint Conference (NRC), Cairo, 2009.*

[155] P. M. Conn, "*Handbook of proteomic methods*," Totowa, NJ: Humana Press, 2003.

[156] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, pp. 910-3,2002.

# نبذة عن الرسالة

يعد التنبؤ بوظائف البروتينات واحدة من اهم المهام المطلوبة فى علم البروتين حيث انها تعمل على فهم نشاط الخلية. فهذا التنبؤ يمكن ان يحدث من خلال تسلسلات البروتين, مواضع البروتينات , تركيب البروتينات او من خلال تحليل شبكات البروتينات كعلم حديث في السنوات السابقة.

فبالرغم من أن الطريقة الحديثة (تحليل شبكات البروتين) تعتبر واحدة من الطرق المتقدمة الا انها تخلو من بعض العوامل التى تؤثر على دقة هذا التنبؤ مثل اولا:- بعض من الروابط المتواجدة بين البروتينات لم يتم اكتشافه او تحديدها بدقة ولكن يمكن ان تكون صدفة او تم اكتشافها بطرق غير دقيقة وغير معبرة. ثانيا:- لم يتم التعبير عن العلاقات بين هذة الوظائف واستغلالها فى عملية التنبؤ. ثالثا:- لم يتم تحديد مؤشرات معينة لكل وظيفة لتكون بمثابة بصمة لهذة الوظيفة.

فى هذة الرسالة تم تقديم العديد من الطرق بافكار جديدة لكى تتغلب على العقبات السابقة . فتم علاج دقة الروابط عن طريق عمل اسلوب جديد ياخذ عدد طرق الاكتشاف ودقة كل طريقة اضافة للوضع الحالى للرابط من المنظور المحلى و الكلى.

و بالنسبة الى العلاقات بين وظائف البروتينات فقد تم اكتشافها عن طريق الروابط بين مجموعة البروتينات علاوة على عدد البروتينات المتطابقة فى كل وظيفة. اما بالنسبة الى الحصول على عوامل مميزة لتلك الوظائف فقد تم عمل محازاة لتسلسلات البروتينات لكى يتم اخذ اعلى متشابهات من الاحماض الامينية لكى تكون مميزة لكل وظيفة.

هذة الطرق تم تطبيقهاعلى فطر الخميرة الذى يتشابه مع الانسان الى حد كبير.ففطر الخميرة يحتوى على 6500 بروتين و تنقسم وظائفة الى ثلاثة وظائف اساسية من حيث الروابط الكيميائية , وضع البروتين فى الخلية و دور البروتين فى الخلية.

بعد تطبيق هذة الاساليب والطرق تم الحصول على نتائج عالية الدقة مما ادى الى التنبؤ بهذة الوظائف بدرجة صحيحة و عالية.

# التنبؤ بوظائف البروتينات الجديدة من الروابط المتواجدة في شبكة البروتينات


إعداد

**المهندس**

## خالد السيد أحمد مصطفى


**رسالة مقدمه إلي كليه الهندسة - جامعه القاهرة**
**كجزء من متطلبات الحصول علي درجه الدكتوراة**

**في**

## الهندسه الحيوية الطبيه والمنظومات


**كلية الهندسة - جامعه القاهرة**

**الجيزه-مصر**

**2011**

# التنبؤ بوظائف البروتينات الجديدة من الروابط المتواجدة في شبكة البروتينات

إعداد


**المهندس**

## خالد السيد أحمد مصطفى


**رسالة مقدمه إلي كليه الهندسة, جامعه القاهرة**
**كجزء من متطلبات الحصول علي درجه الدكتوراة**

**في**

## الهندسه الحيوية الطبيه والمنظومات

تحت إشراف


**أ.د. أبوبكرمحمد يوسف**

الهندسه الحيوية الطبيه والمنظومات
كليه الهندسة
جامعه القاهرة

**أ. د. ياسر مصطفي قدح**

الهندسه الحيوية الطبيه والمنظومات
كليه الهندسة
جامعه القاهرة

**د. ناهد حسين سلومة**
استاذ مساعد بمعهد بحوث اليزر
جامعه القاهرة



**كلية الهندسة - جامعه القاهرة**

**الجيزه-مصر**

**2011**

# التنبؤ بوظائف البروتينات الجديدة من الروابط المتواجدة في شبكة البروتينات

إعداد


المهندس

**خالد السيد أحمد مصطفى**


رسالة مقدمه إلي كليه الهندسة, جامعه القاهرة

كجزء من متطلبات الحصول علي درجه الدكتوراة

**في الهندسه الحيوية الطبيه والمنظومات**

يعتمد من لجنه الممتحنين:

**الأستاذة الدكتورة: سامية عبد الرازق مشالى ( عضو)**

.....

**الأستاذ الدكتور: عبد الله سيد أحمد محمد (عضو)**

.....

**الأستاذ الدكتور: : ياسر مصطفي ابراهيم قدح ( المشرف الرئيسي)**

.....

**الدكتورة: ناهد حسين سلومة ( المشرف)**

.....


كلية الهندسة - جامعه القاهرة

الجيزه-مصر

2011