Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software Department

Complex Prediction from Static and Dynamic Protein-Protein Interaction Networks using Graph Mining Algorithms

A Dissertation

Submitted to the Council of the College of Information Technology at University of Babylon in Partial Fulfillment of the Requirements for the Degree of Doctorate of Philosophy in Information Technology/ Software

By

Soheir Noori Alwan Mohammed

Supervised by

Prof. Dr. Nabeel Hashem Al-A'araji

Prof. Dr. Eman Salih Al-Shamery

2022 A.C. **1443 A.H**

﴿ اللَّه نُورُ السَّمَاوَاتِ وَالْأَرْضِ مَثَلُ نُورِهِ كَمِشْكَاةٍ فِيهَا مِصْبَاحٌ الْمِصْبَاحُ فِي زُجَاجَةٍ الزُّجَاجَةُ كَأَنَّهَا كَوْكَبٌ دُرِّيٌّ يُوقَدُ مِنْ شَجَرَةٍ مُبَارَكَةٍ زَيْتُونَةٍ لَا شَرْقِيَّةٍ وَلَا غَرْبِيَّةٍ يَكَادُ زَيْتُهَا يُضِيءُ وَ لَوْ لَمْ تَمْسَسْهُ نَارٌ نُورٌ عَلَىٰ نُورٍ يَهْدِي اللَّه لِنوُرِهِ منَ يشاَءُ وَيَضْرِبُ أللَّه الْأَمْثَالَ لِلنَّاسِ وَاللَّه بِكُلِّ شَيْءٍ عَلِيم ﴾

صدق الله العظيم

سورة النّور\آية35

Declaration

I declare that this Dissertation, submitted to University of Babylon in partial fulfillment of requirements for the degree of Doctorate of Philosophy in Information Technology / Software has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose sources are appropriately cited in the references.

Signature:

Name: Soheir Noori Alwan

Date: / / 2022

Dedication

I dedicate this work to who was lighting my live, my lovely Mother (may Allah bless her soul and grant her the highest state in the Jannah)

I dedicate this work to my Father, my Husband, my Daughters, my Family, supervisors, and friends.

Acknowledgment

Praise be to Allah. Cherisher and Sustainer of the worlds, I'm entrusted to his Almighty in helping me to finish what I started, and in presenting this work in the best way.

I would like to express my deep appreciation and sincere gratitude to my supervisors: **prof. Dr. Nabeel Al-A'araji** and **Prof. Dr. Eman S. Al-Shamery**, whose dedications to this work have been boundless. Their efforts to ensure the high quality of the material of the work have been invaluable. I am lucky enough to be advised and guided by them.

Sincere appreciation to the members of the Department of Software at University of Kerbala for their help during accomplishing this work.

Special thanks go to Limsoon Wong for his constructive suggestions and perceptive comments.

Finally, special thanks to my family (my mother, my husband and my lovely daughters) and my friends for their help and encouragement during conducting this work.

Soheir Noori

Abstract

Protein complexes play a critical role in understanding the mechanism and organization of cellular processes since they are involved in vital cellular functions. They are combinations of two or more proteins assembled from the proteins in the Protein Interaction Network (PIN).

Prediction of complexes from static and dynamic protein interaction networks is the main aim of the dissertation. Here four contributions are proposed for predicting protein complexes from static and dynamic PINs, two for each. Using the seed expansion model and the topological structure (SETS) of the static PIN (SPIN), the SETS algorithm is proposed to predict overlapping protein complexes with varying densities. Another algorithm based on the Gene Expression (GE) and Core-Attachment (GECA) approach is proposed from weighted SPIN with GE data without eliminating the proteins and their edges that offer no gene expression data. GECA identifies core proteins employing common neighbor techniques and biological information. Moreover, GECA improves the attachment technique by adding the proteins that have low closeness but high similarity to the gene expression of the core proteins.

For dynamic networks, determining an adequate threshold is one of the system's biological challenges. A quartile one (q-one) method is proposed to determine the active time points for each protein according to the features of its expression value. Q-one constructs dynamic protein interaction networks (DPINs) in which a protein is added to the network if it is active in two successive time points. This leads to reduce the number of DPINs to half. This is in addition to its

ability to reveal the dynamicity within the protein-protein interaction (PPI) network and identify essential proteins. By utilizing the q-one method to determine the active time points for each protein, a new algorithm (shared-one-time) is proposed for predicting protein complexes from DPINs whose proteins shared at least one active time point.

Eleven datasets for yeast and human have been exploited to evaluate the proposed methodology base on evaluation measures (F-measure, exact and good matching with reference complexes). As well as, many comparisons are implemented with known and powerful algorithms.

In different datasets, the F-measure of the proposed algorithms is at least 10% and up to 65% better than that of other algorithms. The rate of improvement of predicted complexes that have exactly and good matching with reference complexes, on the other hand, is at least 15% higher than previous approaches, indicating considerable biological significance.

Declaration Associated with this Dissertation

I. Published Articles

1. Soheir Noori, Nabeel Al-A'araji and Eman Al-Shamery. (2021). SETS: A Seed-Dense-Expanding Model-Based Topological Structure for the Prediction of Overlapping Protein Complexes, **Pertanika Journal of Science and Technology**, 29(2), **UNIV PUTRA MALAYSIA PRESS, Scopus(Q3)**, Clarivate Analytics **(emerging)**, **https://doi.org/10.47836/pjst.29.2.35**
2. Soheir Noori, Nabeel Al-A'araji and Eman Al-Shamery. (2021). Identifying Protein Complexes from Protein–Protein Interaction Networks Based on the Gene Expression Profile and Core-Attachment Approach, **Journal of Bioinformatics and Computational Biology**, 19(03), **WORLD SCIENTIFIC PUBL CO PTE LTD**, **Clarivate Analytics (Q3)**, **Impact Factor** Web of Science JCR **1.122**, https://doi.org/10.1142/S0219720021500098.
3. Soheir Noori, Nabeel Al-A'araji and Eman Al-Shamery. (2022). Construction of Dynamic Protein Interaction Network Based on Gene Expression Data and Quartile One Principle, **PROTEINS: Structure, Function, and Bioinformatics**, 90(5), **John Wiley & Sons Inc**, **Clarivate Analytics (Q1)**, **Impact Factor** Web of Science JCR **3.756**, **https://doi: 10.1002/prot.26304**

II. Under Preparing Article

1. Soheir Noori, Nabeel Al-A'araji and Eman Al-Shamery. (2022). Prediction protein Complexes from Dynamic PPI Network.

Table of contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description
BP	**Biological Process**
CC	**Cellular Component**
CN	**Common Neighbor**
CR	**Coverage Rate**
CS	**Closeness Score**
DPIN	**Dynamic Protein Interaction Network**
ECC	**Edge Clustering Coefficient**
F	**F-measure**
GE	**Gene Expression**
GEP	**Gene Expression Profile**
GO	**Gene Ontology**
L	**Large complex**
MF	**Molecular Function**
OS	**Overlapping Score**
P	**Precision**
PCC	**Pearson's Correlation Coefficient**
PC	**Protein Complex**
PPC	**Predicted Protein Complex**
PPI	**Protein-Protein Interaction**
PIN	**Protein Interaction Network**
RPC	**Reference Protein Complex**
R	**Recall**
S	**Small complex**
SPIN	**Static Protein Interaction Network**

Chapter One **General Introduction**

1.1. Introduction

Protein complexes are the key for understanding the mechanism and organization of cell processes as they take part in essential cellular functions (Yong et al., 2012). Most of the functional processes within a cell are executed by protein complexes. Therefore, the identification of protein complexes is an important research problem in systems biology. These complexes are made up of two or more proteins that interact at the same time and in the same location. They are created from proteins in the Protein–Protein Interaction (PPI) network, which plays an important role in regulatory processes, cellular activities, and signaling cascades (Zengyou, 2015).

During the past decade, scientific research has directed more attention to detecting protein complexes and developing numerous algorithms for dealing with them. Most of these algorithms predict a protein complex as a dense area in a protein interaction network (PIN). They do this either by using a density threshold or an objective density function, while other algorithms search for cliques. All these algorithms find cliques and then merge them depending on various criteria for identifying the dense sub-graphs that represent the protein complexes. However, these methods ignore many low-density complexes, even though 40% of the complexes in CYC2008 (Pu et al., 2009), MIPS (Mewes et al., 2006), and Aloy (Aloy et al., 2004) have a density of less than 0.5 (Liu et al., 2011).

Gavin et al. (Gavin et al., 2006) reported that protein complexes are organized as core proteins and attachment proteins. The core represents the proteins that strongly interact with each other, whereas the attachments mark the boundary of the core. Several algorithms based on this technique have also been proposed. These algorithms consist of two phases, namely, identifying the complex core as a dense

sub-graph, and then identifying the attachment proteins as those that interact with more than half the proteins in the core.

Nevertheless, most protein complex detection algorithms employ the topological properties of a graph or mix PIN with other information such as the gene expression (GE) that reveals the dynamic properties of the PIN. Hunter (Chin et al., 2010) and WEC (Keretsu and Sarmah, 2016) demonstrate that a weighted PIN by its gene expression profile (GEP) can improve the quality of protein complex detection. While there has been great progress in computational study of proteome scale cellular networks, the inherent dynamism of protein interactions within these networks is sometimes disregarded. The interactions of Biomolecules have changed across time, environment and different phases of celiac development, making cellular systems highly dynamic and responsive to environmental stimuli (Przytycka et al., 2010). As a result, it is important to transit the analysis of PPI networks from static to dynamic (Yong and Wong, 2015) since the dynamic PIN (DPIN) helps to show how disease progression is reflected in the time-evolving of PIN, and aids in disease identification before clinical symptoms appear (Wang et al., 2014).

The available PIN datasets are static and lack temporal and condition characteristics for protein interactions. As a consequence, dynamic information on PPI and protein complexes are neglected (Jenghara et al., 2018). Because the PIN is static, other information such as gene expression data should be used to develop dynamic networks in order to make use of their dynamic properties. In this area, microarray gene expression data provides useful dynamic information for modeling the activities of proteins at any given time and producing dynamic PIN. The protein is not always active, thus it is important to determine the time point at which a protein demonstrates activity before constructing the dynamic protein interaction network (Zhang et al., 2019).

1.2. Problem Statement

Protein complex prediction in biological networks can be considered as a graph clustering problem. Biological networks can be shown as proteins connected by edges. The edges represent the interactions between the network's proteins. Multiple methodologies are used for processing biological networks. Moreover, a number of acknowledged issues in predicting protein complexes that are relevant to this dissertation are listed below:

1) Most existing algorithms can detect only highly dense regions as protein complexes and ignore low density complexes (Wang et al., 2018b). Further, most of them cannot detect overlapping protein complexes (Zhao and Lei, 2019).
2) The elimination of edges from the PIN when biological information about their proteins is absent may lead to bias in the results. While most core-attachment-based methods employ the same technique in adding attachment proteins, ignoring the dynamic properties of the PIN.
3) The utilization of dynamic networks has recently attracted the attention of the research community (Saha et al., 2019). Nonetheless, constructing an accurate DPIN is still a challenge (Meng et al., 2021). All available approaches use a cautious and relatively high threshold to determine the protein’s active time point. As a result, the dynamic information of genes with expression values lower than particular values is lost (Zhang et al., 2019).
4) Recently, various computational methods for identifying complexes from static PPI networks have been developed by researchers. However, because static PPI networks cannot reflect the dynamics in real cellular systems (Xiao et al., 2019). Many attempts have been made to improve the prediction of

dynamic protein complexes by grouping the static PPI network with gene expression data (Rani et al., 2019).

As a result, the proposed algorithms are put into action to address the issues raised above.

1.3. The Challenges of the Dissertation

Because of the unique characteristics of PPI networks, any clustering approaches devised for these networks should take into account the following distinct network and cluster characteristics (Bhowmick and Seah, 2015):

- The overlapping nature of protein complexes: In recent years, the knowledge that many proteins have numerous activities has supplanted the concept of one gene, one protein and one function. When a protein is involved in many functional modules, protein complexes in a PPI network may overlap. Therefore, PPI clustering approaches must produce overlapping protein complexes.
- Sparse and dense protein complexes: Topological modules may be dense, but functional modules may not be. As a consequence, in order to produce good results, a PPI clustering technique must take into account both dense and sparse protein complexes.
- Covering all PIN: Because protein complexes can be both dense and sparse, it is critical for a clustering algorithm to cover all the proteins in a PPI network.
- PIN with attribution. With the growing of knowledge on protein functions, proteins in a PIN are increasingly being annotated with attributes (such as Gene Ontology (GO) and Gene Expression (GE) to encode information about

their functions, localization, and biological processes. When such annotations are available, clustering algorithms should use them to produce higher-quality results.

- Construction of DPIN. Identifying each protein's active time point by using GE data is crucial to construct DPIN (Zhang et al., 2016).

1.4. The Aim of the Dissertation

Prediction of protein complexes in static and dynamic PPI networks is the main aim of this dissertation. The following are some of the objectives:

1) Predicting overlapping protein complexes from SPIN. The prediction is not only for high-density complexes, but also for low-density ones in a reasonable length of time, hence increasing the network's protein coverage rate.
2) Improving the prediction accuracy by adding biological information to SPIN while keeping proteins and their edges that do not have GE data. This is in addition to bettering the core-attachment strategy by allowing the inclusion of proteins that are related to less than half of the proteins in the core, but have a gene expression pattern that is very comparable to proteins in the core.
3) Determining each protein's active time points according to the features of its expression values and construct dynamic protein interaction networks (DPINs).
4) Predicting overlapping protein complexes from DPINs whose proteins share the same time point.

1.5. The Contributions of the Dissertation

The following are the primary contributions of this dissertation:

1) Proposing a SETS algorithm by using a seed expanding model and the topological structure of the PPI network to predict overlapping protein complexes with different densities within an acceptable execution time and good coverage of PPI network.
2) Proposing a GECA algorithm that employing the core-attachment technique to predict accurate overlapping protein complexes utilizing biological information.
3) Proposing a new technique to construct DPIN by identifying a new threshold to determine the active time points of each protein in SPIN.
4) Proposing a new algorithm for improving the accuracy of predicting protein complexes from DPIN by identifying protein complexes with proteins that have the same active time points.

1.6. Related Work

This section looks at the most up-to-date SPIN and DPIN protein complex detection algorithms. Table 1-1 summarizes all relevant algorithms with additional information.

1. SPIN

In the last few decades, researchers have presented a number of techniques to investigate the SPIN in order to predict protein complexes. Some of these algorithms are described in the next section.

1) MCODE (Bader and Hogue, 2003b) is one of the seed-extension approaches. It identifies overlapping protein complexes in three steps.

Step 1, is based on the core clustering coefficient and assigns a weight to every node in the graph. Step 2, extends from seeds that have a high weight and finds a dense region in the weighted graph. Finally, complexes that are not dense are filtered.

2) The Clique percolation method (CPM) (Palla et al., 2005): It is proposed by Palla et al. The concept of a k-clique cluster, which was defined as the union of all k-cliques (complete subgraphs of size k) that can be accessed from each other through a series of consecutive k-cliques (where adjacency means sharing k-1 nodes). Iterative recursion can be used to obtain all the network's k-cliques, and then create the overlap matrix of these k-cliques. Finally, by analyzing the overlap matrix, a number of k-clique clusters are revealed. Although the algorithm CPM is good at analyzing the overlapping clusters of society and biology, it has the limitation of identifying a restricted number of protein complexes. When a relatively big k value is chosen, the number of protein complexes is reduced.

3) IPCA (Li et al., 2008b) is another algorithm that identifies a dense region in the PPI network as a protein complex. It starts from the seed that has the biggest weight, which is the summation of its weighted edge that represents the number of its common neighbours. IPCA then recursively adds the neighbours of the seed based on two criteria, namely, the shortest path between the seed and the node as well as the probability of its interaction.

4) Markov clustering (MCL) (Vlasblom and Wodak, 2009) is a stochastic flow simulation-based graph clustering technique that has been found to be successful in clustering biological networks. It provides a technique

based on graph transition probabilities and demonstrates a high level of noise tolerance. While not totally parameter-free, changing just one parameter might result in clusters of various granularities.

5) SPICi (Jiang and Singh, 2010) is a fast heuristic clustering algorithm employed to cluster large PPI networks that select the seed having the highest weight. The weight represents the degree of the node and then uses a support function to expand the way that the density of the cluster remains over a user-defined threshold. Otherwise, the cluster is terminated and the nodes in this cluster are removed from the network.

6) ClusterOne (Nepusz et al., 2012) detects overlapping protein complexes by starting from the seed protein having the highest degree and then gradually adding and removing proteins to find a cohesive group of proteins that can be overlapped. Finally, complexes with fewer than three proteins or a density below a specified level are eliminated.

7) PEWCC (Zaki et al., 2013) consists of two primary steps. First, analyzing the trustworthiness of protein interaction data using a novel measurement (PE-measure), which allows reducing the amount of noise in PPI networks. Second, weighted clustering coefficients are used as a metric to determine which subgraph is closest to the maximal clique that represents the protein complex.

8) NCMine (Tadaka and Kinoshita, 2016) defines a near-complete subgraph as a functional module by using the centrality degree as the weight of the nodes, which then iteratively merges these like cliques to define overlapping modules.

9) The weighted Edge-Based Clustering (WEC) (Keretsu and Sarmah, 2016) approach can detect protein complexes that are strongly

interconnected and co-expressed by weighting the edge between two proteins using the Edge Clustering Coefficient (ECC) and the correlation between the proteins' gene expression. This made it possible to predict the protein complex accurately. WEC then attached the proteins if they are connected with half of the core proteins.

10) Weighted COACH (WCOACH) (Kouhsar et al., 2016): COACH is a well-known algorithm for identifying protein complexes in unweighted PINs. It predicts protein complexes without taking into account the semantic similarity of contacts. As a result, incorporating weighted graphs into the COACH technique can improve the accuracy of complicated prediction. A unique algorithm, WCOACH, enables weighted networks of complicated prediction using the following four main steps:

First: Using gene ontology (GO) to weight the PIN.

Second: Detecting cores with high density.

Third: Getting rid of any cores that are redundant.

Finally: Assembling the predicted complex by adding attachment proteins to each core.

11) Core-attachment with gene CAG (Keretsu and Sarmah, 2017) used the core-attachment technique and identified the core as a functional unit by finding only the positive correlation between proteins in the neighborhood and iteratively removing nodes with the minimum degree to maintain the density of the core. They then attached the proteins to the core if they are connected with half of the core proteins.

All the aforementioned algorithms rely on the topological structure of SPIN and most of them use the seed-extension approach to detect protein complexes as a dense subgraphs in static networks.

2. DPIN

Several algorithms for predicting protein complexes from dynamic PPI networks have been proposed. There some of these algorithms:

1) The Three-Sigma Method (Wang et al., 2013): Wang et al. conclude that determining the protein's activity time point is critical in DPIN construction. They proposed the 3-sigma method to calculate an active threshold for each gene based on its expression curve's properties instead of using a global threshold for all proteins. After that, DPINs are built by making a connection between two proteins in a DPSN if these proteins are active and interact with each other in the static PIN.

2) DPC (Li et al., 2014) is proposed to predict dynamic protein complexes. DPC recognizes dense subgraphs as cores and constantly active proteins are found from these. Attachments based on a topological property of "closeness" and dynamic meaning are added to final protein complexes. DPC is tested on yeast data, and the experimental results prove that it outperforms static algorithms.

3) C0-DPC (Xiao et al., 2019): It is a novel approach for detecting dynamic protein complexes based on the core-attachment principle. CO-DPC first uses gene expression profiles and the 3-sigma principle to identify active proteins. It then employs the co-expression principle and PPI networks to build dynamic PPI networks. Second, CO-DPC recognizes local dense subgraphs as the cores of protein complexes. It then groups them with their neighbors, which are connected with half of the core proteins, to

form protein complexes. Finally, the high overlapping protein complexes are filtered out.

4) IFPA (Lei et al., 2019) predicte protein complexes in DPIN based on a flower pollination mechanism. IFPA developed dynamic protein networks that are multi-relation rebuilt. It then uses the core–periphery structure to classify the closely related proteins as cores and employs the IFPA algorithm to bind peripheries to the best cores. IFPA is tested on yeast dataset, and the findings show that IFPA is superior to the previous methods.

Table 1-1: Summary of related work

No	Algorithm name	Weighted network	PPI datasets	Reference datasets	Evaluation metrics
1	**MCODE**	(node) Core clustering coefficient	Gavin MIPS	Gavin MIPS	Sensitivity Specificity
2	**ClusterONE**	(node) Node's degree	Krogan Collins BioGRID Gavin	MIPS SGD	Accuracy Maximum matching ratio (MMR)
3	**NCMine**	(node) degree centrality	HPRD	HPRD	Fisher's exact test P-value
4	**SPICi**	weighted	Biogrid-Yeast STRING- Yeast Biogrid- Human STRING- Human Bayesian-Human	MIPS	Accuracy GO-analysis
5	**IPCA**	(node) sum of the weights of its incident edges	MIPS	MIPS	F-measure
6	**PEWCC**	Edge weighted	PPI-D1 PPI-D2	MIPS Cmplx-D2 Cmplx-D3	Accuracy MMR
7	**WEC**	Edge weighted	Collins Gavin Human PPI	SGD MIPs CORUM	F-measure Sensitivity Positive Predictive Value (PPV) Co-localization Score
8	**CAG**	Edge weighted (PCC)	Collins Gavin	SGD MIPs	F-measure Co-localisation
9	**WCOACH**	Edge weighted (GO)	DIP Krogan MIPS	CYC2008	F-measure CR

			Gavin2002 Gavin2006		
10	**Three-Sigma**		DIP	CYC2008	F-measure
11	**C0-DPC**		Krogan DIP	CYC2008	F-measure
12	**DPC**		DIP	Nucleic Acids Research	F-measure
13	**IFPA**	Edge weighted (co-essentiality, Co-localization, Co-annotation and Co-cluster)	DIP MIPS Krogan	CYC2008	F-measure

1.7. Dissertation Outline

The chapters are ordered as follows:

- The essential principles of the PPI Network are described in Chapter two.
- Chapter three explains the essential processes in developing the suggested methodology for predicting protein complexes using graph mining techniques.
- The experimental results of the proposed methodology are discussed in Chapter four.
- Chapter five summarizes the findings and makes recommendations for future researches.

Chapter Two Protein Interaction Network and Graph Mining

2.1. Introduction

This chapter covers all the concepts about proteins and their networks, as well as how to predict protein complexes from protein interaction networks using graph mining techniques.

2.2. Protein in Bioinformatics

Bioinformatics is a combination of biological and computer science that has the ability to analyze huge biological datasets by using a computer. Before the existence of bioinformatics, there are two scientific terms to describe biological experiments, namely, in Vivo and in Vitro. The former means that the experiment is done on organisms directly, whereas the latter means that it is done in a laboratory. Recently, researchers in bioinformatics use the term ‘in silico’. It means that the experiment is implemented on computer before applying it on the organism or in the laboratory to save money and time (zohairy, 2014).

For many years, bioinformatics has been used in protein research and has made significant contributions to the study of protein, including sequencing, structure, and evolution. Proteins are one of the most versatile compounds in living species. Several methodologies have been used to examine the structure and functional roles of proteins. It includes experimental studies such as protein's characterization, as well as numerous computational techniques for protein analysis (Gromiha, 2010).

Large assemblies of proteins are dubbed "protein machines of cells" by the biochemist Bruce Alberts in an early study (Alberts, 1998). Protein assemblies are made up of highly specialized pieces that work together to carry out practically all

of the biochemical, signaling, and functional operations in cells (Alberts, 1998). Even in the simplest of eukaryotic cells, protein assemblages can number in hundreds. However, our understanding of these protein assemblies, in addition to how they interact to form the "higher level" functional architecture of cells, is still limited. It is therefore critical to make a sincere effort to identify and characterize all protein assemblies in order to understand how the cellular machinery works.

A protein must physically interact with other proteins and biomolecules to be functioning. According to estimates, approximately 80% of proteins in humans do not operate alone, but rather interact to form macromolecular assemblies (Berggård et al., 2007). Protein assemblies are characterized as protein complexes, functional modules, biochemical and signaling pathways based on the functional, spatial, and temporal context of their interactions (Srihari et al., 2017).

Protein complexes are the main types of protein assemblages and are essential functional components in organisms. Complexes are persistent complexes that result from physical protein interactions at a certain time and location (Tang et al., 2011, Spirin and Mirny, 2003, Jung et al., 2010, Kong et al., 2020). Complexes are responsible for a variety of functions within cells, including cytoskeleton formation, cargo transportation, substrate metabolism for energy production, DNA replication, genome protection and maintenance, genes to gene transcription and translation, protein turnover maintenance, and cell protection from internal and external damaging agents (Srihari et al., 2017). Complexes might be permanent or transient. In recent years, several algorithms have been proposed for protein complex prediction. The majority of these algorithms on protein interaction networks (PINs).

2.3. Protein Structure

The Greek word "proteios," which meaning "principal," is where the term "protein" originates. As the name implies, proteins play a crucial role in biological systems. About 3/4 of the dry weight of the body is made up of proteins. Proteins are essential for bodybuilding since they carry out all of the body's key structural and functional functions. Protein structural abnormalities can cause molecular disorders that significantly change metabolic processes.

Proteins are synthesized by polymerization of amino acids through peptide bonds. A dipeptide is made up of two amino acids; a tripeptide of three; a tetrapeptide of four; an oligopeptide of a few amino acids; and a polypeptide of between ten and fifty amino acids. Proteins are lengthy polypeptide chains with more than 50 amino acids, according to convention.

There are 3 amino acids in a tripeptide, although any of the 20 amino acids can make up these 3. A tripeptide can therefore take any of 203 = 8000 distinct permutations and combinations. A typical protein with 100 amino acids has 20100 distinct possible combinations. More atoms exist in this number than there are in the entire cosmos. Thus, despite the fact that there are only 20 amino acids, nature makes a huge variety of distinctly different proteins by altering the order and combination of these amino acids (Vasudevan et al., 2019).

Proteins are structurally organized at primary, secondary, tertiary, and quaternary levels as shown in Figure 2-1. The primary structure of a protein is its polypeptide chain's amino acid sequence; secondary structure is the local spatial arrangement of the atoms that make up the polypeptide's main chain; tertiary structure is the three-dimensional arrangement of the polypeptide chain as a whole; and quaternary structure is the three-dimensional arrangement of the subunits in a multisubunit protein (Everse, 2014).

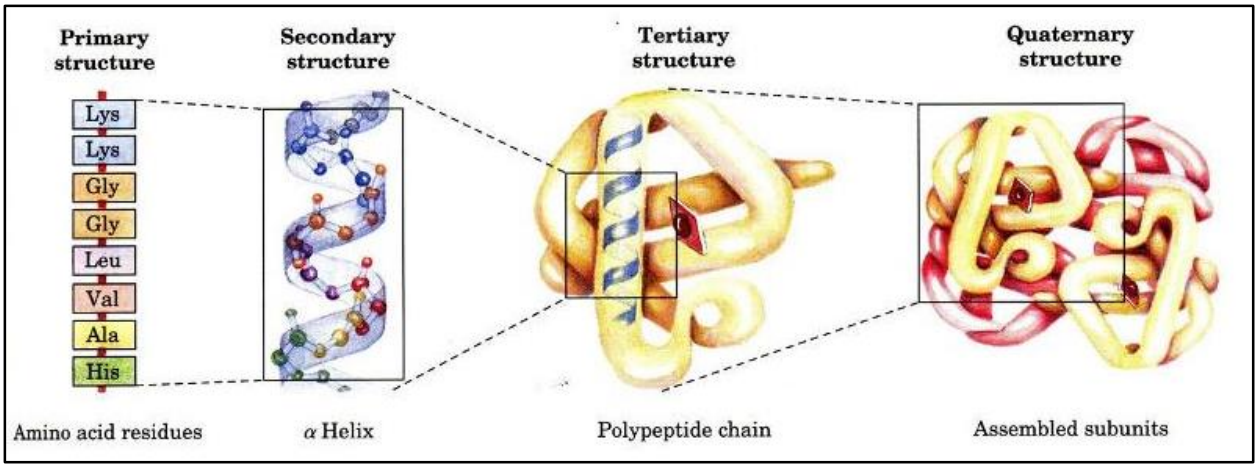


Figure 2-1: Protein structure levels

2.4. Protein-Protein Interaction Datasets

Recent improvements in technology have increased the processing power and the speed with which information is exchanged. This shift the focus from simple procedures to machine learning techniques in order to achieve the highest level of accuracy between experimental data and computational methods. The distance between experiment and theory has narrowed due to the development of high-speed computers with massive storage capacities (Gromiha, 2010).

Advanced experimental techniques such as yeast two-hybrid (Y2H) (Ito et al., 2001), microarrays (Stoll et al., 2005) and tandem affinity purification (TAP) (Puig et al., 2001) have generated a vast amount of known data of PPI datasets (DIP (Xenarios et al., 2002), Krogan (Krogan et al., 2006), Collins (Collins et al., 2007), Gavin (Gavin et al., 2006), BioGRID (Breitkreutz et al., 2007), Homo sapiens (human) (Ma et al., 2017), HPRD (the Human Protein Reference Database) (Peri et al., 2003), HSN (Human Signalling Network) (Liu et al., 2016) and STRING (Jensen et al., 2009)) that provide Protein interactions for yeast and human, in which the

interactions could be annotated or weighted. One of the most frequent ways of representing PPI as complicated sets of binary interactions between proteins as shown in Figure 2-2 is through networks, which are called Protein Interaction Networks (PINs) Figure 2-3.

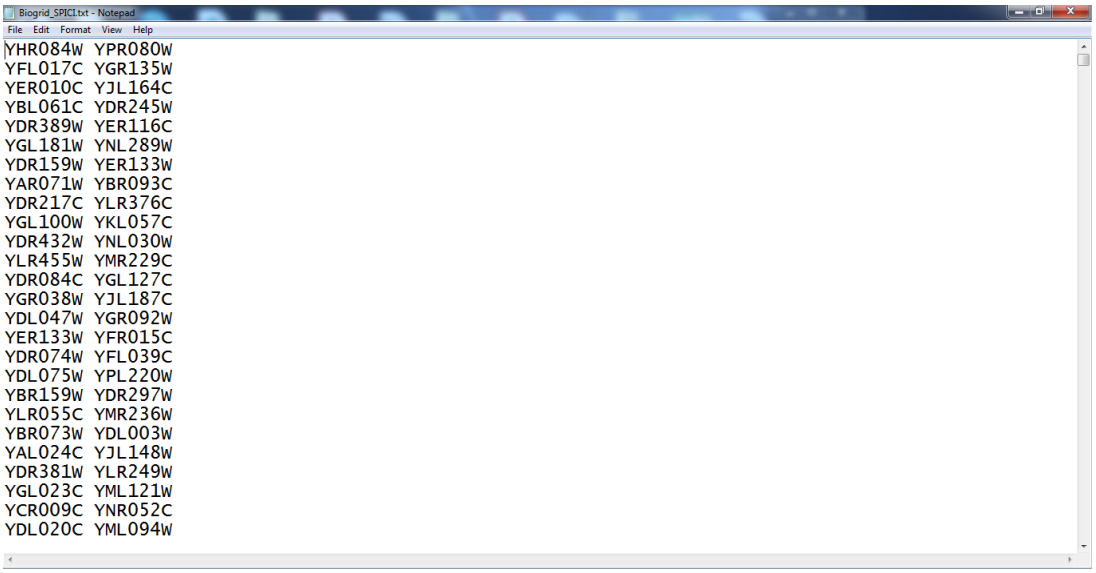


Figure 2-2: BioGRID PPI dataset

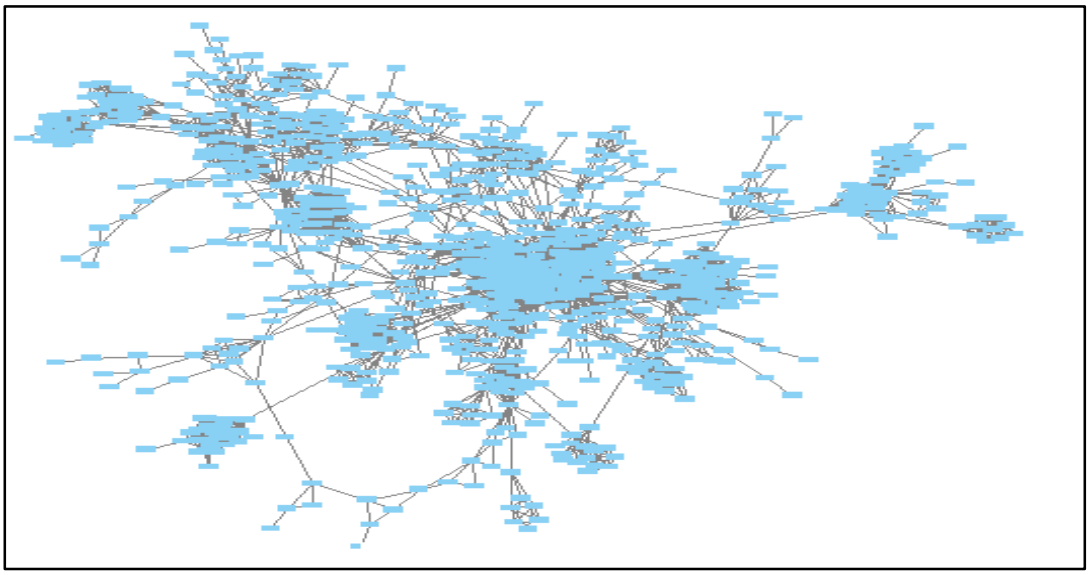


Figure 2-3: Protein Interaction Network of Collins dataset

2.5. Gene Expression Profile

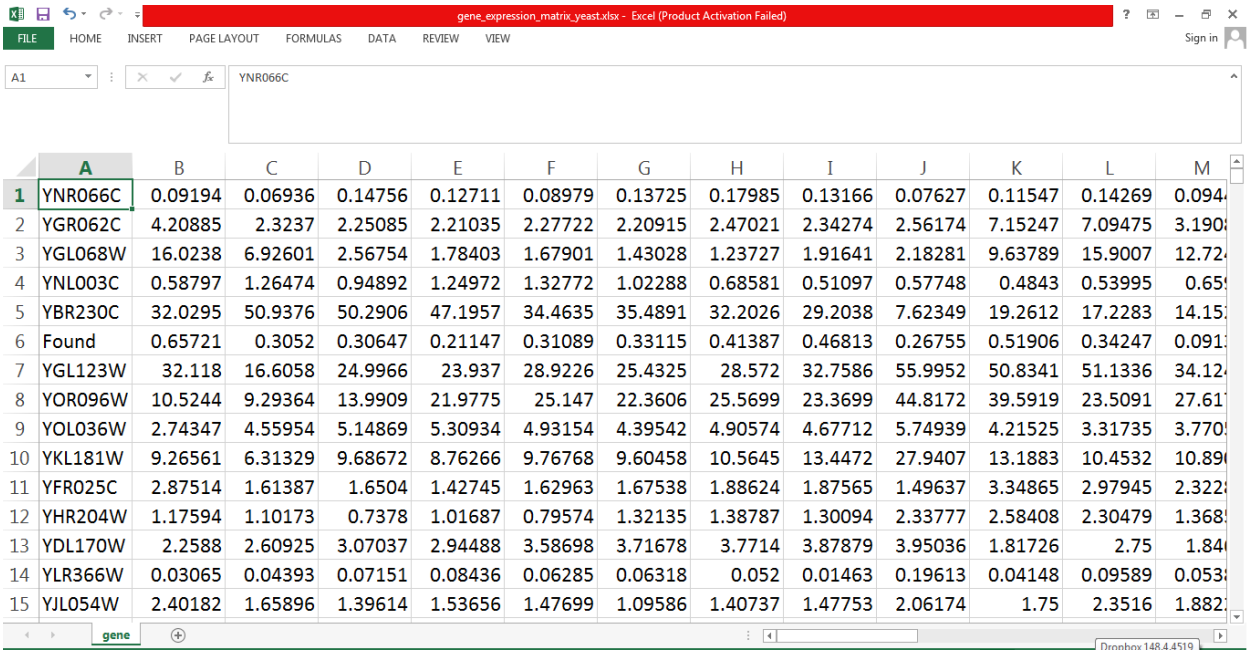
DNA is the genetic material of all organisms on Earth. It is divided up into functional units called genes. Information from a gene is used to build a functional product (protein) in a process called gene expression. The cell reads the sequence of the gene in groups of three bases. Each group of three bases (codon) corresponds to one of 20 different amino acids which are the base unit to build the protein. Proteins determine the phenotype of all organisms which can differ in their composition and functions depending on the number and type of amino acids involved in their synthesis. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function (Academy, 2004).

Gene expression profile contains the values of proteins in different time points or conditions as shown in Figure 2-4. GEP is a t × n matrix of real values to t genes or proteins in n conditions or at n time points. The expression value of protein *p* can be represented as Ge(*p*) = [g(*p*,1), g(*p*,2), …., g(*p*, n)]. The expression values of two genes or proteins X and Y are represented as X = [x_1, x_2, x_3, ….. , x_n] and Y = [y_1, y_2, y_3, …… , y_n]. Pearson's Correlation Coefficient (PCC) as defined in Equation (2-1) is used to measure the similarity between these two GEP patterns. PCC is the most successful way to measure the similarity of the co-expression between genes (Makrodimitris et al., 2020).

$$PCC(X,Y) = \frac{\sum_{i=1}^{n}(x_i - x')(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \ \sum_{j=1}^{n}(y_j - y')^2}} \qquad (2-1)$$

PCC determines the similarity between two genes X and Y. x_i to x_n and y_i to y_n are representing in the expression values of gene X and Y at n time points, x' and y' are representing the mean of the expression value of X and Y genes. The value of PCC is between -1 and 1. 1 represents the perfect correlation between two genes in

a positive direction, whereas -1 represents the perfect correlation but in a negative direction, and 0 represents no correlation between two genes.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	YNR066C	0.09194	0.06936	0.14756	0.12711	0.08979	0.13725	0.17985	0.13166	0.07627	0.11547	0.14269	0.094
2	YGR062C	4.20885	2.3237	2.25085	2.21035	2.27722	2.20915	2.47021	2.34274	2.56174	7.15247	7.09475	3.190
3	YGL068W	16.0238	6.92601	2.56754	1.78403	1.67901	1.43028	1.23727	1.91641	2.18281	9.63789	15.9007	12.72
4	YNL003C	0.58797	1.26474	0.94892	1.24972	1.32772	1.02288	0.68581	0.51097	0.57748	0.4843	0.53995	0.65
5	YBR230C	32.0295	50.9376	50.2906	47.1957	34.4635	35.4891	32.2026	29.2038	7.62349	19.2612	17.2283	14.15
6	Found	0.65721	0.3052	0.30647	0.21147	0.31089	0.33115	0.41387	0.46813	0.26755	0.51906	0.34247	0.091
7	YGL123W	32.118	16.6058	24.9966	23.937	28.9226	25.4325	28.572	32.7586	55.9952	50.8341	51.1336	34.12
8	YOR096W	10.5244	9.29364	13.9909	21.9775	25.147	22.3606	25.5699	23.3699	44.8172	39.5919	23.5091	27.61
9	YOL036W	2.74347	4.55954	5.14869	5.30934	4.93154	4.39542	4.90574	4.67712	5.74939	4.21525	3.31735	3.770
10	YKL181W	9.26561	6.31329	9.68672	8.76266	9.76768	9.60458	10.5645	13.4472	27.9407	13.1883	10.4532	10.89
11	YFR025C	2.87514	1.61387	1.6504	1.42745	1.62963	1.67538	1.88624	1.87565	1.49637	3.34865	2.97945	2.322
12	YHR204W	1.17594	1.10173	0.7378	1.01687	0.79574	1.32135	1.38787	1.30094	2.33777	2.58408	2.30479	1.368
13	YDL170W	2.2588	2.60925	3.07037	2.94488	3.58698	3.71678	3.7714	3.87879	3.95036	1.81726	2.75	1.84
14	YLR366W	0.03065	0.04393	0.07151	0.08436	0.06285	0.06318	0.052	0.01463	0.19613	0.04148	0.09589	0.053
15	YJL054W	2.40182	1.65896	1.39614	1.53656	1.47699	1.09586	1.40737	1.47753	2.06174	1.75	2.3516	1.882

Figure 2-4: Gene Expression Profile

2.6. Protein Interaction Network and Protein Complexes

There are three main biological networks that defined molecules: Gene co-expression network, Protein-protein interaction (PPI) and metabolic network. PPI is widely used by researchers who explain the interaction between proteins such as the building of protein complexes and the activation of one protein by another protein. PPI controls the health and disease mechanism, thus it becomes the basis of studying biological diseases.

PPI plays a central role in many biological functions and its network provides a global view of cellular functionality. Studying PIN is important to identify the pathway of different biological processes and determine the disease causes and its

progression. Machine-learning algorithms have been used to discover novel interactions between proteins. However, hierarchical and k-means are useful clustering methods, but they do not work properly with the PPI network because of their characteristics. Thus, representing PIN as a graph makes the process of finding clusters faster and more accurate (Aittokallio and Schwikowski, 2006).

PIN has been given more attention by biologists due to their part in major diseases (Xu and Li, 2006). Although the availability of PPI data aided the understanding of PPI networks, it is widely known that PPI data is incomplete (Venkatesan et al., 2009). The following are some of the shortcomings in PPI datasets that have a direct impact on protein complex prediction:

- The significant amount of noise in protein interactions.
- The scarcity of protein interactions.
- A lack of contextual information about protein interactions (e.g., temporal and geographical).

These limitations lead to the following three main issues currently facing computational techniques for identifying protein complexes:

- Predicting sparse complexes.
- Predicting complexes with less than four proteins.
- Predicting overlapping complexes (i.e., complexes that share a large number of proteins), especially when they arise in various biological settings.

The PIN is mined for modular subnetworks by algorithms of predicting protein complexes. This modularity is defined topologically as dense subgraphs of proteins separated by less dense network regions. Internal (mutations) and external (chemical offensives) factors disturb the network. Therefore this modularity symbolizes division of action among the complexes and offers robustness. These algorithms, in general, complement experimental methods in a variety of ways.

These strategies have helped to overcome some of the limitations of proteomic research, such as removing false interactions by interaction scoring and boosting actual interactions through prediction of missing interactions. These algorithms have resulted in the addition of new protein complexes and protein interactions to proteomics databases, which has helped to expand our resources and understanding in the field.

2.6.1. The Characteristics of PIN

The purpose of abstracting a system's components to proteins and connections between them is to depict relationships in biological systems as networks. While such representations enable modeling and analysis using abstract computer methods, certain characteristics of such modeling are especially crucial for biological networks (Dančík et al., 2013). The following are some of the characteristics of PPI networks:

1) Small world PIN: PIN exhibits a small world effect, indicating that proteins are highly connected. No matter how large the network is, any node can reach another in a minimal number of steps. This usually means that any two nodes are separated by less than six steps. This is called "six degrees of separation" (Watts, 2004). Because it provides for an efficient and rapid flow of signals within the network, this level of connectedness has crucial biological implications. In addition, biological systems are extraordinarily resilient and can withstand a large number of mutations in single genes or proteins.
2) Power law for PIN: Many proteins in PIN have few connections, while only a few proteins have many connections. These are called hub proteins (proteins with more than fifteen connections) (Tsonis, 2007). Because hub proteins are crucial for the operation and stability of the interactome, it is no surprise that

their disruption in the human is commonly linked to disease (Han et al., 2004, Vidal et al., 2011).

3) Scale free PIN: PIN is a scale-free network. As noted above, the majority of proteins have a few connections to other proteins, while others (hubs) are connected to many proteins. The scale-free nature of PIN provides it with a number of features:
    - Stability: If failures happen randomly and the majority of proteins have a low degree of connection, the chances of a hub being impacted are low. Due to the residual hubs, the network will often not lose connectivity if a hub fails.
    - Changes in scale have no effect on the network properties: No matter what proteins and interactions a PIN contains, the PIN properties remain stable. However, hubs in a PIN are under the small-world effect regardless of PIN's size.
    - Sensitivity to attack: When some hubs are removed from PIN, it becomes a collection of subgraphs. However, the majority of hubs are essential genes/proteins which are cancer related.

Some studies have questioned whether PIN is a scale-free network (Stumpf et al., 2005) because of PPI dataset limitations that produce PIN with missing interactions.

4) PIN is also known for their modularity, which is another important feature. A network's transitivity measures the proteins' inclination to subgraph together. High transitivity indicates that PIN has tightly connected subgraphs. Finding these subgraphs in PIN is critical because they can reflect PCs and functional modules.

2.6.2. The Importance of PIN Analysis

PIN analysis is momentous for the following purposes:

- Understanding of the cause and development mechanisms of disease (Fionda et al., 2009).
- Discovering new protein functions (Fionda et al., 2009).
- Predicting functional and interaction modules (Fionda and Palopoli, 2011).
- Finding the essential proteins that are ideal candidates for further therapeutic target screening (Childs and Larremore, 2021).
- Discovering proteins in bottleneck or critical places that are vital to the PIN's overall integrity (Yu et al., 2007).
- Identify driver proteins that are crucial for the management of the underlying of PIN and gain a better understanding of PIN's functional mechanism (Wuchty, 2014).
- Understanding that the PIN interactions are important for a variety of applications such as development of drugs and therapeutic target identification (Queiroz et al., 2020).

2.6.3. Static and Dynamic Protein Interaction Networks

Massive-scale PPI information produced through high-throughput strategies gives maps of molecular networks for numerous organisms (Uetz et al., 2000). These PPIs allow researchers to build protein interaction networks and promote many computational methods to identify protein complexes. The majority of these techniques rely purely on clustering PINs or combining them with biological data such as gene expression data, functional annotation and functional orthology information (Wang et al., 2018a).

While there has been great progress in computational study of proteomescale cellular networks, the inherent dynamism of protein interactions within these networks is sometimes disregarded (Przytycka et al., 2010). The interactions of Biomolecules are changed across time, environment and different phases of celiac development, making cellular systems highly dynamic and responsive to environmental stimuli (Przytycka et al., 2010). As a result, it is important to transit the analysis of PPI networks from static to dynamic (Yong and Wong, 2015) since the dynamic PIN helps to show how disease progression is reflected in the time-evolving of PIN, and aids in disease identification before clinical symptoms appear (Wang et al., 2014).

The available PIN datasets are static and lack temporal and condition characteristics for protein interactions. As a consequence, dynamic information on PPI and protein complexes are neglected (Jenghara et al., 2018). Because the PIN is static, other information such as gene expression data should be used to develop dynamic networks in order to make use of their dynamic properties. In this area, microarray gene expression data provides useful dynamic information to model the activities of proteins at any given time and produce dynamic PIN. The protein is not always active, so it is important to determine the time point at which a protein demonstrates activity before constructing the dynamic protein interaction network (Zhang et al., 2019).

Identifying each protein's active time point by using GE data is momentous to construct DPIN (Zhang et al., 2016). Figure 2-5 explains the construction of DPIN by determining the active time points for each proteins in SPIN, then the DPIN is constructed for each time point by keeping only the proteins which are active in that time point and they have interacted in SPIN.

(de Lichtenberg et al., 2005) built DPIN over the yeast mitotic cell cycle by defining each protein's active time point. When the level of gene expression in a protein surpasses a certain threshold, it is considered active. To filter noise from gene expression patterns, (Tang et al., 2011) used a global threshold for all proteins. They also created a network of time-series PIN (called TC-PIN). It is illogical to use a single threshold to identify the activity/passivity of all proteins (Wang et al., 2011).

Wang et al. (Wang et al., 2013) devised a three-sigma (3-sigma) method that considers each protein's own distinctive expression curve to identify its active time points instead of using a global threshold for all proteins. After that, DPINs are built. A protein p is considered active at a specific time point if its value is equal to or greater than Pthreshold(p) :

$$P_{threshold}(p) = \mu(p) + 3\sigma(p)\big(1 - F(p)\big) \qquad (2-2)$$

Where the mean of p's gene expression values is $\mu(p)$ and $\sigma(p)$ is its standard deviation. $F(p)$ is defined as follows:

$$F(p) = \frac{1}{1+\sigma^2(p)} \qquad (2-3)$$

The number of protein complexes is large since the algorithms apply for each DPIN. 3-sigma adopts a filtering strategy where the protein complexes are arranged in decreasing order according to their size. The first undeleted protein complex is compared with other smaller undeleted protein complexes according to similarity scores (SS) Equation (2-4). If the SS is greater than 0.65, the large protein complex is retained, whereas the small protein complex is deleted. The results show superiority of 3-sigma method.

$$SS(PC_1, PC_2) = \frac{PC_1 \cap PC_2}{Max(PC_1, PC_2)} \qquad (2-4)$$

Where ($PC_1 \cap PC_2$) is the number of proteins shared between two protein complexes divided by the maximum length of those complexes ($Max(PC_1, PC_2)$).

The 3-sigma method has been largely acknowledged in academic circles as a state-of-the-art strategy for creating DPIN (Ou-Yang et al., 2014) and has better performance in predicting protein complexes. This is proved by previous studies (Zhao et al., 2017). However, three sigma filters proteins with low expression values and does not perform well on low expression levels dataset (Liu et al., 2018).

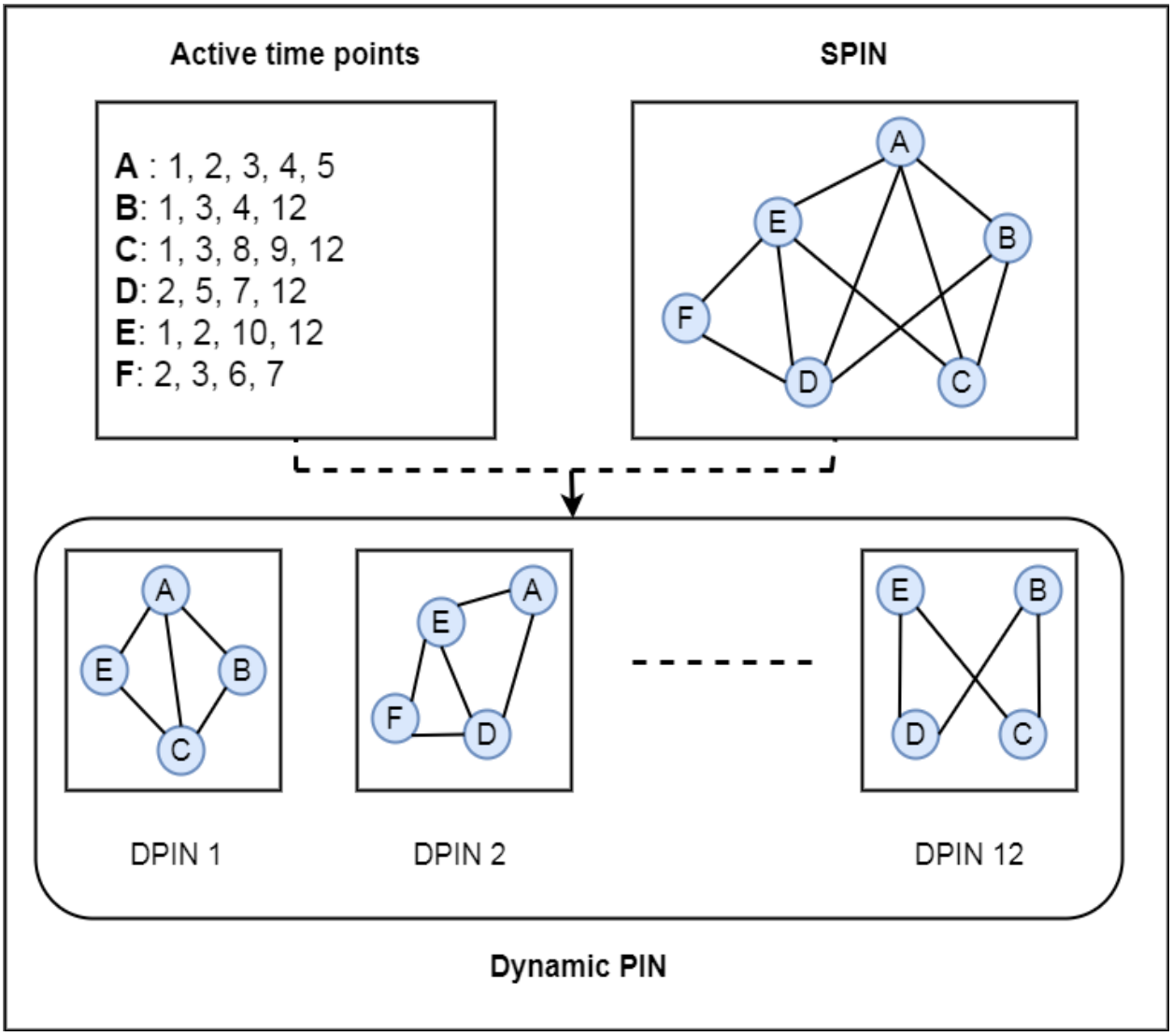


Figure 2-5: Construction of DPIN.

2.7. Computational Methods for Predicting Protein Complexes

During the past decade, scientific research has directed more attention to predicting protein complexes from PIN and developing numerous algorithms for dealing with them. Most of these algorithms involve the following steps to predict protein complexes from PPI datasets:

1) Combining datasets from different sources and setting the degree of confidence in interactions.
2) Using only the high confidence interactions to build PPI network.
3) Generating predicted complexes by searching for subnetworks from PIN.
4) Verifying and assigning the predicted complexes, as well as comparing them to reference complexes.

2.7.1. Module Based Prediction Algorithms

There are two types of modules algorithms that are followed to predict protein complexes (PCs) (Bhowmick and Seah, 2015):

1) Functional-based modules: the algorithm searches for a subgraph or molecular interactions whose proteins have comparable functional features that work together to produce a certain functional goal in a biological process.
2) Topological-based modules: the algorithm searches for dense subgraphs in PIN. In particular, nodes in a subgraph are more likely to connect to other nodes within the same subgraph than nodes outside it. It is worth noting that these modules are unconcerned about the function of individual proteins.

The algorithms are further differentiated at higher levels based on their methodologies for predicting PCs as shown in Figure 2-6. There are three ways in the topological-based approach. The first one is a seed-expanding technique, in which the algorithm searches for dense subgraphs based on a seed and then expands them heuristically. The algorithms use neighborhood information to weight nodes in a graph, then start with the node with the highest weight as a seed and expand by adding neighbors that have a weight which exceeds a certain threshold. Then the complexes with two proteins are ignored and merge complexes that overlap with specific numbers of proteins defined by the algorithm. This concept is being applied by MCODE (Bader and Hogue, 2003b), ClusterONE (Nepusz et al., 2012), SPICi (Jiang and Singh, 2010)and IPCA (Li et al., 2008a), which is the modification of the algorithm DPClus (Altaf-Ul-Amin et al., 2006b).

The second way is a clique merging technique in which the algorithm searches for cliques as researchers believe that a complete connected graph represents the protein complex such as Cliques Finder (CFinder) (Adamcsek et al., 2006), Clique Percolation- Distance Restriction (CP-DR) (Wang et al., 2010), Maximal Clique based (CMC) (Liu et al., 2009) and Local Clique Merging Algorithm (LCMA) (Li et al., 2005). All these algorithms detect cliques and then merge them depending on different criteria to identify the protein complex.

The third way is a network partition technique. The algorithms (MCL (Vlasblom and Wodak, 2009), SPC (Spirin and Mirny, 2003) and Pu et al. (Pu et al., 2007)) use a flow-based technique to divide a PIN into subgraphs. These techniques cover the entire PPI graph, but they are unable to incorporate annotation information, which would allow the algorithm to forecast PCs that are more compatible with reference complexes.

Although having a high coverage score is a benefit for algorithm, the first two strategies have partial coverage for PPI graph since it looks for localized dense areas in PIN.

Functional-based approaches are divided into two parts. Gavin et al. (Gavin et al., 2006) reported that protein complexes are organized as core proteins and attachment proteins as shown in Figure 2-7. The core represents the proteins that strongly interact with each other, whereas the attachments mark the boundary of the core. Several methods based on this technique have also been proposed, such as CORE (Leung et al., 2009), COACH (Wu et al., 2009), and WCOACH (Kouhsar et al., 2016) (which improved COACH by using a GO to weight the PPI network), CAG (Keretsu and Sarmah, 2017) and CO-DPC (Xiao et al., 2019). These algorithms contain two phases, namely, identifying the complex core as a dense sub-graph, and then identifying the attachment proteins as those that interact with more than half the proteins in the core.

The functional homogeneity approach is being used to improve complex prediction. Proteins in a complex are enriched for functions that are the same or comparable. When compared to random collections of dense subgraphs, dense protein clusters with strong functional coherence are more likely to correspond to reference complexes. This concept is being applied by the RNSC (King et al., 2004), DECAFF (Li et al., 2007) and WEC (Keretsu and Sarmah, 2016) algorithms. These algorithms are either divided the biological network into biological related subgraphs, or grown and merged by adding proteins that have strong functional coherence to identified protein complexes.

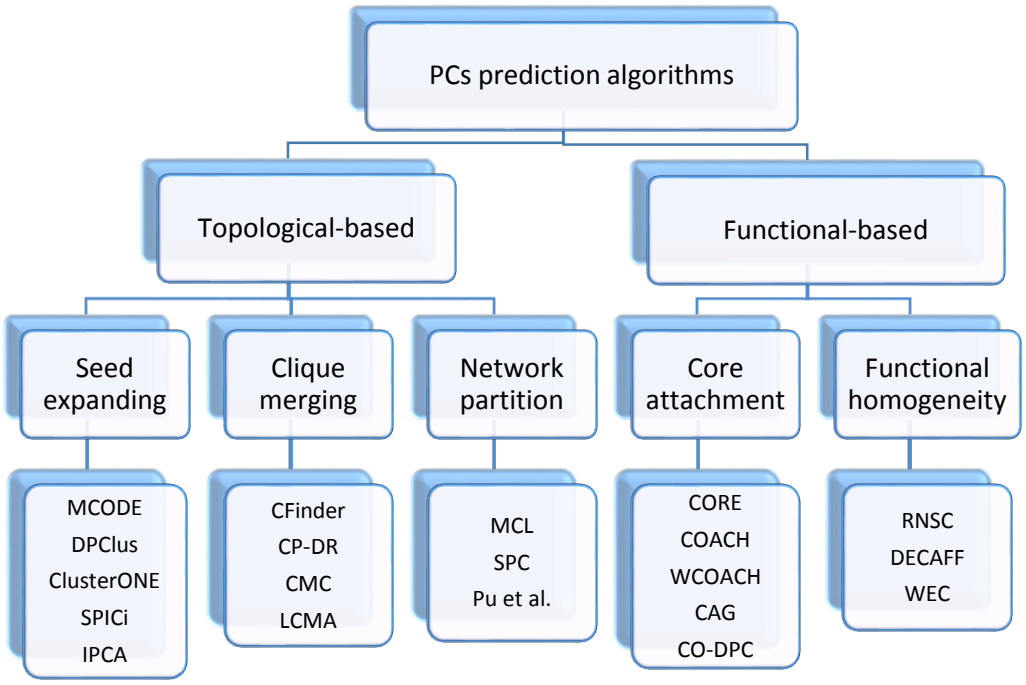


Figure 2-6: Classification of protein complex prediction algorithms

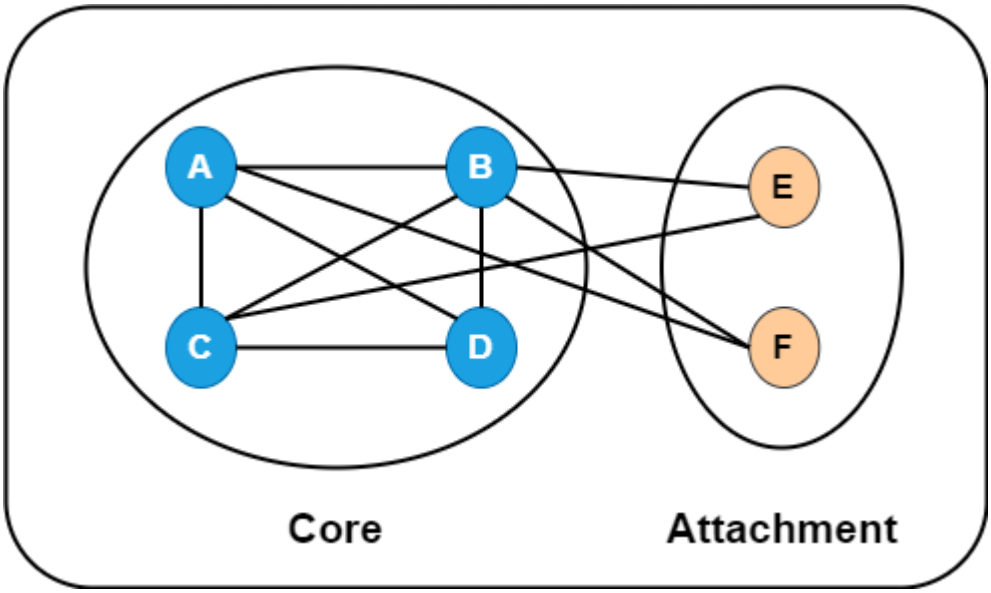


Figure 2-7: Example of core-attachment technique

2.8. Graph Mining

Graph mining is the technique used for analyzing the properties of the graph. The tasks of graph mining are:

- Random walk: It is a technique of traversing from node to node by picking a neighbor at random. This procedure can also be used to find a small region around a seed in large graph.
- Connected regions: Determining the regions that are linked in most big graph is a crucial step.
- PageRank: It is a term that provides answers to the question: "What are the most important nodes in my graph?" It has been effectively used to analyze big networks.
- Diameter: It is the longest path required to connect 90% of all potential node pairs. Our intuition regarding short pathways between nodes is guided by this value.
- Triangles: This means finding connected triples of nodes, and it has a number of applications in large graphs. Finding triangles incident to a node, for example, can suggest whether the graph has a tendency to have interesting groups. It is used in various link prediction, recommendation, and anomaly detection algorithms.
- Node pair's conundrums: For big graphs, explicit all-pairs computations (shortest paths, commuting times and graph kernels) are infeasible. One of the open difficulties in massive graph mining is finding exact scalable techniques for these situations.

2.8.1. Graph Mining Applications

- Social network analysis

  Social network is represented as graph where the nodes are the persons and linkages that correspond to communications or relationships between these diverse people. The social network links can be utilized to find relevant communities, users with certain skill sets and the flow of information in the network.

- Web network analysis

  The World Wide Web is naturally arranged like a graph, with web pages serving as nodes and links serving as edges. The PageRank algorithm is the most well-known application that takes advantage of the web's connection structure (Page et al., 1999). The main premise behind the algorithm is that the number and importance of hyperlinks connecting to a website on the internet may be used to determine its importance.

- Biological and chemical network analysis

  In systems biology research, analyzing biological and chemical networks using graph-mining algorithms has become widespread. Graph mining may aid drug design by revealing chemical and biological features like activity, metabolism, absorption, toxicity, and so on (Liu et al., 2007). As a result, academia and the pharmaceutical sector have increased their efforts in chemistry and biological graph mining in the hopes of drastically reducing drug discovery time and cost (Selvaraj et al., 2021).

2.8.2. Graph Mining in PIN

The biological system can be understood by analyzing a complex system as a whole, not only as discrete components. This can be accomplished by looking at the interactions of its basic components which are best described as networks that are primarily depicted as graphs with hundreds or thousands of nodes (Pavlopoulos et al., 2011).

A PIN is represented as an undirected and unweighted graph G = (P, E), where P is the nodes representative of proteins and E is the edges which represent the interactions between the proteins, or using biological data such as GO and GE to weight the graph G = (P, E, W), where W is the edge's weight as shown in Figure 2-8. This section will briefly outline major measures that have been utilized to predict protein complexes in PIN.

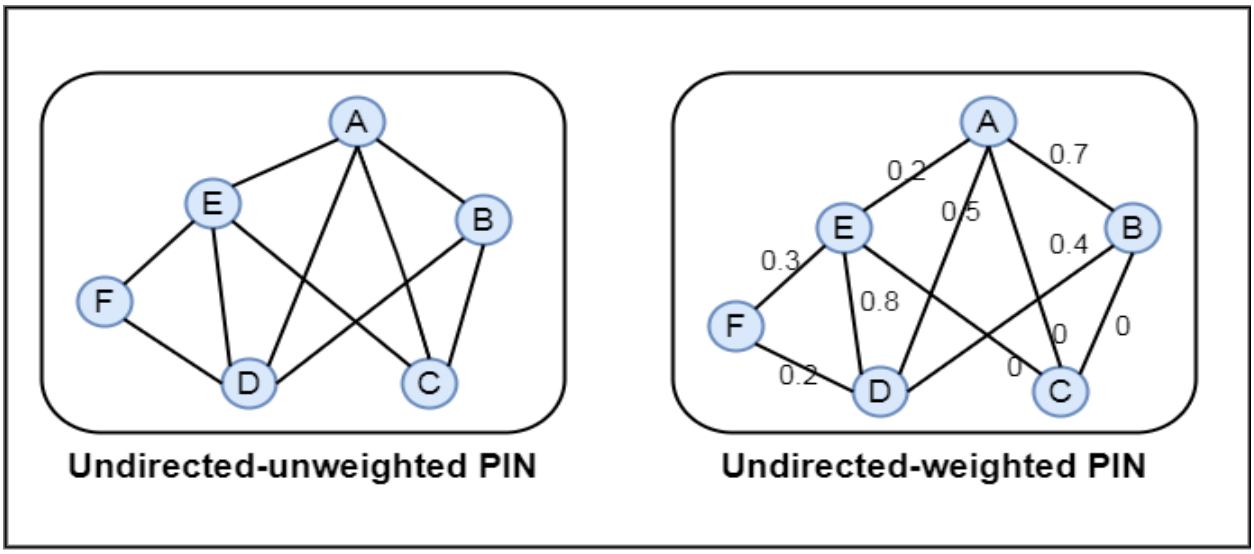


Figure 2-8: PIN's graph representation

- Protein neighbors: The neighbors of protein p are the set of proteins that interact directly with protein p and are denoted as $|N_p|$.
- Protein degree: For each protein p, the degree of p is the summation of its connected edges Equation (2-5).

$$d(p) = \sum_{i} e_i \qquad (2-5)$$

- Common neighbours (CN): The CN between two proteins (p_i and p_j) are the number of proteins that indices to both divided by the square root of the product of the nodes' degrees Equation (2-6).

$$CN = \frac{N_{P_i} \cap N_{P_j}}{\sqrt{d(p_i) * d(p_j)}} \qquad (2-6)$$

- Density: The density of the set of proteins $S \subset P$ is the number of edges among them divided by the number of possible edges between the set proteins (i.e., how close the set to the clique is, ranging between 0 and 1) Equation (2-7).

$$density(S) = \frac{2 * |E|}{|P| * (|P| - 1)} \qquad (2-7)$$

- Edge Clustering Coefficient (ECC): It is used to determine the closeness between two proteins in a network. ECC is defined in Equation (2-8) where Triangle (p_i , p_j) signifies the number of common neighbors between protein p_i and protein p_j divided by the minimum degree of the two proteins.

$$ECC = \frac{Triangle(p_i, p_j)}{\min\{d(p_i) - 1, d(p_j) - 1\}} \qquad (2-8)$$

- Cohesiveness: It determines the likelihood of a collection of proteins to form a protein complex. Cohesiveness is defined in Equation (2-9). Let PC be a protein complex, and let W_{in} be the edge weights accumulated between

proteins in PC. The PC proteins are connected to other proteins in the PIN. Let W_{out} be the edge weights accumulated for each edge between (p_i, p_j), where $p_i \in$ PC and $p_j \notin$ PC. Because of the limitations of experimental procedures, a penalty term (h|PC|) is utilized to indicate the uncertainty in PIN.

$$Coh = \frac{W_{in}(PC)}{W_{in}(PC)+W_{out}(PC)+h|PC|} \qquad (2-9)$$

- Conductance: The proteins inside PC heavily interact with each other, while they interact less with the proteins outside the PC. The conductance of PC is defined in Equation (2-10), where φ(PC) = |{(p_i, p_j) | $p_i \in$ PC, $p_j \notin$ PC}| and *vol*(PC) = $\sum_{P_i \in PC} d(p_i)$.

$$Cond(PC) = \frac{\varphi(PC)}{\min(vol(PC), 2|E| - vol(PC))} \qquad (2-10)$$

- Edge betweeness: It is a well-known metric for assessing the topological properties of numerous networks. In PINs, edge betweeness is defined as the number of shortest pathways passing through an edge divided by the highest number of shortest paths passing through an edge Equation (2-11).

$$Edge - btw = 1 - \frac{|ShortPath_{ij}|}{ShortPath_{max}} \qquad (2-11)$$

- Modularity: Several studies have shown that subgraphs in PIN with high density or modularity frequently correlate to protein complexes. Therefore protein complex detection algorithms have emphasized on subgraph density or modularity (Ren et al., 2013). Other studies prove that virus outbreaks and synchronization are both highly dependent on PIN modularity (Arenas et al., 2006). The modularity of PC in PIN graph G is defined in Equation (2-12),

where I_{PiPj} is equal to one if (p_i, p_j) ∈ E, otherwise it is equal to zero. X_{PiPj} is equal to one if both p_i and p_j are in the same PC, otherwise it is equal to zero.

$$M(G, PC) = \sum_{P_i, P_j \in V} \frac{I_{P_i P_j} - d(p_i) * d(p_j)/2|E|}{1 - X_{P_i P_j}} \qquad (2-12)$$

- Closeness Score (CS): It is a measurement score used to attach the protein *p* to core (C) proteins or preminirary complex as obtained by Equation (2-13) by attaching the protein *p* that connects with half or more of the proteins in the candidate complex or core proteins.

$$CS(C, p) = \frac{N_p \cap |P_C|}{|P_C|} \qquad (2-13)$$

2.9. Evaluation Metrics

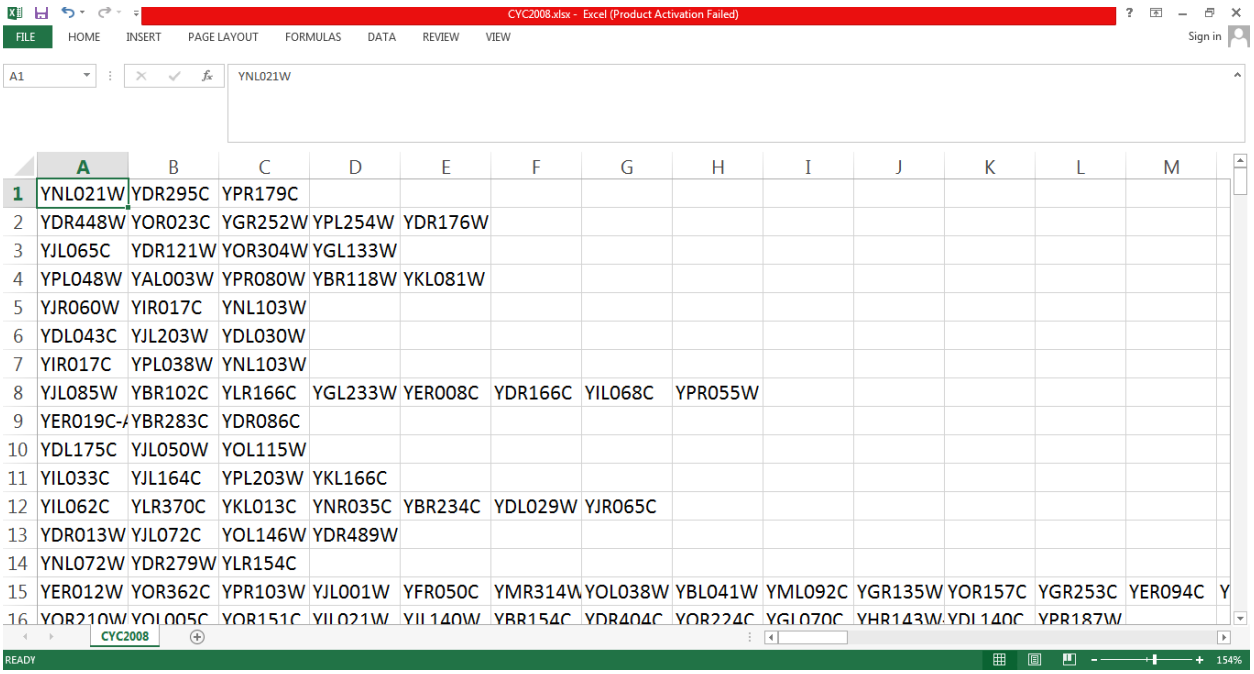
The quality of predicted complexes from each proposed algorithm is evaluated using various metrics.

2.9.1. Reference Datasets

Over the years, many high-quality resources for reference protein complexes have been established for yeast and human. NewMIPS (Mewes et al., 2004) and CYC2008 (Pu et al., 2009), MIPS (Mewes et al., 2004), Aloy (Aloy et al., 2004), SGD (Dwight et al., 2002), and CORUM are some examples of reference complexes as shown in Figure 2-9, each of which has a different number of them.

The overlapping score (OS) is the matching score between the predicted complex (PC) and the reference complex (RC), as expressed in Equation (2-14). PC and RC are considered matched if the OS between both is equal to or greater than a specific threshold.

$$OS(PC, RC) = \frac{|PC \cap RC|^2}{|PC| \times |RC|} \qquad (2-14)$$



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	YNL021W	YDR295C	YPR179C										
2	YDR448W	YOR023C	YGR252W	YPL254W	YDR176W								
3	YJL065C	YDR121W	YOR304W	YGL133W									
4	YPL048W	YAL003W	YPR080W	YBR118W	YKL081W								
5	YJR060W	YIR017C	YNL103W										
6	YDL043C	YJL203W	YDL030W										
7	YIR017C	YPL038W	YNL103W										
8	YJL085W	YBR102C	YLR166C	YGL233W	YER008C	YDR166C	YIL068C	YPR055W					
9	YER019C-A	YBR283C	YDR086C										
10	YDL175C	YJL050W	YOL115W										
11	YIL033C	YJL164C	YPL203W	YKL166C									
12	YIL062C	YLR370C	YKL013C	YNR035C	YBR234C	YDL029W	YJR065C						
13	YDR013W	YJL072C	YOL146W	YDR489W									
14	YNL072W	YDR279W	YLR154C										
15	YER012W	YOR362C	YPR103W	YJL001W	YFR050C	YMR314W	YOL038W	YBL041W	YML092C	YGR135W	YOR157C	YGR253C	YER094C
16	YOR210W	YOL005C	YOR151C	YIL021W	YIL140W	YBR154C	YDR404C	YOR224C	YGL070C	YHR143W	YDL140C	YPR187W	

Figure 2-9: CYC2008 reference dataset

2.9.2. Recall, Precision and F-measure

One of the most commonly used metrics to evaluate any algorithm is recall, precision and F-measure. Let PC = {pc1, pc2, pc3, …. , pc_n} a set of predicted complexes and R = {r1, r2, r3, ….., r_n} a set of reference complexes. The recall (R), precision (P) and F-measure (F) are defined as follows:

$$Precision = \frac{|\{pc \mid pc \in PC, \exists r \in R, \alpha(r, pc) \geq \theta\}|}{|PC|} \qquad (2-15)$$

$$Recall = \frac{|\{r \mid r \in R, \exists pc \in PC, \alpha(r, pc) \geq \theta\}|}{|R|} \qquad (2-16)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (2-17)$$

$\alpha\ (r, pc)$ is either the OS between the reference and the predicted complexes or the Jaccard similarity (J-Sim) as shown in Equation (2-18) between the reference and predicted complexes is based on that of Yong & Wong (Yong and Wong, 2015). θ for OS is determined as 0.255 as in WEC and CAG or 0.2 as in previous studies (Bader and Hogue, 2003a, Altaf-Ul-Amin et al., 2006a), θ for J-Sim is 0.75 for yeast and 0.50 for human or other large complexes greater than 3 in size, while θ is equal to 1 for small complexes.

$$J - Sim(PC, R) = \frac{|\ P_{Pc}\ \cap\ P_R\ |}{|\ P_{Pc}\ \cup\ P_R\ |} \qquad (2-18)$$

2.9.3. Coverage Rate

Coverage rate (CR) evaluates the number of proteins that have been covered by the predicted complexes (Brohee and Van Helden, 2006, Friedel et al., 2008). CR is defined in Equation (2-19), where RC is the set of reference complexes, $maxcom_{ij}$ is the maximal common proteins between the i^{th} reference and j^{th} predicted complex divided by N_i protein numbers in i^{th} reference complex.

$$CR\ = \frac{\sum_{i=1}^{|\ RC|} Max_j\{maxcom_{ij}\}}{\sum_{i=1}^{|\ RC|} N_i} \qquad (2-19)$$

2.9.4. Exact and Good Matching with Real Complexes.

The quality of predicted complexes is evaluated by reporting the number of reference complexes that exactly match with the predicted complexes and that have an OS score greater than or equal to 0.8, excluding the exact match.

2.9.5. Co-localisation Score and Gene Ontology Semantic Similarity Score

The predicted complex might be valid, but might not match any reference complex since the standard set of reference complexes is not complete (Jansen and Gerstein, 2004). The co-localisation score, Equation (2-20), is allowed to modify the quality of predicted complexes that do not match the reference complexes. The basic idea of the co-localisation score is that the proteins in the complex should be found in the same cellular compartment (Jansen et al., 2003) and more than likely to be involved in the same function.

$$CO - localisation\ score = \frac{\sum_i max_j\ N_{ij}}{\sum_i |C_I|} \qquad (2-20)$$

Where N_{ij} is the number of proteins in the C_i complex that is assigned to j, the localisation group, and $|C_i|$ is the number of proteins in C_i complex assigned to localisation.

The gene ontology (GO) annotation is used to evaluate the functional similarity between the proteins of the complex. The functional similarity between proteins is based on measuring the similarity in GO terms that annotate these proteins (Wang et al., 2007). Accordingly, the greater the score of GO semantic similarity, the better the quality of the predicted complex is.

The Procope software tool (Schlicker et al., 2006) as shown in Figure 2-10 is used for co-localisation and GO semantic similarity scores and the data used in the evaluation process is set to 'default'.

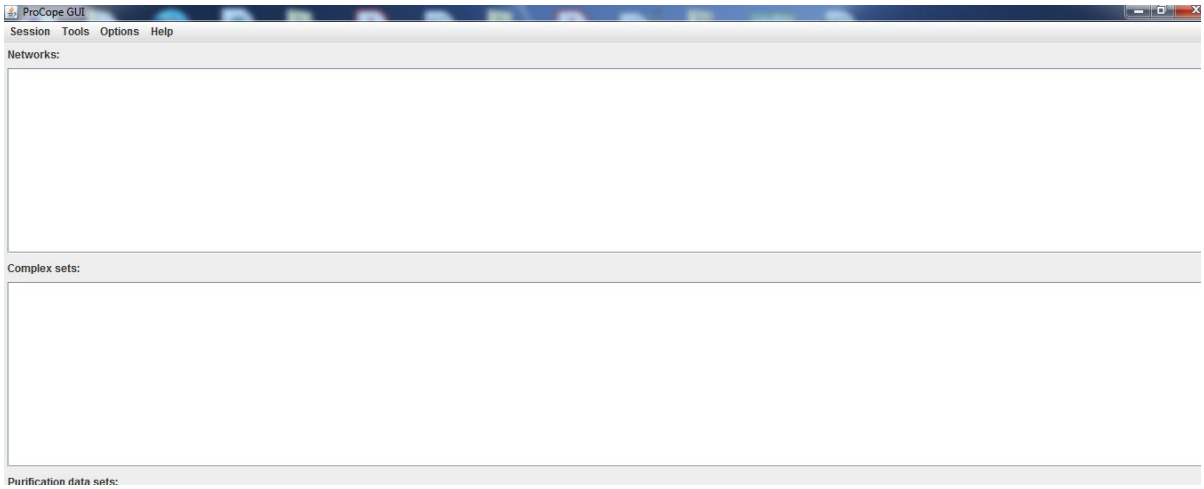


Figure 2-10: ProCope software tool

2.10. Visualization Tool of PIN

Cytoscape software (Shannon et al., 2003) as shown in Figure 2-11 is a free tool which is based on JavaTM technology and used to visualize networks and biological pathways, as well as integrate them with annotations, gene expression, and other biological data. Despite its origins in biological research, Cytoscape has evolved for sophisticated network analysis and other networks visualization.

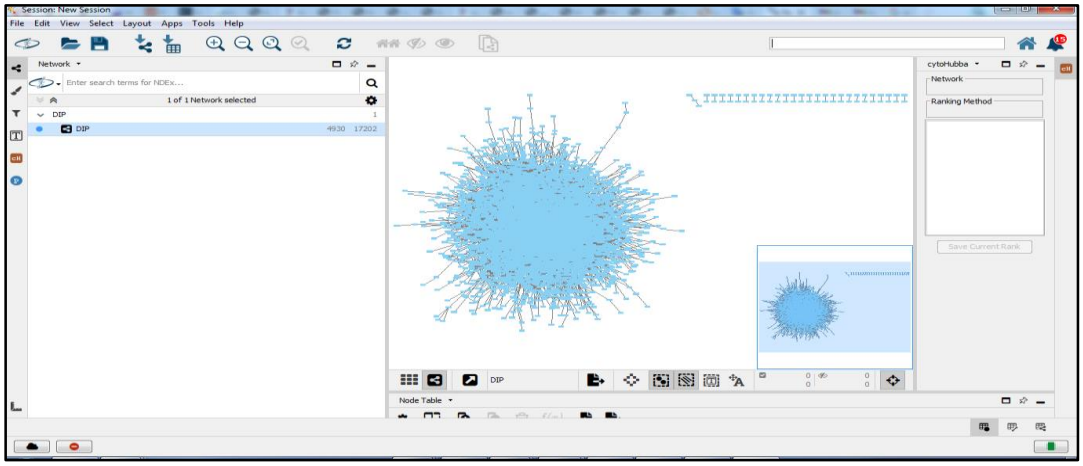


Figure 2-11: Cytoscape software

Chapter Three Proposed Algorithms for SPIN and DPIN

3.1. Introduction

This chapter focuses on the design and implementation of proposed methodology for predicting protein complexes. It is primarily composed of two key strategies, each of which employs a distinct algorithm for predicting protein complexes in a large protein network graph. The first strategy makes use of network clustering and extra biological information to predict protein complexes in a static network. The second strategy, which utilizes a gene expression dataset to detect the active proteins in each network, predicts protein complexes in a dynamic network.

3.2. Architecture of the proposed methodology

The suggested methodology architecture in Figure 3-1 depicts the design and development of the clustering for biological networks to predict protein complexes. It consists of three stages:

1- Network graph creation.

2- Protein complex predicting in the static and dynamic network.

3- Evaluating metrics.

The PPI dataset is turned into a network of nodes and linkages, from which the actual graph is constructed. After that, the algorithms are used to detect protein complexes in enormous datasets of biological networks. Finally, to confirm the quality of identified protein complexes, the evaluation stage is carried out.

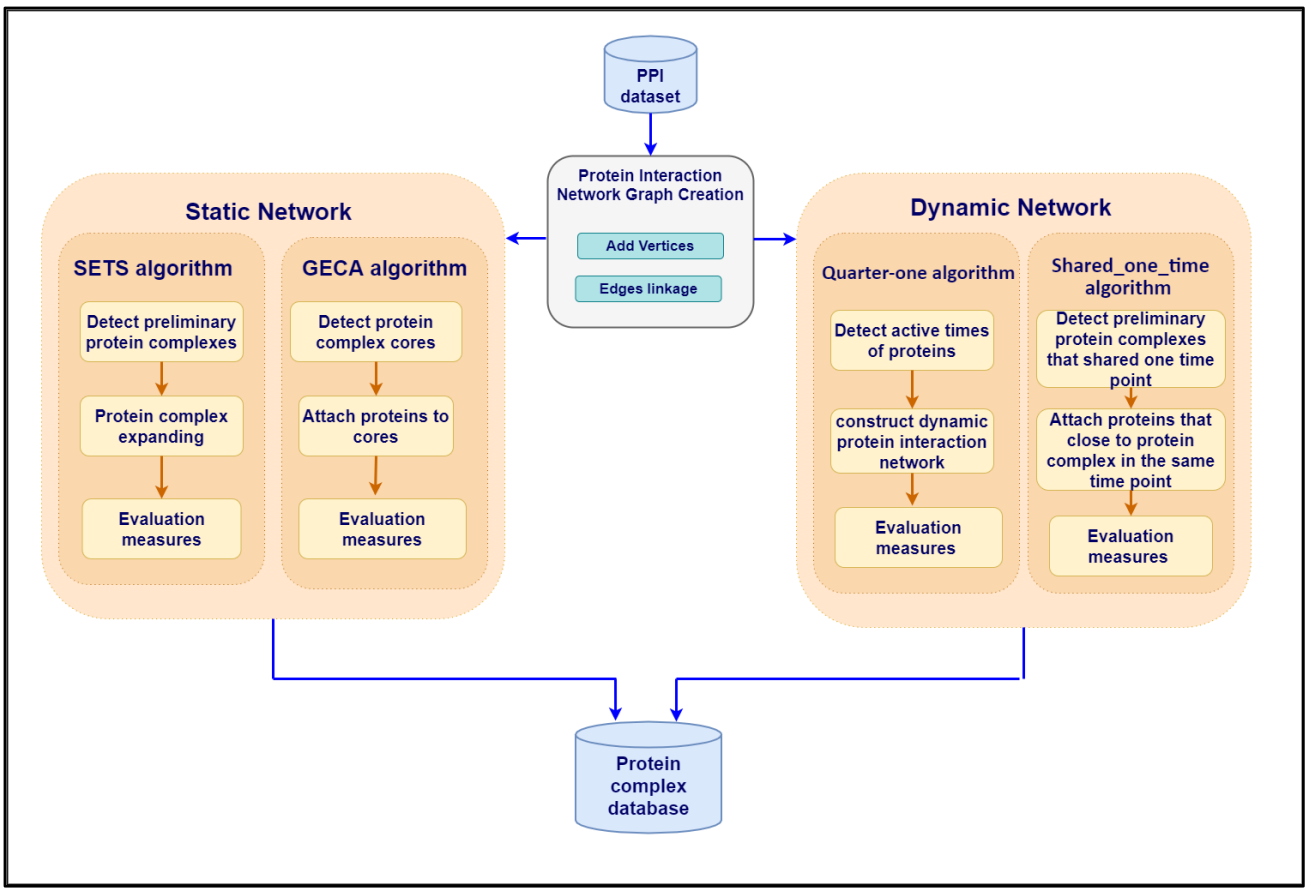


Figure 3-1: Block diagram of proposed methodology

3.3. Network Graph Creation

Generally, PIN is represented as an undirected and unweighted graph G = (P, E), where P is the nodes representative of proteins $\{p_1, p_2, \ldots, p_n\}$ and E is the edges $\{e_1, e_2, \ldots, e_n\}$, which represent the interactions between the proteins as in Figure 3-2. The strategy used to create graph employs the concept of classes as in the following steps:

1. Create class Protein that contains name, neighbours, degree as its attributes. It has Initialize() and Add_neighbour_degree() as its main methods.
2. Create class Graph that contains an instance of class Protein and perform

the main methods Add_Protein(), Add_Edge(), Get_Protein() and Get_Degree() to maintain the interactions between Protein classes.

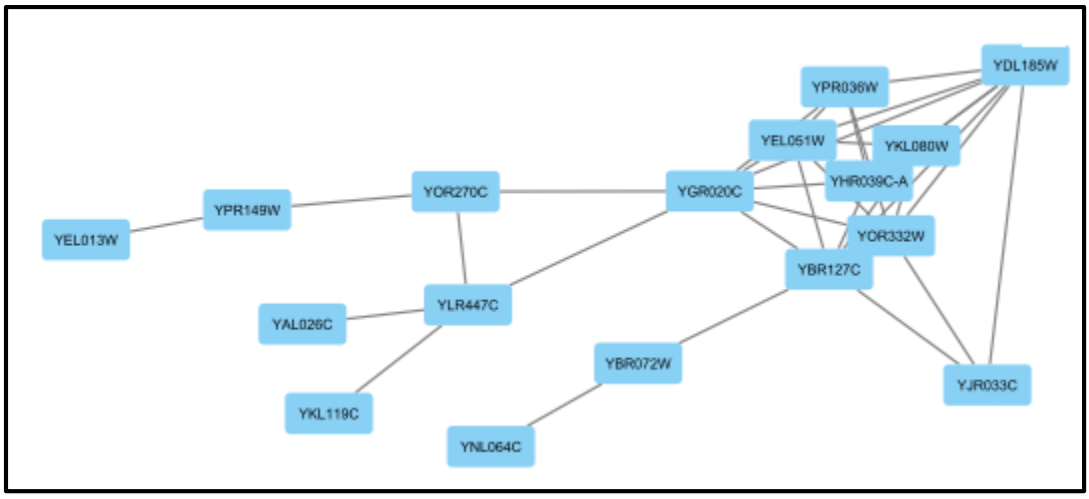


Figure 3-2: Snapshot of protein-protein interaction network

3.4. Protein complex identification

The proposed algorithms aim to predict protein complexes or biological structures by leveraging the features of protein-protein interaction networks, in which the protein complexes are a combination of two or more proteins. As a result, the suggested algorithms are aimed at determining the locations of protein complexes within the protein-protein interaction network.

3.4.1. Identifying protein complexes in static network

The PIN datasets that are available are static and are represented as an undirected graph, where the nodes are proteins and the edges are the interaction between the proteins. Many algorithms have been proposed for the analysis of PPI networks to discover the protein complexes by using topological properties of the

graph or combining the PPI network with other information like the gene expression. Two algorithms have been proposed to predict protein complexes in static network.

3.4.1.1 Seed expanding model and topological structure (SETS) algorithm

Most of the existing algorithms can detect only highly dense regions as protein complexes and ignore low density complexes. Further, most of them cannot detect overlapping protein complexes. Given an undirected and non-weighted graph, the overlapping protein complexes with different densities are predicted based on the seed expanding approach and topological structure of the PPI networks (SETS) in an acceptable amount of time as shown in Figure 3-3. The algorithm accomplishes this through the following steps (Algorithm 1):

1) **Ordering the proteins**

   Those proteins that have a degree of more than 1 are ordered in increasing order of their degree, put in a Set ‘Q’ and label each protein as not_visited.

2) **Building a preliminary complexes**

   The preliminary complex is built starting from the seed. The neighbours that share a specific ratio of common neighbours are added iteratively and they are labelled as visited if it satisfies the predefined threshold of shared neighbours in order to avoid selecting it as a seed in the next iteration. The complex is accepted as a preliminary complex if its density is greater than a predefined threshold; otherwise, all nodes label as not_visited and moved to the next node in the Q with not visited label. After defining the preliminary complex, no nodes will be deleted

from the Q, so that it can get overlapping complexes. The preliminary complex is accepted as a candidate complex if it contains more than three proteins and is not defined before from another seed in order to avoid redundancy (algorithm 1, steps 1–19).

3) **Expanding the preliminary complexes**

Preliminary complex is iteratively expanded according to the closeness score (CS) as obtained by Equation (2-13) by adding its neighbours 'N_{CC}' that connect with half or more of the candidate complex's proteins. The expansion is done in rounds. In each round, the algorithm searches for proteins that satisfy the threshold of closeness score (T_{CS}), then add them to the complex. In the second round, the algorithm searches for proteins, such as those that are connected with the updated complex until no proteins can satisfy T_{CS}. This step assists in the identification of complexes with different densities (algorithm 1, steps 20–31).

4) **Remove redundant complexes**

Redundant complexes are removed by retaining only one of the exactly matched complexes.

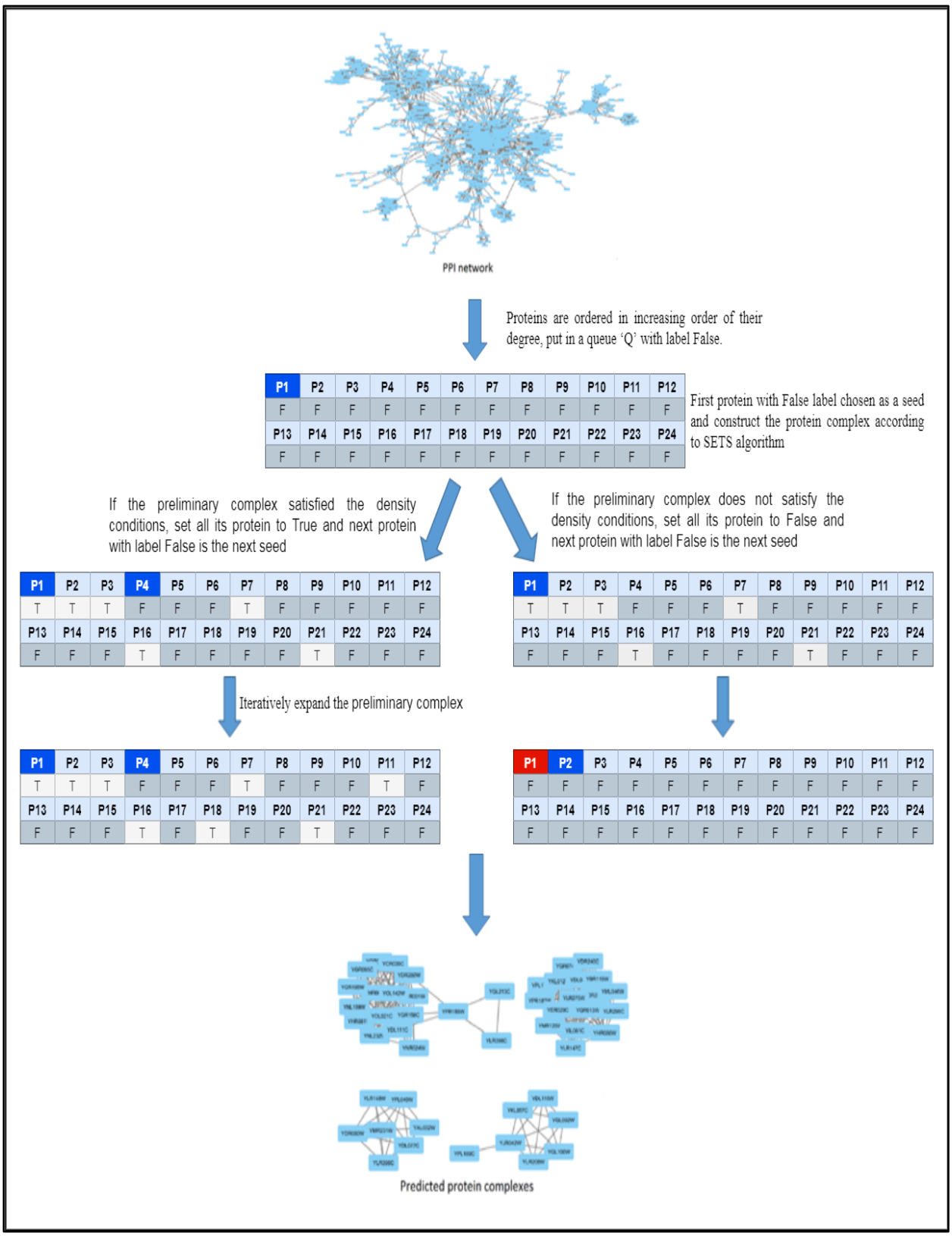


Figure 3-3: SETS algorithm technique

Algorithm 1: SETS

Inputs: Q, the set of ascending ordered proteins with label not-visited.

Output: COMPLEXES, the set of sets of predicted protein complexes.

1. **Let** $T_{CN,}$ DT, and T_{CS} be pre-defined thresholds.
2. **for** each protein *p* in Q **do** /* Building preliminary complexes */
3. **if** *p* is not visited **then**
4. Add *p* to complex set COMP
5. Label *p* node as visited
6. **for** each neighbour *n* of p **do**
7. Find the common neighbours (CN) between *p* and its neighbours
8. **if** $CN \geq T_{CN}$ **then** /* Eq. (2-6) */
9. Add *n* to COMP
10. Label *n* as visited
11. **Endfor**
12. **if** density(COMP) ≥ DT **and** COMP ***is not*** in COMPLEXES **then**
13. Add COMP to COMPLEXES
14. **else**
15. **for** each *p* in COMP **do**
16. Label *p* as not visited
17. **Endfor**

18. **Endif**

19. **Endfor**

20. **for** each COMP in COMPLEXES **do** /* Expanding preliminary complexes */

21. flag = True

22. **While** cp is not empty **or** flag

23. flag = False

24. candidate_protein cp set ← Ø

25. Find neighbours N_{CC} of COMP's proteins

26. **for** each protein *p* in N_{CC} **do**

27. **if** CS (COMP, *p*) ≥ T_{CS} **then** /* Eq. (2-13) */

26. Add *p* to cp

27. **Endif**

28. **Endfor**

29. Add cp's proteins to COMP

30. **EndWhile**

31. **Endfor**

32. Remove redundant complexes

A. SETS Mechanism

An effective technique for determining the relationship between proteins is provided in this algorithm, which is based on common neighbors. A PPI graph is created by constructing a class protein which contains name, degree and visited label that is set to not_visited. Then these proteins are connected with edges according to PPI dataset as shown in Figure 3-4. First, a pre-processing step needs to be undertaken to order the proteins according to their degrees in set Q. The algorithm begins with first protein with label not_visited (P16) to be the seed. Then, add their neighbors where CN with it are greater than or equal to 0.3. Only neighbor P8 satisfy the condition and their label is set to visited, P16 cannot construct a preliminary complex (COMP) since it contains two proteins. Therefore, the label of the seed and P8 are set to not_visited again.

The algorithm moves to the second protein with label not_visited (P14), adds its neighbors [P12, P13, P15] that satisfy the condition of CN and sets their label to visited. It is accepted as a preliminary complex since its density is equal to or greater than 0.5. Later, the preliminary complex is expanded by adding vertices that have closeness score equal to or greater than 0.5. They are P9, P11 and P10. All the nominate attachment proteins interact with at least two proteins out of four in the preliminary complex. The new complex is [p14, p12, P13, p15, p9, p11, p10]. In the second round of expansion, the algorithm checks the neighbors of the new complex, namely, P17, P8, P7, P6 and P3. But none of which interact with three proteins out of six in the complex. The algorithm accepts the complex [p14, p12, P13, p15, p9, p11, p10] as a final complex and moves to the next protein with label not_visited to be the next seed as shown in Figure 3-5.

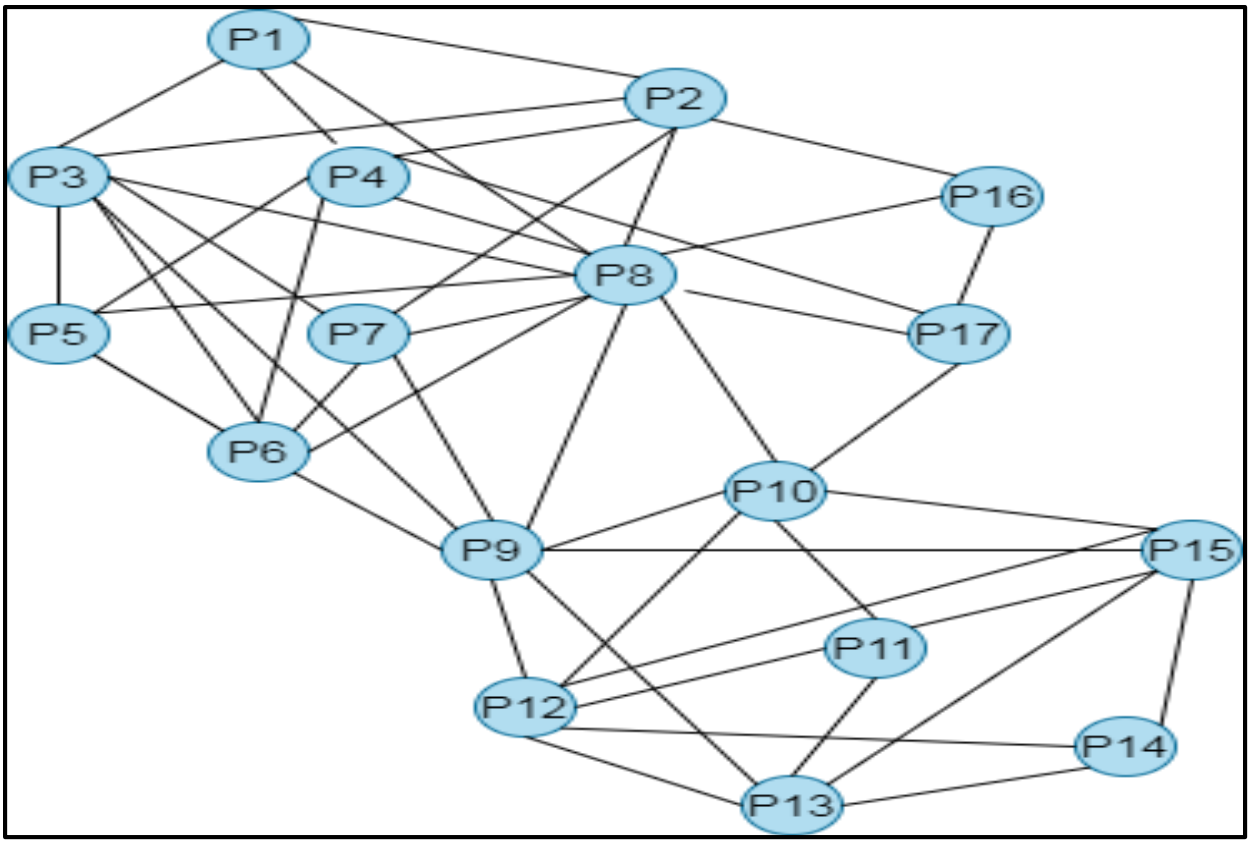


Figure 3-4: PPI network

P16	P14	P5	P17	P11	P1	P7	P13	P10	P12	P2	P15	P6	P4	P3	P9	P8
T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Seed	Neighbors	CN	PrCOMP	Density
P16	P2	0.236	[P16]	
	P8	0.348	[P16, P8]	
	P17	0.289	[P16, P8]	less than three

P16	P14	P5	P17	P11	P1	P7	P13	P10	P12	P2	P15	P6	P4	P3	P9	P8
F	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Seed	Neighbors	CN	PrCOMP	Density	Expand 1	COMP	Expand 2	Final complex
P14	P12	0.471	[P14, P12]					
	P13	0.516	[P14, P12, P13]					
	P15	0.471	[P14, P12, P13, P15]	1.0	P9	[P14, P12, P13, P15, P9]		
					P11	[P14, P12, P13, P15, P9, P11]		
					P10	[P14, P12, P13, P15, P9, P11, P10]	P17	
							P8	
							P7	
							P6	
							P3	[p14, p12, p13, p15, p9, p11, p10]

P16	P14	P5	P17	P11	P1	P7	P13	P10	P12	P2	P15	P6	P4	P3	P9	P8
F	T	T	F	T	F	F	T	T	T	F	T	F	F	F	T	F

Figure 3-5: Trace with example of SETS algorithm

3.4.1.2 Gene Expression and Core-Attachment (GECA) algorithm

SPIN is mixed with gene expression as a biological information to increase the accuracy of the prediction. GECA is a new algorithm to identify core proteins using common neighbour (CN) techniques and biological information (PCC). Here CN captures the topological relation between any node and its neighbours, while PCC captures the dynamic properties of the proteins by using the gene expression over a time series. Moreover, GECA improves the attachment technique by adding the proteins that have low closeness but high similarity to the gene expression of the core proteins as shown in Figure 3-6. In addition, the edges of the network are weighted by the value of PCC that represents the degree of similarity between the gene expression patterns of the two proteins if the gene expression data is available for both proteins. Otherwise the weight is set to zero. In this algorithm, protein complexes are predicted in three steps, namely, core construction, merging cores and adding an attachment, as described below:

Core construction: The proteins in the core of a protein complex constantly interact and the gene expression between them is highly correlated. The procedure begins with a seed *p* as a preliminary core, where neighbours of the seed are added either if the CN $\geq T_{core}$ (T_{core} is the threshold that determines the closeness between the seed and its neighbours), or if the PCC is greater than T_{ppc}, according to the correlation between the two proteins. The core contained the proteins that are either closely connected or correlated in gene expression. The core is accepted if its density is greater than the density threshold T_{dense}, which is set to 0.7. The use of PCC in the core allows the addition of peripheral protein that has an expressed relationship with the seed. The process of core identification is done by applying Algorithm 2.

Algorithm 2: Core_construction

Input: WG, weighted PIN graph by PCC; T_{core}; T_{pcc}

Output: CORES, the set of Cores

1: Let N_p is the set of protein neighbours

2: **for** *p* in WG **do**

3: Temp_Core ⟵ Ø

4: Add *p* to Temp_Core

5: **for** each n ∈ N_p **do**

6: **If** CN (*p*, n) ≥ T_{core} **or** PCC (*p*, n) ≥ T_{pcc} **then**

7: Add n to Temp_Core

8: **Endfor**

9: **If** Temp_Core **not in** CORES **and** |Temp_Core| > 2 **then**

If density (Temp_Core) ≥ T_{dense} **then**

10: Add Temp_Core to CORES

11:**Endfor**

Merged cores: Algorithm 2 identifies the cores which have many overlapping proteins. It is sufficient to merge highly overlapping cores since the density of the cores is equal to or greater than 0.7. When the overlapping scores (OS) between two cores, C1 and C2 are greater than the $T_{filtering}$, which is set to 0.9, the two cores merge. When the core does not overlap with other cores, it is kept as a core in the list of cores.

Adding the attachment*:* Most of the core-attachment-based methods followed the CoAch algorithm procedure, in which a protein is added to the core if it has a direct

connection with half of the core proteins. However, in reference complexes, the proteins in the attachment interact with fewer than half of the core proteins. The process of adding proteins to the core is ineffective and requires further information that can be obtained from the gene expression, which provides a more meaningful biological results. In this algorithm, however, the protein is added to the core so long as it meets one of the following two conditions:

(1) CS is greater than $T_{closeness}$, where N_p represents the direct neighbourhoods of p and P_{core} represents the proteins in Core.

OR

(2) The interWeight Equation (3-1) is equal to or greater than coreWeight Equation (3-2) and node p has at least n connections with the core.

$$interWeight(u, core) = \frac{\sum_e w(e)}{|N_p| \cap |P_{core}|} \qquad (3-1)$$

$$coreWeight(core) = \frac{\sum_e wg(e)}{|P_{core}|} \qquad (3-2)$$

Where $\sum_e w(e)$ in Equation (3-1) is the sum of weights of all interactions between protein p and the core proteins normalized by the number of proteins in the core that protein p interact with. While $\sum_e wg(e)$ in Equation (3-2) is the sum of all weights in cores normalized by the number of the core proteins.

A protein is added if it highly interacts with the proteins in the core, or has a good similarity with them in the gene expression pattern. The aim here is to improve the attachment technique in order to add proteins that interact with less than half of the core proteins but have similar gene expression patterns. Finally, the set of predicted complexes is defined as expressed in Algorithm 3.

Algorithm 3: Adding an attachment

Input: set of CORES

Output: PC, set of Predicted Complexes

1: **for** core in CORES **do**

2: proteins_not_in_core ⟵ Ø

3: **for** protein *p* in core **do**

4: difference = $N_{protein} \setminus$ core /* $N_{protein}$ is *p*'s neighbours */

5: **Endfor**

6: **for** protein in difference **do**

7: **If** (Closeness (core, protein) ≥ $T_{closeness}$) **or** (interWeight(protein, core) ≥ coreWeight(core) **and** ($N_{protein} \cap$ core) > $N_{connection}$) **then**

8: Add protein to protein_not_in_core

9: **Endfor**

10: **if** |protein_not_in_core| > 0 **then**

11: Add protein_not_in_core's proteins to core

12: **Endfor**

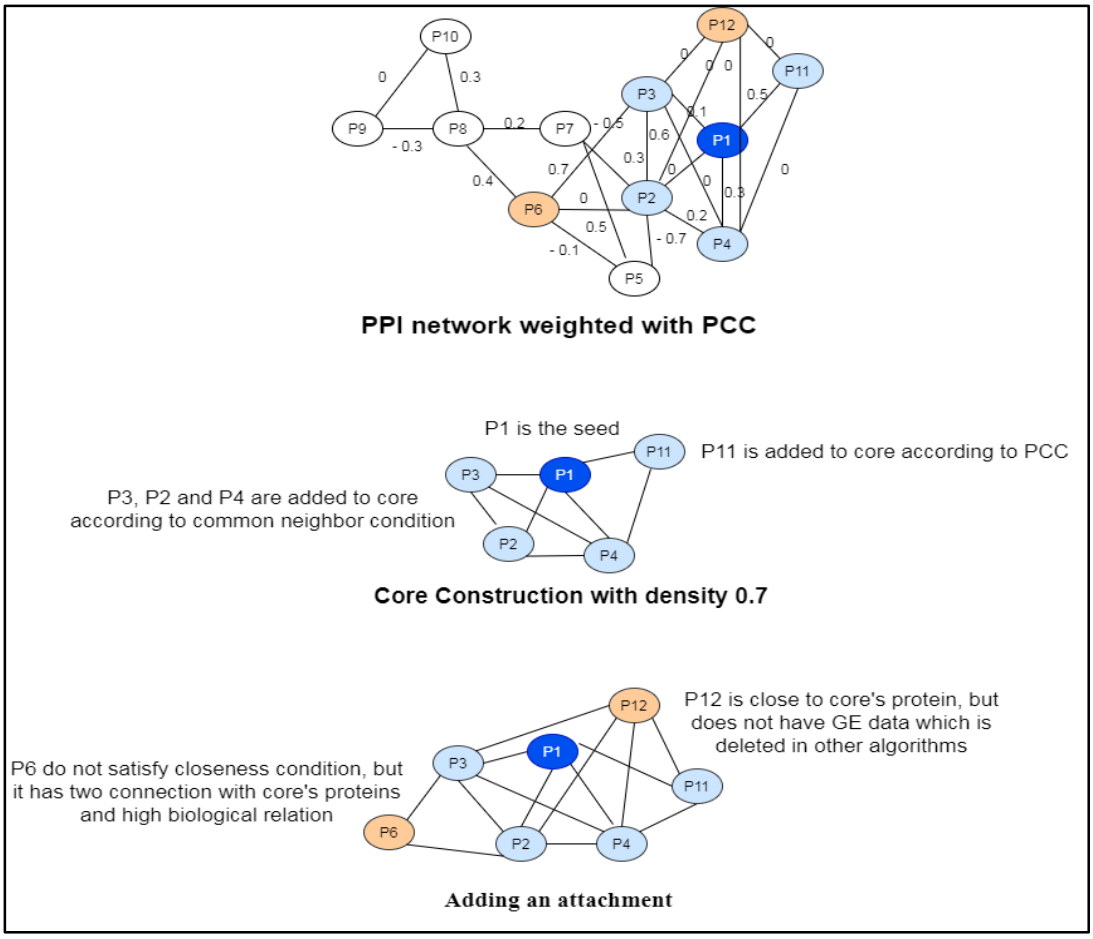
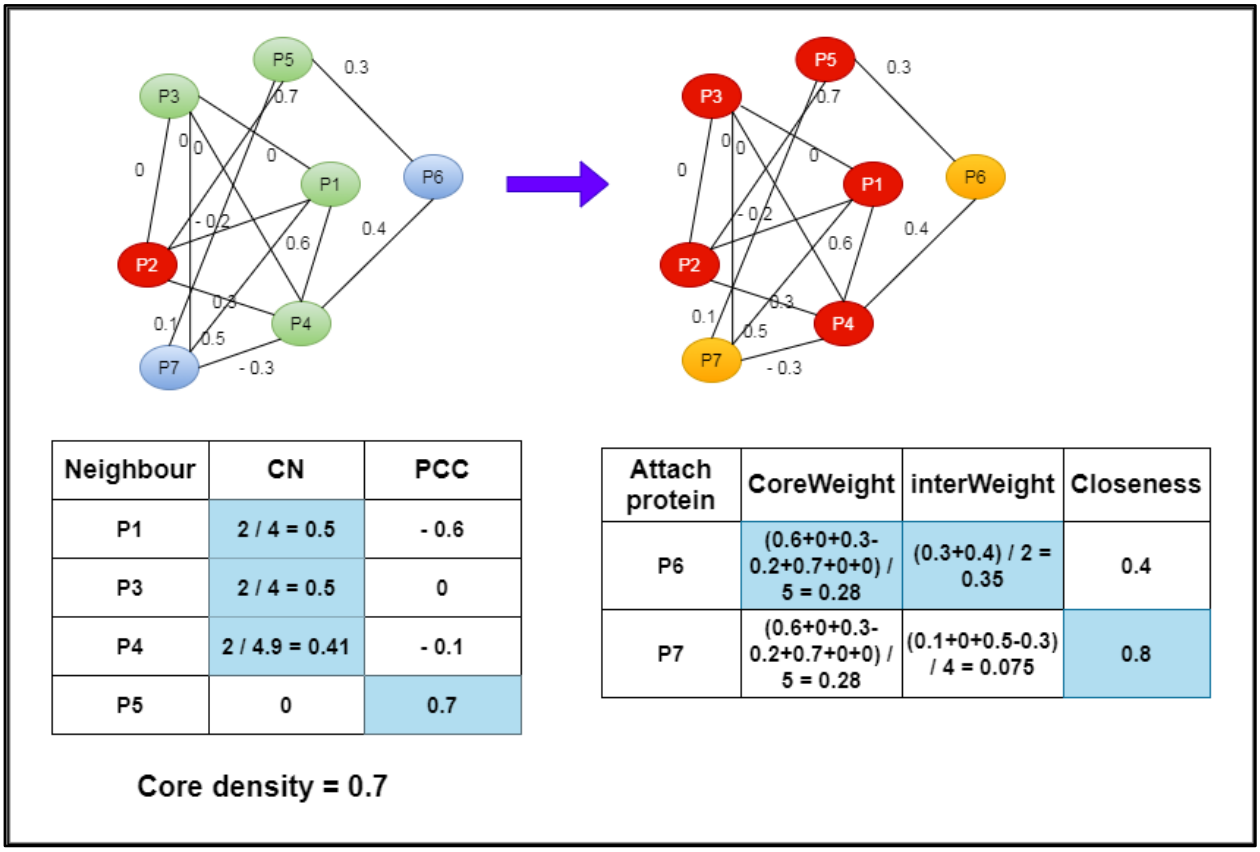


Figure 3-6: GECA algorithm technique

A. GECA Mechanism

A PPI graph is created by constructing a class protein which contains name, degree, and neighbors. Then these proteins are connected with edges according to PPI dataset. Gene expression profile contains the gene expression values in 36 times for each protein. GECA weighs the edges between proteins according to GEP by calculating the PCC between them, if both proteins have gene expression values in GEP. Otherwise weighs the edge with zero. GECA predicts protein complexes from a seed and considers each protein as seed. Figure 3-7 explains an example of prediction, beginning from P2 as a seed and examines its direct neighbor if they

accomplish any one of the core conditions. P1, P3 and P4 are added to seed P2 to construct the core by satisfying the CN condition. P3 does not have expression values in GEP, but GECA does not remove any protein to avoid losing important topological information like P3 that shares neighbors with the seed. P5 does not share any neighbor with the seed but it has high biological relation with it, so it added to the seed. The final constructed core [P2, P1, P3, P4, P5] is accepted as its density is 0.7. After constructing the core, all neighbors of the core proteins P7 and P6 are found. P7 interacts with four core proteins with high closeness, so it is attached to the core. P6 interacts with only two proteins from the core. Yet its interWeight is greater than the CoreWeight. This means a high biological relation with the core protein, so it is attached to the core.



Neighbour	CN	PCC
P1	2 / 4 = 0.5	- 0.6
P3	2 / 4 = 0.5	0
P4	2 / 4.9 = 0.41	- 0.1
P5	0	0.7

Core density = 0.7

Attach protein	CoreWeight	interWeight	Closeness
P6	(0.6+0+0.3-0.2+0.7+0+0) / 5 = 0.28	(0.3+0.4) / 2 = 0.35	0.4
P7	(0.6+0+0.3-0.2+0.7+0+0) / 5 = 0.28	(0.1+0+0.5-0.3) / 4 = 0.075	0.8

Figure 3-7: GECA mechanism

3.4.2. Identifying protein complexes in dynamic network

The available PIN datasets are static and lack temporal and condition characteristics for protein interactions. As a consequence, dynamic information on PPI and protein complexes is neglected. Because the PIN is static, other information such as gene expression data should be used to develop dynamic networks in order to make use of their dynamic properties. In this area, microarray gene expression data provides useful dynamic information to model the activities of proteins at any given time and produce dynamic PIN. A technique is proposed to construct a DPIN and proposed an algorithm to predict protein complexes in DPIN.

3.4.2.1 Construction of DPIN Based on Gene Expression Data and Quartile One Principle

Identifying the active time point of each protein by using GE data is critical to construct DPIN. Three sigma (3-sigma) method has been largely acknowledged in academic circles as a state-of-the-art strategy for creating DPIN. However, three sigma filters proteins with high expression values and does not perform well on low expression levels dataset. A new technique to construct DPIN is proposed to overcome the limitations of the 3-sigma method named quartile one (q-one).

The definition of adequate thresholds for detecting active proteins at different times is one of the thorny and contentious problems in creating DPIN based on gene expression data. As a result, a major goal is determining whether a protein is active or not at a specific time point.

Protein active time points and DPINs construction

The q-one method considers a protein as active at a specific time point if its value is equal to or greater than 25% of the protein's values, which is the first quartile of the protein's gene expression values in all time points. The choice of 25^{th} percentile of a given protein's expression curve is based on analyzing two gene expression data: GSE3431 and GSE4987. The number of genes and the number of their active times are calculated using different percentiles from 25 to 50. The percentile 25^{th} has the highest number of genes active in different time points in both gene expression data. The researchers realize that the genes active times have successive time points and the highest number of successive times is with 25^{th} percentile. The aim is to avoid filtering proteins with low gene expression values. Q-one suggests constructing the dynamic networks by taking into account these successive times. Thus to construct the first dynamic network, the protein is either active at time points one and two or at time points two and three. For the second dynamic network, its proteins are active either at time points three and four or at time points four and five, and so on as shown in Figure 3-8. The interactions in static PIN is kept in DPIN, if both of its proteins are active in the same dynamic network. As a result, the number of constructed DPINs is half the time points of gene expression data. For more analysis, this strategy applies again on two gene expression data and the percentile 25^{th} produces the best cover of genes in each network with a good distribution range and at the same time produces a pre-filtering stage of genes' active times.

Filtering strategy

After constructing DPINs, three algorithms have been used (Markov cluster method (MCL), clique percolation method (CPM) and ClusterONE) to evaluate the efficiency of the q-one method in predicting protein complexes from DPINs. The number of PCs from each algorithm is large since the algorithm is applied for each DPIN. To reduce the number of predicted protein complexes (PPCs), a filtering strategy is adopted based on another reference protein complexes (RPCs) analysis. It calculates the active times for each protein using the q-one method, and then computes the shared active times of each reference protein complex using two gene expression data. The results show that the majority of RPCs with two proteins share at least six time points. RPCs with three proteins, however, share at least two time points, while RPCs with four proteins and above share at least one time point. Based on this analysis, the q-one method filters the predicted PCs from all DPINs. The filtering strategy is based on the size of the protein complex. If the PC has two proteins, it is accepted as a predicted PC if it appears in at least six DPINs. A PC with three proteins are accepted as a predicted PC if it appears at least in two DPINs. Other PCs with more than three proteins are accepted as predicted PCs if they appear in any DPIN.

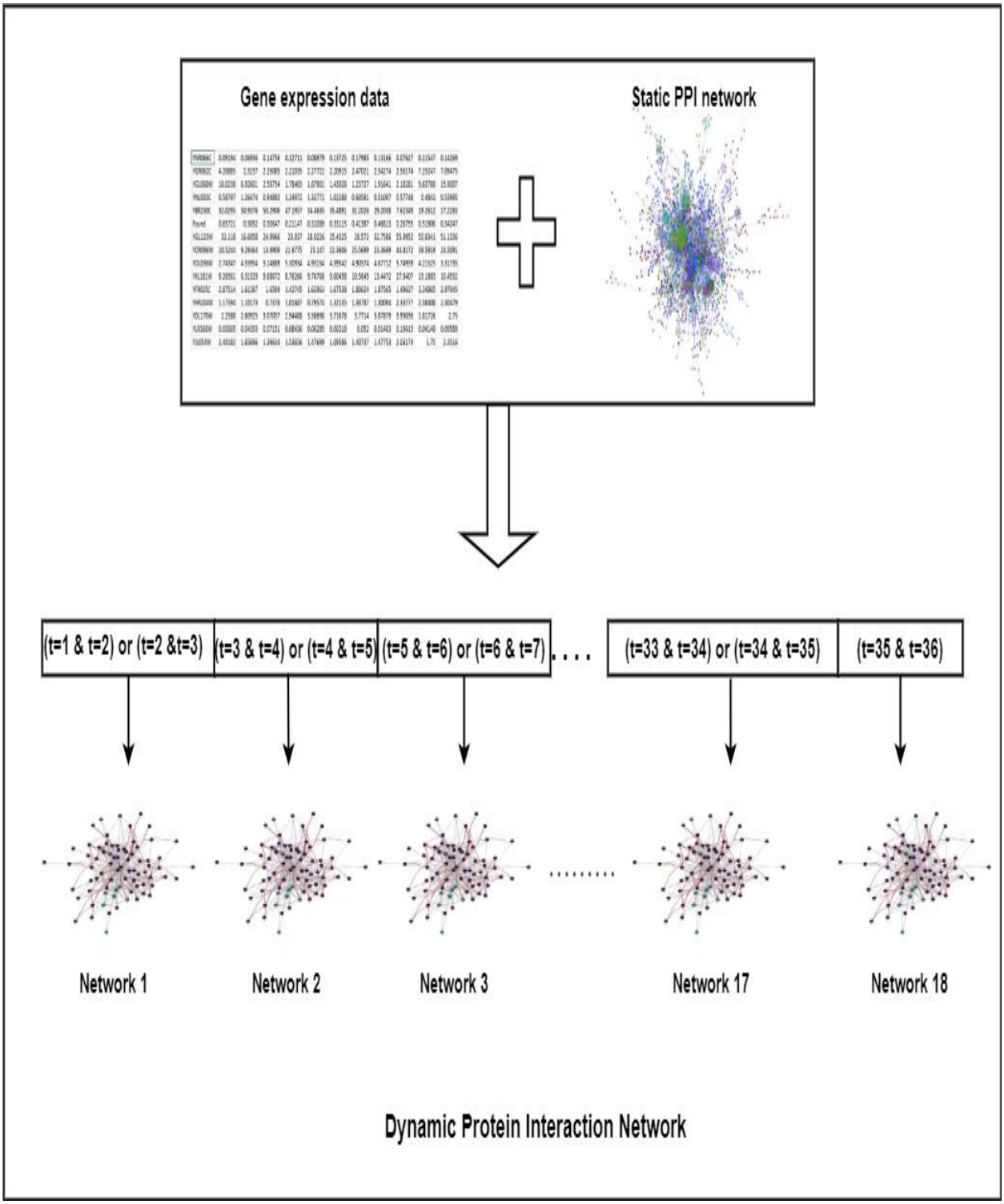


Figure 3-8: Quartile one method for construction of DPINs

A. Mechanism of construction DPINs

The PIN datasets available are static, so other information such as GE data should be utilized to develop dynamic networks. In this area, GEP provides useful dynamic information to model the activities of proteins at any given time and construct dynamic PIN. The protein is not always active, so it is important to determine the time point at which a protein demonstrates activity before constructing the dynamic PIN. Q-one proposed a new method to determine the active times of each protein. For example, Table 3-1 contains the expression values of protein 'YGR062C' in 36 times. First, the quartile one of these values is calculated, which is equal to 2.2407255. Then, the times are determined at which its expression values are equal to or greater than 2.2407255. They are [1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 31, 33, 34, 35 and 36]. After that, the active times of each protein are calculated. A protein is considered active in DPIN-1 if it appears in times one and two, or two and three. Whereas it is considered active in DPIN-2 if it appears in times three and four or four and five, and so on as shown in Table 3-2. As a result, protein appears in following DPINs [1, 4, 5, 6, 7, 8, 10, 11, 12, 13, 17, 18]. The edge in static PIN is kept in DPIN if both of its proteins are active in the same DPIN. Thus from one static PPI and 36 time points of each protein, q-one constructs 18 DPINs. That is, it reduces the number of DPINs to half the number of time points in the gene expression profile.

Table 3-1: gene expression values of 'YGR062C' protein

Protein name : YGR062C					
Time point	**value**	**Activation**	**Time point**	**Value**	**Activation**
1	4.208853	active	**19**	2.258847	active
2	2.323699	active	**20**	2.62311	active
3	2.250851	active	**21**	4.101714	active
4	2.210349		**22**	5.419156	active
5	2.277217	active	**23**	4.60515	active
6	2.20915		**24**	3.831182	active
7	2.470206	active	**25**	5.637363	active
8	2.342738	active	**26**	3.163313	active
9	2.561743	active	**27**	1.863122	
10	7.152467	active	**28**	1.504057	
11	7.094749	active	**29**	1.688016	
12	3.190863	active	**30**	2.015756	
13	4.528129	active	**31**	2.513721	active
14	2.398864	active	**32**	2.203955	
15	2.18623		**33**	5.227473	active
16	2.50685	active	**34**	4.139764	active
17	2.315496	active	**35**	3.558139	active
18	2.019388		**36**	3.717842	active

Note: In 36 time points, the protein "YGR062C" exhibits distinct values. The protein is considered active at that time for each value that is equal to or greater than the quartile one value; otherwise, it is considered inactive.

Table 3-2: DPINs that protein 'YGR062C' appears in

DPIN no.	Successive times	Activation
1	(1∧2) ∨ (2∧3)	*DPIN-1 (*)*
2	(3∧4) ∨ (4∧5)	
3	(5∧6) ∨ (6∧7)	
4	(7∧8) ∨ (8∧9)	*DPIN-4 (*)*
5	(9∧10) ∨ (10∧11)	*DPIN-5 (*)*
6	(11∧12) ∨ (12∧13)	*DPIN-6 (*)*
7	(13∧14) ∨ (14∧15)	*DPIN-7 (*)*
8	(15∧16) ∨ (16∧17)	*DPIN-8 (*)*
9	(17∧18) ∨ (18∧19)	
10	(19∧20) ∨ (20∧21)	*DPIN-10 (*)*
11	(21∧22) ∨ (22∧23)	*DPIN-11 (*)*
12	(23∧24) ∨ (24∧25)	*DPIN-12 (*)*
13	(25∧26) ∨ (26∧27)	*DPIN-13 (*)*
14	(27∧28) ∨ (28∧29)	
15	(29∧30) ∨ (30∧31)	
16	(31∧32) ∨ (32∧33)	
17	(33∧34) ∨ (34∧35)	*DPIN-17 (*)*
18	(35∧36)	*DPIN-18 (*)*

Note: The protein "YGR062C" appears in the first DPIN, if it is active between time points 1 and 2 or between time points 2 and 3. It appears in the second DPIN, if it is active between time points 3 and 4 or between time points 4 and 5, and so on.

3.4.2.2 Shared-one-time algorithm

After constructing a DPIN using the q-one method, a new algorithm is proposed to predict protein complexes in dynamic networks. It consists of three major steps: First, label each protein in the PIN with its active time points using the q-one method. Second, PIN is divided into several subgraphs that satisfy the CN condition and share at least one time point. These are denoted as preliminary protein complexes. Third, attach proteins to the preliminary complex, where the closeness between an attached protein and the preliminary complex is not less than a specific threshold. The attached protein shares at least one time point with the proteins in the preliminary complex as shown in Figure 3-9. The steps of the algorithm are as follows:

1. The active times of each protein are calculated employing the q-one method, and are used as a node label.

2. Add the neighborhood of the seed if they share a specific percentage of common neighbors 'T_{CN}'. They should also share at least one active time with the seed. The preliminary complex is accepted if its density is equal to or greater than the density threshold (DT), which is set to 0.5 (algorithm 4, step 1-10).

3. Find the neighborhood 'N_C' of the preliminary complex's proteins, if the closeness Score Equation (2-10) between a protein and a preliminary complex is larger than a threshold (T_{CS}) and shares at least one active time with the proteins of the preliminary complex, the protein is attached to the preliminary complex to produce a predicted complex (algorithm 4, step 11-17).

4. Redundant complexes have been removed by retaining only one of the exactly matched complexes.

Algorithm 4: shared-one-time

Inputs: Q, set of proteins label with active times using q-one method

Output: COMPLEXES, The sets of predicted protein complexes

1. **for** each protein *q* in Q **do**
2. Add *q* to COMP
3. **for** each neighbour of *q* **do**
4. Find CN between *q* and its neighbours
5. **if** CN ≥ T_{CN} **and** shared-time > 0 **then**
6. Add neighbour to COMP
7. **Endfor**
8. **if** density(COMP) ≥ DT **and** COMP not in COMPLESES **then**
9. Add COMP to COMPLEXES
10. **Endfor**
11. **for** each COMP in COMPLEXES **do**
12. Find neighbours N_C of COMP's proteins
13. **for** each protein n in N_C **do**
14. **if** CS(core, n) ≥ T_{CS} **and** shared-time > 0 **then**
15. Add n to COMP
16. **Endfor**
17. **Endfor**

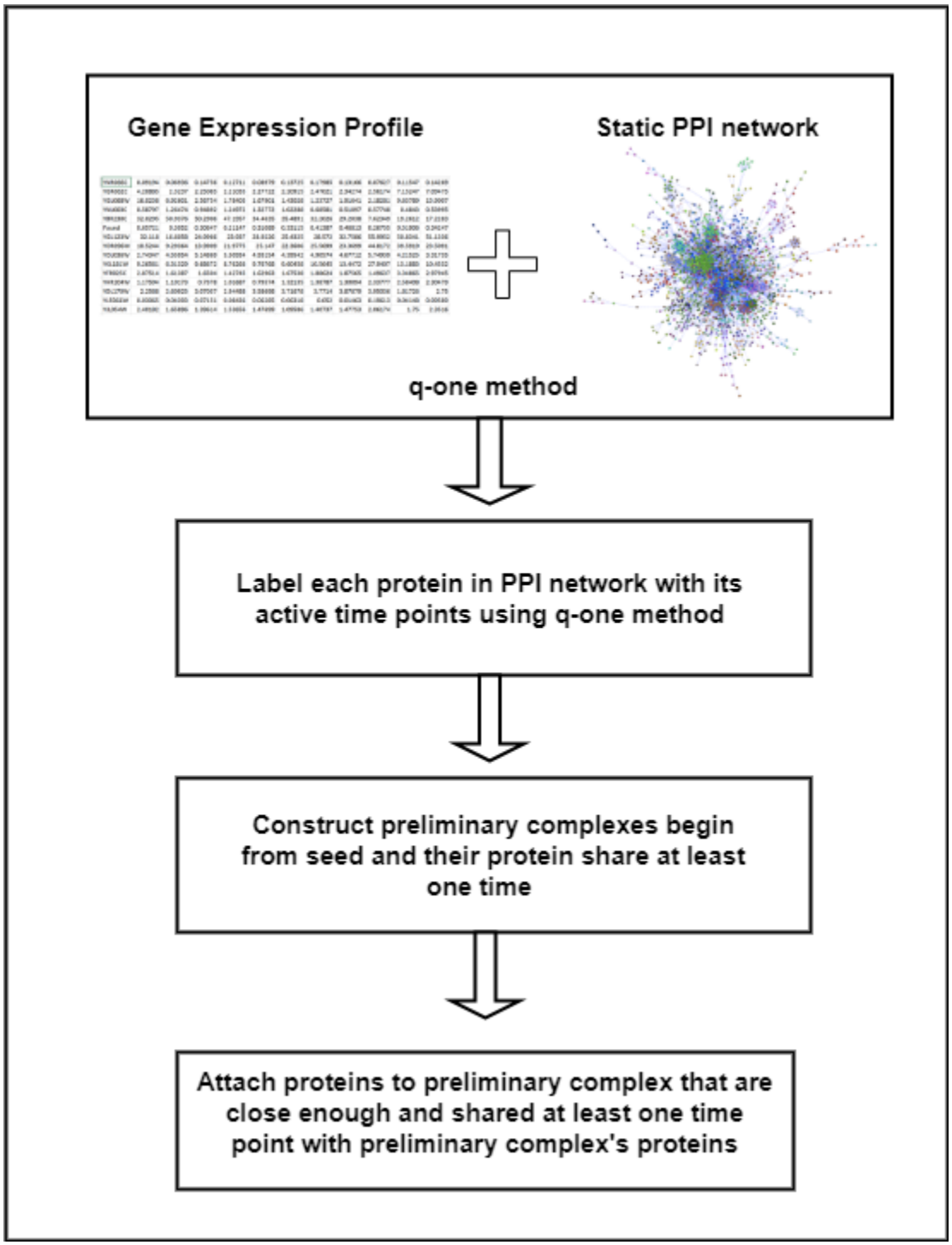


Figure 3-9: Shared-one-time technique

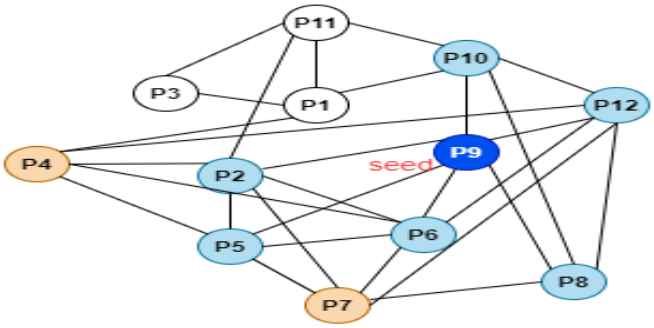
A. Shared-one-time Mechanism

The goal of this algorithm is constructing protein complexes whose proteins share at least one active time point. The PPI network is constructed as in previous algorithms. The q-one method is used to determine the active time points of each protein in PIN and set its label. To build the preliminary complex, the shared-one-time method starts with a seed (P9) and identifies its direct neighbors (P10, P12, P2, P5, P6 and P8). All the neighbors meet the CN criteria of having a value equal to or greater than 0.3. Yet P10 has no time points shared with the seed, which minimizes its chances of being in a preliminary complex, despite having a high percentage of common neighbors. The preliminary complex (P9, P12, P2, P5, P6 and P8) is approved because its density is equal to or greater than 0.5. After constructing the preliminary complex, the algorithm will check its neighbors (P4 and P7) and their closeness to the proteins in the preliminary complex. Even though both of them meet the closeness requirement, only P4 will be attached to the preliminary complex because it shares at least one time point with all its proteins. The complex (P9, P12, P2, P5, P6, P8 and P4) is then approved as a final predicted complex with at least one time point shared by its proteins as shown in Figure 3-10.

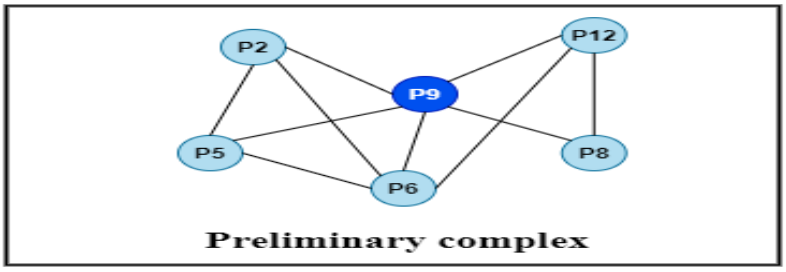
PPI network

Proteins labels

P1: {1,3,7,15,11,17,18
P2: {1,2,8,9,10,12,18}
P3: {2,3,4,8,9,11,15}
P4: {1,2,11,12,15,17,18}
P5: {1,2,7,8,11,15,16,}
P6: {1,2,11,14,17}
P7: {2,8,11,12,15,16,17}
P8: {1,2,3,10,11}
P9: {1,3,4,5,6,7,10,17,18}
P10: {2,11,12,15,16}
P11: {7,11,15,16,18}
P12: {1,2,10,11,12}



Seed	Neighbour	CN	Shared times
P9	P10	0.365	{}
	P12	0.5	{1, 10}
	P2	0.333	{1, 10, 18}
	P5	0.365	{1, 7}
	P6	0.5	{1, 17}
	P8	0.408	{1, 10, 3}



Attached protein	Closeness	Shared times
P4	0.666	P9 ∩ P12 ∩ P2 ∩ P5 ∩ P6 ∩ P8 ∩ P4 = {1}
P7	0.833	P9 ∩ P12 ∩ P2 ∩ P5 ∩ P6 ∩ P8 ∩ P7 = {}

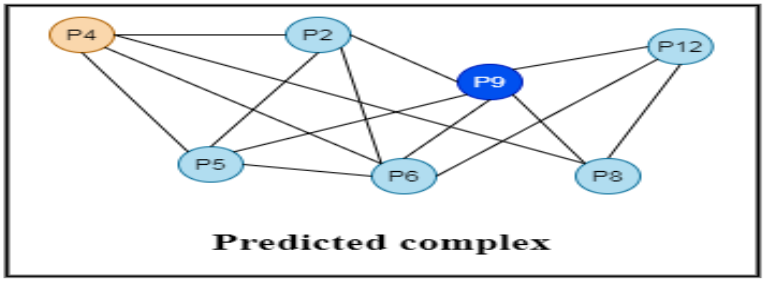


Figure 3-10: Trace with example of shared-one-time mechanism

Chapter Four **Experimental Results and Discussion**

4.1. Introduction

This chapter discussions the results of the proposed methodology and clarifies the comparison results with other algorithms on the datasets used in this study.

4.2. System Requirements

- Processor: 2.40GHz Intel(R) Core(TM) i7-4500U CPU 1.80 GHz. Memory (RAM): 8.00 GB.
- Operating system (OS): Windows 7 and above 64-bit.
- Language: Python 3.8.
- IDE environment: PyCharm 2019.3.1.

4.3. Results of SPIN

This section discusses the experimental result of two proposed algorithms for prediction protein complexes from static protein interaction network.

4.3.1. Results of SETS algorithm

The algorithm SETS is developed to predict overlapping protein complexes in the PPI network in an acceptable execution time, where the complex is predicted using a seed expanding model based on network structure.

A. PPI and reference Datasets

SETS algorithm has been analysed by concentrating on five PPI networks of *Saccharomyces cerevisiae* (yeast) and one network of *Homo sapiens* (human), which

is the combination of data from two databases: HPRD (Human Protein Reference Database) and BioGRID. The PPI datasets of the yeast are Collins and Gavin for ClusterONE algorithm, DIP, Krogan and BioGRID from SPICi algorithm.

Table 4-1 explains the properties of each dataset. Each dataset contains a different number of proteins having a different number of interactions that create variety in the density of network to satisfy the diversity that is required in the PPI networks used with the algorithm. NewMIPS and CYC2008 are used as reference complexes. All datasets are available online from their authors.

Table 4-1: The number of proteins, the number of intersections and the network density in PPI datasets.

Datasets	No. of Proteins	No. of Intersections	Network density
Collins	1622	9074	0.007
Gavin	1855	7669	0.004
Krogan	2675	7084	0.002
DIP	4930	17201	0.001
BioGRID	5361	85866	0.006
Human	15459	144687	0.001

B. Time Complexity

The execution time has been calculated for each dataset in order to analyse the time complexity of the SETS algorithm. SETS receives a set of ordered nodes Q in increasing order that take $O(n^2)$, which is considered a pre-processing step. Let n and m be the number of nodes and its neighbours, respectively. SETS processes each node in Q that has a label set to not-visited and adds its neighbours. This process

takes O(n*m) and reduces it to O(N) since not all nodes are processed. The second part of SETS is expanding for each candidate complex C; this takes O(C*m), where m is the neighbours of complex's proteins. The time complexity of SETS is O(n)+O(C*m).

C. Comparison of SETS with other algorithms

The performance of the algorithm has been compared to those of six others: MCODE, ClusterONE, NCMine, SPICi, IPCA, and PEWCC.

D. Evaluation Metrics

The quality of a predicted complex is evaluated using different metrics which are Recall, Precision and F-measure using α as OS, Coverage Rate (CR), and exact and good complexes that matching with real complexes.

E. Selection of Parameters

The reference datasets are analyzing using every PPI network. Table 4-2 reports the results of the analyzing which contains the number of proteins in each PPI as well as the number of proteins that are in reference complexes but are not in PPIs. The number of complexes in the reference dataset is reported, the reference complexes from proteins that are not in PPI are filtered out and only the complexes that have a length of more than two proteins are retained (first filter). The reference complexes are filtered again (second filter) and only those complexes that have all its proteins in PPI are retained. The CN is calculated between the proteins of the same complex for different thresholds as shown in Table 4-3 and Table 4-4. The number of complexes where at least two of its proteins are satisfied at the threshold is reported and according to the number of complexes that satisfied T_{CN} to the number of

complexes from the second filter, almost 25% of complexes from the second filter, the threshold T_{CN} is set to each PPI according to the result of reference complexes analysing. The F-measure of each dataset with a different threshold proved the accuracy of the selected threshold.

The other two parameters have been used in SETS: DT and T_{CS}. Liu (Liu et al., 2010, Dec. 18-21) analysed the protein complexes of CYC2008, MIPS and Aloy and found that almost 60% of the complexes have a density equal to or more than 0.5. Therefore, DT is set to 0.5 to define complexes that are dense enough to be the preliminary complexes. T_{CS} is assigned according to dataset and set at least 0.5 in order to let only proteins that have a good closeness to the preliminary complex that is to be added. Table 4-5 explains the threshold of each dataset.

Table 4-2: Reference datasets analyzing using PPI dataset

	# proteins In PPI	# proteins in reference but not in PPI	# reference complexes	First filter	Second filter
Collins					
CYC2008	1662	382	236	145	102
NewMIPS	1662	695	328	221	106
Gavin					
CYC2008	1855	439	236	143	86
NewMIPS	1855	724	328	218	90
Krogan					
CYC2008	2675	389	236	169	119
NewMIPS	2675	604	328	249	123
DIP					
CYC2008	4930	138	236	226	191
NewMIPS	4930	194	328	313	231
BioGRID					
CYC2008	5361	6	236	236	231
NewMIPS	5361	31	328	322	301
Human					
CORUM	15459	157	2351	2340	2196

Note: The first column shows the no. of proteins in PPI data. The no. of proteins that are in the reference data but not the PPI data is shown in the second column. The no. of PCs in the reference data is shown in the third column. The no. of PCs in the reference data that are filtered from proteins not in the PPI is shown in the fourth column. The fifth column displays the no of PCs in the reference data that all of its proteins in PPI data.

Table 4-3: T_{CN} values using different yeast PPI datasets

T_{CN}	CYC2008	NewMIPS	F-measure (CYC2008)	F-measure (NewMIPS)
Collins				
0.1	1	2	0.602	0.638
0.2	14	10	0.606	0.638
0.3	**24**	**24**	**0.606**	**0.642**
0.4	36	38	0.588	0.623
0.5	47	48	0.571	0.604
Gavin				
0.1	5	6	0.483	0.51
0.2	17	17	0.499	0.526
0.3	**25**	**29**	**0.53**	**0.563**
0.4	33	36	0.537	0.566
0.5	47	51	0.544	0.565
Krogan				
0.1	12	17	0.6	0.575
0.2	**40**	**43**	**0.589**	**0.583**
0.3	70	68	0.531	0.548
0.4	88	90	0.468	0.494
0.5	99	103	0.421	0.442
DIP				
0.1	**61**	**100**	**0.565**	**0.559**
0.2	133	173	0.516	0.518
0.3	158	201	0.464	0.507
0.4	168	215	0.35	0.407
0.5	170	218	0.194	0.407
BioGRID				
0.1	75	165	0.406	0.416
0.2	**159**	**242**	**0.477**	**0.474**
0.3	199	277	0.45	0.485
0.4	217	292	0.358	0.408
0.5	222	296	0.247	0.297

Table 4-4: TCN values with Human PPI datasets

T_{CN}	CORUM	F-measure (CORUM)
0.1	**1483**	**0.447**
0.2	1948	0.417
0.3	2071	0.325
0.4	2123	0.242
0.5	2145	0.137

Note: The no. of complexes from second filter where at least two of its proteins meet the T_{CN} criteria, almost the threshold that covers 25% of PCs is chosen. The F-measure is then performed with each threshold to support the threshold selection.

Table 4-5: Threshold values to each dataset.

Datasets	T_{CN}	DT	T_{CS}
Collins	0.3	0.5	0.6
Gavin	0.3	0.5	0.5
Krogan	0.2	0.5	0.5
DIP	0.1	0.5	0.6
BioGRID	0.2	0.5	0.7
Human data	0.1	0.5	0.7

F. Quality of Predicted Complexes

The performance of SETS is compared with that of six other algorithms using five datasets for yeast and one dataset for human. All datasets are unweighted except SPICi, which uses weighted networks. All parameters in the algorithms are set to default; in addition, complexes with less than three proteins are ignored. All the algorithms are implemented in the Cytoscape software except SPICi, which is implemented in its web site. The complex is considered matched if the OS with reference complex is greater than or equal to 0.2. SETS has the highest F-measure in all cases and competes with other algorithms in recall and precession as shown in Table 4-6, Table 4-7, Table 4-8, Table 4-9, Table 4-10 and Table 4-11. SETS obtains the highest CR in most cases, except in Collins and BioGRID, where it obtained the second-highest CR. Besides a few exceptions where its prediction ranks behind that of the PEWCC, the exact and good predicted complexes by SETS are the best in most cases as shown in Figure 4-1.

The ProCope software tool is used to evaluate the biological significance of predicted complexes and the data used in the evaluation process is set to ‘default’. The evaluation is based on biological process (BP) and cellular component (CC). SETS detects more complexes that are significant in BioGRID and human datasets as shown in Figure 4-2 and ranks second with regard to the rest of the datasets, competitive with SPICi, IPCA, and ClusterONE algorithms.

SETS predicts overlapping complexes as explained in Table 4-12 that report some of these complexes that have high OS scores with reference complexes and share some of its proteins. The complexes that predicted by SETS is with different densities and not restricted to dense complex as other algorithms that used the topological structure of PPI. SETS has achieved higher F-measure with different densities of PPI network in contrast with other algorithms that its F-measure

decrease when PPI network density decrease. Table 4-13 reports some of the low-density complexes that have high OS scores with reference complexes.

Table 4-6: Performance analysis for Gavin data with CYC2008 and NewMIPS.

	# complex	Recall	Precession	F-measure	CR
Gavin with CYC2008					
SPICi	91	0.36	**0.76**	0.491	0.504
ClusterONE	258	0.508	0.419	0.459	0.633
NCMine	621	0.513	0.393	0.445	0.64
PEWCC	656	0.517	0.402	0.453	0.596
IPCA	464	**0.53**	0.457	0.491	0.626
MCODE	101	0.021	0.05	0.03	0.118
SETS	246	0.475	0.602	**0.531**	**0.656**
Gavin with NewMIPS					
SPICi	91	0.372	**0.736**	0.494	0.248
ClusterONE	258	0.53	0.419	0.468	0.417
NCMine	621	0.549	0.39	0.456	0.422
PEWCC	656	0.552	0.433	0.485	0.392
IPCA	464	**0.573**	0.47	0.516	0.413
MCODE	101	0.021	0.059	0.031	0.045
SETS	246	0.524	0.607	**0.563**	**0.43**

Table 4-7: Performance analysis for Krogan data with CYC2008 and NewMIPS

	# complex	Recall	Precession	F-measure	CR
Krogan with CYC2008					
SPICi	131	0.458	0.641	0.534	0.583
ClusterONE	240	0.492	0.512	0.502	0.598
NCMine	578	0.458	0.433	0.445	0.593
PEWCC	708	**0.525**	0.496	0.51	0.593
IPCA	472	0.517	0.595	0.553	0.599
MCODE	60	0.03	0.117	0.047	0.111
SETS	220	0.479	**0.764**	**0.589**	**0.68**
Krogan with NewMIPS					
SPICi	131	0.479	0.618	0.54	0.352
ClusterONE	240	0.442	0.458	0.45	0.323
NCMine	578	0.479	0.427	0.452	0.362
PEWCC	708	**0.534**	0.476	0.503	0.368
IPCA	472	0.515	0.574	0.543	0.36
MCODE	60	0.021	0.1	0.035	0.038
SETS	220	0.485	**0.732**	**0.583**	**0.391**

Table 4-8: Performance analysis for Collins data with CYC2008 and NewMIPS.

	# complex	Recall	Precession	F-measure	CR
Krogan with CYC2008					
SPICi	106	0.419	**0.736**	0.534	0.69
ClusterONE	203	**0.559**	0.547	0.553	**0.797**
NCMine	377	0.517	0.475	0.495	0.763
PEWCC	426	0.53	0.521	0.525	0.738
IPCA	342	0.542	0.64	0.587	0.751
MCODE	103	0.051	0.107	0.069	0.121
SETS	218	0.521	0.725	**0.606**	0.767
Krogan with NewMIPS					
SPICi	106	0.473	0.726	0.573	0.443
ClusterONE	203	**0.588**	0.542	0.564	0.519
NCMine	377	0.537	0.501	0.518	**0.493**
PEWCC	426	0.546	0.533	0.539	0.479
IPCA	342	0.567	0.705	0.628	0.486
MCODE	103	0.03	0.087	0.045	0.055
SETS	218	0.555	**0.761**	**0.642**	0.488

Table 4-9: Performance analysis for DIP data with CYC2008 and NewMIPS.

	# complex	Recall	Precession	F-measure	CR
Krogan with CYC2008					
SPICi	219	0.555	**0.507**	0.53	0.541
ClusterONE	342	0.436	0.336	0.38	0.466
NCMine	1074	0.542	0.291	0.378	0.497
PEWCC	1544	**0.678**	0.317	0.432	0.582
IPCA	826	0.589	0.318	0.413	0.516
MCODE	50	0.008	0.04	0.014	0.116
SETS	540	0.653	0.498	**0.565**	**0.593**
Krogan with NewMIPS					
SPICi	219	0.573	0.479	0.522	0.334
ClusterONE	342	0.412	0.304	0.35	0.265
NCMine	1047	0.546	0.287	0.376	0.32
PEWCC	1544	**0.683**	0.318	0.434	**0.39**
IPCA	826	0.579	0.311	0.405	0.323
MCODE	50	0.006	0.04	0.011	0.046
SETS	540	0.64	**0.496**	**0.559**	**0.39**

Table 4-10: Performance analysis for BioGRID data with CYC2008 and NewMIPS.

	# complex	Recall	Precession	F-measure	CR
Krogan with CYC2008					
SPICi	440	0.432	0.186	0.26	0.613
ClusterONE	476	0.487	0.265	0.343	0.697
NCMine	3671	0.737	0.123	0.211	0.807
PEWCC	4048	**0.873**	0.196	0.32	**0.872**
IPCA	2718	0.576	0.14	0.226	0.758
MCODE	56	0.008	0.036	0.014	0.073
SETS	633	0.644	**0.379**	**0.477**	0.816
Krogan with NewMIPS					
SPICi	440	0.436	0.18	0.254	0.442
ClusterONE	476	0.488	0.25	0.331	0.496
NCMine	3671	0.695	0.13	0.219	0.551
PEWCC	4048	**0.826**	0.21	0.335	**0.627**
IPCA	2718	0.591	0.138	0.223	0.538
MCODE	56	0.006	0.036	0.01	0.029
SETS	633	0.622	**0.382**	**0.474**	0.561

Table 4-11: Performance analysis for Human dataset

	# complex	Recall	Precession	F-measure	CR
SPICi	NA	NA	NA	NA	NA
ClusterONE	1037	0.223	0.235	0.229	0.33
NCMine	7776	0.552	0.221	0.315	0.459
PEWCC	9036	0.68	0.276	0.393	**0.559**
IPCA	6533	0.463	0.266	0.338	0.455
MCODE	74	0.001	0.014	0.002	0.04
SETS	2026	**0.498**	**0.405**	**0.447**	0.484

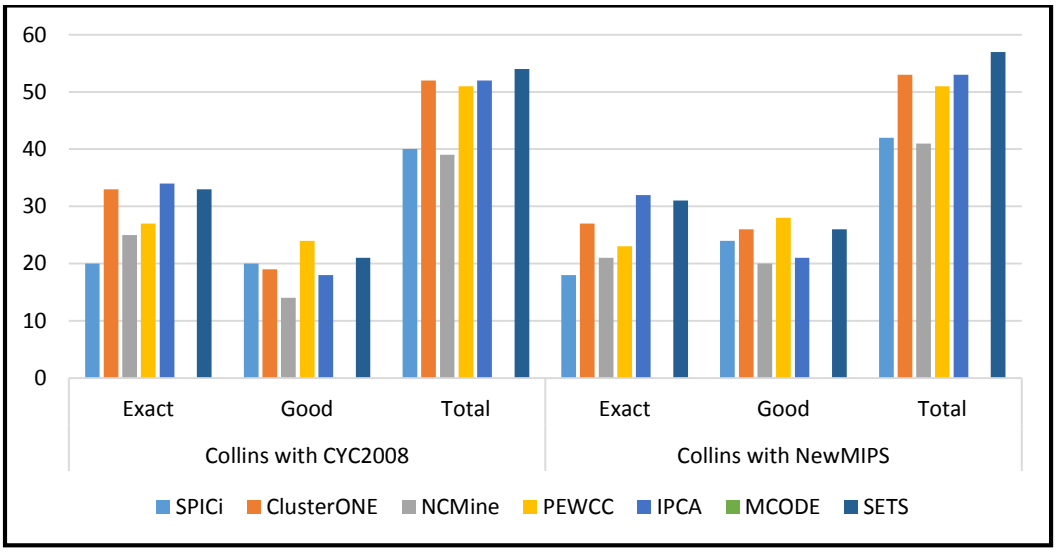


Figure 4-1: The number of exact and good predicted complexes in Collins dataset

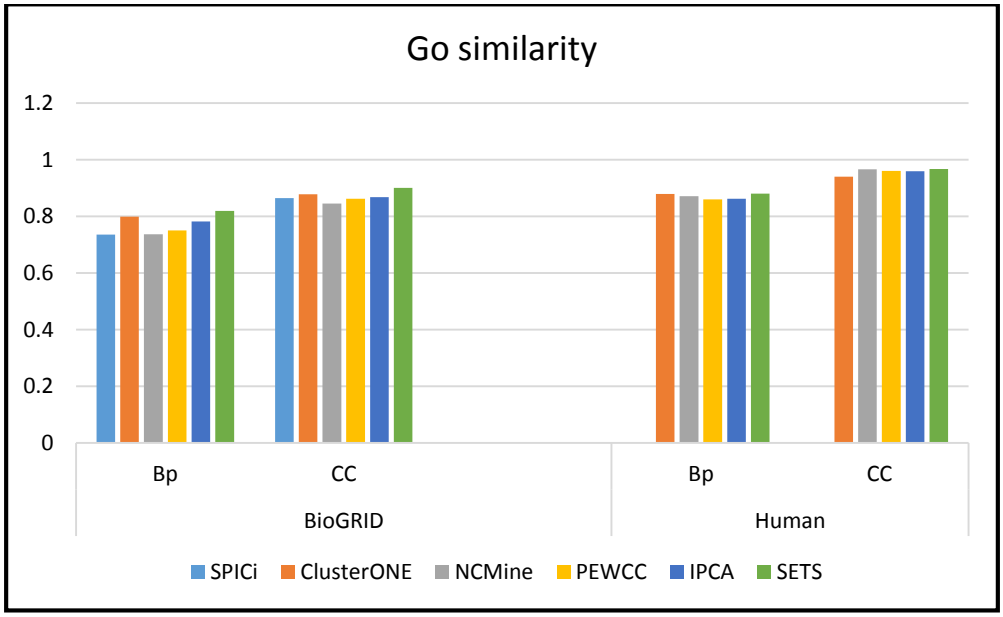


Figure 4-2: The biological significance of predicted complexes in BioGRID and Human

Table 4-12: Overlapping predicted complexes with high OS score from Collins using NewMIPS

Predicted complex_1	Real complex	OS	Predicted complex_2	Real complex	OS	Overlapping proteins
YFL039C	**YFL039C**	1	**YFL039C**	**YFL039C**	0.92	**YNL107W**
YJL081C	**YJL081C**		YDR334W	YDR334W		**YFL039C**
YOR244W	YOR244W		**YJL081C**	**YJL081C**		**YJL081C**
YHR090C	YHR090C		YML041C	YLR385C		**YGR002C**
YNL107W	**YNL107W**		**YNL107W**	YML041C		
YNL136W	YNL136W		YBR231C	**YNL107W**		
YEL018W	YEL018W		YLR085C	YBR231C		
YHR099W	YHR099W		YDR485C	YDR485C		
YPR023C	YPR023C		YDR190C	YLR085C		
YFL024C	YFL024C		YAL011W	YDR190C		
YJR082C	YJR082C		YPL235W	YAL011W		
YDR359C	YDR359C		**YGR002C**	YPL235W		
YGR002C	**YGR002C**			**YGR002C**		
YKL144C	YHR143W-A	0.94	YHR143W-A	YHR143W-A	0.93	**YOR210W**
YOR210W	YKL144C		YOR341W	YOR341W		**YPR110C**
YOR116C	**YOR210W**		**YOR210W**	**YOR210W**		**YNL113W**
YPR190C	YOR116C		**YPR110C**	**YPR110C**		**YBR154C**
YPR110C	YPR190C		**YNL113W**	**YNL113W**		**YPR187W**
YNL113W	**YPR110C**		YOR340C	YOR340C		**YOR224C**
YDL150W	**YNL113W**		YJR063W	YJR063W		
YBR154C	YDL150W		YOR151C	**YBR154C**		
YPR187W	**YBR154C**		**YBR154C**	YNL248C		
YKR025W	**YPR187W**		YNL248C	YDR156W		
YDR045C	YKR025W		YDR156W	**YPR187W**		
YNR003C	YDR045C		**YPR187W**	YPR010C		
YNL151C	YNR003C		YPR010C	**YOR224C**		
YOR207C	YNL151C		**YOR224C**	YJL148W		
YJL011C	YOR207C		YJL148W			
YOR224C	YJL011C					
	YOR224C					

Table 4-13: DIP with Newmips reports low density (D.) complexes with a high OS.

Real complex	Predicted complex	Inter.	D.	(OS)
YKR068C YBR254C YDR472W YDR108W YMR218C YDR246W YML077W YOR115C YGR166W YDR407C	YLR342W YKR068C YBR254C YDR472W YDR108W YMR218C YDR246W YML077W YOR115C YGR166W YDR407C	10	0.36	0.91
Length = 10	Length = 11			
YDL005C YNL236W YPR070W YOL135C YNR010W YDR308C YBR193C YBR253W YNL025C YPR168W YMR112C YGL025C YCR081W YGL151W YOL051W YOR174W YLR071C YGR104C YGL127C YHR041C YHR058C YPL042C YER022W YBL093C YDR443C	YDL005C YNL236W YPR070W YOL135C YNR010W YDR308C YBR193C YBR253W YMR112C YGL025C YCR081W YGL151W YOL051W YOR174W YLR071C YGR104C YHR041C YHR058C YER022W YBL093C YDR443C	21	0.44	0.84
Length = 25	Length = 21			
Q0080 YDR322C-A YPL271W YPR020W YBL099W YML081C-A YDR377W YOL077W-A YDL004W YKL016C YDR298C YGR008C YBR039W YLR295C Q0085 Q0130 YDL130W-A YDL181W YPL078C YJR121W	Q0080 Q0130 YDR322C-A YBL099W YNL315C YPL271W YBR039W YPL078C YLR295C YML081C-A YPR020W YDR377W YDR298C YDL004W YKL016C Q0085 YDL181W YJR121W	17	0.29	0.8
Length = 20	Length = 17			
YIL084C YMR075W YOL004W YMR128W YPL181W YDL076C YPR023C YMR263W YNL330C YPL139C YNL097C	YIL084C YMR075W YOL004W YPL181W YIL035C YDL076C YPR023C YNL330C YMR263W YLR103C YPL139C YNL097C YBR095C	10	0.36	0.7
Length = 11	Length = 13			
YKR068C YBR254C YDR472W YDR108W YDR246W YML077W YOR115C	YLR342W YKR068C YBR254C YDR472W YDR108W YMR218C YDR246W YML077W YOR115C YGR166W YDR407C	7	0.36	0.64
Length = 7	Length = 11			

Note. Inter. is the interaction between predicted and real complexes, D. is the density and OS is the overlapping score between predicted and reference complex.

4.3.2. Results of GECA algorithm

A new method of predicting protein complexes based on the Gene Expression profile and Core-Attachment approach (GECA) is proposed. The GECA identifies core proteins using common neighbour techniques and gene expression profile, and improves the attachment technique by adding proteins that have fewer connections but are more similar in gene expression with core proteins. GECA does not remove any edges from the PPI network as other algorithms do.

A. PPI and reference Datasets

The datasets used in this algorithm comprise the same data used in the WEC and CAG algorithms. There are two datasets from yeast saccharomyces cerevisiae (Collins and Gavin with GEP from Tu et al. (Tu et al., 2005)) and the reference datasets taken from MIPS and SGD. The GEP data does not cover 792 interactions in Collin and 152 interactions in Gavin. The PPI network of human is integrated from HPRD (the Human Protein Reference Database Release 9) and HSN (Human Signalling Network). The GEP of Homo sapiens(Nymark et al., 2007) is downloaded from http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2604. However, the GEP does not cover 654 of PPI network interactions. The reference complexes of Homo sapiens have been obtained from CORUM. Details of the datasets are presented in Table 4-14. The connections that are not covered by GEP are not removed.

Table 4-14: Datasets details

Dataset	
Collins	9074 interactions
Gavin	7669 interactions
Yeast GEP	9335 gene and 36-time points
Homo sapiens PPI network	37437 interactions
Homo sapiens GEP	54675 genes and 27-time points
MIPS	203 complexes
SGD	323 complexes
CORUM	1521 complexes

B. Analysis of reference complexes

GEP is applied to reference complexes (SGD, MIPS and CORUM) to analyse the correlation between proteins in the same complex and determine the number of complexes that have more positive or negative correlations. First, the correlation between each pair in the complex is calculated to identify the complex that contain more positive or negative correlations. Next, the reference complexes are filtered from proteins that are not in PPI or do not have GE information, keeping the complexes that contain more than three proteins. Finally, these complexes are split into a core and an attachment. The core contains proteins that interact with a density equal to or greater than 0.7. The rest of the proteins are considered to be an attachment. The core and attachment are analysed to determine the positive and negative correlations between their proteins according to their connection in the PPI network. The correlation between core and attachment proteins is then calculated. Consequently, the number of cores and attachments that contain positive correlations

is greater than the number of cores and attachments with negative correlations between their proteins. In the yeast data, most of the positive correlations are greater than 0.1. Meanwhile, most of the negative correlations are greater than -0.5. In the human data, most of the positive and negative correlations are low. Table 4-15, Table 4-16, Table 4-17, Table 4-18, and Table 4-19 describe the details concerning this analysis. Moreover, the data analysis made in this study leads to the conclusion that GECA depends on obtaining all the positive correlations as well as having a low negative correlation to add proteins to the core in the yeast data. In the human data, GECA uses only the low positive and low negative correlations to add proteins to the core.

Table 4-15: Analyzing of MIPS using Gavin dataset

# genes in PPI	# shared genes in PPI and MIPS	# complexes in MIPS	#complexes with a size greater than 2
1855	766	203	139
# complexes	**# positive complexes**	**# Negative complexes**	**# Equal complexes**
203	166	24	13

# complexes	# positive complexes			# Negative complexes			# Equal complexes
139	115			6			13
	>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
	58	70	4	76	0	1	

Interactions of the core proteins						
# positive			**# Negative**			**# Equal**
109			17			5
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
52	64	3	62	0	1	

Interactions of the attachment proteins						
# positive			**# Negative**			**# Equal**
38			3			1
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
15	20	6	25	0	0	

Interactions between the core and the attachment		
# positive	**# Negative**	**# Equal**
27	3	5

Note: First, the correlation between complex's proteins is calculated without filtering. Next, the reference complexes are filtered from proteins that are not in PPI or do not have GE information. Finally, these complexes are split into a core and an attachment. In each step, the correlation between all of the complex's proteins is determined, as well as the correlation between each interval's proteins, and the number of PCs whose positive correlation exceeds their negative correlation or vice versa is reported.

Table 4-16: Analyzing of SGD using Gavin dataset

# genes in PPI	# shared genes in PPI and SGD	# complexes in SGD	#complexes with a size greater than 2
1855	775	323	139
# complexes	**# positive complexes**	**# Negative complexes**	**# Equal complexes**
323	233	40	22

# complexes	# positive complexes			# Negative complexes			# Equal complexes
139	113			12			9
	>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
	52	71	6	71	0	3	

Interactions of the core proteins						
# positive			**# Negative**			**# Equal**
107			16			8
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
50	68	5	62	0	3	

Interactions of the attachment proteins						
# positive			**# Negative**			**# Equal**
20			1			2
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
7	14	1	62	0	1	

Interactions between the core and the attachment		
# positive	**# Negative**	**# Equal**
24	3	0

Note: First, the correlation between complex's proteins is calculated without filtering. Next, the reference complexes are filtered from proteins that are not in PPI or do not have GE information. Finally, these complexes are split into a core and an attachment. In each step, the correlation between all of the complex's proteins is determined, as well as the correlation between each interval's proteins, and the number of PCs whose positive correlation exceeds their negative correlation or vice versa is reported.

Table 4-17: Analyzing of MIPS using Collins dataset

# genes in PPI	# shared genes in PPI and MIPS	# complexes in MIPS	#complexes with a size greater than 2
1622	773	203	141
# complexes	**# positive complexes**	**# Negative complexes**	**# Equal complexes**
203	166	24	13

# complexes	# positive complexes			# Negative complexes			# Equal complexes
141	124			6			7
	>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
	53	79	4	79	0	2	

Interactions of the core proteins						
# positive			**# Negative**			**# Equal**
118			8			7
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
54	78	0	65	0	2	

Interactions of the attachment proteins						
# positive			**# Negative**			**# Equal**
39			4			1
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
16	21	3	22	2	0	

Interactions between the core and the attachment		
# positive	**# Negative**	**# Equal**
31	5	3

Note: First, the correlation between complex's proteins is calculated without filtering. Next, the reference complexes are filtered from proteins that are not in PPI or do not have GE information. Finally, these complexes are split into a core and an attachment. In each step, the correlation between all of the complex's proteins is determined, as well as the correlation between each interval's proteins, and the number of PCs whose positive correlation exceeds their negative correlation or vice versa is reported.

Table 4-18: Analyzing of SGD using Collins dataset

# genes in PPI	# shared genes in PPI and SGD	# complexes in SGD	#complexes with a size greater than 2
1622	830	323	149
# complexes	**# positive complexes**	**# Negative complexes**	**# Equal complexes**
323	233	40	22

# complexes	# positive complexes			# Negative complexes			# Equal complexes
149	124			12			6
	>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
	51	79	8	73	0	3	

Interactions of the core proteins						
# positive			**# Negative**			**# Equal**
118			12			9
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
47	81	7	67	0	4	

Interactions of the attachment proteins						
# positive			**# Negative**			**# Equal**
19			4			0
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
5	13	0	9	0	0	

Interactions between the core and the attachment		
# Positive	**# Negative**	**# Equal**
30	2	2

Note: First, the correlation between complex's proteins is calculated without filtering. Next, the reference complexes are filtered from proteins that are not in PPI or do not have GE information. Finally, these complexes are split into a core and an attachment. In each step, the correlation between all of the complex's proteins is determined, as well as the correlation between each interval's proteins, and the number of PCs whose positive correlation exceeds their negative correlation or vice versa is reported.

Table 4-19: Analyzing of CORUM using Human dataset

# genes in PPI	# shared genes in PPI and CORUM	# complexes in CORUM	#complexes with a size greater than 2
5664	1991	1521	1198
# complexes	**# positive complexes**	**# Negative complexes**	**# Equal complexes**
1521	692	553	229

# complexes	# positive complexes			# Negative complexes			# Equal complexes
1198	514			304			168
	>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
	139	579	75	519	52	33	

Interactions of the core proteins						
# positive			**# Negative**			**# Equal**
522			337			71
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
144	509	28	420	46	15	

Interactions of the attachment proteins						
# positive			**# Negative**			**# Equal**
83			45			23
>= 0.5	**0.5 < >= 0.1**	**Equal**	**-0.1 <= > - 0.5**	**< - 0.5**	**Equal**	
24	79	12	79	7	3	

Interactions between the core and the attachment		
# Positive	**# Negative**	**# Equal**
174	119	12

Note: First, the correlation between complex's proteins is calculated without filtering. Next, the reference complexes are filtered from proteins that are not in PPI or do not have GE information. Finally, these complexes are split into a core and an attachment. In each step, the correlation between all of the complex's proteins is determined, as well as the correlation between each interval's proteins, and the number of PCs whose positive correlation exceeds their negative correlation or vice versa is reported.

C. Threshold values

The experiment is conducted with different parameter values. As shown in core_construction algorithm, there are two thresholds where the T_{core} determined the number of common neighbours between the seed and its neighbours, which uses 0.2 based on the SETS algorithm's analysis of reference datasets. T_{pcc} is kept at 0.1 for a positive correlation and equal to or greater than -0.5 for a negative correlation in the yeast data. Meanwhile, in the human data, the T_{pcc} is equal or less than 0.5 for a positive correlation and equal to or greater than -0.5 for a negative correlation. Whereas adding an attachment algorithm has two thresholds, $T_{closeness}$ is kept at 0.7 in the Gavin and Collins data and 0.9 in the human data in order to add only the protein that is closely connected with the core proteins. $N_{connection}$ is kept at 2 in the Collins and Gavin data and 3 in the human data to allow the addition of the protein with low connections to the core but high correlation to the gene expression pattern.

D. Comparison of GECA with other algorithms

The GECA algorithm is compared to four methods. ClusterONE, WEC which weighted the edge between two proteins by the Edge Clustering Coefficient (ECC) and the correlation between the proteins' gene expression, that leads to predict the protein complex accurately and CAG that use the core-attachment technique and identify the core as a functional unit by finding only the positive correlation between proteins in the neighbourhood and iteratively remove nodes with the minimum degree to maintain the density of the core. Then, attach proteins to the core if they are connected with half of the core proteins. WEC and CAG, used the same datasets, but remove unweighted interactions between proteins. Finally, WCOACH use Gene Ontology to weight the PIN and also remove the unweighted edges. Every parameter in all the algorithms is set to default or as the papers on them recommended. In

addition, complexes with less than three proteins are ignored. The next section describes the effects of removing unweighted edges on the result, especially in the Collins data from which the highest ratio of edges are removed.

E. Evaluation metrics

Recall, precision and F-measure

GECA is compared with the other methods using different metrics which are recall, precision and F-measure (using α as OS and J-Sim), co-localisation Score and Gene Ontology Semantic Similarity Score.

F-measure results of yeast and human data

Gavin data: GECA predicts 538 complexes, 238 of which matched with reference MIPS complexes and 271 match with SGD complexes. The maximum predicted complex contains 43 proteins. When using MIPS as reference data, GECA achieves recall, precision and F-measure values of 0.453, 0.442 and 0.448 respectively. GECA achieves the highest F-measure. The other methods (CAG, WEC, WCOACH and ClusterONE) achieve the F-measure values of 0.442, 0.429, 0.408 and 0.314 respectively. When SGD is used as reference data, the F-measure value of GECA is 0.438, followed by CAG with 0.41, WEC with 0.409, WCOACH with 0.38 and ClusterONE having 0.35. GECA achieves the highest F-measure with respect to the Gavin data when using SGD and MIPS as reference complexes.

Collins data: GECA predicts 574 complexes, 302 of which matched with MIPS reference complexes and 292 match with SGD reference complexes. The maximum predicted complex contains 72 proteins. With the Collins data, GECA achieves recall, precision and F-measure values of 0.498, 0.526 and 0.511 respectively when

compared with the MIPS data. The other methods (ClusterONE, WCOACH, WEC and CAG) achieve F-measure values of 0.414, 0.375, 0.499 and 0.463 respectively. The F-measure values of GECA, ClusterONE, WCOACH, WEC and CAG are 0.438, 0.458, 0.323, 0.437 and 0.431 respectively when using SGD as reference data. GECA has the highest F-measure with the Collins data using MIPS as reference data.

Human data: GECA predicts 857 complexes, with 288 matched with the CORUM data and the maximum predicted complex contained 103 proteins. GECA achieves the recall, precision and F-measure values of 0.227, 0.336 and 0.271 respectively. ClusterONE, WCOACH, CAG and WEC achieve F-measure values of 0.149, 0.09, 0.158 and 0.178 respectively.

It is evident that GECA has almost the highest F-measure with all data using the OS score to calculate the recall and precision values, as illustrated in Figure 4-3. When J-Sim is used to calculate the precision and recall values, GECA is able to get the highest recall in all the data. It also got the highest F-measure except with Collins data, where it came second to ClusterONE because the precision of the latter is higher than that of GECA. It predicts only half the number of complexes predicted by GECA. GECA predicts the highest number of small (S) and large complexes (L) in all the data. All the results come from using J-Sim are shown in Table 4-20, Table 4-21, and Table 4-22. Thus, it can be concluded that GECA predicts protein complexes better than its counterparts.

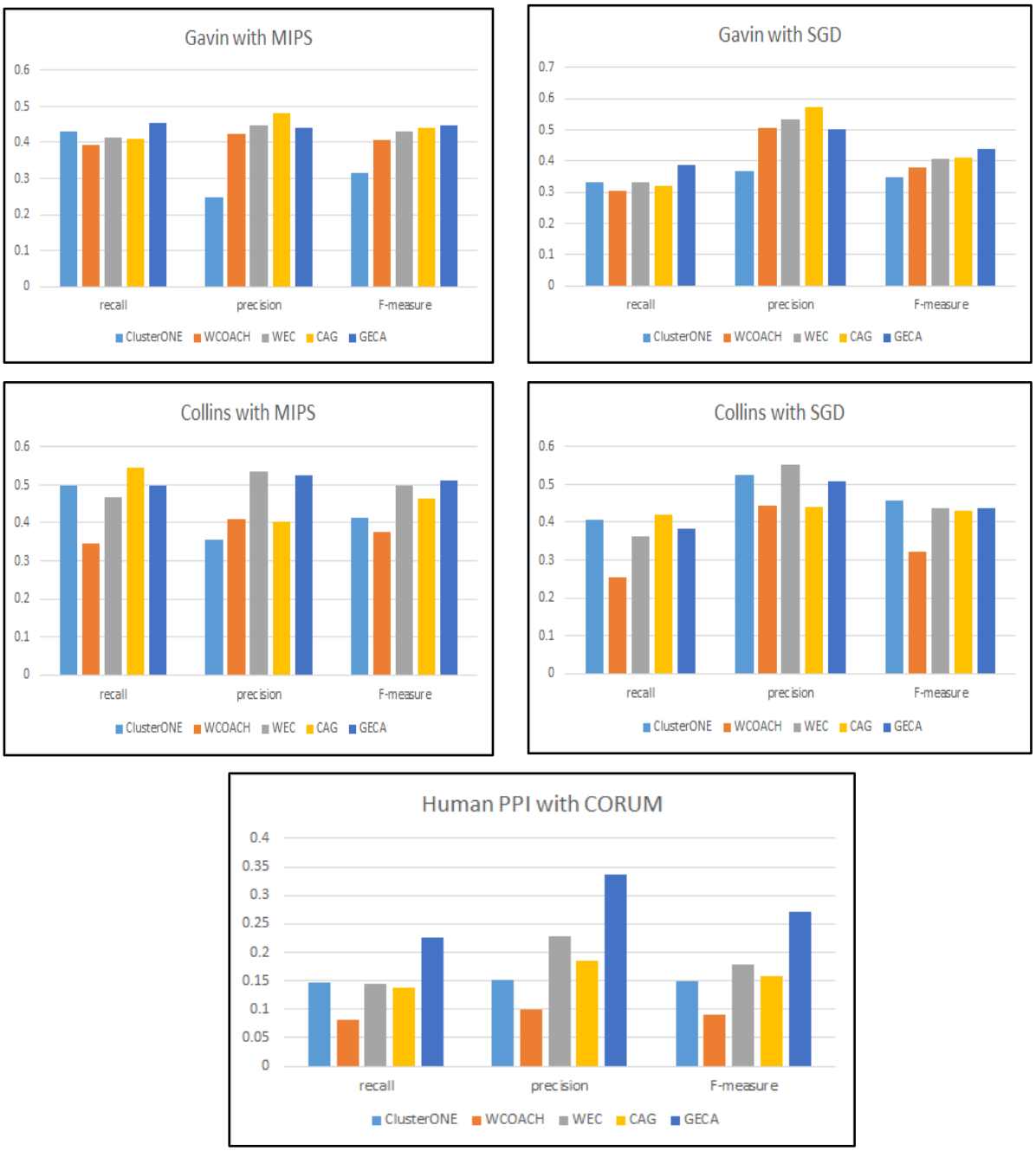


Figure 4-3: Recall, Precision, and F-measure using OS for all PPI networks

Table 4-20: Analysis results for Collins dataset using J-Sim

	Predicted complexes				Reference complexes				
	# complex	**# S**	**# L**	**P**	**# complex**	**# S**	**# L**	**R**	**F**
MIPS									
ClusterONE	30	3	27	**0.148**	32	3	29	0.158	**0.153**
WCOACH	21	0	21	0.057	19	0	19	0.094	0.071
WEC	40	4	36	0.041	34	4	30	0.167	0.066
CAG	39	4	35	0.039	35	4	31	0.172	0.064
GECA	**70**	**5**	**65**	0.122	**40**	**5**	**35**	**0.197**	0.151
SGD									
ClusterONE	49	7	42	**0.241**	53	7	46	0.164	**0.195**
WCOACH	25	1	24	0.068	25	1	24	0.077	0.073
WEC	61	8	53	0.063	54	8	46	0.167	0.092
CAG	76	8	68	0.077	54	8	46	0.167	0.105
GECA	**95**	**9**	**86**	0.166	**60**	**9**	**51**	**0.186**	0.175

Table 4-21: Analysis results for Gavin dataset using J-Sim

	Predicted complexes				Reference complexes				
	# complex	**# S**	**# L**	**P**	**# complex**	**# S**	**# L**	**R**	**F**
MIPS									
ClusterONE	17	**2**	15	0.066	18	**2**	16	0.089	0.076
WCOACH	31	**2**	29	0.061	22	**2**	20	0.108	0.078
WEC	35	**2**	33	0.039	27	**2**	25	0.133	0.06
CAG	33	**2**	31	0.038	24	**2**	22	0.118	0.058
GECA	**48**	**2**	**46**	**0.089**	**30**	**2**	**28**	**0.148**	**0.111**
SGD									
ClusterONE	31	2	29	0.12	30	2	28	0.093	0.105
WCOACH	38	**5**	33	0.075	31	**5**	26	0.096	0.084
WEC	48	3	45	0.053	32	3	29	0.099	0.069
CAG	39	3	36	0.045	28	3	25	0.087	0.06
GECA	**66**	3	**63**	**0.123**	**39**	3	**36**	**0.121**	**0.122**

Table 4-22: Analysis results for Human dataset using J-Sim

	Predicted complexes				Reference complexes				
	# complex	**# S**	**# L**	**P**	**# complex**	**# S**	**# L**	**R**	**F**
ClusterONE	43	1	42	0.039	50	1	49	0.033	0.036
WCOACH	42	0	42	0.029	47	0	47	0.031	0.03
WEC	70	2	68	0.045	67	2	65	0.044	0.044
CAG	49	2	47	0.039	57	2	55	0.037	0.038
GECA	**106**	**5**	**101**	**0.124**	**94**	**5**	**89**	**0.062**	**0.082**

Co-localisation and GO semantic similarity scores

Figure 4-4 reports the co-localisation score for all algorithms using data from Collins and Gavin. As shown, the co-localisation score of GECA is higher than those in other methods, having a value of 0.6925 using Gavin's data and 0.766 using Collins' data. An analysis of GO semantic similarity is done on the basis of the biological process (Bp), Cellular Component (CC) and Molecular Function (MF). Figure 4-5 shows the GO scores for all methods using Gavin, Collins, and Human PPI networks. Here, GECA has achieved the highest scores for all terms using Collins and Human data, except for the Bp and CC terms which use Gavin data. The pleasing results that GECA achieved in co-localisation and GO scores support the biological significance of the predicted complexes.

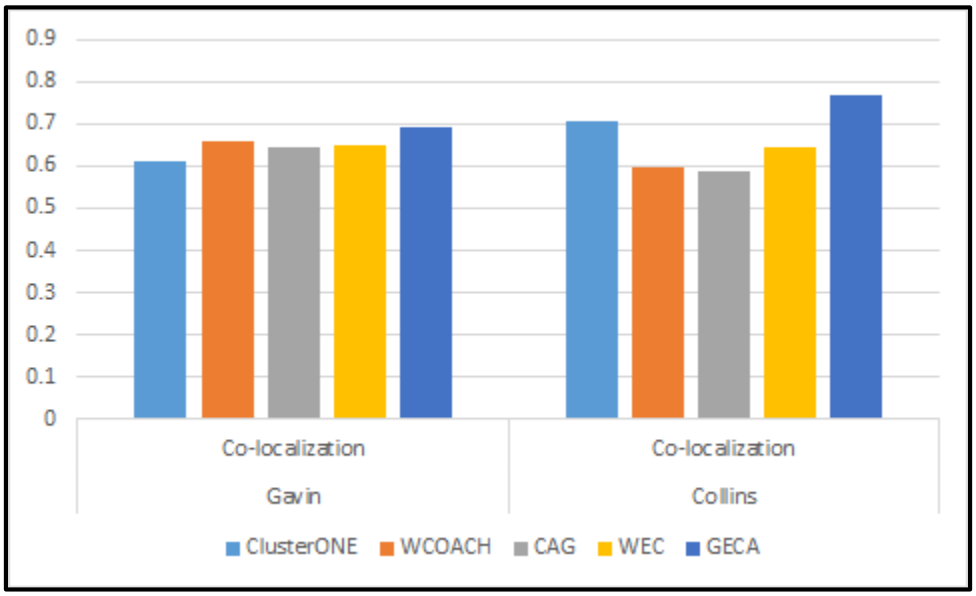


Figure 4-4: Co-localisation score of Gavin and Collins PPI networks

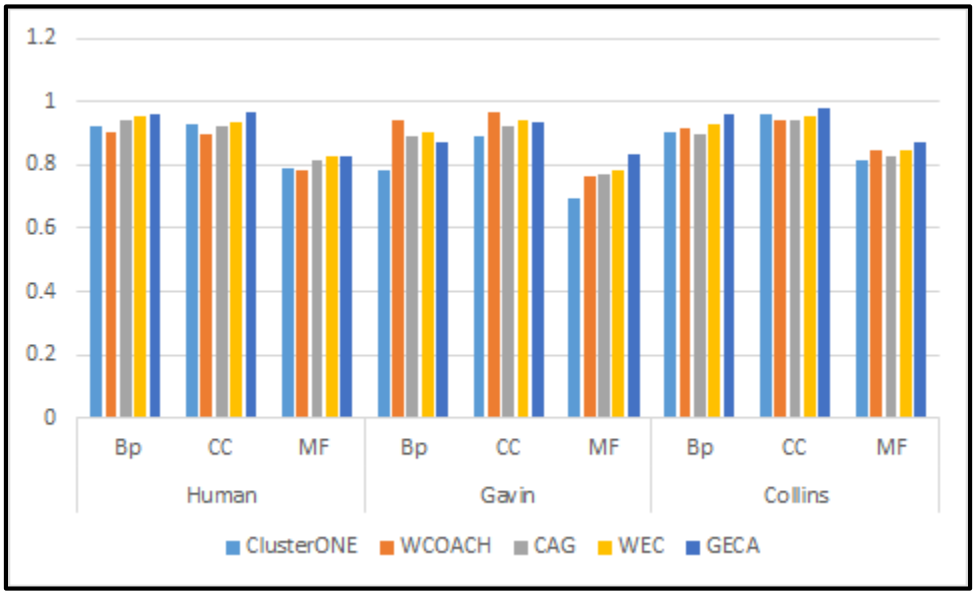


Figure 4-5: GO similarity score of Gavin, Collins, and Human PPI networks.

Good and exact matching of reference complexes

In this section, the quality of predicted complexes is evaluated by reporting the number of reference complexes that match them precisely and the number of reference complexes having an OS score equal to or greater than 0.8, but not matching exactly. GECA achieves the best number of matching complexes in all the reference data as shown in Table 4-23. Hence, it can be concluded that GECA predicts complexes that exactly or has best match reference complexes and outperforms other algorithms in this aspect.

Table 4-23: Good and exact matching with reference complexes

	Gavin with MIPS		**Gavin with SGD**		**Collins with MIPS**		**Collins with SGD**		**Human data**	
	exact	**good**	**exact**	**good**	**exact**	**good**	**exact**	**good**	**exact**	**good**
ClusterONE	7	6	14	11	14	8	33	**14**	4	6
WCOACH	**9**	9	16	**13**	4	10	3	13	2	4
WEC	8	11	16	10	14	12	36	11	6	4
CAG	6	12	12	11	15	9	33	13	4	5
GECA	8	**16**	**19**	**13**	**17**	**16**	**37**	**14**	**7**	**7**
Total	**GECA 24**		**GECA 32**		**GECA 33**		**GECA 51**		**GECA 14**	

F. Analysis of predicted complexes

An important aspect of any algorithm that predicts protein complexes is that it can predict overlapping complexes since that protein can participate in multiple complexes. GECA is notably good at predicting overlapping complexes. In the Collins data with SGD, GECA predicts two overlapping complexes which exactly matched reference ones. The first complex is YNL262W YBR278W YPR175W YDR121W and the second one is YOR304W YGL133W YDR121W YJL065C,

which overlaps with YDR121W. Other examples of overlap between exactly matching predictions and reference complexes in human data are the first complex (TCEB3C TCEB2 TCEB1) and the second (TCEB2 TCEB1 TCEB3B), which overlaps in two proteins (TCEB1 and TCEB2). Table 4-24 presents a further example of overlapping complexes from the Collins data with SGD. The two complexes overlap with five genes and both complexes have an OS greater than 0.8 with reference complex. The predicted complexes have densities of 0.8 and 0.82.

GECA predicts complexes that contain a gene which has only a single connection with a complex, using the gene expression information. Figure 4-6 (a) shows the complex that GECA has predicted exactly when using the Collins data with SGD as reference data. GECA is also able to predict complexes with a density of less than 0.5. Table 4-25 shows the predicted complex with a density of 0.48 and a considerable biological significance. It has an OS of 0.35 with a reference complex in the Collins data with SGD.

Moreover, GECA is analysed to show the effectiveness of using GEP and its influence on the F-measure using the OS score. Figure 4-6 (b) shows that the F-measure of the algorithm is the highest in all the data when using GEP followed by the F-measure of the algorithm when constructing the core according to topological relations and ignoring the addition of protein to the core according to its gene expression pattern. The lowest F-measure for the algorithm is observed when constructing the core which depends on its connection with its neighbours and then attaching the closely connected gene to the core without using the GEP. As a result, applying a gene expression in predicting protein complexes enhances the prediction and increase its accuracy.

The most important aspect of the GECA algorithm is not merely to keep the proteins and their connections that do not have a gene expression pattern, but to determine how removing these proteins can affect the predicted complexes.

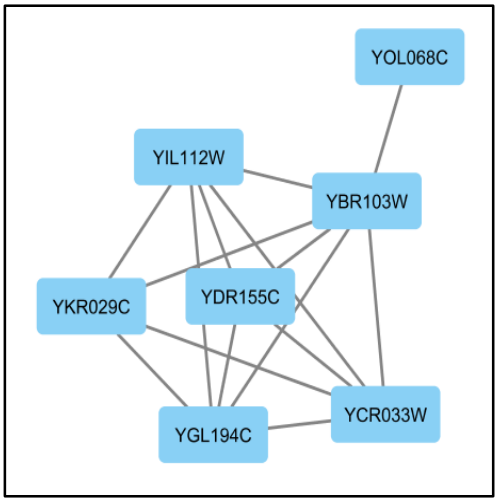
Table 4-26 lists some of the genes which do not have an expression pattern but are found in the predicted complexes in the Collins data with SGD that have an OS greater than 0.8 with a reference complex.

Overall, GECA performs well, if not better than others in all cases with both yeast and human data. It also achieves the highest number of exact and well-predicted complexes in all the data. This indicates that GECA complexes have a high matching with reference complexes. That is, all in all, it increases the accuracy of prediction.

Table 4-24: Overlapping predicted complexes

Predicted complex	Reference **complex**	**Overlapping genes**
YFL049W **YGR275W** YBR289W YDR073W YNR023W **YJL176C** **YMR033W** YHL025W **YPR034W** YPL016W **YOR290C**	**YMR033W** YHL025W YFL049W **YPR034W** YPL129W **YGR275W** YPL016W YBR289W YDR073W YNR023W **YOR290C YJL176C**	**YMR033W** **YPR034W** **YGR275W** **YOR290C** **YJL176C**
YGR056W YCR020W-B YFR037C **YGR275W** YML127W YLR357W **YJL176C** YMR091C **YMR033W** YLR033W YDR303C YKR008W **YPR034W** YHR056C YIL126W YCR052W **YOR290C** YLR321C	YCR020W-B **YGR275W** YML127W YLR357W **YMR033W** YKR008W YPL129W **YPR034W** YPL082C YIL126W YCR052W YLR321C YBL006C YGR056W YFR037C YMR091C YLR033W YDR303C YHR056C **YOR290C**	

(a)



(b)

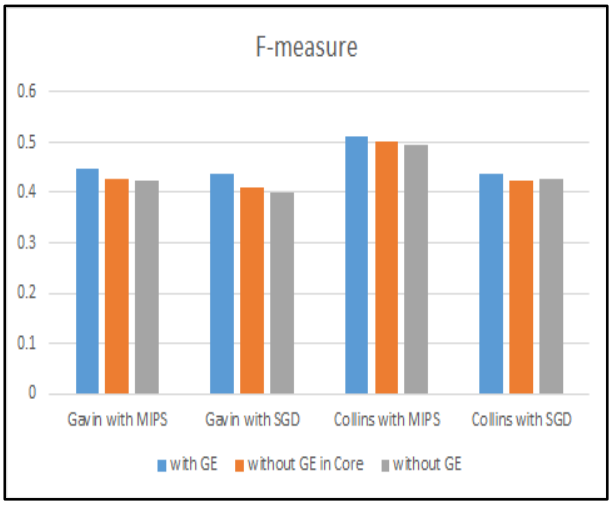


Figure 4-6: (a) Exactly predicted complex with peripheral gene (b) F-measure of GECA with GEP and without GEP

Table 4-25: Predicted complex with density less than 0.5 and overlapping score 0.35 with reference complex.

Predicted complex	Reference complex	Co-localisation	BP	CC	MF
YLR448W YPR107C YMR061W YNL317W YOR179C YBL072C YGL103W YKR002W YOR096W YPL090C YOL127W YDL082W YLL045C YGL076C YHL033C YJL190C YIL133C YFR031C-A YER133W YBR048W YNL222W YBL027W YAL043C YDL083C YKL018W YKL059C YOR063W YPL198W YGR156W YNL178W YMR142C YBR191W YJL177W YDR418W YDR301W YMR194W YLR277C YJR093C YLR441C YER074W YLR029C YDR195W YGR085C YLR115W YNL069C YPL220W	YPR107C YNL222W YLR277C YMR061W YKL018W YKL059C YJR093C YNL317W YGR156W YAL043C YOR179C YDR195W YLR115W YKR002W YER133W YDR301W	0.571	0.867	0.858	0.933

Table 4-26: Gene without GEP but in predicted and reference complexes

Gene	Predicted complex	Reference complex
YHR039C-A	YDL185W **YHR039C-A** YPR036W YGR020C YEL051W YOR332W YBR127C YKL080W	YBR127C YDL185W **YHR039C-A** YPR036W YGR020C YEL051W YOR332W YOR270C YKL080W
YHR056C	YCR052W YGR056W YLR033W YMR091C YIL126W YFR037C YDR303C YCR020W-B YLR357W YMR033W YPR034W YGR275W YML127W YKR008W **YHR056C** YLR321C	YCR052W YMR091C YOR290C YCR020W-B YLR357W YML127W **YHR056C** YGR056W YLR033W YIL126W YFR037C YDR303C YBL006C YPL129W YMR033W YPR034W YGR275W YKR008W YPL082C YLR321C

4.4. Results of DPIN

The experimental findings of DPIN algorithms have been detailed in this part based on the evaluation metrics.

4.4.1. Result of construction DPINs

This section presents the experimental results of creating DPINs using GE data and the quartile one principle.

A. Protein active time points

The q-one technique considers a protein as active at a specific time point if its value is equal to or greater than 25% of the protein's values, which is the first quartile of the protein's gene expression values in all time points. The choice of 25^{th} percentile of a given protein's expression curve is based on analyzing two gene expression data: GSE3431 (Tu et al., 2005) and GSE4987 (Pramila et al., 2006). The number of genes and the number of their active times are calculated using different

percentiles from 25 to 50. Table 4-27 reports that the GSE3431 data which has 6757 genes active in different 27 time points, one gene active in 24 time points and four genes active in different 36 time points and so on. The percentile 25^{th} has the highest number of genes active in different time points in both gene expression data. The researchers realize that the genes active times have successive time points and the highest number of successive times is with 25^{th} percentile.

The goal of this study is to avoid filtering proteins with low gene expression values. Q-one suggests constructing the dynamic networks by taking into account these successive times. To construct the first dynamic network, the protein is either active at time points one and two or at time points two and three. For the second dynamic network, its proteins are active either at time points three and four or at time points four and five, and so on. The interactions in static PIN is kept in DPIN, if both of its proteins are active in the same dynamic network. As a result, the number of constructed DPINs is half the time points of gene expression data. For more analysis, this strategy applies again on two gene expression data, in Table 4-28 GSE3431 shows that 25^{th} percentile contains 32 genes active in two successive time points in different 17 time points, while it has 4 genes active in any two successive times from three successive time points in different 9 PINs and so on. The percentile 25^{th} produces the best cover of genes in each network with a good distribution range and at the same time produces a pre-filtering stage of genes' active times.

The proteins of a complex are grouped and interacted at the same time and place. Reference protein complexes (RPCs) CYC2008 with 408 complexes have been analyzed to determine the number of times shared between their proteins using the 3-sigma and q-one methods. The results show that the number of RPCs that proteins do not share any time using the q-one method is 8 instead of 241 RPCs utilizing the 3-sigma method with the GSE3431 data. Another analysis is done for

RPCs by calculating the number of RPCs whose proteins share the same times. The quartile is employed one threshold only with 36 time points and the q-one method with 18 networks using different percentile from 25 to 50 and utilizing two gene expression data. The number of RPCs whose proteins do not share any time is the lowest using the q-one method with 25th percentile especially with the GSE4987 data Figure 4-7.

Table 4-27: The number of active genes in different time points using GSE3431 and GSE4987 datasets

	GSE3431		GSE4987	
percentile	**#gene**	**#active times**	**#gene**	**#active times**
25	6757	27	6218	37
	1	24	3	38
	4	36	2	39
			1	41
30	6757	25	6222	35
	4	26	2	36
	1	36		
35	6760	23	6219	32
			5	33
40	6758	22	6222	30
			1	31
			1	33
45	6760	20	6217	27
			5	28
			1	29
			1	33
50	6761	18	6223	25
			1	26

Table 4-28: The number of active genes in two successive time points using GSE3431 and GSE4987 datasets

	GSE3431				GSE4987			
percentile	**# active genes**	**times points**	**# active genes**	**DPIN**	**# active genes**	**times points**	**#active genes**	**DPIN**
25	32	17	4	9	2	23	2	13
	274	18	102	10	12	24	27	14
	836	19	803	11	63	25	222	15
	1583	20	2153	12	151	26	847	16
	1745	**21**	2407	13	398	27	1756	17
	1352	22	1106	14	643	28	2089	18
	763	23	171	15	1001	29	1073	19
	172	24	12	16	**1177**	**30**	203	20
	1	25	0	17	1177	31	5	21
	4	35	4	18	879	32		
					480	33		
					209	34		
					32	35		
					6	36		
30	3	13	1	7	1	19	4	12
	24	14	19	8	1	20	58	13
	182	15	183	9	12	21	291	14
	556	16	1015	10	40	22	940	15
	1118	17	2198	11	125	23	1835	16
	1423	18	2303	12	262	24	1997	17
	1442	**19**	911	13	476	25	948	18
	1149	20	126	14	725	26	142	19
	659	21	5	15	967	27	9	20
	205	22	0	16	**1096**	**28**		
	1	35	0	17	1077	29		
			1	18	788	30		
					436	31		
					185	32		
					33	33		
35	3	10	1	5	1	15	3	10
	17	11	4	6	4	16	36	11
	96	12	33	7	11	17	206	12
	316	13	265	8	43	18	664	13
	653	14	1039	9	108	19	1574	14

	1146	15	2357	10	193	20	2056	15
	1295	**16**	2153	11	339	21	1321	16
	1241	17	822	12	540	22	331	17
	953	18	85	13	752	23	33	18
	748	19	3	14	898	24		
	294	20			**994**	**25**		
					955	26		
					743	27		
					431	28		
					175	29		
					35	30		
					2	31		
40	7	9	2	5	2	13	9	9
	25	10	17	6	13	14	65	10
	123	11	116	7	14	15	243	11
	343	12	610	8	59	16	703	12
	712	13	1744	9	126	17	1610	13
	1151	14	2538	10	230	18	2016	14
	1186	**15**	1383	11	394	19	1245	15
	1174	16	337	12	559	20	301	16
	946	17	15	13	778	21	32	17
	668	18			815	22		
	427	19			857	23		
					886	**24**		
					718	25		
					494	26		
					237	27		
					39	28		
					3	29		
45	2	6	1	3	1	9	1	6
	7	7	3	4	2	10	13	7
	56	8	26	5	10	11	35	8
	148	9	162	6	34	12	168	9
	373	10	739	7	74	13	526	10
	702	11	1693	8	161	14	1173	11
	1099	12	2312	9	282	15	1803	12
	1172	**13**	1453	10	417	16	1805	13
	1082	14	362	11	543	17	610	14
	913	15	11	12	727	18	88	15
	792	16			**816**	**19**	2	16

	416	17			770	20		
					736	21		
					635	22		
					384	23		
					425	24		
					187	25		
					20	26		
50	5	4	2	2	2	8	2	5
	8	5	6	3	11	9	7	6
	41	6	31	4	37	10	42	7
	157	7	214	5	71	11	213	8
	396	8	825	6	151	12	559	9
	665	9	1691	7	301	13	1184	10
	1075	10	2059	8	436	14	1803	11
	1157	**11**	1653	9	530	15	1559	12
	1041	12	277	10	725	16	787	13
	936	13	4	11	773	17	70	14
	945	14			805	18		
	336	15			695	19		
					643	20		
					343	21		
					373	22		
					94	23		
					234	24		

Note: GSE3431 shows that 25th percentile contains 32 genes active in two successive time points in different 17 time points, while it has 4 genes active in any two successive times from three successive time points in different 9 PINs and so on.

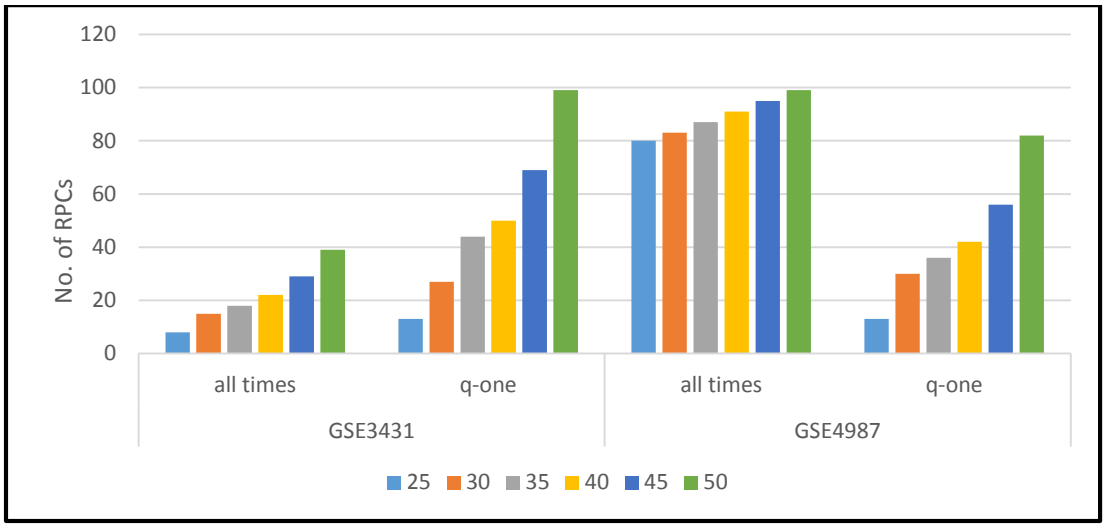


Figure 4-7: No. of RPCs whose proteins do not share any time point

B. PPI and reference Datasets

DIP dataset is used for PPI with 17201 interactions and 4930 proteins. Two gene expression profiles GSE3431 and GSE4987 are employed to construct DPINs. GSE3431 contains 36 time points, while GSE4987 contains 50 time points. CYC2008 contains 408 complexes which are used as reference complexes. All the datasets are the same used by 3-sigma method.

C. Evaluation metrics

Predicted protein complexes (PPCs) from all algorithms using the 3-sigma method and the q-one method are evaluated using precision, recall and F-measure (using α as OS and J-Sim).

In order to explain the difference between the 3-sigma and q-one thresholds, q-one is implemented by using quartile one as a threshold to construct 12 DPINs, using the GSE3431 data and 50 DPINs using the GSE4987 data. Protein complexes from each algorithm are filtered using the 3-sigma filtering strategy (SF) and the

new filtering (NF) strategy proposed by the q-one method for more analysis of two methods.

D. Clustering algorithms results

Improving the quality of algorithms for detecting protein complexes is based on taking the cellular system dynamicity into account (SabziNezhad and Jalili, 2020). Three algorithms MCL, CPM and ClusterONE have been used to evaluate the q-one and 3-sigma methods in predicting protein complexes using DPINs. MCL and ClusterONE are implemented in Cytoscape software using default parameters, while CPM is implemented in python language using default parameters. The ClusterONE and CPM algorithms predict complexes with a size greater than two, while the MCL algorithm predicts complexes with a size greater than one.

GSE3431 results

Figure 4-8 reports the results of three algorithms using the GSE3431 data. In MCL algorithm, q-one achieves the highest F-measure and recall using OS and J-Sim scores with SF strategy and a higher precision, recall and F-measure than 3-sigma using the proposed new filtering strategy. Q-one could predict more precise PCs that have high similarity with RPCs. Q-one with 12 PINs achieves a recall and F-measure higher than 3-sigma using the 3-sigma filtering strategy, and a higher recall, precision and F-measure using the q-one filtering strategy. In addition, it predicts more precise RPCs. In all cases, q-one could achieve the highest F-measure using OS and J-Sim scores with two filtering strategies and the highest number of PCs that match precisely with RPCs as shown in Table 4-29. In the CPM algorithm, q-one achieves a recall and F-measure higher than 3-sigma using SF and NF strategies. Q-one with 12 PINs achieves a recall and F-measure higher than 3-sigma in both filtering strategies. Q-one achieves the highest F-measure in OS score and

came second after q-one with 12 PINs in J-Sim score, but it predicts the highest number of precise RPCs as shown in Table 4-30. In the ClusterONE algorithm, q-one achieves a recall, precision and F-measure higher than 3-sigma using OS and J-Sim to calculate them with NF filtering strategies. Q-one with 12 PINs has achieved a higher recall, precision and F-measure than the 3-sigma with OS and J-Sim using the NF strategy, while it achieves a higher recall and F-measure than the 3-sigma in OS and J-Sim using the SF strategy. Overall, it predicts more precise RPCs in two filtering strategies. Q-one achieves the highest F-measure in OS score and come secondly after q-one with 12 PINs in J-Sim score, but it predicts the highest number of precise RPCs as shown in Table 4-31. In all algorithms and all cases, q-one has achieved high results with 18 PINs and 12 PINs.

GSE4987 results

Figure 4-9 reports the results of three algorithms using the GSE4987 data. In the MCL algorithm, the recall, precision and F-measure of the q-one method are higher than 3-sigma in all cases using the J-Sim score. The results of q-one are almost the same as 3-sigma using an OS score with SF strategy. However, it is higher in precision and F-measure using the NF strategy as shown in Table 4-32. In the CPM algorithm, q-one with SF strategy achieves less precision, recall and F-measure than 3-sigma using OS score, but it achieves a higher precision and F-measure using the J-Sim score. Q-one with NF achieves a higher precision and F-measure than 3-sigma in OS and J-Sim scores as shown in Table 4-33. In the ClusterONE algorithm, q-one achieves a higher recall, precision and F-measure with NF strategy with both scores, while it achieves almost the same results with 3-sigma using OS score but with a higher F-measure using J-Sim score as shown in Table 4-34. In all algorithms, the best results for q-one in recall, precision and F-measure are with 50 dynamic

PINs using OS score, but with the cost of double the number of PPCs, while q-one achieves high results with 25 dynamic PINs using the J-Sim score. In all algorithms and all cases, q-one can predict more PCs that have high matching score with RPCs.

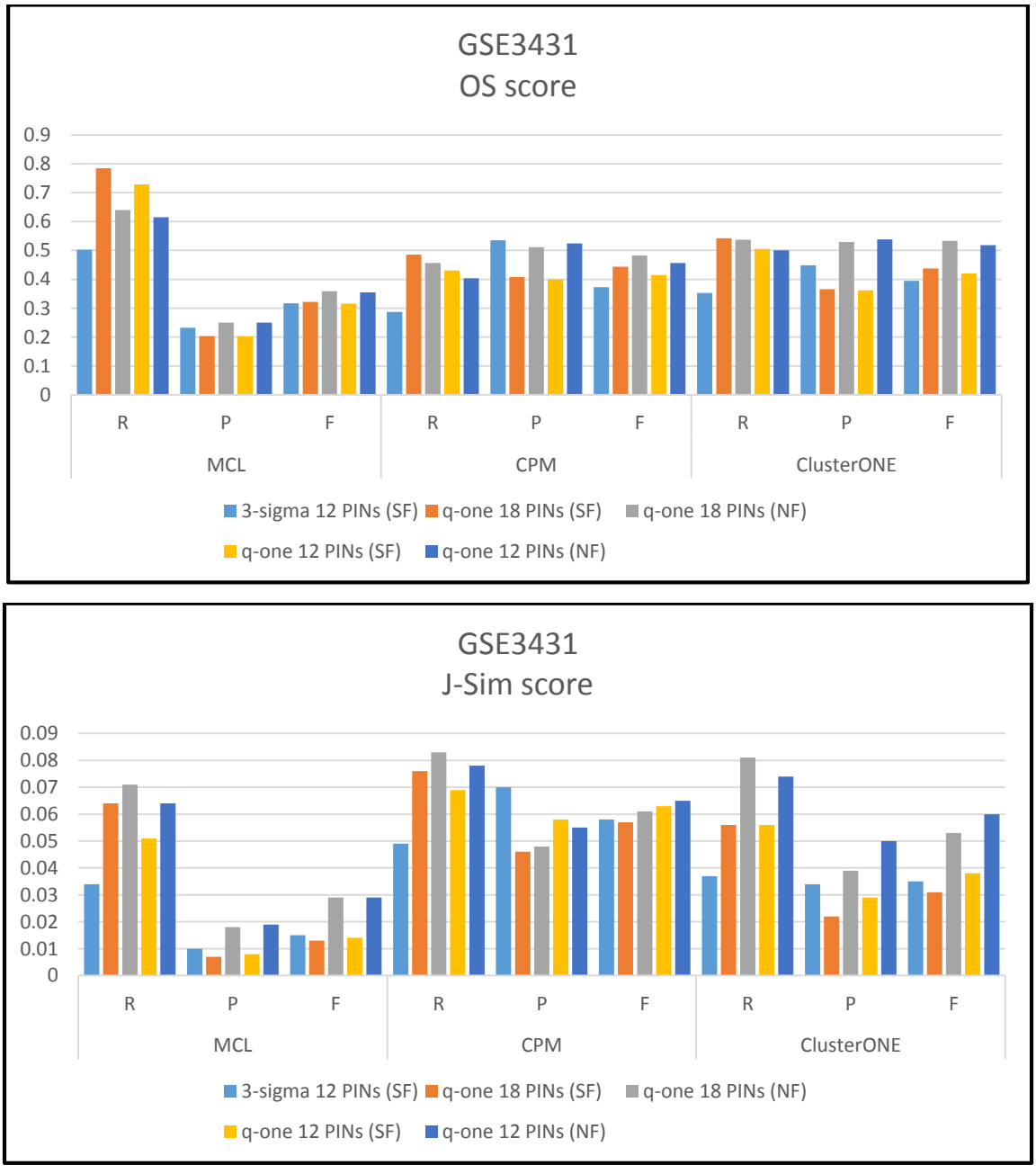


Figure 4-8: Results of three algorithms for two methods with two filtering strategies using GSE3431.

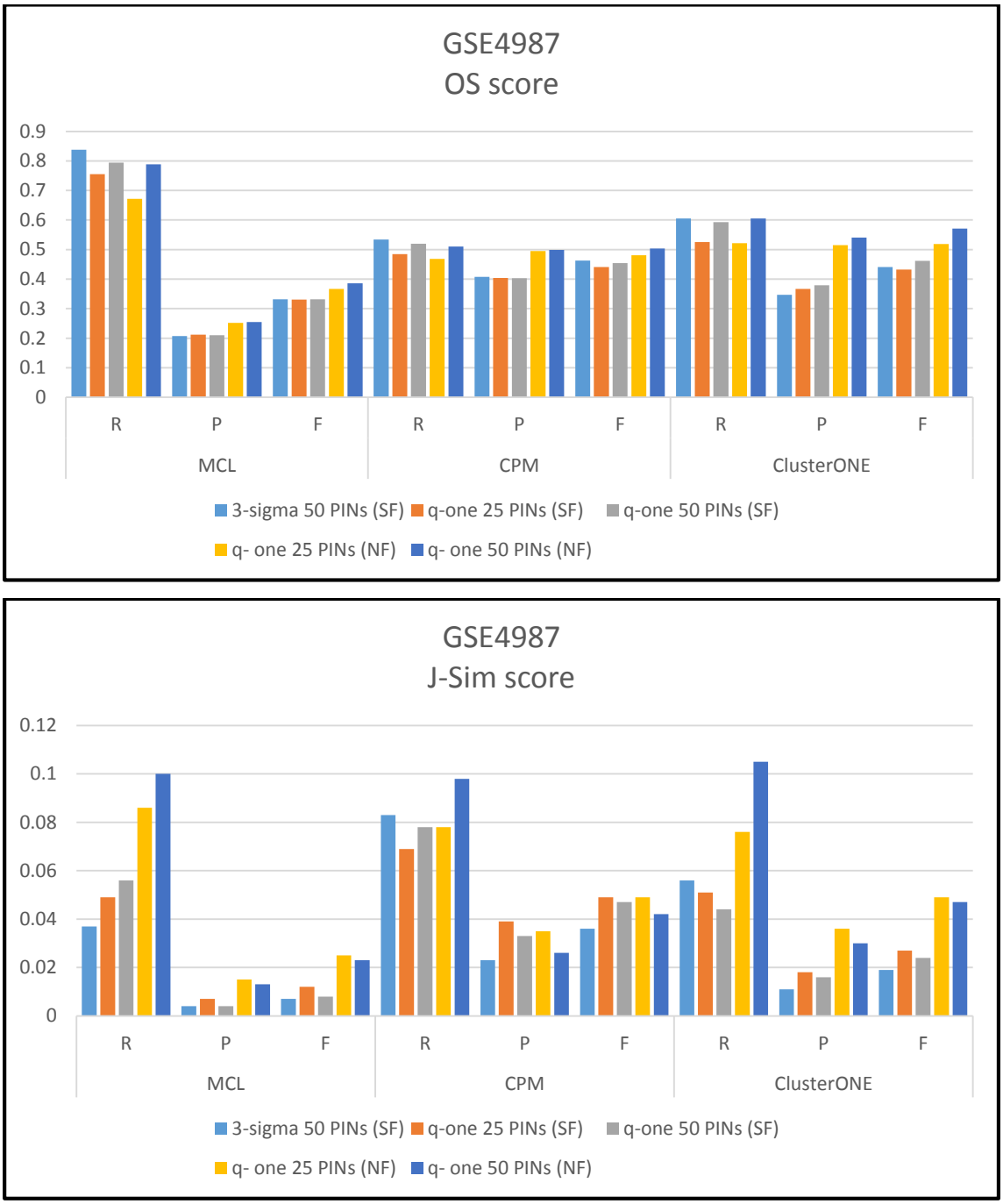


Figure 4-9: Results of three algorithms for two methods with two filtering strategies using GSE4987.

Table 4-29: Analysis results of MCL algorithm using GSE3431

	# PPC	# match	OS					J-sim				
			R	P	F	Exact	High	R	P	F	#RPC	
3-sigma 12 PINs SF	1229	285	0.502	0.232	0.317	17	4	0.034	0.01	0.015	14	
											S	**L**
											6	8
q-one 18 PINs SF	3513	712	0.784	0.203	0.322	36	7	0.064	0.007	0.013	26	
											S	**L**
											7	19
q-one 18 PINs NF	3268	816	0.64	0.25	**0.359**	21	9	0.071	0.018	**0.029**	**29**	
											S	**L**
											5	24
q-one 12 PINs SF	2803	566	0.728	0.202	0.316	31	7	0.051	0.008	0.014	21	
											S	**L**
											5	16
q-one 12 PINs NF	2678	669	0.615	0.25	0.355	18	10	0.064	0.019	**0.029**	26	
											S	**L**
											4	22

Note: R: Recall, P: Precession, F: F-measure, S: Small complex, L: Large complex, PPC: Predicted Protein Complex and RPC: Reference Protein Complex

Table 4-30: Analysis results of CPM algorithm using GSE3431

	# PPC	# match	OS					J-Sim				
			R	P	F	Exact	High	R	P	F	#RPC	
3-sigma 12 PINs SF	301	161	0.287	0.535	0.373	12	4	0.049	0.07	0.058	20	
											S	**L**
											8	12
q-one 18 PINs SF	802	327	0.485	0.408	0.443	17	7	0.076	0.046	0.057	31	
											S	**L**
											11	20
q-one 18 PINs NF	1557	795	0.456	0.511	**0.482**	24	7	0.083	0.048	0.061	**34**	
											S	**L**
											14	20
q-one 12 PINs SF	585	234	0.431	0.4	0.415	17	6	0.069	0.058	0.063	28	
											S	**L**
											11	17
q-one 12 PINs NF	1147	601	0.404	0.524	0.456	20	7	0.078	0.055	**0.065**	32	
											S	**L**
											11	21

Note: R: Recall, P: Precession, F: F-measure, S: Small complex, L: Large complex, PPC: Predicted Protein Complex and RPC: Reference Protein Complex

Table 4-31: Analysis results of ClusterONE algorithm using GSE3431

	# PPC	# match	OS					J-Sim				
			R	P	F	Exact	High	R	P	F	#RPC	
3-sigma 12 PINs SF	475	213	0.353	0.448	0.395	6	6	0.037	0.034	0.035	15	
											S	**L**
											4	11
q-one 18 PINs SF	1293	473	0.542	0.366	0.437	7	9	0.056	0.022	0.031	23	
											S	**L**
											5	18
q-one 18 PINs NF	2255	1193	0.537	0.529	**0.533**	12	14	0.081	0.039	0.053	**33**	
											S	**L**
											4	29
q-one 12 PINs SF	993	359	0.505	0.362	0.421	7	9	0.056	0.029	0.038	23	
											S	**L**
											6	17
q-one 12 PINs NF	1620	871	0.5	0.538	0.518	11	12	0.074	0.05	**0.06**	30	
											S	**L**
											4	26

Note: R: Recall, P: Precession, F: F-measure, S: Small complex, L: Large complex, PPC: Predicted Protein Complex and RPC: Reference Protein Complex

Table 4-32: Analysis results of MCL algorithm using GSE4987

	# PPC	# match	OS					J-Sim				
			R	P	F	Exact	High	R	P	F	#RPC	
3-sigma 50 PINs SF	6071	1258	0.838	0.207	0.332	28	3	0.037	0.004	0.007	15	
											S	**L**
											3	12
q-one 25 PINs SF	3972	841	0.755	0.212	0.331	28	7	0.049	0.007	0.012	20	
											S	**L**
											4	16
q-one 50 PINs SF	5707	1197	0.794	0.21	0.332	30	7	0.056	0.004	0.008	23	
											S	**L**
											7	16
q- one 25 PINs NF	4610	1163	0.672	0.252	0.367	27	9	0.086	0.015	**0.025**	35	
											S	**L**
											9	26
q- one 50 PINs NF	9064	2316	0.789	0.255	**0.386**	41	14	0.1	0.013	0.023	**41**	
											S	**L**
											10	31

Note: R: Recall, P: Precession, F: F-measure, S: Small complex, L: Large complex, PPC: Predicted Protein Complex and RPC: Reference Protein Complex

Table 4-33: Analysis results of CPM algorithm using GSE4987

	# PPC	# match	OS					J-Sim				
			R	P	F	Exact	High	R	P	F	#RPC	
3- sigma 50 PINs SF	1441	588	0.534	0.408	0.463	13	9	0.083	0.023	0.036	34	
											S	**L**
											10	24
q-one 25 PINs SF	882	356	0.485	0.404	0.441	17	9	0.069	0.039	**0.049**	28	
											S	**L**
											12	16
q-one 50 PINs SF	1140	459	0.52	0.403	0.454	19	6	0.078	0.033	0.047	32	
											S	**L**
											13	19
q-one 25 PINs NF	2066	1023	0.468	0.495	0.481	23	7	0.078	0.035	**0.049**	32	
											S	**L**
											12	20
q-one 50 PINs NF	4383	2186	0.51	0.499	**0.504**	30	8	0.098	0.026	0.042	**40**	
											S	**L**
											16	24

Note: R: Recall, P: Precession, F: F-measure, S: Small complex, L: Large complex, PPC: Predicted Protein Complex and RPC: Reference Protein Complex

Table 4-34: Analysis results of ClusterONE algorithm using GSE4987

	# PPC	# match	OS					J-Sim				
			R	P	F	Exact	High	R	P	F	#RPC	
3-sigma 50 PINs SF	2290	794	0.605	0.347	0.441	5	9	0.056	0.011	0.019	23	
											S	**L**
											4	19
q-one 25 PINs SF	1482	544	0.525	0.367	0.432	6	10	0.051	0.018	0.027	21	
											S	**L**
											5	16
q-one 50 PINs SF	1925	729	0.593	0.379	0.462	7	10	0.044	0.016	0.024	20	
											S	**L**
											6	14
q-one 25 PINs NF	2950	1520	0.522	0.515	0.519	13	11	0.076	0.036	**0.049**	31	
											S	**L**
											4	27
q-one 50 PINs NF	5424	2932	0.605	0.541	**0.571**	17	18	0.105	0.03	0.047	**43**	
											S	**L**
											7	36

Note: R: Recall, P: Precession, F: F-measure, S: Small complex, L: Large complex, PPC: Predicted Protein Complex and RPC: Reference Protein Complex

E. The effects of percentile threshold selection and filtering strategy

There are two main factors that affect the results of the q-one method: the percentile threshold selection and filtering strategy. Figure 4-10 shows the effects of percentile threshold selection on F-measure using j-Sim score with two filtering strategies: 3-sigma filtering (SF) and new filtering (NF) suggested by the q-one method with two gene expression data. The results show that 25^{th} percentile produces almost the best f-measure with all algorithms in two gene expression data. Figure 4-11 shows the F-measure of the q-one method with 25 PINs and 50 PINs with different percentiles and how the new method of dynamic network construction enhances the F-measure, and at the same time, it reduces the number of predicted protein complexes and dynamic networks constructed.

The new filtering strategy overcomes the sigma filtering strategy not only by enhancing the recall, precision and F-measure of all algorithms Figure 4-12, but also by the number of exact and well predicted complexes Figure 4-13. The ClusterONE and CPM algorithms predict complexes with a size greater than two, whereas the MCL algorithm predicts complexes with size two. The F-measure with OS and J-Sim scores is calculated for the MCL algorithm with different filtering thresholds for complexes with size two. The results show how the filtering threshold enhances the F-measure and reduces the number of predicted complexes. At the same time, it does not affect the number of small (S) reference complexes that have a J-Sim score equal to one or large (L) reference complexes that have a J-Sim score equal to 0.75 with a predicted complex as shown in Table 4-35. Therefore, the filtering threshold for two-protein complexes is chosen greater than five based on the analysis of reference complexes.

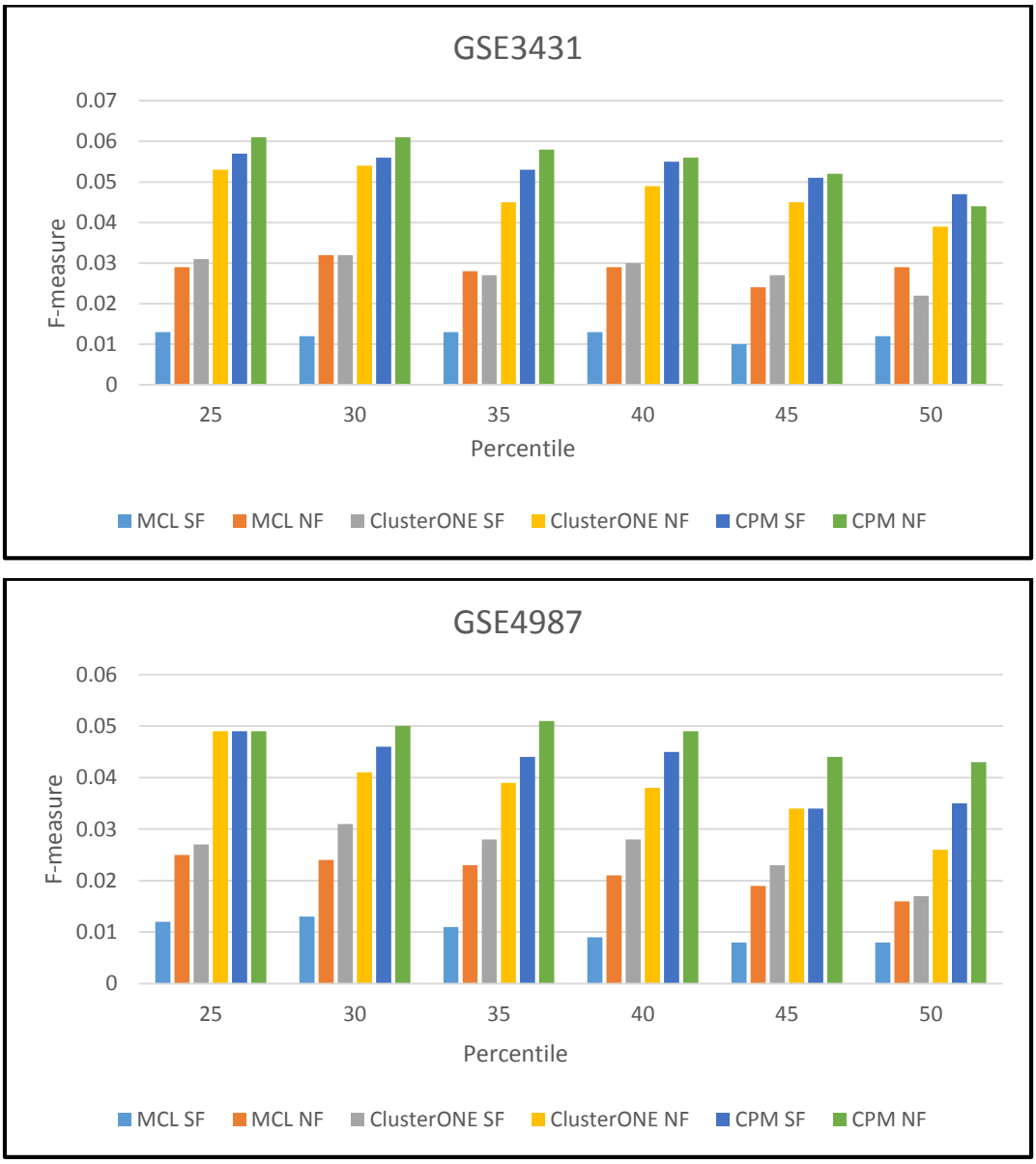


Figure 4-10: F-measure of three algorithms with different percentile thresholds using two filtering strategies.

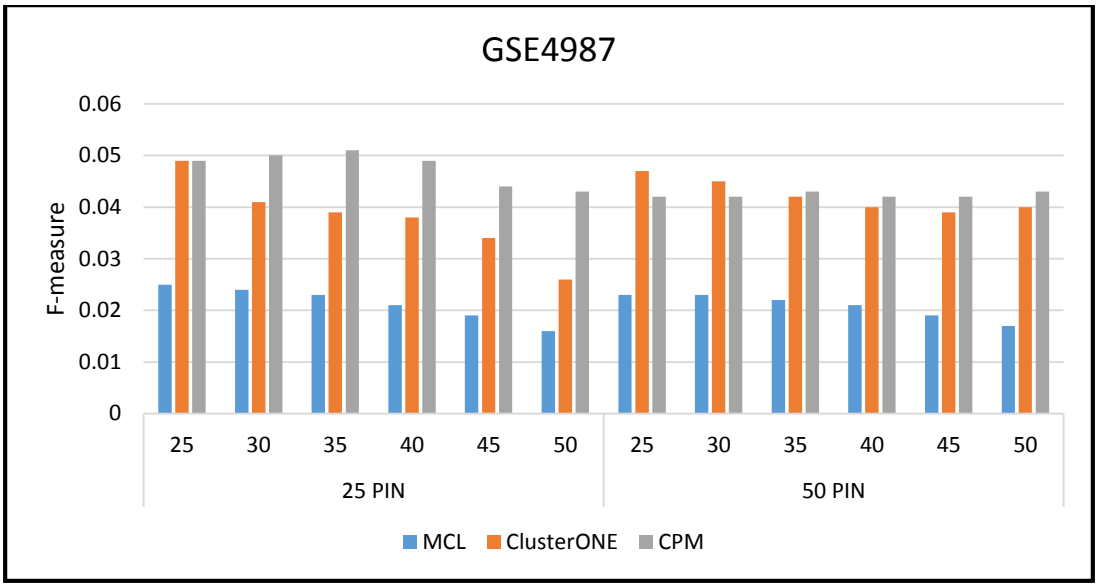


Figure 4-11: F-measure of three algorithms using 25 PINs and 50 PINs with different percentile thresholds.

Table 4-35: F-measure of MCL algorithm with different filtering thresholds for size-two protein complexes.

Threshold	#PPC	#match	F-measure (OS)	F-measure (J-Sim)	S	L
0	4897	1113	0.356	0.021	**5**	**24**
1	4108	956	0.355	0.024		
2	3745	882	0.354	0.026		
3	3519	848	0.355	0.028		
4	3371	828	0.358	0.028		
5	3268	816	0.359	0.029		

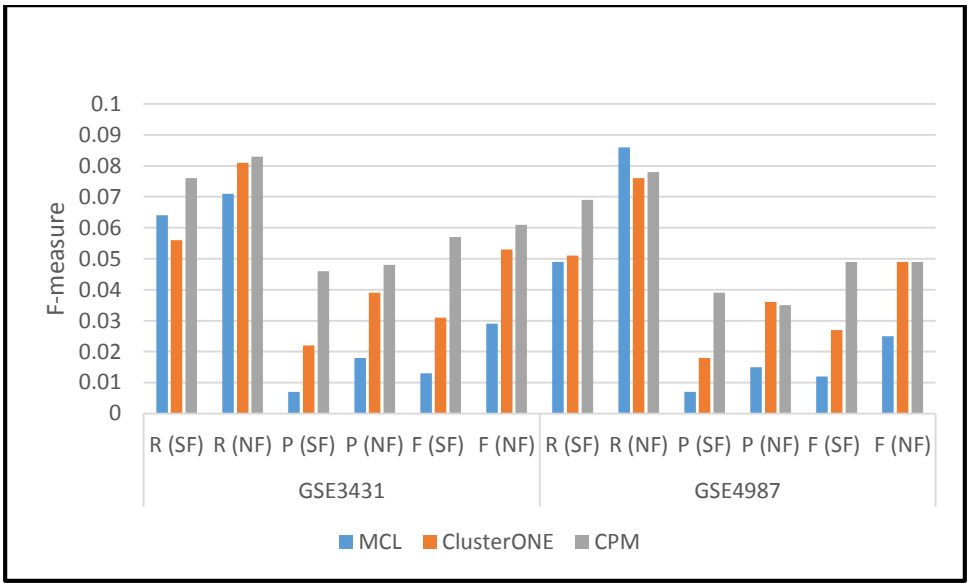


Figure 4-12: Recall (R), precision (P) and F-measure (F) of three algorithms using two filtering strategies.

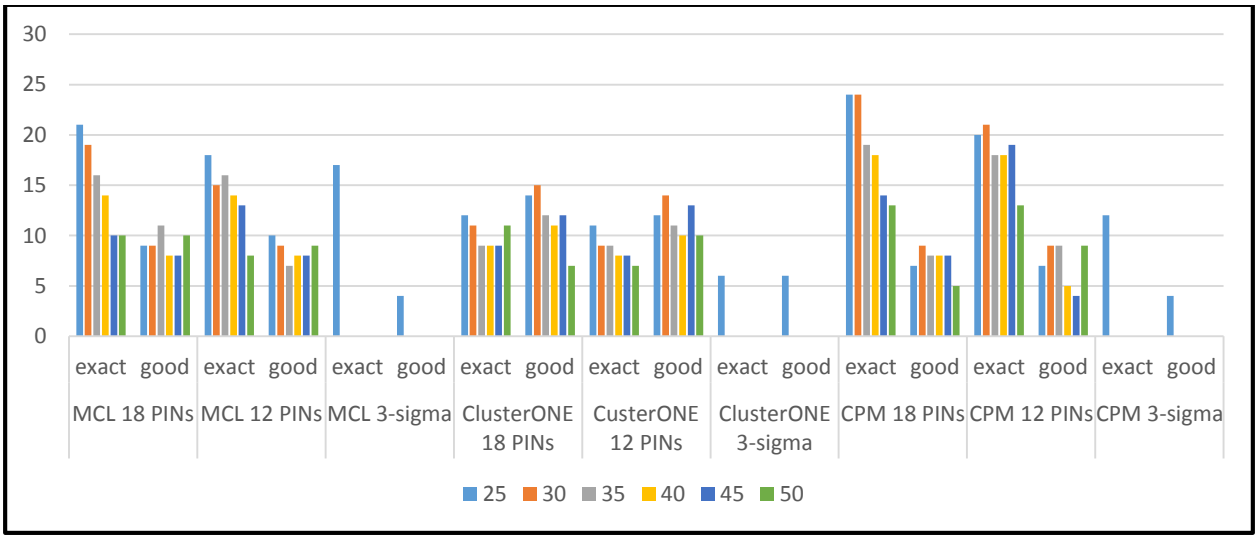


Figure 4-13: Exact and good matching of RPCs with PPCs for three algorithms.

F. Dynamic protein interaction networks analysis

The properties of DPINs have been analyzed in terms of the number of active proteins in each PIN with two GE profiles Figure 4-14. By applying GSE3431, the average number of proteins is about 14% in 3-sigma, 63% in q-one with 18 PINs and 73% in q-one with 12 PINs. However, by applying GSE4987, the average number of proteins is about 37% in 3-sigma, 61% in q-one with 25 PINs and 69% in q-one with 50 PINs. The total number of proteins in all PINs using GSE3431 are 98% in q-one and 50% in 3-sigma from DIP proteins, while it is 92.6% in q-one with 25 PINs, 99.6% in q-one with 50 PINs, and 99.1% in 3-sigma using GSE4987.

Q-one used a good percentage of DIP proteins at every time point which could give the dynamicity of the PPI network. Table 4-36 and Table 4-37 show some examples of this dynamicity. {YML051W, YDR009W} is a PPC which matches exactly with the RPC that appears in network 11 in the MCL algorithm. In network 12, protein YML051W interacts with another protein YPL248C to construct a complex {YML051W, YPL248C} which matches exactly with the RPC. The same PCs appear in the MCL algorithm with GSE4987. The first PC appears in networks 20, 21 and 22, while the second PC appears in network 23. Table 4-37 gives another example of progress in the construction of large protein complexes using the CPM algorithm in which the last PPC exactly matches the reference complex. In the first example, proteins YOR069W and YHR012W, are active proteins in the q-one method only while 3-sigma could not catch them as active proteins using GSE3431 as a gene expression data.

The F-measure and the exact match with the RPCs at every DPIN with GSE3431 and GSE4987 are calculated in all algorithms as shown in Figure 4-15 and Figure 4-16. In GSE3431, the curve of 3-sigma is almost the same as q-one with 12 PINs. However, q-one with 12 PINs could achieve a high F-measure and more

precise PCs, even better than q-one with 18 PINs with the CPM and ClusterONE algorithms. In GSE4987, q-one with 25 and 50 PINs achieves better results than 3-sigma in each DPIN with more exact matches of PPCs with RPCs. The results show the superiority of q-one in F-measure, and at the same time it predicts more precise PCs than the 3-sigma method.

The protein interaction network for yeast contains 1167 essential proteins (EP) gathered from MIPS (Mewes et al., 2006), SGD (Dwight et al., 2002), DEG(Zhang and Lin, 2009), and SGDP (Giaever and Nislow, 2014). Zhong et al. (Zhong et al., 2021) proposed a strategy named JDC for identifying essential proteins based on PIN and gene expression data. Statistics are produced to define the active proteins that are essential. The active protein is considered essential if it has the highest number of appearances in all DPINs. The number of active proteins is calculated in one DPIN and above, two DPINs and above, etc. till the highest number of appearances is reached in all DPINs. In GSE3431, 3-sigma has eight proteins active in eight DPINs, four in EP (50%) and six in JDC (75%). Q-one has eleven proteins active in sixteen DPINs, seven in EP (64%) and nine in JDC (82%). In GSE4987, q-one with 25 PINs has twelve proteins active in twenty one DPINs, two in EP(17%) and eleven in JDC (92%). However, q-one with 50 PINs has one protein active in forty-one DPINs, and it is in JDC and EP. Q-one with 50 PINs has eleven proteins active in forty and forty-one DPINs, six in EP (55%) and ten in JDC (91%). 3-sigma has four proteins active in thirty two DPINs, one in EP and four in JDC, while it has thirteen proteins that are active in thirty one and thirty two DPINs, five in EP(38%) and twelve in JDC (92%). As a result, q-one could identify more essential proteins than the 3-sigma method as shown in Figure 4-17.

The recall, precision and F-measure of the q-one method are the highest in most cases. Q-one provides good coverage for PPI proteins and at the same time

shows the network's dynamicity. Q-one will be a good step to debug the construction of protein complexes and functional models in all time points.

Figure 4-14: The number of active proteins in each DPIN in GSE3431 and GSE4987

Table 4-36: Prediction of small protein complexes that exactly match reference complexes changed over dynamic PINs.

DPIN	PPC1	RPC	DPIN	PPC2	RPC
11	{'YML051W', 'YDR009W'}	{'YML051W', 'YDR009W'}	**12**	{'YML051W', 'YPL248C'}	{'YML051W', 'YPL248C'}
23	{'YLR182W', 'YDL056W'}	{'YLR182W', 'YDL056W'}	**24** **25**	{'YLR182W', 'YER111C'}	{'YLR182W', 'YER111C'}
20 **21** **22**	{'YML051W', 'YDR009W'}	{'YML051W', 'YDR009W'}	**23**	{'YML051W', 'YPL248C'}	{'YML051W', 'YPL248C'}
14	{'YPL046C', 'YJR052W'}	{'YBR114W', 'YPL046C', 'YJR052W', 'YKL112W'}	**15**	{'YPL046C', 'YNL230C'}	{'YPL046C', 'YNL230C'}

Table 4-37: Evaluation of Large PPCs that exactly match RPCs over DPINs.

Network	PPC	Network	PPC	Network	PPC
7	{'YJL154C', 'YOR132W', 'YJL053W'}	8	{'YJL154C', 'YHR012W', 'YOR132W', 'YJL053W'}	9	{'YOR069W', 'YOR132W', 'YJL053W', 'YHR012W', 'YJL154C'}
1	{'YHR119W', 'YAR003W', 'YBR258C', 'YKL018W', 'YLR015W', 'YPL138C'}	2	{'YAR003W', 'YBR258C', 'YBR175W', 'YLR015W', 'YKL018W', 'YPL138C'}	4	{'YHR119W', 'YAR003W', 'YBR258C', 'YBR175W', 'YKL018W', 'YLR015W', 'YDR469W', 'YPL138C'}
2	{'YLR384C', 'YPL101W', 'YHR187W', 'YMR312W'}	3	{'YMR312W', 'YGR200C', 'YPL086C', 'YLR384C', 'YPL101W'}	5	{'YMR312W', 'YGR200C', 'YPL086C', 'YLR384C', 'YHR187W', 'YPL101W'}

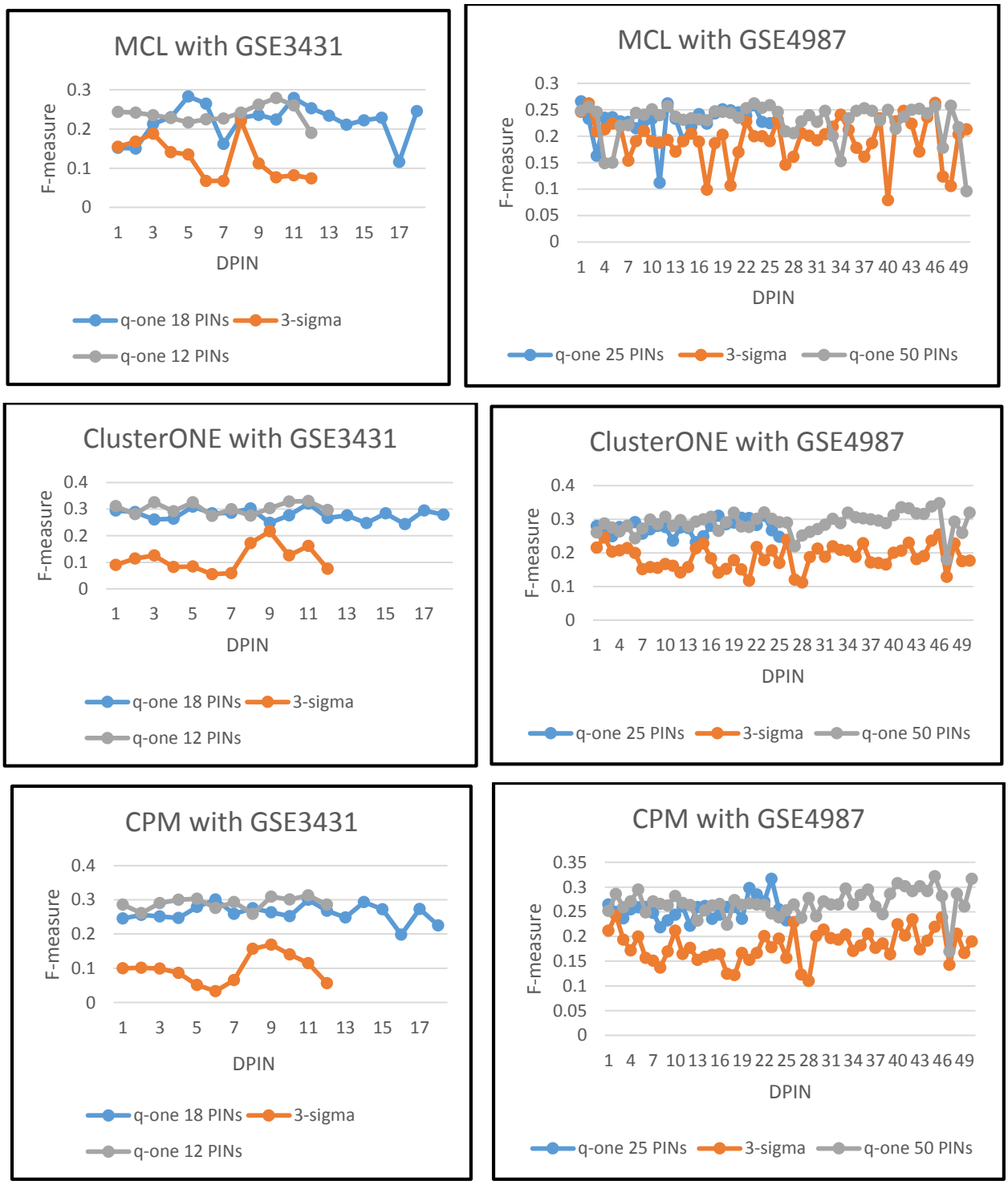


Figure 4-15: F-measure for every DPIN

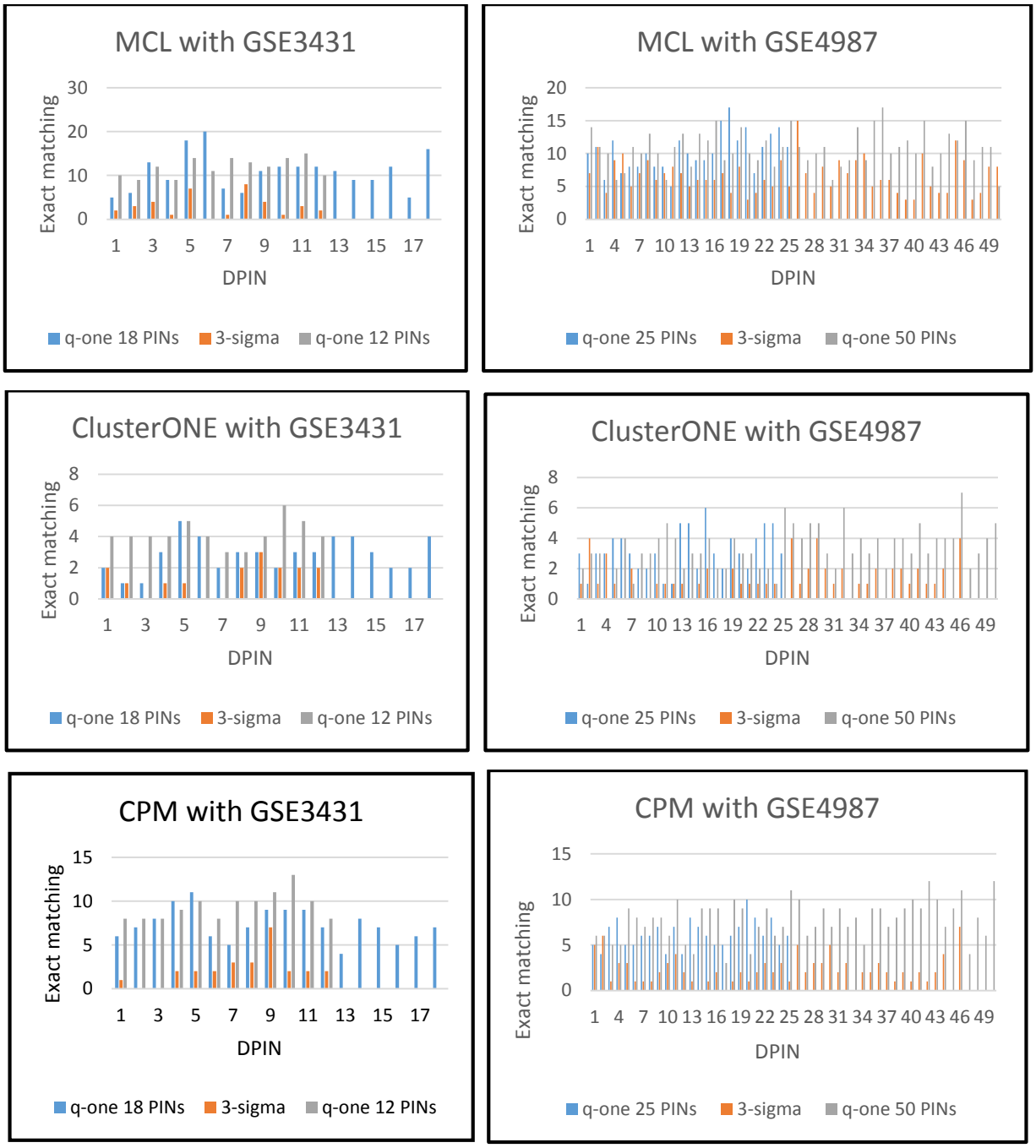


Figure 4-16: Number of exact PPCs in each DPIN

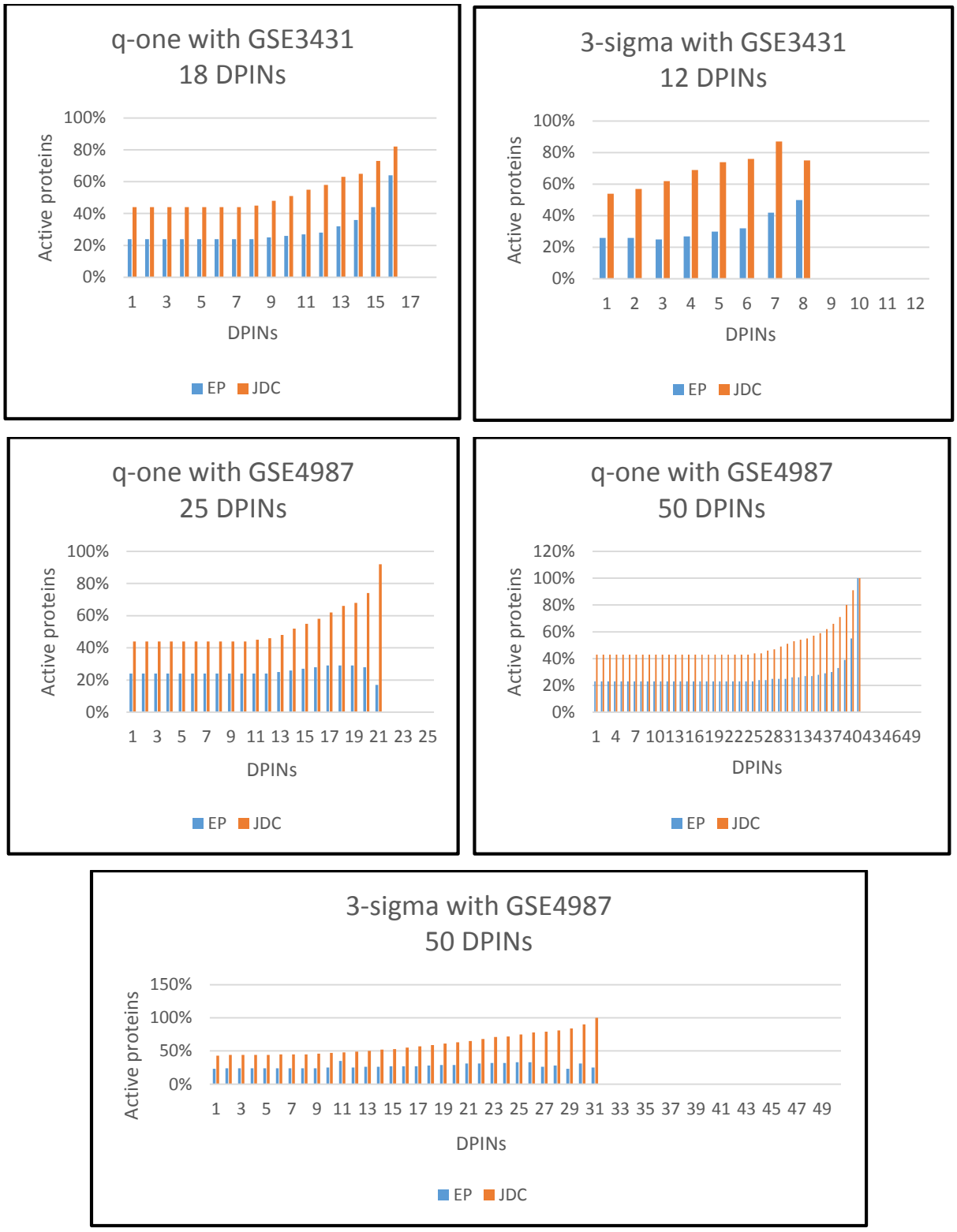


Figure 4-17: Essential proteins for each method using GSE3431 and GSE4987

4.4.2. Results of Shared-one-time Algorithm

The experimental results of shared-one-time algorithm are discussed based on different evaluation metrics.

A. PPI and reference Datasets

Shared-one-time algorithm has been analyzed by concentrating on six PPI networks of *Saccharomyces cerevisiae* (yeast). They are Collins, krogan-extended and Gavin for ClusterONE algorithm, DIP, Krogan and BioGRID from SPICi algorithm. NewMIPS and CYC2008 are used as reference complexes. All datasets are available online from their authors.

B. Comparison of the shared-one-time algorithm with other algorithms

The performance of the algorithm has been compared to those of seven others: ClusterONE, NCMine, SPICi, IPCA, SETS, PEWCC and C0-DPC, which is a novel approach for detecting dynamic protein complexes based on the core-attachment principle. CO-DPC first uses gene expression profiles and the 3-sigma principle to identify active proteins. It employs the co-expression principle and PPI networks to build dynamic PPI networks. Second, CO-DPC recognizes dense local subgraphs as the cores of protein complexes. Then, it attaches them with their neighbors to form protein complexes.

C. Evaluation Metrics

Recall, precision and F-measure (using α as OS) are used to assess the quality of a predicted complex, as well as the number of reference complexes that exactly match the predicted complexes and have an OS score more than or equal to 0.8, excluding

the exact match. The values of the parameters used with each dataset are listed in Table 4-38.

Table 4-38: Threshold values of each parameter

Datasets	T_{CN}	DT	T_{CS}
Collins	0.3	0.5	0.6
Gavin	0.3	0.5	0.5
Krogan	0.2	0.5	0.5
DIP	0.1	0.5	0.6
BioGRID	0.2	0.5	0.7
Krogan-extended	0.1	0.5	0.7

D. Quality of Predicted Complexes

The performance of the algorithm is compared with that of seven other approaches by using six datasets for yeast. All datasets are unweighted except SPICi, which uses weighted networks. Every parameter in all the algorithms is set to default. In addition, complexes with less than three proteins are ignored. All the algorithms are implemented in the Cytoscape software except SPICi, which is implemented in its web site. The complex is considered matched if the OS with the reference complex is greater than or equal to 0.2. The shared-one-time algorithm has the highest F-measure in all cases and competes with other algorithms in recall and precession. The exact and well predicted complexes by the shared-one-time algorithm are the best in most cases. Table 4-39, Table 4-40, Table 4-41, Table 4-42, Table 4-43 , Table 4-44 explain all the results.

The shared-one-time algorithm predicts overlapping complexes as explained in Table 4-45 that reports some of those complexes that have high OS scores equal

to or greater than 0.8 with reference complexes and their proteins share the same active time point. The shared-one-time method depicts the dynamic evolution of protein complex construction. Using the krogan-extended dataset, the algorithm predicts that PPC shares 20 of the RPC's 22 proteins in time point two, and that PPC contains all 22 proteins in time point three. At time point five, PPC has seven of the eight RPC proteins. Further, at time point six, PPC is identical to RPC as shown in Table 4-46. The algorithm could serve as a solid example of how protein complexes evolve over time.

Table 4-39: Performance analysis for Collins data with CYC2008 and NewMIPS.

	# complex	# match	R	P	F	Exact	Good	Total
CYC2008								
SPICi	106	78	0.419	**0.736**	0.534	20	20	40
ClusterONE	203	111	**0.559**	0.547	0.553	33	19	52
NCMine	377	179	0.517	0.475	0.495	25	14	39
PEWCC	426	222	0.53	0.521	0.525	27	24	51
IPCA	342	219	0.542	0.64	0.587	34	18	52
SETS	218	158	0.521	0.725	**0.606**	33	21	54
CO_DPC	1496	584	0.487	0.39	0.433	26	18	44
shared-one-time	522	359	0.542	0.688	**0.606**	**35**	23	**58**
NewMIPS								
SPICi	106	77	0.473	0.726	0.573	18	24	42
ClusterONE	203	110	0.588	0.542	0.564	27	26	53
NCMine	377	189	0.537	0.501	0.518	21	20	41
PEWCC	426	227	0.546	0.533	0.539	23	28	51
IPCA	342	241	0.567	0.705	0.628	32	21	53
SETS	218	166	0.555	0.761	0.642	31	26	57
CO_DPC	1501	766	0.473	0.51	0.491	20	18	38
shared-one-time	522	394	0.573	0.755	**0.652**	**33**	29	**62**

Note: R: Recall, P: Precision, F: F-measure.

Table 4-40: Performance analysis for Gavin data with CYC2008 and NewMIPS.

	# complex	# match	R	P	F	Exact	Good	Total
CYC2008								
SPICi	91	70	0.36	0.76	0.491	14	10	24
ClusterONE	258	108	0.508	0.419	0.459	11	22	33
NCMine	621	244	0.513	0.393	0.445	9	14	23
PEWCC	656	264	0.517	0.402	0.453	11	20	31
IPCA	464	212	0.53	0.457	0.491	**15**	19	34
SETS	246	148	0.475	0.602	0.531	12	25	37
CO_DPC	1052	513	0.483	0.488	0.485	14	28	**42**
shared-one-time	584	354	0.504	0.606	**0.551**	**15**	26	41
NewMIPS								
SPICi	91	67	0.372	0.736	0.494	11	15	26
ClusterONE	258	108	0.53	0.419	0.468	11	19	30
NCMine	621	242	0.549	0.39	0.456	10	16	26
PEWCC	656	284	0.552	0.433	0.485	13	21	34
IPCA	464	218	0.573	0.47	0.516	**17**	25	42
SETS	247	159	0.524	0.607	0.563	13	31	44
CO_DPC	1052	562	0.476	0.534	0.503	13	23	36
shared-one-time	584	358	0.54	0.613	**0.574**	15	31	**46**

Note: R: Recall, P: Precision, F: F-measure.

Table 4-41: Performance analysis for Krogan data with CYC2008 and NewMIPS.

	# complex	# match	R	P	F	Exact	Good	Total
CYC2008								
SPICi	131	84	0.458	0.641	0.534	17	15	32
ClusterONE	240	123	0.492	0.512	0.502	12	15	27
NCMine	578	250	0.458	0.433	0.445	5	17	22
PEWCC	708	351	0.525	0.496	0.51	15	24	39
IPCA	472	281	0.517	0.595	0.553	19	15	34
SETS	220	168	0.479	0.764	0.589	19	20	39
CO_DPC	555	435	0.428	0.784	0.554	16	19	35
shared-one-time	484	382	0.517	0.789	**0.625**	**20**	23	**43**
NewMIPS								
SPICi	131	81	0.479	0.618	0.54	**16**	17	33
ClusterONE	240	110	0.442	0.458	0.45	9	12	21
NCMine	578	247	0.479	0.427	0.452	7	13	20
PEWCC	708	337	0.534	0.476	0.503	10	22	32
IPCA	472	271	0.515	0.574	0.543	13	18	31
SETS	220	161	0.485	0.732	0.583	14	21	35
CO_DPC	554	411	0.405	0.742	0.524	10	18	28
shared-one-time	484	367	0.521	0.758	**0.618**	**16**	23	**39**

Note: R: Recall, P: Precision, F: F-measure.

Table 4-42: Performance analysis for DIP data with CYC2008 and NewMIPS.

	# complex	# match	R	P	F	Exact	Good	Total
CYC2008								
SPICi	219	111	0.555	0.507	0.53	13	8	21
ClusterONE	342	115	0.436	0.336	0.38	7	7	14
NCMine	1074	312	0.542	0.291	0.378	8	10	18
PEWCC	1544	490	0.678	0.317	0.432	**22**	18	40
IPCA	826	263	0.589	0.318	0.413	17	10	27
SETS	540	269	0.653	0.498	0.565	18	14	32
CO_DPC	1147	517	0.521	0.451	0.483	17	13	30
shared-one-time	976	512	0.661	0.525	**0.585**	19	22	**41**
NewMIPS								
SPICi	219	105	0.573	0.479	0.522	11	8	19
ClusterONE	342	104	0.412	0.304	0.35	5	5	10
NCMine	1047	308	0.546	0.287	0.376	5	11	16
PEWCC	1544	491	0.683	0.318	0.434	16	17	33
IPCA	826	257	0.579	0.311	0.405	18	8	26
SETS	540	268	0.64	0.496	0.559	16	17	33
CO_DPC	1141	500	0.491	0.438	0.463	16	8	24
shared-one-time	976	509	0.668	0.522	**0.586**	**20**	20	**40**

Note: R: Recall, P: Precision, F: F-measure.

Table 4-43: Performance analysis for BioGRID data with CYC2008 and NewMIPS.

	# complex	# match	R	P	F	Exact	Good	Total
CYC2008								
SPICi	440	82	0.432	0.186	0.26	4	2	6
ClusterONE	476	126	0.487	0.265	0.343	1	7	8
NCMine	3671	451	0.737	0.123	0.211	4	10	14
PEWCC	4048	792	0.873	0.196	0.32	11	21	**33**
IPCA	2718	381	0.576	0.14	0.226	2	14	16
SETS	633	240	0.644	0.379	0.477	6	**23**	29
CO_DPC	16782	2712	0.763	0.162	0.267	**12**	19	31
shared-one-time	1783	783	0.771	0.439	**0.56**	**12**	21	**33**
NewMIPS								
SPICi	440	79	0.436	0.18	0.254	2	5	7
ClusterONE	476	119	0.488	0.25	0.331	1	6	7
NCMine	3671	478	0.695	0.13	0.219	5	8	13
PEWCC	4048	850	0.826	0.21	0.335	10	20	30
IPCA	2718	374	0.591	0.138	0.223	2	13	15
SETS	633	242	0.622	0.382	0.474	6	21	27
CO_DPC	16763	3298	0.68	0.197	0.305	9	17	26
shared-one-time	1783	799	0.726	0.448	**0.554**	**12**	**24**	**36**

Note: R: Recall, P: Precision, F: F-measure.

Table 4-44: Performance analysis for Krogan_extend data with CYC2008 and NewMIPS.

	# complex	# match	R	P	F	Exact	Good	Total
CYC2008								
SPICi	145	83	0.449	0.572	0.503	13	16	29
ClusterONE	240	105	0.398	0.438	0.417	8	6	14
NCMine	1006	293	0.436	0.291	0.349	3	12	15
PEWCC	1195	452	0.568	0.378	0.454	15	17	32
IPCA	751	293	0.479	0.39	0.43	9	17	26
SETS	431	247	0.538	0.573	0.555	18	18	36
CO_DPC	940	604	0.436	0.643	0.52	16	21	37
shared-one-time	882	562	0.547	0.637	**0.588**	**21**	23	**44**
NewMIPS								
SPICi	145	77	0.463	0.531	0.495	11	17	28
ClusterONE	240	96	0.393	0.4	0.397	7	7	14
NCMine	1006	265	0.442	0.263	0.33	2	10	12
PEWCC	1195	428	0.555	0.358	0.435	13	19	32
IPCA	751	267	0.476	0.356	0.407	7	16	23
SETS	431	229	0.543	0.531	0.537	14	18	32
CO_DPC	944	574	0.424	0.608	0.499	13	18	31
shared-one-time	882	532	0.555	0.603	**0.578**	**17**	21	**38**

Note: R: Recall, P: Precision, F: F-measure.

Table 4-45: Overlapping protein complexes that have high OS score with reference complexes

PPC_1	RPC	PPC_2	RPC	Shared-time	Overlapping proteins
'YJR063W', 'YOR224C', 'YPR110C', 'YJL148W', 'YOR340C', 'YBR154C', 'YDR026C', 'YDR156W', 'YNL248C', 'YPR187W', 'YNL113W', 'YOR210W', 'YOR341W', 'YPR010C'	'YJR063W', 'YOR224C', 'YPR110C', 'YJL148W', 'YOR340C', 'YHR143W-A', 'YBR154C', 'YDR156W', 'YNL248C', 'YPR187W', 'YNL113W', 'YOR210W', 'YOR341W', 'YPR010C'	'YOR207C', 'YOR224C', 'YPR110C', 'YDL150W', 'YNR003C', 'YNL151C', 'YPR187W', 'YBR154C', 'YDR045C', 'YPR190C', 'YER162C', 'YKR025W', 'YOL021C', 'YKL144C', 'YOR210W', 'YNL113W', 'YOR116C', 'YJL011C'	'YOR207C', 'YOR224C', 'YPR110C', 'YDL150W', 'YNR003C', 'YNL151C', 'YHR143W-A', 'YBR154C', 'YDR045C', 'YPR190C', 'YKR025W', 'YKL144C', 'YPR187W', 'YNL113W', 'YOR210W', 'YOR116C', 'YJL011C'	{4, 10}	'YOR224C', 'YPR110C', 'YBR154C', 'YPR187W', 'YNL113W', 'YOR210W'

Table 4-46: dynamic evolution of protein complex construction

PPC	Time	PPC	Time	RPC
'YFR052W', 'YMR191W', 'YOR117W', 'YER021W', 'YHR027C', 'YFR010W', 'YDR427W', 'YPR108W', 'YLR421C', 'YIL075C', 'YDL007W', 'YDR394W', 'YOR259C', 'YHR200W', 'YDL147W', 'YFR004W', 'YDR363W-A', 'YOR261C', 'YDL097C', 'YKL145W', 'YGL048C'	2	'YLR421C', 'YPL004C', 'YER021W', 'YKL145W', 'YPR108W', 'YHR200W', 'YFR010W', 'YDR363W-A', 'YHL030W', 'YOR259C', 'YDR427W', 'YFR052W', 'YGR232W', 'YDR394W', 'YIL075C', 'YDL097C', 'YOR117W', 'YOR261C', 'YHR027C', 'YHR170W', 'YDL147W', 'YDL007W', 'YBR080C', 'YGL048C', 'YMR191W', 'YFR004W'	3	'YFR052W', 'YOR117W', 'YER021W', 'YHR027C', 'YFR010W', 'YDR427W', 'YPR108W', 'YHL030W', 'YLR421C', 'YIL075C', 'YDL007W', 'YDR394W', 'YOR259C', 'YHR200W', 'YDL147W', 'YFR004W', 'YDR363W-A', 'YOR261C', 'YGR232W', 'YDL097C', 'YKL145W', 'YGL048C'
'YFR051C', 'YIL076W', 'YDL145C', 'YPL010W', 'YDR238C', 'YNL287W', 'YGL137W'	5	'YFR051C', 'YIL076W', 'YDL145C', 'YPL010W', 'YDR238C', 'YER122C', 'YNL287W', 'YGL137W'	6	'YFR051C', 'YIL076W', 'YDL145C', 'YPL010W', 'YDR238C', 'YER122C', 'YNL287W', 'YGL137W'

Chapter Five **Conclusion and Future Work**

5.1. Conclusion

The key to exploring cell behavior is understanding the mechanism of protein complexes. For protein complexes prediction, it is critical to build more reliable biological network clustering algorithms. Four algorithms are proposed to predict protein complexes, two for SPIN and two for DPIN:

1. SPIN

- SETS algorithm
  - I. The SETS algorithm has been shown to be effective at predicting overlapping protein complexes of various densities, rather than being limited to dense complexes as previous algorithms are.
  - II. It also has the potential to provide an acceptable execution time, which is less than five minutes for humans and around one minute for yeast datasets.
  - III. SETS achieves high accuracy in all datasets that have different densities with good biological significance of predicted complexes compared to other methods
- GECA algorithm
  - I. It weights SPIN with GE data without removing any unweighted edges from the network as previous methods did. It leads to enhanced results because it can avoid the loss of important nodes that have no gene expression information.
  - II. It applies a new technique to attach the proteins that allowed the addition of a protein that is connected to fewer than half the proteins in the core, but with good similarity in the gene expression pattern to the core proteins.

III. The evaluation showed that GECA performed well, if not best, regarding the F-measure, co-localization score, number of exact PCs, and GO semantic similarity scores. Moreover, the results demonstrated that GECA effectively identifies protein complexes with high biological significance.

2. DPIN

- A strategy named q-one for transforming SPIN into DPIN is proposed.
  I. Q-one employs a strategy to include genes with low expression values with some restrictions in the DPIN construction, while most available methods exclude genes with a low expression level.
  II. The threshold used by q-one is the median of the bottom half of the protein's values. A protein is added to DPIN if its value is equal to or greater than 25% of the protein's values in two successive time points. This strategy reduces the DPINs to half of the time points of gene expression data, which makes the method unique.
  III. Q-one proves that the percentile 25 is optimal or near optimal threshold to identify active proteins.
  IV. It leads to cover most proteins in the PPI network with 30%-35% difference in each DPIN.
  V. The evaluation reveals that the suggested method yields better results than the 3-sigma method and provides significant ramifications for the understanding of the dynamic of the PPI network, as well the ability to build effective DPIN.
- Shared-one-time algorithm

I. Most algorithms that predict dynamic PCs construct DPINs equal to the GEP time points and then apply the algorithm for each DPIN. The shared-one-time algorithm, however, labels each protein with its active time points using the q-one strategy and predicts PCs that share at least one active time point.

II. The algorithm enhances the prediction up to 65% and at the same time depicts the dynamic evolution of protein complex construction.

5.2. Future work

Future work may include the following:

- Improving the prediction accuracy and predicting small protein complexes with two proteins by adding extra biological information to the SPIN, in addition to the gene expression profile.
- Using a bipartite graph to analyze the disease evaluation by applying the GECA algorithm to two GE datasets, one normal and the other cancerous, for the same organ, then analyzing the predicted protein complexes from both datasets.
- Improving the q-one strategy by combining points n-1 and n or n and n+1 into time point n for all time points except the first and last, and comparing the results.

REFERENCES

References

ADAMCSEK, B., PALLA, G., FARKAS, I. J., DERÉNYI, I. & VICSEK, T. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics,* 22**,** 1021-1023.

AITTOKALLIO, T. & SCHWIKOWSKI, B. 2006. Graph-based methods for analysing networks in cell biology. *Briefings in bioinformatics,* 7**,** 243-255.

ALBERTS, B. 1998. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *cell,* 92**,** 291-294.

ALOY, P., BÖTTCHER, B., CEULEMANS, H., LEUTWEIN, C., MELLWIG, C., FISCHER, S., GAVIN, A.-C., BORK, P., SUPERTI-FURGA, G. & SERRANO, L. 2004. Structure-based assembly of protein complexes in yeast. *Science,* 303**,** 2026-2029.

ALTAF-UL-AMIN, M., SHINBO, Y., MIHARA, K., KUROKAWA, K. & KANAYA, S. 2006a. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC bioinformatics,* 7**,** 207.

ALTAF-UL-AMIN, M., SHINBO, Y., MIHARA, K., KUROKAWA, K. & KANAYA, S. 2006b. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC bioinformatics,* 7**,** 1-13.

ARENAS, A., DIAZ-GUILERA, A. & PÉREZ-VICENTE, C. J. 2006. Synchronization reveals topological scales in complex networks. *Physical review letters,* 96**,** 114102.

BADER, G. D. & HOGUE, C. W. 2003a. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics,* 4**,** 2.

BADER, G. D. & HOGUE, C. W. 2003b. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics,* 4.

BERGGÅRD, T., LINSE, S. & JAMES, P. 2007. Methods for the detection and analysis of protein–protein interactions. *Proteomics,* 7**,** 2833-2842.

BHOWMICK, S. S. & SEAH, B. S. 2015. Clustering and summarizing protein-protein interaction networks: A survey. *IEEE Transactions on Knowledge and Data Engineering,* 28**,** 638-658.

BREITKREUTZ, B.-J., STARK, C., REGULY, T., BOUCHER, L., BREITKREUTZ, A., LIVSTONE, M., OUGHTRED, R., LACKNER, D. H., BÄHLER, J. & WOOD, V. 2007. The BioGRID interaction database: 2008 update. *Nucleic acids research,* 36**,** D637-D640.

BROHEE, S. & VAN HELDEN, J. 2006. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics,* 7**,** 488.

CHILDS, L. M. & LARREMORE, D. B. 2021. Network models for malaria: Antigens, dynamics, and evolution over space and time.

CHIN, C.-H., CHEN, S.-H., HO, C.-W., KO, M.-T. & LIN, C.-Y. 2010. A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. *BMC bioinformatics,* 11**,** S25.

COLLINS, S. R., KEMMEREN, P., ZHAO, X.-C., GREENBLATT, J. F., SPENCER, F., HOLSTEGE, F. C., WEISSMAN, J. S. & KROGAN, N. J. 2007. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & Cellular Proteomics,* 6**,** 439-450.

DANČÍK, V., BASU, A. & CLEMONS, P. 2013. Properties of Biological Networks. *In:* PROKOP, A. & CSUKÁS, B. (eds.) *Systems Biology: Integrative Biology and Simulation Tools.* Dordrecht: Springer Netherlands.

DE LICHTENBERG, U., JENSEN, L. J., BRUNAK, S. & BORK, P. 2005. Dynamic complex formation during the yeast cell cycle. *Science,* 307**,** 724-727.

References

DWIGHT, S. S., HARRIS, M. A., DOLINSKI, K., BALL, C. A., BINKLEY, G., CHRISTIE, K. R., FISK, D. G., ISSEL-TARVER, L., SCHROEDER, M. & SHERLOCK, G. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic acids research,* 30**,** 69-72.

EVERSE, S. 2014. *Levels of Protein Organization* [Online]. Available: https://comis.med.uvm.edu/VIC/coursefiles/MD540/MD540-Protein_Organization_10400_574581210/Protein-org/index.html.

FIONDA, V. & PALOPOLI, L. 2011. Biological network querying techniques: analysis and comparison. *Journal of Computational Biology,* 18**,** 595-625.

FIONDA, V., PALOPOLI, L., PANNI, S. & ROMBO, S. E. 2009. A technique to search for functional similarities in protein-protein interaction networks. *International journal of data mining and bioinformatics,* 3**,** 431-453.

FRIEDEL, C. C., KRUMSIEK, J. & ZIMMER, R. Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. Annual International Conference on Research in Computational Molecular Biology, 2008. Springer, 3-16.

GAVIN, A.-C., ALOY, P., GRANDI, P., KRAUSE, R., BOESCHE, M., MARZIOCH, M., RAU, C., JENSEN, L. J., BASTUCK, S. & DÜMPELFELD, B. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature,* 440**,** 631.

GIAEVER, G. & NISLOW, C. 2014. The yeast deletion collection: a decade of functional genomics. *Genetics,* 197**,** 451-465.

GROMIHA, M. M. 2010. *Protein bioinformatics: from sequence to function*, academic press.

HAN, J.-D. J., BERTIN, N., HAO, T., GOLDBERG, D. S., BERRIZ, G. F., ZHANG, L. V., DUPUY, D., WALHOUT, A. J., CUSICK, M. E. & ROTH, F. P. 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature,* 430**,** 88-93.

ITO, T., CHIBA, T., OZAWA, R., YOSHIDA, M., HATTORI, M. & SAKAKI, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences,* 98**,** 4569-4574.

JANSEN, R. & GERSTEIN, M. 2004. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current opinion in microbiology,* 7**,** 535-545.

JANSEN, R., YU, H., GREENBAUM, D., KLUGER, Y., KROGAN, N. J., CHUNG, S., EMILI, A., SNYDER, M., GREENBLATT, J. F. & GERSTEIN, M. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science,* 302**,** 449-453.

JENGHARA, M. M., EBRAHIMPOUR-KOMLEH, H. & PARVIN, H. 2018. Dynamic protein–protein interaction networks construction using firefly algorithm. *Pattern Analysis and Applications,* 21**,** 1067-1081.

JENSEN, L. J., KUHN, M., STARK, M., CHAFFRON, S., CREEVEY, C., MULLER, J., DOERKS, T., JULIEN, P., ROTH, A. & SIMONOVIC, M. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research,* 37**,** D412-D416.

JIANG, P. & SINGH, M. 2010. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics,* 26**,** 1105-1111.

JUNG, S. H., HYUN, B., JANG, W.-H., HUR, H.-Y. & HAN, D.-S. 2010. Protein complex prediction based on simultaneous protein interaction network. *Bioinformatics,* 26**,** 385-391.

References

KERETSU, S. & SARMAH, R. 2016. Weighted edge based clustering to identify protein complexes in protein–protein interaction networks incorporating gene expression profile. *Computational biology and chemistry,* 65**,** 69-79.

KERETSU, S. & SARMAH, R. 2017. Identification of protein complexes in protein-protein interaction networks by core-attachment approach incorporating gene expression profile. *International journal of bioinformatics research and applications,* 13**,** 313-328.

KING, A. D., PRŽULJ, N. & JURISICA, I. 2004. Protein complex prediction via cost-based clustering. *Bioinformatics,* 20**,** 3013-3020.

KONG, P., HUANG, G. & LIU, W. 2020. Identification of protein complexes and functional modules in E. coli PPI networks. *BMC microbiology,* 20**,** 1-9.

KOUHSAR, M., ZARE-MIRAKABAD, F. & JAMALI, Y. 2016. WCOACH: Protein complex prediction in weighted PPI networks. *Genes & genetic systems***,** 15-00032.

KROGAN, N. J., CAGNEY, G., YU, H., ZHONG, G., GUO, X., IGNATCHENKO, A., LI, J., PU, S., DATTA, N. & TIKUISIS, A. P. 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature,* 440**,** 637-643.

LEI, X., FANG, M., GUO, L. & WU, F.-X. 2019. Protein complex detection based on flower pollination mechanism in multi-relation reconstructed dynamic protein networks. *BMC bioinformatics,* 20**,** 63-74.

LEUNG, H. C., XIANG, Q., YIU, S.-M. & CHIN, F. Y. 2009. Predicting protein complexes from PPI data: a core-attachment approach. *Journal of Computational Biology,* 16**,** 133-144.

LI, M., CHEN, J.-E., WANG, J.-X., HU, B. & CHEN, G. 2008a. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC bioinformatics,* 9.

LI, M., CHEN, J.-E., WANG, J.-X., HU, B. & CHEN, G. 2008b. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC bioinformatics,* 9**,** 1-16.

LI, M., CHEN, W., WANG, J., WU, F.-X. & PAN, Y. 2014. Identifying dynamic protein complexes based on gene expression profiles and PPI networks. *BioMed research international,* 2014.

LI, X.-L., FOO, C.-S. & NG, S.-K. 2007. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. *Computational Systems Bioinformatics: (Volume 6).* World Scientific.

LI, X.-L., FOO, C.-S., TAN, S.-H. & NG, S.-K. 2005. Interaction graph mining for protein complexes using local clique merging. *Genome Informatics,* 16**,** 260-269.

LIU, G., WONG, L. & CHUA, H. N. 2009. Complex discovery from weighted PPI networks. *Bioinformatics,* 25**,** 1891-1897.

LIU, G., YONG, C. H., CHUA, H. N. & WONG, L. Decomposing PPI networks for complex discovery. Proteome science, 2011. BioMed Central, 1-11.

LIU, G., YONG, C. H., WONG, L. & CHUA, H. N. Decomposing PPI networks for complex discovery. [Paper presentation]. 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2010, Dec. 18-21 Hong Kong, China. https://doi.org/10.1109/BIBM.2010.5706577. IEEE, 280-283.

LIU, L., SUN, X., SONG, W. & DU, C. 2018. A method for predicting protein complexes from dynamic weighted protein–protein interaction networks. *Journal of Computational Biology,* 25**,** 586-605.

References

LIU, Q., SONG, J. & LI, J. 2016. Using contrast patterns between true complexes and random subgraphs in ppi networks to predict unknown protein complexes. *Scientific reports,* 6**,** 21223.

LIU, W.-C., LIN, W.-H., DAVIS, A. J., JORDÁN, F., YANG, H.-T. & HWANG, M.-J. 2007. A network perspective on the topological importance of enzymes and their phylogenetic conservation. *BMC bioinformatics,* 8**,** 1-12.

MA, C.-Y., CHEN, Y.-P. P., BERGER, B. & LIAO, C.-S. 2017. Identification of protein complexes by integrating multiple alignment of protein interaction networks. *Bioinformatics,* 33**,** 1681-1688.

MAKRODIMITRIS, S., REINDERS, M. J. & VAN HAM, R. C. 2020. Metric learning on expression data for gene function prediction. *Bioinformatics,* 36**,** 1182-1190.

MENG, X., LI, W., PENG, X., LI, Y. & LI, M. 2021. Protein interaction networks: centrality, modularity, dynamics, and applications. *Frontiers of Computer Science,* 15**,** 1-17.

MEWES, H.-W., AMID, C., ARNOLD, R., FRISHMAN, D., GÜLDENER, U., MANNHAUPT, G., MÜNSTERKÖTTER, M., PAGEL, P., STRACK, N. & STÜMPFLEN, V. 2004. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic acids research,* 32**,** D41-D44.

MEWES, H.-W., FRISHMAN, D., MAYER, K. F., MÜNSTERKÖTTER, M., NOUBIBOU, O., PAGEL, P., RATTEI, T., OESTERHELD, M., RUEPP, A. & STÜMPFLEN, V. 2006. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic acids research,* 34**,** D169-D172.

NEPUSZ, T., YU, H. & PACCANARO, A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods,* 9**,** 471-472.

NYMARK, P., LINDHOLM, P. M., KORPELA, M. V., LAHTI, L., RUOSAARI, S., KASKI, S., HOLLMÉN, J., ANTTILA, S., KINNULA, V. L. & KNUUTILA, S. 2007. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC genomics,* 8**,** 62.

OU-YANG, L., DAI, D.-Q., LI, X.-L., WU, M., ZHANG, X.-F. & YANG, P. 2014. Detecting temporal protein complexes from dynamic protein-protein interaction networks. *BMC bioinformatics,* 15**,** 1-14.

PAGE, L., BRIN, S., MOTWANI, R. & WINOGRAD, T. 1999. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.

PALLA, G., DERÉNYI, I., FARKAS, I. & VICSEK, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature,* 435**,** 814-818.

PAVLOPOULOS, G. A., SECRIER, M., MOSCHOPOULOS, C. N., SOLDATOS, T. G., KOSSIDA, S., AERTS, J., SCHNEIDER, R. & BAGOS, P. G. 2011. Using graph theory to analyze biological networks. *BioData mining,* 4**,** 1-27.

PERI, S., NAVARRO, J. D., AMANCHY, R., KRISTIANSEN, T. Z., JONNALAGADDA, C. K., SURENDRANATH, V., NIRANJAN, V., MUTHUSAMY, B., GANDHI, T. & GRONBORG, M. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research,* 13**,** 2363-2371.

PRAMILA, T., WU, W., MILES, S., NOBLE, W. S. & BREEDEN, L. L. 2006. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes & development,* 20**,** 2266-2278.

PRZYTYCKA, T. M., SINGH, M. & SLONIM, D. K. 2010. Toward the dynamic interactome: it's about time. *Briefings in bioinformatics,* 11**,** 15-29.

References

PU, S., VLASBLOM, J., EMILI, A., GREENBLATT, J. & WODAK, S. J. 2007. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics,* 7**,** 944-960.

PU, S., WONG, J., TURNER, B., CHO, E. & WODAK, S. J. 2009. Up-to-date catalogues of yeast protein complexes. *Nucleic acids research,* 37**,** 825-831.

PUIG, O., CASPARY, F., RIGAUT, G., RUTZ, B., BOUVERET, E., BRAGADO-NILSSON, E., WILM, M. & SÉRAPHIN, B. 2001. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods,* 24**,** 218-229.

QUEIROZ, F. C., VARGAS, A. M., OLIVEIRA, M. G., COMARELA, G. V. & SILVEIRA, S. A. 2020. ppiGReMLIN: a graph mining based detection of conserved structural arrangements in protein-protein interfaces. *BMC bioinformatics,* 21**,** 1-25.

RANI, R. R., RAMYACHITRA, D. & BRINDHADEVI, A. 2019. Detection of dynamic protein complexes through markov clustering based on elephant herd optimization approach. *Scientific reports,* 9**,** 1-18.

REN, J., WANG, J., LI, M. & WANG, L. 2013. Identifying protein complexes based on density and modularity in protein-protein interaction network. *BMC systems biology,* 7**,** 1-15.

SABZINEZHAD, A. & JALILI, S. 2020. DPCT: A Dynamic Method for Detecting Protein Complexes From TAP-Aware Weighted PPI Network. *Frontiers in genetics,* 11.

SAHA, S., PRASAD, A., CHATTERJEE, P., BASU, S. & NASIPURI, M. 2019. Protein function prediction from dynamic protein interaction network using gene expression data. *Journal of bioinformatics and computational biology,* 17**,** 1950025.

SCHLICKER, A., DOMINGUES, F. S., RAHNENFÜHRER, J. & LENGAUER, T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics,* 7.

SELVARAJ, C., CHANDRA, I. & SINGH, S. K. 2021. Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. *Molecular Diversity***,** 1-21.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research,* 13**,** 2498-2504.

SPIRIN, V. & MIRNY, L. A. 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences,* 100**,** 12123-12128.

SRIHARI, S., YONG, C. H. & WONG, L. 2017. *Computational prediction of protein complexes from protein interaction networks,* Morgan & Claypool.

STOLL, D., TEMPLIN, M. F., BACHMANN, J. & JOOS, T. O. 2005. Protein microarrays: applications and future challenges. *Current opinion in drug discovery & development,* 8**,** 239-252.

STUMPF, M. P., WIUF, C. & MAY, R. M. 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences,* 102**,** 4221-4224.

TADAKA, S. & KINOSHITA, K. 2016. NCMine: Core-peripheral based functional module detection using near-clique mining. *Bioinformatics,* 32**,** 3454-3460.

TANG, X., WANG, J., LIU, B., LI, M., CHEN, G. & PAN, Y. 2011. A comparison of the functional modules identified from time course and static PPI network data. *BMC bioinformatics,* 12**,** 1-15.

References

TSONIS, A. A. 2007. *Nonlinear Dynamics in Geosciences*, Springer Science & Business Media.

TU, B. P., KUDLICKI, A., ROWICKA, M. & MCKNIGHT, S. L. 2005. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, 310**,** 1152-1158.

UETZ, P., GIOT, L., CAGNEY, G., MANSFIELD, T. A., JUDSON, R. S., KNIGHT, J. R., LOCKSHON, D., NARAYAN, V., SRINIVASAN, M. & POCHART, P. 2000. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, 403**,** 623-627.

VASUDEVAN, D. M., SREEKUMARI, S. & VAIDYANATHAN, K. 2019. *Textbook of biochemistry for medical students*, Jaypee brothers Medical publishers.

VENKATESAN, K., RUAL, J.-F., VAZQUEZ, A., STELZL, U., LEMMENS, I., HIROZANE-KISHIKAWA, T., HAO, T., ZENKNER, M., XIN, X. & GOH, K.-I. 2009. An empirical framework for binary interactome mapping. *Nature methods*, 6**,** 83-90.

VIDAL, M., CUSICK, M. E. & BARABÁSI, A.-L. 2011. Interactome networks and human disease. *cell*, 144**,** 986-998.

VLASBLOM, J. & WODAK, S. J. 2009. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC bioinformatics*, 10**,** 1-14.

WANG, J., LIU, B., LI, M. & PAN, Y. 2010. Identifying protein complexes from interaction networks based on clique percolation and distance restriction. *BMC genomics*, 11.

WANG, J., PENG, X., LI, M., LUO, Y. & PAN, Y. Active protein interaction network and its application on protein complex detection. 2011 IEEE International Conference on Bioinformatics and Biomedicine, 2011. IEEE, 37-42.

WANG, J., PENG, X., LI, M. & PAN, Y. 2013. Construction and application of dynamic protein interaction network based on time course gene expression data. *Proteomics*, 13**,** 301-312.

WANG, J., PENG, X., PENG, W. & WU, F. X. 2014. Dynamic protein interaction network construction and applications. *Proteomics*, 14**,** 338-352.

WANG, J. Z., DU, Z., PAYATTAKOOL, R., YU, P. S. & CHEN, C.-F. 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23**,** 1274-1281.

WANG, R., LIU, G., WANG, C., SU, L. & SUN, L. 2018a. Predicting overlapping protein complexes based on core-attachment and a local modularity structure. *BMC bioinformatics*, 19**,** 1-15.

WANG, R., LIU, G., WANG, C., SU, L. & SUN, L. 2018b. Predicting overlapping protein complexes based on core-attachment and a local modularity structure. *BMC bioinformatics*, 19.

WATTS, D. J. 2004. *Six degrees: The science of a connected age*, WW Norton & Company.

WU, M., LI, X., KWOH, C.-K. & NG, S.-K. 2009. A core-attachment based method to detect protein complexes in PPI networks. *BMC bioinformatics*, 10**,** 169.

WUCHTY, S. 2014. Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences*, 111**,** 7156-7160.

XENARIOS, I., SALWINSKI, L., DUAN, X. J., HIGNEY, P., KIM, S.-M. & EISENBERG, D. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30**,** 303-305.

XIAO, Q., LUO, P., LI, M., WANG, J. & WU, F. X. 2019. A Novel Core-Attachment–Based Method to Identify Dynamic Protein Complexes Based on Gene Expression Profiles and PPI Networks. *Proteomics*, 19**,** 1800129.

XU, J. & LI, Y. 2006. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics,* 22**,** 2800-2805.

YONG, C. H., LIU, G., CHUA, H. N. & WONG, L. Supervised maximum-likelihood weighting of composite protein networks for complex prediction. BMC systems biology, 2012. Springer, S13.

YONG, C. H. & WONG, L. 2015. From the static interactome to dynamic protein complexes: Three challenges. *Journal of bioinformatics and computational biology,* 13**,** 1571001.

YU, H., KIM, P. M., SPRECHER, E., TRIFONOV, V. & GERSTEIN, M. 2007. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology,* 3**,** e59.

ZAKI, N., EFIMOV, D. & BERENGUERES, J. 2013. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC bioinformatics,* 14.

ZENGYOU, H. 2015. *Data mining for bioinformatics applications,* Woodhead Publishing.

ZHANG, J., ZHONG, C., LIN, H. X. & WANG, M. 2019. Identifying protein complexes from dynamic temporal interval protein-protein interaction networks. *BioMed research international,* 2019.

ZHANG, R. & LIN, Y. 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research,* 37**,** D455-D458.

ZHANG, Y., LIN, H., YANG, Z. & WANG, J. 2016. Construction of dynamic probabilistic protein interaction networks for protein complex identification. *BMC bioinformatics,* 17**,** 1-13.

ZHAO, J. & LEI, X. 2019. Detecting overlapping protein complexes in weighted PPI network based on overlay network chain in quotient space. *BMC bioinformatics,* 20.

ZHAO, J., LEI, X. & WU, F.-X. 2017. Predicting protein complexes in weighted dynamic PPI networks based on ICSC. *Complexity,* 2017.

ZHONG, J., TANG, C., PENG, W., XIE, M., SUN, Y., TANG, Q., XIAO, Q. & YANG, J. 2021. A novel essential protein identification method based on PPI networks and gene expression data. *BMC bioinformatics,* 22**,** 1-21.

ZOHAIRY, A. M. A. 2014. *Introduction to Bioinformatics,* academic library.

الخلاصة

تعتبر تعقيدات البروتين هي المفتاح الاساسي لفهم آلية وتنظيم عمليات الخلايا لأنها تشارك في الوظائف الاساسية للخلية. حيث تتكون من تفاعل بروتينين او اكثر في شبكة تفاعل البروتين(PIN) .

تقترح هذه الرسالة أربع مساهمات للتنبؤ بتعقيدات البروتين من شبكة تفاعل البروتينات الثابتة والديناميكية ، اثنان لكل منهما. باستخدام نموذج التوسع من بذرة والهيكل التبولوجي لشبكة تفاعل البروتينات الثابتة (SPIN) ، تم اقتراح خوارزمية SETS للتنبؤ بتعقيدات البروتين المتداخلة بكثافات متفاوتة خلال وقت تنفيذ مقبول. تم اقتراح خوارزمية أخرى تعتمد على التعبير الجيني (GE) لوزن شبكة تفاعل البروتينات دون حذف البروتينات التي لا تقدم بيانات التعبير الجيني. تحدد GECA البروتينات الأساسية التي تستخدم تقنيات الجوار الشائعة والمعلومات البيولوجية. علاوة على ذلك ، تعمل GECA على تحسين تقنية attachment عن طريق إضافة البروتينات ذات التفاعل القليل مع بروتينات core ولكنها تتشابه بشكل كبير بالتعبير الجيني.

ابتكر الباحثون شبكة تفاعل البروتينات الديناميكيية (DPIN) من خلال الجمع بين SPIN وبيانات التعبير الجيني. يعد تحديد عتبة مناسبة للجين النشط أحد التحديات البيولوجية. تم اقتراح طريقة (q-one)لتحديد النقاط الزمنية النشطة لكل بروتين وفقًا لخصائص قيمة تعبيره. الطريقة المقترحة يضاف فيها البروتين إلى الشبكة إذا كان نشطًا في نقطتين زمنيتين متتاليتين, هذا بالإضافة إلى قدرتها على الكشف عن التطور الديناميكي داخل شبكة PPI وتحديد البروتينات الأساسية. من خلال استخدام طريقة q-one لتحديد النقاط الزمنية النشطة لكل بروتين ، تم اقتراح خوارزمية جديدة (shared-one-time) للتنبؤ بتعقيدات البروتين من الشبكات الديناميكية التي تشارك بروتيناتها نقطة زمنية نشطة واحدة على الأقل.

باستخدام أحد عشر من بيانات الخميرة والانسان ، و باستخدام مقاييس التقييم (مقياس F-measure ، المطابقة الدقيقة والجيدة مع التعقيدات الحقيقية) وتقنية المقارنة مع الخوارزميات المعروفة والقوية لتقييم الخوارزميات المقترحة.

باستخدام البيانات المختلفة ، يكون أداء الطرق المقترحة افضل على الاقل بنسبة 10٪ وبنسبة قد تصل إلى 65٪ مقارنة باداء الخوارزميات الأخرى. من ناحية أخرى ، فإن معدل تحسين التعقيدات المتوقعة التي تتطابق تمامًا مع التعقيدات الحقيقية ، أعلى بنسبة 15 ٪ على الأقل من الأساليب السابقة ، مما يشير إلى أهمية بيولوجية كبيرة للتعقيدات المتوقعة بواسطة الخوارزميات المقترحة.

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل

كلية تكنلوجيا المعلومات - قسم البرمجيات

التنبؤ بتعقيدات البروتينات من الشبكات الثابتة والديناميكية باستخدام خوارزميات التنقيب الرسومية

اطروحة مقدمة إلى
مجلس كلية تكنولوجيا المعلومات -جامعة بابل كجزء من متطلبات
نيل درجة الدكتوراه في تكنولوجيا المعلومات - برمجيات

من قبل

سهير نوري علوان محمد

بإشـراف

أ.د. نبيل هاشم كاغد الاعرجي
أ.د. إيمان صالح صكبان الشمري

1443 هـ

2022 م