9-2015

Mining Biological Networks towards Protein complex Detection and Gene-Disease Association

Eileen Marie Hanna

United Arab Emirates University

College of Information Technology

MINING BIOLOGICAL NETWORKS TOWARDS PROTEIN-COMPLEX DETECTION AND GENE-DISEASE ASSOCIATION

Eileen Marie Hanna

This dissertation is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Under the Supervision of Dr. Nazar Zaki

September 2015

Declaration of Original Work

I, Eileen Marie Hanna, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this dissertation entitled "*Mining Biological Networks towards Protein-Complex Detection and Gene-Disease Association*", hereby, solemnly declare that this dissertation is my own original research work that has been done and prepared by me under the supervision of Dr. Nazar Zaki, in the College of Information Technology at UAEU. This work has not previously been presented or published, or formed the basis for the award of any academic degree, diploma or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my dissertation have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation and/or publication of this dissertation.

Student's Signature_____ Date_____

Approval of the Doctorate Dissertation

This Doctorate Dissertation is approved by the following Examining Committee Members:

1) Advisor (Committee Chair): Dr. Nazar Zaki
   Title: Associate Professor
   Department of Intelligent Systems
   College of Information Technology

   Signature _____ Date _____

2) Member: Prof. Amr Amin
   Title: Professor
   Department of Biology
   College of Science

   Signature _____ Date _____

3) Member: Dr. Elarbi Badidi
   Title: Associate Professor
   Department of Intelligent Systems
   College of Information Technology

   Signature _____ Date _____

4) Member (External Examiner): Prof. Periklis Andritsos
   Title: Professor
   Department of Information Systems
   Institution: University of Lausanne, Switzerland

   Signature _____ Date _____

This Doctorate Dissertation is accepted by:

Dean of the College of Information Technology: Professor Omar El Gayar

Signature _____ Date _____

Dean of the College of the Graduate Studies: Professor Nagi T. Wakim

Signature _____ Date _____

Copy ____ of ____

Advisory Committee

1) Advisor: Dr. Nazar Zaki

Title: Associate Professor

Department of Intelligent Systems

College of Information Technology

2) Member: Dr. Salah Bouktif

Title: Associate Professor

Department of Software Development

College of Information Technology

3) Member: Prof. Amr Amin

Title: Professor

Department of Biology

College of Science

Abstract

Large amounts of biological data are continuously generated nowadays, thanks to the advancements of high-throughput experimental techniques. Mining valuable knowledge from such data still motivates the design of suitable computational methods, to complement the experimental work which is often bound by considerable time and cost requirements. Protein complexes, or groups of interacting proteins, are key players in most cellular events. The identification of complexes not only allows to better understand normal biological processes but also to uncover disease-triggering malfunctions. Ultimately, findings in this research branch can highly enhance the design of effective medical treatments. The aim of this research is to detect protein complexes in protein-protein interaction networks and to associate the detected entities to diseases. The work is divided into three main objectives: first, develop a suitable method for the identification of protein complexes in static interaction networks; second, model the dynamic aspect of protein interaction networks and detect complexes accordingly; and third, design a learning model to link proteins, and subsequently protein complexes, to diseases. In response to these objectives, we present, ProRank+, a novel complex-detection approach based on a ranking algorithm and a merging procedure. Then, we introduce DyCluster, which uses gene expression data, to model the dynamics of the interaction networks, and we adapt the detection algorithm accordingly. Finally, we integrate network topology attributes and several biological features of proteins to form a classification model for gene-disease association. The reliability of the proposed methods is supported by various experimental studies conducted to compare them with existing approaches. ProRank+ detects more protein complexes than other state-of-the-art methods. DyCluster goes a step further and achieves a better performance than similar techniques. Then, our learning model shows that combining topological and biological features can greatly enhance the gene-disease association process. Finally, we present a comprehensive case study of breast cancer in which we pinpoint disease genes using our learning model; subsequently, we detect favorable groupings of those genes in a protein interaction network using the ProRank+ algorithm.

Keywords: Protein-protein interactions, protein complex, gene expression, topological features, biological features, gene-disease association.

Title and Abstract (in Arabic)

دراسة الشبكات البيولوجية بهدف الكشف عن المركبات البروتينية وربط الجينات بالأمراض

الملخص

يتم توليد كميات كبيرة من البيانات البيولوجية في الوقت الحاضر وذلك بفضل تقدّم التقنيات التجريبية ذات الإنتاجية العالية. ما زال استخراج معلومات قيّمة من هذه البيانات يحفز تصميم طرق حاسوبية مناسبة، لاستكمال العمل التجريبي المرتهن غالباً بوقت طويل ومتطلبات الكلفة. المركبات البروتينية، أو مجموعات البروتينات المتفاعلة هي لاعب أساسيّ في معظم الأحداث الخلوية. لا يسمح تحديد المركبات بفهم العمليات البيولوجية العادية بشكل أفضل فحسب بل أيضاً بالكشف عن الاختلالات المسببة للأمراض. في النهاية، تستطيع نتائج فرع البحث هذا تحسين تصميم العلاجات الطبية الفعالة إلى حد كبير. هدف هذا البحث الكشف عن المركبات البروتينية في شبكات تفاعل البروتينات مع بعضها البعض وربط الكيانات المكتشفة بالأمراض. إن العمل مقسّم إلى 3 أهداف رئيسية: أولاً، تطوير طريقة مناسبة لتحديد المركبات البروتينية في شبكات تفاعل ثابتة. ثانياً، صياغة الجانب الديناميكي لشبكات تفاعل البروتينات واستبيان المركبات وفقاً لذلك. ثالثاً، تصميم نموذج تعلّم حاسوبي لربط البروتينات وبالتالي المركبات البروتينية، بالأمراض. رداً على هذه الأهداف، نقدّم برورانك+، وهو أسلوب جديد في تحديد المركبات مبنيّ على خوارزمية لتصنيف المركبات وترتيبها ودمجها. ثم ندخل دايكلستر، الذي يستخدم بيانات التعبير الجيني لصياغة ديناميات شبكات التفاعل ونكيّف خوارزمية الكشف وفقاً لذلك. أخيراً، ندمج الصفات الطوبولوجية لشبكة البروتينات في سمات البروتينات البيولوجية لتشكيل نموذج تصنيف لربط الجينات بالأمراض. مصداقية الطرق المقترحة مدعومة بدراسات تجريبية متنوعة أجريت لمقارنتها مع أساليب قائمة. يكشف برورانك+ عن مركبات بروتينية أكثر من أيّ أساليب متطورة أخرى. يذهب دايكلستر خطوة أبعد ويحقق أداءً أفضل من أداء تقنيات مشابهة. ثم يظهر نموذج التعلّم الخاص بنا أن الجمع بين السمات الطوبولوجية والسمات البيولوجية يستطيع أن يحسّن إلى حد كبير عملية ربط الجينات بالأمراض. أخيراً، نقدّم دراسة شاملة لحالة سرطان الثدي نحدد فيها جينات المرض مستخدمين نموذج التعلّم الخاص بنا. من ثم

نحدد التجمعات المناسبة لتلك الجينات في شبكة تفاعل البروتينات مستخدمين خوارزومية برورانك+.

مفاهيم البحث الرئيسية: تفاعل البروتينات مع بعضها البعض، المركبات البروتينية، التعبير الجيني، السمات الطوبولوجية، السمات البيولوجية، ربط الجينات بالأمراض.

Acknowledgements

I am grateful to the administration of the United Arab Emirates University and the Faculty of Information Technology for the support and assistance presented throughout my PhD studies and research.

I would like to express my sincere gratitude to my advisor, Dr. Nazar Zaki, for setting the bar high from the beginning. Dr. Zaki's valuable guidance, encouragement and support were essential during the development of this work. I would also like to thank the members of my research committee for their discussions and recommendations. Special thanks to Prof. Amr Amin for providing the biological data and helping in the case studies.

My deepest appreciations go to my big family for the faithful love and care in all my pursuits. I am indebted to your unconditional support provided during my studies. Finally, I also dedicate this work to the memory of those who inspired it and could not read it.

Dedication

To my precious Tony and Aya-Sofia

To my esteemed parents and sister, Salim, Marie and Tamara

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Scope

From metabolism, signal transduction, transport, cellular organization to most biological processes, proteins are the key players. Their interconnections shape interaction networks which define highly-organized cellular systems (1000 Genomes Project Consortium, 2010). The association of a gene or a complex to a certain biological function broadens our perception of how this function occurs and it consequently allows us to uncover the malfunctions that trigger various diseases. For instance, in terms of a normal phenomenon, happiness can be psychologically defined as "the experience of joy, contentment, or positive well-being, combined with a sense that one's life is good, meaningful, and worthwhile" (Lyubomirsky, 2008). Philosophically, according to Aristotle, "Happiness depends upon ourselves". Genetically, legitimate questions are asked: How far does happiness really depend on "ourselves"? Do genes, the shaping elements of "ourselves", contribute to our happiness? In fact, genetics are linked to individual life satisfaction and particularly to happiness: long and more efficient alleles of the serotonin transporter gene 5-HTTLPR and self-reported life satisfaction are positively associated (De Neve, Christakis, Fowler, & Frey, 2012). In view of that, a recent study (Oswald & Proto, 2013) measures the genetic distance among countries' populations and finds that it is highly correlated with international well-being differences. On the other hand, in terms of identifying disease-related genes, in ancient history, cancers were blamed on the gods (The History of Cancer, 2015). Then, through the middle ages, it was associated to imbalances in the body. Various theories were proposed later ranging from the lymph theory in the 1700s to the trauma theory and the infectious disease

theory in the 17th and the 18th centuries. Accumulated knowledge in genetics throughout the following years allowed more understanding of the disease. In 2014, more than 100 chemical, physical and biological substances were associated by the World Health Organization to cancer (World Health Organization IARC, 2014). In the same year, breast cancer was listed as the highest occurring cancer type in women worldwide (World Health Organization, 2014). The earlier association of the BRCA1 gene to breast cancer (Miki et al., 1994) not only accelerated the design of more efficacious treatments but also allowed the discovery of many key genes and complexes in other cancer types and complex diseases.

Looking at the bigger picture, every genetic finding can be viewed as a puzzle piece that contributes to our comprehension of various molecular functions as well as different disorders. Ultimately, the more we know, the more we are able to improve medical treatments.

1.2 Background

The deoxyribonucleic acid (DNA) is considered the cell's "master molecule" thanks to its essential functional properties (Lodish et al., 2013). It has a double-helix structure composed by two helical strands coiled around a common axis (Watson & Crick, Molecular structure of nucleic acids, 1953). This structure allows transferring genetic characteristics among successive generations and is thus crucial to heredity. The DNA strands consist of four types of nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). They are arranged in such a way that A on one strand is matched with T on the other strand and likewise, C is matched with G. The linear order of nucleotides along each strand defines the genetic information carried by DNA. The informative segments of DNA are divided into functional units called

genes which typically consist of 5,000 to 100,000 nucleotides. Many genes are responsible for making proteins, the primary molecules defining cellular structure and activities. The conversion of DNA into proteins is divided into two processes as presented in Figure 1. The first process is transcription by which the coding portion of a gene is copied into a single-stranded ribonucleic acid (RNA) version of DNA. A large enzyme called RNA polymerase uses DNA as a template and catalyzes the linkage of nucleotides into RNA chain. In eukaryotic cells, RNA is transformed into a smaller messenger RNA (mRNA) molecule which moves to the cytoplasm region of the cell. This is where the second process, known as translation, takes place. A very complex molecular machine called ribosome and composed of both RNA and

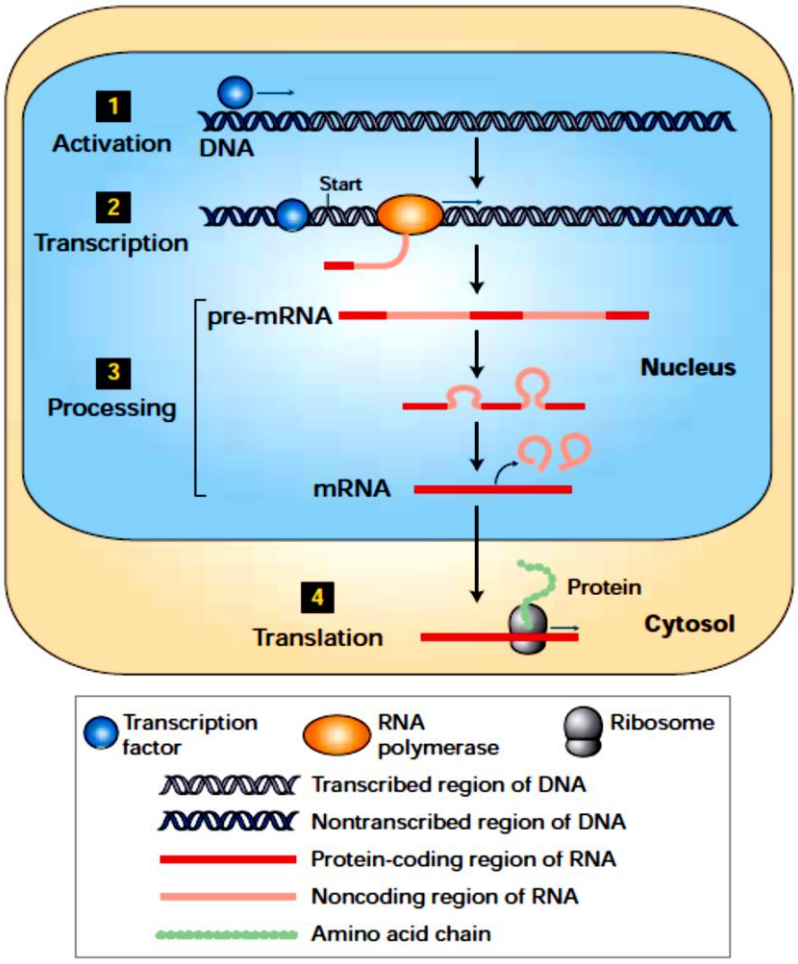

Figure 1: The multi-process conversion of DNA into protein (Lodish et al., 2013).

protein comes into play. The ribosome assembles amino acids to form proteins exactly as inferred by the mRNA sequence. Proteins consist of linear chains in which 20 different amino acids can be combined. Once a chain is created, it folds to form a three-dimensional structure which determines its distinctive function, as shown in Figure 2. The linear amino acid sequence of a protein (primary structure) folds into helices (secondary structure) that pack into globular domains (tertiary structure). Some proteins self-associate into complexes (quaternary structure) which comprise tens to hundreds subunits (supramolecular assemblies). Proteins can exhibit various functions including regulation, structure, movement, catalysis, transport and signaling. All those functions are subject to proper protein folding. The types and amounts of mRNA molecules existing in a cell determine its function. Accordingly, the course of protein formation through transcription and translation critically defines the functions of cells (O'Connor, Adams, & Fairman, 2010). The regulation of those

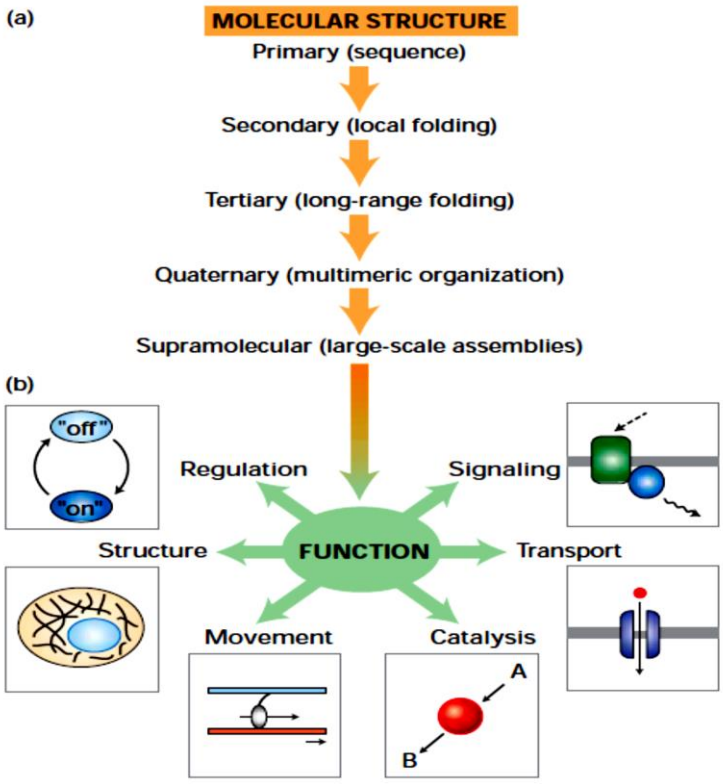

Figure 2: Protein structure and function (Lodish et al., 2013).

processes allows cells to respond to environmental variations. In the same context, we note here that genes which exhibit similar expression patterns across various environmental conditions most likely interact (Baldi & Hatfield, 2002).

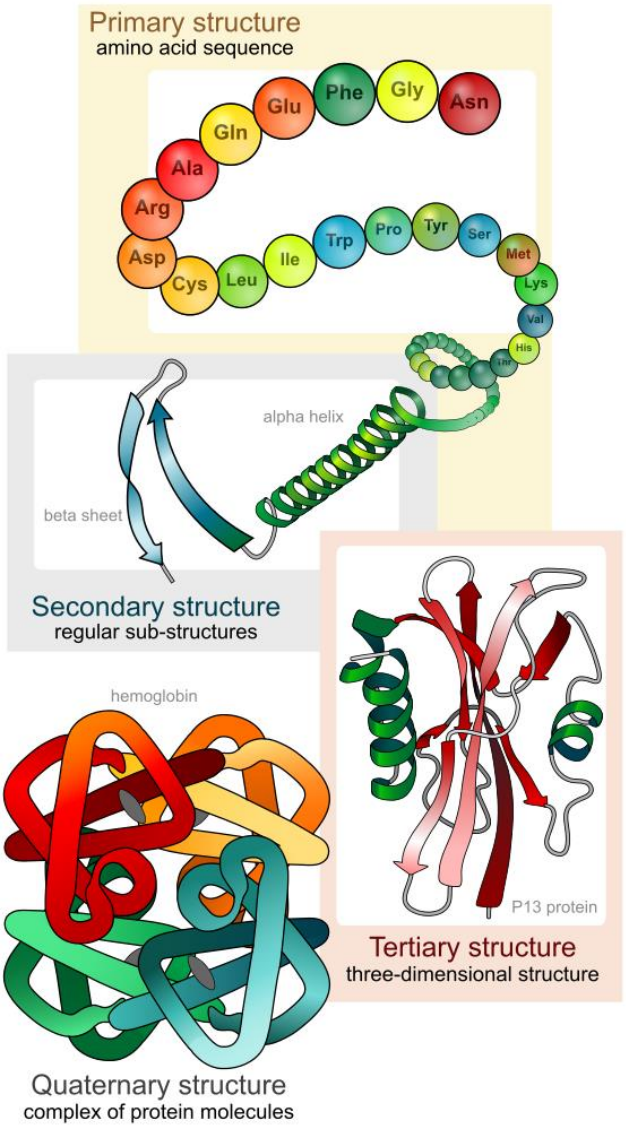

Figure 3: The structural levels of proteins (Wikimedia Commons, 2008).

Figure 3 visualizes the four structure levels of a protein, from amino acid sequence to protein complex. Mutations are errors that occur during DNA replication. It alters the nucleotide sequence by changing its order, deleting or inserting an element, or even inverting it. This can cause the abnormal generation of proteins and can lead to inherited diseases if not controlled properly. For example, the sickle cell disease is caused by a single mutation of the hemoglobin gene by which the 17th nucleotide is changed from T to A (Rees, Williams, & Gladwin, 2010). In view of that, identifying disease-causing mutations and subsequently, disease-related genes, protein and protein complexes is indeed a crucial task towards understanding various disorders and finding suitable ways to possibly avoid or treat them.

1.3 Motivation and Problem Statement

Biological functions are often acquired through collaborations of interacting protein groups referred to as protein complexes (Gavin et al., 2006). High-throughput experimental techniques designed to study protein complexes, such as yeast two-hybrid (Y2H) (Fields & Song, 1989) and tandem affinity purification (TAP-MS) (Collins & Choudhary, 2008) approaches, are generally time-consuming and costly. Moreover, they are susceptible to high error rates (Marto, 2009). In view of that, various computational methods are developed to complement and reduce the efforts required for biological explorations. Ideally, looking at protein-protein interaction (PPI) data, a reliable computational approach can identify proteins and subsequently protein complexes possibly engaged in certain functions or phenotypes, for further experimental examinations. It is believed that the more enrichment with biological information is added to interaction networks and complex-detection algorithms, the

better is the overall quality of the results. In a computational context, a PPI dataset is usually modeled as a single graph in which vertices and edges represent the proteins and their interconnections, respectively. An example of a protein complex is shown in Figure 4 in terms of graph and structural representations based on the work by (Newman, Brändén, & Jones, 1993) and visualized at the Protein Data Bank (Berman et al., 2000).

Given a PPI dataset, the goal is to develop a suitable computational approach that can identify the corresponding protein complexes and subsequently associate the detected entities to diseases. In this direction, several challenges need to be addressed. First, experimentally-generated datasets are usually large. For instance, in the case of human PPIs, the June 2015 release of the BioGRID repository (Stark et al., 2006) contains 186744 non-redundant interactions among 19415 unique proteins. As a result, scalable and efficient methods are required for their analysis. In addition, PPI data may contain false positive (spuriously-detected) and false negative

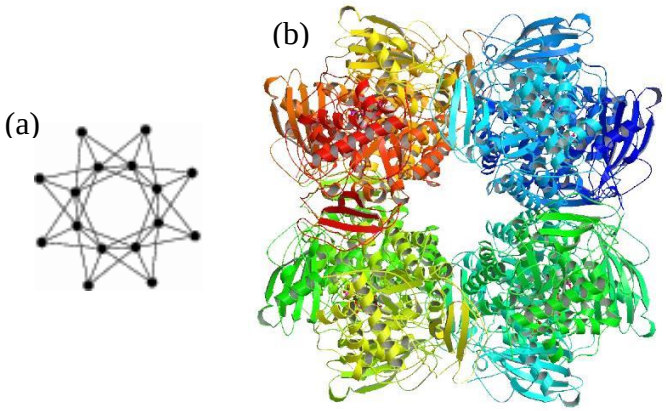

Figure 4: An example of a protein complex. (a) A graph representation in which nodes and edges represent protein and their interactions, respectively. (b) The biological assembly of the complex.

(missing) interactions. Consequently, suitable data cleaning techniques have to be applied prior to analysis in order to ensure the reliability of the results. Moreover, protein interactions do not occur simultaneously. They are subject to temporal, spatial and contextual conditions (Macropol, Can, & Singh, 2009). Accordingly, the comprehensive network representation of a PPI dataset ought to take the dynamic nature of interactions into consideration. These main challenges, among others, constitute the focus points based on which our methodology is designed and developed.

1.4 Research Objectives and Proposed Solutions

The goal of this dissertation is to develop a suitable approach for the identification and association of proteins and protein complexes to diseases. In this direction, the work is divided into three main objectives:

1- Detecting protein complexes in PPI networks.

2- Modeling the dynamic aspect of PPI networks and detecting protein complexes accordingly.

3- Developing a learning model to classify genes as disease-related or not.

To begin, we introduce ProRank+ (Hanna & Zaki, 2014), a protein-complex detection method based on a ranking algorithm which sorts proteins according to their importance in the PPI network; and a merging procedure which refines the detected complexes in terms of their members. When compared to several state-of-the-art approaches, ProRank+ is able to detect more protein complexes with higher quality scores.

Since protein interaction networks are dynamic in nature (Levy & Pereira-Leal, 2008), our second objective is to model the dynamic aspect of PPI networks and to tune ProRank+ accordingly. Recent experimental tools, such as ChIP-chip (Kim & Ren, 2006) and ChIP-seq (Johnson, Mortazavi, Myers, & Wold, 2007), can provide temporal, spatial and contextual information across which PPIs occur. Consequently, advances in computational approaches developed to analyze PPI networks ought to relate to such diversity of information which is currently available. Gene expression datasets consist of quantitative measurements of genes in cellular compartments across different conditions (Lovén et al., 2012). Genome-wide expression levels can now be studied (Secrier & Schneider, 2013). Genes with correlated expressions across subsets of conditions most likely interact (Baldi & Hatfield, 2002). As a result, the integration of gene expression data with PPI information can potentially reveal the processes which underline the formation of protein complexes. In this direction, we present DyCluster, a framework to model the dynamic aspect of protein interaction networks by incorporating gene expression data, through biclustering techniques (Busygin, Prokopyev, & Pardalos, 2008), prior to applying complex-detection algorithms. The experimental studies, including biological applications, show that DyCluster leads to high numbers of correctly detected complexes with better evaluation scores.

The last objective consists of designing a suitable approach for gene-disease association. We propose a learning model which integrates PPI network topology features and biological information collected from various sources. A learning model, here classification model, given training data (data objects whose class label is known), consists of a set of functions that can describe and distinguish data classes. Such model can then be used to predict the class of objects whose class label

is unknown. Given a list of genes, the goal is to maximize the contrast between disease and non-disease classes. Accordingly, we study the topology of the corresponding PPI network to find distinctive positioning of genes in interaction networks. Then, we combine those features with biological data from various sources to uncover potential similarities which characterize each class. The experimental work strongly favors our approach.

1.5 Dissertation Outline

The dissertation is organized as follows. In chapter 2, we survey and discuss state-of-the-art methods related to our research objectives. Chapter 3 introduces our method for the detection of protein complexes in PPI networks. In chapter 4, we present our approach to model dynamic protein interaction networks and subsequently detect complexes. Chapter 5 includes our solution for gene-disease association. A comprehensive case study of the breast cancer disease is presented in chapter 6. Finally, we conclude the dissertation in chapter 7.

Chapter 2: Related Work

This chapter surveys the state-of-the-art approaches related to our research objectives. Computational methods developed for the detection of protein complexes in PPI networks are reviewed in section 2.1. Various ways to model the dynamic aspect of PPI networks and detect protein complexes accordingly are presented in section 2.2. Gene-disease association approaches are discussed in section 2.3. Lastly, the drawbacks of previous approaches that we seek to overcome in our work are summarized in section 2.4.

2.1 Detecting Protein Complexes in Protein-Protein Interaction Networks

In a computational setting, it is generally assumed that protein complexes correspond to dense subgraphs in PPI networks. We hereafter highlight state-of-the-art methods for the detection of complexes in protein interaction networks. The Markov Clustering algorithm (MCL) (Van Dongen, 2001) looks for cluster structures in protein interaction networks using random walks. The search is based on alternations between two main operators: expansion which is given by taking the power of a stochastic matrix using matrix squaring; and inflation which corresponds to taking the Hadamard power of a matrix, i.e. entry-wise, followed by a scaling step, to generate a stochastic matrix. The algorithm deterministically calculates the probabilities of random walks in the network and transforms one set of probabilities into another based on the expansion and inflation operators. The Molecular Complex Detection (MCODE) algorithm (Bader & Hogue, 2003) identifies complexes as dense regions grown from highly-weighted vertices. A vertex is weighted by checking the highest *k*-core in its neighbourhood, i.e. the central most densely

connected subgraph of minimal degree *k*. MCODE then seeds complexes with proteins, considered by their decreasing weights, and include only vertices of weights above a given threshold, at each time. The clustering based on maximal cliques (CMC) method (Liu, Wong, & Chua, 2009) starts by using an iterative scoring scheme to assign weights to protein interactions. Lower scores correspond to less reliable interactions in order to reduce their impact in the detection process. All maximal cliques are generated from the PPI network and then, they are ranked based on their weighted density. Finally, highly overlapping cliques are merged or removed. The Affinity Propagation (AP) algorithm (Frey & Dueck, 2007) uses a distance matrix to propagate messages between nodes until a high-quality set of "exemplars" and corresponding clusters are gradually generated. ClusterONE (Nepusz, Yu, & Paccanaro, 2012) identifies protein complexes through clustering with overlapping neighborhood expansion. A cohesiveness measure is introduced to reflect the notion by which a protein complex is viewed as an entity that is well-separated from the rest of the network and whose members have reliable interconnections. Proteins are considered by their descending order of degrees and a greedy algorithm is applied to generate complexes by joining proteins which are not yet added to any complex. Next, groupings with high overlaps are merged and complexes with less than three proteins or with low density are discarded. The Restricted Neighborhood Search Clustering (RNSC) algorithm, presented in (King, Pržulj, & Jurisica, 2004) and (Pržulj, Wigle, & Jurisica, 2004), is a cost-based local search algorithm that uses the tabu metaheuristic. It seeks to partition proteins into highly-interconnected subsets. The method starts with a random clustering and then, moves nodes from one group to another to improve clustering cost. The RRW algorithm (Macropol, Can, & Singh, 2009) exploits the global structure of a PPI

network using repeated random walks. It moves from one node to another based on the probabilities of the connective edges. CFinder (Adamcsek et al., 2006) finds overlapping and fully-connected complexes based on the clique percolation method. The GIBA tool (Moschopoulos, Pavlopoulos, Schneider, Likothanassis, & Kossida, 2009) clusters the whole network and then, filters the generated clusters in order to only keep the important ones.

Although these state-of-the-art methods offer good solutions to the considered problem, most of them are bound by the assumption that protein complexes only correspond to dense subgraphs in protein interaction networks. As a result, they cannot identify complexes with few members and/or few interactions. That is an important drawback to overcome since for instance, among the 313 protein complexes included in the MIPS catalogue (Mewes et al., 2006), 104 complexes consist of 2 or 3 proteins (approximately 33%). ProRank, introduced in (Zaki, Berengueres, & Efimov, 2012a) and (Zaki, Berengueres, & Efimov, 2012b) is a recent complex-detection method which is not restrained by this density supposition. It is mainly based on a protein ranking algorithm inspired by Google's PageRank algorithm discussed in (Brin & Page, 1998), (Bryan & Leise, 2006), (Ishii & Tempo, 2010) and (Langville & Meyer, 2011). As PageRank sorts web pages according to their level of importance, ProRank applies the same analogy to rank proteins in PPI networks, and subsequently, to identify the "essential" ones which most-likely have central roles in cellular functions. Those proteins are the starting point based on which the detected complexes are formed. In addition, the pairwise similarities of the proteins are computed under the assumption that proteins belonging to the same complex share evolutionary relationships (Kuang, Weston, Noble, & Leslie, 2005). In view of the notable performance of ProRank when compared to previous

approaches, the algorithm is the keystone of our approach introduced hereafter. Five main steps delineate the ProRank algorithm (Zaki, Berengueres, & Efimov, 2012a):

1- Pruning: PPI datasets are usually noisy; they have high false-positive and false-negative rates (Reguly et al., 2006). Accordingly, the first step consists of removing unreliable interactions which could negatively affect the detection process. That is done using the AdjustCD method introduced in (Chua, Sung, & Wong, 2006) and (Chua, Ning, Sung, Leong, & Wong, 2008); a weighting scheme that iteratively calculates the reliability of protein interactions based on the topology of the network and then discards the interactions with scores less than a specified threshold.

2- Filtering: based on the protein interaction network, three types of noisy proteins are identified: bridge proteins which have a disconnected subgraph of neighbors; fjord proteins whose neighbors have a small number of interactions among each other; and shore proteins which have at least one neighbor with significantly few interactions with other proteins. Accordingly, the network vertices are examined for possible memberships in these types. Figure 5 illustrates examples of the three described categories.

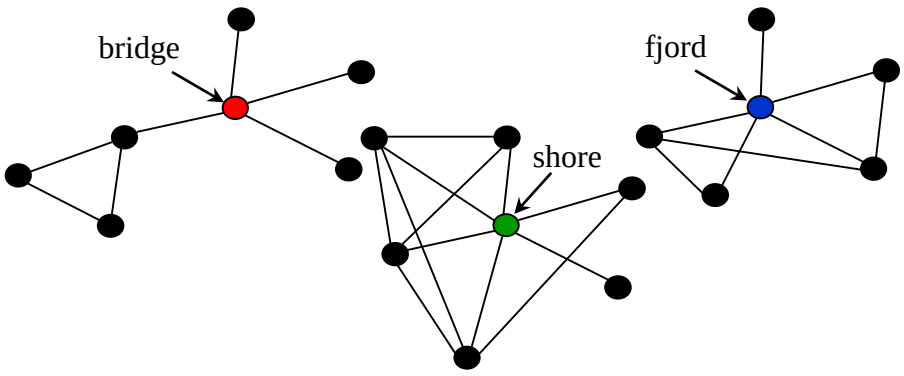

Figure 5: Examples of bridge, fjord and shore proteins in PPI networks.

3- Protein Similarity Calculating: proteins belonging to the same complex are expected to have evolutionary relationships (Kuang, Weston, Noble, & Leslie, 2005). Therefore, the similarity scores among all the proteins in the PPI network are calculated using pairwise alignment.

4- Protein Ranking: in analogy with the PageRank algorithm, a ranking algorithm is applied to order the proteins by their importance in the interaction network.

Given n interacting proteins, we represent their interaction network by a graph $G = (V, E)$ where $V = 1, 2, \dots, n$ is the set of nodes (proteins) and E the set of edges (interactions) among those proteins and $(i, j) \in E$ if protein i interacts with protein j. The goal of the ranking algorithm is to order proteins by their importance in the network. Accordingly, the importance measure x_i of protein i is a real number such that $x_i \in [0,1]$ and $x_i > x_j$ means that protein i is more important than protein j. The value of a protein is given by the sum of contributions of all proteins interacting with it. The importance of protein i is based on the following equation:

$$x_i = \sum_{j \in \tau_i} S^*_{i,j} \, . \, n_j \qquad (1)$$

where $\tau_i \coloneqq \{j : (j, i) \epsilon E\}$ is the index set of the proteins interacting with i, n_j is the number of outgoing links from node j and $S^*$is the normalized similarity matrix computed in the Protein Similarity Calculating step. The total of all values is normalized i.e. $\sum_{i=1}^{n} x_i = 1$. Let the values of x be in the vector form where $x \in [1,0]^n$. The PageRank algorithm can thus be rewritten as:

$$x = S^* x,\ x \in [1,0]^n \text{ and } \sum_{i=1}^{n} x_i = 1 \qquad (2)$$

Note that the vector x is a nonnegative eigenvector corresponding to the eigenvalue 1 of the nonnegative matrix S^*. Nonetheless, for this eigenvector to exist and to be unique, it is essential that the PPI network is strongly connected. To find the eigenvector corresponding to the eigenvalue 1, a modified version of the values is defined as follows. Let m be a parameter such that$m \in (0,1)$, and let the modified interaction matrix $M \in \Re^{n\times n}$ be given by:

$$M \coloneqq (1-m)S^* + \frac{m}{n} l \qquad (3)$$

where l is an $n \times n$ matrix with all entries equal to 1. A typical value of m is 0.15. M is a positive stochastic matrix. Thus, according to Perron theorem 33, this matrix is primitive. Particularly, $|\lambda| = 1$is the unique maximum eigenvalue. Therefore, we apply the following formula to find corresponding eigenvector x:

$$x(k+1) = Mx(k) = (1-m)S^* x(k) + \frac{m}{n} l \qquad (4)$$

where $x(k) \in \Re^{n\times 1}$ and the initial vector $x(0) \in \Re^{n\times 1}$ is a probability vector. Expanding on the convergence rate of this scheme, let $\lambda_1(M)$ and $\lambda_2(M)$ be the largest and the second largest eigenvalues of M in magnitude. Then, by the power method applied to M, the asymptotic rate of convergence is exponential and depends on the ratio $|\lambda_2(M)/\lambda_1(M)|$.Since M is a positive stochastic matrix, we have $\lambda_1(M) =$

1 and $\lambda_2(M) < 1$. Therefore, the structure of the link matrix M leads us to the bound:

$$|\lambda_2(M)| \leq 1 - m \quad (5)$$

5- Complex Detection: the essential proteins are the ones which do not belong to any of the categories defined in step 2. Using the spoke model, those proteins are considered by their decreasing ranking order and each of them is pulled from the interaction network along with its neighbors to form a protein complex. Note that each protein can belong to one complex only.

In addition to those five steps, ProRank discards complexes of less than three members and merges two complexes if more than 50% of the neighbors of each protein belonging to the first complex are in the second complex. Figure 6 shows the workflow of the ProRank algorithm.

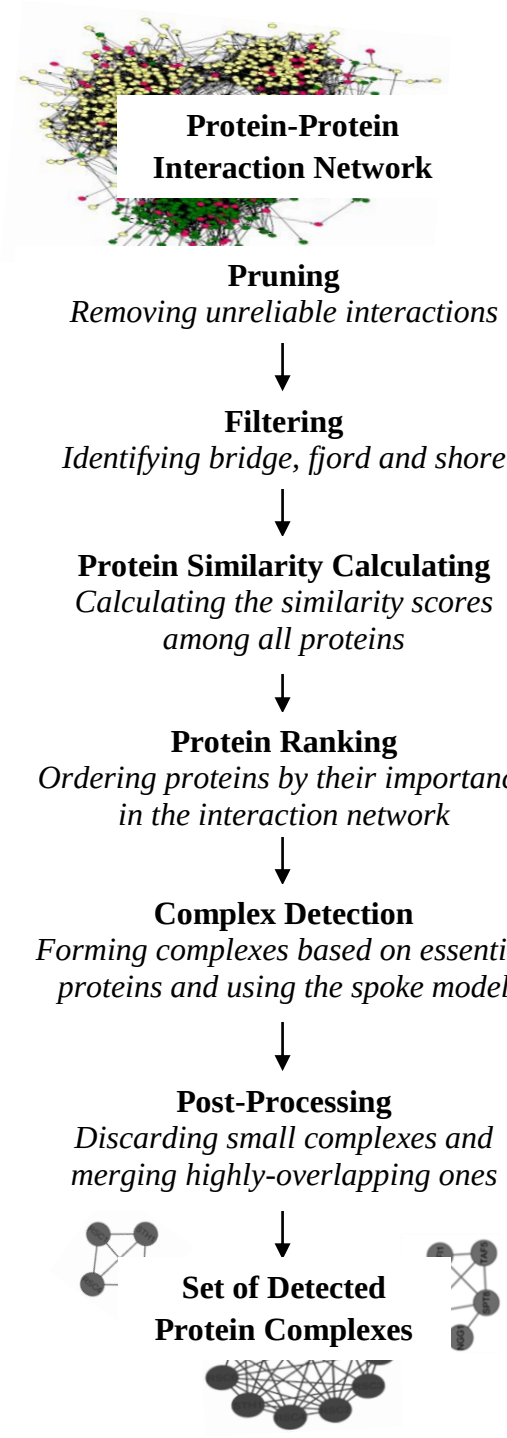

Figure 6: The workflow of the ProRank algorithm which takes as input a protein-protein interaction network and detects the corresponding protein complexes.

2.2 Detecting Protein Complexes in Dynamic Protein-Protein Interaction Networks

PPI networks are dynamic in nature (Levy & Pereira-Leal, 2008). In the direction of acquiring better complex-detection results, computational methods ought to profit from the abundance of biological information provided by advanced experimental techniques to model the dynamics of protein interactions. A single network is usually used to represent a PPI dataset. In contrast, a dynamic PPI network can be visualized by a series of schemes representing snapshots of the network states, corresponding to different stages and/or locations of molecular activities. We will hereafter highlight some of the potential concepts and approaches that can be used to model the dynamics of protein interaction networks.

Gene expression datasets present quantitative measurements of RNA species in cellular compartments across different conditions (Lovén et al., 2012). Genome-wide expression levels can now be generated (Secrier & Schneider, 2013). Time-series gene expression data report quantities of RNA species across various time points in cellular processes. Genes with correlated expressions across subsets of conditions most likely interact. When combined with PPI data to simulate the interaction dynamics, they can potentially reveal the processes which underline the formation of protein complexes. We note here that not all the genes described in one biological dataset may be covered in another dataset. However, combining data from multiple sources is viewed as advantageous since it helps overcome data limitations such as false-positive and false-negative interactions in PPI datasets and low gene coverage in gene expression datasets. For example, PPI and gene expression data combination is done in (Wang, Peng, Xiao, Li, & Pan, 2013) where it is shown that a

just-in-time mechanism elapsing through continuous time points delineates the formation of most protein complexes. The statistical 3-sigma principle is used in (Wang, Peng, Xiao, Li, & Pan, 2013) and (Wang, Peng, Li, Luo, & Pan, 2011) to define the active time points of proteins based on their gene expression levels and thus, to introduce approaches for the identification and refinement of protein complexes. The core-attachment composition of complexes is recently considered in (Li, Chen, Wang, Wu, & Pan, 2014). Relying on gene expression data, the identification of a protein complex is split into two main parts: a static core consisting of proteins expressed throughout the whole cell cycle and short-lived proteins that form a dynamic attachment. The results of these approaches are better than the ones deduced from static networks. Kim et al. (Kim, Han, Choi, & Hwang, 2014) highlight some of the computational methods used to infer dynamic networks from expression data, based on statistical dependence to categorize nodes and/or edges as active or not. These methods include: Bayesian networks (Friedman, Linial, Nachman, & Pe'er, 2000), relevance networks (Remondini et al., 2005), Markov Random Fields (Song, Kolar, & Xing, 2009), ordinary differential equations (Bansal, Belcastro, Ambesi-Impiombato, & Di Bernardo, 2007) and logic-based models (Morris, Saez-Rodriguez, Sorger, & Lauffenburger, 2010).

As it is conditioned by time, the occurrence of a protein interaction is also subject to the co-localization of its interacting partners in cellular components (Park et al., 2011). In fact, unsuccessful interactions caused by inappropriate protein localizations can be pathological. Consequently, subcellular localization annotations (de Lichtenberg, Jensen, Brunak, & Bork, 2005) can be also used to model dynamic PPI networks based on spatial constraints. Indeed, the formation of protein complexes is influenced by the localization settings of proteins as well. As a result, it

is certainly beneficial to incorporate the spatial dynamics in the direction of improving complex-detection approaches. Various methods aim at studying and collecting spatial movements of proteins (Lee, Tan, & Chung, 2010). However, in addition to mathematical modeling techniques, methods to appropriately integrate spatial protein dynamics in PPI networks are still required.

Gene ontology annotations (Ashburner et al., 2000) provide information about genes across different species. They can potentially infer the dynamic aspect of PPI networks (Xu, Lin, & Yang, 2010). As an indicator of interaction probability, various weighting schemes are introduced to assign PPI weights based on the similarity degrees of gene ontology terms between interacting partners. Among these approaches are: SWEMODE (Lubovac, Gamalielsson, & Olsson, 2006), which detects communities within PPI networks based on weighted clustering coefficient and weighted average nearest-neighbors degree measures; and OIIP (Xu, Lin, & Yang, 2010), which identifies protein complexes in PPI networks by assigning node and edge weights based on the size of gene annotations.

By modeling the dynamics of PPI networks, we can potentially: reproduce the mechanisms of protein-complex formation; uncover new biological facts about complexes; overcome limitations existing in most experimental datasets; categorize modules deduced from PPI networks; and finally, increase the accuracy and reliability of the detected results.

2.3 Associating Genes to Diseases

The identification of the genes and the inter-molecular events leading to the formation of diseases remains an essential research area towards the development of effective medical treatments. Based on the assumption that genes related to similar

disorders tend to be functionally associated (Oti & Brunner, 2007) (Wu, Jiang, Zhang, & Li, 2008), existing methods often follow a guilt-by-association (Altshuler, Daly, & Kruglyak, 2000) conjecture by which genes are ranked by their similarity to known disease genes. We hereafter list various existing approached for gene-disease association.

Deng et al. (Deng, Chen, & Sun, 2004) combine physical and genetic interactions of proteins with gene expression networks, protein complex data and domain structures to form an integrated probabilistic model to predict protein functions. They apply the Markovian random field theory. Xu and Li (Xu & Li, 2006) classify genes as disease-related or not based on PPI network topology features, using the k-nearest neighbor algorithm. Ma et al. (Ma, Lee, Wang, & Sun, 2007) apply a method based on Markov Random Field (MRF) on a high-throughput dataset comprising gene expression profiles and protein interaction data. They seek to prioritize disease genes without requiring known candidate genes. Lage et al. (Lage et al., 2007) consider that mutations in members of protein complexes lead to comparable phenotypes. Accordingly, they build a phenome-interactome network by integrating phenotypic data and phenotypic similarities with a high-confidence human protein interaction network. They use a Bayesian classifier to potentially link previously-unknown protein complexes to diseases. Köhler et al. (Köhler, Bauer, Horn, & Robinson, 2008) apply a random walk with restart algorithm (RWR) on a heterogeneous interaction network to prioritize candidate disease genes. They consider that global network-similarity measures reflect the relationships among disease genes better than direct or shortest-paths algorithms. Li and Agarwal (Li & Agarwal, 2009) use pathway data to answer the gene prioritization problem. They examine disease relationships via literature mining to identify disease genes. That is

done by associating diseases to biological pathways where those genes are enriched and linked to diseases based on their shared pathways. Zhang et al. (Zhang et al., 2012) construct a combined classifier on multiple PPI network topology features to identify disease genes. Guan et al. (Guan et al., 2012) create tissue-specific functional networks to prioritize disease genes. Their approach is based on the notion by which tissue-specificity is viewed as an essential factor that highlights the diversity of protein roles in different cell lineages. In other words, forming tissue-specific functional networks can potentially lead to more accurate gene-phenotype associations. Li et al. (Li et al., 2014) introduce novel topological attributes and use support vector machines (SVM) to classify genes as disease-related or not. Chen et al. (Chen B. , Wang, Li, & Wu, 2014) introduce a method based on the Markovian random field theory and Bayesian analysis for gene-disease association. They combine biological data from multiple sources in order to prioritize disease genes.

Although most of the existing approaches perform well, their limitations mainly reside in requiring initial settings of parameters and thresholds in addition to the dependence on a single set of gene features, either topological or biological.

2.4 Summary

The literature offers various solutions to the research problem and objectives that we address in our work. Nevertheless, the research area remains open thanks to the continuously-growing biological knowledge provided by advanced experimental techniques. In view of that, we seek to overcome the limitations of the existing approaches while developing algorithms that are also enriched by the available biological information. Accordingly, the proposed solution for the first objective is not bound by the assumption that protein complexes only correspond to dense

subgraphs in PPI networks. Complexes may also overlap. In addition, post-processing steps are introduced to examine and refine the detected entities based on their overlapping protein members. In the second objective, we model the dynamic aspect of protein interaction networks by incorporating time-series gene expression data. The expressions are analyzed using biclustering techniques (Busygin, Prokopyev, & Pardalos, 2008) which allow the identification of subsets of co-regulated genes across subsets of samples. And in analogy to biological facts, by using these techniques, a gene can belong to multiple clusters or may not fit in any cluster as well. Finally, we present a classification model for the gene-disease association problem which is built based on integrated PPI network topology attributes and various biological features collected from multiple sources. We believe that by combining computationally-conveyed network analysis and experimentally-generated biological information, the gene-disease association process can be greatly enhanced.

Chapter 3: Detecting Protein Complexes in Protein-Protein Interaction Networks

In this chapter, we present our approach for the detection of protein complexes in PPI networks. Section 3.1 revisits some background information of the research objective. In section 3.2, our ProRank+ method is introduced. The performance of ProRank+ is tested and compared to the performance of existing state-of-the-art approaches in section 3.3. The chapter conclusion is in section 3.4.

3.1 Background

The importance of this objective originates from the fact that protein complexes are key players in most cellular processes (Gavin et al., 2006). Designing suitable methods for the detection of protein complexes in protein interaction networks continues to be an intriguing area of research. The more complexes we identify, the better we can perceive normal as well as abnormal molecular events. Given a set of proteins that participate in a process under study and based on the interconnections that they exhibit, biologists use advanced experimental techniques to identify the corresponding protein complexes. Nevertheless, this procedure is often accompanied with extensive time and cost requirements. Computational approaches are consequently developed in order to overcome those drawbacks. Their goal is to narrow down the required experimental work by pinpointing protein groups which presumably correspond to complexes. Among the existing methods, we here recall: the Markov Clustering Algorithm (MCL) (Van Dongen, 2001) which uses random walks to find cluster structures in PPI networks; the Molecular Complex Detection (MCODE) algorithm (Bader & Hogue, 2003) which interprets complexes as dense regions grown from highly-weighted vertices; the clustering based on

maximal cliques (CMC) method (Liu, Wong, & Chua, 2009); the Affinity Propagation (AP) algorithm (Frey & Dueck, 2007) that uses a distance matrix to gradually generate high-quality clusters; the Restricted Neighborhood Search Clustering (RNSC) algorithm, presented in (King, Pržulj, & Jurisica, 2004) and (Pržulj, Wigle, & Jurisica, 2004), which is a cost-based local search algorithm that seeks to partition proteins into highly-interconnected subsets; the RRW algorithm (Macropol, Can, & Singh, 2009) that uses repeated random walks to exploit the structure of PPI networks; CFinder (Adamcsek et al., 2006) which looks for overlapping and fully-connected complexes based on the clique percolation method; and recently, ClusterONE (Nepusz, Yu, & Paccanaro, 2012) which identifies protein complexes through clustering with overlapping neighborhood expansion. Despite the good performance of these methods, most of them are restricted by the assumption that protein complexes only correspond to dense subgraphs in PPI networks. Thus, they are usually unable to detect complexes with few members and/or few interconnections. Our approach for the detection of protein complexes in PPI networks is based on the ProRank method (Zaki, Berengueres, & Efimov, 2012a), presented in Chapter 2, which is not restrained by the complex-density assumption. We consider the network presented in Figure 7 to trace the ProRank algorithm. It is a sub-network generated from the yeast PPI dataset at the Mentha interactome browser (Calderone, Castagnoli, & Cesareni, 2013), version date 05/01/2014. It corresponds to the largest connected portion of the network and includes 235 interactions of scores greater than or equal to 0.99. The yellow nodes correspond to the essential proteins identified by ProRank and the detected protein complexes are presented in Figure 7.

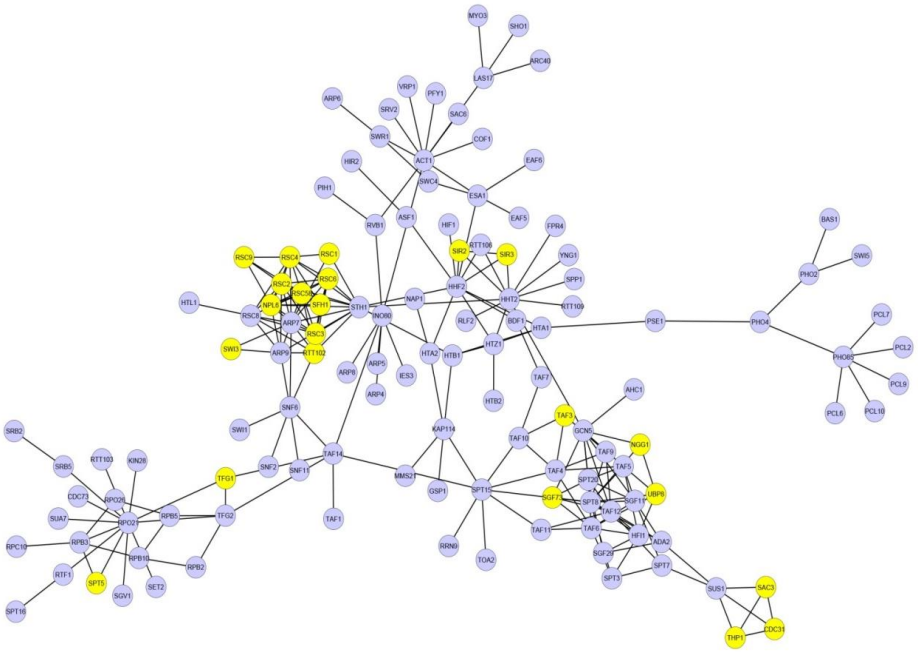

Figure 7: Yeast PPI sub-network. The nodes colored in yellow correspond to essential proteins identified by the ProRank algorithm.

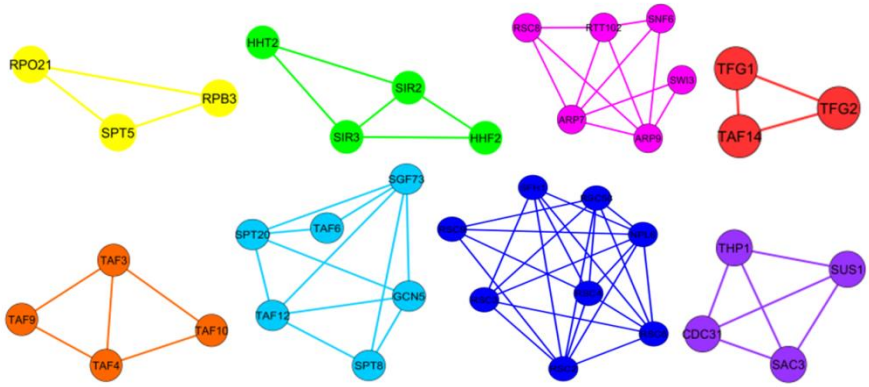

Figure 8: Detected complexes by the ProRank algorithm when applied on the PPI network in Figure 7, under the assumption that a protein can belong to one complex only.

3.2 The ProRank+ Method

Granting that ProRank achieves competitive results when compared to previous approaches, it can be further improved. The pruning, filtering, ranking and complex-detection steps are certainly requisite. In fact, PPI datasets are usually noisy. They have high false-positives (spuriously-detected interactions) and false-negatives (missing interactions) rates which could negatively affect the detection

process. Accordingly, it is important to remove the unreliable edges or even detect the missing ones using proper computational techniques. Categorizing the proteins and forming the detected complexes based on the essential nodes can potentially lead to more accurate results. Moreover, ranking the proteins by their importance in the network and using the spoke model to form protein complexes are all vital to the complex-detection algorithm. Nevertheless, the similarity calculating step can be discarded due to its high computational cost and its low effect on the final results (Zaki, Berengueres, & Efimov, 2012a).

Proteins can participate in multiple cellular functions (Hodgkin, 1998). Hence, a protein can belong to many complexes. For instance, among the 1189 proteins contained in the MIPS catalog of protein complexes (Mewes et al., 2006), 820 proteins (approximately 69%) belong to more than one complex. Similarly, among the 1279 complexes covered by the SGD set (Hong et al., 2008), 332 proteins (approximately 26%) belong to multiple complexes. Consequently, the detected protein complexes are expected to have common members. A detection algorithm which accounts for this fact would most likely lead to more accurate results. This is the first alteration of ProRank. To do that, we explore the formation of protein complexes from various protein seeds by allowing the complexes to overlap. Indeed, the number of detected entities would increase but it is then subject to merging or deleting entities based on their degrees of overlaps. The complexes detected after adding the overlap assumption to ProRank and applying it on the PPI network in Figure 7, are shown in Figure 9.

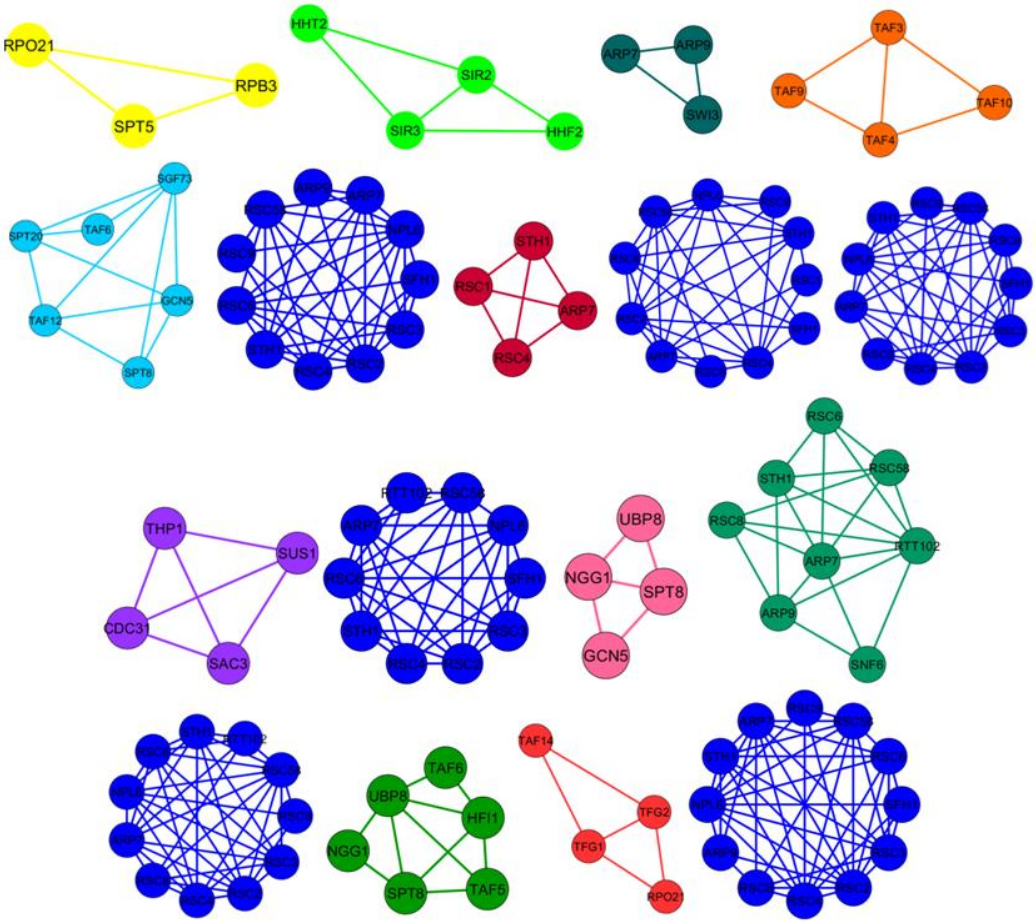

Figure 9: Detected complexes by the ProRank algorithm when applied on the PPI network in Figure 7, under the assumption that a protein may belong to more than one complex.

The results uphold the improvement added by allowing the detected complexes to overlap. However, it can be noticed that the amount of overlaps among some of the detected complexes is relatively high. This was anticipated. Actually, since all essential proteins are now seeds for protein-complex formation, the ones that share numerous neighbors will certainly produce highly similar protein complexes. In order to overcome this limitation and to further improve the quality of the predicted complexes, the following filtering and merging steps are added to the algorithm:

1- Duplicate complexes resulting from the complex-overlap notion are removed.

2- Next, a merging procedure, Merging by Cohesiveness, is applied to explore more variations of the detected complexes. In conformity with the initial considerations of the ProRank method, we rely on the key roles of the essential proteins in the network to establish the merging process. All the detected complexes are matched against each other. Two complexes, *C1* and *C2*, whose percentage of overlapping essential proteins is above a merging threshold, are merged along with their interconnections to form a larger complex *C*. Then, the process uses the cohesiveness measure introduced in (Nepusz, Yu, & Paccanaro, 2012) to assess the quality of the resulting complex and its iterative extensions as follows. The cohesiveness of a complex *C* is given by equation (6):

$$Cohesiveness(C) = \frac{w_{in}(C)}{w_{in}(C) + w_{out}(C) + p} \qquad (6)$$

where $w_{in}(C)$ is the sum of the weights of edges that are entirely contained in *C*, $w_{out}(C)$ is the sum of the weights of edges that connect the proteins belonging to *C* to the rest of the network and *p* is a penalty term reflecting PPI uncertainties. This cohesiveness measure was developed to model the assumption by which a protein complex is viewed as an entity with strongly-interconnected members that is well-separated from the rest of the network. The successive steps of our merging procedure aim at refining merged complex while increasing their cohesiveness measures. For each protein, *prot*, contained in *C*: first, the set of its neighbors, N_{prot}, is formed; then, for each neighbor protein n_{prot} in N_{prot}, the complex $C' = C \cup \{n_{prot}\}$ is

constructed; and if the cohesiveness of *C'* is greater or equal to the cohesiveness of *C*, n_{prot} is added to *C*. After exploring all the proteins initially belonging to *C* in the same manner, the derived complex is added to the final list of detected complexes. The pseudocode of merging two complexes, Merge_by_Cohesiveness, is presented next.

Pseudocode of the Merge-by-Cohesiveness algorithm

```
Merge_by_Cohesiveness (C1, C2, merging_threshold)
ep1 = (set of essential proteins in C1)
ep2 = (set of essential proteins in C2)
if size(ep1) > size(ep2) then
    larger_set = ep1
else larger_set = ep2
end if
ep = ep1 ∪ ep2
if size(ep) > size(larger_set)*merging_threshold then
    C = C1 ∪ C2
    for prot in C do
        N_prot = (set of neighbors of prot)
        for n_prot in N_prot do
            C' = C ∪ {n_prot}
            if Cohesive(C') ≥ Cohesive(C) then
                C = C ∪ {n_prot}
            end if
        end for
    end for
end if
```

3- Additional screening of the generated complexes is applied to remove possible duplicates.

In summary, Figure 10 shows the steps of the ProRank+ algorithm.

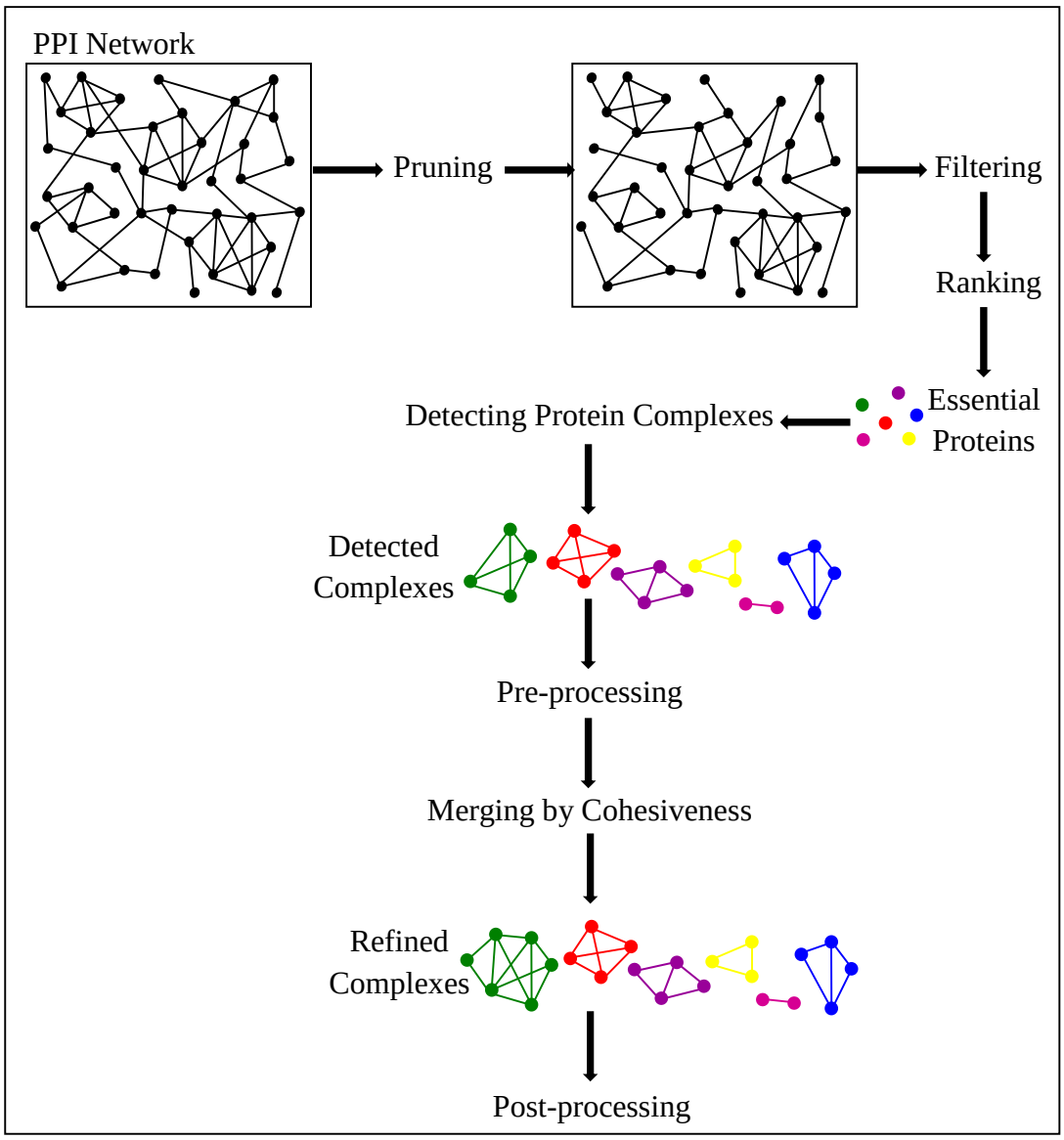

Figure 10: Steps of the ProRank+ algorithm.

3.3 Experimental Study

3.3.1 Datasets and Evaluation Criteria

ProRank+ is tested on five large-scale protein-protein interaction datasets associated to the well-studied yeast microorganism. Four of the datasets consist of weighted protein interactions, they are: Collins (Collins et al., 2007), Krogan core and Krogan extended (Krogan et al., 2006), and Gavin (Gavin et al., 2006). The fifth dataset, BioGRID (Stark et al., 2006), consists of unweighted interactions. The

characteristics of the five datasets used in the experimental work are shown in Table 1.

Dataset	No. of Proteins	No. of Interactions	Network Density	Average no. of neighbors
Collins	1,622	9,074	0.007	11.189
Krogan Core	2,708	7,123	0.002	5.261
Krogan extended	3,672	14,317	0.002	7.798
Gavin	1,855	7,669	0.004	8.268
BioGRID	5,640	59,748	0.004	21.187

Table 1: Characteristics of the five experimental datasets.

The sets of predicted complexes are matched against the MIPS catalog of protein complexes (Mewes et al., 2000). The same datasets and the reference set of complexes are used to evaluate the ClusterONE method and to compare its performance with other approaches. We also adopt the same quality scores applied in (Nepusz, Yu, & Paccanaro, 2012) to assess the quality of our algorithm. In addition, it is important to note that the parameters of the compared algorithms are optimized in such a way to produce best possible results. Given *m* predicted complexes and *n* reference complexes and based on the confusion matrix, $T = [t_{ij}]$, the quality scores cover:

a. The number of complexes in the reference catalog that are matched with at least one of the predicted complexes with an overlap score, *w*, greater than 0.25. The overlap score of two complexes *A* and *B* is calculated based on equation (7).

$$w(A,B) = \frac{|A \cap B|^2}{|A||B|} \qquad (7)$$

b. The clustering-wise sensitivity (S_n) which assesses the matching quality among the reference complexes and the detected ones. It is calculated as shown in equation (8).

$$S_n = \frac{\sum_{i=1}^{n} max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i} \quad (8)$$

c. The clustering-wise positive predictive value (*PPV*) which also reflects the matching quality, mainly in terms of the correctly-matched protein members among the detected complexes. It is computed as shown in equation (9).

$$PPV = \frac{\sum_{j=1}^{m} max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m} \sum_{i=1}^{n} t_{ij}} \quad (9)$$

d. The geometric accuracy (*Acc*) which is the geometric mean of S_n and *PPV*, as shown in equation (10).

$$Acc = \sqrt{S_n \times PPV} \quad (10)$$

e. The maximum matching ratio (*MMR*) which reflects how accurately the predicted complexes represent the reference complexes by dividing the total weight of the maximum matching by the number of reference complexes.

3.3.2 Experimental Settings of ProRank+

The steps of applying ProRank+ on a given dataset, *D*, and their experimental settings are as follows:

1- Pruning: removing unreliable protein interactions from *D* using the AdjustCD method (Chua et al., 2008). This technique assigns weights to the interactions based on the network topology and considers unreliable those whose weights are less than a specified threshold. Here, we experimentally set the pruning threshold to 0.2 for weighted datasets and to 0.45 for unweighted datasets.

2- Filtering: identifying bridge, fjord, and shore proteins which could add noise to the network, as defined in (Zaki, Berengueres, & Efimov, 2012a).

3- Protein Ranking: ordering proteins using a ranking algorithm, in analogy with the PageRank algorithm.

4- Complex Detection: considering all the essential proteins, i.e. those that do not belong to any of the types defined in step 2, as seeds based on which detected complexes are formed using the spoke model. Here, a protein can belong to more than one complex.

5- Pre-processing: filtering the set of predicted complexes by removing possible duplicates generated due to the introduced overlap assumption.

6- Merging by Cohesiveness: two detected complexes, whose overlap is above a merging threshold, here 75%, are merged. The subsequent complex is iteratively extended following the presented merging procedure.

7- Post-processing: filtering the refined set of predicted complexes to remove possibly replicated copies of the same complexes resulting from the previous merging step.

3.3.3 Comparison with Other Methods

ProRank+ is compared to other state-of-the-art methods. They include ProRank (Zaki, Berengueres, & Efimov, 2012a) to highlight the attained improvement, Markov Clustering (MCL) (Van Dongen, 2001), the molecular complex detection (MCODE) algorithm (Bader & Hogue, 2003), the clustering based on maximal cliques (CMC) method (Liu, Wong, & Chua, 2009), the Affinity Propagation (AP) algorithm (Frey & Dueck, 2007), ClusterONE (Nepusz, Yu, & Paccanaro, 2012), the restricted neighborhood search (RNSC) algorithm (King,

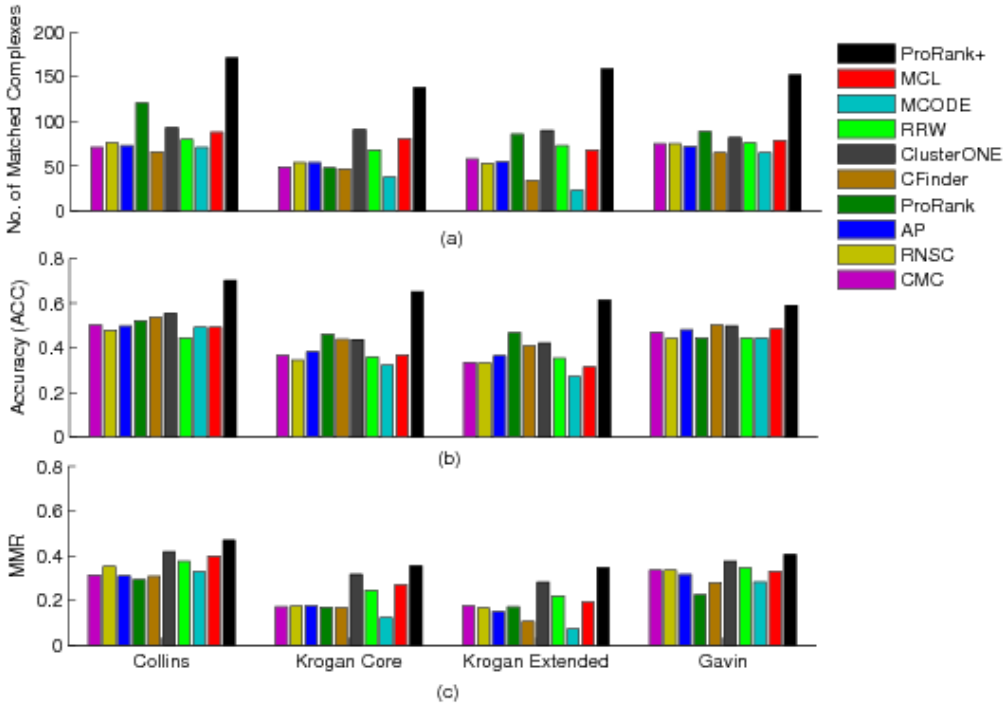

Figure 11: ProRank+ compared to ProRank, MCL, MCODE, CMC, AP, ClusterONE, RNSC, RRW, and CFinder. Here, the four weighted yeast datasets are used: Collins, Krogan core, Krogan extended and Gavin. The comparisons are in terms of (a) the number of clusters that match the reference complexes, (b) the geometric accuracy (Acc) which reflects the clustering-wise sensitivity (Sn) and the clustering-wise positive predictive value (PPV), and (c) the maximum matching ratio (MMR).

Pržulj, & Jurisica, 2004), the RRW algorithm (Macropol, Can, & Singh, 2009), and CFinder (Adamcsek, Palla, Farkas, Derényi, & Vicsek, 2006). The comparisons among the results scored by these approaches (Nepusz, Yu, & Paccanaro, 2012) and those scored by ProRank + are displayed in Figures 11 and 12. Since not all the algorithms can be applied to unweighted datasets, fewer methods for instance were applied on the BioGRID dataset.

The experimental results show that ProRank+ detects a higher number of protein complexes that are matched with the reference set. Note that the number of clusters predicted by ProRank+ is relatively higher than the number of clusters returned by the other methods for Collins, Gavin and BioGRID datasets.

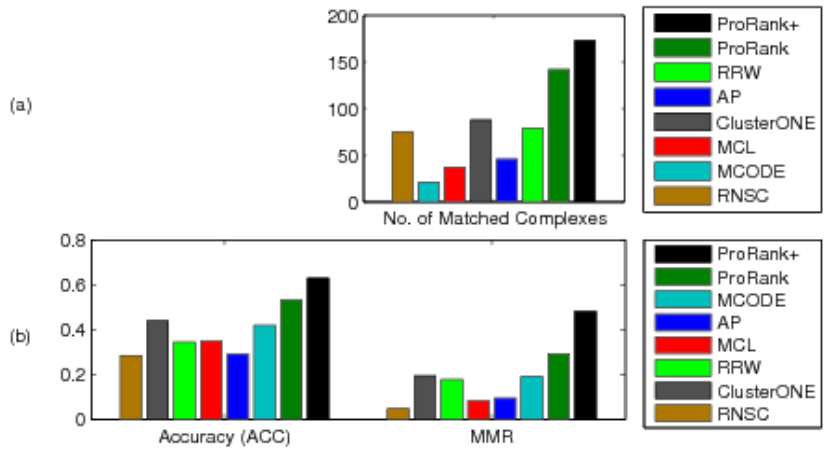

Figure 12: ProRank+ compared to ProRank, MCL, MCODE, AP, ClusterONE, RNSC, and RRW. Here, the un-weighted BioGRID dataset is used. The comparisons are in terms of (a) the number of clusters that match reference complexes, and (b) the geometric accuracy (*Acc*) which reflects the clustering-wise sensitivity (S_n) and the clustering-wise positive predictive value (*PPV*), and the maximum matching ratio (*MMR*).

Nevertheless, the ratio equivalent to the number of matched complexes over the number of detected clusters falls within the same range of the ratio corresponding to the other methods. Added to that, ProRank+ achieves higher clustering-wise sensitivity (S_n), geometric accuracy (*Acc*) and maximum matching ratio (*MMR*) for all the considered datasets. However, it cannot surpass the clustering-wise positive predictive value (*PPV*) of ProRank which was the highest for all datasets. This can be justified by the fact that *PPV* tends to be lower when the overlaps among the detected complexes are substantial. By the *PPV* formula, a complex-detection algorithm that fully succeeds in detecting the reference complexes has a *PPV* value less than or equals to 1 since there is a matching predicted complex for every reference complex, in addition to other predicted complexes that partially overlap with reference complexes. On the other hand, a dummy detection algorithm which distributes the proteins into separate sets of single elements has a *PPV* value equals to 1, which is greater than the *PPV* of the perfect algorithm that is able to detect all

reference complexes. Consequently, *PPV* values must be carefully analyzed since they may not always reflect the competence of a certain method. Moreover, the geometric accuracy (*Acc*) is negatively affected by the predicted complexes that do not match any of the reference complexes. This somehow contradicts the initial purpose of developing methods for the detection of protein complexes which mainly consists of finding previously unknown or undiscovered entities. Accordingly, the *MMR* measure (Nepusz, Yu, & Paccanaro, 2012) is introduced to overcome such limitations by dividing the total weight of the maximum matching with the number of reference complexes. The *MMR* values achieved by ProRank+ are in the favor of the proposed approach. We hereby note that our approach can also be explored using other pruning methods such as the ones introduced in (Zaki, Efimov, & Berengueres, 2013) and (Kritikos, Moschopoulos, Vazirgiannis, & Kossida, 2011).

3.3.4 Testing the Ability of ProRank+ to Detect Small Complexes

Detecting small protein complexes is not a common feature of complex-detection methods. In fact, it is important to identify such complexes in protein interaction networks. For instance, among the 313 protein complexes included in the MIPS catalogue (Mewes et al., 2006), 104 complexes consist of 2 or 3 proteins (approximately 33 %). Most of the approaches which view protein complexes as dense regions in the interaction networks are usually unable to detect complexes of small sizes. Hence, we also test the ability of ProRank+ to detect small protein complexes. We consider the same yeast datasets that are utilized in the previous experiments. The set consisting of the 104 complexes of small sizes in the MIPS catalogue (Mewes et al., 2006) is formed and used as a reference set. The datasets are filtered by the AdjustCD method with a threshold of 0.2. The corresponding results

are shown in Table 2. The table highlights the competency of ProRank+ in detecting small protein complexes in terms of the number of matched complexes as well as the accuracy (*Acc*) and the maximum matching ratio (*MMR*) scores.

Dataset	Predicted Complexes	Matched Complexes	Sn	Acc	MMR
Collins	428	91	0.875	0.935	0.433
Krogan Core	229	34	0.667	0.816	0.163
Krogan Extended	260	78	0.75	0.769	0.217
Gavin	534	57	0.897	0.947	0.293
BioGRID	823	78	0.882	0.9	0.351

Table 2: The results of testing ProRank+ on small complexes.

3.3.5 Testing ProRank+ on Human Protein-Protein Interaction Dataset

When tested on various datasets, weighted and unweighted, ProRank+ is able to detect more complexes than state-of-the-art methods with higher quality scores. Indeed, the method could be very helpful for biologists if it is also tested on Human interactions and proved valuable in detecting known protein complexes of key roles in normal and abnormal cellular functions. Therefore, we apply our method on the Human protein interactions dataset in the BioGRID repository (Stark et al., 2006). The interactions are unweighted, and thus the pruning threshold was set to 0.45. The pruned dataset consists of 3031 interactions. ProRank+ is able to predict 267 protein complexes. We then examine the detected entities for potential mappings with known protein complexes; some of which are presented in Table 3 and highlighted hereafter.

Detected Complex	Proteins Members of the Detected Complex	Matching Percentage
CCT micro-complex	{CCT3, CCT2, CCT8, CCT6A, CCT4, CCT7, CCT5, TCP1}	100 %

Ribosomal protein complex	{RPL32, RPS17, RPSA, RPL10A, RPL12, SLC25A5, RPL7, RPL18, RPL15, RPL21, RPS6, RPS4X, RPL19, RPL14, RPL4, RPS27L, RPS23, RPS26, RPS16, RPL7A, RPS24, RPS13, RPS15A, RPS8, RPS3A, FAU, RPL11, RPL6, RPL9, RPL5, RPS27, RPL17, RPS2, RPS25, RPS20, NOP56, RPS15, RPL23A, RPS10, RPL10L, RPLP0P6, RPS28, RPS5, RPS9, RPL23, RPL18A, RPS3, RPL37A, RPL31, RPL10, RPL8, RPS11, RPL36, RPS19, RPL30, RPL24, RPS21, RPL27, RPS12, RPL29, RPS29, RPS7, RPL22, RPLP0, RPS14, RPL3, RPLP2, RPL27A, RPL13, RPS18, RPS27A}	81.48 %
PA700-20S-PA28 complex	{PSMD8, PSMB2, PSMC3, PSMC4, PSMA4, PSMA1, PSMD1, PSMD7, PSMA2, PSMB6, PSMB7, PSMD3, PSMB1, PSMC1, PSMC5, PSMC2, PSMB4, PSMA6, PSMD6, PSMD14, PSMD12, PSMD11, PSMD13, PSMA7, PSMC6, PSMA5, PSMB3, PSMB5, PSMA8, PSMD2}	83.33 %
SWItch/Sucrose NonFermentable (SWI/SNF) complex	{SMARCA4, SMARCC1, ARID1A, SMARCE1, SMARCC2, SMARCA2, SMARCB1}	60 %

Table 3: Selected complexes detected by ProRank+ when tested on human protein-protein interaction dataset.

1- The CCT micro-complex (Liou & Willison, 1997) which participates in protein folding, assembly and transport. It is fully-detected by ProRank+.
2- The Ribosomal protein complex (Nakao, Yoshihama, & Kenmochi, 2004) is detected with a 81.48 % match. Five additional proteins are detected: SLC25A5, RPS27L, NOP56, RPL10L, and RPLP0P6. Their association with the detected complex may be just noise or, on the contrary, can present biologically meaningful information.

3- The PA700-20S-PA28 complex (Kopp, Dahlmann, & Kuehn, 2001) is detected with a mapping percentage of 83.33%. This complex is a key component of the ATP-dependent proteolytic pathway in eukaryotic cells and is responsible for the degradation of most cellular proteins.

4- A recent publication (Shain & Pollack, 2013) confirms that the mutations of the SWItch/Sucrose NonFermentable (SWI/SNF) complex are ubiquitous in various types of cancer. Accordingly, future research efforts will put more focus on this tumor suppressor complex towards better understanding of cancer diseases and in the direction of developing more effective therapies. The SWI/SNF complex is composed of ten elements distributed as follows: (a) SMARCA2 or SMARCA4, two mutually-exclusive ATPase enzymatic subunits; (b) ARID1A, ARID1B, or PBRM1, three mutually-exclusive subunits associated to functional specificity; (c) core and accessory subunits including SMARCB1, SMARCC1, SMARCC2, SMARCE1, SMARCD1, SMARCD2, or SMARCD3, PHF10, DPF1, or DPF2, DPF3, and ACTL6A or ACTL6B. We map the composition of SWI/SNF with the set of predicted complexes by ProRank+. Our method is able to detect a complex consisting of the elements SMARCA4, SMARCC1, ARID1A, SMARCE1, SMARCC2, SMARCA2, SMARCB1. In comparison with the known structure of SWI/SNF, ProRank+ correctly predicts six members out of ten corresponding to 60 % of its subunits with a relatively low number of false positives.

The above experiment affirms the ability of ProRank + to identify significant and key protein complexes from protein interaction data. In addition, such outcomes

can potentially contain relevant and previously-undiscovered protein complexes or unidentified protein members of certain complexes.

3.4 Conclusion

ProRank+ is an efficient method for detecting protein complexes in protein-protein interaction networks. The detection process is mainly centered on a ranking algorithm that allows the identification of key proteins based on which the corresponding components are formed. It is also tailored by a series of pruning, filtering and merging steps, allowing the refinement of the drawn complexes. Unlike most approaches, the design of our method is not bound by the sole association of protein complexes to dense regions in interaction networks. In addition, ProRank+ takes into account possible overlaps among complexes and this is an important assumption that reflects biological facts. In contrast with other methods, the experimental study underlines the competitive ability of ProRank+ to identify protein complexes. The performance of our algorithm is tested on weighted and un-weighted datasets and using Human protein interaction data as well. The results greatly favor our method.

Chapter 4: Detecting Protein Complexes in Dynamic Protein-Protein Interaction Networks

In this chapter we present our solution for the second research objective: modeling the dynamic aspect of PPI networks and detecting protein complexes accordingly. In section 4.1, we review the motivation of this objective and list the advantages of modeling the dynamics of protein interaction networks. The DyCluster method is introduced in section 4.2. The experimental study and results are presented in section 4.3. Finally, the chapter is concluded in section 4.4.

4.1 Background

Early methods developed for the detection of protein complexes usually model protein-protein interaction data as a static and all-inclusive network. However, protein interactions do not occur at the same time (Macropol, Can, & Singh, 2009), i.e. they are subject to various temporal, spatial and contextual settings. Accordingly, instead of a single network representation, we would rather be looking at a series of snapshots of a PPI network modeled based on either one or a combination of conditions, as shown in Figure 13.

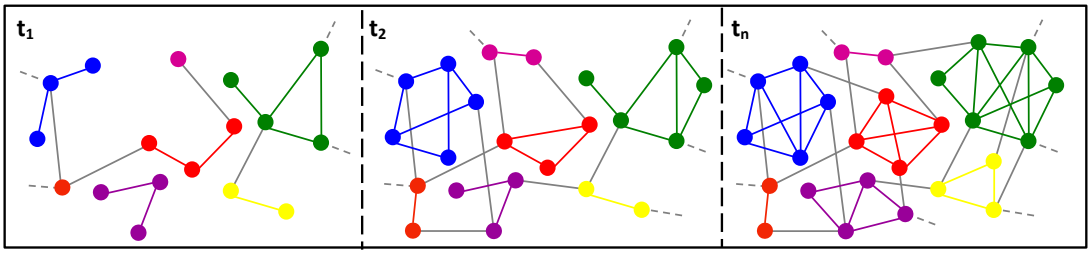

Figure 13: Snapshots of a hypothetical PPI network, showing its dynamics through different temporal, spatial and/or contextual settings. Nodes and edges of the same color belong to the same protein complex.

Novel experimental techniques can currently make such biological information available. Hence, the shift from viewing PPI networks as static to modeling the dynamics of these networks became fundamental (Przytycka, Singh, & Slonim, 2010). Hereafter, we highlight some of the advantages of this transition. First, it is a natural response to advances in experimental methods as it enhances the replication of real biological events. Indeed, the more representative are the models and the methods, the higher the accountability and the accuracy of the produced results. Second, by combining different biological data, we can reach a computational visualization level of protein interaction events that could verify or even contradict biological concepts. Furthermore, previously unknown facts may be learned, such as the characterization of hub proteins (Han et al., 2004) as "party hubs" which interact with their partners at the same time or "date hubs" which connect to their partners at different times and locations. In addition, integrating multiple types of biological information allows overcoming data limitation issues. For instance, PPI datasets are usually susceptible to high error rates (Reguly et al., 2006); they may have missing interactions or may include spuriously-detected ones. Moreover, possible enrichment data that can be used to model the dynamics of PPI networks, such as gene expression profiles (Chen & Yuan, 2006) and gene ontology (Xu, Lin, & Yang, 2010), suffer from low gene coverage in contrast with most PPI datasets, in which the number of interacting proteins is typically very high (Von Mering et al., 2002). The recurrence of information and/or inferences that are drawn from different types of biological data can be seen as a confidence indicator. In view of that, the integration of various datasets, even if not highly-credible, in the direction of modeling PPI dynamics can potentially reduce the effect of false positive and false negative rates, as well as low coverage issues. In contrast with static PPI

networks, the information revealed by dynamic networks is at a higher level of details. For instance, in the problem of identifying protein complexes, most of the presented algorithms do not differentiate between functional modules and protein complexes. That is mainly due to the absence of embedded information in the networks that could guide the search. In fact, complexes are formed by proteins which interconnect at the same time and place, whereas the members of functional modules may interact at different times and places (Spirin & Mirny, 2003). Accordingly, when PPIs are constrained by spatiotemporal conditions inferred from gene expression and gene ontology datasets, for example, the detected components could more likely be categorized as protein complexes or functional modules. Likewise, dynamic PPI modeling may highly contribute to the detection of protein subcomplexes. Various approaches were developed to solve this important research problem, but all based on static networks (Zaki & Mora, 2014). As dynamic modeling can reveal the mechanisms of protein-complex formation and can thus yield better complex-detection approaches, it can also provide the same for the detection of subcomplexes. Finally, since dynamic PPI networks better describe protein interconnections, they can highly lead to better analytical results. The integration of temporal, spatial or contextual biological information with PPI data as a means to reproduce the PPI dynamics, can be viewed as clustering based on temporal, spatial and/or contextual attributes. Hence, proteins and their interactions can be grouped based on the integrated conditions and complex-detection methods shall be applied accordingly, indeed with a generalization capability. Consequently, the reliability of computational approaches is expected to increase.

Based on the listed advantages, the next objective is to model the dynamic aspect of PPI networks and then modify our complex-detection algorithm, ProRank+, accordingly. Our approach is presented hereafter.

4.2 The DyCluster Method

The biological information that could be used to represent dynamic PPI networks include, but are not limited to, gene expression data (Lovén et al., 2012) which report quantitative measurement of RNA species in cellular compartments across various conditions, subcellular localization annotations (de Lichtenberg, Jensen, Brunak, & Bork, 2005) which provide spatial positions of elements in cellular components; and gene ontology annotations (Ashburner et al., 2000) which highlight genes that are present across different species. Time-series gene expression data measure quantities of RNA across different time points in cellular processes. Genes with correlated expressions across various conditions most likely interact. Hence, the combination of time-series gene expression information with PPI data can be used to model the dynamics of the PPI networks. For instance, that is done in (Wang, Peng, Xiao, Li, & Pan, 2013), (Wang, Peng, Li, Luo, & Pan, 2011), (Li, Chen, Wang, Wu, & Pan, 2014) and (Kim, Han, Choi, & Hwang, 2014); as elaborated in the literature review (Chapter 3). Our proposed approach, DyCluster, requires as input a gene expression dataset and a PPI dataset. It consists of five main steps: biclustering gene expression data, extracting biclusters' PPIs, pruning bicluster PPIs, detecting protein complexes and finally, merging and filtering the sets of detected protein complexes. An outline of the method is presented in Figure 14.

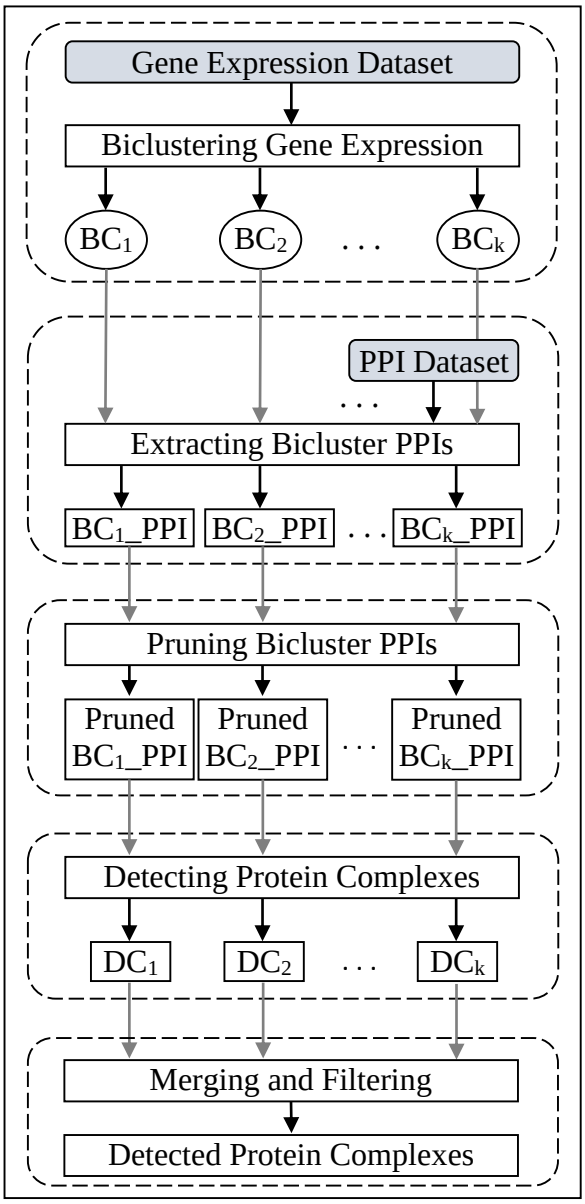

Figure 14: An outline of the DyCluster method.

4.2.1 Biclustering Gene Expression Data

A gene expression dataset shows the expression levels of a typically large number of genes across different environmental conditions, time points, organs, species, etc. It is conventionally represented as a matrix in which rows and columns correspond to genes and their expression levels at different conditions or samples respectively. It is assumed that genes which exhibit similar expression patterns

across various conditions can be functionally-related (Baldi & Hatfield, 2002). The analysis of these datasets is challenging because they are usually unbalanced, i.e. the number of genes is quite larger than the number of conditions (Watson & Berry, 2009). Various approaches are proposed to analyze expression data and to group genes according to their expression patterns; in particular, data mining approaches such as classification and clustering. Classification methods require knowing the label of the resulting classes in advance, which somehow limits the process of data exploration. Nevertheless, several research efforts study the application of such supervised techniques on gene expression data (Asyali, Colak, Demirkaya, & Inan, 2006). Likewise, typical clustering techniques have two drawbacks when applied to gene expression data (Jiang, Tang, & Zhang, 2004): first, each gene must be placed in a cluster even if its similarity with other cluster members is relatively low; second, a gene can belong to one cluster only. Consequently, these techniques cannot account for the fact that a large number of genes can exhibit multiple biological functions (Hodgkin, 1998), and thus can belong to more than one cluster. Besides, clustering spans the whole sample set whereas in reality, the expression levels of a gene cluster may be correlated based on a subset of samples. Thanks to the simultaneous two-dimensional clustering capability which they provide, biclustering techniques present better means to explore expression data (Madeira & Oliveira, 2004). In fact, they allow the identification of subsets of co-regulated genes across subsets of samples. Added to that, in analogy to biological facts, a gene may belong to multiple clusters or may not fit in any cluster in some cases.

A problem formulation of biclustering gene expression data is as follows: Let A be an $n*m$ data matrix, representing a gene expression dataset consisting of n genes measured across m conditions; a_{ij} being a real value corresponding to the

expression level of the gene at row *i* and the condition at column *j*. The goal is to find a set of biclusters *BC(I, J)*; where *I* is a subsets of genes which exhibit similar expression patters across the subset of conditions *J*.

We highlight some of the existing biclustering approaches which are also used later to evaluate the DyCluster method. Biclustering is first applied on gene expression data by Cheng and Church (Cheng & Church, 2000). Their method, CC, consists of a greedy search heuristic to form the biclusters, namely the set covering algorithm, and uses the Mean Square Residue (MSR) measure to assess the biclusters' quality based on a specified threshold. The MSR of a bicluster *BC*, of *I* rows and *J* columns, reflects the degree of coherence between the genes and the conditions which it includes. It is calculated based on equation (6) where bc_{ij}, bc_{iJ}, bc_{Ij} and bc_{IJ} represent the elements in row *i* and column *j*, the row and the column means, and the mean of *BC*, respectively.

$$MSR(BC) = \frac{1}{|I||J|}\sum_{i=1}^{|I|}\sum_{j=1}^{|J|}\left(bc_{ij} - bc_{iJ} - bc_{Ij} + bc_{IJ}\right)^2 \qquad (11)$$

The lower the MSR, the higher is the bicluster coherence. Correlations among genes can be expressed in terms of scaling and shifting patterns. A robustness characteristic of a biclustering algorithm, when applied on expression data, is in its ability to capture both types of patterns. MSR can only detect shifting correspondences among the expression levels of genes (Bozdağ, Kumar, & Catalyurek, 2010). Despite that, it is used by several similar approaches and some variants of this measure are also introduced to identify the scaling patterns (Mukhopadhyay, Maulik, & Bandyopadhyay, 2009). The Order Preserving Sub Matrix (OPSM) algorithm (Ben-Dor, Chor, Karp, & Yakhini, 2003) searches for large submatrices in which genes have the same linear ordering of the samples. The

Iterative Signature Algorithm (ISA) (Bergmann, Ihmels, & Barkai, 2003) uses the signature algorithm to identify self-consistent transcriptional modules consisting of co-expressed genes and the samples corresponding to them. A comprehensive survey of these methods can be found in (Madeira & Oliveira, 2004).

4.2.2 Extracting Bicluster PPIs

Given the set of gene biclusters, $BC = \{BC_1(I_1, J_1), BC_2(I_2, J_2), ..., BC_k(I_k, J_k)\}$, the next step consists of finding the interconnections among the members of each bicluster based on a specified PPI dataset. The interactions involving elements that belong to the set of proteins in each bicluster are extracted.

4.2.3 Pruning Bicluster PPIs

PPI datasets are usually noisy (Marto, 2009). As a result, many methods are developed to prune PPI data and thus to reduce their levels of false positives and false negatives such as (Chua et al., 2008) and (Zaki, Efimov, & Berengueres, 2013). Here, we use the PE method introduced by Zaki et al. to assess the reliability of protein interactions at the level of generated biclusters and to prune the corresponding PPI subsets accordingly. Experiments show that PE-measure is efficient as it reduces the level of noise in protein interaction networks by looking for subgraphs that are closest to maximal cliques, based on the weighted clustering coefficient measures.

4.2.4 Detecting Protein Complexes

Successively, a protein-complex detection method is applied on the pruned biclusters PPIs, disjointedly on each bicluster. Therefore, several sets of identified protein complexes are formed, $DC_1, DC_2, ..., DC_k$.

4.2.5 Merging and Filtering

Merging and filtering the resultant sets of complexes is crucial to the overall accuracy of our approach. However, developing an appropriate post-processing method is challenging because it is subject to various considerations. For instance, in its simplest form, it may consist of matching the detected entities against each other and combining the ones which have an overlap greater than a certain threshold. In contrast, keeping the common members of highly-overlapping entities may also be explored and it might lead to better outcomes. Another approach may think through the core-attachment interpretation of complexes (Gavin et al., 2006) and consider that a repeated subgroup of interacting proteins in several detected groupings may be a potentially correct core, which forms different complexes when linked with various protein attachments. Nonetheless, in our paper, we keep this task for later research stages and we hereby limit the formation of the combined set of complexes to merging based on an overlap threshold and a condition by which members of one complex interact with a certain percentage of members of the other complex; in addition to filtering duplicates. This step finalizes the complex-detection process.

4.3 Experimental Study

4.3.1 Datasets

DyCluster requires a gene expression dataset to model the dynamic aspect of protein interactions and a PPI dataset from which the interconnections among those proteins are extracted. Certainly, the higher the homogeneity of both datasets, namely in terms of the species and the number of common genes that they cover, the better are the expected outcomes. We refer to Gene Expression Omnibus (GEO)

repository (Barrett et al., 2013) from which we select the expression dataset of accession number GSE3431 (Tu, Kudlicki, Rowicka, & McKnight, 2005), entitled "Logic of the yeast metabolic cycle". It reports the expression levels of genes across twelve time intervals in three successive metabolic cycles. Our choice is primarily based on its wide coverage of yeast proteins and potentially, a high number of participants in various cellular processes. The yeast PPI dataset is downloaded from the Database of Interacting Proteins (DIP) (Xenarios et al., 2002) catalogue of experimentally-determined protein interactions. Finally, we compare our results to the CYC2008 catalogue (Pu et al., 2009) containing 408 complexes, as reference set of yeast protein complexes.

4.3.2 Evaluation Scores

The quality scores, used to evaluate our approach, include: (a) the number of complexes in the reference catalogue that are matched with at least one of the predicted complexes with an overlap score, $OS \geq 0.2$; (b) the clustering-wise sensitivity (S_n) and (c) the clustering-wise positive predictive value (*PPV*) used to calculate the matching quality, mainly in terms of the correctly-matched protein members among the detected complexes; (d) the geometric accuracy (*Acc*) which is the geometric mean of S_n and *PPV*; and (e) the maximum matching ratio (*MMR*) which measures the maximal one-to-one mapping between predicted and reference complexes by dividing the total weight of the maximum matching with the number of reference complexes. Note that the same measures are used to evaluate the ProRank+ method, introduced in the previous chapter.

4.3.3 Algorithms

For the gene expression biclustering step, we use three algorithms: OPSM (Ben-Dor, Chor, Karp, & Yakhini, 2003), CC (Cheng & Church, 2000) and ISA (Bergmann, Ihmels, & Barkai, 2003). Here, we note that although efforts are spent in the direction of finding suitable ways to evaluate biclustering approaches (Oghabian, Kilpinen, Hautaniemi, & Czeizler, 2014), comparing their performances is still a challenging task. Added to that, in order to shed the light on the advantage of using gene expression data, we also examined the results of applying the framework using the one-way clustering method k-means (Hartigan & Wong, 1979), based on Pearson's correlation as a distance measure. The parameters settings of these algorithms are presented in Table 4. We used the BicAT tool (Barkow et al., 2006) to visualize and perform the biclustering of the gene expression dataset.

	Parameter Settings
CC	upper limit of MSR: $\delta = 0.5$ threshold for multiple node deletion: $\alpha = 1.2$ number of output biclusters = 10
OPSM	number of passed models for each iteration: $l = 10$
ISA	threshold of genes: $t\_g = 0.5$ threshold of chips: $t\_c = 0.5$ number of starting points = 100
K-means	distance measure: Pearson's correlation number of clusters = 10 number of iterations = 100 number of replications = 1

Table 4: Parameter settings of the biclustering algorithms.

For the step consisting of pruning the PPI data at the biclusters levels, we apply the PE method (Zaki, Efimov, & Berengueres, 2013) with default parameters, specifically, with edges reliability score threshold equals to 0.1. In terms of protein-complex detection methods, we use ProRank (Zaki, Berengueres, & Efimov, 2012a), ProRank+ (Hanna & Zaki, 2014), ClusterONE (Nepusz, Yu, & Paccanaro, 2012) and

CMC (Liu, Wong, & Chua, 2009), MCODE (Bader & Hogue, 2003) and CFinder (Adamcsek et al., 2006). ProRank, ProRank+, ClusterONE and CFinder are applied with default parameters. For CMC, the overlap and the merging thresholds are set to 0.75 and 0.5, respectively. For MCODE, degree cutoff, node score cutoff, k-core and maximum depth from seed are set to 2, 0.2, 2 and 3, respectively.

Added to that, the generated sets of detected complexes are examined and refined as follows: if two complexes have a number of overlapping members greater than 75% of the size of the smaller complex; and if the members of the first complex interact with at least 50% of the members of the second complex, then they are merged.

4.3.4 Results

According to the presented framework, the gene expression dataset, GSE3431, is processed by the three biclustering algorithms, OPSM, CC and ISA, and by the k-means clustering algorithm, one at a time. The PPIs corresponding to the proteins contained in each of the resulting biclusters are extracted from the specified yeast PPI dataset and are pruned using PE technique. The protein complex-detection methods, listed above, are applied on the generated biclusters. Finally, the detected sets of complexes are merged, filtered and matched against the CYC2008 reference catalogue. Table 5 shows the corresponding results in terms of the number of matched protein complexes and the number of detected complexes along with the corresponding evaluation scores. For comparison purpose, the table also includes the results of just applying the detection algorithms on the PPI dataset, excluding the gene expression data.

Method	Biclustering Algorithm	No. of matched complexes	No. of detected complexes	Acc	S_n	MMR	PPV
ProRank	**None**	41	230	0.4715	0.3072	0.1032	0.7237
	OPSM	**78**	335	**0.5911**	**0.4627**	**0.2103**	**0.755**
	CC	**63**	252	**0.5658**	**0.4296**	**0.1804**	**0.7451**
	ISA	**71**	320	**0.564**	**0.4332**	**0.195**	**0.7342**
	K-means	**71**	331	**0.556**	**0.4222**	**0.1896**	**0.7322**
ProRank+	**None**	46	274	0.4788	0.3371	0.1161	0.6801
	OPSM	**81**	397	**0.5982**	**0.5116**	**0.225**	**0.6995**
	CC	**65**	305	**0.5668**	**0.4724**	**0.1947**	**0.6802**
	ISA	**78**	392	**0.5677**	**0.4719**	**0.2231**	**0.683**
	K-means	**78**	424	**0.5687**	**0.4782**	**0.2196**	0.6764
ClusterONE	**None**	76	365	0.6008	0.511	0.2349	0.7064
	OPSM	**89**	929	**0.6426**	**0.5758**	**0.2469**	**0.7172**
	CC	**78**	578	**0.6267**	**0.5465**	0.2036	**0.7186**
	ISA	**87**	890	**0.6015**	**0.5506**	**0.2499**	0.6571
	K-means	**83**	862	**0.6153**	**0.533**	0.2334	0.7102
CMC	**None**	114	4292	0.6587	0.6517	0.347	0.6658
	OPSM	100	1207	0.6159	0.5566	0.2903	**0.6816**
	CC	95	1145	0.5983	0.5264	0.2844	**0.6801**
	ISA	100	1843	0.6041	0.5518	0.3071	0.6614
	K-means	94	1126	0.6088	0.5542	0.2913	**0.6689**
Mcode	**None**	62	168	0.55	0.4271	0.149	0.7082
	OPSM	**71**	475	**0.5695**	**0.4602**	**0.1835**	0.7049
	CC	60	285	0.545	0.4058	**0.1581**	**0.7321**
	ISA	**63**	315	**0.5529**	0.4232	**0.171**	**0.7222**
	K-means	**74**	448	**0.5658**	**0.4583**	**0.1947**	0.6986
CFinder	**None**	116	6381	0.6143	0.5641	0.3776	0.669
	OPSM	94	2079	**0.6187**	0.525	0.2925	**0.7291**
	CC	98	1236	0.5977	0.559	0.3005	0.6391
	ISA	99	2119	0.5738	0.5393	0.3021	0.6104
	K-means	99	1352	0.5988	0.5455	0.3098	0.6574

Table 5: Experimental results of matching the sets of protein complexes, detected by the DyCluster framework, against the CYC2008 reference catalogue.

4.3.5 Case Study

Next, we test the effectiveness of DyCluster on a network of 140 key genes involved in programmed cell death in Rat Apoptosis (RT2 Profiler PCR Array Rat Apoptosis, PARN-012A) and inflammation (RT2 Profiler PCR Array Rat

Inflammatory Cytokines and Receptors, PARN-011A). All the 140 genes are processed using String 9.1 (http://string-db.org/) (Jensen et al., 2009). String is a biological database and web resource of known and predicted protein-protein interactions. All the corresponding proteins and their interactions are retrieved and the network was built. Once the PPI network including 1413 interactions and 140 proteins related to the Rattus norvegicus species is build, several enrichment features available in String 9.1 (features related to KEGG pathway, Reactome Pathway, Molecular function, Pfam domain, InterPro-Domains) are used to generate several sub-networks/groups which were then considered as protein complexes. The idea here is to see whether DyCluster is capable of detecting such groups of biologically related proteins given only the PPI network information.

In this experiment, the gene expression data set, of accession number GSE17384, is downloaded from the GEO (Barrett et al., 2013) repository. It is entitled: "Gene expression data from the LEC rat model with naturally occurring and oxidative stress induced liver tumorigenesis". It reports the variations of gene expression levels in a stepwise manner from the normal liver condition, to chronic oxidative stress-induced hepatitis and liver tumor by time-series microarray analysis. In other words, the study involves a comparison between normal liver tissues and developed liver tumors at different time points. It can potentially reveal genes which participate in the progressive formation of the disease. The OPSM method (Ben-Dor, Chor, Karp, & Yakhini, 2003) is used to bicluster the gene expression data since it shows a relatively good performance in our experimental study.

Then, we examine the results for potential matching with the reference subnetworks/groups generated using String. Table 6 shows the detected components by DyCluster framework, listed by types, along with their matching percentages. The

experimental results thus confirm the potential of our approach in detecting and understanding protein entities of key roles in normal and abnormal cellular functions.

	Detected Component	**Matching Percentage**
InterPro-Domains	Chemokine receptor family	100%
	G protein-coupled receptor, rhodopsin-like	100%
	GPCR, rhodopsin-like, 7TM	100%
	BLC2 family	83.3%
	BLC2-like	83.3%
	Death effector domain	66.7%
	Interleukin-6 receptor alpha, binding	50%
	Death domain	100%
	Apoptosis regulator, Bcl-2, BH2 motif, conserved site	75%
	Chemokine interleukin-8-like domain	60%
KEGG Pathway	Chemokine signaling pathway	40%
	Cytokine-cytokine receptor interaction	32.8%
	NOD-like receptor signaling pathway	31.3%
	Apoptosis	34.4%
	Autoimmune thyroid disease	71.4%
	Huntington's disease	66.7%
	Systemic lupus erythematosus	40%
	Asthma	50%
	Intestinal immune network for IgA production	25%
	Cell adhesion molecules	50%
	Pathways in cancer	70%
Molecular Function	Peptide receptor activity	58.3%
	Receptor activity	52.2%
	Growth factor activity	60%
	C-C chemokine binding	66.7%
	Tumor necrosis factor receptor superfamily binding	40%
	Death effector domain binding	66.7%
	Growth factor binding	50%
	Nucleic acid binding transcription factor activity	75%
	Chemokine activity	77.8%
Pfam Domains	7 transmembrane receptor, rhodopsin family	100%
	Apoptosis regulator proteins, Bcl-2 family	83.3%
	Death effector domain	66.7%
	Interleukin-6 receptor alpha chain, binding	50%
	Small cytokines (intecrine/chemokine), interleukin-8 like	53.3%
	Death domain	100%
	Activation of DNA fragmentation factor	66.7%
	Interleukin-1 family precursors are cleaved by caspase-1	100%

Reactome Pathway	Downstream TCR signaling	100%
	FasL/CD95L signaling	100%
	Exocytosis of platelet alpha granule contents	100%
	IRAK4 is activated by autophosphorylation	75%
	Beta defensins	66.7%
	TRAIL signaling	66.7%
	Interleukin-1 processing	75%
	FASL:FAS Receptor Trimer, FADD complex	100%

Table 6: The detected components by the DyCluster framework when applied on the Rattus norvegicus datasets.

4.4 Conclusion

DyCluster is a framework for the detection of protein complexes in dynamic protein interaction networks modeled by incorporating gene expression data, through biclustering techniques. It responds to the important shift from interpreting PPI data as a single static network to modeling and exploring the dynamic nature of these networks. Our approach is tested using several biclustering techniques and various protein complex detection methods. As the experimental results show, the incorporation of gene expression data in the process of detecting protein complexes in dynamic PPI networks is indeed beneficial, in contrast with the detection of complexes in static networks. On one hand, it can notably increase the correctness and the quality of the results, as it is the case for ProRank, ProRank+ and ClusterONE where the numbers of matched complexes, Acc, Sn, PPV and MMR are higher. On the other hand, biclustering genes based on their expression patterns can significantly reduce the large number of complexes detected by some algorithms, such as CMC and CFinder, while not compromising the quality of the outcomes. The framework models the dynamic aspect of PPI networks by grouping proteins according to the similarities of their expression patterns across subsets of conditions. Moreover, it is not restricted by threshold imposition on gene expression levels. As

mentioned earlier, biclustering approaches are better than conventional clustering methods when it comes to expression data analysis. Nonetheless, the results attained by DyCluster using the k-means clustering algorithm accentuate the improvement which can be gained by incorporating gene expression information to model the dynamics of PPI interactions and to detect protein complexes in PPI networks accordingly. Finally, the produced results in our case study are in favor of the DyCluster framework.

Chapter 5: Gene-Disease Association through Topological and Biological Feature Integration

In this chapter, we present our gene-disease association approach. Background information is presented in section 5.1. The building blocks of our learning model are discussed in section 5.2. The experimental study is shown in section 5.3. A case study in which we apply our approach on the Diabetes Mellitus, Type II disease is presented in section 5.4. The chapter is concluded in section 5.5.

5.1 Background

The huge amounts of information generated using high-throughput experimental techniques continue to motivate the design of suitable methods for valuable biological knowledge mining. In particular, the identification of the genes and the inter-molecular events leading to the formation of diseases remains essential towards the development of effective medical therapies. The association of genes to one disorder accelerates the linkage of key players to other diseases. Full insights about the formation processes of most diseases are still incomplete. Based on the assumption that genes related to similar disorders tend to be functionally associated (Oti & Brunner, 2007) (Wu, Jiang, Zhang, & Li, 2008), existing methods often follow the notion of guilt-by-association (Altshuler, Daly, & Kruglyak, 2000) by which genes are ranked based on their similarity to known disease genes. The literature contains numerous approaches designed to link genes to diseases. We hereafter recall some of them. Deng et al. (Deng, Chen, & Sun, 2004) build an integrated probabilistic model to predict protein functions based on their physical and genetic interconnections, in addition to gene expression networks and known protein complexes. Xu and Li (Xu & Li, 2006) apply the k-nearest neighbor algorithm to

identify disease-related genes based on PPI network topology features. Ma et al. (Ma, Lee, Wang, & Sun, 2007) prioritize disease genes using a method based on Markov Random Field (MRF), applied on gene expression profiles and protein interaction datasets. Lage et al. (Lage et al., 2007) integrate phenotypic data and phenotypic similarities with a high-confidence human protein interaction network and use a Bayesian classifier to link previously-unknown protein complexes to diseases. Köhler et al. (Köhler, Bauer, Horn, & Robinson, 2008) apply a random walk with restart algorithm (RWR) on a heterogeneous interaction network to prioritize candidate disease genes. Li and Agarwal (Li & Agarwal, 2009) answer the gene prioritization problem by looking for disease relationships via literature mining to identify disease genes. Zhang et al. (Zhang, Li, Tai, Li, & Chen, 2012) classify genes based on various PPI network topology features. Guan et al. (Guan et al., 2012) create tissue-specific functional networks to prioritize disease genes. Li et al. (Li et al., 2014) introduce novel topological attributes and use support vector machines (SVM) to classify genes as disease-related or not. Chen et al. (Chen B. , Wang, Li, & Wu, 2014) combine biological data from multiple sources in order to prioritize disease genes; they develop a method based on the Markovian random field theory and Bayesian analysis. Many existing approaches have good performance. However, they mainly require initial setting of parameters and thresholds in addition to the dependence on a single kind of gene features, either topological or biological.

We present a learning model which classifies genes as disease-related or not, based on both topological and biological features. Given a list of genes, the goal is to maximize the contrast between disease and non-disease classes. Accordingly, we study the topology of the corresponding PPI network to find distinctive positioning of genes and we combine biological data from various sources to discover potential

similarities which characterize each class. Our proposed approach scores an area under the receiver operating characteristic (ROC) curve of 0.941 when applied using the Naïve Bayes classifier on a multiple disease dataset.

5.2 Gene Features

Recent advances in experimental technologies, in general, and in next-generation sequencing, in particular, offer great means for the identification of disease-related genes (Mardis, 2008). Large amounts of informative biological data can be easily generated nowadays. Nonetheless, fully perceiving various molecular processes still requires the assistance of computational techniques for the analysis and the integration of heterogeneous data. Based on the characteristics of reference disease genes, the goal is to shortlist other genes which could most likely be related to diseases, for further experimental explorations. In our study, we develop a model to classify genes as disease-related or not according to common PPI network topology attributes and various biological features shared by each type. We believe that by gathering computationally-conveyed network study and experimentally-generated biological information, we can enhance the gene-disease association process.

5.2.1 Topological Features

Mutations in interacting proteins often lead to similar phenotypes. Accordingly, PPI networks in which proteins and their interconnections are represented as nodes and edges respectively, significantly reflect the functional associations among genes (Xu & Li, 2006). Studying PPI networks to extract the topological features can greatly expedite gene-disease association tasks. In view of

that, given a set of genes to classify, we examine the corresponding PPI network and compute the topological features of the nodes, as described in Table 7.

Topological Features	Description
Degree	Number of edges that are adjacent to a node
Eccentricity	The distance from a node to the farthest node from it in the network
Closeness Centrality	The average distance from a node to all other nodes in the network
Betweenness Centrality	The number of times a node appears on shortest paths between nodes in the network
Authority	The value of the information stored at a node
Hub	The quality of a node's links
Modularity Class	The class reflecting how well a network decomposes into modular communities
PageRank	The rank of a node by its importance in the network
Component ID	The number of connected components to a node in the network
Clustering Coefficient	The completeness of the neighborhood of a node in the network
Number of triangles	The number of connected triangles including a certain node
Eigenvector Centrality	The importance of a node in the network based on its connections

Table 7: The topological features of genes and their definitions.

5.2.2 Biological Features

Various experimental observations can be viewed as sources of descriptive evidences that could potentially tell apart disease from non-disease genes (Piro & Di Cunto, 2012). Hence, the more attributes we include, the larger the potential contrast between gene classes. The biological features of genes considered in our study are presented hereafter.

1- Sequence Length: Previous studies show that disease genes tend to have longer sequences (Mushegian et a., 1997). In view of that, the number of amino acids in the canonical gene sequence is examined.

2- Gene Ontology (GO) Terms: The GO project (Ashburner et al., 2000) provides a set of hierarchically-controlled vocabulary that describes gene products in terms of their biological processes, molecular functions and cellular components. Biological processes cover the gene molecular events related to the functioning of integrated living units including cells, tissues, organs, and organisms. Molecular functions delineate the elemental activities of a gene product at the molecular level. Cellular components are the parts of a cell or its extracellular environment in which the gene resides.

3- Topological Domains: The topology and the compartments of proteins in the cell can potentially take part in disease or non-disease gene classification (Ibn-Salem et al., 2014). In particular, the topological domain information which describes the subcellular compartments where each non-membrane region of a membrane-spanning protein is found.

4- Chain: This feature shows the extent of a polypeptide chain in the mature protein following processing. It can also provide an insight on whether a protein is related to disease or not (Park & Park, 2015).

5- Domain: Defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold, protein domains usually correspond to structural domains which fold independently of the rest of the protein chain.

6- Protein Family: Under the assumption that proteins belonging to the same family share common evolutionary origins and thus exhibit similar functions (Wu, Huang, Yeh, & Barker, 2003); we take into account protein family groupings in our classification process.
7- Pathway: Considering the pathways in which genes participate could potentially direct the association of genes to diseases.

5.3 Experimental Study

5.3.1 Data Sources and Feature Collection

We refer to the paper by Goh et al. (Goh et al., 2007) to extract the gene-disease association data which reports 1777 genes linked to 1284 disorders split into 22 types. This data is originally derived from the Online Mendelian Inheritance in Man (OMIM) database (McKusick, 2007). We preprocess the dataset as reported in the paper by Chen et al. (Chen B. , Wang, Li, & Wu, 2014), namely, by removing the genes related to "multiple", "unclassified", "cancer", "neurological" diseases in addition to the disease types of less than 30 gene members. The PPI dataset is extracted from the Human Protein Reference Database (HPRD) (Prasad et al., 2009) which originally comprises 37039 edges. We use the PE method presented in (Zaki, Efimov, & Berengueres, 2013) to assess the reliability of protein interactions and clean the PPI data accordingly. As a result, the final learning dataset consists of 9228 genes out of which 839 are associated to diseases.

We use Gephi (Bastian, Heymann, & Jacomy, 2009), the interactive visualization and exploration platform, to study the PPI network and compute the topological features of the genes as listed in Table 7. Next, we consult the Universal Protein Resource (UniProt) (UniProt Consortium, 2014) to extract the biological

features of the genes. UniProt is a comprehensive resource that captures accurate and consistent information on proteins including, but not limited to, accepted biological ontologies, classifications and cross-references. Sequence length, chain and domain attributes are directly retrieved and added to the learning dataset. Some preprocessing is required for the rest of the biological characteristics which are multi-valued and comprise a large number of possible descriptions. Accordingly and since we are interested in identifying disease genes, we look for distinctive top feature values describing them. For instance, we look for the top 25 GO biological processes, molecular functions and cellular components associated to disease genes in the learning dataset, convert them to Boolean attributes and find the values

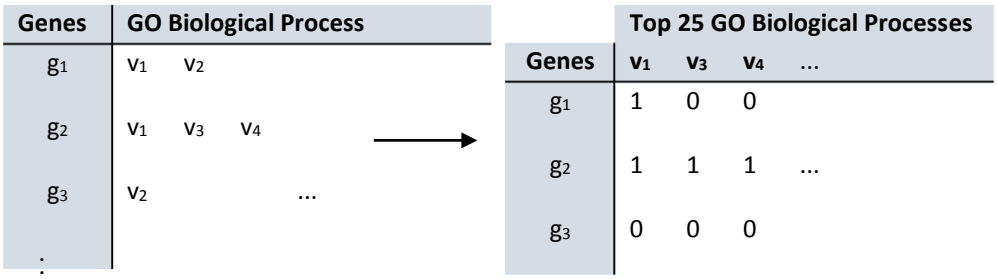

Figure 15: Conversion to Boolean attributes.

The same applies for GO molecular functions, GO cellular components, Protein Family and Pathway. Protein family information is based on PROSITE database of protein domains, families and functional sites (Sigrist et al., 2012). Pathway information is based on the Reactome Pathway Database (Croft et al., 2014). The number of Topological Domains associated to the genes under consideration is relatively lower. For this reason, we pick the top 3 domains and

convert them to Boolean features in the same manner; they are “cytoplasmic”, “lumenal” and “extracellular”. In total, we have 9228 genes with 142 features.

5.3.2 Classification Model and Results

After assembling the components of our learning dataset, we develop a classification model based on the Naïve Bayes classifier (John & Langley, 1995). We use the Weka data mining software (Hall et al., 2009). The generated results are based on default parameters of the Naïve Bayes classifier with a 10-fold cross-validation process. Table 8 and Table 9 represent the confusion matrix and the classification scores, respectively. The ROC curve is an essential indicator of the classification quality (Zweig & Campbell, 1993). It reflects the classifier's ability to distinguish between classes by plotting the true positive rate against the false positive rate across various thresholds. The ROC curve of our Naïve Bayes classification is presented in Figure 16. It corresponds to an area under curve (AUC) of 0.941.

Classified as	**Non-Disease Genes**	**Disease Genes**
Non-Disease Genes	**7858**	531
Disease Genes	149	**690**

Table 8: The confusion matrix showing the number of correctly-classified and the incorrectly-classified instances per class.

Class	**Precision**	**Recall**	**F-Measure**	**AUC**
Non-Disease Genes	0.981	0.937	0.959	0.941
Disease Genes	0.565	0.822	0.67	

Table 9: The classification scores of our gene classification model.

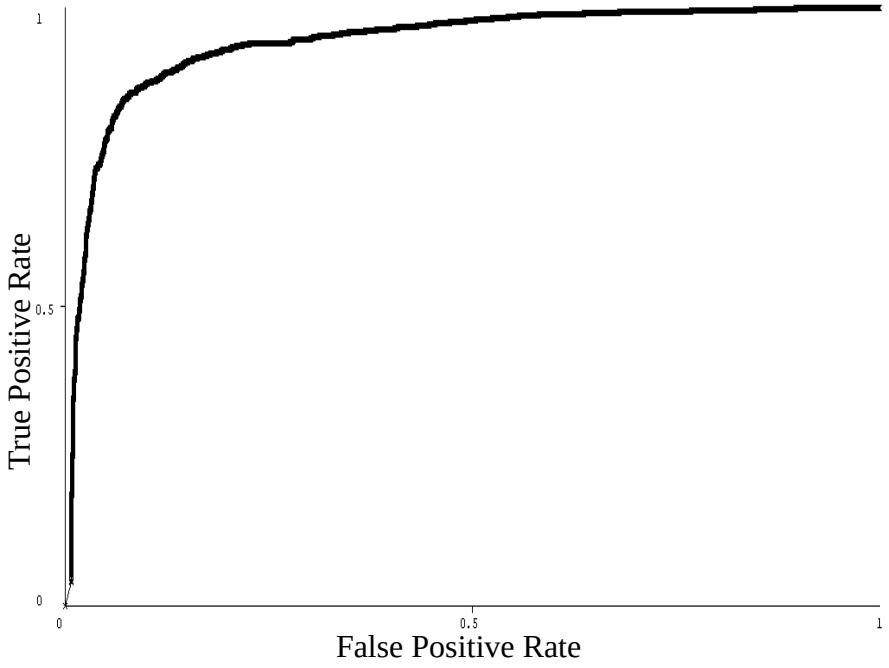

Figure 16: The ROC curve of our learning model, it corresponds to an AUC score of 0.941.

We compare the outcome of our approach to the experimental results presented in (Chen B. , Wang, Li, & Wu, 2014) in which several disease-gene identification methods are applied on the same gene-association dataset. Those methods include: $IMRF_2$ (Chen B. , Wang, Li, & Wu, 2014) which uses the theory of Markov Random Field (MRF) and Bayesian analysis to integrate data from various sources; MRF-Deng (Deng, Chen, & Sun, 2004) which is also based on MRF; the Random Walk with Restart (RWR) algorithm (Köhler, Bauer, Horn, & Robinson, 2008) proposed to identify disease genes by combining multiple PPI networks; and the method by Chen et al. (Chen et al., 2011) who define a Data Integration Rank (DIR) score to find key information in integrated data. DIR has the best performance when compared to previous approaches (Chen et al., 2011). The AUC score comparisons are presented in Table 10 and our approach clearly has better disease-gene association results.

	IMRF$_2$	MRF-Deng	RWR	DIR	Our Model
AUC	0.743	0.551	0.676	0.691	**0.941**

Table 10: AUC score comparison of our model with previous approaches.

5.4 Case Study: Diabetes Mellitus, Type II disease

In order to test the performance of our proposed model, we consider the case study of the Diabetes Mellitus, Type II disease which is a metabolic disorder marked by high blood sugar and a lack of insulin in the body (Chen, Magliano, & Zimmet, 2012). The occurrence of this disease is continuously growing making it one of the major healthcare challenges around the world. We consult the OMIM database (McKusick, 2007), and extract the genes associated to the Diabetes Mellitus, Type II (OMIM: 125853). Next, in terms of PPI data, we refer to HPRD (Prasad et al., 2009). We collect the corresponding topological and biological features of the genes, as described in our approach. Consequently, a learning dataset is formed; it consists of 9166 genes out of which only 23 are related to the Diabetes Mellitus, Type II disease. We consistently use a 10-fold cross-validation Naïve Bayes classifier to build the model. The resultant confusion matrix is shown in Table 11. It corresponds to an AUC score of 0.895.

Classified as	**Non- Diabetes Genes**	**Diabetes Genes**
Non-Diabetes Genes	**9027**	116
Diabetes Genes	8	**15**

Table 11: The confusion matrix showing the number of correctly-classified and the incorrectly-classified instances per class in our Diabetes Mellitus, Type II case study.

The attained results are in favor of the presented model which can identify 15 of the Type II Diabetes genes, equivalent to 65.2%. In comparison, our model has a

better performance than the method by Chen et al. (Chen, Wu, & Jiang, 2013) which predicts 13 Type II Diabetes genes.

5.5 Conclusion

Finding suitable methods for the identification of disease genes remains essential towards understanding how various disorders are formed and ultimately finding appropriate medical treatments. Our solution integrates topological features calculated based on PPI network analysis with various biological information/features of genes stored in multiple databases. Our experimental work verifies our initial hypothesis. By combining computationally-conveyed network study and experimentally-generated biological information, we can enhance the gene-disease association process.

Chapter 6: A Comprehensive Case Study: Breast Cancer

6.1 Background

In this chapter, we present a comprehensive case study on which we apply the approaches introduced in this dissertation. The main question that we would like to answer here is: by applying our gene-disease association model on a specific disease and given that it is able to reliably-identify the disease-related genes, can our complex-detection method, ProRank+, generate substantial groupings of those genes from a PPI network? To answer this question, we consider the breast cancer case study. This disease develops in breast tissues and it is actually the highest occurring cancer type in women worldwide (World Health Organization, 2014). Indeed, it is important to identify the key players as well as the cellular events which lead to the formation of this malady. We start by validating the reliability of our model in identifying the genes related to breast cancer. Once confirmed, we apply the ProRank+ algorithm (Hanna & Zaki, Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure, 2014) to potentially detect protein clusters that can be associated to breast cancer.

6.2 Identifying Genes Related to Breast Cancer

From the OMIM database (McKusick, 2007), we get the list of genes related to breast cancer (OMIM: 114480). We refer to the Human Protein Reference Database (HPRD) (Prasad et al., 2009) and download the up-to-date Human PPI dataset. Then, we compute the topological features and collect the biological attributes to form the corresponding learning data. It consists of 9167 genes out of which 23 breast cancer genes are covered. We use the Naïve Bayes classifier (John &

Langley, 1995) with a 10-fold cross-validation to generate the learning model. The resultant confusion matrix is presented in Table 12.

Classified as	Non- Breast Cancer Genes	Breast Cancer Genes
Non-Breast Cancer Genes	**8973**	171
Breast Cancer Genes	10	**13**

Table 12: The confusion matrix showing the number of correctly-classified and the incorrectly-classified instances per class in our breast cancer case study.

Although the learning dataset is unbalanced, i.e. the number of disease genes is very small in comparison with the number of non-disease genes; our model identifies 13 of the breast cancer genes (56.5%). In comparison, the protein complex-prioritization method by Chen et al. (Chen, Jacquemin, Zhang, & Jiang, 2014) ranks 6 breast cancer genes in the top ten complexes. Considering the classes unbalance and examining the results, we can infer that our model has a relatively good performance.

6.3 Detecting Protein Complexes using ProRank+

Considering the performance of our model, we then apply the ProRank+ algorithm (Hanna & Zaki, 2014) to potentially detect groupings of breast cancer genes in the same PPI network downloaded from HPRD. Various studies noted the fact that genes related to the same or similar diseases are often close to one another in a PPI network, for example (Oti, Snel, Huynen, & Brunner, 2006) and (Oti & Brunner, 2007). In view of that, we refer to the MimMiner (van Driel et a., 2006) tool which uses many text-mining algorithms to compute the similarities among phenotypes contained in the OMIM database (McKusick, 2007). Given a query disease, in our case breast cancer, MimMiner returns the related phenotypes along

with the similarity scores and the causal genes of each phenotype. We select the top 24 disorders similar to breast cancer and generate the set of associated genes to all of them. The considered disorders are listed in Table 13.

OMIM	OMIM Title	Similarity
114480	Breast Cancer	1.0000
176807	Prostate Cancer	0.5108
113705	Breast Cancer, Type 1	0.4996
120435	Colon Cancer, Familial Nonpolyposis, Type 1	0.4560
155720	Melanoma, Uveal	0.4402
151623	Li-Fraumeni Syndrome	0.4383
259500	Osteogenic Sarcoma	0.4205
278700	Xeroderma Pigmentosum, Complementation Group	0.4152
208900	Ataxia-Telangiectasia	0.4100
256700	Neuroblastoma	0.4093
102660	Adamantinoma Of Long Bones	0.4044
603737	Ovarian Germ Cell Cancer	0.4039
180200	Retinoblastoma	0.4030
260350	Pancreatic Carcinoma	0.3909
300068	Androgen Insensitivity Syndrome	0.3904
305700	Germinal Cell Aplasia	0.3877
273300	Testicular Tumors	0.3862
188550	Thyroid Carcinoma, Papillary	0.3853
211410	Breast Cancer, Ductal, 1	0.3845
139300	Gynecomastia, Hereditary	0.3834
158350	Cowden Disease	0.3822
210900	Bloom Syndrome	0.3794
194070	Wilms Tumor 1	0.3783
211980	Lung Cancer	0.3745
151410	Breakpoint Cluster Region	0.3732

Table 13: Top 24 disorders similar to breast cancer, given by MimMiner.

Genes in this set are used as seeds to protein complex-formation by ProRank+ when applied on the Human PPI dataset. Note that protein interactions are pruned using the PE method (Zaki, Efimov, & Berengueres, 2013). Since we are interested in complexes including as much disease genes as possible, we set the minimum complex size generated by ProRank+ to 5. The total number of detected

complexes is 113. Among those entities, 12 are shortlisted since the percentage of disease genes that they include is greater or equal to 30%. Those complexes are presented in Table 14, ordered by their decreasing percentage of breast cancer genes.

Complex No.	Breast Cancer Genes in the Detected Complex	Percentage of Breast Cancer Genes
1	BRCA1, BRCA2, ATM, TP53, RAD51	100%
2	BRCA1, TP53, MSH2, ATM, CHEK2	83.3%
3	BRCA1, BRCA2, TP53, RAD51	80%
4	BRCA1, TP53, MSH2, CHEK2	80%
5	BRCA1, TP53, XPA, ATM, RAD51, CHEK2	75%
6	BRCA1, TP53, RB1	60%
7	BRCA1, BRCA2, TP53, BARD1, ATM, RB1, AR, RAD51, CHEK2	52.9%
8	BRCA1, WT1, BRCA2, TP53, BARD1, ATM, PTEN, PPM1D, RAD51, CHEK2, BCR	37.9%
9	AR, HIP1, RB1	37.5%
10	EPHB2, EGFR, BCR	33.3%
11	BRCA1, EGFR, RB1, PTEN, AR, RNASEL	31.6%
12	BRCA1, AR, RB1	30%

Table 14: Groupings of genes associated to breast cancer and similar phenotypes, detected by ProRank+ and numbered by their decreasing percentage of disease genes that they contain.

Figure 17 displays three of those twelve complexes. They respectively correspond to complex numbers 8, 7 and 5; based on Table 14. In addition, we show the associations of the genes to the phenotypes in which they are involved.

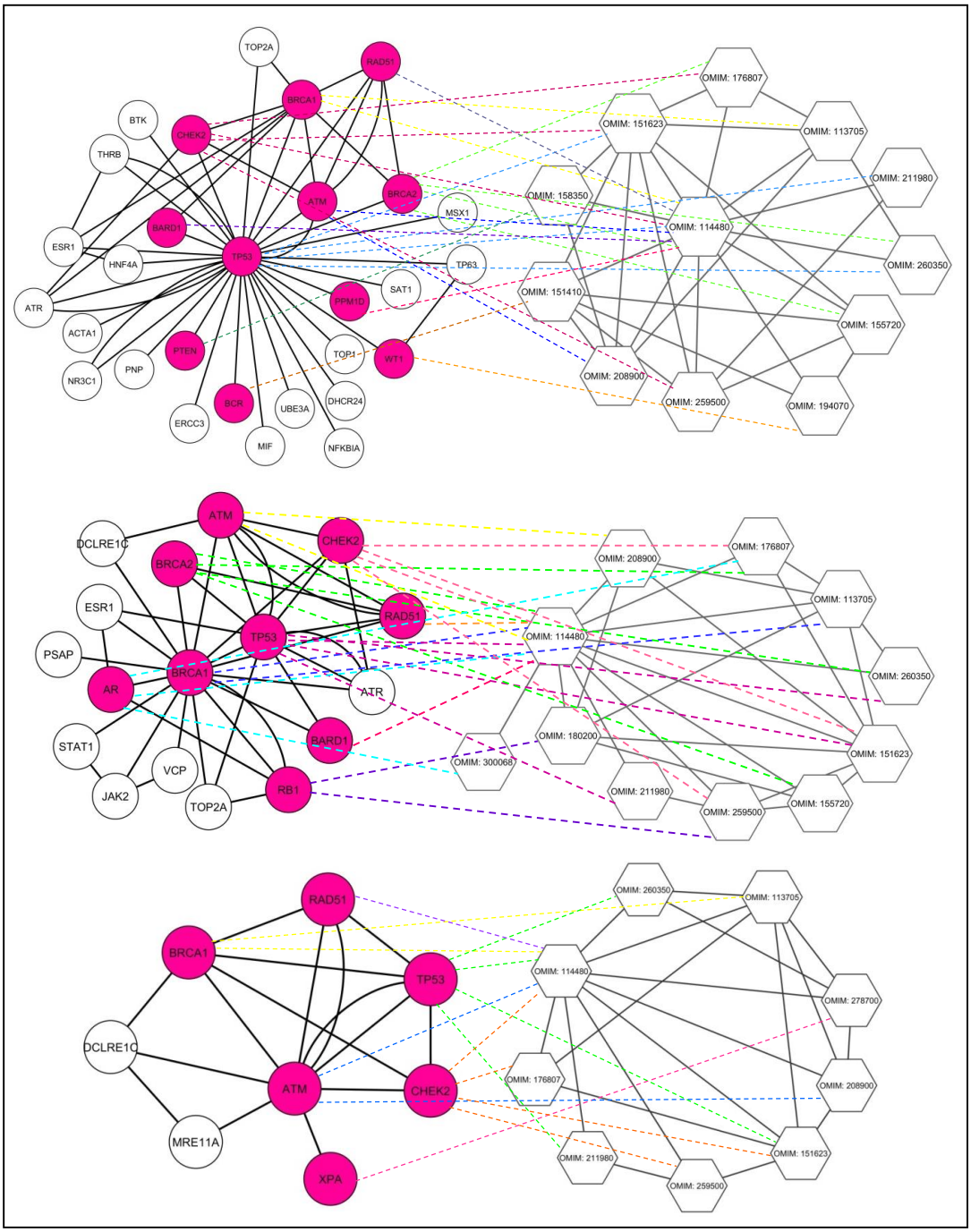

Figure 17: Detected groupings of proteins as detected by the ProRank+ algorithm. Circular nodes correspond to proteins among which the ones associated to breast cancer and similar phenotypes are colored. Hexagonal nodes correspond to phenotypes given by their OMIM numbers. Interactions among the proteins are based on the PPI dataset. Interconnections among phenotypes correspond to their similarities based on MimMiner. The dotted lines correspond to the association of disease genes to various phenotypes.

6.4 Conclusion

Considering the case of breast cancer, our learning model reliably identifies most of the genes related to this disease and the top 24 similar disorders. In view of that, we apply the ProRank+ algorithm and successfully detect groupings of those disease genes in the PPI network. Out of the 113 complexes generated by ProRank+, twelve have more than 30% of their protein members related to breast cancer and comparable diseases. Our results support the usability and the reliability of the presented solutions in this dissertation. Namely, given a set of proteins reported as disease-related, the detected entities by ProRank+ may highly-likely include proteins which can also be related to the considered disorders. Therefore, such proteins can be subject to further experimental examination.

Chapter 7: Conclusion

Looking at a protein-protein interaction network, a reliable computational approach can identify proteins and subsequently protein groupings that are possibly engaged in certain functions or phenotypes, for further experimental examinations. The proposed methodology in this dissertation is divided into three main contributions.

First, we present ProRank+, an effective method for the detection of protein complexes in PPI networks. It is based on a ranking algorithm which orders proteins by their importance in the network. It also applies a merging procedure to refine the detected complexes. In addition, our method accounts for the fact that a protein can participate in multiple cellular functions by belonging to several complexes. The method is tested on weighted and unweighted yeast datasets, as well as human PPI data. When compared to several state-of-the-art approaches, our algorithm is able to detect more complexes with better evaluation scores. Additional examinations and modeling of biological structures and properties of PPI networks and protein complexes could further improve the ProRank+ method.

Second, since protein interactions are usually subject to various temporal, spatial and contextual settings, we introduce a novel way to model the dynamic aspect of PPI networks. Genes which exhibit similar expression patterns across various conditions most likely interact. Hence, we apply biclustering techniques to analyze time-series gene expression data in order to group genes by their expression patterns across different subsets of conditions. Then, we detect protein complexes according to the generated groupings. In terms of experimental results, our framework allows the detection of more protein complexes with higher quality

scores. Our approach can be extended by integrating other biological data with PPI networks such as gene ontology annotations and subcellular localizations to better reproduce the dynamics of protein interaction networks.

Third, we present a classification model which integrates PPI network topology attributes and numerous biological features to identify disease-related genes. The experimental results validate our hypothesis that combining computationally-conveyed network study and experimentally-generated biological information can enhance the gene-disease association process. The approach identifies 65.2% of the genes related to the Diabetes Mellitus, Type II disease which is a better percentage than existing experiments. Based on the attained results, contributions from additional gene features may be examined. Moreover, the presented approach can be extended by exploring various diseases and ultimately linking the outcomes to drug design. Finally, we present a comprehensive study of breast cancer. Our learning model recognizes most of the genes associated to this disease and the top 24 related disorders. Then, we apply the ProRank+ algorithm to detect groupings of those genes in the PPI network.

The generated results are in favor of our approaches, they support their reliability and usability to analyze protein interaction networks and potentially discover previously-unknown biological facts. In terms of future research directions, our plan includes: (1) Integrating additional biological data and refining the modeling of PPI dynamics towards better detection of protein complexes; (2) Working closer to biology to potentially explore specific diseases using our approaches, in the direction of identifying and validating the associations of genes and protein complexes; (3) Computationally and biologically examining the extent at which various attributes used in our approaches influence the association of genes to

diseases; (4) Finally, developing comprehensive and flexible tools which allow the convenient use of the presented methods in this dissertation.

Bibliography

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073.

The History of Cancer. (2015). Retrieved June 14, 2015, from American Cancer Society: http://www.cancer.org/cancer/cancerbasics/thehistoryofcancer/index?sitearea

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., & Vicsek, T. (2006). CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics, 22*(8), 1021-1023.

Altshuler, D., Daly, M., & Kruglyak, L. (2000). Guilt by association. *Nature genetics, 26*(2), 135-138.

Ashburner, M., Ball, C. A., Blake, J. A., (...), & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25-29.

Asyali, M. H., Colak, D., Demirkaya, O., & Inan, M. S. (2006). Gene expression profile classification: a review. *Current Bioinformatics, 1*(1), 55-73.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics, 4*(1), 2.

Baldi, P., & Hatfield, G. W. (2002). *DNA microarrays and gene expression: from experiments to data analysis and modeling.* Cambridge University Press.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & Di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular systems biology, 3*(1).

Barkow, S., Bleuler, S., Prelić, A., Zimmermann, P., & Zitzler, E. (2006). BicAT: a biclustering analysis toolbox. *Bioinformatics, 22*(10), 1282-1283.

Barrett, T., Wilhite, S. E., Ledoux, P., (...), & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research, 41*(D1), D991-D995.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM, 8*, 361-362.

Ben-Dor, A., Chor, B., Karp, R., & Yakhini, Z. (2003). Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology, 10*(3-4), 373-384.

Bergmann, S., Ihmels, J., & Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical review, E 67*(3), 031902.

Berman, H. M., Westbrook, J., Feng, Z., (...), & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research, 28*(1), 235-242.

Bozdağ, D., Kumar, A. S., & Catalyurek, U. V. (2010). Comparative analysis of biclustering algorithms. *In Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology* (pp. 265-274). Niagara Falls, NY: ACM.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems, 30*(1), 107-117.

Bryan, K., & Leise, T. (2006). $25,000,000,000 eigenvector: The linear algebra behind Google. *Siam Review, 48*(3), 569-581.

Busygin, S., Prokopyev, O., & Pardalos, P. M. (2008). Biclustering in data mining. *Computers & Operations Research, 35*(9), 2964-2987.

Calderone, A., Castagnoli, L., & Cesareni, G. (2013). Mentha: a resource for browsing integrated protein-interaction networks. *Nature methods, 10*(8), 690-691.

Chen, B., Wang, J., Li, M., & Wu, F. X. (2014). Identifying disease genes by integrating multiple data sources. *BMC medical genomics, 7*(Suppl 2), S2.

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

Chen, L., Magliano, D. J., & Zimmet, P. Z. (2012). The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives. *Nature Reviews Endocrinology, 8*(4), 228-236.

Chen, Y., Jacquemin, T., Zhang, S., & Jiang, R. (2014). Prioritizing protein complexes implicated in human diseases by network optimization. *BMC systems biology, 8*(Suppl 1), S2.

Chen, Y., Wang, W., Zhou, Y., (...), & Li, J. (2011). In silico gene prioritization by integrating multiple data sources. *PloS one, 6*(6), e21137.

Chen, Y., Wu, X., & Jiang, R. (2013). Integrating human omics data to prioritize candidate genes. *BMC medical genomics, 6*(1), 57.

Cheng, Y., & Church. (2000). Biclustering of expression data. *Ismb, 8*, 93-103.

Chua, H. N., Ning, K., Sung, W. K., Leong, H. W., & Wong, L. (2008). Using indirect protein–protein interactions for protein complex prediction. *Journal of bioinformatics and computational biology*, *6*(03), 435-466.

Chua, H. N., Sung, W. K., & Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, *22*(13), 1623-1630.

Collins, M. O., & Choudhary, J. S. (2008). Mapping multiprotein complexes by affinity purification and mass spectrometry. *Current Opinion in Biotechnology*, *19*(4), 324-330.

Collins, S. R., Kemmeren, P., Zhao, X. C., (...), & Krogan, N. J. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & Cellular Proteomics*, *6*(3), 439-450.

Consortium, U. (2014). UniProt: a hub for protein information. *Nucleic Acids Research*, gku989.

Croft, D., Mundo, A. F., Haw, R., (...), & D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic acids research*, *42*(D1), D472-D477.

de Lichtenberg, U., Jensen, L. J., Brunak, S., & Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *science*, *307*(5710), 724-727.

De Neve, J. E., Christakis, N. A., Fowler, J. H., & Frey, B. S. (2012). Genes, economics, and happiness. *Journal of Neuroscience, Psychology, and Economics*, *5*(4), 193.

Deng, M., Chen, T., & Sun, F. (2004). An integrated probabilistic model for functional prediction of proteins. *Journal of Computational Biology*, *11*(2-3), 463-475.

Fields, S., & Song, O. K. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, *340*, 245-6.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, *315*(5814), 972-976.

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601-620.

Gavin, A. C., Aloy, P., Grandi, P., (...), & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, *440*(7084), 631-636.

Goh, K. I., Cusick, M. E., Valle, D., (...), & Barabási, A. L. (2007). The human disease network. *Proceedings of the National Academy of Sciences, 104*(21), 8685-8690.

Guan, Y., Gorenshteyn, D., Burmeister, M., (...), & Troyanskaya, O. G. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS computational biology, 8*(9), e1002694.

Hall, M., Frank, E., Holmes, G., (...), & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Han, J. D., Goldberg, D. S., Berriz, G. F., (...), & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature, 430*(6995), 88-93.

Hanna, E. M. (2013). Detection of overlapping protein complexes using a protein ranking algorithm. *In Proceedings of the 2013 9th International Conference on Innovations in Information Technology (IIT)* (pp. pp. 233-236). Al Ain: IEEE.

Hanna, E. M., & Zaki, N. (2014). Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. *BMC bioinformatics, 15:204*.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*, 100-108.

Hodgkin, J. (1998). Seven types of pleiotropy. *International Journal of Developmental Biology, 42*, 501-505.

Hong, E. L., Balakrishnan, R., Dong, Q., (...), & Cherry, J. M. (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic acids research, 36*(suppl 1), D577-D581.

Ibn-Salem, J., Köhler, S., Love, M. I., (...), & Robinson, P. N. (2014). Deletions of chromosomal regulatory boundaries are associated with congenital. *Genome Biology, 15*(243).

Ishii, H., & Tempo, R. (2010). Distributed randomized algorithms for the PageRank computation. *Automatic Control, IEEE Transactions on, 55*(9), 1987-2002.

Jensen, L. J., Kuhn, M., Stark, M., (...), & von Mering, C. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research, 37*(suppl 1), D412-D416.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on, 16*(11), 1370-1386.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Montreal: Morgan Kaufmann Publishers Inc.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502.

Kim, T. H., & Ren, B. (2006). Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet, 7*, 81-102.

Kim, Y., Han, S., Choi, S., & Hwang, D. (2014). Inference of dynamic networks using time-course data. *Briefings in bioinformatics, 15*(2), 212-228.

King, A. D., Pržulj, N., & Jurisica, I. (2004). Protein complex prediction via cost-based clustering. *Bioinformatics, 20*(17), 3013-3020.

Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics, 82*(4), 949-958.

Kopp, F., Dahlmann, B., & Kuehn, L. (2001). Reconstitution of hybrid proteasomes from purified PA700–20 S complexes and PA28αβ activator: ultrastructure and peptidase activities. *Journal of molecular biology, 313*(3), 465-471.

Kritikos, G. D., Moschopoulos, C., Vazirgiannis, M., & Kossida, S. (2011). Noise reduction in protein-protein interaction graphs by the implementation of a novel weighting scheme. *BMC bioinformatics, 12*(1), 239.

Krogan, N. J., Cagney, G., Yu, H., (...), & Gerstein, M. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature, 440*(7084), 637-643.

Kuang, R., Weston, J., Noble, W. S., & Leslie, C. (2005). Motif-based protein ranking by network propagation. *Bioinformatics, 21*(19), 3711-3718.

Lage, K., Karlberg, E. O., Størling, Z. M., (...), & Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology, 25*(3), 309-316.

Langville, A. N., & Meyer, C. D. (2011). *Google's PageRank and beyond: the science of search engine rankings.* Princeton University Press.

Lee, Y. H., Tan, H. T., & Chung, M. (2010). Subcellular fractionation methods and strategies for proteomics. *Proteomics, 10*(22), 3935-3956.

Levy, E. D., & Pereira-Leal, J. B. (2008). Evolution and dynamics of protein interactions and networks. *Current opinion in structural biology, 18*(3), 349-357.

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, Y., & Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PloS one, 4*(2), e4346.

Li, Z. C., Lai, Y. H., Chen, L. L., (...), & Zou, X. Y. (2014). Identifying and prioritizing disease-related genes based on the network topological features. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1844*(12), 2214-2221.

Liou, A. K., & Willison, K. R. (1997). Elucidation of the subunit orientation in CCT (chaperonin containing TCP1) from the subunit composition of CCT micro-complexes. *The EMBO journal, 16*(14), 4311-4316.

Liu, G., Wong, L., & Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics, 25*(15), 1891-1897.

Lodish, H., Berk, A., Kaiser, C., (...), & Scott, M. P. (2013). *Molecular Cell Biology.* W. H. Freeman and Company.

Lovén, J., Orlando, D. A., Sigova, A. A., (...), & Young, R. A. (2012). Revisiting global gene expression analysis. *Cell, 151*(3), 476-482.

Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins: Structure, Function, and Bioinformatics, 64*(4), 948-959.

Lyubomirsky, S. (2008). *The how of happiness: A scientific approach to getting the life you want.* Penguin.

Ma, X., Lee, H., Wang, L., & Sun, F. (2007). CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data. *Bioinformatics, 23*(2), 215-221.

Macropol, K., Can, T., & Singh, A. K. (2009). RRW: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC bioinformatics, 10*(1), 283.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 1*(1), 24-45.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics, 24*(3), 133-141.

Marto, J. A. (2009). Protein complexes: the forest and the trees. *Expert Rev. Proteomics, 6*(1), 5-10.

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *The American Journal of Human Genetics, 80*(4), 588-604.

Mewes, H. W., Frishman, D., Gruber, C., (...), & Weil, B. (2000). MIPS: a database for genomes and protein sequences. *Nucleic acids research, 28*(1), 37-40.

Mewes, H. W., Frishman, D., Mayer, K. F., (...), & Stümpflen, V. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic acids research, 34*(suppl 1), D169-D172.

Miki, Y., Swensen, J., Shattuck-Eidens, D., (...), & Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science, 266*(5182), 66-71.

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., & Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry, 49*(15), 3216-3224.

Moschopoulos, C. N., Pavlopoulos, G. A., Schneider, R., Likothanassis, S. D., & Kossida, S. (2009). GIBA: a clustering tool for detecting protein complexes. *BMC bioinformatics, 10*(Suppl 6), S11.

Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2009). A novel coherence measure for discovering scaling biclusters from gene expression data. *Journal of bioinformatics and computational biology, 7*(05), 853-868.

Mushegian, A. R., Bassett, D. E., Boguski, M. S., Bork, P., & Koonin, E. V. (1997). Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proceedings of the National Academy of Sciences, 94*(11), 5831-5836.

Nakao, A., Yoshihama, M., & Kenmochi, N. (2004). RPG: the ribosomal protein gene database. *Nucleic acids research, 32*(suppl 1), D168-D170.

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods, 9*(5), 471-472.

Newman, J., Brändén, C. I., & Jones, T. A. (1993). Structure determination and refinement of ribulose 1, 5-bisphosphate carboxylase/oxygenase from Synechococcus PCC6301. *Acta Crystallographica Section D: Biological Crystallography, 49*(6), 548-560.

O'Connor, C. M., Adams, J. U., & Fairman, J. (2010). *Essentials of cell biology.* Cambridge: NPG Education.

Oghabian, A., Kilpinen, S., Hautaniemi, S., & Czeizler, E. (2014). Biclustering Methods: Biological Relevance and Application in Gene Expression Analysis. *PloS one, 9*(3), e90801.

Oswald, A. J., & Proto, E. (2013). *National Happiness and Genetic Distance.* Warwick University, mimeo.

Oti, M., & Brunner, H. G. (2007). The modular nature of genetic diseases. *Clinical genetics, 71*(1), 1-11.

Oti, M., Snel, B., Huynen, M. A., & Brunner, H. G. (2006). Predicting disease genes using protein–protein interactions. *Journal of medical genet, 43*(8), 691-698.

Park, S., Yang, J. S., Shin, Y. E., (...), & Kim, S. (2011). Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Molecular systems biology, 7*(1).

Park, W. J., & Park, J. W. (2015). The effect of altered sphingolipid acyl chain length on various disease models. *Biological chemistry, 396*(6-7), 693-705.

Piro, R. M., & Di Cunto, F. (2012). Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS Journal, 279*(5), 678-696.

Prasad, T. K., Goel, R., Kandasamy, K., (...), & Pandey, A. (2009). Human protein reference database—2009 update. *Nucleic acids research, 37*(suppl 1), D767-D772.

Pržulj, N., Wigle, D. A., & Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics, 20*(3), 340-348.

Przytycka, T. M., Singh, M., & Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Briefings in bioinformatics*, bbp057.

Pu, S., Wong, J., Turner, B., Cho, E., & Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic acids research, 37*(3), 825-831.

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). Sickle-cell disease. *The Lancet, 376*(9757), 2018-2031.

Reguly, T., Breitkreutz, A., Boucher, L., (...), & Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of biology, 5*(4), 11.

Remondini, D., O'connell, B., Intrator, N., (...), & Cooper, L. N. (2005). Targeting c-Myc-activated genes with a correlation method: detection of global changes in large gene expression network dynamics. *Proceedings of the National Academy of Sciences of the United States of America, 102*(19), 6902-6906.

Secrier, M., & Schneider, R. (2013). Visualizing time-related data in biology, a review. *Briefings in bioinformatics*, bbt021.

Shain, A. H., & Pollack, J. R. (2013). The spectrum of SWI/SNF mutations, ubiquitous in human cancers. *PloS one, 8*(1), e55119.

Sigrist, C. J., De Castro, E., Cerutti, L., (...), & Xenarios, I. (2012). New and continuing developments at PROSITE. *Nucleic acids research*, gks1067.

Song, L., Kolar, M., & Xing, E. P. (2009). KELLER: estimating time-varying interactions between genes. *Bioinformatics, 25*(12), i128-i136.

Spirin, V., & Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences, 100*(21), 12123-12128.

Stark, C., Breitkreutz, B. J., Reguly, T., (...), & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research, 34*(suppl 1), D535-D539.

Tu, B. P., Kudlicki, A., Rowicka, M., & McKnight, S. L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science, 310*(5751), 1152-1158.

UniProt Consortium. (2014). UniProt: a hub for protein information. *Nucleic Acids Research*, gku989.

Van Dongen, S. M. (2001). *Graph clustering by flow simulation.*

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., & Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *European journal of human genetics, 14*(5), 535-542.

Von Mering, C., Krause, R., Snel, B., (...), & Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature, 417*(6887), 399-403.

Wang, J., Peng, X., Li, M., Luo, Y., & Pan, Y. (2011). Active protein interaction network and its application on protein complex detection. *In Bioinformatics*

and Biomedicine (BIBM), 2011 IEEE International Conference on (pp. 37-42). Atlanta: IEEE.

Wang, J., Peng, X., Xiao, Q., Li, M., & Pan, Y. (2013). An effective method for refining predicted protein complexes based on protein activity and the mechanism of protein complex formation. *BMC systems biology, 7*(1), 28.

Watson, J. D., & Berry, A. (2009). *DNA: The secret of life.* Knopf.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. *Nature, 171*(4356), 737-738.

Wikimedia Commons. (2008). Protein structure, from primary to quaternary structure.

World Health Organization. (2014). *World Cancer Report 2014.*

World Health Organization IARC. (2014). *IARC monographs on the evaluation of carcinogenic risks to humans.* Lyon, France.

Wu, C. H., Huang, H., Yeh, L. S., & Barker, W. C. (2003). Protein family classification and functional annotation. *Computational Biology and Chemistry*, *27*(1), 37-47.

Wu, X., Jiang, R., Zhang, M. Q., & Li, S. (2008). Network-based global inference of human disease genes. *Molecular systems biology*, *4*(1).

Xenarios, I., Salwinski, L., Duan, X. J., (...), & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research, 30*(1), 303-305.

Xu, B., Lin, H., & Yang, Z. (2010). Ontology integration to identify protein complex in protein interaction networks. *In BIBM*, (pp. 290-295).

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

Zaki, N., & Mora, A. (2014). A comparative analysis of computational approaches and algorithms for protein subcomplex identification. *Scientific reports, 4*.

Zaki, N., Berengueres, J., & Efimov, D. (2012a). Detection of protein complexes using a protein ranking algorithm. *Proteins: Structure, Function, and Bioinformatics, 80*(10), 2459-2468.

Zaki, N., Berengueres, J., & Efimov, D. (2012b). ProRank: a method for detecting protein complexes. *In Proceedings of the 14th annual conference on Genetic and evolutionary computation* (pp. 209-216). Philadelphia: ACM.

Zaki, N., Efimov, D., & Berengueres, J. (2013). Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC bioinformatics, 14*(1), 163.

Zhang, L., Li, X., Tai, J., Li, W., & Chen, L. (2012). Predicting candidate genes based on combined network topological features: a case study in coronary artery disease. *PloS one, 7*(6), e39542.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry, 39*(4), 561-577.

List of Publications

Conference Papers

Hanna, E. M., & Zaki, N. (2015). Gene-disease association through topological and biological feature integration. *Accepted for presentation at the 11th International Conference on Innovations in Information Technology (IIT'15), Dubai, UAE*.

Hanna, E. M., & Zaki, N. (2015). Detecting Protein Complexes using Gene Expression Biclusters. *In Proceedings of the 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2015), Niagara Falls, Canada*.

Hanna, E. M., & Zaki, N. (2014). Dynamic protein-protein interaction networks and the detection of protein complexes: an overview. In *Proceedings of the 14th International Conference on Bioinformatics & Computational Biology (BIOCOMP'14)*, (p. 1), *Las Vegas, USA.*

Hanna, E. M., & Zaki, N. M. (2014). ProRank+: A Method for Detecting Protein Complexes in Protein Interaction Networks. *In Proceedings of the 5th International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS'14)*, (pp. 239 – 244), *Angers, Loire Valley, France*.

Hanna, E. M. (2013). Detection of overlapping protein complexes using a protein ranking algorithm. *In Proceeding of the 9th International Conference on Innovations in Information Technology (IIT'13)*, (pp. 233 – 236), *Al Ain, Abu Dhabi, UAE*.

Journal Papers

Hanna, E. M., Zaki, N., & Amin, A. (2015). Detecting Protein Complexes in Protein Interaction Networks Modeled as Gene Expression Biclusters. *Under Revision.*

Bouktif, S., Hanna, E. M., Zaki, N., & Abu Khousa, E. (2014). Ant Colony Optimization Algorithm for Interpretable Bayesian Classifiers Combination: Application to Medical Predictions. *PLoS one, 9(2),* e86456 (ISI IF: 3.730).

Hanna, E. M., & Zaki, N. (2014). Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. *BMC Bioinformatics, 15(1),* 204 (ISI IF: 3.024).