MINING SIGNIFICANT PATTERNS BY INTEGRATING BIOLOGICAL INTERACTION NETWORKS WITH GENE PROFILES

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Rami Mohammed Alroobi

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Computer Science

July 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

MINING SIGNIFICANT PATTERNS BY INTEGRATING BIOLOGICAL INTERACTION NETWORKS WITH GENE PROFILES

By

Rami Mohammed Alroobi

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Prof. Saeed Salem
Chair

Prof. William Perrizo

Prof. Simone Ludwig

Prof. Mukhlesur Rahman

Approved:

07/09/2015
Date

Prof. Brian Slator
Department Chair

ABSTRACT

Nowadays, large amounts of high-throughput data are available. Automatic with classical cell biology techniques which are employed in the analysis of cell functions, interactions, and how pathogens can exploit them in disease, are becoming available because of the huge advancements in both Genomics and Proteomics technologies. Analyzing and studying these vast amounts of data will enable researchers to uncover, clarify, and explain some aspects of gene products behavior and characteristics under a very diverse set of conditions. The biological data belong to different types. The integration of several types of data can help reduce the effect of problems each data source has. The focus or our work and among the very important tasks in the bioinformatics field are functional module discovery and discriminative pattern. In functional module discovery, the goal is to find groups of genes that interact to perform different processes in the living organism. Discriminative patterns mining aims at discovering groups of proteins that can be classified as related to a specific phenotype. Understanding what genes, or proteins, are involved in biological phenomena can lead to advancements in related medical and pharmaceutical research. Many research has be done in this area. The two main sources of data used in my work are the gene expression and the protein-protein interaction network. The expression data shows how genes react in several conditions. The interaction network represents real protein cooperations occurring in the living cell. Our research efforts proved to show competitive performance with well established methods as illustrated in this document.

ACKNOWLEDGMENTS

It is a pleasure to express my gratitude to the many people who were abundantly helpful and offered invaluable assistance and made this dissertation possible. It is difficult to exaggerate my gratitude to my Ph.D. advisor, Dr. Saeed Salem; for his encouragement, guidance and his efforts to explain concepts clearly and simply. Special thanks to my supervisory committee, Dr. William Perrizo, Dr. Simone Ludwig and Dr. Mukhlesur Rahman for their support, guidance and helpful suggestions. Without their comments and assistance this dissertation would not have reached this level of accomplishment. The deepest gratitude are due to my beloved parents, Mohammed and Ihsan, for their praying, endless love , care, and support throughout my entire life. This achievement is simply impossible without them. I owe them , and I will continue to owe them for the rest of my life, for every beautiful thing in my life and I wish I could show them how much I love and appreciate them. I would like to thank my wife, Shaymaa, for her encouragement, tolerance, and patience that helped me to continue, overcome obstacles, and finish the doctoral trip. She is my beloved wife and I will just give her a heartfelt thanks. In Addition, my deepest loving feelings go to my three sons, Mohammed, Hasan, and Osama for the lots of happiness and joyful times they brought to my life. Especially during the hardships I faced during this trip. With their smiles and playful nature, I felt always that I am in a different more joyful and less stressful place. I thank my friends, for their faith in me and supporting me in several ways to be as ambitious as I wanted, for helping me defeat hard times, and for all the emotional support, entertainment, and caring they provided. Lastly, I wish to thank my entire family, especially my sisters, for being a constant source of encouragement during my graduate study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

We start this work by the explanation of the data used and the integration efforts of multiple data types, especially gene expression data and interaction networks. After that, a description of research done on discovering functional modules. Then, an introduction of mining discriminative patterns. Lastly, the organization of this dissertation based on the published work.

1.1. Background

Currently, the experiments the biologists are designing and performing every day are producing and accumulating very large amounts of data. This data has to be analyzed and inspected to discover the hidden knowledge concealed within. While traditional analysis and manual interpretation methods can no longer cope with this task and will fall short, the need to employ computing technology to unravel biology's secrets is inevitable. This led to the rise of the bioinformatics field.

As a result of the huge advancements in both Genomics and Proteomics technologies, automatic with classical cell biology techniques are becoming available to be employed in the analysis of cell functions and interactions. Genomics technologies in the fields of DNA sequencing, sequence assembly, and annotation; where biological information is attached to the sequences. In Proteomics technologies, the main focus is about identifying and characterizing proteins, discovering protein structure, discovering proteins functions, and protein-protein interactions.

Analyzing and studying these vast amounts of data will enable researchers to uncover, clarify, and explain some aspects of gene products behavior and characteristics under a very diverse set of conditions. Furthermore, this huge amount of data needs well designed tools and techniques to analyze and understand. Many of these are in the field of computer science. In fact computer science helps in tackling many biological tasks. Tasks such as, comparing sequences, constructing evolutionary trees, detecting patterns in sequences, determining 3D structures from sequences, inferring cell regulation, determining protein function and metabolic pathways, just to name some. Here, scientists create algorithms for biological applications and check their complexity and performance. Currently, many biological databases are available to be accessed via Internet. Databases containing DNA sequences, 3D protein structures, interaction networks,

gene expression datasets, and software tools, that can be used to perform several tasks, are available for researchers. A good listing can be found in [1]. In this work we depend mainly on two sources of data; gene expression data and protein-protein interaction, PPI, networks data.

1.1.1. Gene Expression

The interesting aspect of gene expression data is that it helps researchers, simultaneously, monitor the expression levels of thousands of genes during important biological processes and across collections of related samples [2]. Therefore studying gene expression levels can give deep insights about how the gene reacts in response to different conditions. Gene expression represents the RiboNucleic Acid, RNA, that is produced by the gene. The different amounts of RNA that are produced by multiple genes can give an estimation about protein levels. Gene expression levels, which are often represented numerically, can vary depending on the state the living cell is in. In high-throughput cell biology, one of the important techniques to study the cell mechanics is through gene expression data. A key method for obtaining gene expression data is via microarray. Microarray technologies made it possible to observe the expression levels of tens of thousands of genes in parallel.

1.1.2. Protein Interaction Network

According to [3], a PPI is the physical contact between proteins that occur in a cell or in a living organism in vivo. From all of these protein interactions a PPI network is built. The goal here is to create a network of the interacting proteins, or genes. The nodes of the network are the genes and the edges are the bindings. The protein interactions considered to build the PPI network are the stable interactions which constitute macromolecular protein complexes and cellular machines. PPI networks represent bindings and cooperations occurring in the cell to perform biological functions. As a result, PPI networks have attracted many research efforts. The motivation is that studying these networks will illuminate some information about the systems they represent [4]. A group of genes that are well connected in the PPI network are likely to share similar functions [5]. A gene who is interacting with many other genes in the network in a hub-like behavior could be regarded as very important in biological perspectives on the contrary to a weakly connected gene [6].

1.1.3. Integrating Gene Expression and Protein Networks

In bioinformatics research a lot of work has focused on studying gene expression data and interaction data separately [7, 8]. However, there are some issues that arose from this separation. Gene expression data can contain errors and biases depending on the technologies used. Moreover, expression levels can be affected by experimental conditions. In addition, if two genes are found to be coexpressed that does not mean that they share similar functionality. Also, if they are not coexpressed, that does not mean that they are not functionally similar. Therefore, extracting significant patterns from gene expression alone is still problematic. On the other side, studying PPI networks is illuminating and can lead to interesting biological conclusions but also this should be done with caution. The reason is that PPI networks are still not complete. Some estimates say that available PPI networks of the yeast, which is a well studied organism, and for the human are around 50% and 10% respectively of the complete networks [9]. Consequently, much effort has been invested to integrate different types of available data. The aim is to overcome the problems each type of data has so that deeper insights into the biology of an organism can be captured and researchers gain a closer look of what activities are taking place in multiple biological processes.

In the following we explore two of the main research areas in the field of systems biology which are relevant to our work:

1.2. Discovering Functional Modules

Our work in this field appears in Chapter 2. However, the explanation to follow is necessary to set the stage for what we introduce in another part of this dissertation. Functional Module Discovery is an important research area that has a leading role in incorporating multiple sources of information. Functional modules are typically defined as a group of cellular components and their interactions that can be attributed a specific biological function [10]. Discovering proteins functional modules from a PPI network can help in many venues of research such as understanding the mechanisms regulating cell life, in describing the evolutionary orthology signal [11], in predicting the biological functions of uncharacterized proteins, and, more importantly, for

therapeutic purposes. Many works have shown the interconnection between expression profile similarity and protein interactions [12, 13, 14, 15, 16]. This has motivated researchers towards employing different types of data, such as gene expression data and interaction data, to obtain better conclusions.

In systems biology, an important objective is to mine modules with well intra-connected genes from an interaction network. To improve the quality of discovered modules, other sources of data can be integrated. For instance, when gene profiles are integrated in the module discovery process, the goal becomes discovering connected modules whose participating genes have highly similar gene profiles, i.e. profiles homogeneity. Similarity between gene profiles is data-relevant and can have several meanings. When the profiles are gene expression, similar genes are coexpressed, or more accurately, highly correlated through multiple experimental and environmental conditions. When gene profiles are annotation-based, two genes have high profile similarity if they share a significant number of annotations.

1.2.1. Non-Integrative Approaches

Many of the existing algorithms for modules discovery are based only on one type of data. Some algorithms use only the topological structure of the network. Well known examples appear in the work of [17, 18, 19, 20]. An example is the Molecular COmplex DEtection, MCODE [21]. MCODE detects dense connected modules where every node in the module has to have a degree that exceeds a predefined limit. Another group of methods consider only clustering the gene expression data. K-Means and Statistical-Algorithmic Method for Bicluster Analysis, or SAMBA, [22] are among these. When solely using PPI structure, results will bear the problems and limitations the PPI has. Gene expression data also has the problems aforementioned.

1.2.2. Integrative Approaches

For the problems above, integrating interaction networks with gene expression/profiles data holds much promise [23]. Gene expression can help reduce the effect of the missing links that should be available in the network [24]. Moreover, based on the expression data that was used in the experiments, these modules can be related to specific kinds of biological conditions such

as diseases. The integration process has been greatly successful in discovering phenotype-specific modules [25]. Integrative modules discovery methods can be categorized into different categories. Here, we consider one categorization which is based on dense versus non-dense functional modules discovery.

- **Density Based**

  Many of the algorithmic approaches aim at mining all subnetworks of proteins with densities exceeding a predefined threshold and are homogeneous over the gene profiles space. Dense Module Enumeration, or DME, by [26] aims at enumerating dense modules from a weighted interaction network by integrating gene expression data with the interaction network. The discovered modules have to satisfy density and homogeneity constraints defined over a number of data sets. The method only reports locally maximal solutions of which direct supermodules violate the condition. The DME approach is criticized for being sensitive to noise in the gene expression data. In addition, the discovered modules must satisfy a density constraint. Another approach is the Densely Connected Biclustering, DECOB, algorithm [27]. The authors try to reduce the search space of modules by employing the loose anti-monotone property of the densely connected biclusters. The authors applied the algorithm on yeast and human gene expression data. The study showed that Gene Ontology, GO, specific clusters of modules have more accurate prediction abilities of functional relations that exist between genes in the detected modules.

- **Non-Density Based**

  Another category of clustering biological networks methods integrates the interaction network and the gene profiles via constructing a distance matrix between genes by combining their network-based distance and gene profiles distance. The Co-clustering algorithm is a pioneer method in this area [28]. The Co-clustering algorithm assigns a distance value between two genes based on their network-based distance and their gene-expression profiles distance. Once the distance matrix is computed, a hierarchical clustering is employed to extract the functional modules. In order to reduce the number of produced clusters, an

associated statistical measure is computed. This method, and other similar methods like [29], have the advantage of discovering modules with high node homogeneity in the profile space. However, these methods tend to produce modules with low density. Moreover, this approach requires the definition of a proper network-based similarity function and a careful way to combine the distances which can be complicated. Another example is the algorithm of Module Analysis via Topology of Interactions and Similarity SEts, MATISSE, [30]. Other works can be found in [31, 32, 33, 34]. In all the mentioned studies the genes in the interaction graph and the profile matrix have unique labels.

When the aim is to mine dense functional modules, a major challenge is that it is not clear how to choose a good density threshold. Choosing a low density value would result in a drastic increase in the computational complexity of the method and a significant increase in the number of modules that satisfy the density threshold. On the other hand, choosing a high density threshold may result in missing many of the important subnetworks. This is true because large percentage of the curated biological complexes for both Yeast and Human have low density. This can be referred to the fact that the current biological interaction networks are far from being complete and suffer from a high false-positive rate [9]. The work done in [35] aims to mine all the dense maximal cohesive subnetworks. A density threshold is enforced so the reported patterns are dense leading to the problems mentioned above when only dense modules are targeted. In our work [36], we extended mining functional modules to mine maximal cohesive subnetworks that show similar expression profiles in multiple datasets. We have introduced an algorithm to achieve this objective in [36] that overcomes the density problems in addition to introducing the the idea of inter-module cohesiveness. More details are to come.

1.3. Mining Discriminative Gene Patterns

The details of our contribution in this area is fully explained in Chapter 3. However, the following background highlights some aspects of this research area. Integrating protein-protein interaction data with gene expression for the task of discriminative biomarker discovery has recently gained more attention. Interestingly, subnetwork biomarkers have shown to be more classificatory

powerful than single-gene markers and have been found to be biologically meaningful. When subnetworks are mentioned we simply mean sets of connected genes. Moreover, these subnetworks have to show differential expression over the gene expression profiles. These subnetworks biomarkers can be considered as phenotype-specific and they can be employed for microarray gene expression classification. This clustering of genes may result in discovering new functions for some of the genes within the discovered pattern, or even discovering functions for genes that are not known to have specific roles inside the biological activities occurring in the organism. This occurs when a subnetwork can be proved to be in relation to a certain phenotype.

In [23], Chuang et al. proposed an approach for mining subnetwork biomarkers. This approach follows a greedy search that starts by growing a gene into a pattern, set of genes, by adding neighboring genes that maximizes an objective function. This work was the first to utilize idea of differentially expressed subnetworks. The authors found that subnetwroks have more predictive abilities for cancer metastasis than single genes.

Several other approaches have been proposed for finding better network markers [37, 38, 39]. In these studies, the aim was to find semi-densely connected components of genes that are differentially expressed in a way that can be employed to distinguish between the disease and no disease conditions. In [37], the authors goal was to mine subnetwork biomarkers by utilizing the idea of covering subnetworks. While the work in [38] aimed at enumerating all dense subnetworks that are dysregulated in subsets of samples. In [39] the authors average the genes expression values of a subnetwork to determine the activity of the subnetwork. This activity level is employed as a score for the biomarker subnetwork.

In the work done by Suthram et al.[40] the authors proposed an algorithm for integrating multiple disease gene expression datasets with modules extracted from the PPI network. The similarity between two diseases was based on modules dysregulation score in the disease data. Diseases were clustered based on these similarities. Their results show that diseases with high correlation share common drugs for treatment. In this direction, we introduce an algorithm for mining **D**ysregulated **P**henotype-**R**elated subnewtworks of genes, the **DPR**s. Here, the annotation

of the gene expression datasets is accomplished using the Unified Medical Language System, UMLS, approach. Our contribution in this venture is explained in Chapter 3.

1.4. Dissertation Overview

This document is formatted as paper based where main chapters are based on published work.

Chapter 2 revolves about introducing the problem of discovering maximal cohesive subnetworks and patterns. This work is based on the publication:

- Rami Alroobi, Syed Ahmed, and Saeed Salem. Mining maximal cohesive induced subnetworks and patterns by integrating biological networks with gene profile data. Journal of Interdisciplinary Sciences: Computational Life Sciences, 5(3):211-224, 2013.

- This work is an extension of the research in the publication:
  Saeed Salem, Rami Alroobi, Syed Ahmed, and Mohammad Hossain. Discovering maximal cohesive subgraphs and patterns from attributed biological networks. In Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), BIBM'12, pages 203-210, 2012.

In chapter 3, the main interest is to discover patterns of proteins that can discriminate between different phenotypic conditions. This work is based on the publication:

- Rami Alroobi and Saeed Salem. Discovering dysregulated phenotype-related gene patterns. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2014, pages 524-532, 2014.

In chapter 4, we summarize the contributions presented in this dissertation.

CHAPTER 2. MAXIMAL COHESIVE PATTERNS DISCOVERY[1]

The systematic integration of gene profile data with interaction data yields significant patterns. Here, we add our contribution that employs the integrative model by introducing the problem of mining maximal cohesive subnetworks that satisfy user-defined constraints defined over the gene profiles of the reported subnetworks. In addition, we introduce the problem of finding maximal cohesive patterns which are sets of cohesive genes. The quality of the discovred subnetworks was assessed by performing experiments on Yeast and Human datasets.

2.1. Contribution

The proposed algorithm in this work follows a pattern-growth approach to enumerate the set of maximal cohesive induced subgraphs that satisfy the user-defined constraints that are defined over the profiles of the genes in the subgraphs. Since there is no minimum density threshold, the proposed algorithm is able to discover cohesive subgraphs that have low density. Moreover, the algorithm can find maximal cohesive patterns that might not be connected. Afterwards, connected sub-components can be extracted, where these sub-components have kind of intra-cohesiveness. To summarize, we have made the following contributions in this work:

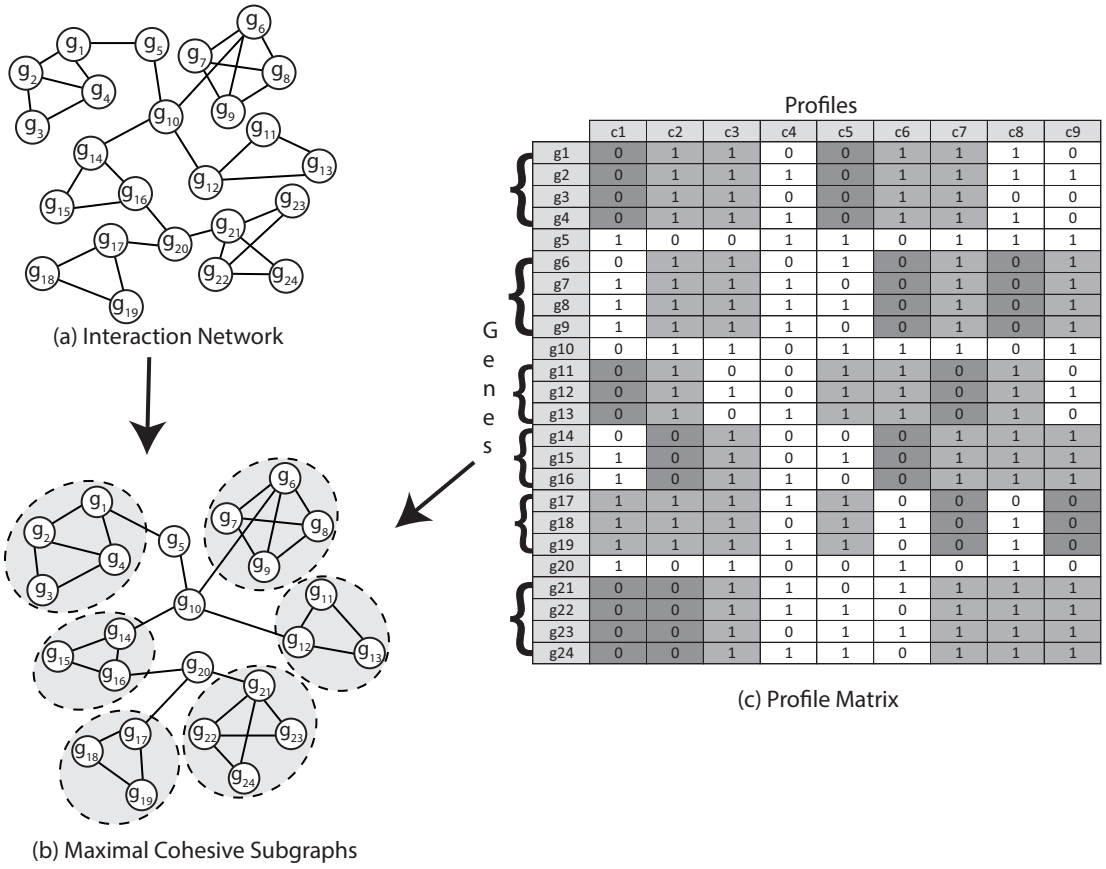
- We introduce and propose algorithms for the problem of mining maximal cohesive induced subgraphs and maximal cohesive patterns.
- We show that by effectively integrating constraints from additional data sources such as phenotypic and evolutionary profiles with protein interaction networks, the search can be guided to discover interesting patterns.
- We performed experimental analysis on Yeast and Human interaction networks with different profile datasets. The experimental results proved the effectiveness of the proposed approach by assessing the overlap of the discovered subnetworks with known biological complexes

[1]The material in this chapter was co-authored by Rami Alroobi, Syed Ahmed, and Saeed Salem. Rami Alroobi and Saeed Salem were responsible of developing the idea. Rami Alroobi prepared the data used, did the analysis, and drafted and revised the chapter. Syed Ahmed helped in preparing the chapter writing. In addition to contributing to the idea, Saeed Salem served as proofreader and checked the correctness of the mathematical formulation.

and pathways. Moreover, GO enrichment analysis show that the discovered subnetworks are biologically significant.

2.2. Problem Description

A protein interaction network is modeled as an undirected graph. In this section, we introduce some preliminary graph definitions that are used throughout this chapter. We then describe the problem of mining maximal cohesive induced subgraphs and the problem of mining maximal cohesive patterns.



	c1	c2	c3	c4	c5	c6	c7	c8	c9
g1	0	1	1	0	0	1	1	1	0
g2	0	1	1	1	0	1	1	1	1
g3	0	1	1	0	0	1	1	0	0
g4	0	1	1	1	0	1	1	1	0
g5	1	0	0	1	1	0	1	1	1
g6	0	1	1	0	1	0	1	0	1
g7	1	1	1	1	0	0	1	0	1
g8	1	1	1	1	1	0	1	0	1
g9	1	1	1	1	0	0	1	0	1
g10	0	1	1	0	1	1	1	0	1
g11	0	1	0	0	1	1	0	1	0
g12	0	1	1	0	1	1	0	1	1
g13	0	1	0	1	1	1	0	1	0
g14	0	0	1	0	0	0	1	1	1
g15	1	0	1	0	1	0	1	1	1
g16	1	0	1	1	0	0	1	1	1
g17	1	1	1	1	1	0	0	0	0
g18	1	1	1	0	1	1	0	1	0
g19	1	1	1	1	1	0	0	1	0
g20	1	0	1	0	0	1	0	1	0
g21	0	0	1	1	0	1	1	1	1
g22	0	0	1	1	1	0	1	1	1
g23	0	0	1	0	1	1	1	1	1
g24	0	0	1	1	1	0	1	1	1

Figure 1: An interaction network, the gene profiles data, and the maximal cohesive subgraphs.

2.2.1. Graphs and Gene Profiles

A graph $G = (\mathcal{V}, E)$, consists of a set of vertices $\mathcal{V} = \{g_1, g_2, \cdots, g_n\}$, and a set of edges $E \subseteq \mathcal{V} \times \mathcal{V}$. The size of a graph G, denoted $|G|$, is the cardinality of the edge set (i.e., $|G| = |E|$).

The vertex set and edge set of a graph G are denoted by $\mathcal{V}(G)$ and $\mathcal{E}(G)$, respectively. Here we give the definition of **Subgraphs and Induced S ubgraphs**. A graph $G' = (V', E')$ is a subgraph of $G = (\mathcal{V}, E)$, denoted as $G' \subseteq G$, if $V' \subseteq \mathcal{V}$ and $E' \subseteq E$. A subgraph $G' = (V', E')$ of G is said to be induced if for $x, y \in V'$, there is an edge between x and y in G' if and only if $(x, y) \in E(G)$. In other words, the set of edges in the induced subgraph, E', include all the edges in G whose endpoints are in V'. The subgraph G' induced by the vertex set V' is written as $G[V']$.

For **Gene Profiles**, given a set of genes $\mathcal{V} = \{g_1, g_2, ..., g_n\}$ and a set of conditions $\mathcal{C} = \{c_1, c_2, ..., c_m\}$, a Gene Profile Matrix $X \in \mathcal{R}^{n \times m}$, where $X = (x_{ij})_{i \in \mathcal{G}, j \in \mathcal{C}}$, is the gene attributes matrix such that x_{ij} is the attribute value of the i^{th} gene in j^{th} condition. Also the i^{th} row of the matrix X is the profile of the i^{th} gene.

2.2.2. Cohesive Constraints

We are interested in interacting genes that show high profile similarity. The profile similarity can be based on the co-expression of the profiles, or the number of common profiles that the genes have. In this work, we incorporate only anti-monotone constraints because they can be effectively integrated in the mining algorithm.

Definition 1 (Anti-monotone constraint)**.** *A constraint C is anti-monotone if for a set of genes, $V \subseteq \mathcal{V}$, the following condition is satisfied:*

$$C(V) = 1 \implies C(V') = 1 \,, \forall\, V' \subseteq V$$

This constraint is often referred to as the down-closure property in the pattern mining literature.

In the following we show how to employ the anti-monotone cohesive constraints with discrete profiles.

Here we explain **Constraints for Discrete Profiles**. Gene attribute data can take discrete values. For example the i^{th} entry in the gene profile can represents the dysregulation of the gene in the i^{th} condition. For example a value of 1 represents up-regulation, 2 for down-regulation, and 0 for no regulation. For discrete data we employ the following constraint [26]: Let $N_i(V)$ be the

number of attributes in which all the genes in V have the value i. Let α_i be the minimum number of attributes in which all the genes in V have a value of i. A cohesive constraint is defined by a set of user-defined thresholds, $\{\alpha_0, \alpha_1, ..., \alpha_{k-1}\}$, where k is the number of distinct values in the profile data.

Given a set of thresholds, a set of genes V is cohesive if for every distinct value, the number of common attributes that have a given value is higher than the threshold for the corresponding value. More formally, for a set of genes V: $C(V) = 1 \iff N_i(V) \geq \alpha_i, \forall\, i,\ 0 \leq i \leq k-1$,

Figure 1 shows a simplified interaction network and gene profile data. For a constraint thresholds, $\{\alpha_0, \alpha_1\} = \{2, 4\}$ which would enforce that a cohesive pattern has at least 2 attributed in which all the values are 0, and at least 4 attributes of values 1. The set of genes $V = \{g_1, g_2, g_3, g_4\}$ is cohesive because $N_0(V) = 2$ (c_1, c_5) and $N_1(V) = 4$ (c_2, c_3, c_6, c_7). Therefore, the subgraph induced by V, $G[V]$, is a cohesive subgraph.

2.2.3. Maximal Cohesive Induced Subgraphs

We will define and propose an algorithm for the problem of mining maximal cohesive induced subgraphs. In this work, we consider induced subgraph since the attribute are attached to nodes and we add the entire set of edges between a set of nodes. All the subgraphs that we mine are connected.

Definition 2 (Cohesive Induced Subgraph)**.** *An induced subgraph $G' = (V', E')$ is cohesive, if the set of genes of the subgraph, V', satisfy the user-defined cohesive constraint ($C(V') = 1$).*

Definition 3 (Maximal Cohesive Induced Subgraph)**.** *An induced subgraph is maximal cohesive if it is cohesive, connected, and none of its super-subgraphs is cohesive, i.e., G' is a maximal cohesive induced subgraph if $C(V(G')) = 1$ and $C(V(G)) = 0$ for all $G \supseteq G'$. In other words, if a cohesive connected induced subgraph cannot be grown by adding any neighboring vertex without violating the cohesive constraint, then the subgraph is maximal cohesive.*

Algorithm: Mining Maximal Cohesive Induced Subgraphs

Input:
$\mathcal{G}$:Interaction Network
$\mathcal{X}$:Profile Matrix
$\mathcal{A}$:Thresholds

Output:
$\mathcal{M}_{\mathcal{H}}$:List of Maximal Cohesive Induced Subgraphs (MCSs)

1. $V' \subseteq V$ ▷Set of cohesive single genes.
2. $\mathcal{S} = \emptyset$ ▷Set of visited subgraphs
3. $\mathcal{M}_{\mathcal{H}} = \emptyset$ ▷Set of MCSs
4. **for each** $g_i \in V'$:
5. $G_i \leftarrow Subgraph(g_i)$ ▷Subgraph that has a single gene
6. genMCS($G_i, \mathcal{A}, \mathcal{M}_{\mathcal{H}}, \mathcal{S}$)
7: **end for**
8. return $\mathcal{M}_{\mathcal{H}}$

Seed Extension

genMCS($G, \mathcal{A}, \mathcal{M}_{\mathcal{H}}, \mathcal{S}$)
1: **if** $G \in \mathcal{S}$
2: return
3: **else**
4: $\mathcal{S} = \mathcal{S} \cup G$
5: **end if**
6: **if** $\exists M \in \mathcal{M}_{\mathcal{H}}$: $[G \subseteq M$ and $\mathcal{X}(G)$ *Subsumed by* $\mathcal{X}(M)]$
7: return
8: maxFlag = TRUE
9: **for each** $g_i \in$ in $N(G)$:
10: G' = newSubgraph(G, g_i)
11: **if isCohesive**($G', \mathcal{A}$)
12: maxFlag = false
13: **genMCS**($G', \mathcal{A}, \mathcal{M}_{\mathcal{H}}, \mathcal{S}$)
14: **end if**
15: **end for**
16: if maxFlag is TRUE
17: $\mathcal{M}_{\mathcal{H}} = \mathcal{M}_{\mathcal{H}} \cup G$

Figure 2: Mining Maximal Cohesive Subgraphs.

Now we introduce the **Problem Definition** of this work. Given a graph G, a gene profile matrix X, and an anti-monotone constraint C, the problem of mining the set of **Maximal Cohesive Induced Subgraphs** is to find the set:

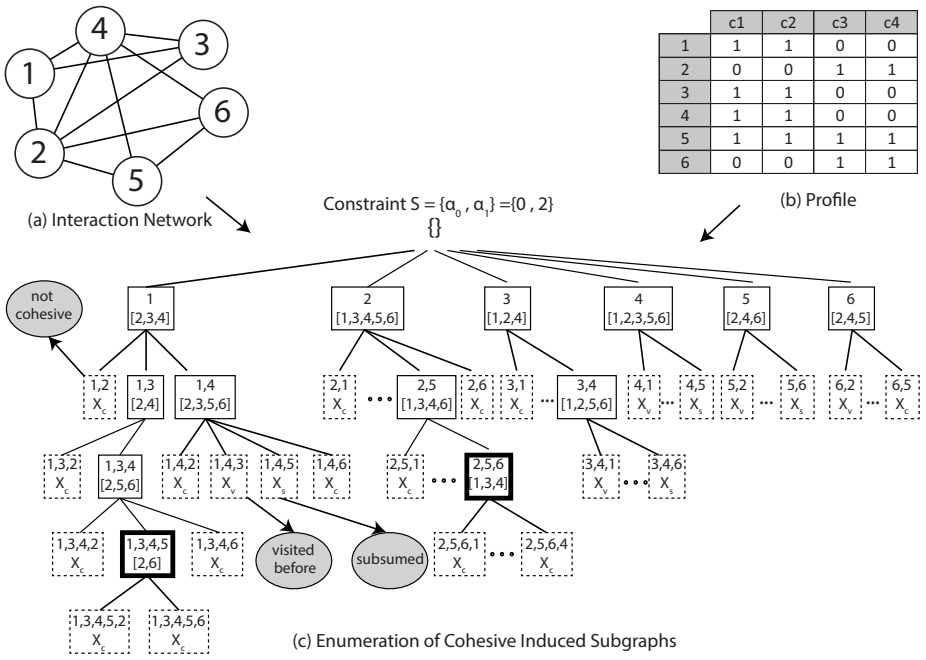
$$\mathcal{M}_{\mathcal{H}} = \{G_1, G_2, G_3, \cdots, G_{|\mathcal{M}_{\mathcal{H}}|}\}$$

such that every $G_i \in \mathcal{M}_{\mathcal{H}}$ is a maximal cohesive induced subgraph.

Figure 1 is an illustrating example of finding all maximal cohesive induced subgraphs from a simple interaction network consisting of 24 genes, and a profile matrix with nine conditions. Figure 1(a) is an interaction network, Figure 1(c) is the gene profile data, In this example, for a subgraph to be cohesive, $N_1(V(G')) \geq 4$, and $N_0(V(G')) \geq 2$.

The **Algorithm** of the this work is explained as following. We propose an algorithm for discovering the maximal cohesive induced subgraphs by integrating two types of data sources; the Protein-Protein Interaction (PPI) Network and a Gene Profiles Matrix. The proposed algorithm is shown in Figure 2. The algorithm takes as input the interaction network (G), profile matrix (X), and a set of thresholds, $\mathcal{A}$. The algorithm adopts a *depth first search* approach of the search space defined by all the cohesive induced subgraphs. We employ different pruning strategies to avoid visiting branches that will not result in cohesive patterns or that would result in redundant patterns. First the algorithm prunes the genes which are not cohesive (line 1) since they can not be in any cohesive pattern; this is important for discrete profile data when an individual gene does meet the constraint. Due to the anti-monotonicity of the cohesive constraint, pruning these genes will not result in any missing patterns because these genes cannot be in any cohesive patterns. There are two sets $\mathcal{S}$, and $\mathcal{M}_{\mathcal{H}}$ to maintain the sets of visited subgraphs and maximal cohesive subgraphs, respectively (lines 2 and 3). Then starting with every gene node as a seed, the algorithm recursively tries to extend the seed (line 6) by calling the genMCS procedure.

In this work we employ 3 pruning strategies. In **Pruning Strategy I**, the genMCS procedure (Seed Extension) starts by checking if the subgraph, G, has been seen before (line 1). if



	c1	c2	c3	c4
1	1	1	0	0
2	0	0	1	1
3	1	1	0	0
4	1	1	0	0
5	1	1	1	1
6	0	0	1	1

Figure 3: An example showing the proposed enumeration approach to discover maximal cohesive induced subgraphs. The example, also, illustrates the different pruning strategies employed. Here the cohesive thresholds are set to $\alpha_0 = 0$ and $\alpha_1 = 2$

yes, the procedure returns (line 2) because a similar subgraph has been visited before and thus the exploration of the the entire branch can be eliminated. If the pattern has not been seen before, the algorithm adds it to the visited set, *S* (line 4), and continues checking other conditions. Since the graph has unique nodes, the set *S* stores the canonical codes (representations) of the visited subgraphs. The canonical code of a subgraph is simply the string representation of the ordered labels (ids) of the nodes in the graph.

After that, **Pruning Strategy II** is executed. If the subgraph, *G*, is a subset of any of the already discovered maximal cohesive subgraphs, and the common profiles of the genes in this subgraph is subsumed by the profiles of the already discovered maximal cohesive subgraph, then we can safely prune the current node. If the current subgraph *G* passed these two conditions we extend the subgraph with neighboring nodes, line 10. In **Pruning Strategy III**, after extending the subgraph *G* with one of its neighbors, the algorithm checks if the newly generated subgraph (G') is cohesive (line 11). If the subgraph G' is cohesive, then the subgraph *G* is not maximal and the

algorithm recursively extends G'. If the subgraph G cannot be extended by any of its neighboring nodes while maintaining the constraint (cannot generate a cohesive supergraph), then the subgraph G is maximally cohesive and it is added to the set $\mathcal{M}_{\mathcal{H}}$ (line 17). At this point we do not have to check whether the subgraph G is already in the maximal set. The reason for this is that before the algorithm started extending the subgraph G, it checked if the same subgraph has been visited before and if it is subsumed by another maximal subgraph and none of the conditions was satisfied.

Figure 3 illustrates how the algorithm traverses the search space of the cohesive subgraphs. The discovered maximal cohesive patterns are inside the boxes with bold-line boundaries. We will try to highlight where the pruning strategies are employed to prune entire branches of the search space. The algorithm starts by extending the induced subgraph that has one node $G[\{1\}]$. This subgraph has not been seen before and also is not subsumed by any already discovered maximal subgraph. Therefore, we try extending it. The first neighbor to add is node 2 and we get a subgraph induced by the nodes $(1,2)$. The profiles of nodes 1 and 2 are not cohesive and thus we prune this subgraph, *i.e.*, $C(V) = 0$ for $V = \{1,2\}$. Therefore the branch is not explored further. We next extend the subgraph to get a subgraph with node set 1 and 3. The subgraph ($G[V]$, $V = \{1,3\}$) is cohesive and we recursively call the genMCS procedure. From this subgraph we get the supergraph with $V = \{1,3,4\}$ from which we get the supergraph with $V = \{1,3,4,5\}$. The subgraph with $V = \{1,3,4,5\}$ cannot be extended with any of its neighbors and thus we add it to the set of maximal subgraphs. So far, we have seen the effect of the cohesive pruning strategy. Next we will explain when **pruning strategy I** is employed. When the algorithm extend the subgraph ($V = \{1,4\}$ to get the supergraph ($V = \{1,4,3\}$, it checks in the set of visited subgraph and finds that the subgraph has been visited before, it then prunes the entire search branch. The subsume pruning strategy is data-dependent but we will discuss it in context of binary profile data. If the pattern being explored is a subset of an already discovered maximal cohesive pattern and the profile of this pattern is subsumed by the profile of a discovered maximal pattern, so this branch is pruned. For example, when the recursive procedure explores the subgraph ($V = \{3,4,5\}$), it finds that is a subset of an already discovered maximal subgraph ($V = \{1,3,4,5\}$). In addition, the common

profile of the subgraph $V = \{3,4,5\}$ which is $[1,1,0,0]$ is subsumed by the common profile of $V = \{1,3,4,5\}$ which is $[1,1,0,0]$. The reason we can safely prune a subsumed subgraph (say G is subsumed by G_m) is that maximal supergraph that could be generated from G cannot be a supergraph of G_m, otherwise, G_m cannot be maximal.

To summarize, we are employing three pruning strategies to reduce the search space for maximal cohesive induced subgraph:

1. If the subgraph G is not cohesive, i.e., $C(G) = 0$, the search branch is pruned. Nodes that are pruned by this strategy are denoted by X_c.

2. If the subgraph G has been seen before, i.e., $G \in \mathcal{S}$, then the search branch is pruned. Nodes that are pruned by this strategy are denoted by X_v.

3. If the subgraph G is a subsumed by any of the already discovered maximal cohesive induced subgraphs $\mathcal{M}_{\mathcal{H}}$, the search branch is pruned. Nodes that are pruned by this strategy are denoted by X_s.

2.2.4. Maximal Cohesive Patterns (MCPs)

The second problem we address in this paper is mining Maximal Cohesive Patterns (MCPs).

Definition 4 (Cohesive Pattern)**.** *Let $V \subseteq \mathcal{V}$ be a pattern consisting of a set of genes, $V = \{g_{i1}, g_{i2}, \cdots, g_{ik}\}$. We say that the pattern V is cohesive if all the genes satisfy a user-defined constraint C, i.e., $C(V) = 1$.*

Definition 5 (Maximal Cohesive Pattern)**.** *A pattern is a maximal cohesive pattern if it is cohesive and none of its super-patterns is cohesive, i.e., V is maximal cohesive pattern if $C(V) = 1$ and $C(V') = 0$ for all $V' \supseteq V$.*

Note that there is no connectivity constraints in the definition of a maximal cohesive pattern. Thus, the subgraph G' induced by the vertices of a pattern V can be disconnected; this subgraph is denoted as $G[V]$. The connected components of the induced subgraph can be written as:

$$G_V^{CC} = \{G_1, G_2, \cdots, G_l\}$$

Where l is the number of connected components in the induced subgraph,$G[V]$. Due to the anti-monotonicity property of the constraint, each connected subgraph G_i is cohesive.

Now we introduce the **Problem Definition**. Given a graph G, a gene profile matrix X, and a cohesiveness constraint C, the problem of discovering **Maximal Cohesive Patterns** is to find the set:

$$\mathcal{M}_{\mathcal{P}} = \{V_1, V_2, V_3, \cdots, V_{|\mathcal{M}_{\mathcal{P}}|}\}$$

such that every $V_i \in \mathcal{M}_{\mathcal{P}}$ is a maximal cohesive pattern.

Here we illustrate the **Algorithm** to mine the set of maximal cohesive patterns, the algorithm adopts a similar approach to the enumeration algorithm introduced in the *genMax* algorithm [41] for mining maximal frequent itemsets. In the original genMax algorithm, new items are added to the pattern as long as the new pattern maintains a frequency threshold. The proposed approach differs from the original genMax algorithm in that any pattern will be grown by adding new genes only if the new pattern satisfies the user constraints. Therefore, we allow for the integration of general anti-monotone constraints. For every reported maximal cohesive pattern (P), the algorithm extracts the connected components (G_P^{CC}) of the induced subgraph. We use a depth-first search to extract the connected components from the induced subgraph [42].

Table 1: A listing of the organisms for which the orthologs are used to create the evolutionary conserved profile for the Yeast.

Organism
Schizosaccharomyces pombe (Fission yeast)
Strongylocentrotus purpuratus (The purple sea urchin)
Xenopus tropicalis (Frog)
Arabidopsis thaliana
Drosophila melanogaster (Fruit fly)
Danio rerio (Zebrafish)
Homo sapiens (Human)
Musca domestica (House fly)
Mus musculus (House mouse)
Rattus norvegicus (Brown rat)

2.3. Experiments

To assess the effectiveness of the proposed method in discovering interesting patterns, we performed an experimental evaluation of our method using two datasets for both Yeast and Human.

2.3.1. Yeast Data

We used the high confidence Yeast protein-protein interactions, referred to YeastHC. The network was obtained from both literature-curated and high-throughput sources [43]. The YeastHC network contains 9857 interactions between 4008 genes. Considering literature-curated data is important because reporting the interactions has gone through several stages that are governed by many factors such as domain expertise, additional independent controls, prior contextual supporting information, and peer review. All of these factors reduce the probability of false positives. The YeastHC network includes data from a number of major interaction databases such as: BIND [44], BioGRID [45], DIP [46], MINT [47], and MIPS [48].

For the phenotype data, we took the growth phenotype profiles for Yeast mutants under 21 experimental conditions [49]. Here two different phenotype states are considered: 1 indicating growth and 0 indicating growth defect. Another profile dataset for the Yeast is the evolutionary conserved profile. Using the Inparanoid eukaryotic ortholog database [50], we built the gene profile using 10 different organisms illustrated in Table 1. A gene has a 1 in i^{th} profile, if the yeast gene has an orthologous gene in the i^{th} organism with 100% inparanoid score. Moreover, we used the gene expression profiles from [51]. This dataset contains 173 conditions. These conditions represent the response of yeast cells to different environmental changes. Some examples of the conditions are, heat shock, amino acid starvation, nitrogen depletion, and H_2O_2 exposure.

2.3.2. Yeast Complex Prediction

To investigate how well the extracted maximal cohesive subgraphs match known protein complexes in Yeast, we used the CYC2008 catalog that is comprised of 408 manually curated annotated protein complexes [52]. We used the overlap score proposed by [21] to measure to what extent a given maximal cohesive subgraph matches a known complex.

The overlap score is defined as follows: $w = \frac{c^2}{a*b}$, where a is the size of the maximal cohesive subgraph, b is the size of the complex, and c is the size of common proteins in both, i.e., size of the intersection.

A maximal cohesive subgraph has an overlap of 1 if it completely matches a known complex. If the overlap score of a given subgraph is higher than an overlap-threshold, we say that the subgraph matches the complex. We define precision (P) as the ratio of the number of matching maximal cohesive subgraphs to the number of all discovered maximal cohesive subgraphs. Since proteins participate in more than one biological complex, a maximal cohesive subgraph can match more than one complex and a complex can be matched by several subgraphs. An overlap threshold of 0.2, as suggested in [21], was used in our experiments.

We used the following cohesive constraints, $\mathcal{A} = \{\alpha_0, \alpha_1\}$, which ensures that $N_0(V(G)) \geq \alpha_0$ and $N_1(V(G)) \geq \alpha_1$. Table 2 shows the results of the proposed algorithm on the Yeast data, α_1 indicates the number of common growth conditions while α_0 indicates the number of common growth defect conditions, (M) is the number of the discovered maximal cohesive subgraphs, (MC) represents the number of matched complexes, (AMS) is the average maximal cohesive subgraph size, (DY) is the average density of the discovered subgraphs, and (AOS) is the average overlap score. The AOS score is computed as the sum of the maximum overlap scores for the matched subgraphs divided by the number of matched subgraphs.

For example, in the first row of Table 2, our proposed algorithm identified 20 candidate Yeast complexes that show growth in at least four experimental conditions and growth defect in at least four experimental conditions . Those 20 subnetworks have an average size of 4.9 genes and an average density of 0.66. Out of the 20 reported subnetworks, 13 matched known complexes.

In the case of the yeast response to multiple environmental changes, we investigated three regulation situations. Up-regulation which is represented by 1 in the profile data, Down-regulation (value of 2) and No regulation (value of 0). Through our analysis, we found that gene expression values were mostly down regulated. This result can be explained by the fact that most of these environmental conditions are stress related that reduce yeast survival chances. Table 3 shows

Table 2: Analysis of the Maximal Cohesive Subgraphs discovered from Yeast interaction data with Phenotype Profiles.

α_1	α_0	**M**	**AMS**	**DY**	**MC**	**AOS**	**P**
4	4	20	4.9	0.66	13	0.35	0.65
4	5	19	4.9	0.66	12	0.36	0.63
4	6	15	5.2	0.66	10	0.36	0.67
5	4	14	4.5	0.69	7	0.36	0.50
5	5	14	4.6	0.70	7	0.36	0.50
5	6	10	5.0	0.71	5	0.33	0.50
6	4	11	4.1	0.72	5	0.42	0.45
6	5	11	4.1	0.72	5	0.42	0.45
6	6	6	5	0.66	3	0.42	0.50

Table 3: Analysis of the Maximal Cohesive Subgraphs discovered from the Yeast dataset of Environmental Changes. Both α_0 and α_1 were set to 0.

α_2	**M**	**AMS**	**DY**
50	2524	9.41	0.31
51	1671	8.75	0.33
52	1111	8.10	0.35
53	728	7.49	0.37
54	500	6.88	0.40
55	329	6.36	0.41

that there are a large number of cohesive subgraphs that are down-regulated in many conditions (ranging from 50 to 55). However, the overlap with known complexes is not significant.

2.3.3. Human Data

We used the human HPRD network [53]. It contains 36888 interactions among 9453 human proteins. For the gene expression profiles, we used the expression patterns of 79 human tissues [54]. The gene is considered expressed if it is classified as present in both of the duplicated measurements. We also compiled a dataset of the Human genes orthologous evolutionary-conserved profiles. For all Human genes in the HPRD network, we extracted the orthologous genes in 63 different organisms from the Inparanoid eukaryotic ortholog database [50], such as the house mouse, the horse, and the common chimpanzee. This profile dataset is referred to as HE1.

Table 4: Analysis of the Maximal Cohesive Subgraphs discovered from the Human Evolutionary Conserved dataset (HE1).

α_0	α_1	**M**	**AMS**	**DY**	**MC**	**AOS**	**P**
3	38	140	3.55	0.66	61	0.32	0.54
3	39	117	3.42	0.69	67	0.32	0.55
3	40	94	3.36	0.70	62	0.33	0.56
4	37	101	3.46	0.65	46	0.32	0.43
4	38	79	3.38	0.67	44	0.33	0.43
4	39	64	3.28	0.70	44	0.34	0.50
4	40	50	3.30	0.71	43	0.35	0.54
5	38	54	3.38	0.68	33	0.34	0.43
5	39	40	3.27	0.71	35	0.35	0.47
5	40	33	3.27	0.72	34	0.35	0.45

Moreover, we extracted the expression profiles for 17 disease from the NCBI database [55]. We only considered the Affymetrix platforms, HGU-133A, HGU-Plus2, and HGU-95V2. This disease data was integrated with the HPRD interaction network. The objective we were looking for is to report disease related maximal cohesive subnetworks.

Some examples of the diseases are Bipolar Disorder, Renal Cell Carcinoma, Diabetic Nephropathy, and Colorectal Cancer.

2.3.4. Human Complex Prediction

For the Human complexes, we used the HPRD catalog that is comprised of 1521 manually curated annotated protein complexes [53]. It is important to mention that more than 30% of Human complexes have a density below 0.4. Table 4 show the results of the proposed algorithm on the HE1 dataset. For the Human interaction network with the evolutionary conservation profiles (HE1, which has 63 profiles), in Table 4, the algorithm, for example, searched for maximal cohesive subgraphs such that all genes have orthologous genes in at least 40 other organisms and at the same time do not have orthologous genes in at least 4 organisms ($\alpha_1 = 40, \alpha_0 = 4$). *Check the row in the bottom of the middle section of Table* 4. In addition, it is clear from Table 4 that our conserved maximal subgraphs matched many known complexes. For example, the proposed algorithm identified 140 subgraphs with ($\alpha_0 = 3$, $\alpha_1 = 38$). These 140 predicted subnetworks have

an average size of 3.55 genes and an average density of 0.66. In addition, out of the 140 reported subnetworks, 61 complexes were matched by the reported subnetworks.

Table 5: Analysis of the Maximal Cohesive Subgraphs discovered from the Human Tissues dataset (HT) .

α_0	α_1	**M**	**AMS**	**DY**	**MC**	**AOS**	**P**
5	30	2719	8.61	0.27	83	0.26	0.07
5	35	627	6.33	0.37	60	0.26	0.14
5	40	243	5.01	0.46	48	0.26	0.22
10	30	251	4.07	0.55	73	0.28	0.31
10	35	96	3.46	0.61	54	0.29	0.33
10	40	39	3.18	0.66	25	0.29	0.31
10	45	12	3.08	0.65	11	0.30	0.33
15	30	53	3.39	0.63	41	0.28	0.42
15	35	12	3.08	0.65	12	0.28	0.50
15	40	3	3	0.67	3	0.33	0.66

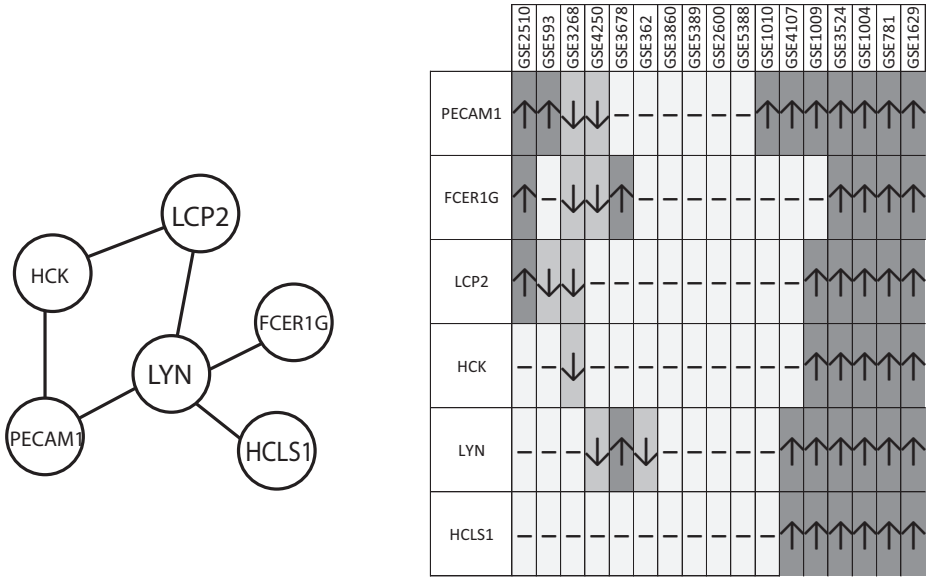
For the Human interaction network with the Human tissues profiles, in Table 5, we used different threshold values for α_1 and α_0 to assess their effect on the quality of the reported subnetworks. In one of the experimental settings, the proposed algorithm searched for subgraphs (subnetworks) that are consistently expressed in at least 30 tissues and consistently not expressed in at least 10 tissues *i.e.*, $\alpha_1 = 30, \alpha_0 = 10$. Table 5 shows that the density (DY), the average overlap score (AOS), and the complex prediction precision (P) are enhancing as the condition becomes more strict. In fact, even in the case of large average module sizes like 8.61 which have relatively low average densities, see the first row of Table 5, the algorithm is still capable of matching 83 known complexes. Knowing that the matching criteria employed in this work highly penalizes large size subnetworks.

For the Human interaction network with the Human disease data. The expression data was discretized to represent three regulation conditions; Up-regulated, Down-regulated, and Un-regulated. A cut-off value of 1.5 fold change was used for discretizing the expression values. In this case we are only interested in Up and Down regulated maximal cohesive subnetworks. The thresholds for the number of diseases that a subnetwork is Up and Down regulated in are

Table 6: Analysis of the Maximal Cohesive Subgraphs discovered from the Human Disease dataset (HD).

α_1	α_2	**M**	**AMS**	**DY**	**MC**	**AOS**	**P**
4	1	30	4.20	0.55	9	0.18	0.37
5	1	10	4.00	0.59	16	0.12	0.60
6	1	3	3.66	0.76	10	0.11	0.33
4	0	68	5.22	0.52	48	0.14	0.60
5	0	17	4.35	0.57	23	0.11	0.60
6	0	5	3.60	0.70	11	0.11	0.60

denoted by α_1, and α_2, respectively. Results for varying thresholds are shown in Table 6. For example, the proposed method reported 17 subgraphs (5^{th} row) that are Up regulated in at least 5 of the 17 diseases. When we make the condition more strict and require the subnetwork to be Down regulated in at least 1 disease (2^{nd} row), the number of cohesive subnetworks drops to 10. When we used different thresholds such as $\alpha_1 = 4$ and $\alpha_2 = 0$, the algorithm reported 68 cohesive subnetworks. Here, we lowered the overlap threshold to 0.1 to see how disease patterns can overlap with protein complexes. Nearly 60% of them have matched known human protein complexes.



(a) Discovered Cohesive Subgraph (b) Heatmap for the Discovered Subgraph

Figure 4: An example of the one of the produced cohesive subgraphs that matched one of the human protein complexes. The heatmap shows the diseases that this subgraph is up-regulated in.

Figure 4(a) shows one of the maximal cohesive subgraphs. This subgraph has a number of biologically interesting properties. One of the them is that this subgraph has matched one of the manually curated human protein complexes. Namely, the **the phospholipase C-gamma2** complex [56]. Moreover, the produced subgraph is up-regulated in four diseases, check the last four column in the profile data in Figure 4(b). Three of the diseases are related to different types of tumors, the *GSE*781 : *Kidney Carcinoma* , *GSE*3524 : *Tumor invasion in Oral Squamous Cell Carcinoma*, *GSE*1004 :*Dystrophin-deficient of human muscle* which found to cause organisms be prone to develop muscle tumors. The fourth disease the *GSE*1629 is related to abnormalities in the pulpal tissue. Furthermore, the subgraph in Figure 4 is involved in several Biological Processes such as, mast cell activation and leukocyte activation, and Molecular functions such as protein tyrosine kinase activity. The subgraph is also related to two biological pathways, namely, the *Fc epsilon RI signaling* pathway and the *Fc gamma R-mediated phagocytosis* pathway.

Table 7: GO enrichment analysis of the maximal cohesive patterns discovered from the Yeast dataset of Evolution Conserved Profiles.

α_1	#MCPs	ACS	ASSIC	ER	COV
5	252	956.0	56.1	1.0	0.37
6	210	880.4	49.9	1.0	0.37
7	120	815.5	45.5	1.0	0.34
8	45	758.8	42.2	1.0	0.30
9	10	708.7	39.67	1.0	0.24
10	1	664.0	37.0	1.0	0.14

2.3.5. Gene Ontology Enrichment of MCPs

In order to assess the biological significance of the extracted MCPs, we performed Gene Ontology enrichment analysis (GO) on the reported MCPs. For enrichment analysis, we adopted the performance measures used in [30].

1. **Enrichment** (ER) is computed as the ratio of MCPs that are enriched for at least one GO term to the total number of MCPs.

2. **Coverage** (COV) is the number of GO terms that are enriched in any of the MCPs divided by the number of all GO terms in the interaction network.

We have used the high-throughput GoMiner tool [57] for Go term enrichment analysis with an FDR-corrected p-values of 0.05. We have also collected some topological properties for the reported MCPs. ACS denotes the average number of genes in the reported MCPs and ASSIC is the average size of the subgraphs in the MCPs. An MCP may contain one or more subgraphs.

For **Yeast MCPs**, Table 7 shows the analysis of GO enrichment on the Yeast evolutionary conserved dataset. We varied the constraints from 5 to 10 with increments of 1. A constraints of 5 indicates that each reported MCP has at least 50% attributes (species) in which all the genes have a value 1. The total number of GO terms for the genes in the yeast interaction network is 1495.

It is clear that as we relax the constraint the average number of the genes in the reported MCPs (ACS) increases and so does the average size of the subgraphs in the MCPs (ASSIC).

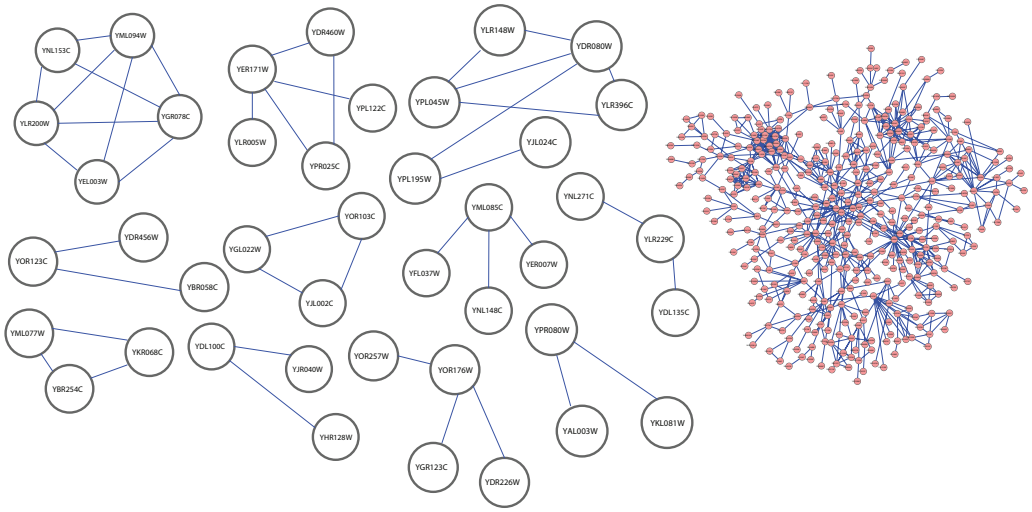


Figure 5: A maximal cohesive pattern from the Yeast Conserved dataset and its connected components. Only subgraphs with at least 3 genes are shown.

Moreover, the GO coverage of the reported MCPs increases since we get larger MCPs with more genes which are enriched. The Enrichment ratio is always 1.0, indicating that all the reported MCPs are enriched. This may be helped by the fact that all the MCPs are relatively large.

Figure 5 shows an MCP that was discovered with $\alpha_1 = 7$. The MCP has 716 genes and 246 GO terms were enriched in this pattern. Among the GO terms that are significantly enriched in this patterns are: RNA_modification, DNA_strand_elongation, and DNA_conformation_change.

With $\alpha_1 = 10$, there is only one MCP with 664 genes. The genes in this MCP have orthologous genes in all the 10 species. The largest component in this MCP has 368 genes. There are 212 GO terms that were significantly enriched in this patterns; among them were: mRNA metabolic process, response to DNA damage stimulus, and DNA strand elongation.

For **Human MCPs**, Table 8 shows the analysis of GO enrichment analysis and topological properties of the reported MCP on a subset of the Human evolutionary conserved dataset. This subset contains only the 17 closest species to the human. This dataset is referred to as HE2. We varied the constraints from 15 to 17 with increments of 1. A constraints of 16 indicates that

Table 8: GO enrichment analysis of the maximal cohesive patterns discovered from the Human dataset of Evolution Conserved Profiles.

α_1	#MCPs	ACS	ASSIC	ER	COV
15	136	1797.9	80.0	1.0	0.15
16	17	1675.8	71.9	1.0	0.09
17	1	1564.0	65.4	1.0	0.04

each gene in the reported MCP has orthologous genes in at least 16 species. The total number of GO terms for the genes in the Human HPRD interaction network is 7962. Similar to the case in the Yeast evolutionary conserved dataset, we observe similar trends in terms of the topological properties and GO enrichment analysis in the Human dataset.

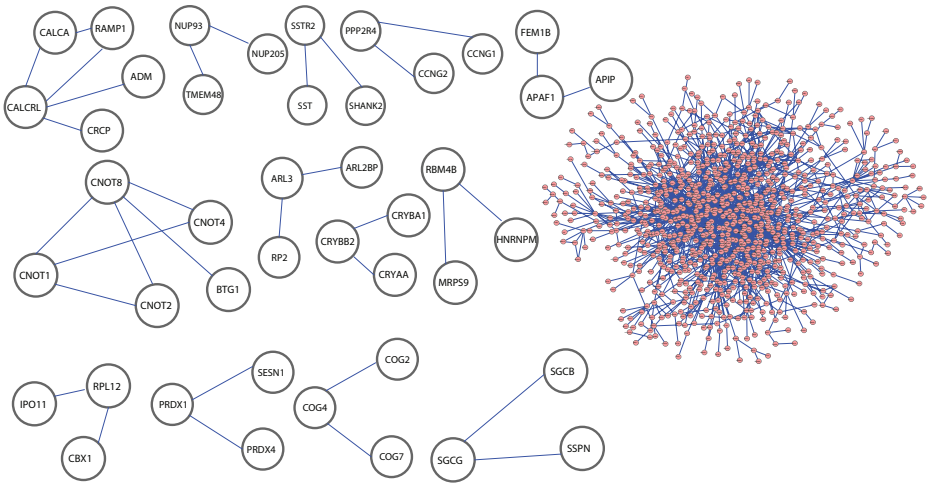


Figure 6: A maximal cohesive pattern from the Human HE2 Conserved Profile dataset and its connected components.

Figure 6 shows an MCP that was discovered with $\alpha_1 = 14$. The MCP has 1009 genes with the largest component of 966 genes and and there were 860 GO terms enriched in this pattern.

The GO term annotation analysis of this MCP shows that 625 biological processes were enriched including: cell cycle, regulation of cell death and proliferation, regulation of apoptosis, and mRNA metabolic process and processing.

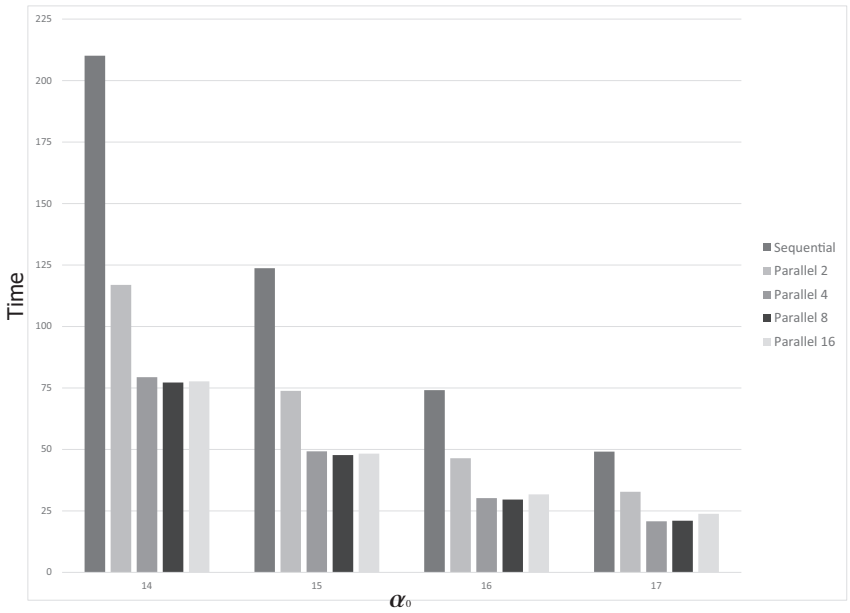


Figure 7: The effect of parallel execution on the human tissue data. Here the value of α_1 is equal to 17.

2.3.6. Running Time

Although the search space is exponential in terms of the number of genes in the network, the running time of the algorithm largely depends on the number of discovered patterns and the number of search branches explored. On the Human dataset(9465 genes and 37039 interactions) with the Human Tissue profiles (79 profiles), it took only 1.4 seconds to run the maximal cohesive subgraph algorithm with ($\alpha_0 = 10, \alpha_1 = 35$) to generate 251 subgraphs and it took only 0.77 second with ($\alpha_0 = 10, \alpha_1 = 45$) to generate 12 subgraphs. For the human evolutionary conserved profile (63 profiles), the algorithm took 1.4 seconds (generated 50 patterns) with ($\alpha_0 = 4, \alpha_1 = 40$), and 1.9 seconds (generated 140 patterns) with ($\alpha_0 = 3, \alpha_1 = 38$).

On the Yeast dataset(4008 genes and 9857 interactions) with the Phenotype profiles (21 profiles), it takes only 60 seconds to run the maximal cohesive subgraph algorithm with ($\alpha_0 = 0, \alpha_1 = 2$) to generate 75 subgraphs. For the evolutionary conserved profile (10 profiles), the algorithm takes 499 seconds with ($\alpha_0 = 1, \alpha_1 = 4$) to generate 128 subgraphs.

In this part we explore the effect of the **Parallel Execution** of the algorithm. The machine used in this experiment has an Intel Xeon processor with four cores of 2.4GHz and 8GB of memory.

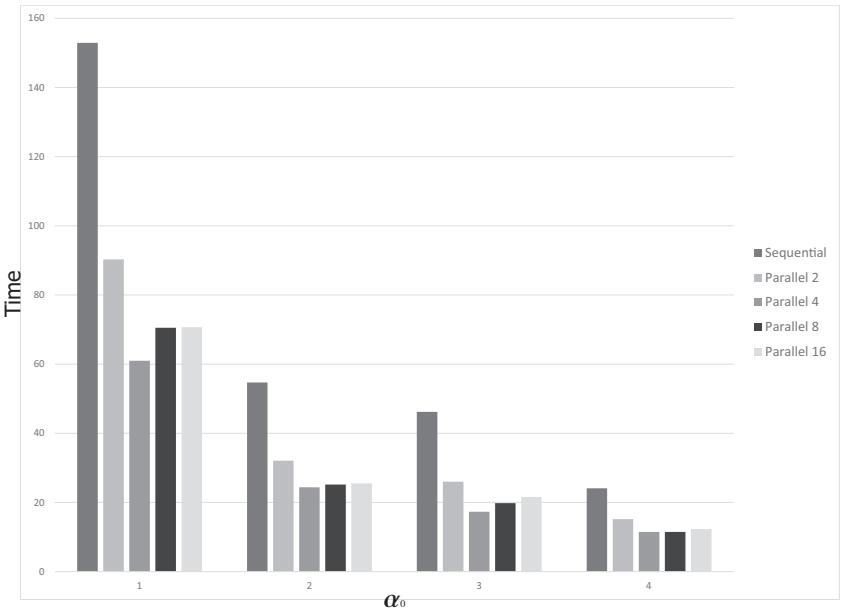


Figure 8: The effect of parallel execution on the yeast phenotype data. The value of α_1 is set to 1.

The algorithm starts with the main thread which is responsible for reading the data. Then a group of threads are spawn to perform the remaining parts of the process. In this experiment we considered using 2, 4, 8 , and 16 threads as can be seen from figures 7 and 8. Any race conditions are guarded against with locking so that the check and modification of the shared resources are done safely and accurately. From the figures, it is clear the performance advantage that we gain with employing the parallel paradigm. Moreover, the effect of changing the values of αs can be seen. In general, the running time increases as we relax the constraints. Furthermore, it is interesting to see that the performance gain starts to fade even with the increase of the number of threads over 4 threads. This note can be referred to the fact that the machine used in the experiments has only 4 cores. As a result, the burden imposed for scheduling these threads starts to harm the overall performance of the algorithm when the number of threads exceeds the number of available cores.

2.4. Conclusion

In this work, we have proposed an enumeration-based method to discover maximal cohesive induced subnetworks by integrating constraints defined over the gene profiles with the subnetworks enumeration process. We have employed three pruning strategies that significantly reduce

the search space explored. We also have proposed an algorithm for discovering maximal cohesive patterns (MCPs). We assessed the biological significance of the reported patterns by investigating the overlap with known complexes and also by GO enrichment analysis. Experimental results on Human and Yeast datasets show that the proposed methods discover biologically significant patterns. Many of these subnetworks match known complexes.

CHAPTER 3. DISCOVERING DYSREGULATED PHENOTYPE-RELATED GENE PATTERNS[2]

Recent research showed that differentially expressed, or *dysregulated*, patterns of interacting proteins exhibit more interesting properties with respect to many complex phenotypes. In this work we follow an integrative approach by combining the physical protein-protein interaction, PPI, network with gene expression data for a number of diseases and phenotypes. In this study, we propose an algorithm for mining *D*ysregulated *P*henotype-*R*elated interacting genes, *DPR*s. Experimental results on 88 Human gene expression datasets that were annotated by employing UMLS mapping demonstrate the effectiveness of the algorithm in discovering biologically and statistically significant DPRs.

3.1. Contribution

In this work we propose an approach for mining dysregulated patterns by integrating expression datasets and PPI network. The problem is similar to the problem of discovering subnetwork biomarkers for gene expression classification [23, 37]. However, instead of classifying samples, we mine patterns, sets of genes, that distinguish between the datasets that are labeled with a UMLS concept.

- Propose the discovery of Dysregulated Phenotype-Related Patterns, DPRs. For the discovery process, we follow the greedy pattern growth approach that was introduced in [23]. Instead of starting the algorithm with single genes, we start with seeds of multiple genes.

- We have created dysregulation profiles for genes in the protein-protein interaction network. A gene dysregulation profile captures the dysregulation of the gene in the 88 datasets. The use of physical network adds credibility to the mined gene patterns.

[2]The material in this chapter was co-authored by Rami Alroobi and Saeed Salem. Rami Alroobi was responsible of developing the idea. Rami Alroobi prepared the data used, did the analysis, and drafted and revised the chapter. Saeed Salem served as proofreader and checked the correctness of the mathematical formulation.

- We demonstrate the biological relevance of the DPRs to the studied phenotypes by illuminating the biological relevance through Gene Ontology, KEGG pathway enrichment, and the DPRs overlap with known protein complexes.

- Additionally, we show that the DPR patterns have high classification power by using these patterns as classification features.

3.2. Problem Description

We introduce some of the definitions that will be used throughout this chapter.

The **Graphs** is defined as the following: A graph $G = (V, E)$, consists of a set of vertices $V = \{g_1, g_2, \ldots, g_n\}$, and a set of edges $E \subseteq V \times V$. The vertex set and edge set of a graph G are denoted by $V(G)$ and $E(G)$, respectively.

For **Dysregulation Profile**, given a set of genes, $V = \{g_1, g_2, \ldots, g_n\}$ and a set of data sets $C = \{C_1, C_2, \ldots, C_m\}$, Let $X \in R^{n \times m}$ represents the Gene Dysregulation Matrix, such that x_{ij} is the dysregulation value of the i^{th} gene in j^{th} experiment. For a gene, g_i, assume that $X_i \in R^m$ is the gene dysregulation profile. Moreover, we have a class label, $L = \{l_1, l_2, \cdots, l_m\}$, where $l_i \in \{1, 0\}$ indicates whether the i^{th} experiment has the phenotype.

The **Pattern Activity** is defined similar to, but different than [23], the pattern dysregulation activity of a set of genes in S, is the average of dysregulation profiles of these genes and is defined as:

$$X_S = \frac{1}{|S|} \times \sum_{g_i \in S} X_i$$

For **The Objective Function**, given that $L = \{l_1, l_2, \ldots, l_m\}$, where $l_i \in \{L_+, L_-\}$, and the objective function, $f(X_S)$, we are considering in this work is the Mutual Information, $I(X_S; L)$, then:

$$f(X_S) = \sum_{d_s \in X_S} \sum_{l \in L} p(d_s, l)\ log\left(\frac{(p(d_s, l)}{(p(d_s)p(l)}\right)$$

3.3. Algorithm Description

The proposed algorithm follows a similar search strategy as the work done by [23], however, in this work we are not looking to find biomarkers for classifying gene expression samples.

Algorithm: Mining *DPR*s:

Input:
$\mathcal{G}$: PP Interaction Network
$\mathcal{X}$: Dysregulation Attribute Matrix
$\mathcal{L}$: Labels vector
$\mathcal{S}$: List of seeds

Output:
$\mathcal{P}_a$: Context related patterns

1.**for each** $p \in \mathcal{S}$:
2. $T_n \leftarrow Nbr(p)$
3. **While** $True$:
4. $g_i \leftarrow findMaxMI(p, \mathcal{L}, T_n, \mathcal{X})$
5. $p' \leftarrow p \cup g_i$
6. **if(** $MI(p', \mathcal{L}) > MI(p, \mathcal{L})$ **):**
7. $p \leftarrow p'$
8. $T_n \leftarrow Nbr(p)$
9. **else**:
10. *break*
11. End if
12. **End While**
13. **if (** $p \in \mathcal{P}_a$):
14. **next**
15. **if (** $\exists P \in \mathcal{P}_a : p \subseteq P$):
16. **next**
17. **else**:
18. $\mathcal{P}_a \leftarrow \mathcal{P}_a \cup p$:
19. **End if**
20.**End for**
21. **return** $\mathcal{P}_a$

Figure 9: The Algorithm for mining Dysregulated Phenotype-Related patterns.

The first phase is **Seeding and Filtering Phase**. In the algorithm in Figure 9, we start from seed patterns. These seeds can be 1-gene seeds, as in [23], or h-genes connected and cohesive seeds. In this work we tried both scenarios, however, we illustrate only the results using 3-genes connected seeds because they gave better results. The seed connectivity condition supports the logic of being biologically important based on the PPI network. In addition to being connected,

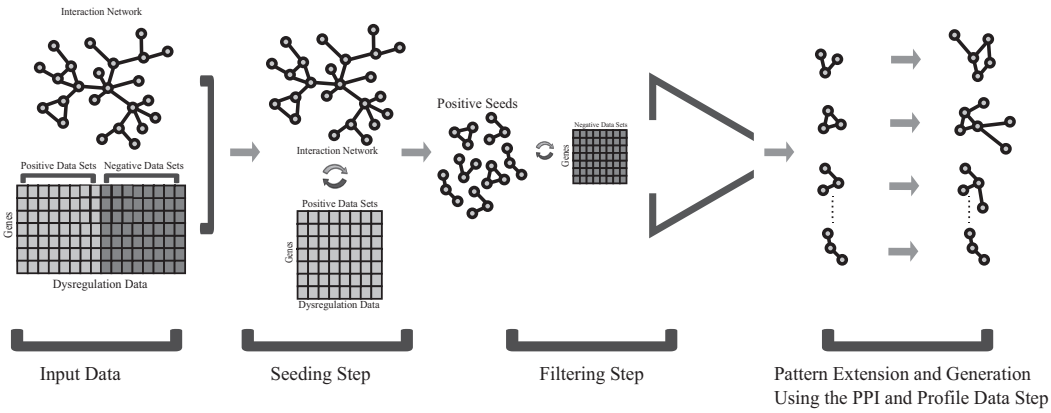


Figure 10: General overview of the approach presented in this work. After generating the seeds, a filtering step is done. Then, seeds are grown by adding neighboring genes. The same steps are performed by exchanging positive with negative contexts.

the seed has to be cohesive in terms of the dysregulation profiles of its member genes. The nodes of the seed have to share at least k dysregulation profiles either in the positive or negative context. The filtering step is required to filter out all unpromising seeds. A seed has to show high dysregulation concentration in one context and not in the other. Any positive seed that shows strong dysregulation on the negative side will be discarded. The same is performed for the negative seeds having strong dysregulation in the positive side. By strong dysregulation we mean that seed genes have high concentration of dysregulation indicators, i.e. density of ones, in the datasets. After the seeding and filtering steps are completed, an extension step will be carried out to produce the final patterns. The second phase is the **Extension and Pattern Generation Phase**. In the extension process, every seed in $\mathcal{S}$ is expanded with one of its neighboring genes that maximizes a mutual information-based objective function. The extension process is repeated until no enhancements is possible. More specifically, we extend each seed by adding new neighboring genes that satisfy two conditions:

- The path distance between the seed and the new gene must not exceed a predefined length which we set to 2. This is to limit the number of candidate genes that the seed can be extended with and thus enhance the performance of our algorithm in terms of running time.

- Intuitively, from the previous condition many candidate genes could be found for every seed to choose from. The new gene, among all candidate genes, to be selected for seed extension should maximize our objective function, which is the Mutual Information score. This is the role of the function $findMaxMI(p, \mathcal{L}, T_n, \mathcal{X})$, where $\mathcal{L}$ is the labels vector in the algorithm shown in Figure 9.

Figure 9 shows the pseudo code of the algorithm. The extension process is repeated for each seed (line 1). The extension process continues as long as adding new genes enhances the pattern's score (line 6). Every time the pattern is extended (line 7) the pattern score is identified and the set of neighbors is updated (line 8). When no more enhancement is possible (line 10), we check to see if the pattern has not been reported before (line 13) since multiple seeds can generate the same pattern. Moreover, we check if the pattern is not subsumed by a previously reported pattern (line 15).

3.4. Experiments

3.4.1. Data Preprocessing

The algorithm is based on an integrative approach through combining two sources of biological data. One source is the protein-protein interaction network. The PPI network we considered in this research is the HPRD network[58], release 9, April 2010. After preprocessing, this PPI has $9,453$ genes and $36,888$ interactions. The other source of data is a group of 88 Human gene expression datasets. The datasets are acquired from the NCBI Gene Expression omnibus [55]. The 88 datasets had to meet two criteria: 1) each dataset has at least eight samples and 2) the samples should have a clear categorization into two classes (e.g., control vs. patient or tumor vs. non-tumor) so that the expression dysregulation can be calculated. The platforms we used were the GPL96 (Affymetrix HG-U133A) where we have 46 of the datasets belong to this platform and the GPL570 (Affymetrix HG-U133-Plus-2) to which the remaining 42 belong. The datasets we chose are related to different phenotypes. For every dataset, the top dysregulated genes whose p-value of the expression dysregulation is below 0.05 and appear in the PPI network are only considered.

Table 9: The 88 datasets used in the study.

Dataset ID	Dataset Description	No. of Samples
GSE2280	Prediction of lymphatic metastasis from primary squamous cell carcinoma of the oral cavity	27
GSE362	Normal human muscle	30
GSE1650	chronic obstructive pulmonary disease Study	30
GSE1786	Vastus lateralis biopsies from healthy trained and sedentary males	24
GSE1551	dermatomyositis	23
GSE474	Obesity and fatty acid oxidation	24
GSE2779	Gene expression profile of normal v early MDS v non-MDS anemia bone marrow CD34 cells	28
GSE5388	Adult postmortem brain tissue (dorsolateral prefrontal cortex) from subjects with bipolar disorder and healthy controls	61
GSE5389	Adult postmortem brain tissue (orbitofrontal cortex) from subjects with bipolar disorder and healthy controls	21
GSE674	Normal Muscle-Female , Effect of Age	30
GSE3868	Gene profiling of primary cultures from human prostate tumors	30
GSE2006	Comparative microarray between normal and essential thrombocythemia platelets	14
GSE1751	Human blood expression for Huntington's disease versus control	31
GSE41804	Hepatic gene expression of HCV related Hepatocellular carcinoma and non-cancerous tissue with Il28B rs8099917 TT genotype and TG/GG genotype	40
GSE23117	Gene expression in minor salivary gland of patients with Sjogren's syndrome (SS) and control	15
GSE21935	Comparison of post-mortem tissue from Brodman Brain BA22 region between schizophrenic and control patients	42
GSE21138	Gene Expression Profiles in BA46 of Subjects with Schizophrenia and Matched Controls	59
GSE9576	Gene expression profiling of classical midgut carcinoid primary tumors and liver metastasis	12
GSE5547	Host Susceptibility to H. ducreyi Infection is Associated with Unique Transcript Profiles in Tissue and Dendritic Cells	24
GSE5281	Alzheimer's disease and the normal aged brain (steph-affy-human-433773)	161
GSE7014	Expression data from DM1, DM2 and Normal Adult Skeletal Muscle Biopsies	36
GSE4757	Alzheimers disease: neurofibrillary tangles (Rogers-3U24NS043571-01S1)	20
GSE13911	Expression data from primary gastric tumors (MSI and MSS) and adjacent normal samples	69
GSE5563	Gene expression profile of VIN lesions in comparison to controls	19
GSE27562	Expression data from human PBMCs from breast cancer patients and controls	162
GSE36668	Expression data from serous ovarian carcinomas, serous ovarian borderline tumors and surface epithelium scrapings from normal ovaries	12
GSE9348	Expression data from healthy controls and early stage CRC patient's tumor	82
GSE4107	Expression profiling in early onset colorectal cancer	22
GSE7803	Human pre-invasive and invasive cervical squamous cell carcinomas and normal cervical epithelia	41
GSE3726	Prognostic gene signatures can be measured with samples stored in RNAlater	104
GSE590	USF1 haplotype comparison	10
GSE593	Uterine Fibroid and Normal Myometrial Expression Profiles- U133 Arrays	10
GSE1577	T-ALL and T-lymphoblastic lymphoma	29
GSE2117	CALM-AF10 T-ALL	23
GSE473	PGA Human CD4+ Lymphocytes	175
GSE3167	Classification of carcinoma in situ lesions in human bladder cancer	60
GSE1615	Theca cell gene expression	26
GSE2712	Clear cell sarcoma of the kidney (CCSK)	35
GSE1518	Human endothelium exposed to shear stress and pressure	8
GSE1045	Estradiol Treated Breast Cancer Cells Expressing Mutant Estrogen Receptors	12
GSE2189	Human lung cancer (A549) teatement with MGd	18
GSE923	Pseudomonas aeruginosa infection of Calu-3 human lung epithelial cells	19
GSE2719	Gene expression of human soft tissue sarcoma	54
GSE3268	Squamous Lung Cancer, Paired Samples	10
GSE1869	Ischemic and Nonischemic CM and NF Hearts	37
GSE1595	Human bladder smooth muscle cells - effect of stretch in vitro	8
GSE6008	Human ovarian tumors and normal ovaries	103
GSE1849	Differential Gene Expression in Pulmonary Artery Endothelial Cells Exposed to Sickle Cell Plasma	65
GSE3320	Gene expression profile of small airway epithelium of normal non-smokers and normal smokers	11
GSE2510	Expression profiling in preadipocytes in obese Pima Indians/humans	56
GSE5788	Expression data from T-cell prolymphocytic leukemia (TPLL) and normal T cells	14
GSE3860	Comparison of Hutchinson Gilford Progeria Syndrome fibroblast cell lines to control fibroblast cell lines	36
GSE1420	Barrett's esophagus, Barrett's-associated adenocarcinomas and normal esophageal epithelium	24
GSE3297	Laparoscopic Donor Nephrectomy Gene Expression Profiling Compared to Healthy Control Kidneys	12
GSE1059	Myometrial cells expressing CREB, CREM alpha, CREM tau2alpha, ATF2 or the ATF2-small gene	18
GSE3365	Comparison of PBMCs in Inflammatory Bowel Disease	127
GSE2549	Malignant pleural mesothelioma	54
GSE1297	Incipient Alzheimer's Disease: Microarray Correlation Analyses	31
GSE4570	Expression profiling reveals novel pathways in the transformation of melanocytes to melanomas	8
GSE2429	Atypical Ductal Hyperplasia	8
GSE1474	Caco-2 and T84 cells stimulated with with flagellin, lymphotox beta or TNF alpha	24
GSE44971	Gene expression data from pilocytic astrocytoma tumour samples and normal cerebellum controls	58
GSE20086	Heterogeneity of gene expression in stromal fibroblasts of human breast carcinomas and normal breast	12
GSE35493	Pediatric rhabdoid tumors of kidney and brain show many differences in gene expression but share dysregulation of cell cycle and epigenetic effector genes	71
GSE37258	Expression data of the iPSCs derived from foreskin fibroblast cells of normal person and KS patient	18
GSE5230	Epigenetics of gene expression in human hepatoma cells	16
GSE16515	Expression data from Mayo Clinic Pancreatic Tumor and Normal samples	52
GSE29796	Transcriptional Differences between Normal and Glioma-Derived Glial Progenitor Cells Identify a Core Set of Dysregulated Genes	72
GSE3744	Human breast tumor expression	47
GSE8514	Expression data from normal adrenal gland and aldosterone-producing adenoma	15
GSE41328	Colorectal adenocarcinomas and matched normal colonic tissues	20
GSE19429	Expression data from bone marrow CD34+ cells of MDS patients and healthy controls	200
GSE4619	Gene expression profiling of CD34+ cells from MDS patients and normal controls	66
GSE15960	Expression data from human colonic epithelial cells normal (N), adenoma (AD) or colorectal cancer (CRC) tissues	18
GSE10927	Human adrenocortical carcinomas (33), adenomas (22), and normal adrenal cortex (10), on Affymetrix HG-U133-plus-2 arrays	65
GSE9171	Expression profiles of human glioblastoma frozen tumors and cell lines	30
GSE4567	Endothelial cell culture with Chapel Hill Ultrafine particle	8
GSE5040	Polyamides alleviate transcription inhibition associated with long GAA.TTC repeats in Friedreichs ataxia	24
GSE8762	Lymphocyte gene expression data from moderate stage HD patients and controls	22
GSE46449	Expression data from Patients with Bipolar (BP) Disorder and Matched Control Subjects	88
GSE4883	Simvastatin has an anti-inflammatory effect on macrophages via upregulation of Kruppel-like factor-2	9
GSE18842	Gene expression analysis of human lung cancer and control samples	91
GSE13471	Expression data from human normal pre-frontal cortex, liver, and colon tissues and colon tumors	18
GSE4183	Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature	53
GSE5081	Expression data from Helicobacter positive and negative human gastritis samples	32
GSE50161	Expression data from human brain tumors and human normal brain	130
GSE5764	Analysis of microdissected invasive lobular and ductal breast carcinomas in relation to normal ductal and lobular cells	30
GSE11151	Gene expression data from different types of renal tumors and normal kidneys	67

The value of the (ij) entry of the dysregulation matrix indicates whether gene i is dysregulated, 1, in dataset j or not, zero. The next step is to create an attribute dysregulation matrix out of the aforementioned vectors. The size of the attribute matrix is n $\times$ 88, where n is the number of genes in the PPI network.

3.4.2. Dataset Phenotypic Annotation

A Unified Medical Language System, UMLS, Metathesaurus [59] approach was used to determine the biological context for every dataset according to what Medical Subject Headings, MeSH, terms the dataset contains. For the datasets we considered in this work, we made sure that they are included in the PubMed database and they have PubMed identifier, PMIDs. The datasets are transformed into their corresponding PMIDs. The BioPython package is used to check that the datasets with these PMIDs have MeSH terms and retrieve the MeSH terms. Hence, the result is a map from datasets to MeSH terms. Then a mapping process is done from MeSH into UMLS concepts. The illustration in Figure 11 shows the process for one of the datasets used in this work.

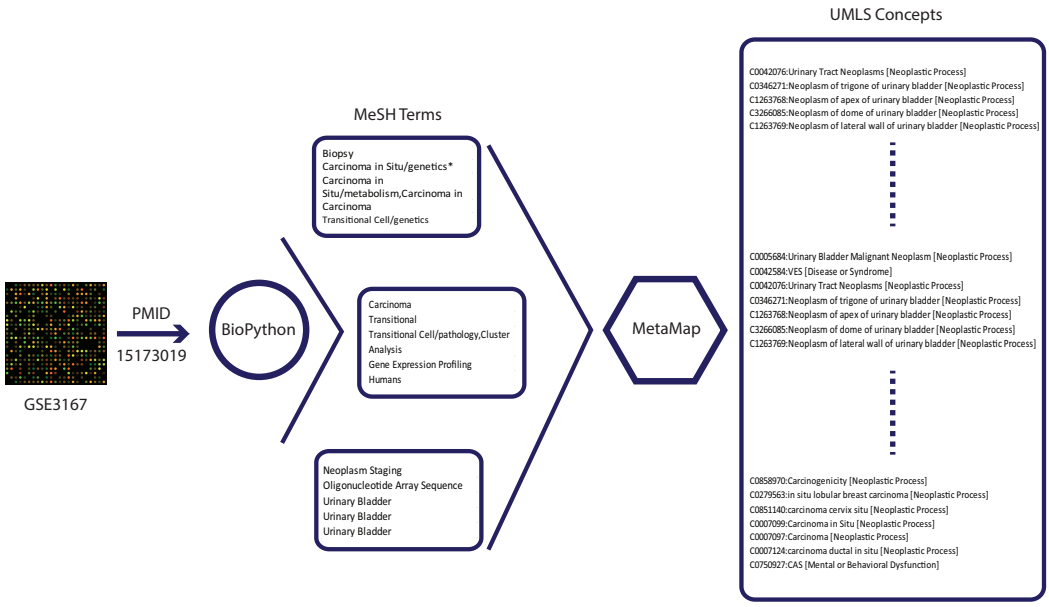


Figure 11: The approach used to map the MeSH terms into their corresponding UMLS concepts. The dataset used in this example has the ID GSE3167 and it is about carcinoma in situ lesions in human bladder cancer.

The dataset is for human bladder cancer. The dataset has an ID of GSE3167. This ID is mapped into its PMID, i.e. 15173019. In the algorithm implementation, this PMID is given to methods of the BioPython package. These methods obtain the MeSH terms attached to the dataset under processing. This dataset has 15 MeSH terms. After acquiring the MeSH terms, these terms are submitted to the code that runs the MetaMap tool. This tool maps the MeSH terms into the UMLS concepts. The result was 162 concepts. The usage of UMLS approach we considered in this study has many advantages, for example: Vocabularies integrated in the UMLS Metathesaurus include many resources such as Gene Ontology, OMIM, and the Digital Anatomist Symbolic Knowledge Base. Also, UMLS concepts are not only inter-related, but may also be linked to external resources such as GenBank. The UMLS knowledge sources are updated quarterly. The UMLS system has major components which are the Metathesaurus, the Semantic Network, and lexical resources including the SPECIALIST lexicon and programs [60, 61]. The Semantic Network provides high-level categories used to categorize every concept. The SPECIALIST part is used for generating the lexical variants of biomedical terms. In the UMLS, the knowledge is arranged into concepts. Similar terms are grouped into concepts. These concepts are linked by different types of relations. These relations can be symbolic such as "is kind of" or "part of", e.g. the NF2 concept is *part of* Tumor Suppressor genes, and statistical which were concluded from co-occurrence of MeSH terms in the Medical Literature Analysis and Retrieval System Online, MEDLINE, database. The UMLS database has millions of biological related concepts that span diseases, treatments, and other phenotypes in different levels of details that start from molecules to entire organisms. During the mapping process, several settings can be used, especially, to specify the semantic types we are interested in. These types provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. Moreover, the UMLS approach helps to overcome two significant problems of retrieving machine-readable information:

1. The variety of names used to express the same concept, and

2. The absence of a standard format for distributing terminologies.

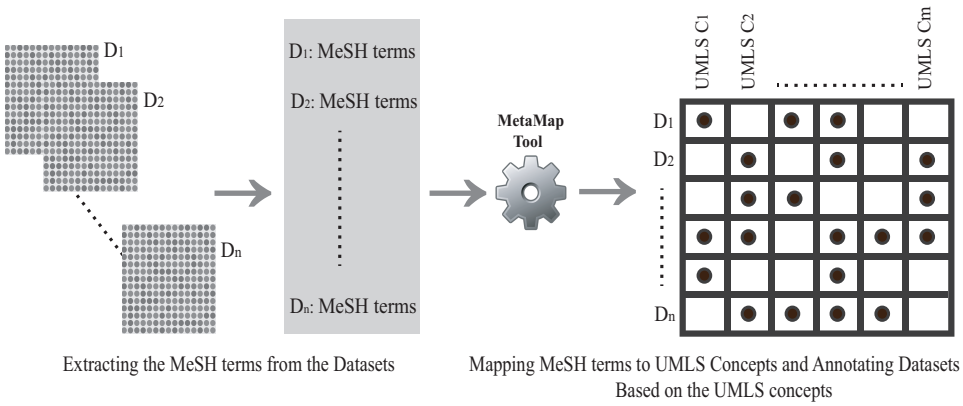


Figure 12: An illustration of how the datasets are annotated with UMLS concepts. The dots in the matrix mean that the dataset, D_i, has the concept, C_j. Each column in the matrix can be used as the class label for the datasets.

MeSH vocabularies are maintained by the National Library of Medicine, NLM, and used as means for indexing PubMed articles. To accomplish the task of mapping the MeSH terms to UMLS concepts, we employed the MetaMap tool [62]. The biomedical resources are vast such as databases for sequences, model organisms, biomedical literature ... etc. These different resources have a common aspect which is the terminology. This terminology is employed for integration among these resources but does not imply any sort of vocabularies standrization. Therefore, the mapping to UMLS is an effort towards establishing the standard for integrating information in different resources.

Figure 12 shows the steps of annotating datasets with a phenotype. In this study, we limited ourselves to only UMLS semantic types that are explicitly related to biological functions, stress conditions, and tissues while in the mapping process. The reason behind this limitation is that we are mainly interested in discovering gene patterns that are associated with phenotypic contexts. After processing all the 88 datasets to extract the UMLS concepts using the MetaMap tool; for every UMLS concept, we create a profile of which datasets are related to this concept. Moreover, every dataset has a profile of which UMLS concepts it contains. In addition, every semantic type has its own profile which shows what datasets the semantic type appears in. The datasets comprise

the dimension of the attribute dysregulation matrix. For each concept/phenotype, we can divide the datasets into two contexts: A positive context, where the datasets in this context have a particular UMLS concept that we are studying, and a negative context where the remaining datasets lack the same concept. Figure 10 shows the entire approach we developed in this work. We applied our approach to 88 Human gene expression datasets. These datasets are related to different phenotypes. According to the mapping process, we found that only a limited number of UMLS terms have large-enough number of datasets annotated with every one of these terms. Table 10 displays the six UMLS concepts/phenotypes that annotate a large number of datasets and were used in the study. The semantic types abbreviations are explained as **dsyn**: Disease or Syndrome. **neop**: Neoplastic Process. **cgab**: Congenital Abnormality. **cmod**: Cell or Molecular Dysfunction. **mbod**: Mental or Behavioral Dysfunction.

Table 10: The UMLS terms used in this study. Third column shows the number of datasets annotated with the corresponding term.

UMLS Term	**Semantic Type**	**# datasets**
Acute Gastroenteritis	dsyn	28
Carcinoma	neop	14
Cell or Molecular Dysfunction	comd	16
Congenital Abnormality Disease or Syndrome	cgab	25
Mental or Behavioral Dysfunction	mobd	30
Neoplastic Process	neop	63

By choosing UMLS terms with relatively large number of annotated datasets, we tried to reduce the gap between the number of datasets having the phenotype and the remaining datasets lacking the same phenotype. Hence, we have a balance between the datasets in the different categories.

Table 11: An illustration of the distribution of the gene patterns along with average pattern size, column $\bar{V}$, and average pattern density, the $\bar{\sigma}$ column.

Phenotype Class	**#S+ patterns**	**#S- patterns**	$\bar{V}$	$\bar{\sigma}$
Acute Gastroenteritis	1092	1861	7.9	0.29
Carcinoma	135	1764	7.5	0.30
Cell or Molecular Dysfunction	351	1891	8.0	0.28
Congenital Abnormality Disease or Syndrome	404	1493	8.1	0.27
Mental or Behavioral Dysfunction	1103	2227	7.7	0.29
Neoplastic Process	1351	278	8.45	0.27

3.4.3. Reported DPRs

For every phenotype class we generated two groups of seeds, i.e. the positive and negative seeds, depending on which part of the dysregulation matrix, see Figure 1, was used to produce the seeds. The seeds are 3-gene patterns that are cohesive in their profiles according to the condition we set. Interestingly, the seed can be one gene, which we tried in the analysis, however, we found that the 3-gene seeds gave better results. In our work, a total of $13,950$ patterns were discovered from all phenotype classes. Throughout our work, every 3-gene seed has to show homogeneous dysregulation in at least 6 datasets and these seeds are used to produce the patterns we report in Table 11. Not all seeds will produce *DPR*s, because some patterns are subsumed in other already found *DPR*s. The third and fourth columns of Table 11 show the average size of the reported *DPR*s and the average density respectively.

3.4.4. Functional Enrichment Analysis

To show how significant the patterns that our algorithm produces, we performed an enrichment analysis focusing on two aspects. The first aspect is to what extent the reported patterns were functionally homogeneous when tested against the Gene Ontology biological process terms. If the pattern was enriched in a GO term with p-value less than 0.01, then this pattern will be added to the group of enriched patterns. The analysis was performed using the DAVID tool [63] using

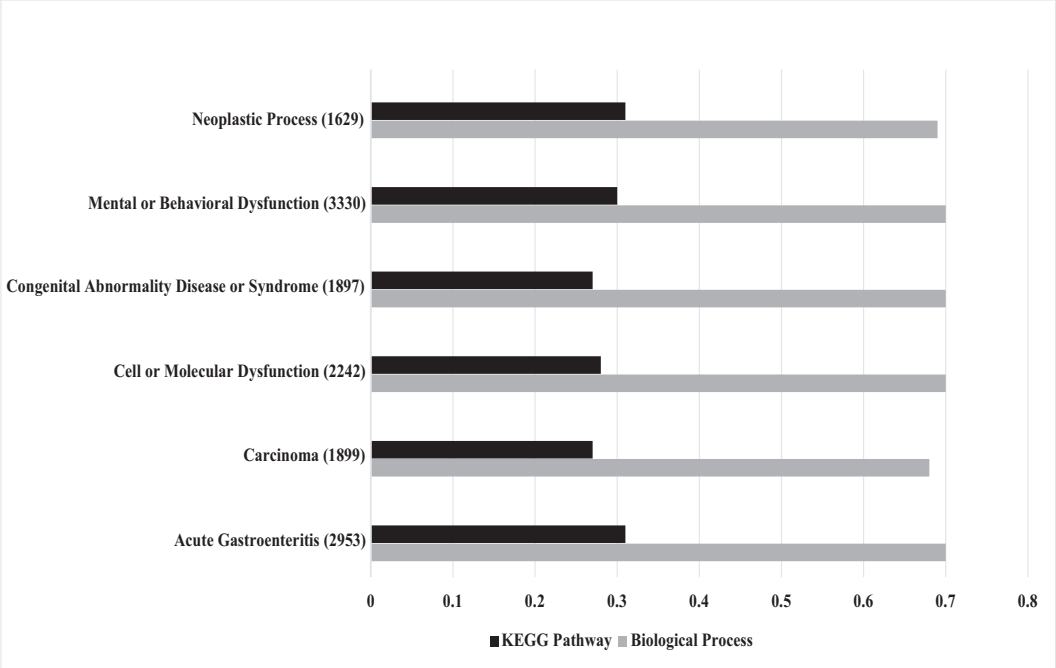


Figure 13: The enrichment analysis for the reported patterns.

Table 12: Examples of GO terms that are enriched in the reported DPRs.

Phenotype Class	GO -Term	GO Explanation	P-Value
Acute Gastroenteritis	GO:0019059	Initiation of viral infection	8.1e-03
	GO:0019058	Viral infectious cycle	4.3e-04
	GO:0002541	Activation of plasma proteins involved in acute Inflammatory response	5.3e-12
Carcinoma	GO:0002858	Regulation of natural killer cell mediated Cytotoxicity directedagainst tumor cell target	2.5e-06
	GO:0012502	Induction of programmed cell death	3.1e-07
	GO:0000718	Nucleotide-excision repair, DNA damage removal	1.8e-10
	GO:0002253	Activation of immune response	4.3e-11
Cell or Molecular Dysfunction	GO:0044092	Negative regulation of molecular function	3.6e-06
	GO:0006974	Response to DNA damage stimulus	2.7e-06
	GO:0016481	Negative regulation of transcription	7.1e-08
Congenital Abnormality Disease or Syndrome	GO:0001701	In utero embryonic development	9.9e-04
	GO:0035113	Embryonic appendage morphogenesis	8.7e-06
	GO:0055008	Cardiac muscle tissue morphogenesis	5.9e-07
Mental or Behavioral Dysfunction	GO:0016079	Synaptic vesicle exocytosis	1.1e-05
	GO:0051969	Regulation of transmission of nerve impulse	1.9e-07
	GO:0031644	Regulation of neurological system process	2.3e-07
	GO:0006836	Neurotransmitter transport	6.5e-09
Neoplastic Process	GO:0070647	Protein modification by small protein conjugation or removal	8.7e-05
	GO:0045596	Negative regulation of cell differentiation	2.1e-06
	GO:0031399	Regulation of protein modification process	8.1e-09

Table 13: Examples of KEGG pathways that are enriched in the reported DPRs.

Phenotype Class	KEGG-Term	KEGG Explanation	P-Value
Acute Gastroenteritis	hsa05130	Pathogenic Escherichia coli infection	2.5e-03
	hsa05110	Vibrio cholerae infection	1.2e-05
Carcinoma	hsa05210	Colorectal cancer	1.7e-05
	hsa05340	Primary immunodeficiency	1.1e-06
Cell or Molecular Dysfunction	hsa04110	Cell cycle	4.7-e04
	hsa04210	Apoptosis	2.6e-06
Congenital Abnormality	hsa04810	Regulation of actin cytoskeleton	5.3e-09
Disease or Syndrome	hsa04010	MAPK signaling pathway	2.2e-06
Mental or Behavioral	hsa04722	Neurotrophin signaling pathway	1.3e-04
Dysfunction	hsa04080	Neuroactive ligand-receptor interaction	8.6e-05
Neoplastic Process	hsa05221	Acute myeloid leukemia	7.5e-04
	hsa04062	Chemokine signaling pathway	1.5e-08

the default settings. DAVID provides information about biological terms enriched in a gene list relative to all annotated genes of the organism. Figure 13 shows the summary of the enrichment analysis we performed in this work. Out of the 13950 reported patterns for all classes in the study, the average enrichment was around 70%. The other aspect we considered is the enrichment in terms of the KEGG pathways. The average KEGG pathway enrichment was 29%.

3.4.5. Interesting GO Terms and KEGG Pathways

During the enrichment analysis part, we found evidences about how the patterns discovered by our algorithm are related to the context at hand. This is accomplished by studying the biological significance in more depth and try to highlight if there are biological processes, or KEGG pathways that are enriched in the discovered patterns. Interestingly, we were able to shed some light on such cases. For example, in the reported patterns that are related to the Acute Gastroenteritis class. We found many GO terms that are in strong relation to this class either in terms of cause or symptoms. GO terms such as, “initiation of viral infection”, “viral infectious cycle”, and “activation of plasma proteins involved in acute inflammatory response”. Moreover, some interesting KEGG pathways are also enriched in the discovered patterns such as ”Pathogenic Escherichia coli infection”. The Escherichia coli bacteria is one of the main reasons for that type of disease [64]. Another phenotype example is the Mental or Behavioral Dysfunction. Some of the illustrative GO terms showing the importance of our discovered patterns are, “regulation of neurological system process”, “forebrain

development", "transmission of nerve impulse", and "regulation of synaptic transmission". Furthermore we found some KEGG pathways enriched in this phenotype as well. Some examples of that are the "Neuroactive ligand-receptor interaction" and "Neurotrophin signaling" pathways. Additional examples of the interesting GO terms that are enriched in the patterns discovered by the algorithm are shown in Table 12. On the other hand, Table 13 presents some of the important KEGG pathways that are associated with the reported *DPR*s.

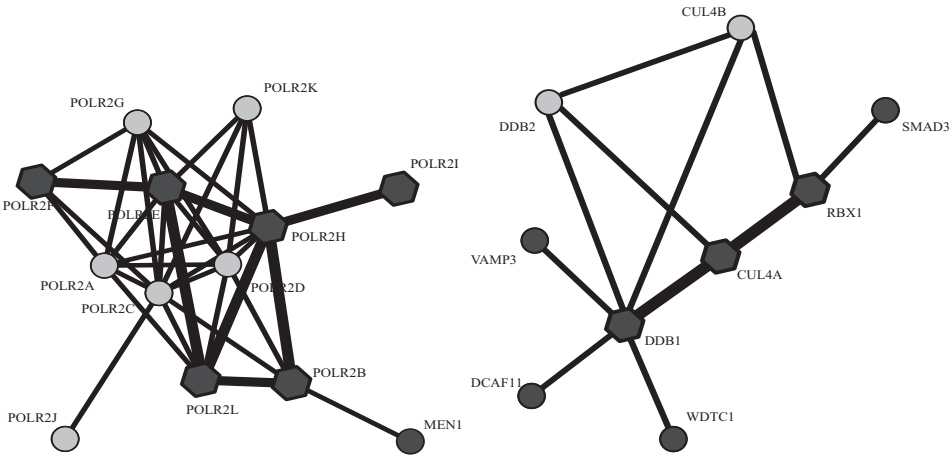


Figure 14: Two examples of patterns from the phenotypes Mental or Behavioral Dysfunction, left, and Carcinoma, right, that overlapped with two protein complexes. Hexagonal nodes are the common proteins. Dark circles are for pattern nodes and light ones are for the complex's.

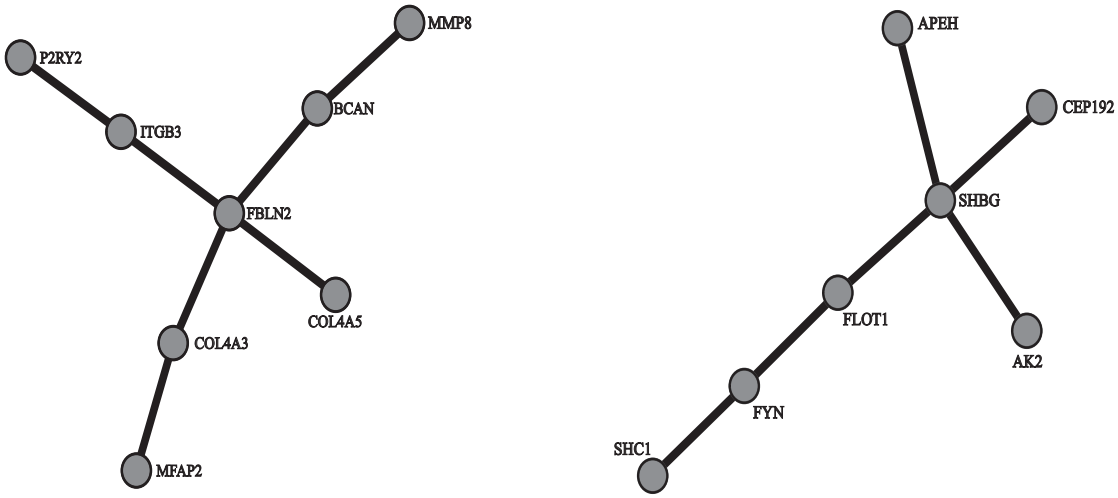
3.4.6. Complex Prediction Analysis

To further investigate the biological significance of the reported patterns, we studied the strength of these patterns as biological protein complexes candidates. The curated protein complexes are groups of connected proteins that are biologically proven to be interacting to perform a function inside the living organism. To achieve this goal, we obtained two sets of Human protein complexes. The first one is the CORUM [65] catalog that includes 1703 protein complexes after cleaning. The average number of proteins in a complex of this catalog is 4.7 proteins. In addition, we used the HPRD [58] catalog that is comprised of 1521 manually curated annotated protein complexes. The average size of a complex is similar to the case of the CORUM complexes.

The importance of the discovered *DPR*s is illustrated by the ability of these *DPR*s to match known protein complexes. Complex match does not necessarily mean that a *DPR* and a complex have 100% of proteins in common. However, they have a relatively high percentage of common genes. For the complex matching we employed the overlap score proposed in [21]. Two examples, that overlap with known complexes, are shown in Figure 14. The left part of Figure 14 shows a *DPR* that is related to the Mental or Behavioral Dysfunction phenotype. Interestingly, this pattern was enriched in pathways of neurological importance such as *hsa05016:Huntington's disease*. The matched complex is the *RNA polymerase II core complex*. The other example, right part of Figure 14, is a bout a pattern discovered during the analysis of the Carcinoma phenotype. Some GO terms that are enriched in the pattern can be linked to DNA damage and repair. Examples of such terms are *GO:0006281 DNA repair* and *GO:0006974 response to DNA damage stimulus*. This pattern has matched the complex *the Ubiquitin E3 ligase (DDB1, DDB2, CUL4A, CUL4B, RBX1)*.

3.4.7. Statistical Significance Analysis

To highlight that the discovered patterns are statistically significant and that generating random patterns would not achieve the biological importance of the patterns reported by the algorithm, we generated 5000 random patterns for every size of the sizes found in the identified *DPR*s. Hence, in any of the phenotype classes if the reported patterns have 20 different sizes, we generate 100000 random patterns for that class. Then, we tried to find the percentage of random patterns that have scores greater than or equal to the scores of the identified *DPR*s of the same size, i.e. the p-value. The average p-value, which signifies our work, was zero. Moreover, we used the t-test [66] to measure the deviation between the scores of the random patterns and the identified patterns. The average p-value in this measure was 8e-91. This proves the importance of the reported patterns statistically. On the other hand and to show that random patterns cannot sustain the biological quality of the patterns discovered by this work, we created a random protein-protein interaction network by permuting the edges between the genes of the HPRD network. We applied our algorithm using this random network and analyzed the enrichment of the reported patterns. The average enrichment, for the biological processes for example, was below 7% while



(a) (b)

Figure 15: Examples of some of the interesting patterns related to the phenotypes studied in this work. The pattern in (a) is enriched with the 'Acute Gastroentirits' GO term. The pattern in (b) is enriched with 'Mental or Behavioral Dysfunction'.

the enrichment for our patterns was around 70% as shown in Figure 13. This result confirmed the biological importance of our approach and that the reported patterns are by no means random.

3.4.8. Examples of Interesting DPRs

To emphasize the importance of this work, we illustrate that by some examples of the discovered patterns. These patterns are in strong relation with the phenotype class used to generate them. Figure 15 shows two examples of such patterns.For instance, in the Acute Gastroenteritis class, the pattern shown in Figure 15(a), is composed of the genes: **ITGB3, BCAN, FBLN2, P2RY2, COL4A3, MFAP2, COL4A5,** and **MMP8**. Employing the GO enrichment analysis for this pattern showed that several biological processes are associated with the genes of this pattern are related to the inflammatory activities. Examples of the enriched GO terms are: "activation of plasma proteins involved in acute inflammatory response", "acute inflammatory response", "regulation of inflammatory response", "inflammatory response", and "regulation of acute inflammatory response". Another example pattern for the same class, not shown in Figure 15, showed links to tumor biological processes such as "positive regulation of response to tumor cell", and "regulation

of natural killer cell mediated cytotoxicity directed against tumor cell target". In another phenotype class, Mental or behavioral Dysfunction, one of the identified patterns, has the genes **FLOT1, SHBG, APEH, FYN, AK2, CEP192,** and **SHC1** that is shown in Figure 15(b) was strongly linked to several neurological processes such "central nervous system neuron differentiation", "neuron projection morphogenesis", and "central nervous system neuron axonogenesis".

Table 14: The classification power of the DPRs illustrated by different classifiers algorithms against single gene markers.

Phenotype Class	DPR patterns vs. Single Genes Classification Results					
	#DPRs	Random Forest	Naïve Bayes	#SingleGene	Random Forest	Naïve Bayes
		F-Measure	F-Measure		F-Measure	F-Measure
Acute Gastroenteritis	2953	0.95	0.84	2953	0.62	0.70
Carcinoma	1899	0.98	0.87	1899	0.78	0.90
Cell or Molecular Dysfunction	2242	0.93	0.80	2242	0.73	0.72
Congenital Abnormality Disease or Syndrome	1897	0.93	0.92	1897	0.72	0.86
Mental or Behavioral Dysfunction	3330	0.91	0.83	3330	0.65	0.73
Neoplastic Process	1629	0.92	0.93	1629	0.71	0.89

3.4.9. Classification Performance of DPRs

In this part of our work we assessed the classification power of the reported patterns and to what extent they can be considered as good markers for a phenotype class. To accomplish that, we considered that every pattern is a feature. For every pattern, the activity vector was calculated based on the attribute dysregulation matrix. For instance, in the Acute Gastroenteritis phenotype, our algorithm has reported 2953 patterns/features. Using the UMLS annotation approach we categorized 28 datasets out of the 88 datasets to be marked with that phenotype. We denote these 28 sets as the positive group. The remaining 60 datasets were denoted as the negative group. To perform the classification experiments, we used two classifiers provided by the WEKA tool [67]; namely, Random Forest, and Naive Bayes. In Table 14 we show the F-measure results of these classifiers. The F-measure is $\frac{2 \times Precision \times Recall}{Precision + Recall}$ and it is a measure for evaluating accuracy. Precision

is the ratio of true positives to the total number of all datasets classified as positive/phenotype. Recall is the ratio of true positives to the total number of positive/phenotype datasets. We used the default settings of the WEKA tool for each of the classification algorithms, i.e. 10-fold cross validation. For comparison, we ranked the genes based on the information gain score. We selected the top ranked genes as single markers for classification. For each dataset, the number of single gene markers selected is equal to the number of DPRs reported by the proposed algorithm. From Table 14, the average F-measure across all phenotypes of the Random Forest classifier, built using the DPRs as attributed, is above 90%. Furthermore, Table 14 shows the clear advantage of using the *DPR*s as markers over single gene markers.

3.5. Conclusion

In this work we developed an approach for discovering Dysregulated Phenotype-Related patterns of connected genes, *DPR*s. The approach proved to be effective in many facets. Biologically, the functional enrichment analysis showed that the *DPR*s discovered by our algorithm are significant and they are strongly linked to the phenotypes considered in the study. This is achieved by examining which GO terms and KEGG pathways are enriched in the discovered *DPR*s. Moreover, the reported *DPR*s have overlap with manually curated protein complexes which are biological-proven to be protein modules involved in common functions in the living organism. Additionally, we studied the classification performance of the reported *DPR*s. The results indicate that the *DPR*s are good markers for the studied phenotypes. Furthermore, we assessed the statistical significance of the discovered patterns and the results confirmed that the reported *DPR*s are far from being discovered by chance. Discovering phenotype-related gene sets can help the researchers in the medical field by assisting them focus on small number of genes that are related to the phenotype or disease they are studying.

CHAPTER 4. CONCLUSIONS

In this last part of the dissertation we conclude the work presented so far and summarize our contributions.

In Chapter 2, we accomplished number of contributions when introducing the problem of mining maximal cohesive patterns. We have proposed an algorithm to tackle the problem of mining maximal cohesive induced subgraphs and maximal cohesive patterns. On that, we have illustrated the effectiveness of integrating constraints from several data sources, such as phenotypes and evolutionary profiles, with protein interaction networks. With this integration the search can be guided to discover interesting patterns. The performed experimental analysis on Yeast and Human datasets has proved the quality of the proposed approach by assessing the overlap of the discovered subnetworks with known biological complexes and pathways. Moreover, GO enrichment analysis showed that the discovered subnetworks are biologically significant.

In Chapter 3, we propose an approach for mining dysregulated patterns by integrating gene expression datasets and PPI network. The problem we address here is similar to the problem of discovering subnetwork biomarkers for gene expression classification. However, instead of classifying samples, we mine sets of genes that distinguish between the datasets that are labeled with Unified Medical Language Systems, UMLS, concepts. The approach we developed aimed at the discovery of Dysregulated Phenotype-Related Patterns, DPRs. The process follows the greedy pattern growth approach. For that goal, we have created dysregulation profiles for genes in the protein-protein interaction network. A gene dysregulation profile captures the dysregulation of the gene in the 88 datasets. We employed the physical interaction network in our work to add significance to the mined gene patterns. We have demonstrated the biological relevance of the reported DPRs to the studied phenotypes by illuminating the biological context through Gene Ontology enrichment, KEGG pathway enrichment, and the overlap with known protein complexes. Moreover, The reported DPR patterns have proved to have high classification power when these patterns are used as classification features.

BIBLIOGRAPHY

[1] Christian Burks. Molecular biology database list. *Nucleic Acids Research*, 27(1):1–9, 1999.

[2] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, Nov 2004.

[3] Javier De Las Rivas and Celia Fontanillo. Proteinprotein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, 2010.

[4] Luke Hakes, John Pinney, David Robertson, and Simon Lovell. Protein-protein interaction networks and biology – what's the connection? *Nature Biotechnology*, 26:69 – 72, 2008.

[5] Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(88), 2007.

[6] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[7] Junzhong Ji, Aidong Zhang, Chunnian Liu, Xiaomei Quan, and Zhijun Liu. Survey: Functional module detection from protein-protein interaction networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(2):261–277, 2014.

[8] Radha Shyamsundar, Young Kim, John Higgins, Kelli Montgomery, Michelle Jorden, Anand Sethuraman, Matt van de Rijn, David Botstein, Patrick Brown, and Jonathan Pollack. A dna microarray survey of gene expression in normal human tissues. *Genome Biology*, 6(3):R22, 2005.

[9] Glen Traver Hart, Arun Ramani, and Edward Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120.1–120.9, 2006.

[10] Leland Hartwell, John Hopfield, Stanislas Leibler, and Andrew Murray. From molecular to modular cell biology. *Nature*, 402:C47 – C52, 1999.

[11] Pavol Jancura, Eleftheria Mavridou, Enrique Carrillo-de Santa Pau, and Elena Marchiori. A methodology for detecting the orthology signal in a ppi network at a functional complex level. *BMC bioinformatics*, 13(Suppl 10):S18 – S31, 2012.

[12] Hui Ge, Zhihua Liu, George Church, and Marc Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nature Genetics*, 29:482–486, Dec 2001.

[13] Andreas Hahn, Jorg Rahnenfuhrer, Priti Talwar, and Thomas Lengauer. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6(1):112–123, 2005.

[14] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

[15] Ulrik de Lichtenberg, Lars Juhl Jensen, Sren Brunak, and Peer Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, 2005.

[16] Shinichiro Wachi, Ken Yoneda, and Reen Wu. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, 21(23):4205–4208, 2005.

[17] Clara Pizzuti and Simona Rombo. Algorithms and tools for proteinprotein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.

[18] Michelle Girvan and Mark Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[19] Stijn Van Dongen. A new cluster algorithm for graphs. *Technical Report No. INS-R0012*, 2000.

[20] Stijn Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008.

[21] Gary Bader and Christopher Hogu. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.

[22] Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(Suppl 1):S136–S144, 2002.

[23] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140 – 150, October 2007.

[24] Andreas Hahn, Jorg Rahnenfuhrer, Priti Talwar, and Thomas Lengauer. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6(1):112 – 123, 2005.

[25] Rami Alroobi and Saeed Salem. Discovering dysregulated phenotype-related gene patterns. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB 2014, pages 524–532, 2014.

[26] Elisabeth Georgii, Sabine Dietmann, Takeaki Uno, Philipp Pagel, and Koji Tsuda. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, 25(7):933–940, 2009.

[27] Recep Colak, Flavia Moser, Jeffrey Shih-Chieh Chu, Alexander Schnhuth, Nansheng Chen,

and Martin Ester. Module discovery by exhaustive search for densely connected, co-expressed regions in biomolecular interaction networks. *PLoSe ONE*, 5(10):e13348, 2010.

[28] Daniel Hanisch, Alexander Zien, Ralf Zimmer, and Thomas Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(Suppl 1):S145–S154, 2002.

[29] Yuanyuan Tian, Richard Hankins, and Jignesh Patel. Efficient aggregation for graph summarization. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 567–580, New York, 2008.

[30] Igor Ulitsky and Ron Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1(1):8, 2007.

[31] Ioannis Maraziotis, Konstantina Dimitrakopoulou, and Anastasios Bezerianos. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinformatics*, 8(1):408, 2007.

[32] Young-Rae Cho, Woochang Hwang, and Aidong Zhang. Efficient modularization of weighted protein interaction networks using k-hop graph reduction. In *Sixth IEEE Symposium on BioInformatics and BioEngineering*, BIBE 2006, pages 289–298, Oct 2006.

[33] Hongchao Lu, Baochen Shi, Gaowei Wu, Yong Zhang, Xiaopeng Zhu, Zhihua Zhang, Changning Liu, Yi Zhao, Tao Wu, Jie Wang, and Runsheng Chen. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochemical and Biophysical Research Communications*, 345(1):302 – 309, 2006.

[34] Ronald Jansen, Dov Greenbaum, and Mark Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12(1):37–46, 2002.

[35] Flavia Moser, Recep Colak, Arash Rafiey, and Martin Ester. Mining cohesive patterns from

graphs with feature vectors. In *Proceedings of the Ninth SIAM International Conference on Data Mining*, SIAM 2009, pages 593–604, 2009.

[36] Rami Alroobi, Syed Ahmed, and Saeed Salem. Mining maximal cohesive induced subnetworks and patterns by integrating biological networks with gene profile data. *Interdisciplinary Sciences: Computational Life Sciences*, 5(3):211–224, 2013.

[37] Salim Chowdhury and Mehmet Koyuturk. Identification of coordinately dysregulated subnetworks in complex phenotypes. In *Pacific Symposium on Biocomputing*, pages 133–144, 2010.

[38] Phuong Dao, Recep Colak, Raheleh Salari, Flavia Moser, Elai Davicioni, Alexander Schnhuth, and Martin Ester. Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics*, 26(18):i625–i631, 2010.

[39] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS Computional Biology*, 4(11):e1000217, Nov 2008.

[40] Silpa Suthram, Joel Dudley, Annie Chiang, Rong Chen, Trevor Hastie, and Atul Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology*, 6(2):e1000662, 2010.

[41] Karam Gouda and Mohammed Zaki. Genmax: An efficient algorithm for mining maximal frequent itemsets. *International Journal of Data Mining and Knowledge Discovery*, 11(3):223–242, Nov 2005.

[42] Thomas Cormen, Charles Leiserson, Ronald Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.

[43] Nizar Batada, Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz,

Laurence Hurst, and Mike Tyers. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biology*, 4(10):e317, 2006.

[44] Gary Bader, Doron Betel, and Christopher Hogue. Bind: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1):248–250, 2003.

[45] Teresa Reguly, Ashton Breitkreutz, Lorrie Boucher, Bobby-Joe Breitkreutz, Gary Hon, Chad Myers, Ainslie Parsons, Helena Friesen, Rose Oughtred, Amy Tong, Chris Stark, Yuen Ho, David Botstein, Brenda Andrews, Charles Boone, Olga Troyanskya, Trey Ideker, Kara Dolinski, Nizar Batada, and Mike Tyers. Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae. *Journal of Biology*, 5(4):11, 2006.

[46] Ioannis Xenarios, Lukasz Salwnski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. DIP: The Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002.

[47] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. Mint: A Molecular INTeraction database. *FEBS Letters*, 513(1):135 – 140, 2002.

[48] Hans-Werner Mewes, Dmitrij Frishman, Ulrich Güldener, Gertrud Mannhaupt, Klaus Mayer, Martin Mokrejŝ, Burkhard Morgenstern, Martin Münsterkötter, and Sean Rudd. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, 30(1):31–34, 2002.

[49] Aimëe Marie Dudley, Daniel Maarten Janse, Amos Tanay, Ron Shamir, , and George McDonald Church. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular Systems Biology*, 1:2005.0001 – 2005.0011, 2005.

[50] Gabriel Östlund, Thomas Schmitt, Kristoffer Forslund, Tina Köstler, David Messina, Sanjit Roopra, Oliver Frings, and Erik Sonnhammer. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(suppl 1):D196–D203, 2010.

[51] Audrey Gasch, Paul Spellman, Camilla Kao, Orna Carmel-Harel, Michael Eisen, Gisela Storz, David Botstein, and Patrick Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*, 11(12):4241–4257, December 2000.

[52] Shuye Pu, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, 2009.

[53] Suraj Peri, Daniel Navarro, Troels Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, Nandan Deshpande, Shubha Suresh, Vidya Niranjan, Naveen Talreja, Mary Joy, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, Chandan KumarSinha, Krishna Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32(suppl 1):D497–D501, 2004.

[54] Andrew Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael Cooke, John Walker, and John Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

[55] Eric Sayers, Tanya Barrett, Dennis Benson, Evan Bolton, Stephen Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Lewis Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David Lipman, Zhiyong Lu, Thomas Madden, Tom Madej, Donna Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Kim Pruitt, Gregory Schuler, Edwin Sequeira, Stephen Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana Tatusova, Lukas Wagner, Yanli Wang, John Wilbur, Eugene Yaschenko, and Jian Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39:D38 – D51, 2011.

[56] Barbara Gross, Steven Melford, and Stephen Watson. Evidence that phospholipase C-y2 interacts with SLP-76, Syk, Lyn, LAT and the Fc receptor y-chain after stimulation of the collagen receptor glycoprotein VI in human platelets. *European Journal of Biochemistry*, 263(3):612–623, 1999.

[57] Barry Zeeberg, Weimin Feng, Geoffrey Wang, May Wang, Anthony Fojo, Margot Sunshine, Sudarshan Narasimhan, David Kane, William Reinhold, and Samir Lababidi. Gominer: A resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, 2003.

[58] Suraj Peri, Troels Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, Nandan Deshpande, Shubha Suresh, Vidya Niranjan, Naveen Talreja, Mary Joy, Minal Menezes, Dipanwita Roy Choudhury, Neelanjana Ghosh, Sreenath Chandran, Sujatha Mohan, Chandra Kiran Jonnalagadda, Chandan KumarSinha, Krishna Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32(suppl 1):D497–D501, 2004.

[59] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[60] Guy Divita, Allen Browne, and Russell Loane. dTagger: a POS tagger. *American Medical Informatics Association Annual Symposium Proceedings*, pages 200 – 203, 2006.

[61] Kaihong Liu, Wendy Chapman, Rebecca Hwa, and Rebecca Crowley. Heuristic Sample Selection to Minimize Reference Standard Training Set for a Part-Of-Speech Tagger. *Journal of the American Medical Informatics Association*, 14(5):641–650, September 2007.

[62] Alan Aronson and Francois Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[63] Da Huang, Brad Sherman, Qina Tan, Jack Collins, Gregory Alvord, Jean Roayaei, Robert Stephens, Michael Baseler, Clifford Lane, and Richard Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183, 2007.

[64] Junkal Garmendia, Gad Frankel, and Valrie Crepin. Enteropathogenic and enterohemorrhagic escherichia coli infections: Translocation, translocation, translocation. *Infection and Immunity*, 73(5):2573–2585, 2005.

[65] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Werner Mewes. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, 38(suppl 1):D497–D501, 2010.

[66] Ronald Aylmer Fisher. Applications of student's distribution. *Metron*, 5:90–104, 1925.

[67] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. The weka data mining software: An update. *SIGKDD Exploration Newsletter*, 11(1):10–18, 2009.