Graduate School ETD Form 9
(Revised 01/14)

# PURDUE UNIVERSITY
## GRADUATE SCHOOL
## Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Chayaporn Suphavilai

Entitled
Computational Development of Regulatory Gene Set Networks for Systems Biology Applications

For the degree of Master of Science

Is approved by the final examining committee:

Jake Y. Chen

Shiaofen Fang

Mohammad Al Hasan

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Jake Y. Chen

Approved by Major Professor(s): _____

_____

Approved by: Shiaofen Fang 03/28/2014

Head of the Department Graduate Program Date

# COMPUTATIONAL DEVELOPMENT OF REGULATORY GENE SET NETWORKS FOR SYSTEMS BIOLOGY APPLICATIONS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Chayaporn Suphavilai

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2014

Purdue University

Indianapolis, Indiana

# ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Dr. Jake Y Chen for the support of my MS research, for his guidance, motivation and extensive knowledge.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Shiaofen Fang and Dr. Mohammad Al Hasan for their encouragement, insightful comments, and questions.

My sincere thanks also goes to Dr.Liugen Zhu for helping me on network visualization implementation and Dr. Xiaogang Wu for his help and guidance.

I would like to thank all my labmates in Discovery Informatics and Computing Laboratory.

Last but not the least, I would like to thank my family: Chaisuree Suphavilai, Chusri Bulyalert and Phornchai Suphavilai for their encouragement and support.

TABLE OF CONTENTS

Page

LIST OF TABLES ..... v

LIST OF FIGURES ..... vi

ABSTRACT ..... viii

CHAPTER 1. INTRODUCTION ..... 1

1.1 Introduction ..... 1

1.2 Constructing a Regulatory Gene Set Network ..... 6

1.3 Pathway and Annotated Gene-set Electronic Repository (PAGER) ..... 8

1.4 Contribution Summary ..... 10

CHAPTER 2. METHODS ..... 13

2.1 Design ..... 13

2.2 Constructing a Regulatory Gene Set Network ..... 13

2.2.1.1 Gene Set Data Sources ..... 14

2.2.1.2 Gene Regulation Data Sources ..... 15

2.2.2 Constructing Gene Set Networks ..... 16

2.2.2.1 Co-membership Gene Set Networks ..... 16

2.2.2.2 Regulatory Gene Set Networks ..... 17

2.2.2.3 Hypergeometric Distribution ..... 19

2.2.2.4 Benjamini–Hochberg Procedure ..... 19

2.2.3 A Disease Specific Gene Set Network ..... 20

2.2.4 Network Analysis ..... 20

2.2.4.1 Degree Centrality ..... 21

2.2.4.2 Betweenness Centrality ..... 21

2.2.4.3 Closeness Centrality ..... 22

Page

2.2.4.4 Network Comparison ..... 23
2.3 Pathway and Annotated Gene-set Electronic Repository (PAGER) ..... 23
2.3.1 Data Sources ..... 24
2.3.1.1 Gene Set Data Sources ..... 24
2.3.1.2 Gene Interaction and Gene Regulation Data Sources ..... 24
2.3.2 Database Design ..... 25
2.3.3 Implementation ..... 26
CHAPTER 3. RESULTS ..... 28
3.1 Constructing a Regulatory Gene Set Network ..... 28
3.1.1 Gene Set Networks ..... 28
3.1.2 Comparing the KEGG Regulatory and the KEGG Co-membership Network ..... 32
3.1.3 Comparison of the KEGG Regulatory and the KEGG Co-enrichment Networks ..... 38
3.1.4 A Disease Specific Regulatory Gene Set Network ..... 41
3.2 Pathway and Annotated Gene-set Electronic Repository (PAGER) ..... 46
3.2.1 Data Integration ..... 46
3.2.2 Gene Set and Gene Set Networks ..... 48
3.2.3 PAGER Features ..... 49
3.2.4 PAGER Use cases ..... 54
3.2.4.1 Searching Genes and Gene Sets by Terms ..... 54
3.2.4.2 Searching Gene Sets by a List of Genes ..... 57
3.2.4.3 Constructing Gene Set networks ..... 59
3.2.4.4 Generating Expanded Gene Set Networks ..... 61
3.2.4.5 Viewing Gene Networks of Genes inside a Gene Set ..... 62
3.2.4.6 Constructing Disease Specific Gene Set Networks ..... 63
CHAPTER 4. CONCLUSION ..... 65
REFERENCES ..... 69

# LIST OF TABLES

Table Page

Table 2.1 Human gene regulation data sources ..... 16

Table 2.2 List of gene set data sources ..... 25

Table 3.1 Summary of co-membership gene set networks and regulatory gene set networks for five gene set collections ..... 29

Table 3.2 Top 10 most significant regulatory edges in the KEGG regulatory gene set network where gene set 1 regulates gene set 2 ..... 31

Table 3.3 Top 10 most significant co-membership edges in the KEGG co-membership gene set network ..... 32

Table 3.4 Top 10 highest degree centrality pathway of KEGG co-membership gene set network (left) and KEGG regulatory gene set network (right) ..... 34

Table 3.5 Top 10 highest degree centrality pathway of the KEGG exclusive regulatory gene set network ..... 38

Table 3.6 Top 10 highest outdegree centrality pathways (right) and top 10 highest indegree centrality pathways (left) of the KEGG exclusive regulatory gene set network 38

Table 3.7 Top 10 highest degree centrality gene sets among 261 AD gene sets ..... 43

Table 3.8 Top 10 highest closeness centrality (out) gene sets ..... 45

Table 3.9 Top 10 highest betweenness centrality gene sets ..... 45

Table 3.10 Gene set data sources ..... 47

# LIST OF FIGURES

Figure Page

Figure 1.1 PAGER screenshots ..... 8

Figure 2.1 PAGER ER Diagram ..... 26

Figure 3.1 Correlation between degree centrality of the KEGG co-membership network and degree centrality of the KEGG regulatory network. ..... 34

Figure 3.2 (M) KEGG co-membership network and (R) KEGG regulatory network. ..... 36

Figure 3.3 R-M is KEGG exclusive regulatory network. Node colors represents different classes of pathways. ..... 37

Figure 3.4 (A) Number of edges in a KEGG co-enrichment network which shared with a KEGG co-membership and random networks. (B) Number of edges in a KEGG co-enrichment network that shared with a KEGG regulatory and random networks. ..... 41

Figure 3.5 AD regulatory gene set network ..... 42

Figure 3.6 A sub-network of AD regulatory gene set network which contains only the top 10 highest DC gene sets from Table 3.7. ..... 44

Figure 3.7 PAGER work flow ..... 49

Figure 3.8 PAGER work flow ..... 50

Figure 3.9 PAGER gene set network features ..... 52

Figure 3.10 PAGER gene network features ..... 53

Figure 3.11 PAGER home page ..... 55

Figure 3.12 PAGER results of searching by “non small cell lung” ..... 55

Figure 3.13 PAGER results of searching by “BRAF” ..... 55

Figure 3.14 PAGER displays a list of result gene sets ..... 56

Figure 3.15 PAGER displays a list of gene sets in Gene Set Box ..... 57

Figure Page

Figure 3.16 PAGER displays a list of gene sets which relate to the list of 94 genes..... 58
Figure 3.17 PAGER displays a regulatory gene set network of 28 gene sets ..... 59
Figure 3.18 PAGER displays a co-membership gene set network of 28 gene sets..... 60
Figure 3.19 PAGER displays a detail of Non small cell lung cancer gene set..... 62
Figure 3.20 A gene network of Non small cell lung cancer gene set..... 63
Figure 3.21 A gene network of Non small cell lung cancer gene set..... 64

# ABSTRACT

Suphavilai, Chayaporn. M.S., Purdue University, May 2014. Computational Development of Regulatory Gene Set Networks for Systems Biology Applications. Major Professor: Jake Chen.

In systems biology study, biological networks were used to gain insights into biological systems. While the traditional approach to studying biological networks is based on the identification of interactions among genes or the identification of a gene set ranking according to differentially expressed gene lists, little is known about interactions between higher order biological systems, a network of gene sets. Several types of gene set network have been proposed including co-membership, linkage, and co-enrichment human gene set networks. However, to our knowledge, none of them contains directionality information. Therefore, in this study we proposed a method to construct a regulatory gene set network, a directed network, which reveals novel relationships among gene sets. A regulatory gene set network was constructed by using publicly available gene regulation data. A directed edge in regulatory gene set networks represents a regulatory relationship from one gene set to the other gene set. A regulatory gene set network was compared with another type of gene set network to show that the regulatory network provides additional information. In order to show that a regulatory gene set network is useful for

understand the underlying mechanism of a disease, an Alzheimer's disease (AD) regulatory gene set network was constructed.

In addition, we developed Pathway and Annotated Gene-set Electronic Repository (PAGER), an online systems biology tool for constructing and visualizing gene and gene set networks from multiple gene set collections. PAGER is available at http://discern.uits.iu.edu:8340/PAGER/. Global regulatory and global co-membership gene set networks were pre-computed. PAGER contains 166,489 gene sets, 92,108,741 co-membership edges, 697,221,810 regulatory edges, 44,188 genes, 651,586 unique gene regulations, and 650,160 unique gene interactions. PAGER provided several unique features including constructing regulatory gene set networks, generating expanded gene set networks, and constructing gene networks within a gene set.

However, tissue specific or disease specific information was not considered in the disease specific network constructing process, so it might not have high accuracy of presenting the high level relationship among gene sets in the disease context. Therefore, our framework can be improved by collecting higher resolution data, such as tissue specific and disease specific gene regulations and gene sets. In addition, experimental gene expression data can be applied to add more information to the gene set network. For the current version of PAGER, the size of gene and gene set networks are limited to 100 nodes due to browser memory constraint. Our future plans is integrating internal gene or proteins interactions inside pathways in order to support future systems biology study.

# CHAPTER 1. INTRODUCTION

## 1.1 Introduction

Today a large amount of high throughput data from biological experiments has been produced. High-throughput sequencing and gene profiling techniques have transformed biological research by enabling comprehensive understanding of a biological system [26]. In consequence, many differentially expressed gene lists were generated for different kinds of experiments such as disease gene expression profile and drug sensitivity gene expression profile. This situation presented a new challenge of extracting meaning from the long list of genes. The traditional approach to study biological network is based on identification of interaction among genes or proteins. However, several biological processes often coordinately function together and higher level relationships among biological processes have not been well studied yet. Therefore, networks of gene sets have been proposed in order to reveal the higher level relationships among gene sets where a gene set is a group of genes that belong to the same pathway, share common biological processes, or belong to the same disease.

Pathway analysis is a method for gaining insight into the underlying biology of differentially expressed genes and proteins. It analyzes a list of differentially expressed genes and returns significant pathways, which highly relate to the gene list. A pathway

can be considered as a gene set which contains all genes or proteins in the pathway. Therefore, in some scenarios, pathway analysis and gene set analysis are the same analysis. However, in some cases the analysis of pathway and the analysis of gene set can be different. For example, a biological pathway consists of a series of interactions among molecules, but a gene set is a set of genes which might not have information of interactions among its genes. In this case, if an analysis requires information of the interactions among molecules, the analysis might not appropriate to analyze gene sets.

Typically, researchers use several techniques of pathway analysis to reveal a novel insight into their gene lists. Pathways or gene sets information were provided by several data sources such as Pathway Interaction Database (PID)[35] and MSigDB [28]. These experimental validated pathways or gene sets were used as a knowledge base. The first simple approach of pathway analysis is evaluating the number of genes from a differentially expressed gene list which are found in a particular pathway. The more advanced method is calculating the gene-level statistic for all genes in a pathway and aggregating the gene-level statistics into a single pathway-level statistic. In some studies, a pathway is considered as a gene set, which is a group of genes that share common biological function or regulation. For example, Gene set enrichment analysis (GSEA) [36] is a method to calculate gene set-level statistics. Researchers input their gene expression data from the experiment, GSEA then calculates gene-level statistics and use the gene-level statistics to calculate an enrichment score for each gene set. In addition, using gene set-level statistics has an advantage because significant analysis at single gene level

suffers from a limited number of sample and noise [14]. Gene set based methods have also been developed to investigate phenotypic changes at the pathway level. For example, genetic perturbations have been implicated for initiation and progression of cancer and these perturbations are most likely reflected by the altered expression of sets of genes or pathways [11]. The next generation of pathway analysis also incorporates pathway topology and interaction of genes or proteins inside a pathway. By using both gene expression data and pathway topology, researchers obtain a better ranking of gene sets corresponding to their differentially expressed gene lists.

In systems biology study, two levels of biological network, gene network and gene set network, were used to gain insights into biological systems. The traditional approach to studying complex biological networks is based on the identification of interactions among genes or the identification of gene set ranking according to differentially expressed gene lists. Little is known about interactions between higher order biological systems, a network of gene sets. Several studies proposed methods to construct gene set networks. Yong Li, et al. constructed a pathway crosstalk network and a linkage network in order to understand the relationship between pathways [27]. Dikla, et al. proposed a methodology for gleaning patterns of interactions between biological processes by analyzing protein-protein interactions, transcriptional co-expressions and genetic interactions [10]. In addition, a gene set network can be used for understanding diseases. For example, Zhi-Ping, et al. proposed a network-based systems biology approach to detect the crosstalks among AD related pathways and the dysfunctions in the six brain regions of AD patients

[29]. Kelder, et al. proposed an analysis approach to study interactions between pathways in a mouse by integrating gene and protein interaction networks, biological pathway information and high-throughput data [24].

Recently, Jignesh, et al. proposed methods to construct multi-edge gene set networks to reveal insights into global relationships between biological themes or gene sets [32]. The multi-edge network consists of three types of edges: co-membership, linkage, and co-enrichment. Co-membership gene set networks (M) connect gene sets if there is a significant number of shared genes between the two gene sets. Linkage gene set networks (L) connect a pair of gene sets if there is a significant number of gene or protein interactions between the unique genes of the two gene sets. Co-enrichment gene set networks (E) connect gene sets if there is a significant number of experiments where the unique genes of the two gene sets are enriched together.

Even though several types of gene set networks for humans have been proposed, to our knowledge, none of them contains directionality information in a gene set network. Therefore, in this study we proposed a method to construct a regulatory gene set network (R), a directed network, which reveals novel relationships among gene sets together with directionality information. A regulatory gene set network was constructed by using publicly available gene regulation data. A directed edge in regulatory gene set networks represents a regulatory relationship from one gene set to the other gene set. The significant value of each edge was computed by using hypergeometric distribution and was corrected for multiple comparison using BH-procedure [3]. By comparing a regulatory

gene set network (R) with a linkage gene set network (L), which is constructed from protein interactions discarding directionality of the interactions, a regulatory gene set network contains higher resolution of knowledge. Our hypothesis is a regulatory gene set network can reveal novel gene set relationships and provide complement knowledge to the existing types of gene set networks such as co-membership and linkage. Therefore, a regulatory gene set network can facilitate system biology to understand biological phenomena through global regulatory gene set network construction and disease specific regulatory gene set network construction.

Moreover, several tools have been developed to enable identification of gene set relationships. For example, Sudhir, et al. developed HPD database to enable study of human pathway crosstalk networks [7]; Huang, et al. developed PAGED, a pathway and gene-set enrichment database to enable molecular phenotype discoveries [19]; and Jignesh, et al. developed Metanet, a web tool for constructing multi-edge gene set networks which consist of three types of edges: co-membership, link-age, and co-enrichment, separately for each gene set collection [32]. While HPD focused only on human pathway networks, Metanet and PAGED allowed researchers to construct a gene set network. As previously described, a gene set represents a set of genes belonging to the same pathway, the same biological process, and the same disease. Therefore, Metanet and PAGED provided more coverage and flexibility of gene set network construction. However, Metanet constructed a gene set network based only a single gene set collection, preventing researchers from searching for relationships among gene sets from different

data sources; and PAGED does not allow researchers to visualize a network. In this study, we developed Pathway and Annotated Gene-set Electronic Repository (PAGER) in order to enable researchers to identify relationships among gene sets from different data sources. PAGER also supported constructing a regulatory gene set network (R), which is a new type of gene set network we proposed, and supported constructing gene interaction and gene regulatory network of genes inside a gene set.

In summary, this study consists of two parts, constructing a regulatory gene set network and developing PAGER to allow users to construct several types of gene and gene set networks. A method for constructing a regulatory gene set network was presented, and a method for developing PAGER to allow users to construct several types of gene and gene set networks was described later.

### 1.2 Constructing a Regulatory Gene Set Network

Gene set data and gene regulation data were collected from several reliable data sources. Five co-membership gene set networks were separately construct for each gene set collection including KEGG [23], Reactome, Go biological process, GO cellular component, and GO molecular function. Because a co-membership gene set network is a common type of a gene set network that has been constructed in several studies [27,32], a co-membership gene set network was used as a baseline to compare with our new type of gene set network, a regulatory gene set network. A regulatory gene set network connects a pair of gene sets if there is a significant number of gene regulations between the unique genes of the gene sets. A regulatory gene set network and a co-membership network

were compared in order to show that regulatory gene set network can reveal novel gene set relationships and provide complement knowledge to the co-membership gene set network. The directionality information provided by a regulatory gene set network was used to search for significant gene sets in the network.

Our KEGG regulatory gene set network was validated with the KEGG co-enrichment gene set network obtained from Jignesh, et al. [32]. The KEGG co-enrichment gene set network can be used for validation because it was constructed from several differentially expressed gene lists obtained from several biological experiments. Because a pair of gene set in regulatory network are connected if there is a significant number of gene regulations from one gene set to the other gene set, these gene sets should be enriched together. Therefore, the number of shared edges between the co-enrichment network and the regulatory network should be significantly high.

In addition, a gene set network was often constructed as a global network. In a global gene set network, all gene sets in a gene set collection were presented as nodes in the network. However, a gene set network specific to a disease can also be constructed in order to understand the underlying mechanism of the disease. In this study, Alzheimer's disease (AD) was chosen to use as a case study for our regulatory gene set network. In contrast to constructing a global network for each gene set collection, all gene sets from different collections were combined into a single collection. We selected AD related gene set form the combined gene set collection and constructed an AD specific regulatory gene

set network. Finally, we showed that the AD specific regulatory gene set network is useful for understanding AD.

### 1.3 Pathway and Annotated Gene-set Electronic Repository (PAGER)

PAGER is an online platform for searching gene sets and constructing gene set networks to reveal insights into biological systems. Gene set data were collected to increase the coverage of PAGER. This first version of PAGER contains 166,489 gene sets integrated from 10 different gene set data sources. A global regulatory and a global co-membership gene set networks were pre-constructed. PAGER also allowed users to construct gene interaction and gene regulation networks of genes inside a gene set.

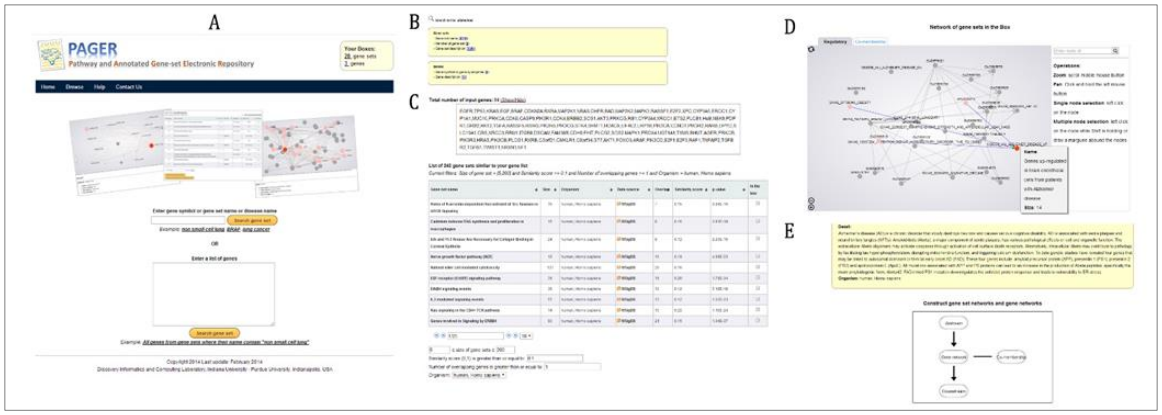



Figure 1.1 PAGER screenshots

On PAGER home page (Figure 1.1A), users can search for genes and gene sets by terms such as a disease name and a gene symbol. The result gene sets can be added into Gene Set Box, a space for users to save their gene sets. A status of Gene Box and Gene Set Box were display on the top right of the page. When users enter terms, the number genes and gene sets which matched, in different aspects, to the terms were displayed in overall

result page (Figure 1.1B). PAGER also allows users to search for gene sets which related to their gene lists. Users entered a list of genes, then PAGER returned a list of related gene sets and p-values (Figure 1.1C). The result gene sets can be sorted by name, size of gene set, number of genes found in the list, and p-value. Users can add gene sets to Gene Set Box and construct and visualize a co-membership and a regulatory gene set networks of gene sets in Gene Set Box (Figure 1.1D). Users can select a gene set node in the networks to see more detail (Figure 1.1E). On a gene set detail page, users can see detail of a gene set and construct a gene interaction network and a gene regulation network of genes inside a gene set.

Furthermore, PAGER allows users to expand gene set networks from a gene set of interest in the existing network. The ability to expand gene set networks has not provided before in other existing tools. For each selected gene set, PAGER enables users to construct three types of expanded gene set network, including networks of upstream, downstream, and co-membership gene sets. Upstream gene sets are gene sets which regulate the selected gene set; downstream gene sets are gene sets which are regulated by the selected gene set; and co-membership gene sets are gene sets which share significant number of genes with the selected gene set. Constructing expanded gene set network allows users to find more gene set related to their study. For example, users constructed gene set networks of 10 gene sets and were interested in one of the 10 gene sets which was regulated by many gene sets. In this example, PAGER allows users to further search for downstream gene set of the gene set of interest. Because PAGER enables users to construct gene

networks, users can also construct expanded gene networks from a gene as well. Three types of expanded gene networks provided by PAGER are networks of upstream, downstream, and sibling genes. Upstream genes are genes which regulate the selected gene; downstream gene are gene which are regulated by the selected gene; and sibling genes are gene which interact with the selected gene.

Finally, we presented six use cases in order to show how PAGER is useful for systems biology study. The six use cases are searching genes and gene sets by a disease name, searching gene sets by a list of genes, constructing a regulatory and a co-membership gene set networks, generating expanded gene set networks for a particular gene set, constructing a gene interaction and a gene regulation networks of a gene set, and constructing disease specific gene set networks

## 1.4 Contribution Summary

- We proposed a method to construct a new type of gene set network, a regulatory gene set network, which reveals novel relationships among gene sets together with directionality information.
- A directed edge in regulatory gene set networks represents a regulatory relationship from one gene set to the other gene set, which has never been reveal before by other types of gene set networks.
- By comparing and validating a regulatory gene set network with other existing types of gene set networks, we showed that a regulatory gene set network provides complement knowledge and useful information.

- A regulatory gene set network facilitates understanding biological phenomena and underlying mechanism of a disease through global and disease specific regulatory gene set network constructions.
- We developed Pathway and Annotated Gene-set Electronic Repository (PAGER) to enable users to construct and visualize regulatory and co-membership gene set networks from multiple gene set collections.
- PAGER contains 166,489 gene sets, 92,108,741 co-membership edges, 697,221,810 regulatory edges, 44,188 genes, 651,586 unique gene regulations, and 650,160 unique gene interactions.
- Typically, a systems biology tool supports either gene networks construction or gene set networks construction. In contrast, PAGER allows users to construct gene interaction and gene regulation networks of genes inside a gene set. This feature enable users to include both levels of biological networks, gene and gene set networks, into their study.
- PAGER enables users to construct three types of expanded gene set network including networks of upstream, downstream, and co-membership gene sets. Constructing expanded gene set networks allows users to find more important gene sets related to their study.
- PAGER also allows users to construct expanded gene networks from a gene including networks of upstream, downstream, and sibling genes.

- PAGER provides an interactive visualization tool for users to study gene and gene set networks and offers spaces, Gene Box and Gene Set Box, for users to store their genes and gene sets.

# CHAPTER 2. METHODS

## 2.1 Design

This study was separated into two parts. The first part is to construct a new type of gene set network, a regulatory gene set network. In the first part of methods, we presented a framework to construct regulatory gene set networks and a method to show that regulatory networks can reveal novel gene set relationships. We also proposed a method to construct disease specific regulatory gene set network to reveal novel insights into diseases. The second part of this study is developing Pathway and Annotated Gene-set Electronic Repository (PAGER) in order to enable identification of gene set relationships for systems biology. PAGER is an integrated platform on which users search for gene sets and construct co-membership and regulatory gene set networks.

## 2.2 Constructing a Regulatory Gene Set Network

A gene set network is a network where a node represent a set of genes and an edge represents a relationship among gene sets. Gene set networks can be used for explaining biological complexity by revealing high level relationships between biological processes. Typically, gene set networks are undirected networks. In this study, we proposed a regulatory gene set network, which is a new type of network. It is a directed network

where a directed edge represents regulatory relationship from one gene set to the other gene set. A regulatory gene set network was constructed by using publicly available gene regulation data obtained from several sources. Our hypothesis is regulatory gene set network can reveals novel insights into the complex of biological processes. A co-membership gene set network is another type of network we constructed. It can be a basement network because it is construct from annotated gene sets collected from sources, which provide experimental validation data.

Gene set data and gene regulation data were collected from different publicly available sources. We filtered only high quality gene regulation data in order to construct a regulatory gene set network. Next, regulatory and co-membership gene set networks were constructed separately for each of five gene set collections. Hypergeometric distribution was used to calculate significant value for each edge in both types of gene set networks. We validated our KEGG gene set networks with the KEGG co-enrichment network obtained from the study of Jignesh, et al. [32] Finally, we construct a regulatory network specific to Alzheimer's disease (AD).

#### 2.2.1.1 Gene Set Data Sources

A gene set is a set of genes which relate to the same biological concepts. Gene set can represent several concepts; for example, genes are in the same pathway, genes express in the same specific condition, genes are regulated by the same transcription factor or miRNA, and genes relate to the same disease. We collected five collections of gene sets including KEGG [23], Reactome [22], Go biological process, GO cellular component, and GO

molecular function [1]. Gene set data of Reactome, GO Biological Process, GO Cell Component, and GO Molecular Function were downloaded from MSigDB [28]. Different types of gene ID obtained from different sources were mapped to NCBI official gene symbols. The total number of gene set from five gene set collections is 2,825 and the total number of genes is 2,304.

#### 2.2.1.2 Gene Regulation Data Sources

Gene regulation data was used for constructing regulatory gene set networks. Human gene regulations were collected from String [13], TRANSFAC [30], TRED [20], and Spike [33]. The total number of gene regulations after combining and filtering data from the four data sources is 22,127. Different types of gene ID obtained from different sources were mapped to NCBI official gene symbols. In order to select only high quality human gene regulation data, different criteria were used to filter gene regulations for different data sources (Table 2.1). For String, gene regulations which have score greater than or equal to 800 were collected. Gene regulations which have binding site quality less than or equal to 5 from TRANSFAC were collected. For TRED, gene regulations which are not obtained from computational predicted method were collected. All genes regulations provided by Spike were collected because they are from pathways.

Table 2.1 Human gene regulation data sources

| Data Source | Description | Download date | Publication | Criterion |
|---|---|---|---|---|
| Spike (2012) | Gene regulations from pathway | 03/01/2013 | NAR, 2011 | Collect all gene regulations |
| TRANSFAC 7.4 (Public version) | Transcription Factor binding site and genes | 2009 | NAR, 2003 | Binding site quality <= 5 |
| TRED | Transcriptional Regulatory Element Database | 03/01/2013 | NAR, 2007 | Remove all computational predicted records |
| String 9.05 | Protein interaction database | 09/04/2013 | NAR, 2013 | Score >= 800 |

### 2.2.2 Constructing Gene Set Networks

In this study, two types of gene set networks we constructed are co-membership and regulatory gene set networks. In a co-membership gene set network, two gene sets were connected if there is a significant number of shared genes. In contrast, in a regulatory gene set network, a pair of gene sets were connected if there is a significant number of gene regulations between their unique genes.

#### 2.2.2.1 Co-membership Gene Set Networks

Co-membership gene set network is a typical type of gene set network which have been constructed in several studies [27,32]. Different studies used different methods to calculate the significant value; for instance, using Fisher's exact test to computing a p-value and computing an experimental p-value by generating random gene sets. In this study, hypergeometric distribution was used to calculate a significant value for each co-membership edges. The following described steps to construct each edge in a co-membership gene set network.

1. Count the number of genes inside both gene sets, GS1 and GS2.
2. Count the number of shared genes between GS1 and GS2.
3. Calculate p-value by using hypergeometric distribution

$$p - value = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (2.1)$$

where N is the total number of genes; n is the number of genes in GS1; K is the number of genes in GS2; and k is the number of shared genes.

4. Adjust calculated p-values for multiple hypotheses in order to control false discovery rate by using the Benjamini–Hochberg procedure with p-value ≤ 0.05 [3].
5. Connect a pair of gene sets if the edge was rejected by the Benjamini–Hochberg procedure.

Co-membership networks were separately constructed for each gene set collections. We used the co-membership gene set network as a basement network to compare with regulatory gene set networks in order to show that the regulatory gene set networks reveal novel relationships, which were not found before in the co-membership network.

#### 2.2.2.2 Regulatory Gene Set Networks

Our goal for this study is to construct a regulatory gene set network. It is a directed network where a node is a gene set and a directed edge represents a regulatory relationship from one gene set to the other. Because a regulatory gene set network is a directed network, it provides more insights into how different biological processes

function together. In addition, it facilitates different methods in network analysis; for example, finding sink and source gene sets. Publicly available regulation data was used to construct a regulatory gene set network. The following described steps to construct each directed edge in the network.

1. Count the number of gene regulations where genes in GS1 regulate genes outside GS1
2. Count the number of gene regulations where genes outside GS2 regulate genes in GS2
3. Remove shared genes from GS1 and GS2
4. Count the number of gene regulations where the remaining genes in GS1 regulate the remaining genes in GS2.
5. Calculate p-value by using hypergeometric distribution using Equation 2.1 where N is the total number of gene regulations; n is the number of gene regulations where genes in GS1 regulate genes outside GS2; K is the number of gene regulations where genes in GS2 are regulated from genes outside GS2; and k is the number of gene regulations from genes in GS1 to GS2.
6. Adjust calculated p-values for multiple hypotheses in order to control false discovery rate by using the Benjamini–Hochberg procedure with p-value ≤ 0.05.
7. Connect a pair of gene sets with directed edge pointing from GS1 to GS2 if the edge is rejected by the Benjamini–Hochberg procedure.

Regulatory networks were separately constructed for each gene set collection. In order to compare the regulatory network, a directed network, with the co-membership network, an undirected network, the regulatory gene set network was converted to an undirected network by discarding the direction of edges and removing loop.

#### 2.2.2.3 Hypergeometric Distribution

The hypergeometric distribution is a discrete probability distribution that describes the probability of $k$ successes in $n$ draws without replacement from a finite population of size $N$ containing exactly $K$ successes [34]. Equation 2.1 shows the probability mass function (pmf) of the hypergeometric distribution. The pmf was used for calculating p-value of all edges in gene set networks.

#### 2.2.2.4 Benjamini–Hochberg Procedure

Benjamini–Hochberg procedure or BH procedure is a method to control the False Discovery Rate for multiple comparisons [3]. False Discovery Rate (FDR) is the expected percent of false predictions in the set of predictions. For this study, the set of predictions is a set of edges in a gene set network. The q-value that we use for this study is 0.05, so we expected that 95 percent of predicted edges to be correct.

Consider testing $E_1, E_2, \dots, E_m$ based on the corresponding p-values $P_1, P_2, \dots, P_m$. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered p-values, and denote by $E_i$ the null hypothesis corresponding to $P_{(i)}$. Let $k$ be the largest $i$ for which $P_{(i)} \leq \frac{i}{m} q$ then reject all $E_i$ where

$i = 1, 2, \ldots, k$. Therefore, two gene sets were connected if their incident edge was rejected by the BH procedure.

### 2.2.3 A Disease Specific Gene Set Network

A disease specific regulatory gene set network was constructed in order to show that the network can help researchers to understand a disease. Alzheimer's disease (AD) was chosen to use as a case study. First, an AD gene list was obtained from Alzgene database [4]. Then the list was used to select AD related gene sets from all five gene set collections. Finally, an AD specific regulatory gene set network was constructed by using the AD related gene sets.

In order to find AD related gene sets, we counted the number of genes in each gene set which were found in AD gene list. We treated an AD gene list as a new gene set, an AD gene set. We used the same method as constructing a co-membership network to calculate a p-value for each gene set. Only gene sets which shared a significantly high number of genes with the AD gene set were selected. 261 gene set were selected to be AD related gene sets and were used to construct an AD specific regulatory gene set network.

### 2.2.4 Network Analysis

After co-membership and regulatory gene set networks were constructed, we calculated several types of centrality values for each gene set in both networks. igraph software package for R [8] was used in order to compute all network values.

#### 2.2.4.1 Degree Centrality

Degree centrality represents the number of edges incident upon a node. Degree centrality value of each gene set in a gene set network is equal to degree of a gene set normalized by the number of total gene sets in the network. A gene set which has high degree centrality is likely to be an important gene set because it acts like a hub in the network. In addition, a degree centrality of each node can be used for comparing two different types of networks, co-membership and regulatory gene set networks. In order to perform this comparison, Pearson's correlation coefficient of degree centrality between co-membership and regulatory gene set networks was calculated.

In addition, because regulatory gene set network is a directed network, indegree centrality and outdegree centrality were also calculated for future analysis.

#### 2.2.4.2 Betweenness Centrality

Betweenness centrality of a node is defined by the number of times a node acts as a bridge along the shortest path between two other nodes. In gene set network, a gene set which has high betweenness value is likely to be a part of several biological critical paths. igraph was used to calculate betweenness for each gene set in the networks. The betweenness of a gene set, $v$, is defined by

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $V$ is a set of gene sets in a network; $\sigma_{st} = \sigma_{ts}$ denote the number of shortest paths from $s \epsilon V$ to $t \epsilon V$ and $\sigma_{ss} = 1$ by convention; $\sigma_{st}(v)$ denote the number of shortest paths from $s$ to $t$ that some $v$ lies on [5].

In case of a regulatory gene set network, which is a directed network, directionality was considered while determining the shortest paths.

2.2.4.3 Closeness Centrality

Closeness centrality of a node is defined by the number of steps required to access every other nodes from a given node. A node with lower total distance to all other nodes is more central; and the higher closeness value the closer to other gene sets. In a gene set network, if a gene set, which has high closeness value, is disturbed, it is likely that higher number of gene sets will be affected.

igraph was used to calculate closeness for each gene set in the networks. The betweenness of a gene set, $v$, is defined by

$$C_C(v) = \frac{1}{\sum_{t \in V} d_G(v,t)}$$

where $d_G(v,t)$ denote the distance between vertices $v$ and $t$ and $d_G(v,v) = 0$ [5]. Normalization is performed by multiplying the raw closeness by $n-1$, where $n$ is the number of gene sets in a network.

In case of a regulatory gene set network, which is a directed network, directionality was considered while determining the shortest paths. In addition, in order to see the effect of a particular gene set, we considered only outgoing paths from a given gene set.

#### 2.2.4.4 Network Comparison

Pearson's correlation coefficient (r) measures the linear correlation between two variables $X$ and $Y$. We compared regulatory and co-membership gene set network by calculating Pearson's correlation coefficient of their degree centrality values. The value of Pearson's correlation coefficient is between +1 and -1, where 1 is total positive correlation suggesting that the two types of networks highly correlated in degree centrality aspect.

## 2.3 Pathway and Annotated Gene-set Electronic Repository (PAGER)

In the previous section, we explained the framework to construct the regulatory gene set network. In order to enable users to construct gene set network corresponding to their studies, we developed PAGER, an online platform for searching gene sets and constructing gene set networks. First, more gene set data were collected to increase the coverage of PAGER. Global regulatory and global co-membership gene set networks were pre-constructed. In addition, two types of gene networks provided by PAGER are gene interaction and gene regulation networks. The gene networks are networks within a gene set node and are useful for users to study about a particular gene set. In addition, interactive network visualization has been developed for displaying both gene and gene set networks. Users can also search for a particular gene set within a gene set network and select multiple gene sets in order to construct a new network. Moreover, users can

search for gene sets related to their gene list and construct co-membership and regulatory gene set networks of the related gene sets.

### 2.3.1 Data Sources

#### 2.3.1.1 Gene Set Data Sources

In order to implement high coverage pathway and gene set repository, we collected more gene set data from additional sources. The total number of gene sets is 187,076 obtained from 10 different sources (Table 2.2). 166,489 gene sets obtained from GAD have size of 1. We counted a disease-gene relationship obtained from GAD as a gene set because it was collected from different publications. The disease associated genes were not grouped together because it is obtained from different experiments. Gene set information from GAD enables users to find more disease related genes. In this study, gene set data was collected by several methods including directly download from the database and implement Java web crawler to retrieve information. Different types of gene ID were mapped into the NCBI official gene symbols.

#### 2.3.1.2 Gene Interaction and Gene Regulation Data Sources

Human gene regulation data from Table 2.1 was imported into a database. Additional human gene interaction data was downloaded from String 9.01 to support gene interaction network construction.

Table 2.2 List of gene set data sources

| Source Name | Description | Download Date | Number of Gene Sets |
|---|---|---|---|
| GAD [2] | Genetic Association Database | 8/26/2013 | 166,489 |
| GWAS Catalog [17] | A Catalog of Published Genome-Wide Association Studies provided by NHGRI | 8/27/2013 | 1,574 |
| GeneSigDB [9] | GeneSigDB: a manually curated database and resource for analysis of gene expression signatures | 8/23/2013 | 3,515 |
| MSigDB [28] | The Molecular Signatures Database is a collection of annotated gene sets for use with GSEA software | 8/26/2013 | 10,295 |
| NGS Catalog [38] | NGS Catalog: A database of next generation sequencing studies in humans | 8/26/2013 | 69 |
| OMIM [15] | Online Mendelian Inheritance in Man | 8/27/2013 | 4,409 |
| PharmGKB [37] | The Pharmacogenomics Knowledge Base | 8/26/2013 | 102 |
| Protein Lounge | Bioinformatics portal which integrates protein information, databases and research tools for researchers and students | 2009 | 393 |
| Spike [33] | SPIKE is a database of highly curated human signaling pathways | 9/5/2013 | 28 |
| WikiPathway [24] | WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community | 8/26/2013 | 202 |

### 2.3.2 Database Design

Database schema was designed after all gene set, gene regulation, gene interaction data were collected (Figure 2.1). In current version, subscription function has not available yet. Some tables which are important for gene set network construction including: GENESET is a table of gene set detail; GS_GS_OVERLAP is a table of co-membership edges; GS_GS_REG is a table of regulatory edges. GENE_GENE_INT is a table of gene interactions; and GENE_GENE_REG is a table of gene regulations. The number of total co-membership

edges discarding p-value is 92,108,741 and the number of total regulatory edges discarding p-value is 697,221,810.

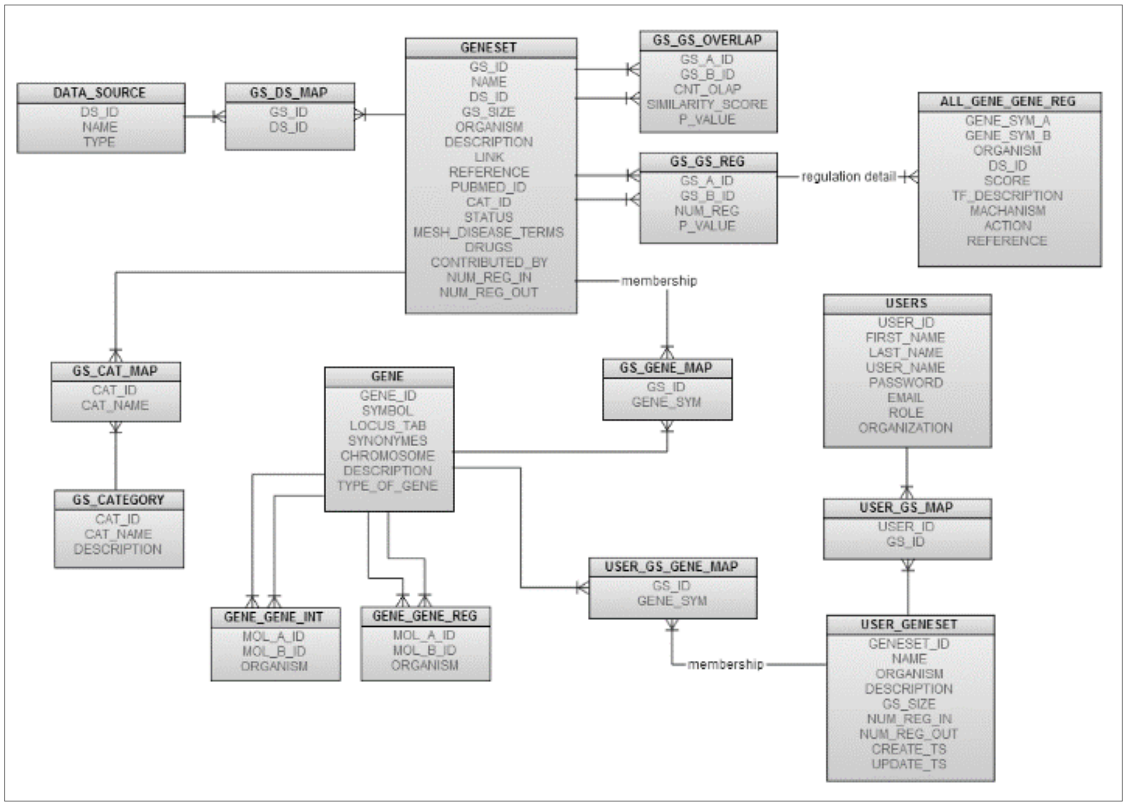



Figure 2.1 PAGER ER Diagram

### 2.3.3 Implementation

In order to implement PAGER, several programming languages and web technologies were used. The implementation followed Model-View-Controller software pattern and a programming framework was used for expandability. PAGER was implemented by using PHP language version 5 and Codeigniter version 2.1.3 [12] which is an application development framework for building web sites.

PAGER data was stored in Oracle 12g database maintained by Indiana University and was connected to PHP server by Oracle Instant Client software [21]. P-value for each edge in gene set networks was computed on-the-fly by using hypergeometric function provided by PDL [31], a PHP library for mathematics. For gene and gene set networks visualization in PAGER, cytoscape.js, an open-source graph library [6], and jQuery was used for implementing interactive networks.

# CHAPTER 3. RESULTS

## 3.1 Constructing a Regulatory Gene Set Network

### 3.1.1 Gene Set Networks

Gene set networks reveal insights into relationships among different biological processes. In this study, for each gene set collection, we constructed a co-membership gene set network, a well study type of gene set network, and a regulatory gene set network, a new type of gene set network. Co-membership gene set networks connect a pair of gene sets if there is a significant number of shared genes. Therefore, constructing co-membership gene set network requires only curated gene set definition from gene set data sources. A co-membership gene set network can be a baseline for comparison with other types of gene set networks to evaluate gaining of novel insights. In this study, the goal is to construct a regulatory gene set network, a new type of gene set network, by using publicly available gene regulations (Table 3.1). Regulatory gene set networks connect a pair of gene sets when there is a significant number of gene regulations between the genes of the two gene sets. In order to construct gene set regulatory network, we collected gene regulation data from 4 different data sources including String 9.05 [13], TRANSFAC [30], TRED [20] and Spike [33].

Two types of gene set network have different interpretations. When a pair of gene sets in co-membership gene set network are connected, the interpretation depends on types of gene sets. For pathway gene sets, co-membership networks represent pathway crosstalk; for GO gene sets, an edge in co-membership networks represents protein moonlighting or gene sharing. In contrast, when a pair of gene sets in regulatory gene set networks are connected, they are connected by a directed edge. The directed edges in a regulatory gene set network present a possibility of one gene set regulates other gene set.

Table 3.1 Summary of co-membership gene set networks and regulatory gene set networks for five gene set collections

| Collection | Number of gene sets | co-membership edges | Regulatory edges | Regulatory relationships | Shared edges |
|---|---|---|---|---|---|
| KEGG | 186 | 2,230 | 4,452 | 3,274 | 1,461 |
| Reactome | 674 | 15,859 | 25,569 | 20,917 | 7,437 |
| GO BP | 825 | 33,055 | 32,607 | 27,513 | 10,354 |
| GO CC | 223 | 4,186 | 1,446 | 1,122 | 793 |
| GO MF | 396 | 3,178 | 2,620 | 2,404 | 503 |

A pair of gene sets can have a loop if both incoming edge and outgoing edge are significant, so there can be two edges for a pair of gene sets. Therefore, the number of regulatory edges is always grater or equal to number of regulatory relationships because each pair of gene sets either has or do not have regulatory relationship.

For regulatory gene set networks, GO Biological Process (32,607 edges and 825 nodes), Reactome Pathway (25,569 edges and 674 nodes) have the highest proportion of edges to nodes among the five gene set collections (Table 3.1). In addition, by considering the

proportion of edges to nodes among three GO collections, GO Biological Process has the highest proportion, while GO Cellular Component (1122 edges and 223 nodes) and GO Molecular Function (2404 edges and 396 nodes) have relatively low proportion. These results suggested that pairs of biological processes are more likely to have regulatory relationships.

The low percentages of shared edges indicated that regulatory gene set networks provide complementary knowledge to co-membership gene set networks. It is worth to note that regulatory gene set networks were constructed from high quality gene regulation data, which were collected from high coverage data sources. Therefore, regulatory gene set networks unlikely depend on the number and the quality of experimental data.

Considering both co-membership and regulatory networks of KEGG pathway gene sets, the most significant edge of the KEGG regulatory gene set network is a regulatory relationship from “Cell cycle” to “Cytokine-cytokine receptor interaction” with significant value 6.46E-75 (Table 3.2). While a co-membership edge between “Cell cycle” to “Cytokine-cytokine receptor interaction” has relatively lower significant value, 0.029. In addition, only 4 of the top 10 most significant regulatory edges were found in the KEGG co-membership network. These findings suggested that the regulatory gene set network reveals additional knowledge to the co-membership gene set network.

For the KEGG regulatory gene set network, 7 of the 10 most significant regulatory edges are from “Cell cycle” gene set to other 7 KEGG pathway gene sets, including “Cytokine-

cytokine receptor interaction", "Pathways in cancer", "Toll-like receptor signaling pathway", "Focal adhesion" and "Leishmania infection" (Table 3.2). These results suggested that changing in "Cell cycle" pathway likely affects other pathways. These results were also corresponding to the fact that cell cycle is the complex series of phenomena by which cellular material is duplicated and divided. If cell cycle pathway does not appropriately function, several pathways can be affected such as Pathways in Cancer.

Table 3.2 Top 10 most significant regulatory edges in the KEGG regulatory gene set network where gene set 1 regulates gene set 2

| Gene set 1 name | Gene set 2 name | P-value |
|---|---|---|
| Cell cycle | Cytokine-cytokine receptor interaction | 6.46E-75 |
| Cell cycle | Pathways in cancer | 5.31E-55 |
| Cell cycle | Toll-like receptor signaling pathway | 1E-42 |
| Cell cycle | Focal adhesion | 2.22E-36 |
| Cell cycle | Leishmania infection | 2.63E-33 |
| Hedgehog signaling pathway | Basal cell carcinoma | 3.08E-33 |
| p53 signaling pathway | Cytokine-cytokine receptor interaction | 3.38E-33 |
| RIG-I-like receptor signaling pathway | Toll-like receptor signaling pathway | 3.39E-33 |
| Cell cycle | Hematopoietic cell lineage | 8.26E-33 |
| Cell cycle | Jak-STAT signaling pathway | 1.08E-31 |

For co-membership network, the KEGG pathway gene sets of "Alzheimer's disease", "Parkinson's disease" and "Huntington's disease" have significant co-membership edges link them together (Table 3.3). The three co-membership edges connecting the neurodegenerative diseases were in the top 10 most significant co-membership edges suggesting that the three neurodegenerative diseases are highly related. In addition, 5 edges of the top 20 co-membership edges were connecting cancer related pathway gene

sets. The 5 edges are “Pathways in cancer” gene set connects to 5 cancer gene sets including “Small cell lung cancer”, “Pancreatic cancer”, “Melanoma”, “Colorectal cancer and “Prostate cancer”.

Table 3.3 Top 10 most significant co-membership edges in the KEGG co-membership gene set network

| Gene set 1 name | Gene set 2 name | P-value |
|---|---|---|
| Dilated cardiomyopathy | Hypertrophic cardiomyopathy (HCM) | 2.9E-134 |
| Oxidative phosphorylation | Parkinson's disease | 5E-132 |
| Huntington's disease | Parkinson's disease | 2.2E-124 |
| Alzheimer's disease | Parkinson's disease | 8.9E-113 |
| Drug metabolism - cytochrome P450 | Metabolism of xenobiotics by cytochrome P450 | 2.5E-110 |
| Alzheimer's disease | Huntington's disease | 1.5E-106 |
| Alzheimer's disease | Oxidative phosphorylation | 1.5E-101 |
| Huntington's disease | Oxidative phosphorylation | 1.02E-96 |
| Pathways in cancer | Small cell lung cancer | 2.94E-91 |
| Arrhythmogenic right ventricular cardiomyopathy (ARVC) | Hypertrophic cardiomyopathy (HCM) | 1E-89 |

### 3.1.2 Comparing the KEGG Regulatory and the KEGG Co-membership Network

In the co-membership gene set network of KEGG, an edge between pathways can be considered as pathway crosstalk, a well-studied phenomenon [7,27,32]. In contrast to the co-membership network, the regulatory network revealed regulatory relationships between pathways, for example, dysfunction of one pathway might affect function of another pathways. In order to show that regulatory gene set network provide complementary information to the co-membership gene set network, we compare co-membership gene set network and regulatory gene set network of the KEGG pathways.

The KEGG pathway data downloaded from MSigDB version 3.1 contains 186 pathways which are considered as gene sets and 872 genes. For the KEGG co-membership gene set network, 2,230 co-membership edges were constructed where the total number of possible pairs of gene sets is 17,205. For the KEGG regulatory gene set network, 4,452 regulatory edges were constructed where the total number of possible pairs of gene sets is 34,410 because the regulatory network is a directed network. However, if we count only the regulatory relationship between a pair of gene sets by discarding the direction, there were 3,274 relationships. Directionality information in the KEGG regulatory gene set network was discarded in order to compare with the co-membership gene set network which is an undirected network. 1,461 edges were found in both the KEGG co-membership gene set network and the KEGG regulatory gene set network. The proportion of shared edges to co-membership edges is 65.52% (1461/2230) and the proportion of shared edges to regulatory relationships is 44.62% (1461/3274).

Besides comparing the two networks by counting shared edges, degree centrality (DC) of each gene set node in both networks were calculated (Table 3.4). In the KEGG co-membership network, “Pathways in cancer” gene set has the highest value of degree centrality, 0.39, while in the KEGG regulatory gene set network, “Cell cycle” gene set has the highest value of degree centrality, 1.96, and the highest outdegree centrality, 0.65. The gene set which has the highest indegree centrality, 0.41, is “Pathways in cancer”. Note that degree centrality value of regulatory gene set network can be greater than 1 because regulatory network is a directed network.

Table 3.4 Top 10 highest degree centrality pathway of KEGG co-membership gene set network (left) and KEGG regulatory gene set network (right)

| **KEGG co-membership network** | | **KEGG regulatory network** | |
|---|---|---|---|
| ***Name*** | ***DC*** | ***Name*** | ***DC*** |
| Pathways in cancer | 0.39 | Cell cycle | 1.96 |
| MAPK signaling pathway | 0.36 | T cell receptor signaling pathway | 1.55 |
| T cell receptor signaling pathway | 0.36 | Chemokine signaling pathway | 1.49 |
| Chemokine signaling pathway | 0.36 | ErbB signaling pathway | 1.48 |
| Natural killer cell mediated cytotoxicity | 0.36 | p53 signaling pathway | 1.48 |
| Fc epsilon RI signaling pathway | 0.34 | Pathways in cancer | 1.47 |
| Progesterone-mediated oocyte maturation | 0.34 | Bladder cancer | 1.46 |
| GnRH signaling pathway | 0.34 | Neurotrophin signaling pathway | 1.45 |
| Colorectal cancer | 0.33 | Chronic myeloid leukemia | 1.43 |
| Prostate cancer | 0.33 | Focal adhesion | 1.43 |

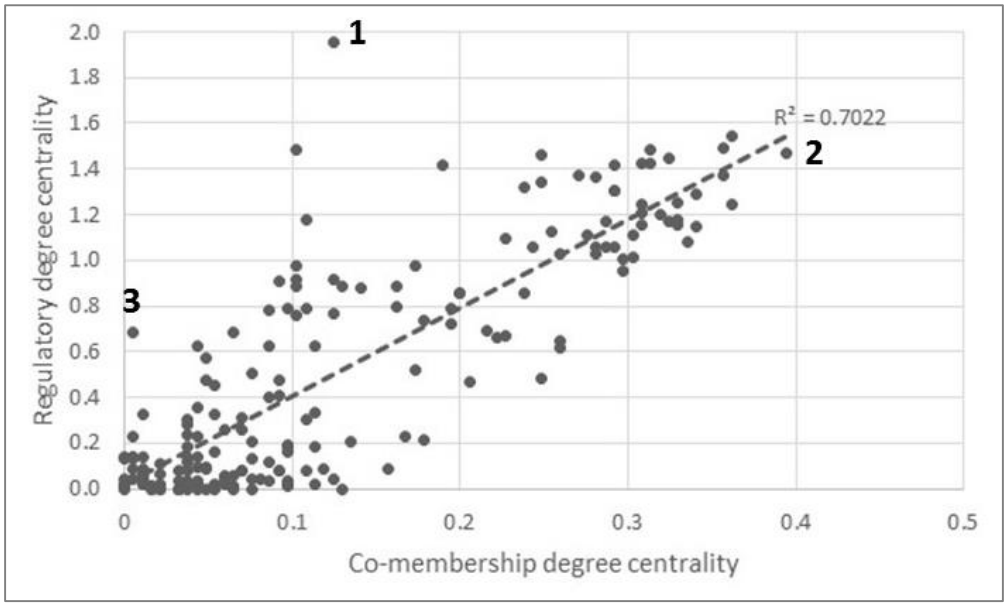



Figure 3.1 Correlation between degree centrality of the KEGG co-membership network and degree centrality of the KEGG regulatory network.

After a degree centrality value for each gene set was calculated, we calculated a correlation between degree centrality of the KEGG co-membership network and degree centrality of the KEGG regulatory network. The correlation coefficient is 0.84 and R-squared value is 0.70 (Figure 3.1). This result suggested that the gene set, which is important in a co-membership gene set network, is likely to be important in a regulatory gene set network. In addition, three interesting outliers were found in Figure 3.1. Pathway number 1, which has regulatory DC = 1.96 and co-membership DC = 0.12, is “Cell cycle”, suggesting that Cell cycle pathway does not tend to share genes with other pathways, but tends to regulate other pathways. Pathway number 2, which has regulatory DC = 1.47 and co-membership DC = 0.39 is “Pathways in cancer”, suggesting that Pathways in cancer shared high number of genes with several pathways and its genes also regulate the unique genes of other pathways. Pathway number 3, which has regulatory DC = 0.68 and co-membership DC = 0.005, is “Maturity onset diabetes of the young”. “Maturity onset diabetes of the young” pathway only shares 6 genes with “Type II diabetes mellitus pathway”. While the correlation between degree centrality of the KEGG co-membership network (M) and degree centrality of the KEGG regulatory network (R) is high, 0.84, the topology of the networks are different (Figure 3.2). These results suggested that two types of networks can be used to explain different biological phenomenon. The networks were layout using Edge-Weighted Spring Embedded Layout according to p-value in Cytoscape software.

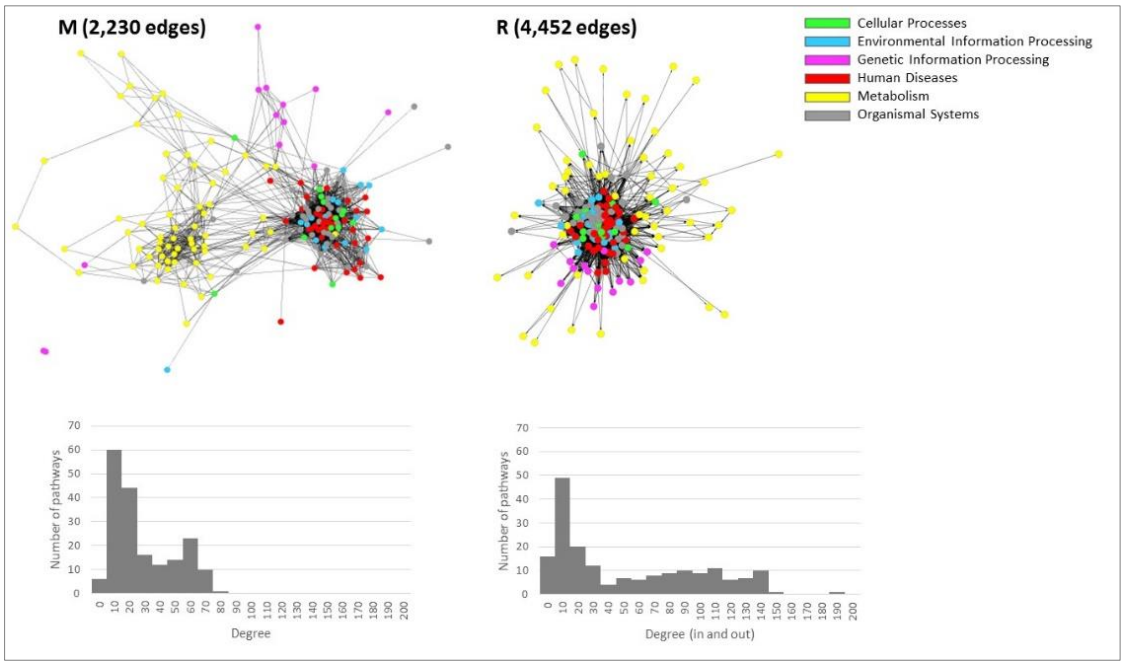



Figure 3.2 (M) KEGG co-membership network and (R) KEGG regulatory network.

We further constructed a KEGG exclusive regulatory gene set network (R-M) which contains only exclusive edges in the KEGG regulatory gene set network (Figure 3.3). The correlation of degree centrality between the KEGG exclusive regulatory gene set network (R-M) and the KEGG regulatory gene set network (R) is 0.81. The correlation of degree centrality between the KEGG exclusive regulatory gene set network (R-M) and the KEGG co-membership gene set network (M) is 0.44 which is relatively low comparing to the correlation between degree centrality of the KEGG co-membership network (M) and degree centrality of the KEGG regulatory network (R). These results suggested that constructing an exclusive regulatory gene set network reveals important gene sets which are not likely to be revealed by constructing a co-membership gene set network.

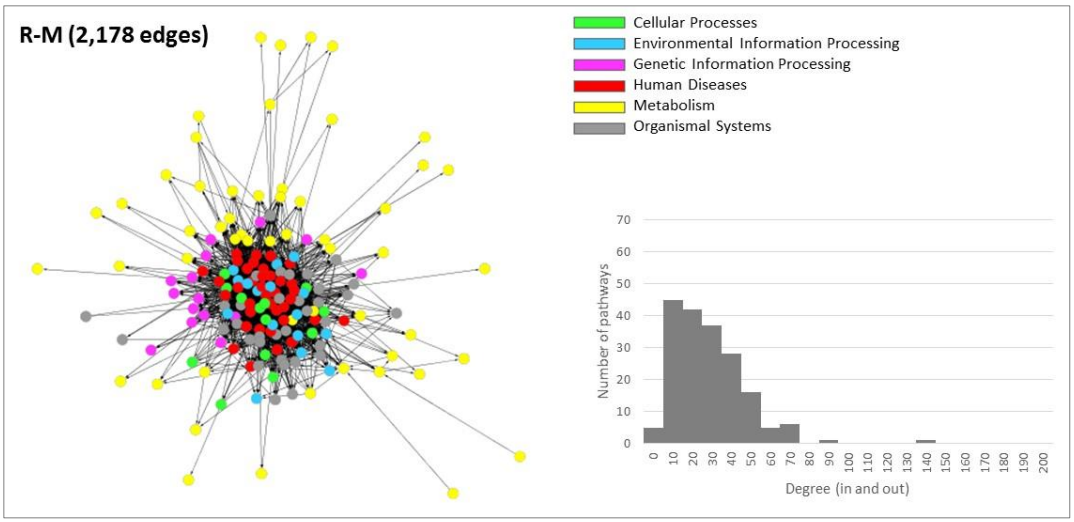



Figure 3.3 R-M is KEGG exclusive regulatory network. Node colors represents different classes of pathways.

A degree centrality of each gene set in the KEGG exclusive regulatory network was calculated by considering both indegree and outdegree (Table 3.5). The “Cell cycle” gene set still has the highest value of degree centrality, 0.78. In addition, degree centrality of the top 10 gene sets in the KEGG co-membership gene set network (M DC) are relatively low; however, degree centrality of the top 10 gene sets in the KEGG regulatory gene set network (R DC) are relatively high. These results also suggested that constructing exclusive regulatory gene set network reveals important gene sets which are not likely to be revealed by constructing co-membership gene set network.

An indegree centrality and an outdegree centrality were also calculated for each gene set in the KEGG exclusive regulatory gene set network (Table 3.6). The directionality information from a regulatory gene set network revealed “sink” and “source” gene sets in addition to “hub” gene sets.

Table 3.5 Top 10 highest degree centrality pathway of the KEGG exclusive regulatory gene set network

| Name | R-M DC | M DC | R DC |
|---|---|---|---|
| Cell cycle | 0.78 | 0.12 | 1.96 |
| p53 signaling pathway | 0.59 | 0.10 | 1.48 |
| TGF-beta signaling pathway | 0.44 | 0.11 | 1.18 |
| Bladder cancer | 0.43 | 0.25 | 1.46 |
| Cytokine-cytokine receptor interaction | 0.42 | 0.19 | 1.42 |
| Chronic myeloid leukemia | 0.37 | 0.31 | 1.43 |
| Small cell lung cancer | 0.37 | 0.27 | 1.37 |
| Huntington's disease | 0.36 | 0.13 | 0.89 |
| Jak-STAT signaling pathway | 0.35 | 0.24 | 1.32 |
| Non-small cell lung cancer | 0.34 | 0.29 | 1.42 |

Table 3.6 Top 10 highest outdegree centrality pathways (right) and top 10 highest indegree centrality pathways (left) of the KEGG exclusive regulatory gene set network

| Name | DC (out) | Name | DC (in) |
|---|---|---|---|
| Cell cycle | 0.54 | Cell cycle | 0.24 |
| p53 signaling pathway | 0.44 | Hematopoietic cell lineage | 0.21 |
| TGF-beta signaling pathway | 0.37 | Cytokine-cytokine receptor interaction | 0.20 |
| Bladder cancer | 0.32 | Systemic lupus erythematosus | 0.17 |
| Small cell lung cancer | 0.25 | Leishmania infection | 0.17 |
| Chronic myeloid leukemia | 0.24 | Basal cell carcinoma | 0.17 |
| Jak-STAT signaling pathway | 0.24 | Cell adhesion molecules (CAMs) | 0.17 |
| Huntington's disease | 0.24 | Graft-versus-host disease | 0.17 |
| Non-small cell lung cancer | 0.23 | Viral myocarditis | 0.16 |
| Wnt signaling pathway | 0.22 | p53 signaling pathway | 0.15 |

### 3.1.3 Comparison of the KEGG Regulatory and the KEGG Co-enrichment Networks

Two types of networks, co-membership gene set network (M) and regulatory gene set network (R), were constructed for KEGG pathways. Our hypothesis is regulatory gene set networks can reveal novel knowledge for understanding systems biology. The KEGG co-membership gene set network was used as a basement and to compare with the KEGG

regulatory gene set network. We presented that several relationships between gene sets were found exclusively in regulatory gene set network. In order to validate both KEGG co-membership network and KEGG regulatory network, we compared these two gene set networks with the KEGG co-enrichment network (E) obtained from the study of Jignesh, et al. [32] In order to construct co-enrichment network, Jignesh, et al. integrated experimental gene lists. Two gene sets were connected if the unique genes of the two gene sets are consistently enriched together across many experimentally derived gene lists. The definition of the co-enrichment gene set network suggested that edges found in the KEGG regulatory network should also been found in the KEGG co-enrichment network.

Because our KEGG regulatory gene set network is a directed network, while the KEGG co-enrichment gene set network is an undirected network, the regulatory network was converted to an undirected network. The total number of edges in the co-enrichment network is 1,556 and the total number of edges in the converted regulatory network is 3,274. The KEGG regulatory network and the KEGG co-enrichment network were compared. We found that the total number of edges found in both co-enrichment and regulatory networks is 1,050 which is equal to 67.48% of the total number of edges in the co-enrichment network. We also compared the KEGG co-membership network with the KEGG co-enrichment network and found that the total number of edges are found in both co-enrichment and co-membership networks is 914 which is equal to 58.74% of the total number of edges in the co-enrichment network.

In order to calculate significant value of the number of shared edges, the KEGG co-enrichment network was compared with random networks. We randomly generated 1,000 networks using 187 gene sets from KEGG. In order to calculate significant value of the number of shared edges between KEGG co-enrichment network (E) and KEGG co-membership network (M), each of the 1,000 random network has 2,230 edges, which is equal to the number of edges found in the KEGG co-membership network (M). Then Fisher's exact test was used for calculating p-value for the number of shared edges; and the p-value is < 2.2e-16 (Figure 3.4A). Next, for calculating significant value of the number of shared edges between KEGG co-enrichment network (E) and the converted KEGG regulatory network (R), each of the 1,000 random network contains 3,274 edges which is equal to the number of edges found in the converted regulatory network (R). Then Fisher's exact test was used for calculating p-value for the number of shared edges and the p-value < 2.2e-16 (Figure 3.4B)

For the number of shared edges between the 2,230 random network and KEGG co-enrichment network (Figure 3.4A); the average and median is 197; the minimum is 162; and the maximum is 236. For the number of shared edges between the 3,274 random network and KEGG co-enrichment network (Figure 3.4B); the average and median is 289; the minimum is 240; and the maximum is 333.

The number of shared edges between the KEGG co-enrichment network and the KEGG regulatory network is significantly high. This result is corresponding to the fact that a pair of gene sets which have a significant regulatory relationship should be connected with a

co-enrichment edge. The number of shared edges between the KEGG co-enrichment network and the KEGG co-membership network is also significantly high. This result is corresponding to the fact that a pair of gene sets which has high number of shared genes should be connected with a co-enrichment edge.

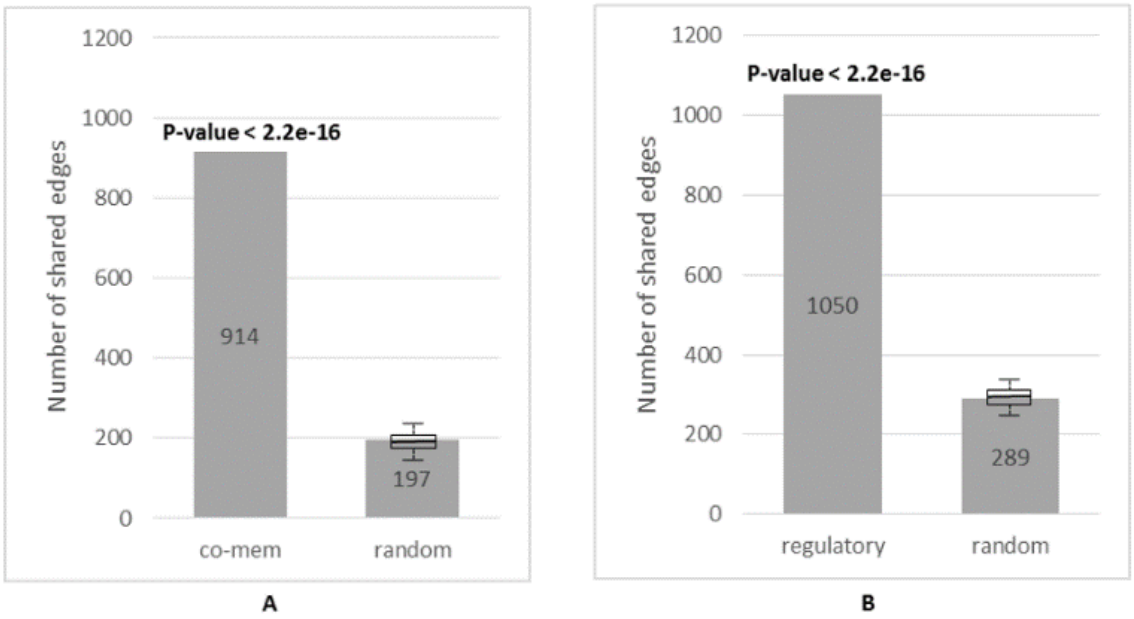



Figure 3.4 (A) Number of edges in a KEGG co-enrichment network which shared with a KEGG co-membership and random networks. (B) Number of edges in a KEGG co-enrichment network that shared with a KEGG regulatory and random networks.

### 3.1.4 A Disease Specific Regulatory Gene Set Network

We already presented that a regulatory gene set network provides additional insights into a co-membership gene set network. We also validated the KEGG regulatory gene set network and the KEGG co-membership gene set network using the KEGG co-enrichment gene set [32]. Furthermore, a regulatory gene set network specific to Alzheimer's disease (AD) was constructed. First, 347 AD associated genes were obtained from Alzgene database [29]. These genes were used to select gene sets, from all five collections, which

are associated with AD. We counted the number of shared genes between the AD gene list and each gene set in the five collections. The same method as constructing co-membership networks was used to calculate significant value of the number of shared genes. 261 out of 2,314 gene sets in the five collections have significant number of shared genes. For the 261 AD gene sets, 42 gene sets are from KEGG, 59 gene sets are from Reactome, 37 gene sets are from GO Molecular Function, 105 gene sets are from Go Biological Process, and 18 gene sets are from GO Cellular component. Among the 261 AD gene sets, 2 gene sets, “Alzheimer's disease” from KEGG and “amyloid precursor protein metabolic process” from GO Biological Process, were annotated that they are associated with AD. Next, a regulatory gene set network specific to AD was constructed by counting number of gene regulation between every pair of the 261 AD gene sets (Figure 3.5)

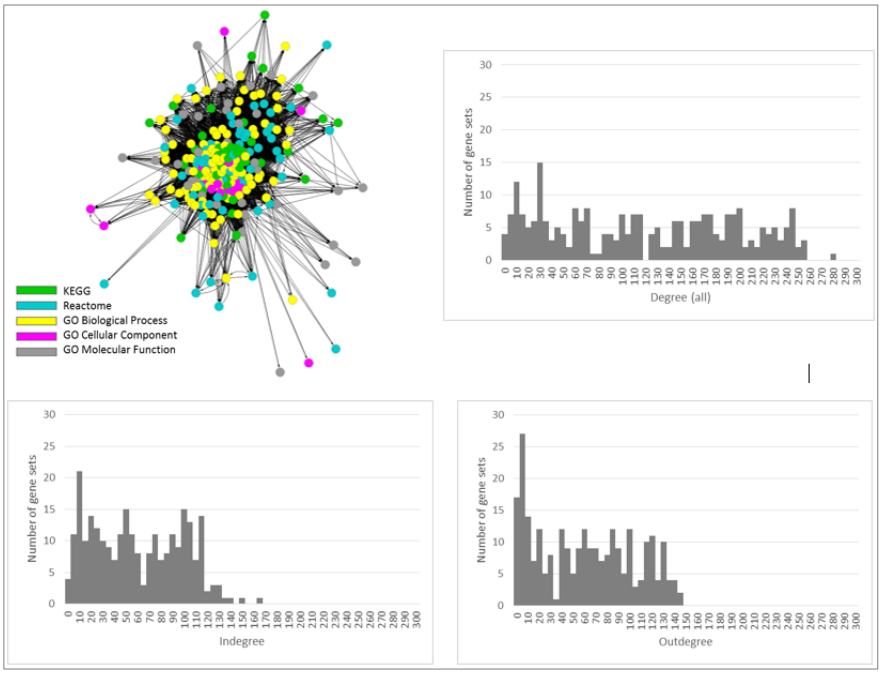



Figure 3.5 AD regulatory gene set network

The AD regulatory gene set network contains 261 gene sets and 15,178 regulatory edges. A node color represents a collection of gene sets. Green represents KEGG; blue represents Reactome; yellow represents GO Biological Process; pink represents GO Cellular Component; and gray represents GO Molecular function. Three charts presented degree distribution (top), indegree distribution (bottom left), and outdegree distribution (bottom right). Top 10 highest degree centrality gene sets were investigated (Table 3.7). “signal transduction” from GO biological process has the highest value of degree centrality (DC) suggesting that signal transduction process is very important in AD. By searching on PubMed [16], there were more than two thousand publications discussing about a relationship between signal transduction abnormality and Alzheimer’s disease. In addition, Liu, et al.[29] studied AD related KEGG pathways and reported top 5 pathways for each of 6 brain regions and “Cytokine-cytokine receptor interaction”, which has the second highest degree centrality, are among the top 5 of 4 brain regions.

Table 3.7 Top 10 highest degree centrality gene sets among 261 AD gene sets

| Name | Collection | DC | DC in | DC out |
|---|---|---|---|---|
| signal transduction | GO biological process | 1.07 | 0.62 | 0.45 |
| Cytokine-cytokine receptor interaction | KEGG gene sets | 0.97 | 0.53 | 0.44 |
| intracellular signal transduction | GO biological process | 0.97 | 0.47 | 0.49 |
| protein metabolic process | GO biological process | 0.96 | 0.43 | 0.54 |
| T cell receptor signaling pathway | KEGG gene sets | 0.95 | 0.44 | 0.51 |
| cellular protein metabolic process | GO biological process | 0.94 | 0.40 | 0.54 |
| receptor binding | GO molecular function | 0.94 | 0.44 | 0.49 |
| cellular macromolecule metabolic process | GO biological process | 0.94 | 0.40 | 0.54 |
| apoptotic process | GO biological process | 0.93 | 0.44 | 0.49 |
| programmed cell death | GO biological process | 0.93 | 0.44 | 0.49 |

A sub-network of AD regulatory gene set network which contains only the top 10 highest DC gene sets was also constructed (Figure 3.6). The sub-network is almost fully connected, indicating that gene sets, which have high DC, tend to be connected together to form a highly connected network.

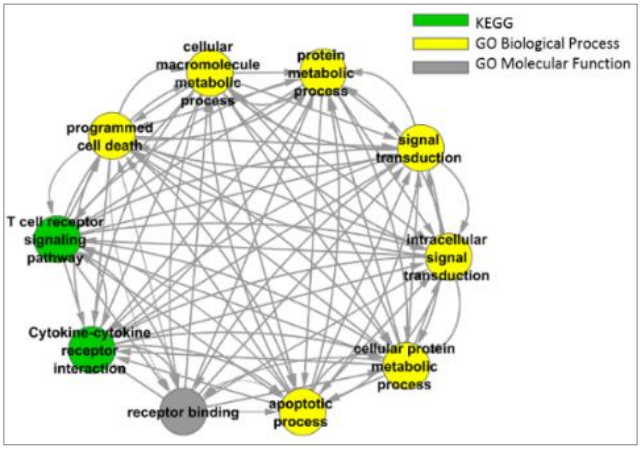



Figure 3.6 A sub-network of AD regulatory gene set network which contains only the top 10 highest DC gene sets from Table 3.7.

Furthermore, closeness and betweenness of each gene sets in the AD regulatory gene set network were computed (Table 3.8). Three gene sets from GO biological process collection have the highest closeness centrality, 0.184, when considered only outgoing edges. The three gene sets are “cellular macromolecule metabolic process”, “cellular protein metabolic process”, and “protein metabolic process”. These results suggested that inappropriately functions of these three gene sets likely affect high number of gene sets or biological processes in AD context.

Table 3.8 Top 10 highest closeness centrality (out) gene sets

| Name | Collection | Closeness (out) |
|---|---|---|
| cellular macromolecule metabolic process | GO biological process | 0.184 |
| cellular protein metabolic process | GO biological process | 0.184 |
| protein metabolic process | GO biological process | 0.184 |
| regulation of apoptotic process | GO biological process | 0.183 |
| regulation of programmed cell death | GO biological process | 0.183 |
| Genes involved in Signaling by TGF-beta Receptor Complex | Reactome | 0.183 |
| regulation of cellular metabolic process | GO biological process | 0.183 |
| regulation of metabolic process | GO biological process | 0.182 |
| apoptotic process | GO biological process | 0.182 |
| programmed cell death | GO biological process | 0.182 |

For betweenness centrality, “Pathways in cancer” from KEGG, “system development” from GO biological process, and “Leishmania infection” from KEGG have the highest betweenness value 1,225.04, 1,168.99, and 1,146.79, respectively. These results suggested that the three gene sets are likely on critical paths of biological functioning in AD context.

Table 3.9 Top 10 highest betweenness centrality gene sets

| Name | Collection | Betweenness |
|---|---|---|
| Pathways in cancer | KEGG | 1225.04 |
| system development | GO biological process | 1168.99 |
| Leishmania infection | KEGG | 1146.79 |
| signal transduction | GO biological process | 983.64 |
| Genes involved in Metabolism of lipids and lipoproteins | Reactome | 859.10 |
| cell proliferation | GO biological process | 823.84 |
| Small cell lung cancer | KEGG | 780.05 |
| cytoplasm | GO cellular component | 767.20 |
| Adipocytokine signaling pathway | KEGG | 746.16 |
| lipid metabolic process | GO biological process | 731.03 |

## 3.2 Pathway and Annotated Gene-set Electronic Repository (PAGER)

A regulatory gene set network reveals insights into systems biology. It provided additional and higher resolution knowledge to the existing types of gene set networks as we presented in the previous sections. It is useful to build a tool for researchers to generate their own gene set networks, which relate to their studies. Therefore, we developed Pathway and Annotated Gene-set Electronic Repository (PAGER), which provided a platform for users to search for gene sets by terms or by a list of genes, construct co-membership and regulatory gene set networks, and construct gene interaction and gene regulation networks of each gene set. PAGER is available at http://discern.uits.iu.edu:8340/PAGER/.

### 3.2.1 Data Integration

In order to make PAGER to have high coverage gene set data, gene sets were collected from GAD [2], GWAS Catalog [17], NGS Catalog [38], GeneSigDB [9], MSigDB [28], OMIM [15], PharmGKB [37], Protein Lounge, Spike [33], and WikiPathway [24] (Table 3.10). The total number of gene sets is 187,076.

According to Table 3.10, gene sets from GAD are all size of 1 because GAD provided disease-gene relationship information. Genes, which belong to the same disease, were not combined into a gene set because they were from different experiments or studies. However, the data integrated from GAD is useful for searching disease related genes in order to find disease related gene sets from other sources within PAGER.

Table 3.10 Gene set data sources

| Source Name | Description | Download Date | Number of Gene Sets |
|---|---|---|---|
| GAD | Genetic Association Database | 8/26/2013 | 166,489 |
| GWAS Catalog | A Catalog of Published Genome-Wide Association Studies provided by NHGRI | 8/27/2013 | 1,574 |
| GeneSigDB | GeneSigDB: a manually curated database and resource for analysis of gene expression signatures | 8/23/2013 | 3,515 |
| MSigDB | The Molecular Signatures Database is a collection of annotated gene sets for use with GSEA software | 8/26/2013 | 10,295 |
| NGS Catalog | NGS Catalog: A database of next generation sequencing studies in humans | 8/26/2013 | 69 |
| OMIM | Online Mendelian Inheritance in Man | 8/27/2013 | 4,409 |
| PharmGKB | The Pharmacogenomics Knowledge Base | 8/26/2013 | 102 |
| Protein Lounge | Bioinformatics portal which integrates protein information, databases and research tools for researchers and students | 7/1/1905 | 393 |
| Spike | SPIKE is a database of highly curated human signaling pathways | 9/5/2013 | 28 |
| WikiPathway | WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community | 8/26/2013 | 202 |

Gene regulation data were collected from four sources, which are the same as the four sources in the previous section, was integrated into PAGER. However, all gene regulations were integrated into PAGER without using any filter. The total number of unique gene regulations is 651,568. In addition, gene interactions were downloaded from String [13] in order to construct a gene interaction network inside each gene sets. The total number of

unique gene interactions is 650,160. Note that gene regulations were also counted as gene interactions regardless the direction.

### 3.2.2 Gene Set and Gene Set Networks

In order to construct a global regulatory gene set network which consists of all gene sets in PAGER, the 22,127 filtered gene regulations were used. For each gene set, the number of gene regulations from genes inside the gene set to genes outside the gene set and the number of gene regulations from genes outside the gene set to genes inside the gene set were counted. The number of gene regulations between two gene sets were counted for every pair of gene sets in PAGER. The total number of regulatory edges discarding the significant value is 697,221,810. A global co-membership gene set network was constructed by counting the number of shared genes among every pair of gene sets. The total number of co-membership edges discarding the significant value is 92,108,741. Due to the large number of edges and the limitation of time, the significant value of each edge was computed on-the-fly using hypergeometric distribution when users query for a particular gene set network. In addition, because gene interactions and gene regulations were integrated into PAGER, users can construct gene networks within each gene set in the gene set networks. In other words, a gene network is a network within a node of a gene set network.

### 3.2.3 PAGER Features

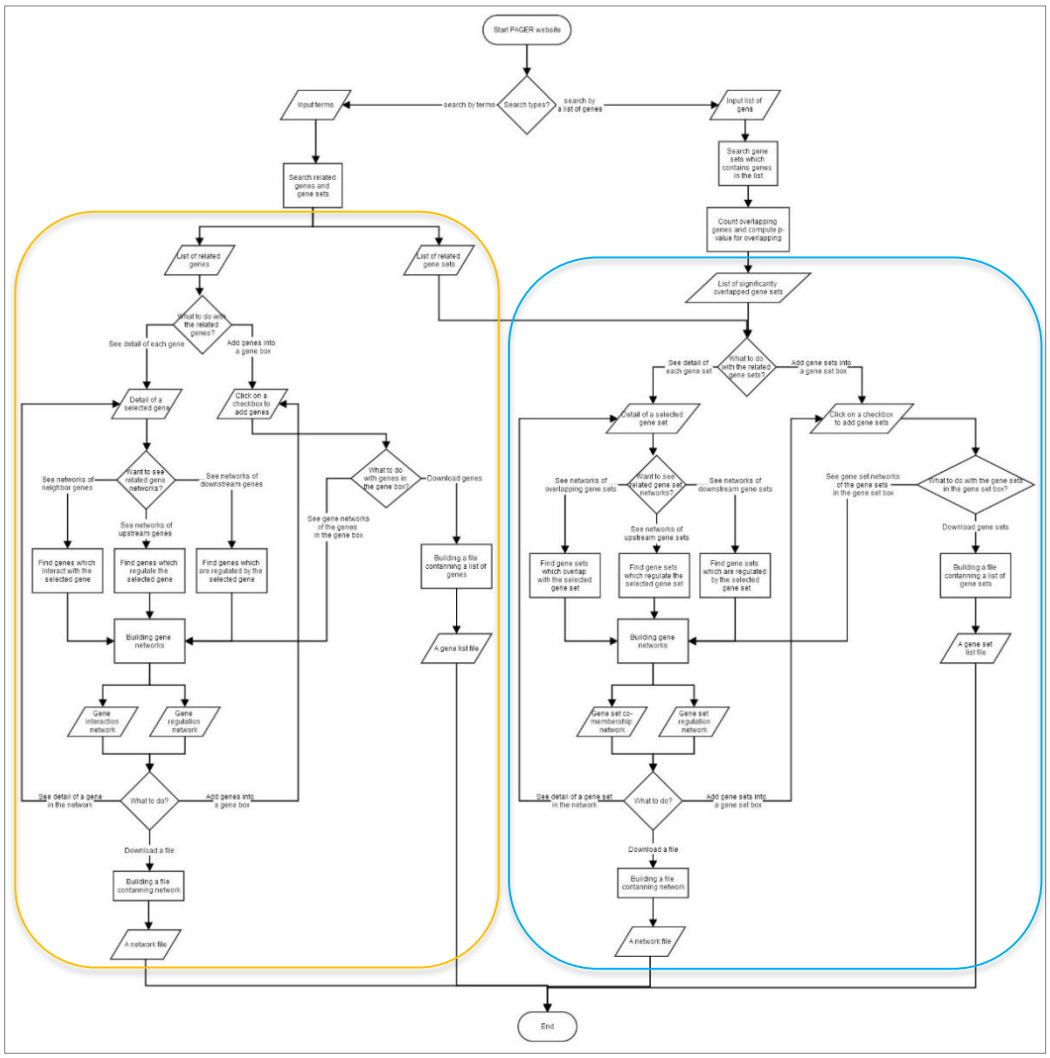



Figure 3.7 PAGER work flow

An overall work flow of PAGER was presented in Figure 3.7. The top part showed two types of searching, search by terms and search by a list of genes (Figure 3.8). When users search by terms, PAGER returns both genes and gene sets which relate to the terms. When users search by a list of genes, PAGER searches for gene set which contains the genes in the list and calculates p-value of the number of shared genes between the gene

list and gene sets available in PAGER. Functions related to gene networks were in an orange rectangle and functions related to gene set networks were in a blue rectangle.

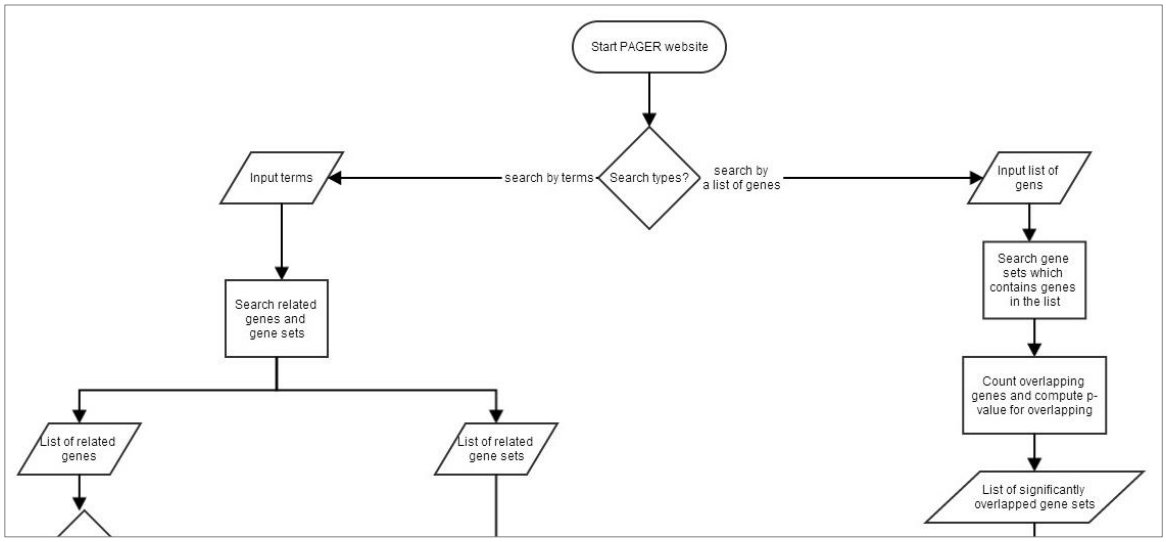



Figure 3.8 PAGER work flow

The main feature of PAGER is enabling users to construct regulatory gene set networks (Figure 3.9). Users can search for gene sets either by terms or a list of genes. Users can add gene sets into Gene Set Box, which is a space for saving gene sets. Both regulatory and co-membership gene set networks can be constructed for gene sets inside Gene Set Box. In addition, for a particular gene set, users can construct gene set networks of upstream, downstream, or co-membership gene sets. Upstream gene sets are gene sets which regulate the current gene set; downstream gene sets are gene sets which are regulated by the current gene set; and co-membership gene sets are gene sets which share genes with the current gene set. Therefore, PAGER does not limit users to construct gene set networks of only gene sets in the Gene Set Box. PAGER allows users to expand

the gene set networks by constructing gene set networks of upstream, downstream, and co-membership gene sets.

In addition to gene set networks, after users entered searching terms, PAGER returns both related genes and gene sets. The chart in Figure 3.10, which is a larger version of blue rectangle area in Figure 3.7, explained features of PAGER which related to gene networks. For instance, after searching, users can add genes into Gene Box, which is an area for user to save their genes, and build gene regulation and gene interaction networks of the genes in their Gene Box. In addition, users can construct gene set networks of either downstream, upstream, and sibling genes of a particular gene.

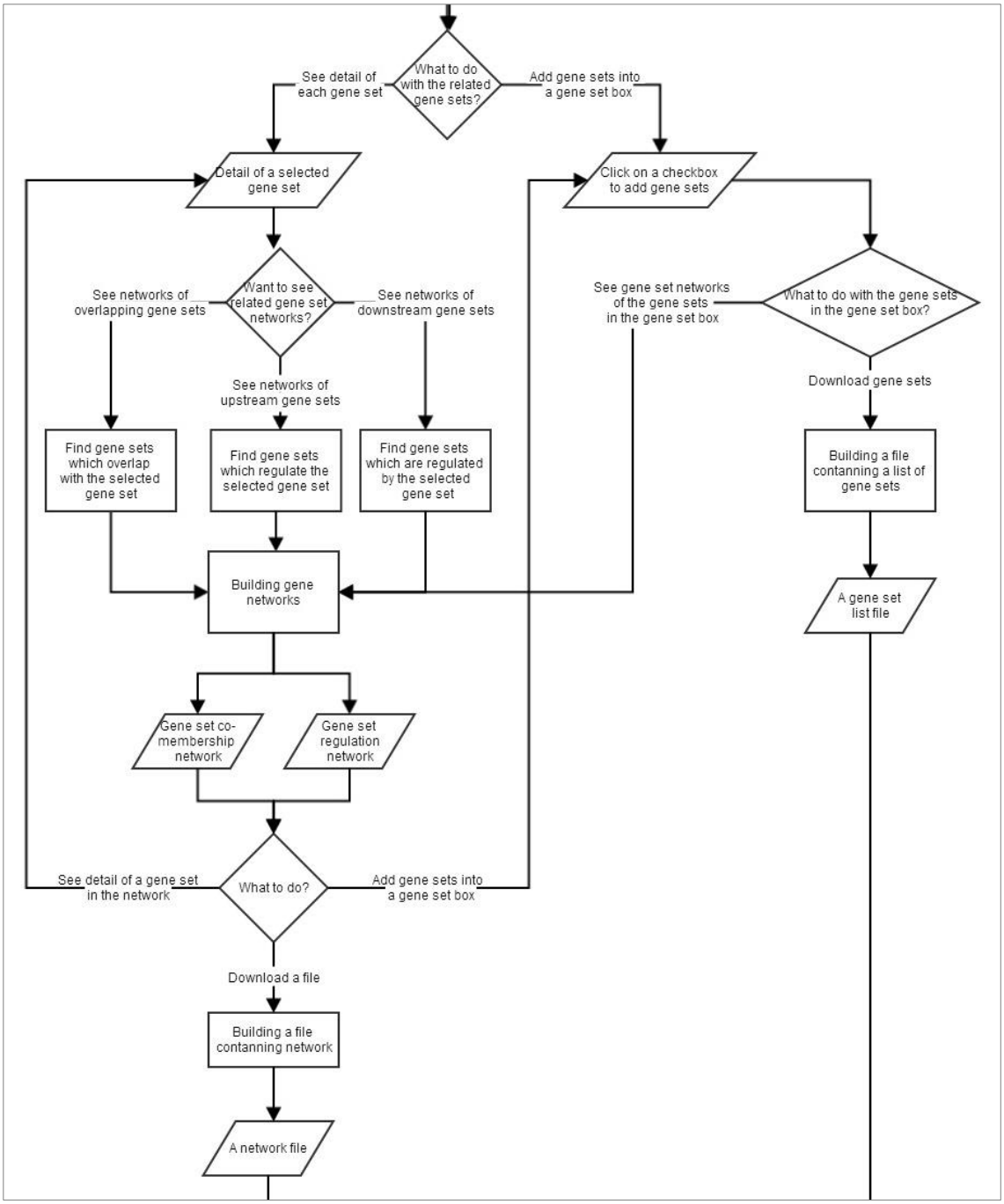



Figure 3.9 PAGER gene set network features

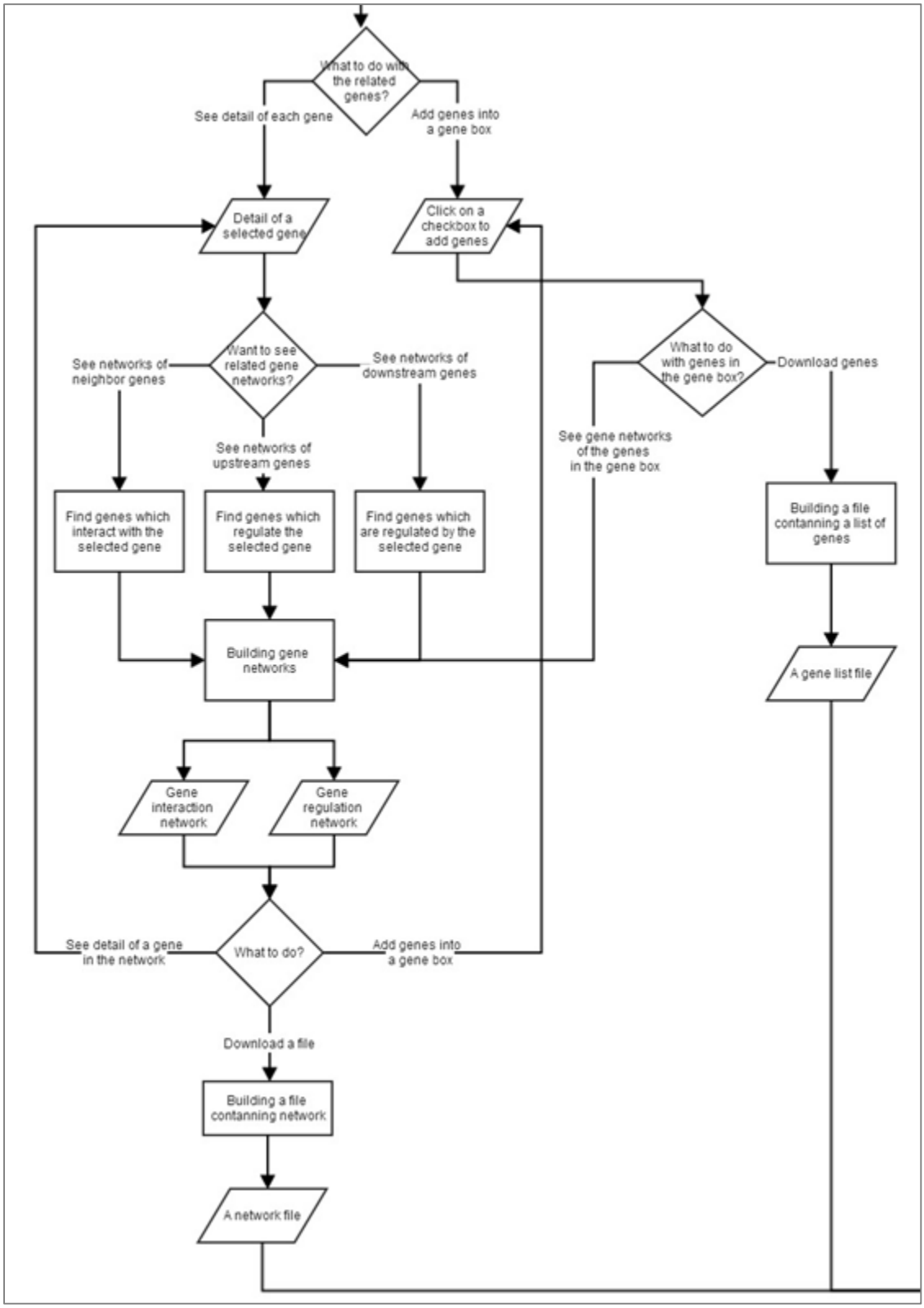



Figure 3.10 PAGER gene network features

### 3.2.4 PAGER Use cases

This section described different ways of using PAGER to study systems biology. Five use cases, including searching gene and gene sets by terms, searching gene sets by a list of genes, constructing gene set networks, generating expanded gene set networks, and viewing gene networks of genes inside a gene set were presented together with screenshots.

#### 3.2.4.1 Searching Genes and Gene Sets by Terms

Users go to PAGER home page (Figure 3.11) and enter searching terms such as a disease name or a gene symbol. For this use case, we entered “non small cell lung” to search for non-small cell lung cancer related gene sets. PAGER returned a list of results (Figure 3.12). The list contains genes and gene sets which relate to the searching terms in different aspects. In this case, “non small cell lung” matched with names of 160 gene sets and descriptions of 720 gene sets. “non small cell lung” is not a name or a symbol of a gene, so PAGER returned 0 for a member of gene set line. However, if users entered “BRAF”, which is a gene symbol, PAGER returned 501 gene sets which contain BRAF gene (Figure 3.13). The next step is clicking on the 160 gene sets which relate to “non small cell lung”. PAGER displayed a list of the 160 gene sets (Figure 3.14). The gene sets can be sorted by name, size, organism, or data source. We filtered only gene sets whose sizes is between 5 and 500 and are from humans. Gene sets can be added to Gene Set Box for further analysis. Checkboxes in the left most column were checked if gene sets were already in the Gene Set Box.

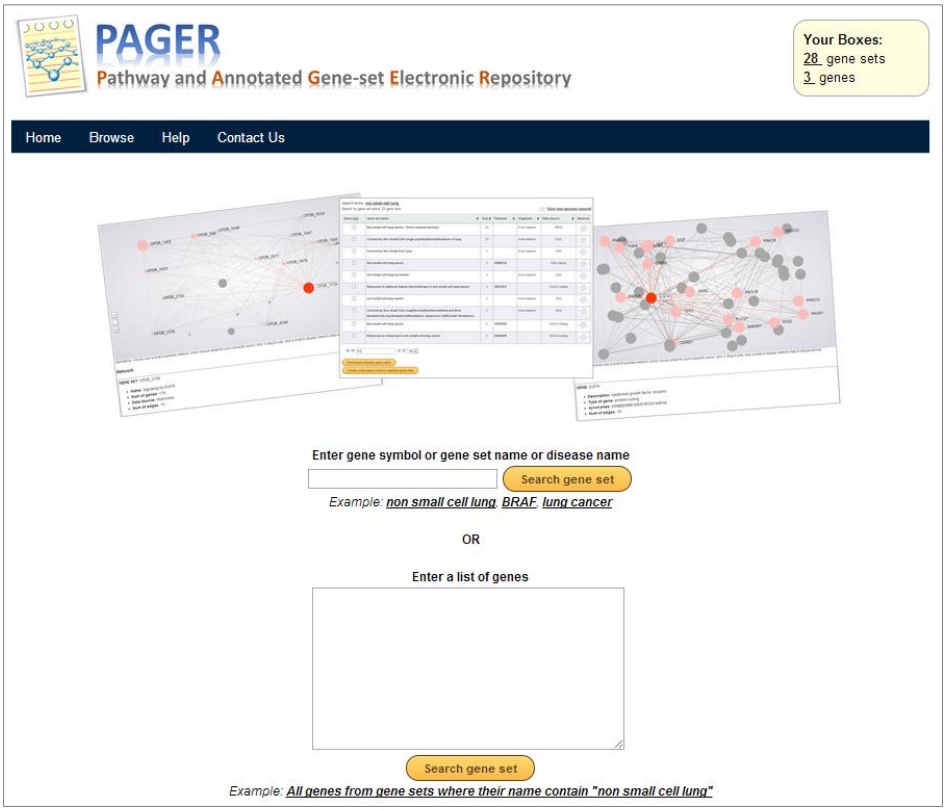



Figure 3.11 PAGER home page

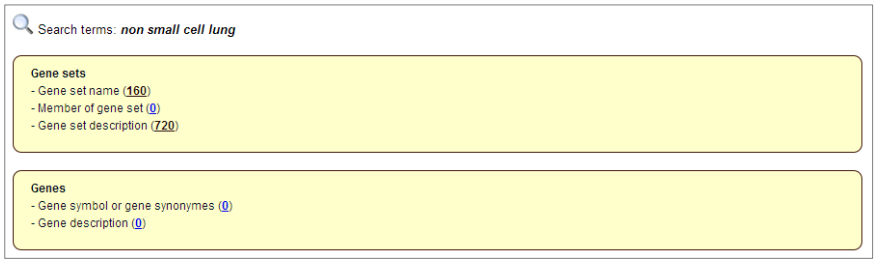



Figure 3.12 PAGER results of searching by “non small cell lung”

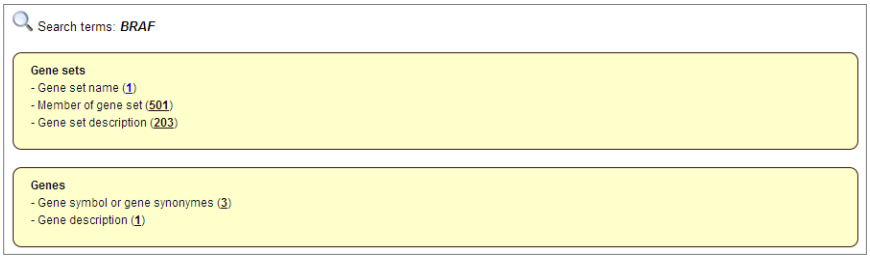
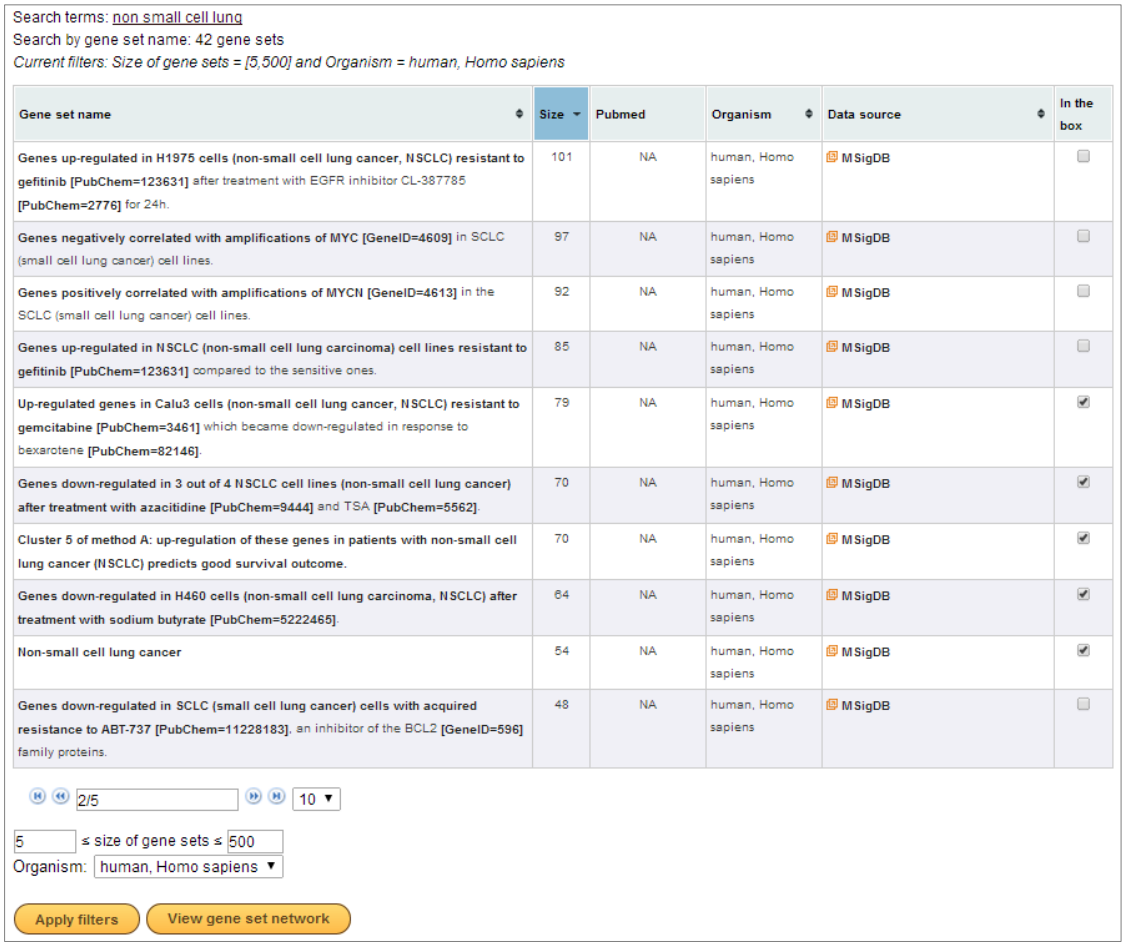



Figure 3.13 PAGER results of searching by “BRAF”



Search terms: non small cell lung
Search by gene set name: 42 gene sets
*Current filters: Size of gene sets = [5,500] and Organism = human, Homo sapiens*

| Gene set name | Size | Pubmed | Organism | Data source | In the box |
|---|---|---|---|---|---|
| **Genes up-regulated in H1975 cells (non-small cell lung cancer, NSCLC) resistant to gefitinib [PubChem=123631]** after treatment with EGFR inhibitor CL-387785 **[PubChem=2776]** for 24h. | 101 | NA | human, Homo sapiens | MSigDB | ☐ |
| **Genes negatively correlated with amplifications of MYC [GeneID=4609]** in SCLC (small cell lung cancer) cell lines. | 97 | NA | human, Homo sapiens | MSigDB | ☐ |
| **Genes positively correlated with amplifications of MYCN [GeneID=4613]** in the SCLC (small cell lung cancer) cell lines. | 92 | NA | human, Homo sapiens | MSigDB | ☐ |
| **Genes up-regulated in NSCLC (non-small cell lung carcinoma) cell lines resistant to gefitinib [PubChem=123631]** compared to the sensitive ones. | 85 | NA | human, Homo sapiens | MSigDB | ☐ |
| **Up-regulated genes in Calu3 cells (non-small cell lung cancer, NSCLC) resistant to gemcitabine [PubChem=3461]** which became down-regulated in response to bexarotene **[PubChem=82146]**. | 79 | NA | human, Homo sapiens | MSigDB | ☑ |
| **Genes down-regulated in 3 out of 4 NSCLC cell lines (non-small cell lung cancer) after treatment with azacitidine [PubChem=9444]** and TSA **[PubChem=5562]**. | 70 | NA | human, Homo sapiens | MSigDB | ☑ |
| **Cluster 5 of method A: up-regulation of these genes in patients with non-small cell lung cancer (NSCLC) predicts good survival outcome.** | 70 | NA | human, Homo sapiens | MSigDB | ☑ |
| **Genes down-regulated in H460 cells (non-small cell lung carcinoma, NSCLC) after treatment with sodium butyrate [PubChem=5222465]**. | 64 | NA | human, Homo sapiens | MSigDB | ☑ |
| **Non-small cell lung cancer** | 54 | NA | human, Homo sapiens | MSigDB | ☑ |
| **Genes down-regulated in SCLC (small cell lung cancer) cells with acquired resistance to ABT-737 [PubChem=11228183]**, an inhibitor of the BCL2 **[GeneID=596]** family proteins. | 48 | NA | human, Homo sapiens | MSigDB | ☐ |

2/5 10

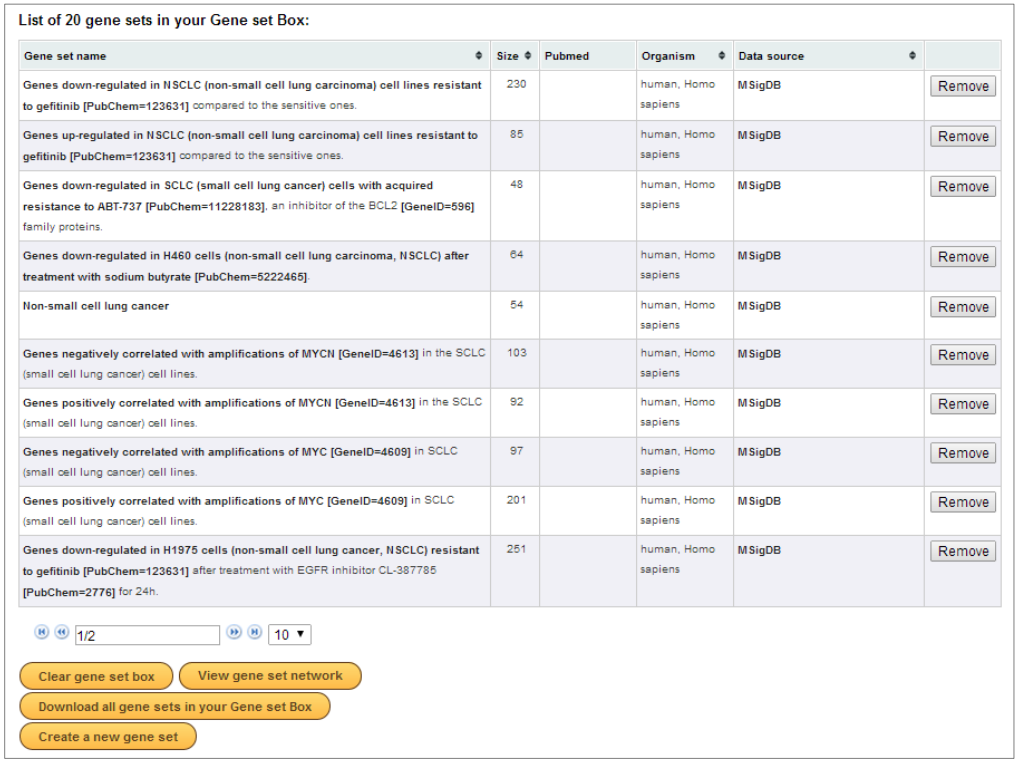
5 ≤ size of gene sets ≤ 500
Organism: human, Homo sapiens

Apply filters View gene set network

Figure 3.14 PAGER displays a list of result gene sets

20 gene sets were added into the Gene Set Box. The status of Gene and Gene Set Box were presented on the top right of the page (Figure 3.11). We clicked in the status box to see all gene sets in Gene Set Box (Figure 3.15). On this page, users can view co-membership and regulatory gene set networks of gene sets in the box, remove gene sets, download gene sets, and create a new gene set by combining all genes from gene sets in the Gene Set Box.

This use case showed that PAGER is useful for searching genes and gene sets related to terms. In case of searching for genes, the searching steps are the same as searching for gene sets. Users can add genes into Gene Box and create gene interaction and gene regulation networks.



List of 20 gene sets in your Gene set Box:

| Gene set name | Size | Pubmed | Organism | Data source | |
|---|---|---|---|---|---|
| **Genes down-regulated in NSCLC (non-small cell lung carcinoma) cell lines resistant to gefitinib [PubChem=123631]** compared to the sensitive ones. | 230 | | human, Homo sapiens | MSigDB | Remove |
| **Genes up-regulated in NSCLC (non-small cell lung carcinoma) cell lines resistant to gefitinib [PubChem=123631]** compared to the sensitive ones. | 85 | | human, Homo sapiens | MSigDB | Remove |
| **Genes down-regulated in SCLC (small cell lung cancer) cells with acquired resistance to ABT-737 [PubChem=11228183]**, an inhibitor of the BCL2 **[GeneID=596]** family proteins. | 48 | | human, Homo sapiens | MSigDB | Remove |
| **Genes down-regulated in H460 cells (non-small cell lung carcinoma, NSCLC) after treatment with sodium butyrate [PubChem=5222465]**. | 64 | | human, Homo sapiens | MSigDB | Remove |
| **Non-small cell lung cancer** | 54 | | human, Homo sapiens | MSigDB | Remove |
| **Genes negatively correlated with amplifications of MYCN [GeneID=4613]** in the SCLC (small cell lung cancer) cell lines. | 103 | | human, Homo sapiens | MSigDB | Remove |
| **Genes positively correlated with amplifications of MYCN [GeneID=4613]** in the SCLC (small cell lung cancer) cell lines. | 92 | | human, Homo sapiens | MSigDB | Remove |
| **Genes negatively correlated with amplifications of MYC [GeneID=4609]** in SCLC (small cell lung cancer) cell lines. | 97 | | human, Homo sapiens | MSigDB | Remove |
| **Genes positively correlated with amplifications of MYC [GeneID=4609]** in SCLC (small cell lung cancer) cell lines. | 201 | | human, Homo sapiens | MSigDB | Remove |
| **Genes down-regulated in H1975 cells (non-small cell lung cancer, NSCLC) resistant to gefitinib [PubChem=123631]** after treatment with EGFR inhibitor CL-387785 **[PubChem=2776]** for 24h. | 251 | | human, Homo sapiens | MSigDB | Remove |

1/2 10

Clear gene set box | View gene set network

Download all gene sets in your Gene set Box

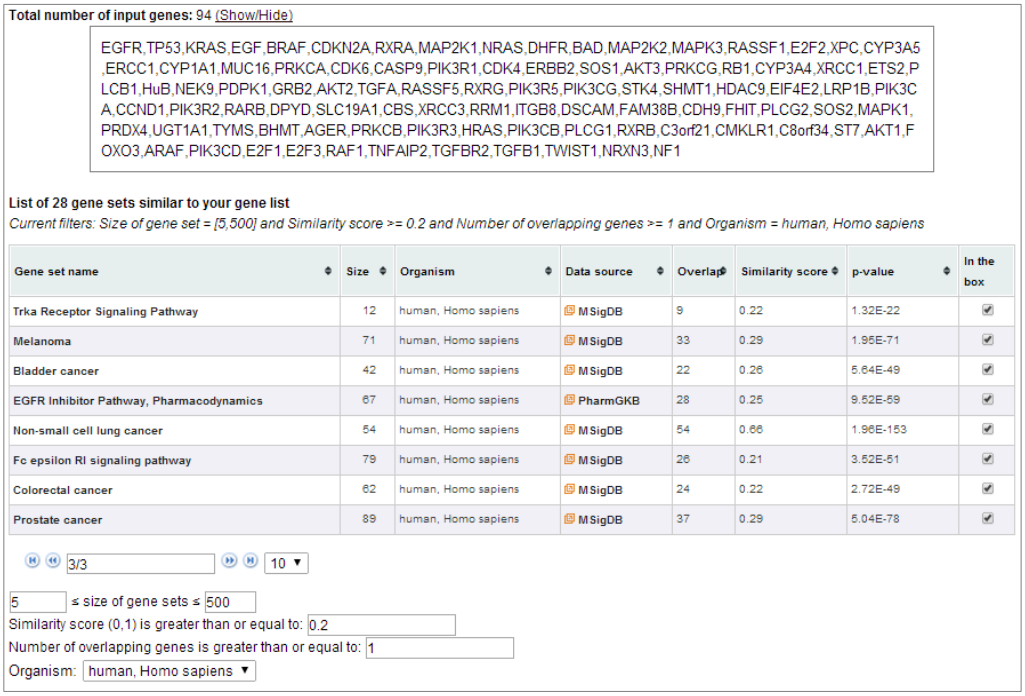
Create a new gene set

Figure 3.15 PAGER displays a list of gene sets in Gene Set Box

### 3.2.4.2 Searching Gene Sets by a List of Genes

On PAGER home page (Figure 3.11), users can enter a list of genes obtained from their experiment or other data sources in order to search for related gene sets. In this use case, a list of 94 non-small cell lung cancer genes were entered. PAGER displayed gene sets which related to the list of 94 genes and 28 gene sets were displayed after we applied

some filters (Figure 3.16). PAGER counted the number of shared genes and calculated a significant value for each gene set. Users can click on a gene set name to see more detail about a gene set or add gene sets into Gene Set Box as in the previous use case. Note that the significant value of each gene set was computed on-the-fly by using hypergeometric function provided by the additional PHP library. Therefore, it is possible that if the number of shared genes or the size of gene set is very high, the hypergeometric function cannot compute p-value because of number overflow.



**Total number of input genes:** 94 (Show/Hide)

EGFR,TP53,KRAS,EGF,BRAF,CDKN2A,RXRA,MAP2K1,NRAS,DHFR,BAD,MAP2K2,MAPK3,RASSF1,E2F2,XPC,CYP3A5,ERCC1,CYP1A1,MUC16,PRKCA,CDK6,CASP9,PIK3R1,CDK4,ERBB2,SOS1,AKT3,PRKCG,RB1,CYP3A4,XRCC1,ETS2,PLCB1,HuB,NEK9,PDPK1,GRB2,AKT2,TGFA,RASSF5,RXRG,PIK3R5,PIK3CG,STK4,SHMT1,HDAC9,EIF4E2,LRP1B,PIK3CA,CCND1,PIK3R2,RARB,DPYD,SLC19A1,CBS,XRCC3,RRM1,ITGB8,DSCAM,FAM38B,CDH9,FHIT,PLCG2,SOS2,MAPK1,PRDX4,UGT1A1,TYMS,BHMT,AGER,PRKCB,PIK3R3,HRAS,PIK3CB,PLCG1,RXRB,C3orf21,CMKLR1,C8orf34,ST7,AKT1,FOXO3,ARAF,PIK3CD,E2F1,E2F3,RAF1,TNFAIP2,TGFBR2,TGFB1,TWIST1,NRXN3,NF1

**List of 28 gene sets similar to your gene list**

*Current filters: Size of gene set = [5,500] and Similarity score >= 0.2 and Number of overlapping genes >= 1 and Organism = human, Homo sapiens*

| Gene set name | Size | Organism | Data source | Overlap | Similarity score | p-value | In the box |
|---|---|---|---|---|---|---|---|
| Trka Receptor Signaling Pathway | 12 | human, Homo sapiens | MSigDB | 9 | 0.22 | 1.32E-22 | ☑ |
| Melanoma | 71 | human, Homo sapiens | MSigDB | 33 | 0.29 | 1.95E-71 | ☑ |
| Bladder cancer | 42 | human, Homo sapiens | MSigDB | 22 | 0.26 | 5.64E-49 | ☑ |
| EGFR Inhibitor Pathway, Pharmacodynamics | 67 | human, Homo sapiens | PharmGKB | 28 | 0.25 | 9.52E-59 | ☑ |
| Non-small cell lung cancer | 54 | human, Homo sapiens | MSigDB | 54 | 0.66 | 1.96E-153 | ☑ |
| Fc epsilon RI signaling pathway | 79 | human, Homo sapiens | MSigDB | 26 | 0.21 | 3.52E-51 | ☑ |
| Colorectal cancer | 62 | human, Homo sapiens | MSigDB | 24 | 0.22 | 2.72E-49 | ☑ |
| Prostate cancer | 89 | human, Homo sapiens | MSigDB | 37 | 0.29 | 5.04E-78 | ☑ |

3/3 10

5 ≤ size of gene sets ≤ 500

Similarity score (0,1) is greater than or equal to: 0.2

Number of overlapping genes is greater than or equal to: 1

Organism: human, Homo sapiens

Figure 3.16 PAGER displays a list of gene sets which relate to the list of 94 genes

This use case showed that PAGER is useful for searching gene sets related to a list of genes together with significant values. These results will be useful for constructing co-

membership and regulatory gene set networks which are corresponding a particular list of genes obtained from biological experiments.

### 3.2.4.3 Constructing Gene Set networks

Users can construct co-membership and regulatory gene set networks of their selected gene sets. In this use case, the 28 gene sets from the previous case study were added into Gene Set Box. On the Gene Set Box page (Figure 3.15), co-membership and regulatory gene set networks were constructed and displayed by clicking “View gene set network” button (Figure 3.17, 3.18).

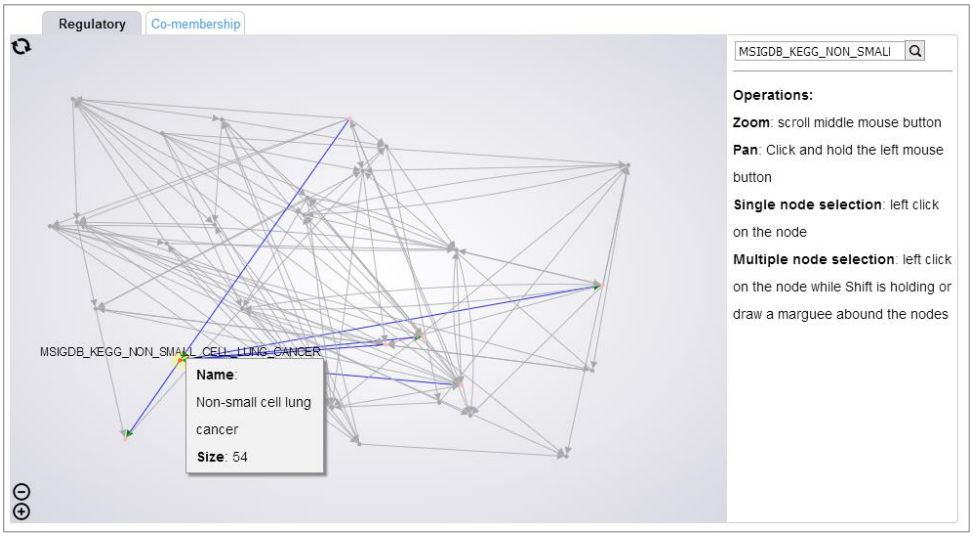



Figure 3.17 PAGER displays a regulatory gene set network of 28 gene sets

An instruction of using gene set network visualization was displayed on the right side. Users move a mouse over a gene set to see more detail about the gene set and see its neighbors (Figure 3.17), and move a mouse over an edge to see the detail of the edge. Users drag a mouse to cover gene set nodes for multiple selection. On the network page,

users can also add or remove gene set from Gene Set Box. Users select a gene set and click the link on the right hand side to see more detail about the gene set. There are two taps for displaying a regulatory gene set network, a directed network, and a co-membership gene set network, an undirected network.

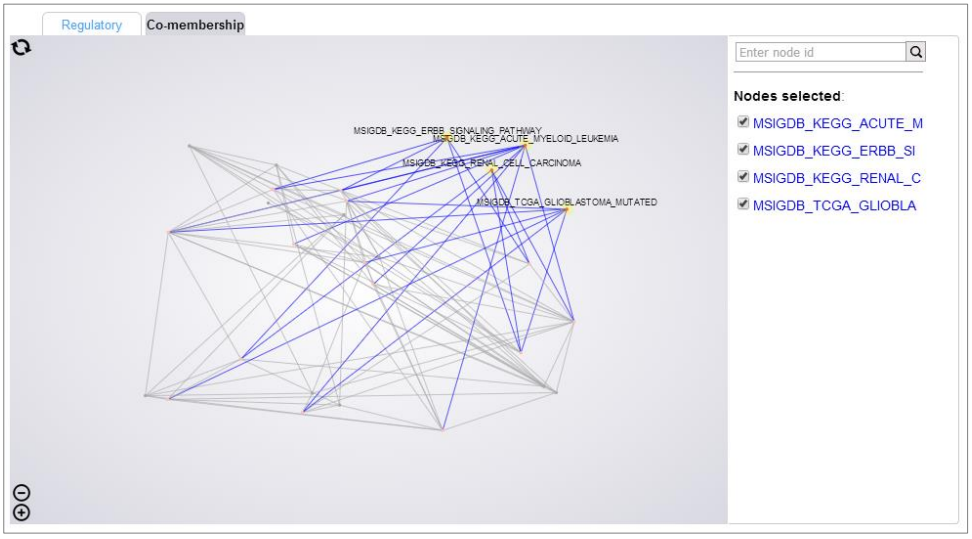



Figure 3.18 PAGER displays a co-membership gene set network of 28 gene sets

This use case showed the main contribution of PAGER, constructing gene set networks corresponding to searching terms or a list of genes. Co-membership and regulatory network were constructed from the pre-computed global gene set networks. This feature make PAGER different from several existing works. MetaNet [32] allows users to construct gene set networks base on a list of genes. However, it allows users to construct a network of gene sets obtained from only one data source. HPD [7] allows users to only search for human pathway using a list of proteins and provide a similarity matrix for the results pathways. PAGED [19] allows users to search for pathways and gene sets by terms or a list

of genes. It provides a downloadable file for co-membership gene set network of the results gene sets. Our PAGER provides a more flexible tool for users to construct gene set networks. Users search and select only gene sets in which they are interested from multiple data sources and view interactive gene set networks on browser.

#### 3.2.4.4 Generating Expanded Gene Set Networks

When a gene set was selected by dragging a mouse cover a node, a link to see more detail about a gene set was displayed on the right hand side (Figure 3.18). A non small cell lung cancer gene set from MSigDB was selected; and Figure 3.19 showed the detail of the gene set. On this detail page, users click on a diagram to construct expanded networks. The expanded networks are networks of upstream, downstream, and co-membership gene sets.

This use case showed another contribution of PAGER, constructing expanded gene set networks. Other existing tools did not provide a way for users to expand a gene set network. By expanding gene set networks, users are not limited to see only a network of their selected gene sets.

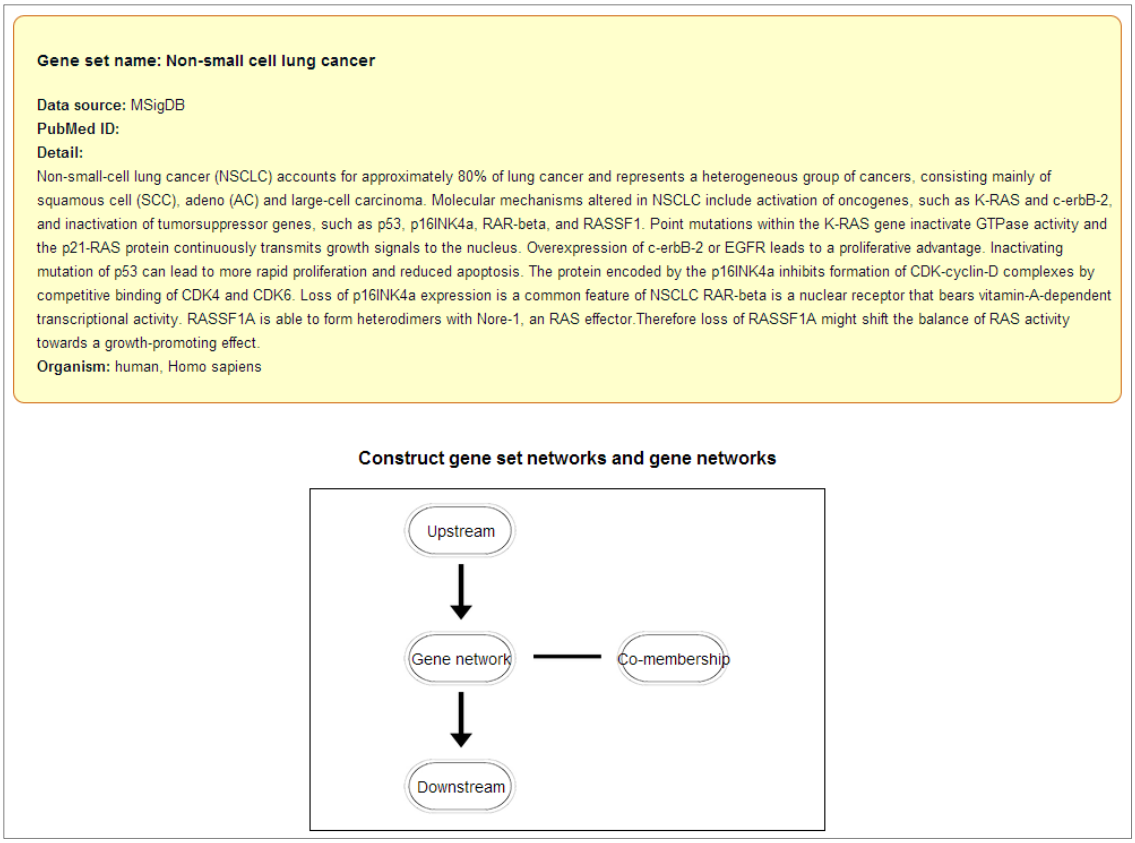



Figure 3.19 PAGER displays a detail of Non small cell lung cancer gene set

### 3.2.4.5 Viewing Gene Networks of Genes inside a Gene Set

In addition to the expanded gene set network, PAGER enables gene interaction and gene regulation network within a gene set. From Figure 3.19, users click on the central button in the diagram to see gene networks of a particular gene set (Figure 3.20). Typically, system biology tools allow users to construct either gene networks or gene set networks. To our knowledge, there is no tool that supports constructing both gene set networks and gene networks for a gene set.

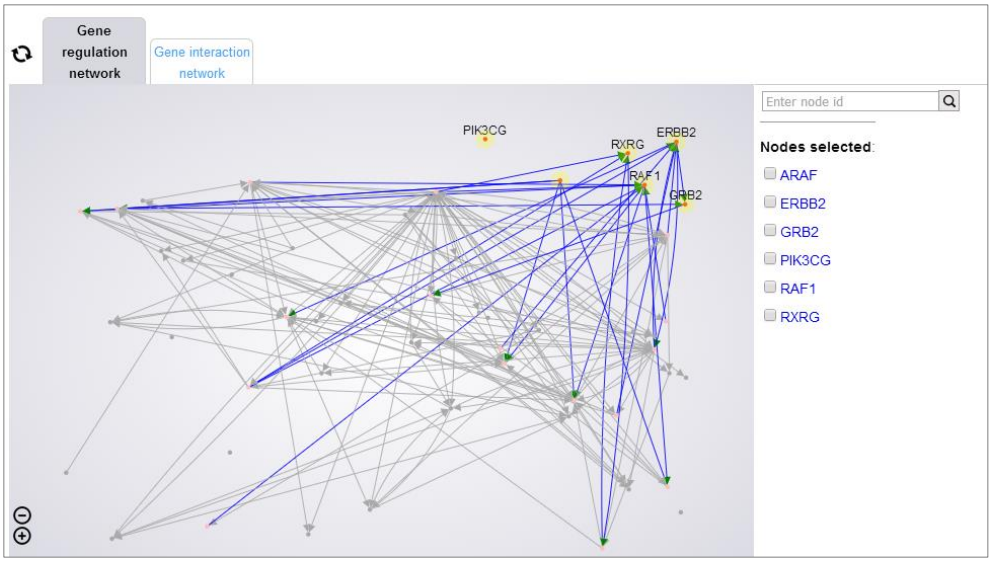



Figure 3.20 A gene network of Non small cell lung cancer gene set

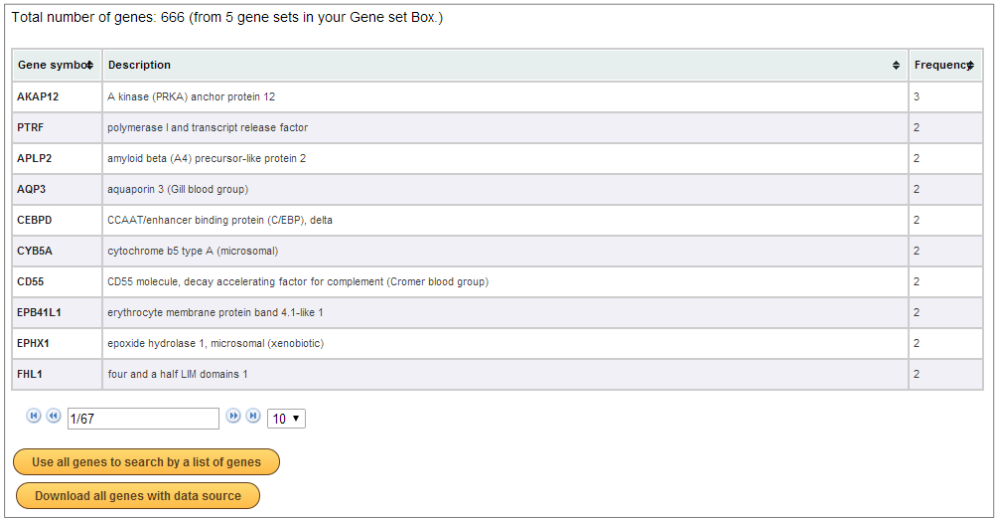
This use case showed that it is useful for researchers to see both levels of system biology networks, gene set level and gene level. For example, users search for gene sets by a list of gene from their experiment, construct gene set networks of the related gene sets to find significant gene sets, and construct gene networks of the significant gene sets to find significant genes.

### 3.2.4.6 Constructing Disease Specific Gene Set Networks

This use case showed how PAGER can be used to construct disease specific gene set networks. In this scenario, a list of disease related genes was not required. We searched for non-small cell lung cancer related gene sets by using “non small cell lung”. The terms matched with the names of 160 gene sets. 5 human gene sets were added to Gene Set Box. On the Gene Set Box page, users create a new gene set by clicking on Create a new gene set button. PAGER displayed a list of gene and the number of gene sets which

contain a particular gene (Figure 3.21). AKAP12 gene has the highest frequency suggesting that it is an important gene among the five gene sets in Gene Set Box. On this page, when users click on Use all genes to search by a list of genes, PAGER automatically generate a list of genes and fill in the text area in PAGER home page (Figure 3.11).

This use case showed that PAGER allows users to search for gene sets which contain a disease name, combine and rank the genes of the result gene sets, use a new list of disease related gene to search for disease related gene sets, and construct disease related gene set networks. This feature is different from direct constructing gene set networks from result gene sets because searching by a disease name returns only gene sets whose names contain the disease name. However, for a particular disease, some gene sets which relate to the disease do not contain the disease name. To our knowledge, only PAGED [19] provide this feature. However, it still lack of network visualization, a regulatory gene set network, and gene networks within a gene set.


Total number of genes: 666 (from 5 gene sets in your Gene set Box.)

| Gene symbol | Description | Frequency |
|---|---|---|
| AKAP12 | A kinase (PRKA) anchor protein 12 | 3 |
| PTRF | polymerase I and transcript release factor | 2 |
| APLP2 | amyloid beta (A4) precursor-like protein 2 | 2 |
| AQP3 | aquaporin 3 (Gill blood group) | 2 |
| CEBPD | CCAAT/enhancer binding protein (C/EBP), delta | 2 |
| CYB5A | cytochrome b5 type A (microsomal) | 2 |
| CD55 | CD55 molecule, decay accelerating factor for complement (Cromer blood group) | 2 |
| EPB41L1 | erythrocyte membrane protein band 4.1-like 1 | 2 |
| EPHX1 | epoxide hydrolase 1, microsomal (xenobiotic) | 2 |
| FHL1 | four and a half LIM domains 1 | 2 |

1/67 10

Use all genes to search by a list of genes

Download all genes with data source

Figure 3.21 A gene network of Non small cell lung cancer gene set

# CHAPTER 4. CONCLUSION

To study complex biological networks, the traditional approach is based on the identification of interactions among internal components of gene sets. Little is known about relationships among higher order biological processes. Several types of the high level network, a gene set network, have been proposed, including co-membership, protein linkage, and co-enrichment gene set networks. In this study, we proposed a method to construct a new type of gene set network, a regulatory gene set network. A regulatory gene set network reveals novel relationships among gene sets together with directionality information. This study consists of two parts, constructing a regulatory gene set network and developing PAGER to allow users to construct several types of gene and gene set networks.

In the first part, a regulatory gene set network and a co-membership gene set network were constructed for each gene set collection. We showed that a regulatory gene set network provides complementary information to a co-membership gene set network, which is commonly constructed by several studies, by presenting the low percentage of shared edges between the two networks. We compared degree centrality of the KEGG co-membership network and degree centrality of the KEGG regulatory network and found that the correlation is relatively high, suggesting that the gene set which is important in

a co-membership gene set network is likely to be important in a regulatory gene set network. However, the different topology of the networks suggests that the two networks can be used to explain different phenomenon in biological systems. To validate both the KEGG co-membership network and the KEGG regulatory network constructed in this study, the two networks were compared with KEGG co-enrichment network obtained from Jignesh, et al. [32]. We found that the number of shared edges between the KEGG co-enrichment network and the KEGG regulatory network and the number of shared edges between the KEGG co-enrichment network and the KEGG co-membership network are significantly high. These results are corresponding to the facts that a pair of gene sets which have a strong regulatory relationship and share a significant number of genes should be connected with co-enrichment. Finally, a regulatory gene set network specific to Alzheimer's disease was constructed. We showed that the network is useful for understanding the underlying mechanism of the disease.

After we found that a regulatory gene set network is useful for systems biology study, PAGER was implemented. PAGER is an online platform for searching gene sets and constructing gene set networks to reveal insights into biological systems. PAGER contains 166,489 gene sets integrated from 10 different gene set data sources. The total number of unique genes is 44,188. Gene regulations and gene interactions were collected from 4 different data sources. The total number of unique gene regulations is 651,586 and the total number of unique gene interactions is 650,160. Human gene regulations were collected and filtered for constructing a regulatory gene set network. PAGER provides pre-

computed global regulatory and global co-membership gene set networks allowing users to construct networks of their gene sets. PAGER has several unique features which have not been provided before by other existing tools. First, PAGER not only allows users to construct two types of gene set networks, PAGER enables users to construct a gene interaction network and a gene regulation network of genes inside a gene set. Second, PAGER allows users to construct three types of expanded gene set networks including networks of upstream, downstream, and co-membership gene sets. Constructing expanded gene set networks enables users to find more important gene sets related to their study. Third, because PAGER offers gene networks, users can also construct expanded gene networks from a gene including networks of upstream, downstream, and sibling genes. Finally, PAGER provides an interactive visualization tool for users to study gene and gene set networks and offers spaces, Gene Box and Gene Set Box, for users to store their genes and gene sets.

In conclusion, we provided a method to construct a regulatory gene set network and methods to construct both global and disease specific gene set networks, which enable future systems biology and translational bioinformatics research. The underlying gene regulation data were collected from high quality and high coverage data sources, so directed edges in a regulatory gene set network do not tend to depend on the number and the quality of experimental data. The directionality information from a regulatory gene set network enables finding of source gene sets and sink gene sets which might be important for drug discovery or drug repositioning. PAGER offers several tools to enable

systems biology and further analysis of gene lists obtained from high throughput experiments. Users can use PAGER to search for genes and gene sets, construct co-membership gene set networks and regulatory gene set networks, and construct gene interaction networks and gene regulation networks in order to understand the underlying processes of disease or drugs.

However, in this study, we have not yet considered tissue specific or disease specific gene regulations, so the disease specific network might not have high accuracy of presenting the high level relationship among gene sets. Therefore, our framework can be improved by collecting higher resolution data, such as tissue specific and disease specific gene regulation data and gene set data. In addition, experimental gene expression data can be used to obtain gene set ranking and can be further applied to add more information to the gene set network. For the current version of PAGER, the sizes of gene and gene set networks are limited. PAGER cannot display a network which has more than 100 nodes, due to browser memory constraint. Our future plans are to integrate tissue specific information and to integrate internal genes or proteins interactions inside pathways in order to support future systems biology study. However, with this first version of PAGER, we enabled researchers to construct several types of networks, especially a regulatory gene set network, to reveal novel insights into complex biological systems.

REFERENCES

# REFERENCES

[1] Ashburner, Michael, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, and Janan T Eppig. 2000. "Gene Ontology: tool for the unification of biology." Nature genetics no. 25 (1):25-29.

[2] Becker, Kevin G, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. 2004. "The genetic association database." Nature genetics no. 36 (5):431-432.

[3] Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological):289-300.

[4] Bertram, Lars, Matthew B McQueen, Kristina Mullin, Deborah Blacker, and Rudolph E Tanzi. 2007. "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database." Nature genetics no. 39 (1):17-23.

[5] Brandes, Ulrik. 2001. "A faster algorithm for betweenness centrality*." Journal of Mathematical Sociology no. 25 (2):163-177.

[6] Centre, Donnelly. Cytoscape.js 2014. Available from http://cytoscape.github.io/cytoscape.js.

[7] Chowbina, Sudhir R, Xiaogang Wu, Fan Zhang, Peter M Li, Ragini Pandey, Harini N Kasamsetty, and Jake Y Chen. 2009. "HPD: an online integrated human pathway database enabling systems biology studies." BMC bioinformatics no. 10 (Suppl 11):S5.

[8] Csardi, Gabor, and Tamas Nepusz. 2006. "The igraph software package for complex network research." InterJournal, Complex Systems no. 1695 (5).

[9] Culhane, Aedín C, Markus S Schröder, Razvan Sultana, Shaita C Picard, Enzo N Martinelli, Caroline Kelly, Benjamin Haibe-Kains, Misha Kapushesky, Anne-Alyssa St Pierre, and William Flahive. 2012. "GeneSigDB: a manually curated database and resource for analysis of gene expression signatures." Nucleic acids research no. 40 (D1):D1060-D1066.

[10] Dotan-Cohen, Dikla, Stan Letovsky, Avraham A Melkman, and Simon Kasif. 2009. "Biological process linkage networks." PloS one no. 4 (4):e5313.

[11] Edelman, Elena J, Justin Guinney, Jen-Tsan Chi, Phillip G Febbo, and Sayan Mukherjee. 2008. "Modeling cancer progression via pathway dependencies." PLoS computational biology no. 4 (2):e28.

[12] EllisLab, Inc. CodeIgniter User Guide Version 2.1.3 2013. Available from http://ellislab.com/codeigniter/user-guide/index.html.

[13] Franceschini, Andrea, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, and Christian von Mering. 2013. "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration." Nucleic acids research no. 41 (D1):D808-D815.

[14] Francesconi, Mirko, Daniel Remondini, Nicola Neretti, John M Sedivy, Leon N Cooper, Ettore Verondini, Luciano Milanesi, and Gastone Castellani. 2008. "Reconstructing networks of pathways via significance analysis of their intersections." BMC bioinformatics no. 9 (Suppl 4):S9.

[15] Hamosh, Ada, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. 2005. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic acids research no. 33 (suppl 1):D514-D517.

[16] Health, US National Library of Medicine and National Institutes of. PubMed 2014. Available from http://www.ncbi.nlm.nih.gov/pubmed.

[17] Hindorff, Lucia A, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. 2009. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proceedings of the National Academy of Sciences no. 106 (23):9362-9367.

[18] Huang, Da Wei, Brad T Sherman, and Richard A Lempicki. 2009. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." Nucleic acids research no. 37 (1):1-13.

[19] Huang, Hui, Xiaogang Wu, Madhankumar Sonachalam, Sammed N Mandape, Ragini Pandey, Karl F MacDorman, Ping Wan, and Jake Y Chen. 2012. "PAGED: a pathway and gene-set enrichment database to enable molecular phenotype discoveries." BMC bioinformatics no. 13 (Suppl 15):S2.

[20] Jiang, C, Zhenyu Xuan, Fang Zhao, and Michael Q Zhang. 2007. "TRED: a transcriptional regulatory element database, new entries and other development." Nucleic acids research no. 35 (suppl 1):D137-D140.

[21] Jones, Christopher. Installing PHP and the Oracle Instant Client for Linux and Windows 2012. Available from http://www.oracle.com/technetwork/articles/technote-php-instant-084410.html.

[22] Joshi-Tope, G, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, and Lisa Matthews. 2005. "Reactome: a knowledgebase of biological pathways." Nucleic acids research no. 33 (suppl 1):D428-D432.

[23] Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research no. 28 (1):27-30.

[24] Kelder, Thomas, Lars Eijssen, Robert Kleemann, Marjan van Erk, Teake Kooistra, and Chris Evelo. 2011. "Exploring pathway interactions in insulin resistant mouse liver." BMC systems biology no. 5 (1):127.

[25] Kelder, Thomas, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. 2012. "WikiPathways: building research communities on biological pathways." Nucleic acids research no. 40 (D1):D1301-D1307.

[26] Khatri, Purvesh, Marina Sirota, and Atul J Butte. 2012. "Ten years of pathway analysis: current approaches and outstanding challenges." PLoS computational biology no. 8 (2):e1002375.

[27] Li, Yong, Pankaj Agarwal, and Dilip Rajagopalan. 2008. "A global pathway crosstalk network." Bioinformatics no. 24 (12):1442-1447.

[28] Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. 2011. "Molecular signatures database (MSigDB) 3.0." Bioinformatics no. 27 (12):1739-1740.

[29] Liu, Zhi-Ping, Yong Wang, Xiang-Sun Zhang, and Luonan Chen. 2010. "Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains." BMC systems biology no. 4 (Suppl 2):S11.

[30] Matys, Vea, Ellen Fricke, R Geffers, Ellen Gößling, Martin Haubrock, R Hehl, Klaus Hornischer, Dagmar Karas, Alexander E. Kel, and Olga V. Kel-Margoulis. 2003. "TRANSFAC®: transcriptional regulation, from patterns to profiles." Nucleic acids research no. 31 (1):374-378.

[31] Meagher, Paul. PDL 2014. Available from http://www.phpmath.com/build02/PDL/docs/index.php.

[32] Parikh, Jignesh R, Yu Xia, and Jarrod A Marto. 2012. "Multi-edge gene set networks reveal novel insights into global relationships between biological themes." PloS one no. 7 (9):e45211.

[33] Paz, Arnon, Zippora Brownstein, Yaara Ber, Shani Bialik, Eyal David, Dorit Sagir, Igor Ulitsky, Ran Elkon, Adi Kimchi, and Karen B Avraham. 2011. "SPIKE: a database of highly curated human signaling pathways." Nucleic acids research no. 39 (suppl 1):D793-D799.

[34] Rice, John. 2006. Mathematical statistics and data analysis: Cengage Learning.

[35] Schaefer, Carl F, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. 2009. "PID: the pathway interaction database." Nucleic acids research no. 37 (suppl 1):D674-D679.

[36] Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, and Eric S Lander. 2005. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences of the United States of America no. 102 (43):15545-15550.

[37] Whirl-Carrillo, M, EM McDonagh, JM Hebert, L Gong, K Sangkuhl, CF Thorn, RB Altman, and Teri E Klein. 2012. "Pharmacogenomics knowledge for personalized medicine." Clinical Pharmacology & Therapeutics no. 92 (4):414-417.

[38] Xia, Junfeng, Qingguo Wang, Peilin Jia, Bing Wang, William Pao, and Zhongming Zhao. 2012. "NGS Catalog: A database of next generation sequencing studies in humans." Human mutation no. 33 (6):E2341-E2355.