NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM TO ANALYZE PATHWAY ACTIVITIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZERRİN IŞIK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

FEBRUARY 2011

Approval of the thesis:

NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM TO ANALYZE PATHWAY ACTIVITIES

submitted by **ZERRİN IŞIK** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering**

Prof. Dr. Mehmet Volkan Atalay
Supervisor, **Computer Engineering Dept., METU**

Assoc. Prof. Dr. Rengül Çetin Atalay
Co-supervisor, **Molecular Biology and Genetics Dept., Bilkent University**

Examining Committee Members:

Prof. Dr. Cevdet Aykanat
Computer Engineering Dept., Bilkent University

Prof. Dr. Mehmet Volkan Atalay
Computer Engineering Dept., METU

Assoc. Prof. Dr. Ferda Nur Alpaslan
Computer Engineering Dept., METU

Assoc. Prof. Dr. Özlen Konu
Molecular Biology and Genetics Dept., Bilkent University

Assoc. Prof. Dr. Tolga Can
Computer Engineering Dept., METU

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ZERRİN IŞIK

Signature :

ABSTRACT

NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM TO ANALYZE PATHWAY ACTIVITIES

Işık, Zerrin
Ph.D., Department of Computer Engineering
Supervisor : Prof. Dr. Mehmet Volkan Atalay
Co-Supervisor : Assoc. Prof. Dr. Rengül Çetin Atalay

February 2011, 122 pages

Current approaches integrating large scale data and information from a variety of sources to reveal molecular basis of cellular events do not adequately benefit from pathway information. Here, we portray a network structure based pathway enrichment system that fuses and exploits model and data: signalling pathways are taken as the biological models while microarray and ChIP-seq data are the sample input data sources among many other alternatives. Our model- and data-driven hybrid system allows for quantitatively assessing the biological activity of a cyclic pathway and simultaneous enrichment of the significant paths leading to the ultimate cellular response.

Signal Transduction Score Flow (SiTSFlow) algorithm is the fundamental constituent of the proposed network structure based pathway enrichment system. SiTSFlow algorithm converts each pathway into a cascaded graph and then gene scores are mapped onto the protein nodes. Gene scores are transferred to *en route* of the pathway to form a final activity score describing behaviour of a specific process in the pathway while enriching the gene node scores. Because of cyclic pathways, the algorithm runs in an iterative manner and it terminates when the node scores converge. The converged final activity score provides a quantitative measure to assess the biological significance of a process under the given experimental conditions. The conversion of cyclic pathways into cascaded graphs is performed by using a linear time multiple source Breadth First Search Algorithm. Furthermore, the proposed network structure based pathway enrichment system works in linear time in terms of nodes and edges of given pathways.

In order to explore various biological responses of several processes in a global signalling network, the selected small pathways have been unified based on their common gene and process nodes. The merge algorithm for pathways also runs in linear time in terms of nodes and edges of given pathways.

In the experiments, SiTSFlow algorithm proved the convergence behaviour of activity scores for several cyclic pathways and for a global signalling network. The biological results obtained by assessing of experimental data by described network structure based pathway enrichment system were in correlation with the expected cellular behaviour under the given experimental conditions.

Keywords: Signal Transduction, Score Flow Algorithm, Multiple Source BFS, Pathway Enrichment, Cyclic Pathways, Biological Graph Merge, Microarray, Chip-seq

ÖZ

YOLAKLARIN AKTİVİTESİNİN ANALİZ EDİLMESİ İÇİN AĞ TABANLI YOLAK ZENGİNLEŞTİRME SİSTEMİ

Işık, Zerrin
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Mehmet Volkan Atalay
Ortak Tez Yöneticisi : Doç. Dr. Rengül Çetin Atalay

Şubat 2011, 122 sayfa

Moleküler tabanlı hücre olaylarını ortaya çıkarmak için farklı kaynaklardan gelen geniş kapsamlı veri ve bilgileri birleştiren yaklaşımlar biyolojik yolak bilgisinden yeterli derecede faydalanmamaktadır. Bu çalışmada veriyi ve modeli kaynaştıran ve kullanan ağ tabanlı yolak zenginleştirme sistemi tanımlıyoruz: yolaklar biyolojik modeler olarak kullanılırken, mikrodizi ve ChIP-seq verileri ise girdi verisi olarak alınmıştır. Model ve veri tabanlı olan melez sistemimiz döngüsel yolakların biyolojik aktivitelerini nicel olarak değerlendirmesine olanak tanır ve temel hücresel tepkilere yol açan anlamlı patikaların eşzamanlı olarak zenginleştirmesini sağlar.

Sinyal Aktarımlı Skor Akışı (SiTSFlow) algoritması geliştirilen ağ tabanlı yolak zenginleştirme sisteminin temel yapıtaşıdır. SiTSFlow algoritması her yolağı kademeli bir çizgeye dönüştürür ve gen puanları protein düğümlerine değer olarak verilir. Biyolojik süreçlerin tepkilerini ifade eden son aktivite puanı, gen puanlarının yolak içindeki topolojik akışa göre aktarılmasıyla oluşturulur. Döngüsel yolaklar nedeniyle, algoritma yinelemeli olarak çalışır ve düğümlerin puanları yakınsadığı zaman sonlanır. Verilen deneysel koşullarda anlamlı olan biyolojik süreçlerin değerlendirmesinde, bu yakınsamış son aktivite puanı niceliksel bir ölçüt sağlamaktadır. Döngüsel yolakları kademeli çizgeye dönüştürme işlemi doğrusal zamanda çalışan çok kaynaklı sığ öncelikli arama (Breadth First Search) algoritması ile gerçekleştirilmektedir. Ayrıca, geliştirilen ağ tabanlı yolak zenginleştirme sistemi de yolakların içerdiği düğüm ve kenar sayısına göre doğrusal zamanda çalışmaktadır.

Seçilen küçük sinyal yolakları ortak gen ve süreçler taban alınarak, evrensel sinyal ağındaki farklı süreçlerin verdiği çeşitli biyolojik tepkileri araştırmak için birleştirilmektedir. Yolakları birleştirme algoritması da yolakların içerdiği düğüm ve kenar sayısına göre doğrusal zamanda çalışmaktadır.

Yapılan deneylerde, SiTSFlow algoritması biyolojik aktivite puanlarının yakınsama durumunu döngüsel yolaklarda ve evrensel sinyal ağında ispatlamıştır. Deneysel verilerin geliştirilen ağ tabanlı yolak zenginleştirme sistemi ile değerlendirilmesiyle elde edilen biyolojik sonuçlar, verilen koşullar için beklenen hücresel tepkilerle ilişkilidir.

Anahtar Kelimeler: Sinyal Aktarımı, Skor Akışı Algoritması, Çok Kaynaklı BFS, Yolak Zenginleştirme, Döngüsel Yolaklar, Biyolojik Çizge Birleştirme, Mikrodizi, Chip-seq

To my little son and dearest family

ACKNOWLEDGMENTS

I am deeply indebted to many people who helped me to complete my Ph.D. at the Department of Computer Engineering in METU. First of all, I would like to thank my supervisor Volkan Atalay who encouraged me to work on projects in which I have had most interest. I am thankful to him for giving useful advice, for sharing experiences, and most importantly for teaching me what is required to be a successful academician! I am grateful to my co-superviser, Rengül Çetin Atalay. She taught me the techniques for understanding and interpretation of biological phenomena. Whenever I stuck into a problem in our method, she always pointed me to the right direction to continue my research. I would like to thank Cevdet Aykanat for his very useful suggestions during development of our method. I also thank my thesis committee members, Özlen Konu and Tolga Can, for their helpful suggestions and comments during my study. I have to thank Tülin Erşahin for helping during interpretation of our results in biological perspective.

I'd like to thank to each member of my department who always behave as a big family. I am grateful to my friends, Ayşe Gül Yaman, Oral Dalay, Sinan Saraç, Özlem Erdaş, Özge Öztimur, Serdar Çiftçi, Mine Yoldaş, Hilal Kılıç, and Selma Süloğlu. I would give a special thank my office mate Ayşe Gül for supporting me during tough times.

I am grateful to Scientific and Technological Research Council of Turkey (TÜBİTAK) who provided me full Ph.D. fellowship and the travel grant.

Finally, a special thank you goes to my family. They have always given me their unconditional love and supported me in my life and education. I'd like to give my heartfelt thanks for *my mother* and *my sister*, they have been always there and helped me whenever I need. I am grateful to each member of *Erğin* family, they accepted me as a part of the family and provided enjoyable life in Ankara. I want to give a very special thank my dear son, *Arda*, for being very patient despite his young age. From now on, I promise to play whenever you want!

TABLE OF CONTENTS

ABSTRACT . . . iv

ÖZ . . . vi

ACKNOWLEDGMENTS . . . ix

TABLE OF CONTENTS . . . x

LIST OF TABLES . . . xiii

LIST OF FIGURES . . . xvi

LIST OF ABBREVIATIONS . . . xxi

CHAPTERS

- 1 INTRODUCTION . . . 1
  - 1.1 Motivation . . . 1
  - 1.2 Problem Definition . . . 2
  - 1.3 State of Pathway Enrichment in Literature . . . 3
  - 1.4 Contributions . . . 4
  - 1.5 Organization of the Thesis . . . 6
- 2 BACKGROUND INFORMATION ON BIOLOGICAL AND COMPUTATIONAL ASPECTS . . . 7
  - 2.1 Biological Pathways . . . 7
  - 2.2 Microarray Technology and Analysis . . . 8
    - 2.2.1 Analysis Methods . . . 10
    - 2.2.2 Pathway Enrichment . . . 11
    - 2.2.3 Network Construction . . . 13
  - 2.3 ChIP-Sequencing Data and Analysis . . . 14
    - 2.3.1 ChIP Technologies . . . 14
    - 2.3.2 Raw Data Processing . . . 16

2.3.3 Analysis Tools 17

2.4 Graph Models and Computational Approaches 20

2.4.1 Notation 20

2.4.2 Graph Traversal 21

2.4.2.1 Depth First Search 21

2.4.2.2 Breadth First Search 22

2.4.3 Cycle Detection 23

3 NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM . 25

3.1 System Overview 25

3.2 Data Processing 27

3.3 Data Integration 28

3.4 Pathway Scoring by SiTSFlow Algorithm 29

3.5 Gene Knockout and Its Evaluation by SiTSFlow Algorithm 34

3.6 Calculation of Significance and Sensitivity of Activity Scores 35

3.7 Computational Complexity of SiTSFlow Algorithm 37

3.8 Convergence Analysis of SiTSFlow Algorithm 37

4 EXPERIMENTAL RESULTS OF NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM 42

4.1 Data Sets 42

4.2 KEGG Pathways 45

4.3 Application of SiTSFlow with HeLa Cells Under Oxidative Stress . . 46

4.4 Application of SiTSFlow with Estradiol Treated MCF7 Cells 50

4.5 Application of SiTSFlow with Estrogen Receptor Beta Treated U2OS Cells 54

4.6 Scores from Individual Data Scorings 58

4.7 Effect of Gene Knockout on Pathway Enrichment 60

4.8 Comparison of Initial and Final Scores 66

4.9 Cytoscape Plug-in 70

4.10 Comparison with Enrichment Tools 74

5 TOWARDS CONSTRUCTING GLOBAL SIGNALLING NETWORK . . . 78

5.1 Method Overview 78

5.2 Merge Algorithm 79
5.3 Calculation of Significance and Sensitivity of Activity Scores 83
5.4 Computational Complexity 83
5.5 Experimental Results 85
5.6 Discussion 88
6 CONCLUSION 93
REFERENCES 99

APPENDIX

A SCREENSHOTS OF EMPLOYED PATHWAYS 107
VITA 121

LIST OF TABLES

TABLES

Table 3.1 The activity scores of nodes and *Anti-apoptosis* process. The scores are calculated after running of 1st and 10th iterations of SiTSFlow algorithm. The initial and converged scores are represented in *1st iteration* and *10th iteration* columns of table, respectively. 39

Table 4.1 The details of employed data sets. *Peak Detection* column represents total number of significant peak regions identified in *peak detection* phase. *Gene Mapping* column represents total number of neighboring genes found in *gene mapping* phase. Microarray analysis column represents total number of genes in a chip after completing *microarray analysis* phase. 45

Table 4.2 The details of selected pathways from KEGG PATHWAY Database. The total number nodes, edges, and processes contained in each pathway are listed. . . 46

Table 4.3 Activity scores of biological processes for control and oxidative stress samples in HeLa cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face. 48

Table 4.4 Activity scores of biological processes for control and E2 samples in MCF7 cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face. 51

Table 4.5 Activity scores of biological processes for ERb and E2 samples in U2OS cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face. 55

Table 4.6 Comparison of activity scores based on integrated gene scores, only microarray scores, and only ChIP-seq scores for several signalling pathways for Erb and E2 experiments in U2OS cells. Significant activity score of each process is marked by bold face. 59

Table 4.7 The original scoring results for Akt pathway by using control (C), Gly12Asp (D), and Gly12Val (V) samples of KRas data. The significance value of each score is specified by the α_{value} at the right column of its score. 65

Table 4.8 Gene knockout results for *Akt knockout*, *p53 knockout*, and *Akt-Erk double-knockout* pathways by using control (C), Gly12Asp (D), and Gly12Val (V) samples of KRas data. The significance value of each score is specified by the α_{value} at the right column of its score. 65

Table 4.9 An example for gene score file. It shows some examples from gene rank scores of ChIP-seq control sample for MCF7 cells. 72

Table 4.10 SPIA results on KRas data set using control vs. G12Val mutation. FDR and $FWER$ were calculated for P_G. 75

Table 4.11 GSEA results on KRas data set using control vs. G12Asp mutation. 75

Table 4.12 GSEA detailed scores for *Reactome Apoptosis*. 76

Table 5.1 Activity scores of processes in global signalling network for control and oxidative stress samples in HeLa cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face. 89

Table 5.2 Activity scores of processes in global signalling network for control and E2 samples in MCF7 cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face. 90

Table 5.3 Activity scores of processes in global signalling network for ERb and E2 samples in U2OS cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face. 91

Table 5.4 Activity scores of processes in global signalling network for control (C), Gly12Asp (D), and Gly12Val (V) samples of KRas data. The significance value of each score is specified by α_{value} and σ_{value} at the right column of its score. Significant activity score of each process is marked by bold face. 92

LIST OF FIGURES

FIGURES

Figure 2.1 Microarray analysis steps. Experimental design is performed based on the biological question. RNA extraction, RNA labeling and hybridization steps are then performed in laboratory environment. Image analysis, quantification of gene expression and normalization are the main steps of data pre-processing. Application of significance tests, clustering and prediction are performed in computational analysis step. Finally, a differentially expressed gene set is constructed. Pathway analysis could be applied to explore the biological function of this gene set. Alternatively, network construction might be performed based on this gene set. 9

Figure 2.2 Comparison of ChIP-chip and ChIP-seq technologies. A ChIP sample is prepared to represent genome samples for protein-DNA bindings. In ChIP-chip technology, the ChIP sample is hybridized to a microarray that investigates entire genome by using probes. In ChIP-seq technology, the ChIP sample is sequenced from both ends to construct millions of short reads using massively parallel sequencing. Control samples are used to remove data biases in computational analysis (adapted from Ji et al. [1]). 15

Figure 2.3 An example of peak region at chromosome 14 between 23700178 and 23700299 positions. The genes ISGF3G and RNF31 are in TSS-upstream and TES-downstream of this peak region. ChIP-seq data is STAT1 data set [2]. 19

Figure 3.1 Diagram of the proposed system. Transcriptome and ChIP-seq data are combined to obtain integrated scores of genes. In the pathway scoring phase, pathways activated under experimental conditions are identified by exploring scores of each pathway. SiTSFlow algorithm computes the activity score of each process that is represented as output. 26

Figure 3.2 Score flow of integrated microarray and ChIP-seq scores to a target process, *Anti-apoptosis*, for HeLa cells under oxidative stress condition. The blue number on each node represents self-score of the gene. Red and green edges represents activation and inhibition properties, respectively. The out-score of a parent node is distributed to all of its children according to the magnitude of their self-scores. An activation edge directly partitions the out-score of the parent between the children nodes. However, a negative score is transferred by the inhibition edges to the children nodes. 30

Figure 3.3 Activity score calculation for the cyclic target process (*Anti-apoptosis*) by using integrated gene scores. The number on each node (gene) represents self-score of the gene. Red and green edges represent activation and inhibition properties, respectively. The nodes and edges belong to a cycle are marked by yellow color. 41

Figure 4.1 Activity scores of genes and processes in original Akt pathway by using KRas control sample. Down-regulated and up-regulated genes or processes are represented in color tones of green and red, respectively. 63

Figure 4.2 Cytoscape view of *Apoptosis* and *Cell cycle* processes in p53 knockout pathway by using KRas data. Activity scores of *Apoptosis* and *Cell cycle* processes for control sample (A), for G12D (B), and for G12V (C) in Colo741 cells are given in Table 3. Down-regulated and up-regulated genes or processes are represented in color tones of green and red, respectively. 64

Figure 4.3 Convergence graphics of activity scores for the biological processes of Apoptosis, Jak-STAT signalling, and Pathway in cancer pathways by using control sample in HeLa cells. 67

Figure 4.4 Convergence graphics of activity scores for the biological processes of Apoptosis, Jak-STAT signalling, and Pathway in cancer pathways by using control sample in MCF7 cells. 68

Figure 4.5 Convergence graphics of activity scores for the biological processes of Apoptosis, Jak-STAT signalling, and Pathway in cancer pathways by using control sample in U2OS cells. 69

Figure 4.6 The screenshot of *Data Panel* of Cytoscape. It displays node (a) and edge (b) attributes of hsa04630 - Jak-STAT signalling pathway after running of SiTS-Flow algorithm. 72

Figure 4.7 The screenshot of Cytoscape plug-in. It displays the activity scores of the nodes in hsa04630 - Jak-STAT signaling pathway using HeLa cells under oxidative stress condition. The genes and processes are represented by circles and rectangles, respectively. Node scores are represented in color tones of green, yellow, and red. The edges carry the out-score of a parent node to its children by multiplying out-score by 1 or -1 weight for activation or inhibition, respectively. The final activity of each process is given under its name. 73

Figure 4.8 Enrichment plot of *Reactome Apoptosis*. The profile of the Running ES Score and positions of gene set members on the rank ordered list given in Table 4.12. 76

Figure 4.9 Gene expression sample of HeLa cells under oxidative stress was mapped onto Jak-STAT signalling pathway by using *kegArray* tool. Green and orange colors indicate down-regulation and up-regulation values, respectively. 77

Figure 5.1 A simple example to explain merge algorithm for given pathways. In the pre-processing phase, the nodes having several clones sharing the same gene name are identified as *duplicated nodes* for both pathways. If there exists such duplicated nodes e.g., X in Pathway 1, only one node represents all relations of such duplications. The common nodes e.g., B between two input pathways are then identified. The unification phase operates the union of graph nodes and edges based on the common node(s). After performing unification, the constructed pathway is given as the output. 80

Figure 5.2 An example to explain conflicting edges problem in unification of two given graphs. Node A and B are common nodes between $\mathcal{G}_1$ and $\mathcal{G}_2$. Before performing unification, we should check the edge type between these nodes, since both nodes will appear in $\mathcal{G}_{merge}$ and type of relation between these nodes should be identical. Therefore, user decides which edge type will be assigned as the final edge relation of node A and B in $\mathcal{G}_{merge}$. 81

Figure 5.3 The screenshot of the global signalling network that contains 450 nodes, 650 edges, and 24 biological processes. 84

Figure A.1 Screenshot of original *Apoptosis* pathway from KEGG PATHWAY Database. 108

Figure A.2 Screenshot of original *Cell cycle* pathway from KEGG PATHWAY Database. 109

Figure A.3 Screenshot of original *ErbB signalling* pathway from KEGG PATHWAY Database. 110

Figure A.4 Screenshot of original *Focal Adhesion* pathway from KEGG PATHWAY Database. 111

Figure A.5 Screenshot of original *Insulin signalling* pathway from KEGG PATHWAY Database. 112

Figure A.6 Screenshot of original *Jak-STAT signalling* pathway from KEGG PATHWAY Database. 113

Figure A.7 Screenshot of original *MAPK signalling* pathway from KEGG PATHWAY Database. 114

Figure A.8 Screenshot of original *mTOR signalling* pathway from KEGG PATHWAY Database. 115

Figure A.9 Screenshot of original *P53 signalling* pathway from KEGG PATHWAY Database. 116

Figure A.10 Screenshot of original *Pathways in cancer* pathway from KEGG PATHWAY Database. 117

Figure A.11 Screenshot of original *Regulation of actin cytoskeleton* pathway from KEGG PATHWAY Database. 118

Figure A.12 Screenshot of original *TGF-β signalling* pathway from KEGG PATHWAY Database. 119

Figure A.13 Screenshot of original *Wnt signalling* pathway from KEGG PATHWAY Database. 120

LIST OF ABBREVIATIONS

$\Delta(x)$	Expression difference between the levels of 12 and 48 hours of E2 samples in MCF7 cells.
$\Delta_{E2}(x)$	Expression difference between the levels ERb-doxy and control of MCF7 cells.
$\Delta_{ERb}(x)$	Expression difference between the levels ERb-doxyE2 and control of MCF7 cells.
e	Total number of edges in $\mathcal{E}$ set.
ε	Error threshold for convergence criteria.
$\mathcal{G}$	A graph consisting of node set V and edge set E .
$\Delta_{G12A}(x)$	Expression difference between the samples of Gly12Asp and control in KRas data.
$\Delta_{G12V}(x)$	Expression difference between the samples of Gly12Val and control in KRas data.
$\mathcal{G}_{merge}$	Union graph constructed by the merge algorithm.
GSEA	Gene set enrichment analysis.
H_n	Null hypothesis designed for assessment tests.
$inAdj(x)$	In-adjacency list of node x.
n	Total number of nodes in $\mathcal{V}$ set.
N	Total number of iterations performed in permutation procedure.
n_i	Number of reads counted in each window i.
NSBA	Network structure based analysis.
$outAdj(x)$	Out-adjacency list of node x.
$RP(x)$	Rank product of individual rank scores of gene x.
R_{true}	Comparison measure used in assessment tests.
$r(x)$	Order of the score of gene x when all of the scores are ordered in the ascending order.

$R(x)$	Rank of gene x.
α_{value}	Significance value obtained in permutation procedure.
SGSA	Significant gene set analysis.
s_{1i}	Number of reads extracted from the ChIP sample.
s_{2i}	Number of reads in the control sample.
σ_{value}	Sensitivity value obtained in permutation procedure.
SiTSFlow	Signal Transduction Score Flow algorithm developed in the course of thesis.
$S^{k}_{out}(x)$	Out-score of a node x at iteration k.
$S_{tot}(p)$	Final (converged) activity score of a process p.
$S(x)$	Self-score of the gene x obtained by transformation of $RP(x)$ value.
TF	Transcription factor.
$\mathcal{V}_0$	Set of nodes with zero in-degree.
σ^2	Variance of activity scores computed in permutation procedure .
w	Length of non-overlapping windows which is used in ChIP-seq data analysis.

CHAPTER 1

INTRODUCTION

1.1 Motivation

High-throughput biological experiments are designed to analyze biological responses of thousands of genes or proteins under specific experimental conditions. Due to huge data producing capacity within short time, massive output of high-throughput biological experiments is described as *large scale data*. Gene expression by microarray, proteomics and metabolomics data by mass spectrometry, and protein-DNA interaction by chromatin immunoprecipitation sequencing are the popular types of large scale data sources. In the last decade, large scale biological data sets have become publicly available for whole genomes and for several species. The ultimate goal of bioinformatics as a research field is to analyze and integrate different types of large scale data sources to understand complex biological phenomena.

Two examples of popular large scale data sources are microarray gene expression and chromatin immunoprecipitation (ChIP) sequencing experiments that provide large scale transcriptome data about the biological responses of genes for specific experimental conditions. Microarray gene expression experiments enable to access the expression profiles of several genes simultaneously under a particular condition. The result of traditional microarray analysis methods is generally list of significant genes that are assumed to be related with the particular condition of the experiment. The second transcriptome data source is combination of the chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq) technologies that detect the location of DNA binding sites which lead to explore functional elements in the entire genome. Thus, huge amount of DNA-protein association data provided by ChIP-seq experiments may help to understand the observed changes in gene regulations for the entire genome.

A biological network represents several types of experimental interactions in the form of node-edge structured graph. Biological networks have several types: gene regulatory, protein-protein interaction, signal transduction, and metabolite. Pathway is a term used to indicate a subset of biological networks. A process in a pathway describes a small specific unit performing well-defined biological event(s). In graph structure, processes are represented by leaf nodes that might be referred as *final nodes* of a graph. A pathway might contain several processes that generally work collectively. For example, *Apoptosis* is the process of cell death and it may occur with coordination of several pathways. In the computer science terminology, pathways might be described as directed cyclic graphs. A particular set of pathways are *signalling pathways* that represent an abstract information about the collective working mechanism of proteins and other chemical compounds to transfer special cell signals. Signal transduction between biological processes provides a cascaded topology for signalling pathways. Therefore, cascaded topology facilitates the modeling of a biological process with a dynamic nature. Pathway analysis yields molecular level interaction information about genes.

Large scale transcriptome data projects instant biological responses of genes during the experiment, while pathways model dynamic flow of various biological events that work collectively. When large scale transcriptome data and pathways are combined, such a data and model integration is called *pathway enrichment*. If a pathway enrichment is realized for a disease treatment experiment, both process interactions at the molecular level and gene responses of applied treatment are combined, hence enrichment simplifies the understanding of complex disease progression mechanisms. Thus, computational enrichment approaches might yield more realistic and successful drug designs based on *in silico* experiments.

1.2 Problem Definition

Transcriptome data analysis methods provide a list of significant genes that are assumed to be related with experimental conditions or disease treatments. At this point of the analysis, the essential issue is to determine how to map a set of significant genes identified in transcriptome experiments onto pathway models. In order to upgrade the analysis to system level, pathway enrichment methods contemplate to incorporate pathway topological information and transcriptome data.

The main problem addressed in this thesis is the incorporation of large scale biological data and pathway models to evaluate biological processes or pathways which are activated under given experimental conditions. Merging of several small pathways constructs a comprehensive network which would effectively explain the relations between various biological processes previously computed independently. Therefore, the second problem addressed in this thesis is the unification of small biological pathway models to provide a broader perspective to complex biological phenomena.

1.3 State of Pathway Enrichment in Literature

Pathway enrichment tools provide interpretation of the gene expression profiles by either identifying major genes or pathways based on traditional statistical tests or allowing visualization of gene expression data on molecular pathways. Pathway enrichment methods can be classified into three types of approaches: Significant Gene Set Analysis, Gene Set Enrichment Analysis, and Network Structure Based Analysis.

The pioneer enrichment approach Significant Gene Set Analysis (SGSA) is based on identification of significant function annotations. SGSA takes differentially expressed genes as input and then iteratively checks the existence of significant functional annotations for the input genes by using public annotation databases. p-value for each identified functional annotation is computed by applying known statistical methods. The main drawback of this approach is the dependence of the enriched annotations to the initially given differentially expressed gene list, since the method used in analysis of significant genes and cutoff thresholds highly affect the result of final enrichment analysis.

Gene Set Enrichment Analysis (GSEA) calculates an enrichment score (ES) according to the matching of input genes in a pre-built gene list. The input genes list is ordered according to a difference metric, such as fold-change or gene expression difference. Pathway database of GSEA has several pre-built gene lists that contain an ordered gene list in which genes are member of previously known pathways. Matching between input list and pre-built gene lists is performed to find out their correlations. If the input gene set is correlated with a biological process or pathway, the input genes appear usually in the top or bottom of the pathway gene list and ES will be very high. The main benefit of GSEA is the usage of all genes from a microarray experiment without applying a gene selection method and cutoff threshold. However, the drawback of GSEA is the dependence of the enrichment score to the ordering of the input gene list. Furthermore, GSEA approach does not incorporate pathway topology.

Network Structure Based Analysis (NSBA) approach aims to compute a sort of pathway activity score by utilizing network topology and differentially expressed gene set information. Pathway interactions and gene information are integrated by applying different probabilistic approaches. However, the dependence on the given differentially expressed gene list also exists in the initialization step of some NSBA methods, since the genes not having high expression levels might have more interesting biological functions in the work flow of a pathway topology. Additionally, none of the recent methods has managed to derive a quantitative measure for assessing biological activities of specific cellular processes in a pathway.

1.4 Contributions

Current approaches integrating large scale data and information from a variety of sources to reveal molecular basis of cellular events do not adequately benefit from pathway information. Here, we portray a network structure based pathway enrichment system that fuses and exploits model and data: signalling pathways are taken as the biological models while microarray and ChIP-seq data are the sample input data sources among many other alternatives. Our model- and data-driven hybrid system allows to quantitatively assess biological activity of a specific cellular process simultaneously identifying significant paths leading to the process. The fundamental constituent of network structure based pathway enrichment system is the *Signal Transduction Score Flow (SiTSFlow)* Algorithm. We first convert a signalling pathway into a cascaded graph structure and then map the individual gene scores onto the nodes. The gene scores are transferred over the nodes by traversing the path until a pre-defined target biological process is attained. The score flow simulates signal transduction inside the cell. Because of cyclic pathways, we carry out iterations and when the scores converge, a final activity score is assigned to the pre-defined target biological process. The final activity score provides a quantitative measure to assess the biological significance of a process under the given experimental conditions. Transcriptome data is integrated by taking the rank products of individual scores of the employed data sources.

Our hybrid system based on pathways and transcriptome data is a novel approach to quantitatively evaluate biological activities of cyclic signalling pathways as well. SiTSFlow algorithm shows convergence behavior for biological cyclic graphs. Several gene knockout operations have been performed on a manually curated pathway. In order to observe the effects of gene knockout operations on the final activity scores of processes, SiTSFlow algorithm was run on the new knockout pathways as well.

In order to explore various biological responses of several processes in a global network, the selected small signalling pathways have been merged based on their common nodes e.g., genes and processes. As the result of iterative unification operations, a global signalling network for the human cell was constructed and it has been assessed by using SiTSFlow algorithm. In order to test the statistical significance and sensitivity of each final activity score, several permutation tests are designed and performed. Furthermore, SiTSFlow algorithm was implemented as Cytoscape plug-in to interactively visualize pathways and perform systematic analysis in a well known environment.

The main contributions of this thesis are as follows:

- Development of a network structure based pathway enrichment system incorporating pathway topological information and transcriptome data;
- Development of a signal transduction score flow algorithm to assess biological activity of a process in a signalling pathway;
- Merge of several signalling pathways to effectively analyze the biological activities in a global signalling network of the human cell;
- A visualization and analysis tool including signal transduction score flow algorithm in Cytoscape environment.

From the computer science perspective, we have achieved several contributions. Instead of identifying of cycles in a pathway, we convert each pathway into a cascaded or levelized graph form by using a linear time multiple source Breadth First Search Algorithm. The time complexity of a cycle identification algorithm is higher than the linear time levelization algorithm, thus we do not aim to detect cycles in a pathway. The proposed SiTSFlow algorithm has also linear time complexity, therefore it is very suitable to run on pathways of bigger sizes having more than 1000 nodes. In our experiments, unification of small size pathways results in a broader global network composed of 450 nodes and 650 edges. Thus, application of the proposed network structure based pathway enrichment system on a global signalling network has been successfully performed.

1.5 Organization of the Thesis

We present a brief introduction to analysis of transcriptome data and biological networks, and importance of pathway enrichment methods in the analysis of large scale data in this chapter. Chapter 2 gives basic computational and biological background and literature information. Information about biological pathways, analysis methods for microarray and ChIP-seq technologies, and computational approaches for graph models reported in literature are discussed in Chapter 2. Chapter 3 describes the details of the proposed network structure based pathway enrichment system and it corresponds three papers published during the course of this study [3, 4, 5]. Data processing steps and details of SiTSFlow algorithm are explained in Chapter 3. Chapter 4 provides experimental results of proposed system on several data sets. The results are discussed from both biological and computational perspectives. Chapter 5 corresponds to graph merge algorithm for unification of individual pathways. Constructed global network is analyzed by using several data sets. Biological results obtained in this global network are discussed in more detail. Chapter 6 concludes the thesis and gives some future directions for pathway assessment and enrichment procedures.

CHAPTER 2

BACKGROUND INFORMATION ON BIOLOGICAL AND COMPUTATIONAL ASPECTS

In this chapter, we first present biological pathways that constitute models of the proposed system. Characteristics of large scale transcriptome data is then explained. Several computational analysis methods and tools for transcriptome data are given in detail. Finally, graph models and algorithms are discussed from the computer science perspective.

2.1 Biological Pathways

There are several types of biological networks. Protein-protein interaction networks represent interconnection between the proteins during the biological working mechanism of the cell. On the other hand, pathways are the abstract representations of gene interactions and chemical reactions within the cell. Hence, pathways deal with molecular and signalling levels of working mechanism of cellular processes. Signalling pathway is a special type of pathways that captures functional relationships between the genes, chemical compounds and biological activities. They are usually represented by directed graphs. Nodes of the graph represent a gene, gene product, chemical compound, small molecule or biological activity. Edges represent functional relations between the nodes. There are three conceptual types of edges: activation, inhibition, and neutral. The rest of relations on edges might be transformed into one of these main relation types.

There are several publicly available biological pathway resources. Kyoto Encyclopedia of Genes and Genomes (KEGG) is the pioneer study for online pathway databases [6]. It contains the collection of manually drawn pathways which represent the knowledge on interaction and reaction networks. The major focus of KEGG database is for yeast, mouse, and human metabolic and signalling pathways. Reactome is another open access, manually curated, peer-reviewed pathway database containing cell metabolic and signalling pathways [7]. It contains pathways for 22 species including human, rat, and mouse. Pathway representation in Reactome is based on the reaction definition that describes many biological events for example, binding, activation, and degradation. Information in the database can only be modified by expert biologist researchers. The are other commercial databases containing cell signalling pathways, such as BioCarta [8], Ingenuity Pathways Knowledge Base [9], Ariadne ResNet [10]. Most of the databases provide download facility for pathway graph - relation data. Although each pathway database uses its own data format, there are some common formats e.g., SBML, BioPAX that become widespread as the data standard in the pathway databases.

For biological analysis, various computational approaches are applied to explore system behavior in complex networks. Global properties of the pathways are identified by performing topological analysis of the network. Functional units, such as hub, minimal cut, loop, or motif are predicted by applying classical graph theory approaches, since these units determine the global behavior of a network. On the other hand, local and more specific behaviors of the system are identified by performing dynamical analysis of the pathways. Dynamical analysis requires complex reaction parameters, initial conditions, and differential equations as contrast to topological analysis [11], therefore it is applicable only on small-sized networks. Application of dynamical analysis to large pathways is computationally very expensive and inefficient.

2.2 Microarray Technology and Analysis

Microarray technology is based on nucleic acid hybridization method that provides information about which genes are active in a tissue under certain experimental conditions. In a gene expression profiling experiment, expression levels of thousands of genes are monitored to explore the effects of a specific treatment or disease on gene expression. Gene expression profiling can be applied to identify genes whose expression has changed by a high amount between for example in a cancer tissue by comparing its gene expression level with a normal tissue.

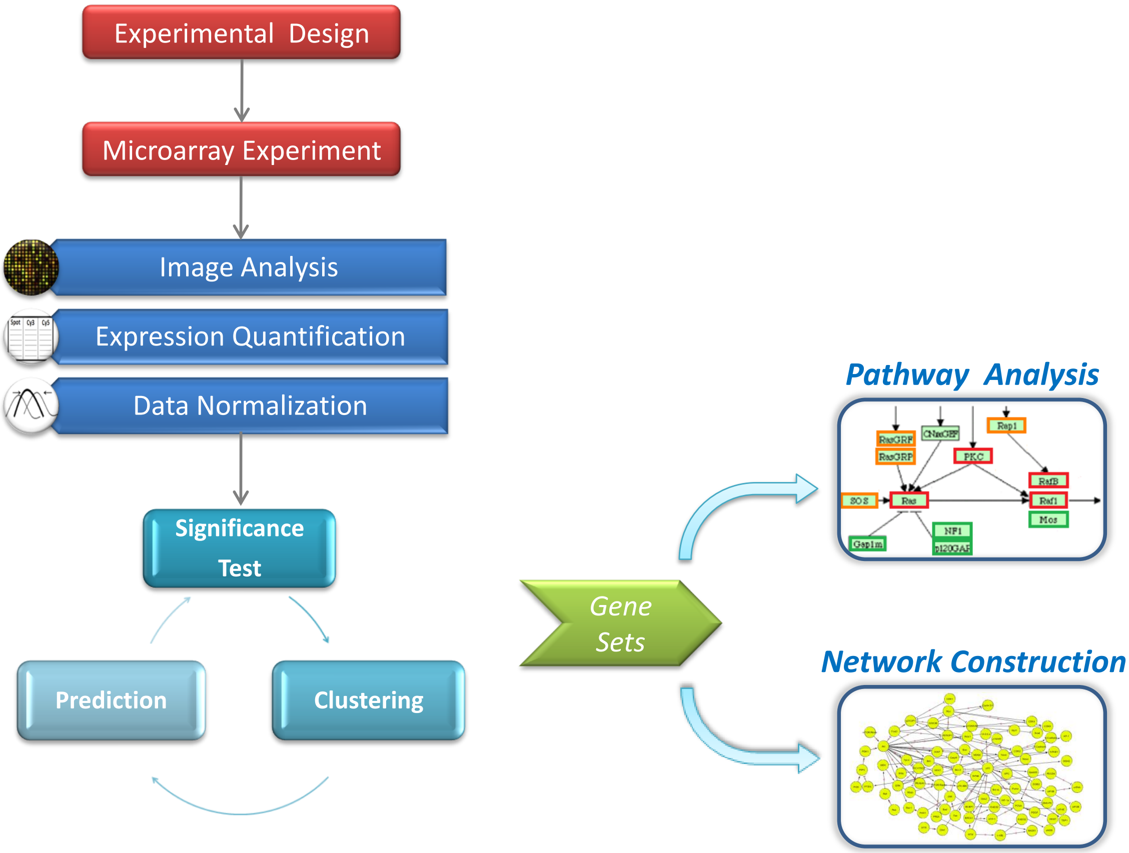


Figure 2.1: Microarray analysis steps. Experimental design is performed based on the biological question. RNA extraction, RNA labeling and hybridization steps are then performed in laboratory environment. Image analysis, quantification of gene expression and normalization are the main steps of data pre-processing. Application of significance tests, clustering and prediction are performed in computational analysis step. Finally, a differentially expressed gene set is constructed. Pathway analysis could be applied to explore the biological function of this gene set. Alternatively, network construction might be performed based on this gene set.

Experimental design is the first step of a microarray experiment as shown in Figure 2.1. Every experiment considers the biological question asked in the design step. Therefore, the setup of each experiment is specific to its conditions e.g., ribonucleic acid (RNA) samples, replicate number, cell type etc. After properly completing of microarray experiment in laboratory environment, the pre-processing step of microarray data analysis is applied. Scanned image containing many colored spots is first processed and converted into a raw data. This raw data is quantified by applying spot discrimination and summarization processes. Raw numerical data is normalized to remove channel variability and array heterogeneity. Then, statistical analysis steps can be applied on the normalized array data.

2.2.1 Analysis Methods

Normalized gene expression data can be analyzed in several ways to extract useful biological information. Much of the analysis research has focused on identification of differentially expressed genes or a gene set sharing similar expression profile.

Statistical analysis methods, such as *t*-test, Fisher-exact test, analysis of variance (ANOVA), False Discovery Rate (FDR) are applied to generate differentially expressed genes by comparing two or more samples. Threshold parameters to select significant gene sets have been set to very strict values e.g., 0.001, eventually, these methods apply a kind of *over-representation analysis* for microarray data. Alternatively, several clustering methods e.g., hierarchical clustering, *k*-means clustering, or Self Organizing Maps (SOM) can be applied for extracting expression patterns across samples. The genes in the same cluster may not be differentially expressed genes, however identification of genes representing similar expression behaviors under the same experimental conditions might be more challenging, since these set of genes might be functioning in a specific biological pathway. Analysis result obtained from either a clustering or a significance test approach is the *significant gene set*. The gene set can form a basis for the network reconstruction or pathway analysis.

The main difficulty in microarray analysis is the biological interpretation of a significant gene list. Understanding functions of individual genes on a list of significant genes is difficult especially when it is done by a human expert. Functionally related genes in a ranked list (ordered by gene expression levels) may not be located on top or bottom of the list.

For instance, a cluster contains genes with similar expression profiles, all genes in this cluster may not have a function on the same biological pathway. Therefore, recent efforts focus on the discovery of biological pathways rather than individual gene function [12]. Even though only small numbers of differentially expressed genes appear in a pathway, they would be associated with a specific biological event that could related with the conditions of the microarray experiment. Therefore the output of a microarray experiment is utilized either in pathway enrichment analysis or network reconstruction process.

2.2.2 Pathway Enrichment

Generally, enrichment tools aim to provide interpretation of the gene expression profiles by either identifying major genes or pathways based on traditional statistical tests or allowing visualization of gene expression data on molecular pathways. Pathway enrichment methods can be classified into three categories:

1. Significant Gene Set Analysis
2. Gene Set Enrichment Analysis
3. Network Structure Based Analysis

Significant Gene Set Analysis (SGSA) is the first attempt for gene enrichment. SGSA takes differentially expressed genes as input and then iteratively checks the existence of significant genes in function annotation databases. *p*-value of enrichment analysis is computed by known statistical methods e.g., Chi-square, Fisher's exact test, Hypergeometric distribution etc. *p*-value for enrichment analysis represents the number of genes in input list that match known function annotations as compared to random assignment of function annotations. Onto-Express [13], GoMiner [14], EASE [15], and FatiGO [16] apply SGSA to associate initially identified differentially expressed genes to known functional terms. There are two main drawbacks of SGSA methods: the first one is that the output enriched annotations highly depend on initially given differentially expressed gene list. The method for the identification of differentially expressed gene set and cutoff threshold settings highly affect the result of enrichment analysis. The second drawback is providing large amount of output annotations that should be post-processed by a human expert to find out exact answers for the asked questions in the experiment.

Gene Set Enrichment Analysis (GSEA) method determines if the members of an input gene list exist in predefined pathway gene sets. Input gene list is ordered by a measure of expression i.e., fold-change measure or p-value of t-test. If the input gene exists in the pathway of interest, the enrichment score (ES) is increased; if the gene does not exist in the pathway, the score is decreased. Finally, if the input gene set was correlated with a biological process or pathway, the input genes appear usually in the top (or bottom) of the pre-defined pathway gene set and ES will be very high. In order to compute p-values for the ES for a pathway, the input gene list is shuffled randomly and ES is calculated thousands of times. The main benefit of GSEA tools is the usage of all genes from a microarray experiment without applying a gene selection method and cutoff threshold. Thus even the genes not having very significant expression changes may provide a contribution for the enrichment analysis. Some example tools adopting this strategy are GSEA [17], FatiScan [18], PAGE [19], and Go-Mapper [20]. However, ES highly depends on the ordering of the input gene list and if the order of this gene list is reversed, the same enrichment may not be observed towards the bottom of the list. Many genes placed at the top of the ordered input list may not exist in a pathway, therefore ES gets very low value. The genes having high or low expression measures highly affect ES calculation, since ES is highly affected by the extreme points i.e., up or down regulated genes of the ordered list. However, the genes in the middle of the ordered list may have more interesting biological functions in the cell environment. Both SGSA and GSEA methods do not consider genes in the middle of the gene list. Furthermore, over-representation analysis limits the accurate identification of perturbed pathways for a specific experiment, since these methods could not incorporate relations of known gene interactions in a pathway.

Network Structure Based Analysis (NSBA) methods have been developed to estimate the effects of specific experimental perturbations i.e., gene expression changes on the biological process of pathways. In this context, there are a couple of methods to compute a sort of pathway activity score by utilizing network topology and differentially expressed gene set information. Signalling Pathway Impact Analysis (SPIA) method was developed by Tarca et al. to estimate the impact of experimental perturbations on pathways [21]. Biological perturbation is the alteration of gene or pathway function by applying various environmental changes. SPIA firstly computes the over-representation of selected genes in a pathway, then identifies the perturbation amount of that pathway by forwarding gene expression changes through network topology. The method combines these two inputs into one global probability value, P_G, which provides a measure to rank pathways based on their perturbation amounts. Signalling pathways can be used to interpret phenotype descriptions of complex diseases. Efroni et al. performed oncogenic phenotype prediction by incorporating gene expression and network topological information [22]. Their method computes a pathway activity score by taking average likelihood of the pathway's individual interactions that are activated at given gene states. Lee et al. developed a new classification method based on identification of pathway activities by using gene expression samples of each patient [23]. For each pathway, an activity level is computed from the gene expression levels of specific conditions, this pathway activity score is then used to build classifiers for predicting the disease phenotypes. Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) was developed by Vaske et al. to indicate patient-specific gene activities by integrating pathway and gene information [24]. A gene is modeled by a probabilistic graph model i.e., factor graph which uses set of interconnected variables to represent expression, activity, and products of the gene. PARADIGM aims to identify which pathway activities are changed in a patient by applying a probabilistic inference. The problem of dependence of the results to the given differentially expressed gene list still exists in the initialization step of some NSBA methods. The genes that are not differentially expressed may also have interesting biological functions related with critical pathways. Hence, the entire gene information extracted from transcriptome data should be incorporated with topological pathway information. Additionally, none of the recent methods has managed to derive a quantitative measure for assessing biological activities of specific cellular processes that are specific to a disease or treatment applied in the experiment.

2.2.3 Network Construction

Construction of a biological network requires to learn network structure *de novo* from the expression values of the genes. The approaches used to construct networks include Boolean networks [25, 26], Bayesian networks [27, 28], and differential equation models [29, 30]. Boolean network approach constructs an abstract gene network in which a gene state is set to either 0 or 1. Bayesian network provides a graph structure based on conditional probabilities of genes given in microarray data. Differential equation model creates a gene network by computing a set of differential equations considering the gene rate changes. By applying these approaches, the global properties of a biological network are predicted. However the prediction process of huge networks is computationally very expensive. Additionally, the quality of constructed network is highly dependent on the quality, experimental design and noise of microarray data. So, the predicted gene network may contain incorrect gene regulations. The use of several biological data sources e.g., protein-protein interactions, sequences of the binding site of the genes, literature etc. empowers *de novo* prediction quality of the proposed approaches. However, the main objective of this thesis is not developing a *de novo* network construction. The proposed system aims to assess existing biological pathways to provide an easier interpretation method for the biological pathways under the effect of experimental conditions formed by transcriptome data.

2.3 ChIP-Sequencing Data and Analysis

2.3.1 ChIP Technologies

DNA binding factors e.g., histones and transcription factors and their associated cofactors e.g., coactivators and corepressors are the dynamic regulators responsible for utilizing genomic information by controlling the transcriptional gene regulation. However, we are still missing genome-wide mapping of their binding sites. A global binding map would allow us to determine which, when, and how genes might be regulated by these factors at a genomic scale. Chromatin Immunoprecipitation (ChIP) technology is applied to identify whether proteins e.g., transcription factors are associated with a specific genomic region of a living cell or tissue. ChIP followed by either genome tiling array analysis (ChIP-chip) or massively parallel sequencing (ChIP-seq) enables transcriptional regulation to be studied on a genome-wide scale.

ChIP-chip provides genome-wide localization analysis of DNA binding factors, cofactors, and histone marks. The experimental design steps of ChIP-chip and ChIP-seq technologies are shown in Figure 2.2. ChIP-chip integrates specific immunoprecipitation of genomic DNA fragments that are associated with specific proteins or histone marks (ChIP) and DNA microarray analysis (chip). Specific chromatin fragments are isolated using antibodies specific to a feature of interest. Then isolated fragments are amplified to produce fluorescently labelled DNA. After performing hybridization to DNA microarrays, the microarray probes are mapped to the genome to produce genomic coordinates. However, there are several technical

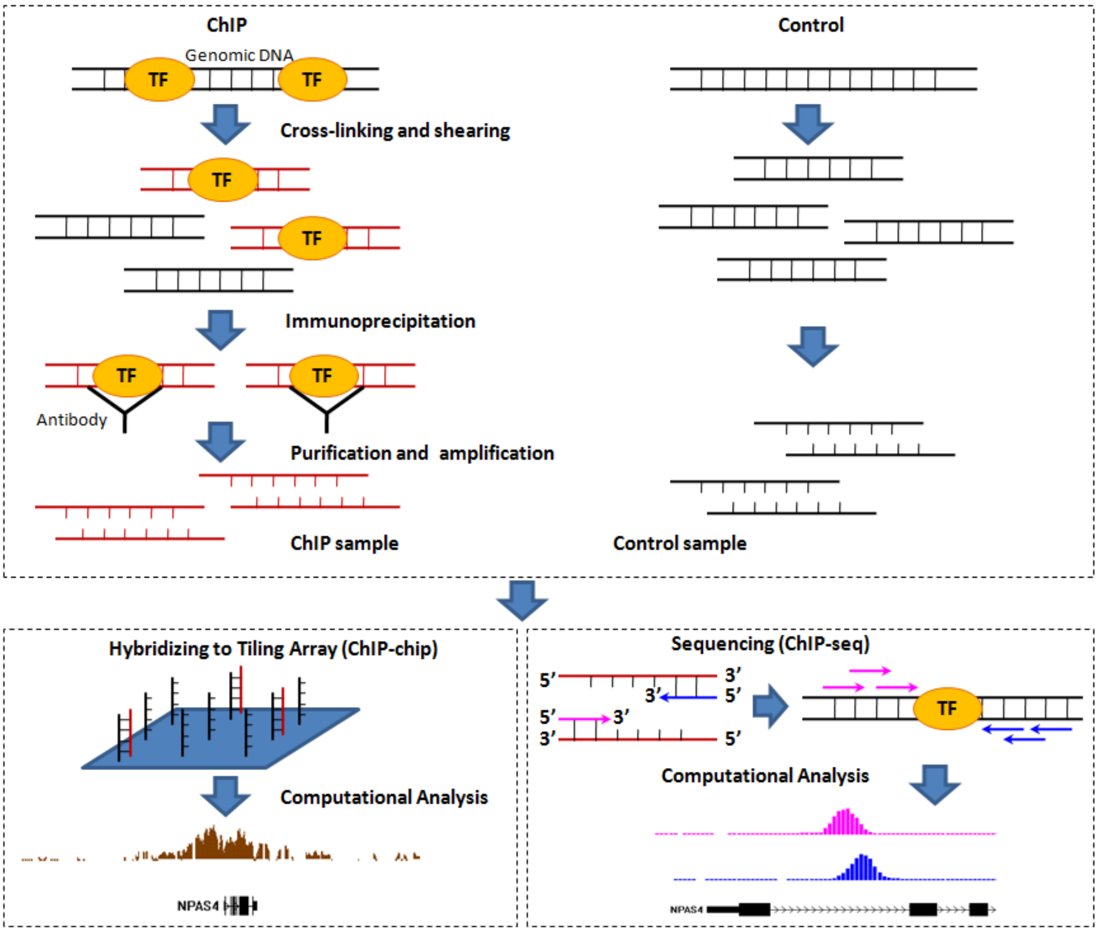


Figure 2.2: Comparison of ChIP-chip and ChIP-seq technologies. A ChIP sample is prepared to represent genome samples for protein-DNA bindings. In ChIP-chip technology, the ChIP sample is hybridized to a microarray that investigates entire genome by using probes. In ChIP-seq technology, the ChIP sample is sequenced from both ends to construct millions of short reads using massively parallel sequencing. Control samples are used to remove data biases in computational analysis (adapted from Ji et al. [1]).

challenges related with whole genome ChIP-chip analysis. These are potential bias introduced by a global polymerase chain reaction (PCR) amplification step, low resolution and low sensitivity, high input material requirements in most approaches, uninformative results on repetitive sequences, and expensive microarrays. The main reasons for limited usage of ChIP-chip technique by researchers are the complexity of raw data (thousands of cofactors, histone marks) and variety of cell types and cellular conditions.

On the other hand, in ChIP-seq technology the conventional ChIP assays are combined with the Illumina Genome Analyzer using massively parallel Solexa DNA sequencing technology. This technology permits high resolution, highly sensitive, and less expensive genome-wide mapping of protein-DNA associations as shown in Figure 2.2. Therefore, ChIP-chip is replaced by ChIP-seq in genomic scale discovery of transcription factor binding sites. ChIP technique permits a library of target DNA binding sites of given transcription factor. Solexa Sequencing identifies isolated DNA sites from ChIP. This massively parallel sequence analysis provides analysis of interaction pattern of any protein with DNA. The Illumina Genome Analyzer identifies the sequences of ChIP-isolated DNA fragments to mark and quantify the sites bound by a protein of interest.

2.3.2 Raw Data Processing

Mining information from the huge data sets generated by these high-throughput technologies is a very complex task. Computational analysis steps of a ChIP-chip experiment are data exploration, normalization, binding region detection, providing gene annotation, and finding enriched sequence motifs. In the past few years, a number of tools performing each step have been developed. We briefly explain some of these tools in the following paragraph.

Quantile normalization is widely applied in the tiling array analysis [31]. MA2C which is a model-based normalization approach based on the guanine-cytosine content of probes, is developed for two-color tiling arrays [32]. Tilescope is a web-based data processing software to analyze tiling arrays [33]. The approaches for detecting binding regions using normalized array data are hidden Markov models [34, 35], moving windows based methods [36, 37], hierarchical mixture models [38], regression and kernel deconvolution methods [39, 40, 41]. Ringo is a R-Bioconductor package for ChIP-chip analysis [42]. The popular motif discovery tools are MEME [43] and Gibbs Motif Sampler [44]. The tools Galaxy and CEAS have been developed to retrieve gene annotations [45, 46].

The major analysis steps of a ChIP-seq experiment are aligning reads to the reference genome and finding read enriched regions. The locations containing high number of DNA fragment reads are called as *peak* or *read enriched* regions. The predicted peak regions are used for motif discovery and annotation retrieval analysis. ELAND software was developed to align millions of reads to the reference genome allowing up to two errors per match [47]. SeqMap, fast sequence mapping software, is developed for ChIP-seq read mapping [48]. It is the first algorithm allowing insertion or deletion detection. In order to estimate FDR for one or two-sample ChIP-seq data, a Poisson model is used. In the post-processing step, it uses the advantage of the separation between the forward strand and reverse strand reads to refine binding region boundaries. By applying boundary refinement step, it can greatly improve the resolution of binding region detection. Recently, other (RMAP, SOAP, ZOOM) read alignment tools have also been developed to align reads generated by ChIP-seq [49, 50, 51].

2.3.3 Analysis Tools

Recently, in order to detect peak regions, comprehensive tools with easier user interface have been developed: GeneTrack [52], QuEST [53], SISSRs [54], and CisGenome [1]. Regions having high sequencing read density are called as *peaks* in ChIP-seq data. Given the aligned reads as input, the pioneer ChIP-seq analysis tools used their own analysis pipelines to detect DNA-binding regions.

GeneTrack applies a Gaussian smoothing procedure to represent signals with a continuous curve across the genome, a peak region is then identified by finding maximum point of the curve. GeneTrack tool does not compute a false discovery rate (FDR) estimation. QuEST uses a Gaussian kernel density estimation approach to identify DNA-binding regions. It generates peaks by utilizing of main attributes of the data, such as directionality of reads and size of fragments. By comparing original and negative control samples, QuEST computes FDR estimation. One drawback of QuEST tool is that it does not convert peak region scores into p-values. SISSRs uses the direction of reads to estimate the average length of DNA fragments. It combines the fragment length, read directionality, and background model to bound the binding sites within tens of base pairs. In the case of only a ChIP-seq sample is available, the method uses a Poisson model to estimate FDR.

CisGenome is designed to provide all essential needs of ChIP data analysis: visualization, data normalization, peak detection, false discovery rate computation, gene-peak association, and motif analysis [1]. It is a standalone system that biologists can use to analyze their own data on their personal computers. CisGenome incorporated a new version of TileMap [35] as the internal ChIP-chip peak caller. Motif discovery Gibbs motif sampler is provided for *de novo* motif discovery [44]. CisModule is provided for novel cis-regulatory module discovery [55]. Given a genome and a list of binding regions, CisGenome provides a function to generate matched genomic control regions.

Computational analysis of raw ChIP-seq data sets employed in this thesis is performed by using CisGenome software. Therefore we give more detail about the analysis steps of ChIP-seq data in CisGenome software. The analysis starts by providing the DNA-reads as the input to the software. CisGenome accepts mapped reads of SeqMap. Peak detection unit identifies the peak regions having sufficient DNA-binding reads with small FDRs. Therefore, the genome is divided into non-overlapping windows with length w (i.e., 100 base pairs (bp)) for FDR computation from a only one ChIP sample. The number of reads (n_i) within each window i is counted. Poisson model is used to model binding regions. The background model for read counts is modeled by negative binomial distribution. For this purpose, negative binomial distribution is fitted to the number of windows with a small number of reads (two or fewer). Then the estimated null distribution is used to compute the FDR estimates for each level of n_i. Observed n_i is compared with the expected read counts estimated by the null model, then the ratio between the two count is reported as FDR estimate. When two samples exist (negative control and ChIP sample) for an experiment, the genome is divided into non-overlapping windows length of w. For each window i, the number of reads extracted from the ChIP sample (s_{1i}), the number of reads in the control sample (s_{2i}), and the total read number ($n_i = s_{1i} + s_{2i}$) are computed. The expected sampling ratio between the ChIP and the negative sample for non-binding regions is estimated by using the windows containing small number of reads: $r_0 = \sum s_{1i} / \sum s_{2i}$. Then the windows are grouped according to n_i. For each group ($n = 0, 1, 2, , \ldots ,$), the observed distribution of s_{1i} is compared to its expectation model by Binomial distribution ($n, p_0 = r_0/(1 + r_0)$). Finally, FDR value is computed by using s_{1i}/n_i ratio. Final step of peak detection operation is scanning of entire genome with a sliding window of width w to detect all windows with FDR smaller than a cutoff. If there exists overlapping windows, they are merged into one region. Figure 2.3 shows an example of a peak region at human chromosome 14 between 23700178 and 23700299 genome positions which is identified from STAT1 data set [2] by using CisGenome software. The genes ISGF3G and RNF31 are in TSS-upstream and TES-downstream of this peak region.

The significant peak regions with small FDRs might be used in the gene-peak association or motif analysis. In this thesis, gene-peak association unit is applied to identify *neighboring genes* of significant peak regions. By using gene-peak association unit, the neighboring location e.g., 10000 bp to both upstream and downstream sides of each significant peak region is scanned. If there exists a gene in this distance range, it is marked as a neighboring gene. The constructed set of such genes is reported as the output of the computational analysis of ChIP-seq data.

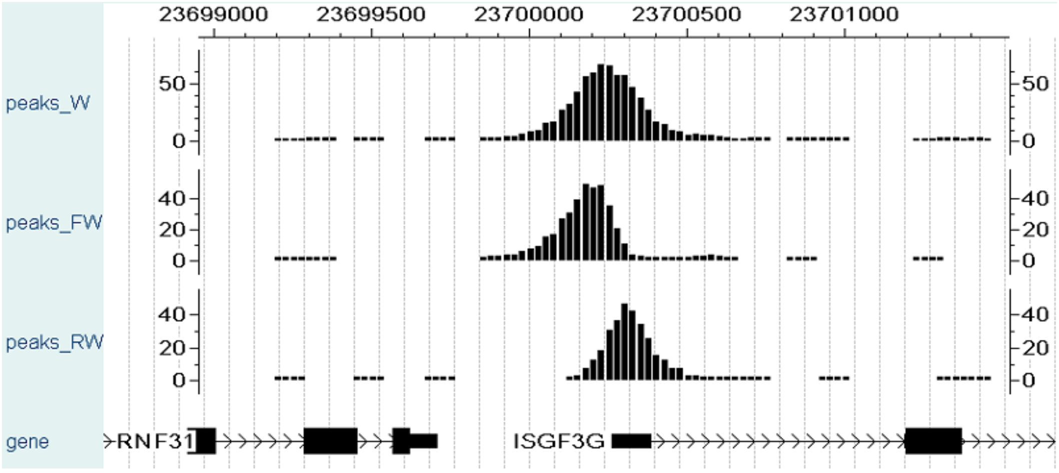


Figure 2.3: An example of peak region at chromosome 14 between 23700178 and 23700299 positions. The genes ISGF3G and RNF31 are in TSS-upstream and TES-downstream of this peak region. ChIP-seq data is STAT1 data set [2].

2.4 Graph Models and Computational Approaches

Biological pathway is one type of graph model representing cellular events by an abstract form. Various graph modeling and analysis approaches originated from classical graph theory have been applied for *de novo* discovery or modeling of pathways. In the context of this thesis, a pathway constitutes the fundamental structure for the underlying biological events. Nodes and edges of a pathway mainly determine dynamic elements of the event. The flow direction and attribute of an event is provided by directed and signed edges. Due to nature of biological phenomena some pathways may contain cycles in the graph. Therefore, a biological pathway used in this thesis is represented by a directed cyclic graph model. Here, we give a general graph notation and then discuss graph traversal approaches and explicit cycle identification methods.

2.4.1 Notation

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a structured model which is composed of set of nodes $\mathcal{V}$ and edges $\mathcal{E}$. The total number of nodes and edges are represented by n and e, respectively. In a biological graph, a node represents gene, gene product, chemical compound, small molecule or biological activity. An edge represents functional relations between the nodes. If an edge $e = (x, y)$ is directed from node x to node y; then x and y are called the tail and head of the edge, respectively. The graph composed of directed edges is called *directed graph*. A *cyclic graph* contains one or more cycles, meaning that some of nodes are connected in a closed chain. A *directed acyclic graph* (DAG) is a type of directed graph without any cycles. There are three conceptual types of edges: activation (+), inhibition (-), and neutral. Other complex biological relations on edges might be transformed into one of these main relation types. Edges are associated with a weight or number in a weighted graph. Edges are represented by various data structures.

- *Adjacency Matrix* is defined by an *nxn* matrix in which:

$$a_{xy} = \begin{cases} 1 & \text{if there is a directed edge from } x \text{ to } y \\ 0 & \text{otherwise} \end{cases}$$

If we have a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{S})$:

$$s_{xy} = \begin{cases} s & \text{if there is a directed edge from } x \text{ to } y \\ 0 & \text{otherwise} \end{cases}$$

Adjacency matrix representation uses $O(n^2)$ storage space and operations on matrix generally run in $O(n^2)$ time.

- *Adjacency List* method uses an array $A[1 \dots n]$ to keep pointers to lists of adjacent nodes. For example, $A[t]$ points to the list of nodes adjacent to t. This type of storage provides flexibility and requires $O(n + e)$ space.

In this thesis, due to space and time efficiency *Adjacency List* notation is applied for node-edge relation representation. More specifically, we define $outAdj(x)$ to denote the out-adjacency list of a node x, that is $outAdj(x) = \{y : (x, y) \in \mathcal{E}\}$. Similarly, $inAdj(x)$ denotes the in-adjacency list of node x, that is $inAdj(x) = \{y : (y, x) \in \mathcal{E}\}$.

2.4.2 Graph Traversal

Traversal algorithms identify the *visiting* order of each node in a graph. Algorithm starts from a root node and then visits all nodes which are reachable from the root node. The visiting order information might be used as input of other graph search problems, e.g., topological labeling, graph connectivity, shortest path, cycle identification etc.

2.4.2.1 Depth First Search

Depth First Search (DFS) is one of the basic recursive traversal algorithms. DFS can be applied on acyclic graphs and trees. Initially all nodes are marked as *unvisited*. DFS visits the neighbors of a selected node recursively, then it continues the selection of new neighbors in deeper direction of the graph until reaching a node has not any unvisited neighbor. This recursive algorithm records a visiting path from the root to the currently processed node. The time complexity of DFS is linear in terms of total node and edge numbers. A call to DFS is made exactly once for each newly visited vertex, DFS is called $O(n)$ times. For a call of each vertex v, the number of operations executed is equal to the number of edges incident on and it is the length of $Adjacency(v)$. So, the for-loop runs in totally $O(e)$. Therefore final time complexity of DFS becomes $O(n + e)$.

Algorithm 1 : $DFS(v)$

```
visited($v$)
pre-visited($v$)
for each node $x \in Adjacency(v)$ do
  if NOT visited($x$) then
    parent[$x$] = $v$
    $DFS(x)$
post-visited($v$)
```

2.4.2.2 Breadth First Search

Another traversal algorithm is Breadth First Search (BFS) which starts with an unvisited node u and spans to its children, i.e., first visiting u, then all children of u, then the children of those children and so on. The difference of DFS and BFS algorithm is based on the selection of next neighbor of a visited node. DFS visits neighbors recursively, it burrows deeper into of selected neighbor node until reaching a goal node. However, BFS visits each node in order of their breadth, another words it broadens visited nodes of a graph. Generally, BFS algorithm uses queue data structure to store traversal information. The traversal information obtained with BFS algorithm may be useful in solving of problems based on the distance between specific nodes, e.g., path-length finding.

When queue data structure is used in BFS algorithm, each node in graph is enqueued and dequeued at once. Each queue operation takes $O(1)$ and the total time spent for queue operations in while-loop takes $O(n)$. The for-loop is run $degree(u)$ times for every node v, so the total time spent for scanning adjacency lists is $O(e)$. The total time complexity of BFS becomes $O(n + e)$.

In the scope of this thesis, classic BFS algorithm is modified to identify visiting order of a graph. The multiple source BFS algorithm gives a level number to each node by propagating visiting orders of nodes starting from initial nodes. By using this level information, a pathway is transformed into well defined cascade graph.

Algorithm 2 : $BFS(u, \mathcal{G})$

```
unvisited($u \in \mathcal{G}$)
Create an empty queue $Q$
$ENQUEUE(Q, s)$
while $Q \neq \emptyset$ do
  $u = DEQUEUE(Q)$
  for each node $v \in Adjacency(u)$ do
    if NOT visited($v$) AND $v \notin Q$ then
      $visited(v)$
      $ENQUEUE(Q, v)$
```

2.4.3 Cycle Detection

General graph traversal algorithms, DFS and BFS, can run on acyclic graphs and trees. However, there may be cycles on biologic signal transduction pathways. The detection of cycles in a graph is the essential work to apply a graph-based search or scoring algorithm on that graph.

The simplest way for detecting cycles in a directed graph is to modify the classic DFS algorithm. The basic idea of this modification is to apply a node coloring scheme that provides identification of back edges. If there exists an edge for which a node is visited a second time before all of its neighbors have been visited; that edge is called as a back edge and this graph must contain a cycle. For this purpose, initially all nodes in graph are marked white. A node is marked grey while its neighbors are being explored. If a node with grey color is encountered, there is a back edge in the graph. We mark nodes with black color when its all neighbors are completely examined.

The pseudo-code of *ModifiedDFS* algorithm is given in Algorithm 3. Existence of back edges is checked in *DFS* function. *CycleDetect* function traverses each vertex of a given graph and uses *DFS* function to check back edges. Therefore, the time complexity of this two-phase algorithm is combination of *CycleDetect* and *DFS* functions. Complexity of *DFS* function is $O(n + e)$ which is the same with the classic DFS algorithm. *CycleDetect* function runs for each node in $\mathcal{G}$, so it iterates $O(n)$ times. Final time complexity of cycle detection algorithm is $O(n^2 + ne)$.

Although, there are several cyclic pathways in our model set, we have not aimed to apply a cycle detection algorithm. We need to provide convergence of node activity scores even if in cyclic graphs. Therefore our ultimate goal is to develop an algorithm which satisfies score convergence criteria for all nodes in a cyclic graph. Besides, time complexity of a basic cycle detection algorithm might be quadratic on the nodes of a graph. Therefore, we applied a graph levelization approach to cyclic pathways. Signal transduction score flow algorithm is iteratively processed all nodes for each level. After running several iterations on entire graph, activity scores of nodes successfully converged. Hence, we have managed to develop a linear time score flow algorithm converging in limited number of iterations on a cyclic graph.

Algorithm 3 : ModifiedDFS

```
function boolean CycleDetect ($\mathcal{G}$)
for each node $v \in \mathcal{G}$ do
    $color(v) = white$
for each node $v \in \mathcal{G}$ do
    if $color(v) == white$ then
        if $DFS(v)$ then
            $return$(TRUE) {Cycle exists}
$return$(FALSE) {No cycle exists}
end function

function boolean DFS ($\mathcal{G}, v$)
$color(v) = grey$
for each node $x \in Adjacency(v)$ do
    if $color(x) == grey$ then
        $return$(TRUE) {back edge detected}
    if $color(x) == white$ then
        $DFS(\mathcal{G}, x)$
$color(v) = black$
$return$(FALSE)
end function
```

CHAPTER 3

NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM

3.1 System Overview

The network structure based pathway enrichment system fuses and exploits biological data and model effectively benefiting from topological information brought in by pathway models. The fundamental constituent of proposed system is the *Signal Transduction Score Flow (SiTS-Flow)* algorithm that is based on flowing of individual gene scores obtained from transcriptome data on the biological pathway models. A pathway is converted into a cascaded graph structure and the individual gene scores are mapped onto the nodes of the graph. Gene scores are transferred to *en route* of the biological pathway to form a final activity score describing biological behavior of a specific process in the pathway. Diagram of our system is shown in Figure 3.1. The proposed system consists of two main phases: *data integration* and *pathway scoring*. Initially, we perform the integration of large scale heterogeneous transcriptome data. Individual score of a gene is obtained by taking products of the rank scores extracted from microarray expression and ChIP-seq data. In pathway scoring phase, signalling pathways selected from KEGG PATHWAY Database or user created networks are used as the models of SiTSFlow algorithm. Each node transmits scores to its child nodes and by traversing the path, this flow continues until a node representing the pre-defined target biological process is met. The output of the algorithm is *final activity score* of a process that provides the identification of significant biological events related with the given input transcriptome data. By this way, user can find out related paths that would respond biological questions enquired at the design stage of microarray and ChIP-seq experiments.

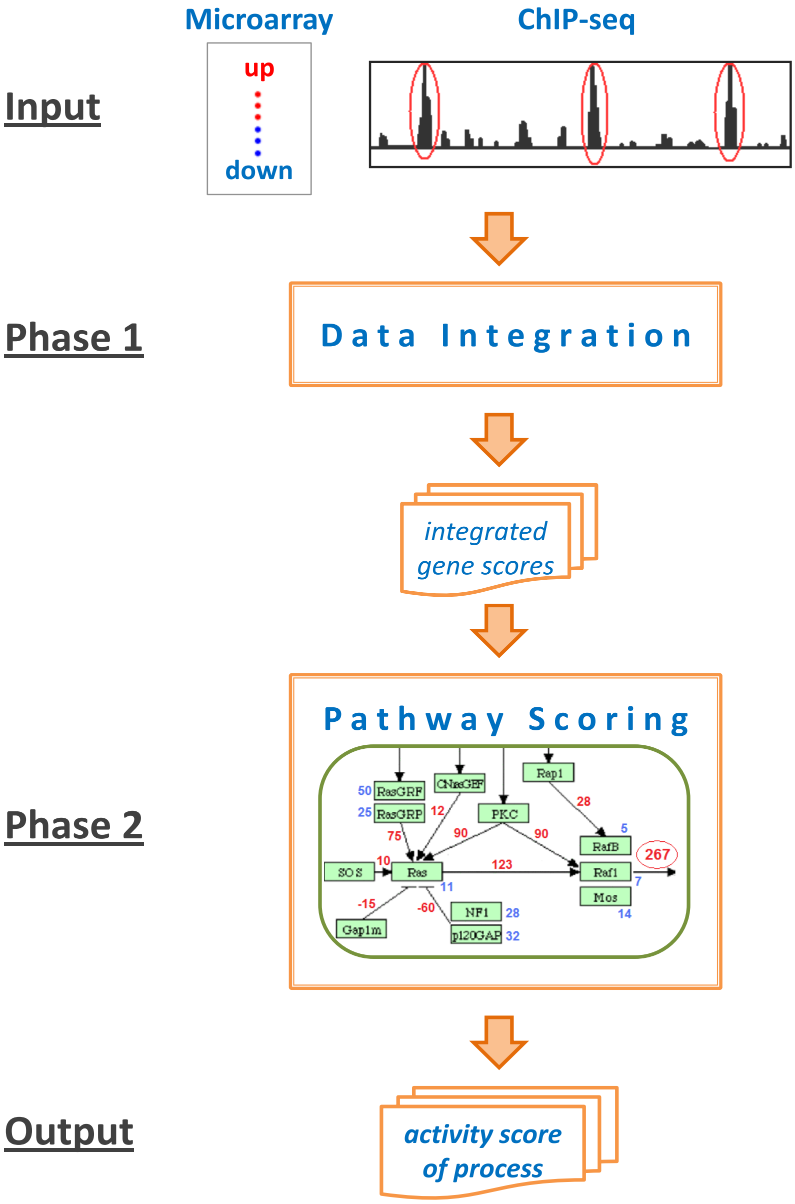


Figure 3.1: Diagram of the proposed system. Transcriptome and ChIP-seq data are combined to obtain integrated scores of genes. In the pathway scoring phase, pathways activated under experimental conditions are identified by exploring scores of each pathway. SiTSFlow algorithm computes the activity score of each process that is represented as output.

SiTSFlow algorithm has an iterative structure, hence it is applicable to cyclic biological pathways as well. For this purpose, original cyclic pathway is converted into cascaded graph topology by applying a linear-time graph cascading algorithm. We perform iterations of the algorithm over the cascaded graph until the convergence of individual node scores. The proposed iterative score computation algorithm has successfully managed to provide convergence of activity scores for every cyclic pathway.

3.2 Data Processing

Data processing is the initial operation before applying other phases of the proposed system. In this section, the pre-processing steps of employed data sources are explained in detail. The proposed system was experimented on three different data sets: HeLa cells under oxidative stress, Estradiol (E2) treated MCF7 cells, and Estrogen Receptor (ER) beta treated U2OS cells. We have applied data processing for each data set. Processing is composed of three basic operations:

1. Peak detection in ChIP-seq data
2. Gene mapping in ChIP-seq data
3. Microarray analysis

Computational analysis of transcriptome data frequently requires using the order rank of scores, such as read count of peak regions in the case of ChIP-seq data and expression value in the case of microarray data. If $r(x)$ indicates the order of the score x when all of the scores are ordered in the ascending order, then rank of x, $R(x)$ is given by

$$R(x) = \frac{r(x)}{TS}, \tag{3.1}$$

where TS is the total number of scores. $R(x)$ score ranges from 0 to 1.

Analysis of ChIP-seq data involves *peak detection* and *gene mapping* operations. For this purpose, CisGenome framework was used to perform these analysis stages on ChIP-seq data [1]. In the first phase of the analysis, we run peak detection method of CisGenome tool to detect the significant peak regions in raw data. Peak detection method essentially searches the entire genome with a sliding window (width=100, slide=25) and determines regions with read counts greater than 10. Our ultimate goal in ChIP-seq analysis is to identify the genes that correspond to neighboring regions of the significant peak regions. This phase of the analysis is called as *gene mapping*. In other words, the distance between transcription start site (TSS) and a peak region is set to ±10000 base pairs, and then the genes within this distance range are marked as *neighboring genes*. $r(x)$ is set to 1 for the gene x, which is located in the neighboring region of the most significant peak region. Hence, $R(x)$ of gene x is very close to 0.

On the other hand, microarray analysis starts by processing of microarray .cel file that contains the image data of gene expressions. A normalization operation is then applied on the raw expression data. The genes might be represented by multiple copies, i.e., probes in microarray chip. Therefore, such copies are unified into one single expression value by taking median of all copies. All these operations are performed on R-Bioconductor environment. Finally, if it is applicable, the expression difference between control and experiment is calculated and this difference value is converted into a rank score by applying Equation 3.1. If a gene x has a high differential expression value, $R(x)$ of this gene becomes very close to 0.

3.3 Data Integration

We initiated the proposed system with the integration of large scale heterogeneous data. Gene scores are calculated by the product of individual ranks extracted from various heterogeneous data sources. Breitling et al. had used this technique to identify genes which were differentially expressed under different conditions [56]. The ranks are assumed to be independent among the experiments. We adapt the rank product method to combine individual ranks of different biological measurements.

$$RP(x) = \prod_{s=1}^{N} R_s(x), \tag{3.2}$$

where $R_s(x)$ is the rank value of gene x coming from the data source s, and N is the total number of heterogeneous data sources. In order to integrate rank scores of genes extracted from individual ChIP-seq and gene expression data set, we apply Equation 3.2 to obtain the product of individual ranks. In Equation 3.2, $R_1(x)$ and $R_2(x)$ represent the individual ranking values of ChIP-seq and microarray experiments for the gene x, respectively. For example, if a gene x has a high differential (up or down) expression value in a microarray experiment and it is significant in the other data sources, $RP(x)$ value will be very close to 0. Therefore, this score can be interpreted as the p-value of gene x by considering individual rankings. $RP(x)$ is normalized to provide better interpretation during the scoring of pathways as given in Equation 3.3.

$$S(x) = (1 - RP(x)) * 100. \tag{3.3}$$

In the rest of the paper, $S(x)$ is referred to as the *self-score* of the gene x.

3.4 Pathway Scoring by SiTSFlow Algorithm

At the *pathway scoring* phase, activity scores for pathways, which control biological processes are computed. For this purpose, we use KEGG pathways as the model to derive cell signalling scoring. A KEGG pathway is converted into a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by using KEGG Markup Language (KGML) files of KEGG PATHWAY Database. A node in the graph represents a gene product, or a target process linking current signal to another KEGG pathway. Edges represent the relations (i.e., activation, inhibition) between the nodes. In $\mathcal{G}$, let $outAdj(x)$ denote the out-adjacency list of node x, that is $outAdj(x) = \{y : (x, y) \in \mathcal{E}\}$. Let $inAdj(x)$ denote the in-adjacency list of node x, that is $inAdj(x) = \{y : (y, x) \in \mathcal{E}\}$. If an edge (x, y) from node x to y is labeled as activation, the total score of node x is then directly transferred to node y. If edge (x, y) is inhibition, the total score of node x is transferred with a negative value to as a score of node y (Figure 3.2). If gene x has no self-score, $S(x)$ is set to zero.

In order to consider processing order of the genes in actual pathway map, the directed graph is converted into a cascade form by applying multiple source Breadth First Search (BFS) algorithm which effectively propagates BFS levels starting from nodes of zero in-degree. Algorithm 5 displays BFS-based cascading algorithm used for this conversion. The crucial point of cascading algorithm is that $\mathcal{G}$ should include at least one node having zero in-degree as a start node. In the initialization phase of Algorithm 5, the nodes having zero in-degree are marked with *BLACK* color and their level are set to 0 and they are enqueued. Other nodes are marked with *WHITE* color. The ordinary BFS algorithm is run in the levelization phase. Every time a node x is dequeued and its neighbors are processed until obtaining an empty queue. For every node y in $outAdj(x)$ with *WHITE* color, its level is stored, it is marked with *BLACK* color and enqueued. The level order information of each node is returned as the output of Algorithm 5.

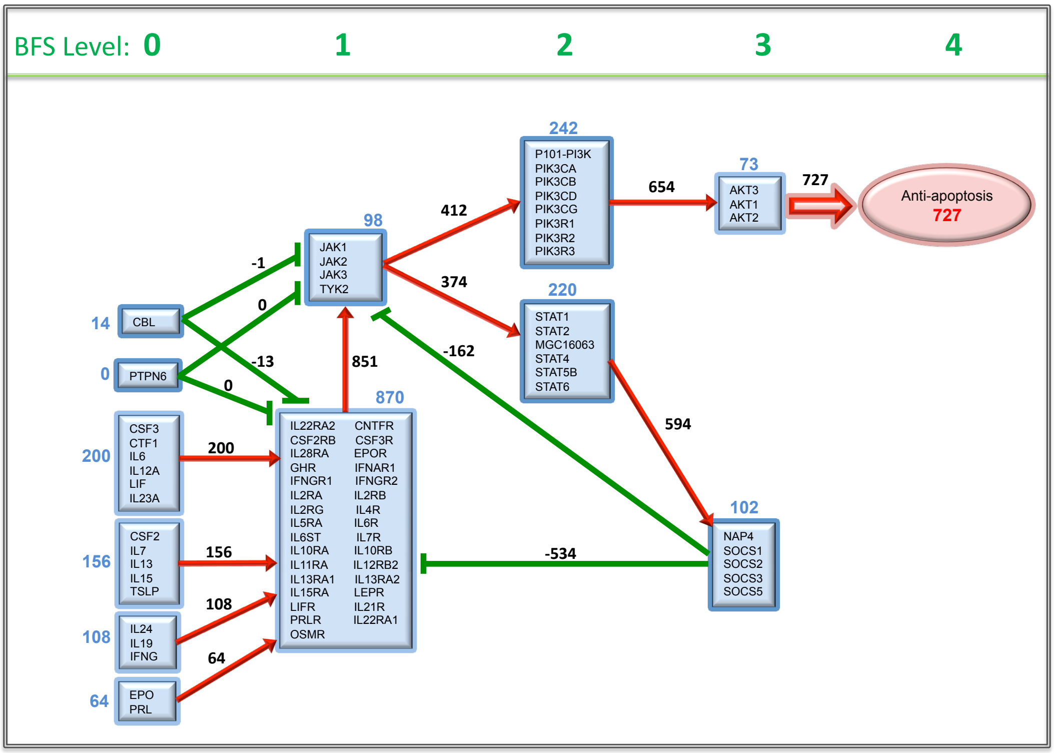


Figure 3.2: Score flow of integrated microarray and ChIP-seq scores to a target process, *Anti-apoptosis*, for HeLa cells under oxidative stress condition. The blue number on each node represents self-score of the gene. Red and green edges represents activation and inhibition properties, respectively. The out-score of a parent node is distributed to all of its children according to the magnitude of their self-scores. An activation edge directly partitions the out-score of the parent between the children nodes. However, a negative score is transferred by the inhibition edges to the children nodes.

Algorithm 4 : Signal Transduction Score Flow Algorithm

```
Input:
Directed graph $\mathcal{G}$ stored in-adjacency and out-adjacency list format
$Score$: indicates self-score of each node calculated by our system
$outScore$: contains out-score of each node
$outAdj(x)$: out-adjacency list of node $x$
$sign$ : keeps edge types: activation (1) or inhibition (-1)
$\mathcal{P} = \{p\}$: set of biological processes
$\mathcal{T}(p)$: set of target nodes representing process $\mathcal{P}$ in $\mathcal{G}$
Levelization information $\mathcal{V}_0, \mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_{L-1}$ obtained by running Algorithm 5.
Initialization:
for each vertex $x \in \mathcal{V}$ do
    $outScore(x) = Score(x)$
    $totOutSelfScore(x) = 0$
    for each vertex $y \in outAdj(x)$ do
        $totOutSelfScore(x) = totOutSelfScore(x) + Score(y)$
Score Computation:
while not converged do
    for each level $\ell = 0, 1, 2, \dots, L-1$ do
        for each vertex $x \in \mathcal{V}_\ell$ do
            for each vertex $y \in outAdj(x)$ do
                $outScore(y) = outScore(y) + sign(x, y) * outScore(x) * \frac{Score(y)}{totOutSelfScore(x)}$
Output:
for each biological process $p \in \mathcal{P}$ do
    $TotalScore(p) = 0$
    for each target node $t \in \mathcal{T}(p)$ do
        $TotalScore(p) = TotalScore(p) + outScore(t)$
return $\{TotalScore(p)\}_{p \in \mathcal{P}}$
```

Algorithm 5 : BFS-based algorithm for cascading graph $\mathcal{G}$

```
Input:
Directed graph $\mathcal{G}$ stored in-adjacency and out-adjacency list format
$outAdj(x)$: out-adjacency list of node $x$
Initialization:
for each vertex $x \in \mathcal{V}$ do
  if $indegree(x) = 0$ then
    $color(x) = BLACK$
    $d(x) = 0$
    $ENQUEUE(Q, x)$
  else
    $color(x) = WHITE$
Levelization:
while $Q \neq \emptyset$ do
  $x = DEQUEUE(Q)$
  for each vertex $y \in outAdj(x)$ do
    if $color(y) = WHITE$ then
      $color(y) = BLACK$
      $d(y) = d(x) + 1$
      $\mathcal{V}_{d(y)} = \mathcal{V}_{d(y)} \cup \{y\}$
      $ENQUEUE(Q, y)$
return $(\mathcal{V}_0, \mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_{L-1})$
```

Let $\mathcal{V}_0, \mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_{L-1}$ denote the node levels of this cascade form of $\mathcal{G}$, where $\mathcal{V}_0$ denotes the set of nodes with zero in-degree. This cascade form enables us to solve the score convergence problems of some cyclic graphs. The proposed algorithm adopts an iterative process which updates the score of the nodes in a level-wise fashion. At each iteration of the algorithm, the nodes of the graph are processed in level order, i.e., nodes in level ℓ are processed before the nodes in level $\ell + 1$. The processing of a node refers to transferring its score to the nodes in its out-adjacency list. At iteration k, a node x transfers its $S^k_{out}(x)$ to each node y in its out-adjacency list according to following equation:

$$f^k(x, y) = sign(x, y) * S^k_{out}(x) * \frac{S(y)}{\sum_{z \in outAdj(x)} S(z)}. \tag{3.4}$$

As seen in Equation 3.4, the out-score of node x is divided among the nodes in $outAdj(x)$ according to the self-scores of those nodes. The distribution of out-score of a parent node is called *partitioned score transfer* method that divides the score of effector (parent) node on the children according to the score of the child node. Thus, the nodes with small self-scores will get small share of $S^k_{out}(x)$ compared to the nodes having large self-scores. Note that the edge type between x and y is defined by $sign(x, y)$ where $sign(x, y) = 1$ denotes activation and -1 denotes inhibition. Hence, the out-score of a node x is updated at each iteration k by summing up the out-score transfers from the nodes in its in-adjacency list as:

$$S^{k}_{out}(x) = S(x) + \sum_{z \in inAdj(x)} f^{k}(z, x). \tag{3.5}$$

Algorithm 4 describes general steps of the biological activity score computation for each pathway. In Algorithm 4, the for-loop inside the initialization for-loop computes the sum of the self-scores of the nodes in out-adjacency of each node, which is equal to the denominator term of Equation 3.4. The scheme adopted in the while-loop of the score computation phase enables in-place accumulation of the contributions of the out-score of a given node x to the out-scores of the nodes in its adjacency list. Thus the scheme avoids the need for maintaining a flow value (see Equation 3.4) for each edge of graph $\mathcal{G}$. The reason of the iterative algorithm is the existence of cyclic signalling pathways in KEGG database, because out-scores of the nodes in a cycle need to be computed many times to get the stable node scores in the cycle. For this purpose, we execute the while-loop until obtaining converged out-scores for all nodes in the graph. The convergence on out-score of a node x is defined as:

$$S^{k}_{out}(x) - S^{k-1}_{out}(x) \leq \varepsilon, \tag{3.6}$$

where ε is the error threshold for convergence criteria and set to 10^{-6}. Note that the proposed algorithm does not necessitate expensive cycle detection process in graph $\mathcal{G}$. Instead, we pass over the entire graph level by level (as indicated in pseudo-code) to achieve the converged out-scores for all nodes.

The graph $\mathcal{G}$ represents an overall pathway containing one or more biological processes. In $\mathcal{G}$ different biological processes are represented by a different subset of target nodes where the distinguishing property of a target node is having zero out-degree. Let $\mathcal{P}$ denote the set of biological processes in $\mathcal{G}$ and let $\mathcal{T}(p)$ denote the subset of target nodes representing biological processes $p \in \mathcal{P}$. Total activity score for a biological process p is computed by taking the sum of all possible biological processes leading to p in $\mathcal{G}$, which is the target biological process linking current pathway to the other pathways in KEGG database.

$$S_{tot}(p) = \sum_{t \in \mathcal{T}(p)} S_{out}(t). \tag{3.7}$$

$S_{tot}(p)$ might be referred to as the *final activity score* of a process p.

3.5 Gene Knockout and Its Evaluation by SiTSFlow Algorithm

There exists genes with high activity scores and they can be evaluated as hub-proteins in a pathway. We furthermore assess the lethality of hub-proteins for the life cycle of a cell by using SiTSFlow algorithm and gene knockout operations on a network. The proposed hypothesis is that the scores of target processes would be highly affected by the deletion of particular hub-proteins. For this purpose, we have constructed Akt pathway called as *original Akt pathway* by integrating of known Akt related genes in literature. The constructed pathway contains 83 gene nodes, 6 target process, DNA repair, Translation, Migration, Angiogenesis, Apoptosis, and Cell Cycle nodes, and 160 edges, 105 activation type, and 55 inhibition type.

We selected Akt pathway related microarray data set (called as *KRas data*) from GEO database (GSE12398). There are certain biological reasons to choose KRas data from the literature. The first reason is that, activating mutations in the small guanosine triphosphate-binding protein Ras, such as G12Asp and G12Val mutations, lead to constitutive downstream signalling and transfection of cell lines with the mutant Ras makes them tumorigenic [57]. Indeed, these Ras mutations have different effects on carcinoma cells due to their structural differences [58].

Another literature review has been performed to explore related processes and genes with KRas data. The serine or threonine kinase Akt is a critical signalling node promoting cell survival and it has been shown to be constitutively expressed in a variety of human tumors [59, 60, 61]. Activated Akt is known to regulate cell survival, cell cycle, translation, DNA repair, apoptosis, migration, invasion and angiogenesis processes [62]. Akt and Ras pathways have been shown to interact with each other and activating mutations in both pathways promote tumor cell growth [63, 64]. The transcription factor (TF) p53 is another important hub-protein in cell signalling regulating pathways, such as apoptosis, cell cycle, and DNA repair. p53 promotes apoptosis through its pro-apoptotic targets Bcl2, Puma, Noxa, and Bax [65, 66]. Based on these experimental results obtained in these studies, we decided to remove Akt, p53, and both Akt-Erk genes from the original Akt pathway. After application of knockout operations, new pathways are called as *Akt knockout*, *p53 knockout*, and *Akt-Erk double-knockout*. The meaning of *double-knockout* is that indicated genes and their adjacent edges are simultaneously deleted from the pathway. In order to identify the most affected biological processes from knockout operations, the scores obtained for knockout pathways are compared to the scores of original Akt pathway.

The application of a knockout operation is explained as follows.

1. Select a gene x from original Akt pathway.
2. Delete the node of x and all edges originating from node x, called this new graph as x *knockout*.
3. Run SiTSFlow algorithm on new x *knockout* graph.
4. Compute the activity score of each process of x *knockout* graph.

3.6 Calculation of Significance and Sensitivity of Activity Scores

Permutation tests are generally designed to determine whether the observed result e.g., final activity score, calculated by a proposed method is different between control and experiment classes of input data. For this purpose, a comparison measure to represent difference between the sample classes and a null hypothesis are designed. In order to evaluate significance and sensitivity of final activity scores obtained by SiTSFlow algorithm, a comparison measure is constructed as follows:

$$R_{true} = \frac{S_{tot}(p_{control})}{S_{tot}(p_{exp})}, \tag{3.8}$$

where $S_{tot}(p_{control})$ and $S_{tot}(p_{exp})$ are the total activity scores of a biological process p obtained with original control and experiment data, respectively. R_{true} value is crucial to identify which experimental condition has more effect on the activity of a specific process. In permutation procedure, R_{true} is computed after every shuffle operation on data and R_{true} is referred as *ratio score*. The null hypothesis H_n is designed as follows:

$$H_n = \frac{S_{tot}(p_{control})}{S_{tot}(p_{exp})} = 1. \tag{3.9}$$

The procedure for a permutation test considering the actual graph structure is performed as follows.

1. Randomly select a gene, shuffle data of that gene between control and experiment classes. Perform shuffling until reaching 50% of entire data set.
2. Run SiTSFlow algorithm on new shuffled data set.

3. Compute the ratio score of a process for control and experiment classes by using shuffled data set.

4. Repeat steps 1, 2, and 3 for N times.

The new ratio scores obtained with permutation test generate new resampled data. This resampling provides to estimate the sampling distribution under the condition that H_n is false. The position of the R_{true} value on the resampled distribution is determined to assign a significance value, α_{value}, for the calculated activity score.

$$\alpha_{value} = \frac{TH}{N+1}, \tag{3.10}$$

where TH represents how many times the ratio score of resamples is greater than or equal to R_{true}, and N is the total number of iterations performed in permutation procedure and set to 1000.

We also performed a sensitivity analysis to determine how the described system is affected by the variation of inputs. This technique observes the effects of parameter change on the outputs of the model. We used new ratio scores obtained with permutation test explained above to compute sensitivity of R_{true}. For this purpose, the sample variance of new ratio scores is calculated.

$$\sigma^2 = \frac{\sum_{i=1}^{M} R_{newi} - \mu}{N-1}, \tag{3.11}$$

where R_{new} represents the new ratio score obtained with new shuffled data, μ is the mean of R_{new} samples, and N is the total number of iterations performed in permutation procedure and set to 1000. We applied Equation 3.12 to convert variance σ^2 value into a *sensitivity value*.

$$\sigma_{value} = \frac{\sigma^2}{R_{true}}. \tag{3.12}$$

σ_{value} indicates that how much an activity score is affected by 50% shuffling of input data. If a process has a consistent activity score, σ_{value} should be very small, i.e., between 0-1. In other words, even if for high percent shuffling on input data, a consistent activity score should able to preserve its original value differentiated in small variances.

3.7 Computational Complexity of SiTSFlow Algorithm

In order to consider processing order of the genes in a pathway $\mathcal{G}$, the BFS-based cascading Algorithm 5 transforms the directed graph into a cascade form. In the initialization step of Algorithm 5, initial or root nodes of the graph are identified and put in a processing queue. The initialization for-loop iterates over entire node set, so it runs $O(\mathcal{V})$ times. In the levelization step of Algorithm 5, the topological order of each node starting from the initial nodes is explored. The while-loop runs until the queue becomes empty, therefore it iterates over entire node set and runs $O(\mathcal{V})$ times. The for-loop of this part runs for each edge of the graph, so total time spent in this part is $O(\mathcal{E})$. Final running time for Algorithm 5 is in linear-time in the size of the pathway $\mathcal{G}$, that is $O(\mathcal{V} + \mathcal{E})$.

Algorithm 4 performs the score flow computation on given cascade form of $\mathcal{G}$. The initialization for-loop of Algorithm 4 makes a single scan over all nodes and edges of $\mathcal{G}$, so it runs for $O(\mathcal{V} + \mathcal{E})$ times. The while-loop of Algorithm 4 runs until obtaining converged node activity scores. The inside for-loop processes each node once thus processing each edge only once by following the topological ordering, hence it takes $O(\mathcal{V} + \mathcal{E})$ time. The entire SiTSFlow algorithm can be considered as a linear-time ($O(\mathcal{V}+\mathcal{E})$) algorithm if constant number of iterations suffices for convergence. Experimental results given in Chapter 4 proves that small number of iterations are needed for convergence.

Linear-time SiTSFlow algorithm may be applicable even if for huge cyclic graphs. It does not aim to detect cycles in such a huge graph, since the detection of cycles in such graphs is very expensive process. The cascaded graph form enables to solve the score convergence problems of some cyclic graphs and it runs in linear time. We have managed to obtain converged node activity scores even if cyclic graphs by using cascaded graph forms and applying an iterative score flow algorithm.

3.8 Convergence Analysis of SiTSFlow Algorithm

We explain in this section how the proposed algorithm provides the convergence of activity scores for every cyclic pathway. During the development of SiTSFlow algorithm, one objective was to provide convergence of activity score of each node even for a cyclic pathway. The aim was not direct identification of cycles in a graph. For this purpose, we applied a graph levelization method to cyclic pathways in the first phase of system (Algorithm 5). By using levelization information, the nodes are processed from first level to last level. An example of iterative score computation in a cyclic pathway is given in Figure 3.3. The nodes and edges that belong to a cycle are marked by yellow color (Figure 3.3a). After completion of first iteration of scoring stage over entire graph, the activity scores of nodes and *Anti-apoptosis* process are given in Figure 3.3a. After first iteration, the scores of nodes which are member of the cycle are not stable yet. After running 10 iterations of scoring for-loop, activity scores of nodes converged and while-loop finished. The converged score of nodes and *Anti-apoptosis* process are represented in Figure 3.3b. The comparison of activity scores of nodes and *Anti-apoptosis* process is given Table 3.1. The activity scores of initial nodes having zero BFS-level show a constant curve feature during 10 iterations, since these nodes are not actual member of the cycle. However, the nodes in other BFS-levels are directly affected by the cycle, since their activity score curves demonstrate up and down characteristics until reaching stable values. Due to nature of given cycle, the converged scores are lower than the scores of first iteration. The usage of the first iteration scores may cause some misleading biologic interpretations about the activity of *Anti-apoptosis* process, so the iterative algorithm should be essentially applied.

The crucial point of iterative algorithm providing convergence is the use of the *partitioned score transfer* method between the nodes. This method divides the score of effector (parent) node on the children according to the score of the child node. In other words, each child node receives a partitioned score from the parents based on its self-score $S(y)$. In other words, the nodes with small self scores do not share the same parent score with the nodes of high scores. Hence the out-score of a parent node ($S_{out}(x)$) is distributed to all of its children according to the magnitude of their self-scores $S(y)$. When partitioned score transfer method is combined with the iterative algorithm, the proposed system yields more approximate activity scores for each iteration. Finally, the out-score of each node reach to a saddle point in which convergence is satisfied. Therefore, there are three necessary conditions for the score convergence: given graph has at least one starting node having zero in-degree, target process nodes have zero out-degree, and *partitioned score transfer* is applied between the nodes.

Table 3.1: The activity scores of nodes and *Anti-apoptosis* process. The scores are calculated after running of 1st and 10th iterations of SiTSFlow algorithm. The initial and converged scores are represented in *1st iteration* and *10th iteration* columns of table, respectively.

Node Name	**BFS-Level**	**Activity Scores**	
		1st iteration	**10th iteration**
CBL	0	14	14
PTPN6	0	0	0
CSF3	0	200	200
CSF2	0	156	156
IL24	0	108	108
EPO	0	64	64
JAK1	1	1380	786
IL22RA2	1	1289	851
P101-PI3K	2	965	654
STAT1	2	877	594
AKT3	3	1038	727
NAP4	3	102	696
Anti-apoptosis	4	1038	727

The convergence of SiTSFlow algorithm can be explained based on the convergence control statement given in Equation 3.6. We rewrite this control statement by using Equation 3.5.

$$S(x) + \sum_{z \in inAdj(x)} f^k(z,x) - S(x) - \sum_{z \in inAdj(x)} f^{k-1}(z,x) \leq \varepsilon. \tag{3.13}$$

Equation 3.13 can be expanded by using Equation 3.4 and then it is arranged.

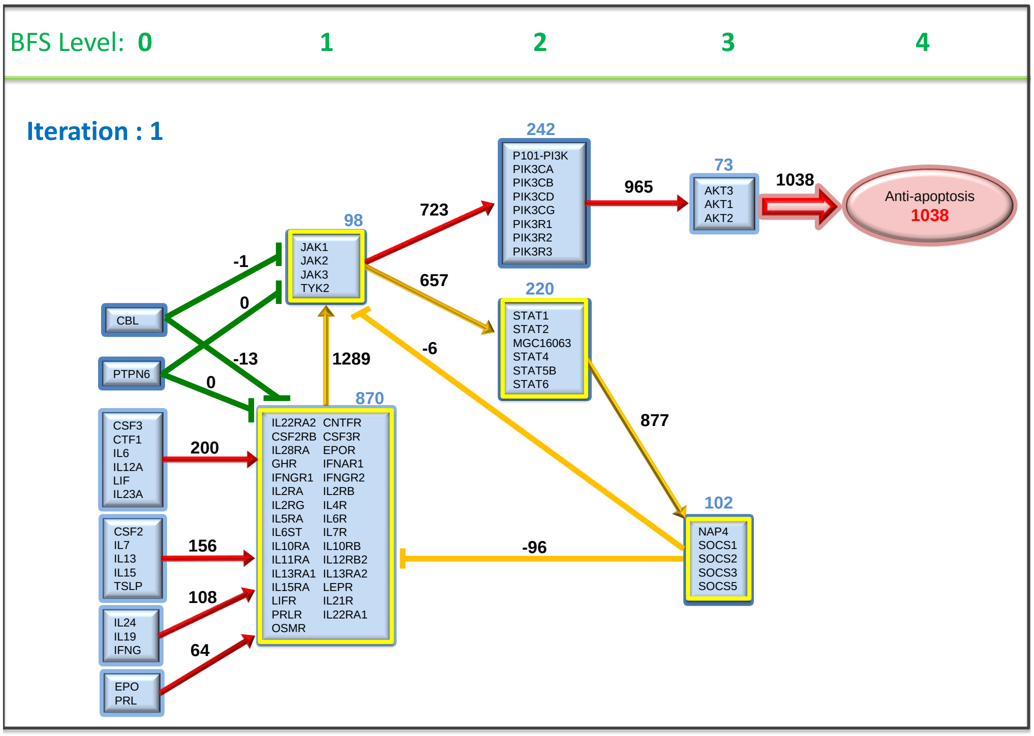
$$\sum_{z \in inAdj(x)} \Big[sign(z,x) * \frac{S(x)}{\sum_{m \in outAdj(z)} S(m)}\Big] * \Big[S^k_{out}(z) - S^{k-1}_{out}(z)\Big] \leq \varepsilon. \tag{3.14}$$

The first term of the summation given in Equation 3.14 is a constant term and it does not change during iterating of the algorithm. However, the second term of the summation should converge to threshold ε. It is clear that we could expand the right side of the summation, similar to Equation 3.13, until reaching the root nodes in level 0.

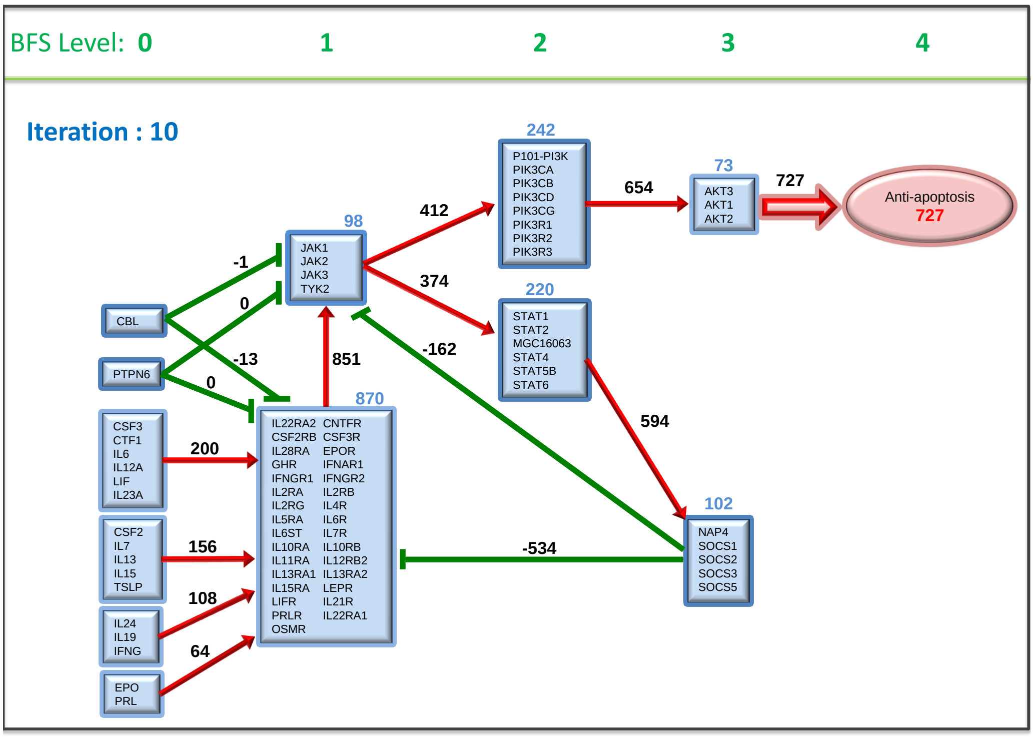
$$\sum_{z \in inAdj(x)} \Big[sign(z,x) * \frac{S(x)}{\sum_{m \in outAdj(z)} S(m)}\Big] * \Big[\sum_{y \in inAdj(z)} f^k(y,z) - \sum_{y \in inAdj(z)} f^{k-1}(y,z)\Big] \leq \varepsilon. \tag{3.15}$$

Let assume that node y is one of the root node of the given pathway and it is also the parent of node z. The root nodes of a given pathway place in the level 0 and they have zero in-degrees. Therefore, the out-score of a parent node, $S_{out}(y)$, is always equal to its self-score, and it is partitioned between its children. The partitioned score of parent y to child z is given by $f(y,z)$. The out-score of the parent node is not affected by the iterations of the algorithm, so it is assumed as a constant term. Thus, for the level 0, the terms $f^k(y,z)$ and $f^{k-1}(y,z)$ become equal and Equation 3.15 converges to ε.

The iterative score computation algorithm has successfully managed to provide convergence of activity scores in a cyclic graph. Even for huge graphs, the iterative algorithm provides the score convergence. The experimental results of score convergence are explained in detail in Chapter 4.



(a) The out-scores of nodes after first iteration.



(b) The converged out-scores of nodes after ten iterations.

Figure 3.3: Activity score calculation for the cyclic target process (*Anti-apoptosis*) by using integrated gene scores. The number on each node (gene) represents self-score of the gene. Red and green edges represent activation and inhibition properties, respectively. The nodes and edges belong to a cycle are marked by yellow color.

CHAPTER 4

EXPERIMENTAL RESULTS OF NETWORK STRUCTURE BASED PATHWAY ENRICHMENT SYSTEM

This chapter provides experimental results of network structure based pathway enrichment system on several data sets. Described system was implemented on various KEGG pathways with three different sets of microarray and its complimentary ChIP-seq data obtained from HeLa cells under oxidative stress, Estradiol (E2) treated MCF7 cells, and Estrogen Receptor beta treated U2OS cells. We applied SiTSFlow algorithm on manually curated Akt pathway with transcriptome data from Colo741 cells transfected by two KRas mutations and experimented gene knockout operations on the curated Akt pathway. The state of the art methods were also experimented with our data sets. KRas expression data was applied on both SPIA and GSEA methods. Similarly, the data of HeLa cells under oxidative stress was applied on *kegArray* tool. Finally, the comparisons of technical capabilities of these pathway enrichment tools are provided as well. The experimental results are discussed in both biological and computational perspectives.

4.1 Data Sets

We experimented SiTSFlow algorithm on four different data sets: `HeLa cells under oxidative stress`, `Estradiol (E2) treated MCF7 cells`, `Estrogen Receptor (ER) beta treated U2OS cells`, and `KRas data`.

First data set of `HeLa cells under oxidative stress` was obtained from NCBI GEO database (GSE14283, GSE4301). The ChIP-seq data by Kang et al. is performed to determine transcription regulation role of OCT1 transcription factor (TF) on HeLa cells under oxidative stress condition [67]. Raw ChIP-seq data of OCT1 TF includes approximately 3.8 million reads. After performing peak detection phase, we identify 5080 putative peak regions for OCT1 ChIP-seq data. Then gene mapping phase is applied, and finally, 268 neighboring genes are identified as significant. The rank value of each significant gene is computed by using Equation 3.1. The microarray data set related with OCT1 TF was selected from HeLa cells having control and oxidative stress experiments [68]. In the microarray analysis part, we compute fold-change ratio of two channel data for control and oxidative stress experiments. Fold-change value of each gene is converted into a rank value by using Equation 3.1. Total number of genes ranked in microarray chip is 12854 and all of them are used during data integration part.

Data set of `E2 treated MCF7 cells` was obtained from NCBI GEO database (GSE19013, GSE11352). The ChIP-seq data by Hu et al. is performed to determine transcription regulation role of estrogen receptor (ER) transcription factor on MCF7 breast cancer cell line [69]. ER is a hormonal transcription factor that plays important roles in breast cancer. It functions primarily through binding to the regulatory regions of target genes containing the consensus ERE motifs. In order to identify ER target genes and redefine the ERE motifs we perform ChIP-Seq analysis of ER in MCF7 breast cancer cell line. After completing peak detection phase for ChIP-seq data, we identify 1906 putative peak regions. As the result of gene mapping phase, 485 neighboring genes are identified as significant. The rank value of each significant gene is computed by using Equation 3.1. We selected a microarray data set experimented on MCF7 breast cancer cells as well [70]. Experiments are performed on Affymetrix U133 Plus 2.0 GeneChip. The aim of microarray experiment is to identify E2-responsive genes in the ER positive MCF7 breast cancer cell line. Therefore, the samples are collected at 12, 24, and 48 hours. In the microarray analysis part, raw data is normalized by Robust Multi-array Average (RMA) pre-processing method [71]. Then, expression levels of 12 and 48 hours are compared to observe time dependent expression changes under E2 effect.

$$\Delta(x) = x_{48h} - x_{12h}, \tag{4.1}$$

where x_{12h} and x_{48h} represent gene expression samples collected at 12 and 48 hours, respectively. $\Delta(x)$ value of gene x is converted into a rank value by using Equation 3.1. Total number of genes ranked in microarray chip is 20271.

Data set of `ER beta treated U2OS cells` was also selected from NCBI GEO database (GSE21790) [72]. In order to understand how ER beta regulates genes, Vivar et al. identify genes regulated by the unliganded (doxy) and liganded (doxy E2) forms of ER beta in U2OS cells by applying ChIP-seq experiments. Unliganded form of ER beta is set as *control* sample for the ChIP-seq analysis. Similarly, liganded form of ER beta is set as *experiment* sample. After completing peak detection phase for ChIP-seq data, we identify 4400 and 9869 putative peak regions for the unliganded and liganded form of ER beta, respectively. As the result of gene mapping phase, 851 and 116 neighboring genes are identified as significant for the unliganded and liganded form of ER beta, respectively. The rank value of each significant gene is computed by using Equation 3.1. We used the Illumina beadchip microarray data included in GSE21790 data set. The experiments are performed for 3 conditions: ER beta transfected without doxycycline (nodoxy), ER beta transfected with doxycycline (ERb-doxy), and ER beta transfected with doxycycline and E2 treated (ERb-doxyE2). In the microarray analysis part, raw data is analyzed by using R-Bioconductor “lumi” package. The expression levels of ERb-doxy and ERb-doxyE2 are compared with that of control sample (i.e. noDoxy) to observe E2 dependent expression changes of the genes.

$$\begin{aligned} \Delta_{ERb}(x) &= x_{Doxy} - x_{noDoxy} \\ \Delta_{E2}(x) &= x_{DoxyE2} - x_{noDoxy}, \end{aligned} \tag{4.2}$$

where x_{noDoxy}, x_{Doxy}, and x_{DoxyE2} represent the gene expression samples of control, ERb-doxy, and ERb-doxyE2 experiments, respectively. $\Delta_{E2}(x)$ (i.e., E2-liganded) and $\Delta_{ERb}(x)$ (i.e., control-unliganded) values of each gene are converted into the rank values by using Equation 3.1. Total number of genes ranked in microarray chip is 25186.

`KRas data` set was used during the gene knockout operations. We could not find Ras gene related ChIP-Seq experiment from public databases, therefore this data set only contains microarray experiments. In this experiment, the adenocarcinoma cell line Colo741 is selected to produce stable transfectants for two mutant forms of KRas (Gly12Asp and Gly12Val) and experiment control [73]. In the microarray analysis part, we performed a row-wise normalization on raw data. In order to consider the expression effects of Gly12Asp and Gly12Val mutations over control sample, we compute differences between control and mutated expression levels of genes.

$$\Delta_{G12A}(x) = x_{control} - x_{G12A} \tag{4.3}$$
$$\Delta_{G12V}(x) = x_{control} - x_{G12V},$$

where x_{G12A}, x_{G12V}, and $x_{control}$ represent gene expression samples of Gly12Asp, Gly12Val, and control experiments, respectively. Then the ranking scores of the genes according to their expression changes (i.e., $\Delta_{G12A}(x)$ and $\Delta_{G12V}(x)$) are computed by applying Equation 3.1. Total number of genes ranked in microarray chip is 20098.

Table 4.1 provides summary information about total number genes identified in peak detection and gene mapping phases applied for ChIP-seq data analysis. After performing microarray analysis phase, the total number of remaining genes in a chip is given in the rightmost column of Table 4.1.

Table 4.1: The details of employed data sets. *Peak Detection* column represents total number of significant peak regions identified in *peak detection* phase. *Gene Mapping* column represents total number of neighboring genes found in *gene mapping* phase. Microarray analysis column represents total number of genes in a chip after completing *microarray analysis* phase.

Experiment	**Peak Detection**	**Gene Mapping**	**Microarray Analysis**
HeLa cells under oxidative stress	5080	268	12854
E2 treated MCF7 cells	1906	485	20271
ER beta treated U2OS cells	9869	851	25186
KRas data	-	-	20098

4.2 KEGG Pathways

Pathways are set as the model to derive cell signalling scoring by applying SiTSFlow algorithm. Therefore, we selected several signalling pathways from KEGG PATHWAY Database: Apoptosis, Cell cycle, ErbB signalling, Focal adhesion, Insulin signalling, Jak-STAT signalling, MAPK signalling, mTOR signalling, Pathways in cancer, P53 signalling, Regulation of actin cytoskeleton, TGF-β signalling, and Wnt signalling pathways. Table 4.2 summarizes the total number of nodes, genes, and processes contained for each pathway.

Table 4.2: The details of selected pathways from KEGG PATHWAY Database. The total number nodes, edges, and processes contained in each pathway are listed.

Pathway Name	# of Nodes	# of Edges	# of Processes
Apoptosis	67	71	3
Cell cycle	112	80	3
ErbB signalling	71	93	6
Focal adhesion	66	94	5
Insulin signalling	69	91	5
Jak-STAT signalling	26	35	4
MAPK signalling	136	189	5
mTOR signalling	31	35	4
Pathways in cancer	223	275	6
P53 signalling	69	95	7
Regulation of actin cytoskeleton	76	87	3
TGF-β signalling	65	54	4
Wnt signalling	69	79	4

4.3 Application of SiTSFlow with HeLa Cells Under Oxidative Stress

Gene ranking scores obtained from microarray and ChIP-seq experiments of HeLa cells under oxidative stress were integrated to compute the self-score of each gene. These gene self-scores were mapped onto several pathways selected from KEGG PATHWAY Database: Pathways in cancer, Cell cycle, P53 signalling, Insulin signalling, Regulation of actin cytoskeleton, Jak-STAT signalling, Apoptosis, TGF-β signalling, MAPK signalling, mTOR signalling, and Wnt signalling. These pathways have 2-6 target cellular processes and include several cycles. Therefore, SiTSFlow algorithm might run 5-10 times over the entire cyclic graph until verifying the convergence threshold.

When the total activity scores of target biological processes were compared, *MAPK signalling* process in Regulation of actin cytoskeleton pathway produced a score of 4551 under the oxidative stress condition (Table 4.3). If the confidence threshold of α_{value} was set to 0.1, there were only 5 significant processes (*Apoptosis*, *Resistance to chemotherapy*, *Focal Adhesion*, *Survival*, and *Regulation of autophagy*) out of 45 target processes. Based on σ_{value} assessment criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data. The significant biological processes were specific to biological function of a given pathway, which is more in correlation with the cellular machinery. The response of a cell to a condition either normal or stressed was expected to be differential; therefore as a result of our analysis, some of the target processes were activated whereas others were down-regulated. When compared with the previous results of our study [3], the new cycle computation algorithm computed more realistic activity scores provided with significance σ_{value} and sensitivity α_{value} values.

In the biological perspective, under oxidative stress condition, gene expression responses of HeLa cells indicated a decrease in *Apoptosis* (given in first row of Table 4.3), *Resistance to chemotherapy* (in second row), and *Focal Adhesion* (in fifteenth row) processes and an increase in *Survival* (in eleventh row) and *Regulation of Autophagy* (in eighth row) processes. This indicated that as a response to oxidative stress, HeLa cells stimulate autophagy opposed to apoptosis to increase cell survival. It has been shown that hypoxia induces cells to assemble cytoplasmic stress granules as a major adaptive defense mechanism, so that apoptosis is inhibited and survival is enhanced through induction of autophagy [74]. However, when autophagy is prolonged, it can switch from being a cell-survival mechanism to a cell-death mechanism and this can render cells sensitive to chemotherapy [75].

Table 4.3: Activity scores of biological processes for control and oxidative stress samples in HeLa cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face.

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		Control Sample	Oxidative Stress	α_{value}	σ_{value}
hsa05200 Pathways in cancer	Apoptosis	**201**	137	0.089	0.048
	Resistance to chemotherapy	**84**	71	0.001	0.021
	Block of differentiation	**632**	602	0.396	0.003
	Proliferation	3214	**3833**	0.257	0.026
	Evading apoptosis	2586	**2747**	0.459	0.040
	Sustained angiogenesis	2056	**2740**	0.140	0.030
hsa04150 mTOR signalling	Cell growth	**169**	40	0.287	0.003
	Regulation of autophagy	86	**143**	0.001	0.002
	VEGF signalling	363	**403**	0.235	0.003
	Differentiation	**51**	46	0.496	0.114
hsa04210 Apoptosis	Survival	131	**266**	0.103	0.310
	Apoptosis	**1437**	1260	0.200	0.011
	Degradation	**663**	440	0.174	0.059
hsa04810 Regulation of actin cytoskeleton	MAPK signalling	3937	**4551**	0.261	0.021
	Focal Adhesion	**325**	179	0.107	0.064
	Adherens junction	**1079**	876	0.386	0.085
hsa04110 Cell cycle	Apoptosis	151	**209**	0.386	0.324
	DNA biosynthesis	554	**684**	0.263	0.022
	S-phase proteins	124	**133**	0.463	0.844
hsa04010 MAPK signalling	Proliferation	**2825**	2676	0.413	0.020
	Cell cycle	**606**	593	0.476	0.063
	Apoptosis	334	**454**	0.260	0.076
	p53 signalling	108	**116**	0.499	0.128

Continued on next page

Table 4.3 – continued from previous page

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		Control Sample	Oxidative Stress	α_{value}	σ_{value}
hsa04115 P53 signalling	Apoptosis	506	**557**	0.336	0.033
	DNA repair and damage prevention	375	**402**	0.387	0.009
	Cell cycle arrest	**446**	226	0.239	0.117
	Inhibition of angiogenesis and metastasis	**322**	267	0.334	0.196
	Inhibition of IGF1 / mTOR pathway	**142**	79	0.156	0.150
	P53 negative feedback	330	**410**	0.347	0.123
hsa04630 Jak-STAT signalling	Anti-apoptosis	2143	**2613**	0.179	0.023
	Cell cycle	**391**	301	0.319	0.136
	Ubiquitin mediated proteolysis	710	**683**	0.469	0.012
	MAPK signalling	**363**	158	0.191	0.361
hsa04910 Insulin signalling	Apoptosis	39	**58**	0.316	0.024
	Glucose homeostasis	193	**311**	0.388	0.672
	Lipid homeostasis	**487**	431	0.384	0.012
	Protein synthesis	**1467**	1214	0.162	0.012
hsa04350 TGF-β signalling	Cell cycle	**135**	103	0.464	0.001
	MAPK signalling	**83**	39	0.262	0.475
	Apoptosis	**25**	24	0.336	0.001
	Ubiquitin mediated proteolysis	**400**	358	0.413	0.025
hsa04310 Wnt signalling	Proteolysis	**447**	377	0.206	0.016
	Cell cycle	475	**481**	0.481	0.064
	Gene transcription	739	**866**	0.247	0.023
	Cytoskeletal change	**155**	135	0.351	0.042

4.4 Application of SiTSFlow with Estradiol Treated MCF7 Cells

ER is a hormonal transcription factor that plays important roles in breast cancer. It functions primarily through binding to the regulatory regions of target genes containing the consensus ERE motifs. By using integrated gene scores obtained from ER treated MCF7 cells, we applied SiTSFlow algorithm to several KEGG pathways: Pathways in cancer, Cell cycle, P53 signalling, Insulin signalling, Regulation of actin cytoskeleton, Jak-STAT signalling, Apoptosis, TGF-β signalling,, MAPK signalling, mTOR signalling, Wnt signalling, ErbB signalling, and Focal adhesion pathways (see Table 4.4). These pathways have 2-7 target cellular processes and include several cycles. The algorithm might run 5-8 times over the entire cyclic graph until verifying the convergence threshold. If the confidence threshold of α_{value} was set to 0.1, there were only 6 significant processes (*Resistance to chemotherapy*, *Glucose homeostasis*, *Ubiquitin mediated proteolysis*, *Apoptosis*, *Degradation*, and *Cell cycle*) out of 53 target processes. If we consider σ_{value} criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data.

We observed an increase in *Proliferation* process (given in seventh row of Table 4.4) in response to E2 treatment in the estrogen-receptor positive MCF7 breast cancer cell line. This is in correlation with the proliferative effect of E2 on MCF7 cells as demonstrated by previous studies [76, 77, 78]. In agreement, *Cell cycle* process (given in eighth row) was increased significantly in MAPK signalling. E2 treatment was shown to increase E-cadherin in ER-alpha over-expressed ERalpha-negative cell lines and to become more proliferative and less invasive [79]. There was a significant increase in *Glucose Homeostasis* process (given in twentieth row) in E2 treated cells. This is in correlation with the data demonstrating the regulatory role of estrogen stimulated ERalpha on metabolic homeostasis and lipid metabolism [80, 81]. *Ubiquitin mediated proteolysis* process (given in seventeenth row) was significantly down-regulated in Jak-STAT signalling pathway in E2 treated cells. It is known that proteasomal degradation functions to limit E2-induced transcription through down-regulating ERalpha levels upon E2 binding [82]. Although the molecular mechanism of this receptor degradation is not known, our analysis suggests that Jak-STAT signalling might be involved. Furthermore, *Resistance to chemotherapy* process (given in second row) was significantly increased in E2 treated cells, consistent with the resistance of ER-positive cells like MCF7 to paclitaxel, probably through a mechanism involving Bcl-2, compared to ER-negative cell lines [83].

Table 4.4: Activity scores of biological processes for control and E2 samples in MCF7 cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face.

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		Control Sample	E2 Experiment	α_{value}	σ_{value}
hsa05200 Pathways in cancer	Apoptosis	309	**400**	0.223	0.080
	Resistance to chemotherapy	30	**97**	0.001	0.010
	Block of differentiation	356	**584**	0.301	0.001
	Proliferation	4223	**4427**	0.354	0.002
	Evading apoptosis	3622	**3796**	0.397	0.004
	Sustained angiogenesis	**2511**	2071	0.242	0.012
hsa04010 MAPK signalling	Proliferation	2902	**3504**	0.205	0.024
	Cell cycle	381	**718**	0.079	0.021
	Apoptosis	161	**354**	0.122	0.015
	p53 signalling	57	**124**	0.177	0.032
	Wnt signalling	38	**128**	0.455	0.006
hsa04210 Apoptosis	Survival	**284**	245	0.343	0.059
	Apoptosis	1612	**2251**	0.122	0.041
	Degradation	527	**957**	0.084	0.144
hsa04630 Jak-STAT signalling	Anti-apoptosis	2415	**3603**	0.197	0.092
	Cell cycle	**922**	826	0.419	0.086
	Ubiquitin mediated proteolysis	**1466**	607	0.069	0.104
	MAPK signalling	345	**560**	0.334	0.408
hsa04910 Insulin signalling	Apoptosis	**73**	64	0.339	0.014
	Glucose homeostasis	678	**1051**	0.077	0.019
	Lipid homeostasis	**475**	338	0.377	0.049
	Protein synthesis	**1182**	977	0.346	0.030
	Proliferation	87	**385**	0.164	0.008

Continued on next page

Table 4.4 – continued from previous page

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		Control Sample	E2 Experiment	α_{value}	σ_{value}
hsa04350 TGF-β signalling	Cell cycle	222	**238**	0.453	0.013
	MAPK signalling	82	**192**	0.300	0.008
	Apoptosis	13	**17**	0.001	0.001
	Ubiquitin mediated proteolysis	427	**449**	0.304	0.001
hsa04115 P53 signalling	Apoptosis	1127	**1274**	0.269	0.016
	DNA repair and damage prevention	409	**505**	0.267	0.041
	Cell cycle arrest	578	**644**	0.490	0.017
	Inhibition of angiogenesis and metastasis	299	**370**	0.196	0.029
	Inhibition of IGF1 / mTOR pathway	**123**	100	0.378	0.431
	P53 negative feedback	**463**	424	0.472	0.039
hsa04810 Regulation of actin cytoskeleton	MAPK signalling	5251	**5638**	0.358	0.013
	Focal Adhesion	**457**	410	0.436	0.143
	Adherens junction	664	**813**	0.250	0.037
hsa04110 Cell cycle	Apoptosis	173	**216**	0.314	0.123
	DNA biosynthesis	623	**638**	0.263	0.009
hsa04150 mTOR signalling	Cell growth	**41**	23	0.495	0.506
	Regulation of autophagy	**235**	207	0.383	0.002
	VEGF signalling	**329**	311	0.453	0.002
hsa04310 Wnt signalling	Proteolysis	619	**662**	0.278	0.004
	Cell cycle	768	**834**	0.372	0.028
	Gene transcription	826	**906**	0.402	0.029
	Cytosketal change	168	**204**	0.453	0.145
hsa04012 ErbB signalling	Degradation	198	**259**	0.486	0.272
	Adhesion migration	120	**121**	0.385	0.325
	Protein synthesis	227	**258**	0.484	0.231

Continued on next page

Table 4.4 – continued from previous page

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		Control Sample	E2 Experiment	α_{value}	σ_{value}
	Apoptosis	**1389**	1368	0.496	0.065
hsa04510	FA-turnover	**1108**	1073	0.480	0.047
	Cell survival	**306**	264	0.453	0.195
Focal Adhesion	Cell motility / FA formation	852	**868**	0.487	0.073
	Cell proliferation	1025	**1460**	0.163	0.061

4.5 Application of SiTSFlow with Estrogen Receptor Beta Treated U2OS Cells

ER beta has potent anti-proliferative and anti-inflammatory properties, suggesting that ER beta-selective agonists might be a new class of therapeutic and chemo-preventative agents. To understand how ER beta regulates genes, the experiments were performed for unliganded (ERb) and liganded (E2) forms of ER beta [72]. Unliganded and liganded form of ER beta were set as control and main experiment for the ChIP-seq analysis, respectively. After completing computational analysis of microarray and ChIP-seq data, gene ranks are integrated to construct self-score of each gene. We applied SiTSFlow algorithm by using self-scores obtained from ER beta treated U2OS cells to the same KEGG pathways with the previous experiment (see Table 4.5). When the total activity scores of target biological processes were compared, *MAPK signalling* process in Regulation of actin cytoskeleton pathway produced a score of 7011 under Erb condition (Table 4.5). If the confidence threshold of p-value was set to 0.1, there were only 5 significant processes (*Resistance to chemotherapy*, *Sustained angiogenesis*, *MAPK signalling*, *Cell cycle*, *Regulation of autophagy*) out of 56 target processes. If we consider σ_{value} criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data.

In E2 treated U2OS cells expressing ER Beta, *Resistance to chemotherapy* process (given in second row of Table 4.5) was significantly increased as in the E2 treated MCF7 cells, consistent with the resistance of ER-positive cells to paclitaxel compared to ER-negative cell lines [83]. In agreement with the induced autophagy with 2-methoxyestradiol in MCF7 cells, we observed an increase in *Regulation of autophagy* process (given in seventeenth row) through mTOR pathway [84]. Both *MAPK signalling* (given in fifteenth row) and *Cell cycle* processes (given in eighth row) were decreased. In MCF7 cells, it was previously shown that hyperactive MAPK down regulates ERalpha expression and inhibition of this hyperactive MAPK restores ERalpha expression [85, 86]. Therefore, we suggested that in E2 treated cells, reduced *MAPK signalling* may induce ERalpha signalling as well. The observed increase in *Sustained angiogenesis* process (given in sixth row) in E2 treated cells was also shown that E2 increases the expression of key angiogenic proteins, VEGF and TSP-1, through transcriptional activation [87, 88].

Table 4.5: Activity scores of biological processes for ERb and E2 samples in U2OS cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face.

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		ERb	E2	α_{value}	σ_{value}
hsa05200 Pathways in cancer	Apoptosis	289	**338**	0.356	0.083
	Resistance to chemotherapy	55	**69**	0.001	0.054
	Block of differentiation	645	**760**	0.157	0.010
	Proliferation	4307	**5305**	0.163	0.024
	Evading apoptosis	3285	**4125**	0.232	0.039
	Sustained angiogenesis	2668	**3425**	0.075	0.015
hsa04010 MAPK signalling	Proliferation	2888	**3144**	0.352	0.017
	Cell cycle	**780**	493	0.079	0.053
	Apoptosis	**575**	417	0.134	0.032
	p53 signalling	231	**258**	0.360	0.033
	Wnt signalling	**154**	116	0.133	0.022
hsa04630 Jak-STAT signalling	Anti-apoptosis	**3373**	3188	0.384	0.011
	Cell cycle	1025	**1058**	0.475	0.019
	Ubiquitin mediated proteolysis	1021	**1188**	0.400	0.024
	MAPK signalling	**701**	613	0.107	0.006
hsa04150 mTOR signalling	Cell growth	23	**60**	0.392	0.817
	Regulation of autophagy	69	**173**	0.095	0.001
	VEGF signalling	283	**346**	0.393	0.040
	Differentiation	11	**62**	0.245	0.389
hsa04910 Insulin signalling	Apoptosis	56	**59**	0.437	0.030
	Glucose homeostasis	724	**1034**	0.162	0.055
	Lipid homeostasis	554	**563**	0.470	0.006
	Protein synthesis	1146	**1262**	0.291	0.013

Continued on next page

Table 4.5 – continued from previous page

Pathway Name	Biological Process	Activity Scores of Target Process		Significance Scores	
		ERb	E2	α_{value}	σ_{value}
hsa04210 Apoptosis	Proliferation	405	**429**	0.364	0.026
	Survival	218	**288**	0.141	0.047
	Apoptosis	2307	**2347**	0.122	0.008
	Degradation	**937**	935	0.463	0.017
hsa04115 P53 signalling	Apoptosis	922	**1035**	0.282	0.014
	DNA repair and damage prevention	541	**644**	0.267	0.032
	Cell cycle arrest	457	**472**	0.437	0.007
	Inhibition of angiogenesis and metastasis	**396**	378	0.429	0.048
	Inhibition of IGF1 / mTOR pathway	**154**	103	0.378	0.264
	P53 negative feedback	**517**	395	0.198	0.028
	Exosome mediated secretion	**53**	24	0.425	0.317
hsa04110 Cell cycle	Apoptosis	309	**377**	0.314	0.029
	DNA biosynthesis	**824**	742	0.286	0.004
	S-phase proteins	111	**118**	0.496	3.215
hsa04810 Regulation of actin cytoskeleton	MAPK signalling	**7011**	6589	0.267	0.004
	Focal Adhesion	**671**	605	0.221	0.004
	Adherens junction	**1285**	987	0.254	0.048
hsa04350 TGF-β signalling	Cell cycle	246	**265**	0.364	0.004
	MAPK signalling	**195**	145	0.469	0.034
	Apoptosis	**37**	31	0.322	0.021
	Ubiquitin mediated proteolysis	**398**	325	0.207	0.017
hsa04310 Wnt signalling	Proteolysis	**493**	468	0.365	0.006
	Cell cycle	**563**	495	0.386	0.030
	Gene transcription	1113	**1243**	0.210	0.054
	Cytoskeletal change	**498**	361	0.453	0.005

Continued on next page

Table 4.5 – continued from previous page

Pathway Name	**Biological Process**	**Activity Scores of Target Process**		**Significance Scores**	
		ERb	E2	α_{value}	σ_{value}
hsa04012	Degradation	212	**249**	0.367	0.029
	Adhesion migration	**121**	106	0.495	0.203
ErbB signalling	Protein synthesis	**226**	186	0.149	0.012
	Apoptosis	1696	**1752**	0.449	0.017
hsa04510	FA-turnover	**1289**	1205	0.390	0.011
	Cell survival	157	**386**	0.278	0.026
Focal Adhesion	Cell motility / FA formation	628	**654**	0.491	0.018
	Cell proliferation	**1316**	1232	0.364	0.018

4.6 Scores from Individual Data Scorings

We compared the final activity scores of the processes based on only microarray or ChIP-seq rank scores, since we aimed to investigate the effect of individual data scores. So far, SiTS-Flow algorithm has used integrated gene scores to compute activity score of processes. However, computation of activity scores based on individual ranks might help us to understand which data source is more useful to explore activated biological events under experimental conditions. For this purpose, SiTSFlow algorithm is run by considering only microarray or ChIP-seq rank scores obtained from Erb and E2 experiments in U2OS cells. The summary of results are given in Table 4.6 which contains four different pathways: *Pathways in cancer*, *Jak-STAT signalling*, *MAPK signalling*, and *mTOR signalling*.

It is clear that, the activity scores obtained with microarray ranks are dominant on the integrated gene score results, since microarray data contains approximately 25000 genes that number is much more than 850 genes extracted from ChIP-seq data. However, ChIP-seq rank scores also provided significant activity scores for some processes, such as *Resistance to chemotherapy*, *Sustained angiogenesis* (given in second and sixth rows of Table 4.6). It proved that very small number of data extracted from ChIP-seq experiment is also very valuable to evaluate biological activities of processes. Therefore, integration of microarray and ChIP-seq rank scores provides more significant activity scores during evaluation of biological activities.

Table 4.6: Comparison of activity scores based on integrated gene scores, only microarray scores, and only ChIP-seq scores for several signalling pathways for Erb and E2 experiments in U2OS cells. Significant activity score of each process is marked by bold face.

Pathway Name	**Biological Process**	**Integrated scores**			**Only microarray ranks**			**Only ChIP-seq ranks**		
		Erb	E2	α_{value}	Erb	E2	α_{value}	Erb	E2	α_{value}
hsa05200 Pathways in cancer	Apoptosis	289	**338**	0.356	289	**338**	0.361	1	1	0.001
	Resistance to chemotherapy	55	**69**	0.001	55	**69**	0.001	1	1	0.001
	Block of differentiation	645	**760**	0.157	645	**756**	0.157	1	**14**	0.001
	Proliferation	4307	**5305**	0.163	4299	**5273**	0.174	39	**90**	0.213
	Evading apoptosis	3285	**4125**	0.232	3285	**4101**	0.247	1	**39**	0.123
	Sustained angiogenesis	2668	**3425**	0.075	2666	**3393**	0.067	17	**166**	0.075
hsa04630 Jak-STAT signalling	Anti-apoptosis	**3373**	3188	0.384	**3372**	3195	0.375	8	**51**	0.339
	Cell cycle	1025	**1058**	0.475	1025	**1079**	0.471	3	68	0.248
	Ubiquitin mediated proteolysis	1021	**1188**	0.400	1021	**1122**	0.397	2	**160**	0.247
	MAPK signalling	**701**	613	0.107	**701**	616	0.082	**2**	1	0.496
hsa04010 MAPK signalling	Proliferation	2888	**3144**	0.352	2888	**3197**	0.343	1	**25**	0.260
	Cell cycle	**780**	493	0.079	**782**	493	0.116	1	**38**	0.196
	Apoptosis	**575**	417	0.134	**577**	417	0.135	1	**16**	0.283
	p53 signalling	231	**258**	0.360	231	**258**	0.344	1	**16**	0.196
	Wnt signalling	**154**	116	0.133	**154**	118	0.150	1	1	0.452
hsa04150 mTOR signalling	Cell growth	23	**60**	0.392	23	**60**	0.400	1	1	0.001
	Regulation of autophagy	69	**173**	0.095	69	**173**	0.107	1	1	0.001
	VEGF signalling	283	**346**	0.393	283	**322**	0.404	1	**79**	0.001
	Differentiation	11	**62**	0.245	11	**62**	0.245	1	1	0.001

4.7 Effect of Gene Knockout on Pathway Enrichment

The proteins residing at central positions in network topology and having many interactions with other proteins are called *hub-proteins*. Our aim during the gene knockout operations was to prove the lethality of such hub-proteins for the life cycle of the cell. We expected that the scores of target processes in a signalling cascade would be affected by the deletion of such hub-proteins. For this purpose the Akt pathway was manually created by using known gene interactions in literature. Original Akt pathway is scored by applying SiTSFlow algorithm based on control sample, G12Asp, and G12Val mutation samples of KRas data see Figure 4.1. While applying knockout operation, the selected knockout gene and its connecting edges to its neighbors are removed from the original Akt graph. The activity score of each target process in the new pathways i.e., *Akt knockout*, *p53 knockout*, and *Akt-Erk double-knockout* is calculated by using SiTSFlow algorithm and same samples of KRas data. In order to identify the most affected biological processes from knockout operations, the scores obtained for knockout pathways are compared to the scores of original Akt pathway.

According to scoring results, the most affected biological process after performing of Akt, p53 and Akt-Erk double knockout operations was *Apoptosis* (see results in Table 4.7 and Table 4.8). As expected, final activity score of *Apoptosis* process was reduced in G12Asp and G12Val mutations compared to control sample in the original scoring. The score decrease of *Apoptosis* process was more prominent in p53 knockout pathway. In both Akt knockout and Akt-Erk double-knockout pathways, final activity score of *Apoptosis* increased and this result was consistent with the anti-apoptotic, proliferation-stimulating role of Akt gene. Comparing Akt-Erk double-knockout pathway to Akt knockout pathway indicated score increase in *Apoptosis*, which was supporting the survival promoting role of Erk gene. p53 knockout pathway resulted in very high decrease (-65.1%) in *Apoptosis* in control sample, so it proves that p53 is the most important regulator of *Apoptosis* process (see Figure 4.2-a). In Akt-knockout pathway, the activity score of *Angiogenesis* was reduced for all samples compared to original scores of Akt pathway. G12Val mutation of Ras has been shown to induce MAPK, invasion and angiogenesis and to be more tumorigenic than G12Asp mutation [89, 90]. Our analysis showed that *Angiogenesis* was higher in G12Asp mutation of Ras in BRAF mutated colorectal cancer cells (compare scores given in D and V columns of Akt knockout section in the first row of Table 4.8). *Angiogenesis* was increased in G12Asp and G12Val mutations compared to control sample. This result is also in correlation with the increase in PI3K activity in G12Asp mutated cells. On the other hand, *Cell cycle* process had no any activation in all pathways, except p53 knockout pathway, in which the final activity score of *Cell cycle* was reduced in mutations compared to control sample (Figure 4.2).

As a novel outcome of our analysis, we could infer that BRAF mutation could be associated with G12Asp mutation of Ras and the co-existence of these two mutations can enhance angiogenesis and render colorectal carcinoma cells more aggressive [73, 91, 92]. Furthermore, by using SiTSFlow, we showed that the processes like *Angiogenesis* and *Apoptosis* were regulated similarly in both mutations of Ras, but through different genetic combinations. This strengthens the importance and the necessity of integrating genetic networks and target processes and visualizing the signal transduction score flow as a whole with the interactions of genes leading to the target processes.

In a typical microarray experiment, genes are ranked according to their differential expression between the analyzed samples, such as tumor vs. healthy or drug-treated vs. untreated. However, the differentially expressed gene analysis cannot truly present the changes in cellular processes, since these processes are regulated by parallel or alternative signalling pathways that are interconnected to each other. For example, given the high score of the survival-promoting genes Akt and NFKB1 in a gene list of the analyzed Ras data, one would expect a low score in apoptosis. Analyzing the same data with tools that analyze gene sets, such as Gene Set Enrichment Analysis, will indicate an increase in apoptosis. With the visualization of the signalling network that is scored with SiTSFlow algorithm, it is possible to see not only that apoptosis has a high score but also which genes indeed are regulated in colorectal cancer cells with mutated BRAF and Ras so that these processes are affected, since a slight increase in most of the genes regulating a process can have a more prominent effect on a target process than a great increase in a single gene (Figure 4.2-c). Moreover, SiTSFlow algorithm can be used to predict process-level and global impacts of single or multiple gene knockouts. Use of our algorithm as a tool for *in silico* knockout analysis enables analysis and interpretation of the effect of genes of interest on a diverse range of cellular processes. In addition, it can be used to analyze the effects of knockout two genes from a single pathway, such as Akt and mTOR, or from parallel or alternative pathways, such as Akt and Erk at the same time, providing a useful tool for the development of combination drug therapies based on molecular mechanism of cancer cells.

SiTSFlow algorithm allows the visualization of the impact of inhibiting the targeted kinases not only on the first downstream proteins of their related signalling pathways, but on the global transcriptome and the various cellular processes, such as *Apoptosis* or *Angiogenesis*. It is possible to visualize the side-effects of inhibiting one protein, since its influence on target processes other than the expected ones is demonstrated as well. It would be of great value to be able to predict the drug combination that can not only increase the activity of *Apoptosis* in cancer cells but also decrease the activity of *Angiogenesis* process. These *in silico* analyses can suggest hypothesis on the molecular mechanism of action of the drug of interest and predict synergistic effect of different kinase inhibitors.

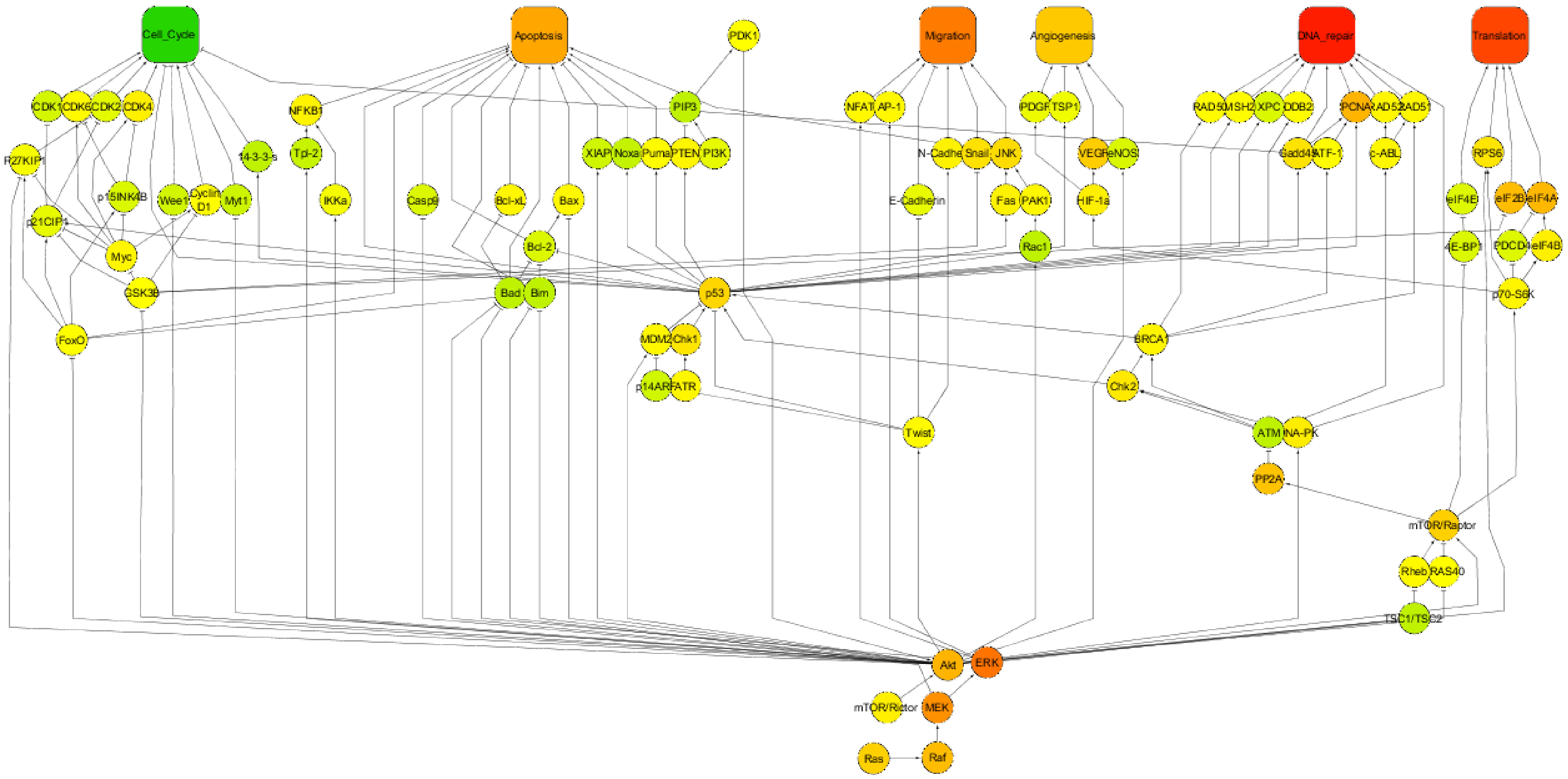


Figure 4.1: Activity scores of genes and processes in original Akt pathway by using KRas control sample. Down-regulated and up-regulated genes or processes are represented in color tones of green and red, respectively.

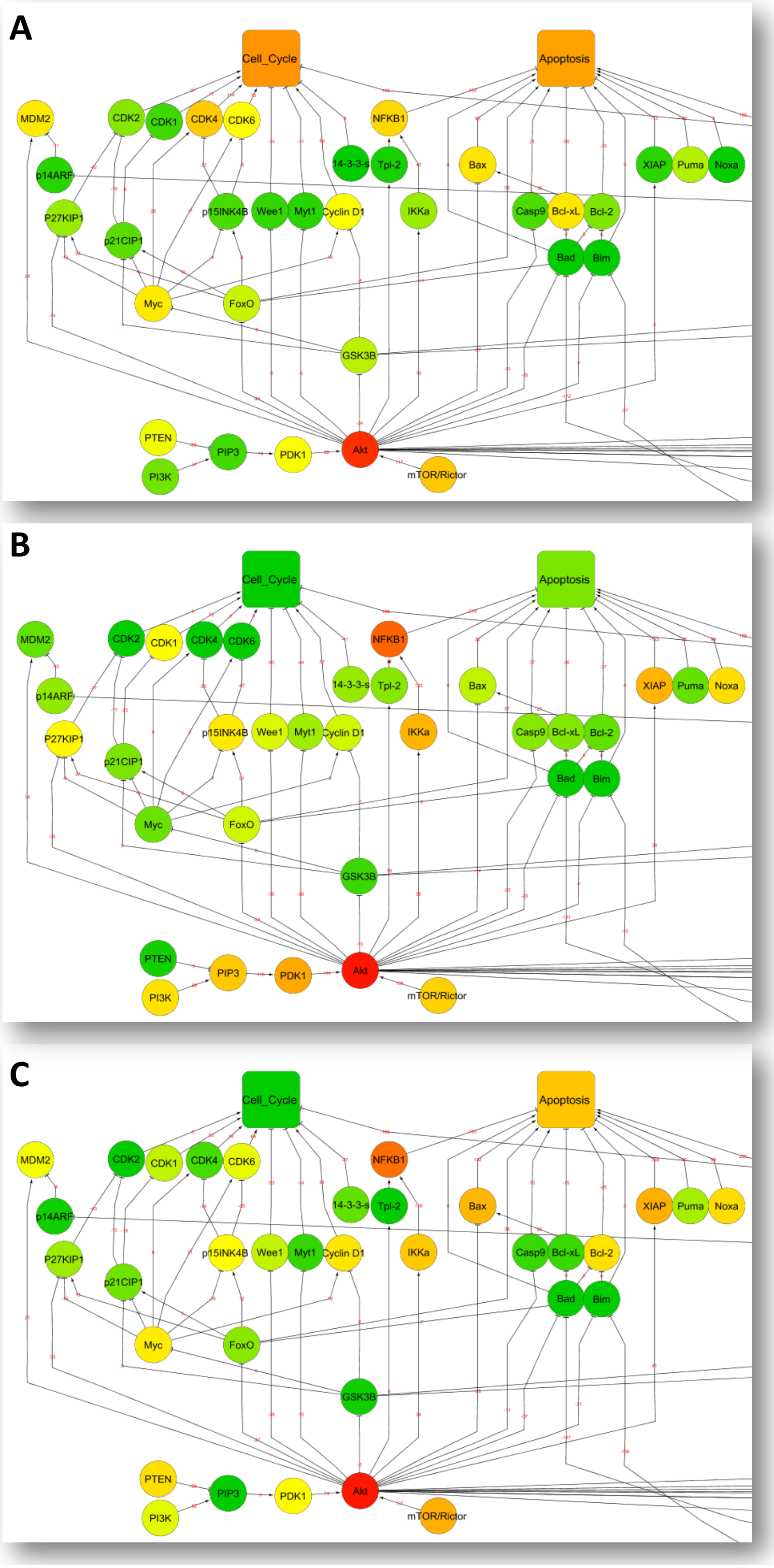


Figure 4.2: Cytoscape view of *Apoptosis* and *Cell cycle* processes in p53 knockout pathway by using KRas data. Activity scores of *Apoptosis* and *Cell cycle* processes for control sample (A), for G12D (B), and for G12V (C) in Colo741 cells are given in Table 3. Down-regulated and up-regulated genes or processes are represented in color tones of green and red, respectively.

Table 4.7: The original scoring results for Akt pathway by using control (C), Gly12Asp (D), and Gly12Val (V) samples of KRas data. The significance value of each score is specified by the α_{value} at the right column of its score.

Biological Process	**Original Akt**					
	C	α_C	D	α_D	V	α_V
Angiogenesis	281	0.12	515	0.24	366	0.14
Apoptosis	427	0.55	393	0.56	408	0.56
Cell cycle	0	0.05	0	0.02	0	0.07
DNA repair	1059	0.04	1380	0.17	1415	0.18
Migration	611	0.37	679	0.41	795	0.31
Translation	864	0.08	567	0.10	672	0.19

Table 4.8: Gene knockout results for *Akt knockout*, *p53 knockout*, and *Akt-Erk double-knockout* pathways by using control (C), Gly12Asp (D), and Gly12Val (V) samples of KRas data. The significance value of each score is specified by the α_{value} at the right column of its score.

Biological Process	**Akt Knockout**						**P53 Knockout**						**Akt-Erk Knockout**					
	C	α_C	D	α_D	V	α_V	C	α_C	D	α_D	V	α_V	C	α_C	D	α_D	V	α_V
Angiogenesis	272	0.14	480	0.25	349	0.13	292	0.12	534	0.21	372	0.12	269	0.10	479	0.24	348	0.14
Apoptosis	510	0.55	550	0.55	553	0.56	149	0.45	34	0.34	119	0.38	615	0.53	648	0.53	655	0.56
Cell cycle	0	0.06	0	0.03	0	0.10	161	0.11	0	0.04	0	0.15	0	0.05	0	0.02	0	0.08
DNA repair	1027	0.07	1352	0.23	1364	0.24	1011	0.05	1251	0.13	1397	0.06	1027	0.07	1358	0.18	1364	0.22
Migration	790	0.36	516	0.37	649	0.31	867	0.55	567	0.55	672	0.49	781	0.44	514	0.41	647	0.29
Translation	596	0.07	674	0.05	786	0.14	601	0.05	674	0.09	798	0.17	476	0.08	496	0.04	583	0.13

4.8 Comparison of Initial and Final Scores

This section explains the comparison of initial activity scores and final i.e., converged, scores of biological processes. The signalling pathways in KEGG database mostly contain several cycles. Due to signal transfer regulations of biological events, cycles are used in many times in a signalling pathway. Therefore, SiTSFlow algorithm has been developed to provide convergence of activity score of each node for cyclic pathways.

The activity score convergence graphics of processes for three pathways: *Apoptosis*, *Jak-STAT signalling*, and *Pathway in cancer*; and for three data sets are given in Figure 4.3, Figure 4.4, and Figure 4.5, respectively. Generally, all activity scores converge after running of SiTSFlow algorithm for 5-10 iterations. The score curve of each biological process is very similar for all data sets. In other words, the convergence behavior of a process is similar even if for different data sets. This fact proves that convergence of an activity score is only dependent to cycle structure of a pathway.

The score convergence curves of biological processes are changed according to whether their parent nodes are member of a cycle or not. In other words, if a node presents in a cyclic path, its children would be definitely affected by the score convergence phase of this node. For example, *Apoptosis* pathway contains three biological processes: *Degradation*, *Apoptosis*, and *Survival*. For three data sets, *Survival* process has a constant activity score. However, *Degradation* and *Apoptosis* processes represent increasing score trends. This shows that while the parent nodes of *Survival* process does not belong to a cycle, the parent nodes of *Degradation* and *Apoptosis* processes are members of cyclic paths. *Apoptosis* process is the most activated process for its pathway, since it has always highest score between other processes, for three data sets. Due to nature of cycles in *Apoptosis* pathway, the converged scores are higher than the scores of first iteration. However, the usage of the first iteration scores may cause some misleading biologic interpretations about the activity of processes in *Apoptosis* pathway.

Jak-STAT signalling pathway is composed of four different processes. Based on their activity score graphics, all of processes are member of cycles in this pathway, since their score curves are not constant. This pathway has the smallest pathway based on its total node and edge numbers. However, the convergence of activity scores in this pathway gets 6-10 iterations, that is the longest run of SiTSFlow algorithm for all experiments. *Anti-apoptosis* process is

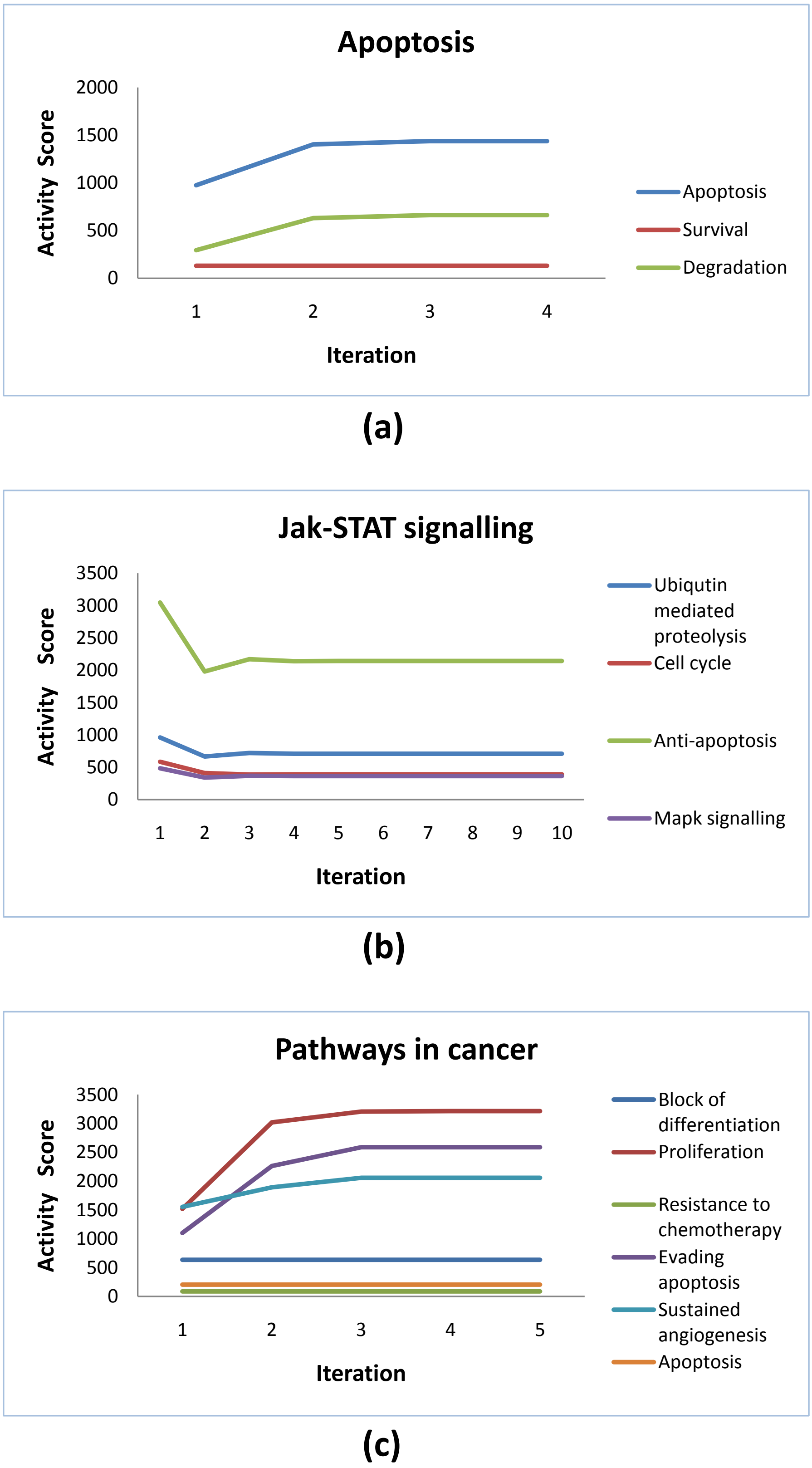


Figure 4.3: Convergence graphics of activity scores for the biological processes of Apoptosis, Jak-STAT signalling, and Pathway in cancer pathways by using control sample in HeLa cells.

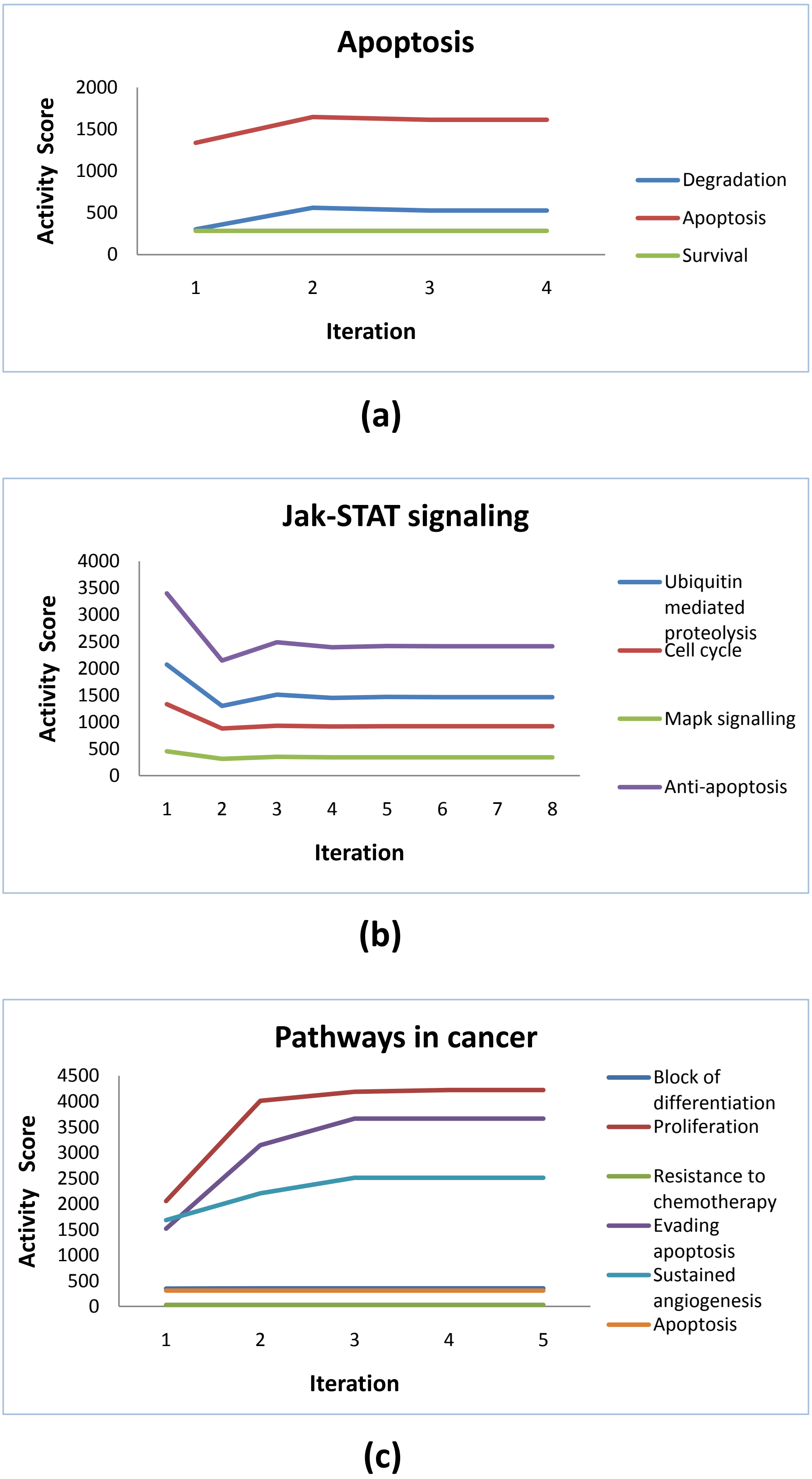
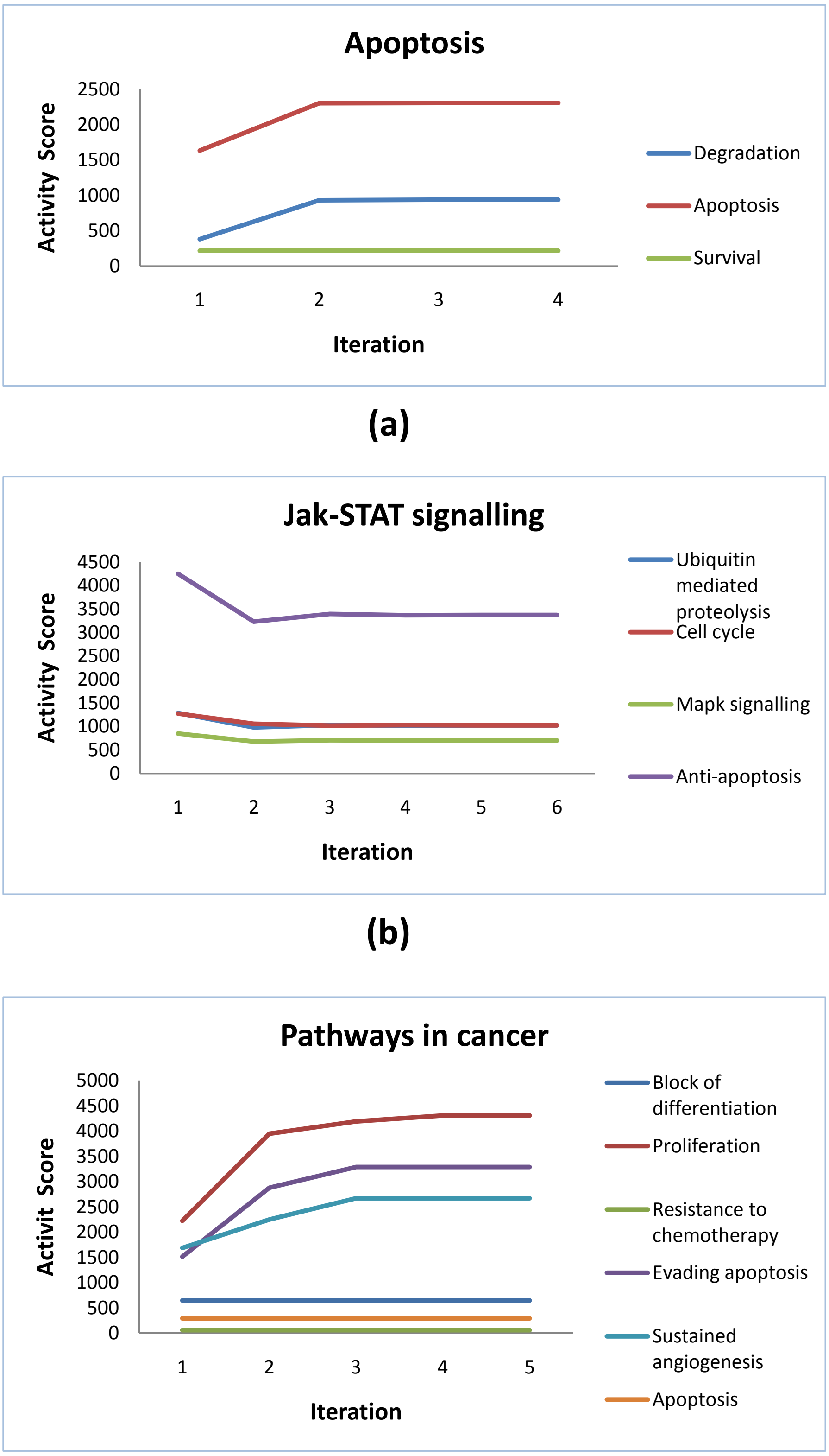


Figure 4.4: Convergence graphics of activity scores for the biological processes of Apoptosis, Jak-STAT signalling, and Pathway in cancer pathways by using control sample in MCF7 cells.



(a)

(b)

(c)

Figure 4.5: Convergence graphics of activity scores for the biological processes of Apoptosis, Jak-STAT signalling, and Pathway in cancer pathways by using control sample in U2OS cells.

the most activated process for this pathway due to its highest score for three data sets. Due to nature of cycles in *Jak-STAT signalling* pathway, the converged scores are lower than the scores of first iteration.

Pathway in cancer contains six different processes. The activity score curves of *Proliferation*, *Evading apoptosis*, and *Sustained angiogenesis* processes have increased for all data sets. However, *Block of differentiation*, *Resistance to chemotherapy*, and *Apoptosis* processes have constant activity score curves. *Proliferation* process is the most activated process for this pathway due to its highest score for three data sets. Due to nature of cycles in *Pathway in cancer* pathway, the converged scores are usually higher than the scores of first iteration.

The total number of iterations for convergence depends on the number of nodes in cycles. If the number of cyclic paths and the nodes contained in such paths is large, the convergence time of algorithm becomes high. This fact was proved in *Jak-STAT signalling* pathway, since almost all of nodes in that pathway present in cyclic paths. Although, in terms of total number of nodes and edges, *Jak-STAT signalling* pathway is the smallest pathway, it has the highest convergence time among other pathways. Finally, we might derive that convergence of scores depends on the number nodes present in cyclic paths rather than total number of nodes and edges of a pathway.

4.9 Cytoscape Plug-in

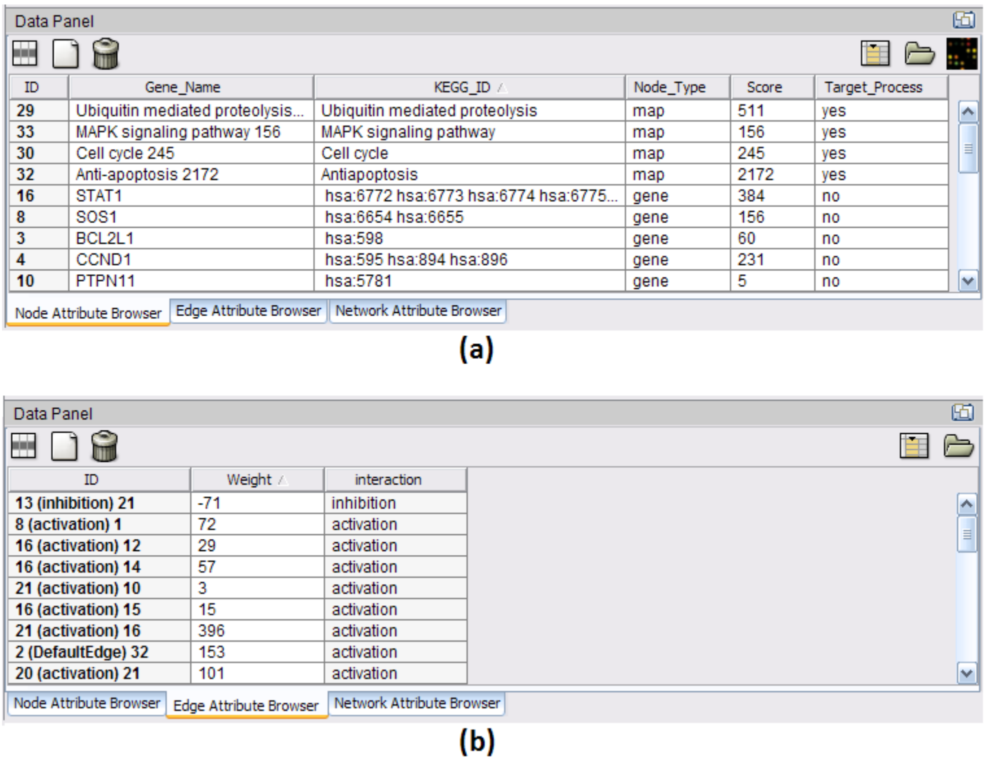
SiTSFlow algorithm assists end users to obtain quantitative measure to identify the most effected cellular process under the experimental setup. We have implemented the transduction score flow algorithm as Cytoscape plug-in to allow users to interactively visualize pathways and perform systematic analysis in a well known environment [93]. There are various functional plug-ins in open source Cytoscape software platform. The main plug-in categories in Cytoscape platform are as follows: analyzing existing networks, inferring new networks, functional enrichment of networks, and importing networks and attributes. If a user implements its algorithm in a Java based environment that would easily run on Cytoscape platform. The Java classes of developed plug-in can access the core data structures and windows of Cytoscape API. The pre-built classes and their methods in core API provide easy development of visual interfaces especially for complicated networks. The programmer might assign any type of shape, color, or value to each node, similarly sets weight, arrow, name features of each edge. Such node and edge properties might be dynamically updated according to the results of the user's algorithm. If someone wants to make its plug-in publicly available, who can upload .jar file of its plug-in in Cytoscape web site as well. Briefly, Cytoscape provides very comprehensive software platform for visualize and analyze very complex network structures. Therefore, we have decided to implement the SiTSFlow algorithm in Java-based Cytoscape environment.

In our Cytoscape plug-in, user can load the original pathways by using online KGML database of KEGG PATHWAY. Additionally, user can create a simulated network structure. The target processes or genes are marked by the user. An example to explain the attributes of nodes and edges is shown in Figure 4.6. The data panel given in Figure 4.6-a is *Node Attribute Browser*. Each node in the graph should contain unique node id, name (process or gene name), and KEGG Id (i.e., hsa:5595). The type of a node might be set to "gene" or "map" representing the processes. The target process flag of a node is set to "no" for genes, and "yes" for process. The score of each node is initially set to zero. The genes and target processes are represented by a circle and rectangle node shapes, respectively (see Figure 4.7). The data panel given in Figure 4.6-b is *Edge Attribute Browser*. Each edge has a unique id, weight (initially set to zero), and interaction type, i.e., "activation" or "inhibition".

In order to find out final activity scores of target processes and genes, a gene score file should be loaded to the environment. The score file should be a tab delimited text file. Each line of the score file contains three attributes: Entrez id of gene, name, and floating point formatted score. An example for gene score file is given in Table 4.9. After uploading of the gene score file, the signal transduction score flow algorithm is run over the given graph until obtaining convergence of node scores. After termination of scoring algorithm, final activity scores of genes and processes are mapped to the original graph. The activity scores of nodes are represented by different color tones of green, yellow and red colors. The score scale for 0-200, 201-900, and 901-3000 are represented by green, yellow and red colors respectively. This coloring scheme was designed to provide visualization facility for the significant paths and nodes in the graph. In order to analyze the final activity scores of genes and processes, user can save the final activity scores of each node in a tab delimited text file.

Table 4.9: An example for gene score file. It shows some examples from gene rank scores of ChIP-seq control sample for MCF7 cells.

Entrez Gene ID	Gene Name	Rank Score
1415	CRYBB2	0.470183
4437	MSH3	0.341743
4521	NUDT1	0.949541
4591	TRIM37	0.850917
5605	MAP2K2	0.908256



ID	Gene_Name	KEGG_ID	Node_Type	Score	Target_Process
29	Ubiquitin mediated proteolysis...	Ubiquitin mediated proteolysis	map	511	yes
33	MAPK signaling pathway 156	MAPK signaling pathway	map	156	yes
30	Cell cycle 245	Cell cycle	map	245	yes
32	Anti-apoptosis 2172	Antiapoptosis	map	2172	yes
16	STAT1	hsa:6772 hsa:6773 hsa:6774 hsa:6775...	gene	384	no
8	SOS1	hsa:6654 hsa:6655	gene	156	no
3	BCL2L1	hsa:598	gene	60	no
4	CCND1	hsa:595 hsa:894 hsa:896	gene	231	no
10	PTPN11	hsa:5781	gene	5	no

(a)

ID	Weight	interaction
13 (inhibition) 21	-71	inhibition
8 (activation) 1	72	activation
16 (activation) 12	29	activation
16 (activation) 14	57	activation
21 (activation) 10	3	activation
16 (activation) 15	15	activation
21 (activation) 16	396	activation
2 (DefaultEdge) 32	153	activation
20 (activation) 21	101	activation

(b)

Figure 4.6: The screenshot of *Data Panel* of Cytoscape. It displays node (a) and edge (b) attributes of hsa04630 - Jak-STAT signalling pathway after running of SiTSFlow algorithm.

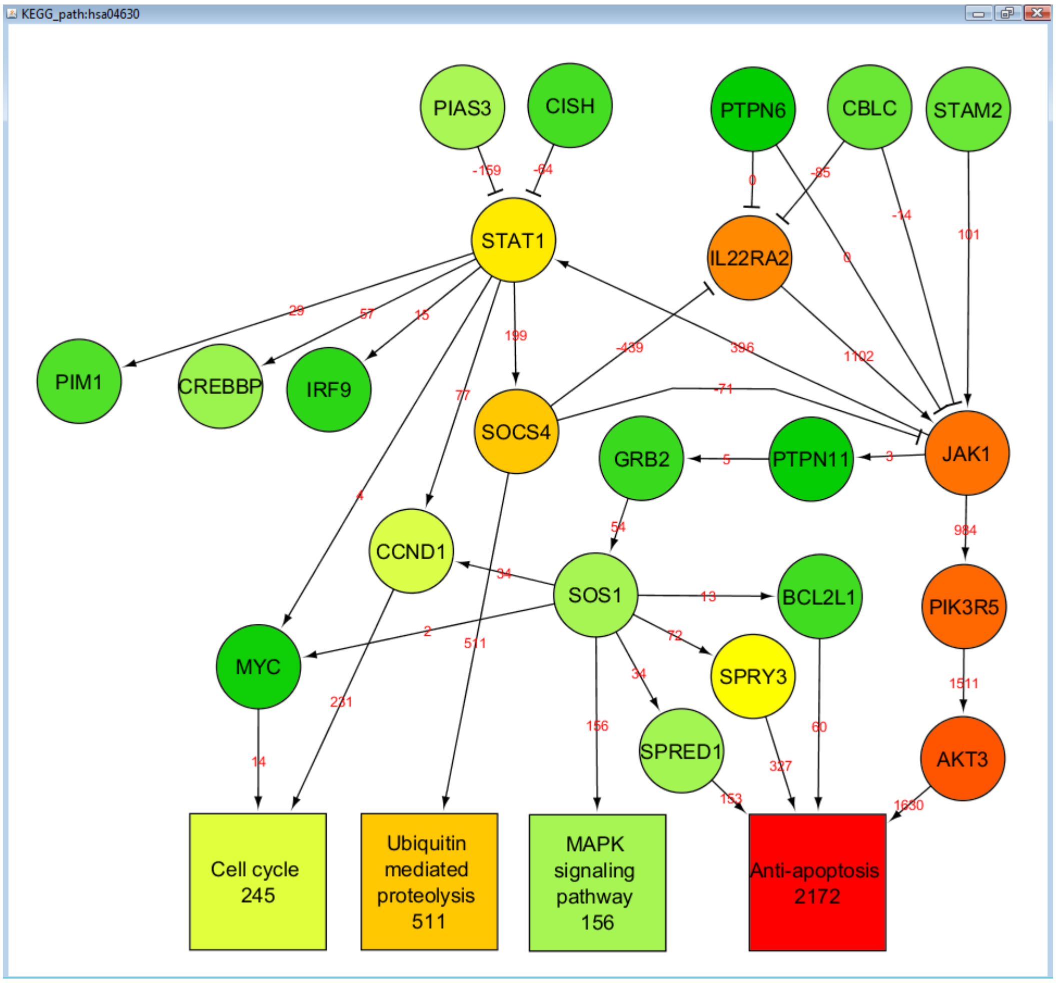


Figure 4.7: The screenshot of Cytoscape plug-in. It displays the activity scores of the nodes in hsa04630 - Jak-STAT signaling pathway using HeLa cells under oxidative stress condition. The genes and processes are represented by circles and rectangles, respectively. Node scores are represented in color tones of green, yellow, and red. The edges carry the out-score of a parent node to its children by multiplying out-score by 1 or -1 weight for activation or inhibition, respectively. The final activity of each process is given under its name.

4.10 Comparison with Enrichment Tools

We have compared the performance of several state of the art methods with that of network structure based pathway enrichment system. Signalling Pathway Impact Analysis (SPIA) method that was also one of the NSBA methods [21]. SPIA method combines the over-representation of differentially expressed genes in a pathway and the perturbation measure of that pathway which is computed by propagating gene expression changes across the network topology. The outputs of SPIA are: a general probability value P_G and status (i.e., activation or inhibition) information about the pathway. It does not provide an activity score for each node or process in the pathway. We applied SPIA method by using KRas data set since SPIA accepts only microarray data type. The fold-change ratios of control vs. G12Asp and control vs. G12Val sets were computed and given as the input of the SPIA method. However, using a 5% cutoff of the FDR adjusted p-values, the method was unable find any significant pathway in KEGG database. (see Table 4.10). Whereas based on the results obtained by SiTSFlow algorithm applied for gene knockout operation, the most affected biological processes in KRas data set were *Apoptosis* and *Angiogenesis*. Eventually, SPIA could not identify such a pathway related with the significant process identified by SiTSFlow algorithm. However, SiTSFlow algorithm provides the activity scores for all target biological processes of a given pathway, rather than giving single pathway impact score, since a pathway may contain several biological processes working for different cellular procedures. Our algorithm is based on the simulation of gene signal transduction inside the cell. Gene signals are provided by integrated scores not based on only differentially expressed genes information. The activity score computation for each process is performed by score signal transduction following the network topology strictly.

In order to compare performance of our system with a well-known gene set enrichment method, we applied GSEA on KRas data set. The original KRas gene expression data was given as the input to GSEA. The samples were compared with t-test statistics and the sorting of genes was performed based on the p-values computed in t-test. Two sets were constructed to use during the GSEA: control vs. G12Asp and control vs. G12Val. Based on GSEA results, only one gene set was significantly enriched at FDR < 25% threshold: *Reactome Apoptosis* pathway (see Table 4.11). The GSEA scores and enrichment plot of *Reactome Apoptosis* pathway are given in Table 4.12 and Figure 4.8, respectively. Consequently, the popular method

GSEA could not identify the other pathways related with KRas data.

In order to highlight the novelties of our system in transcriptome data analysis, we also applied *kegArray* tool [6] to gene expression sample of HeLa cells under oxidative stress over Jak-STAT signalling cascade (Figure 4.9). Several tools, similar to kegArray, map only expression data over pathways; however, they could not assign a score to the target biological process. However, our system provides better representation to observe responses of biological processes to given experimental conditions.

Finally, none of the approaches explains how they manage the pathway activity score computations for cyclic signalling pathways, since SiTSFlow algorithm shows convergence behavior for cyclic pathways as well.

Table 4.10: SPIA results on KRas data set using control vs. G12Val mutation. FDR and $FWER$ were calculated for P_G.

Kegg Pathway Name	P_{NDE}	P_{PERT}	P_G	P_{FDR}	P_{FWER}	**Status**
Alzheimer's disease	0.001	0.582	0.009	0.34336	0.63019	Activated
Vibrio cholerae infection	0.002	0.765	0.014	0.34336	0.98222	Activated
Pathogenic Escherichia coli infect.	0.007	0.591	0.027	0.34336	1	Activated
Chemokine sig. path.	0.127	0.033	0.027	0.34336	1	Inhibited
RIG-I-like receptor sig. path.	0.201	0.025	0.031	0.34336	1	Inhibited
Epithelial cell sig.	0.008	0.616	0.033	0.34336	1	Inhibited
Focal adhesion	0.011	0.580	0.038	0.34336	1	Inhibited
mTOR sig. path.	0.025	0.259	0.039	0.34336	1	Activated
Prion diseases	0.021	0.414	0.049	0.35096	1	Inhibited

Table 4.11: GSEA results on KRas data set using control vs. G12Asp mutation.

Gene Set Name	ES	NES	**Nom** *p-val*	**FDR** *q-val*	**FWER** *p-val*
Reactome Apoptosis	-0.84	-1.68	0.000	0.188	0.144
Reactome Intrinsic Pathway for Apoptosis	-0.84	-1.62	0.000	0.276	0.373
Pujana Brca1 Pcc Network	-0.69	-1.59	0.006	0.322	0.561
Reactome Activation of Bh3 only Proteins	-0.90	-1.58	0.007	0.328	0.676
Krige ResponseE to Tosedostat 6hr up	-0.82	-1.56	0.012	0.377	0.802

Table 4.12: GSEA detailed scores for *Reactome Apoptosis*.

Probe	Description	Rank in List	Rank Score	Running ES	Core Enrich.
842	CASP9	7	0.913	-0.037	No
572	BAD	16	0.462	-0.119	No
27113	BBC3	24	0.221	-0.203	No
581	BAX	27	0.173	-0.221	No
598	BCL2L1	32	0.144	-0.266	No
355	FAS	35	0.106	-0.288	No
10018	BCL2L11	37	0.057	-0.298	No
999	CDH1	47	-0.104	-0.418	No
596	BCL2	56	-0.285	-0.512	No
331	XIAP	80	-3.391	-0.609	Yes
5599	MAPK8	81	-3.755	-0.360	Yes
5366	PMAIP1	82	-5.409	8.15E-09	Yes

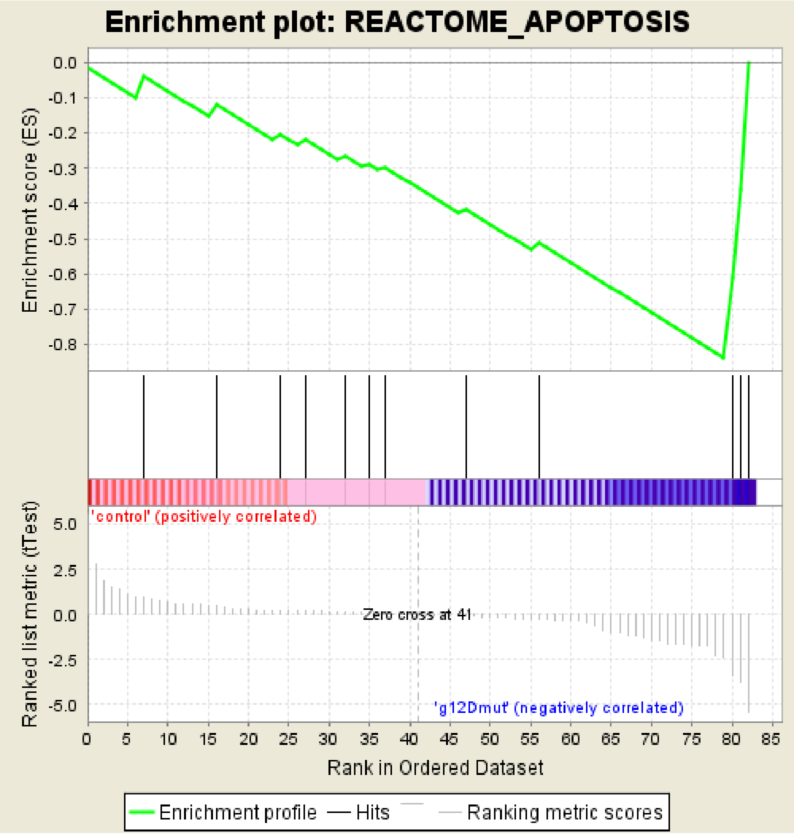


Figure 4.8: Enrichment plot of *Reactome Apoptosis*. The profile of the Running ES Score and positions of gene set members on the rank ordered list given in Table 4.12.

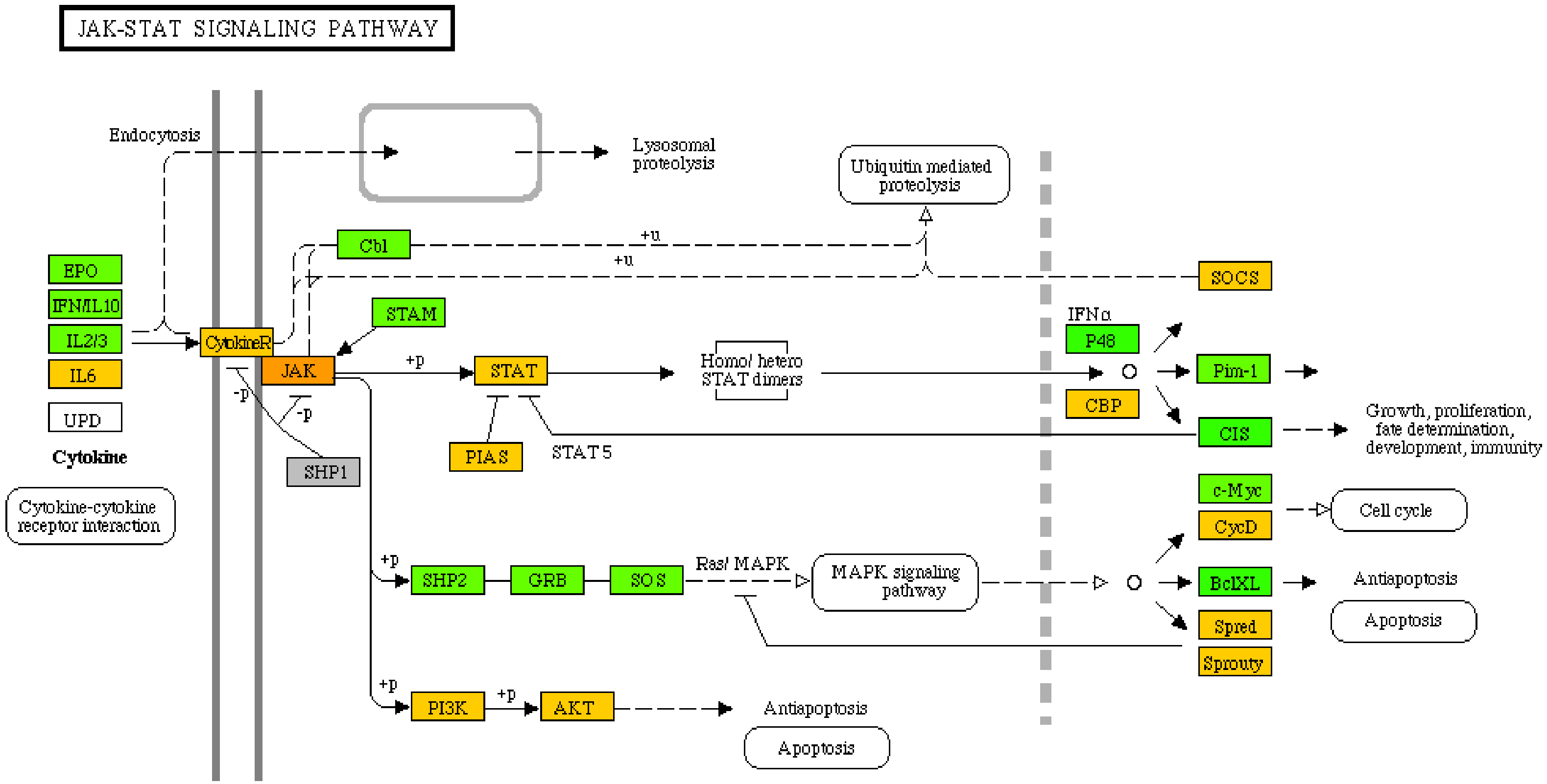


Figure 4.9: Gene expression sample of HeLa cells under oxidative stress was mapped onto Jak-STAT signalling pathway by using *kegArray* tool. Green and orange colors indicate down-regulation and up-regulation values, respectively.

CHAPTER 5

TOWARDS CONSTRUCTING GLOBAL SIGNALLING NETWORK

In this chapter, we describe and explain the merge algorithm developed for constructing a global signalling network. Features of this constructed network and its response to some of the data sets are also discussed.

5.1 Method Overview

The proposed method includes the merge algorithm applied for unification of small signalling pathways. Exploring various biological responses in a global network might be an interesting case, since a broader view of cell signalling mechanism would provide better interpretation for the questions asked during experiments. Therefore, we merged several small signalling pathways based on their common nodes. Merge algorithm is composed of two phases: *pre-processing* and *unification*. A simple example for merge algorithm is given in Figure 5.1. A pathway might contain several copies of a gene, and these copies are called as *clones*. In the pre-processing phase, the nodes having several clones are identified as *duplicated nodes* for both pathways. In this example, *Pathway 1* contains gene *X* as the duplicated node. If there exists such duplicated nodes in a pathway, they are represented by only a single node which encapsulates all relations of a duplication. In the proposed algorithm, unification phase might be considered as a variation of taking union of graph nodes and edges. Hence, in the unification phase, common nodes between two input pathways are identified. Both gene and process nodes might be marked as common nodes, so unification is performed for these node types. In the example shown in Figure 5.1, both pathways contain gene *B* and it constitutes the start point of union operation. While common nodes and their relations are preserved and transferred to new merged graph, remaining nodes and edges are also added to new merged graph. We iteratively applied this pairwise merge scheme for all pathways at hand. Finally, this global signalling network for the human cell is assessed by using SiTSFlow algorithm.

5.2 Merge Algorithm

The merge operation of several signalling pathways is performed by running iteratively Algorithm 6. At an iteration of Algorithm 6, two input pathways are unified into a new graph, while at the subsequent iteration of algorithm the unified graph and another pathway is merged. This pairwise merge strategy terminates when all input pathways are unified into a broader network.

The input of Algorithm 6 is two pathways represented by $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$. The output of Algorithm 6 is a union graph $\mathcal{G}_{merge}$ that is the unified version of $\mathcal{G}_1$ and $\mathcal{G}_2$. In order to reduce space complexity, we use hash tables indexed by gene identifiers to store entire node and edge relations of each input graph. For this purpose, $g1Hash$ and $g2Hash$ hash tables keep node and edge lists for $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively.

Algorithm 6 describes general steps of merge algorithm for given two input pathways: $\mathcal{G}_1$ and $\mathcal{G}_2$. At the *pre-processing phase*, Algorithm 7 is called which removes duplicated nodes of a given graph. The nodes having several clones that share the same gene name are identified as *duplicated nodes*. Identification of duplicated nodes is performed by checking Entrez gene identifier of each node in the graph. If there exists nodes having the same Entrez identifier, these are marked as *duplicated nodes*. A new graph is constructed by using only one node that encapsulates all relations of such duplication. Algorithm 7 performs identification and elimination of duplicated nodes by using a hash table. When there is a new node x, we check if it already exists or not in the hash table. If node x already exists in table, this new node x is marked as a duplicated node. The nodes in out-adjacency list of duplicated node x are added to that of already existing node x. When Algorithm 7 terminates, a new graph $\mathcal{G}_{new}$ is constructed by using unique node and edge relations from hash table. This $\mathcal{G}_{new}$ is returned as the output of Algorithm 7.

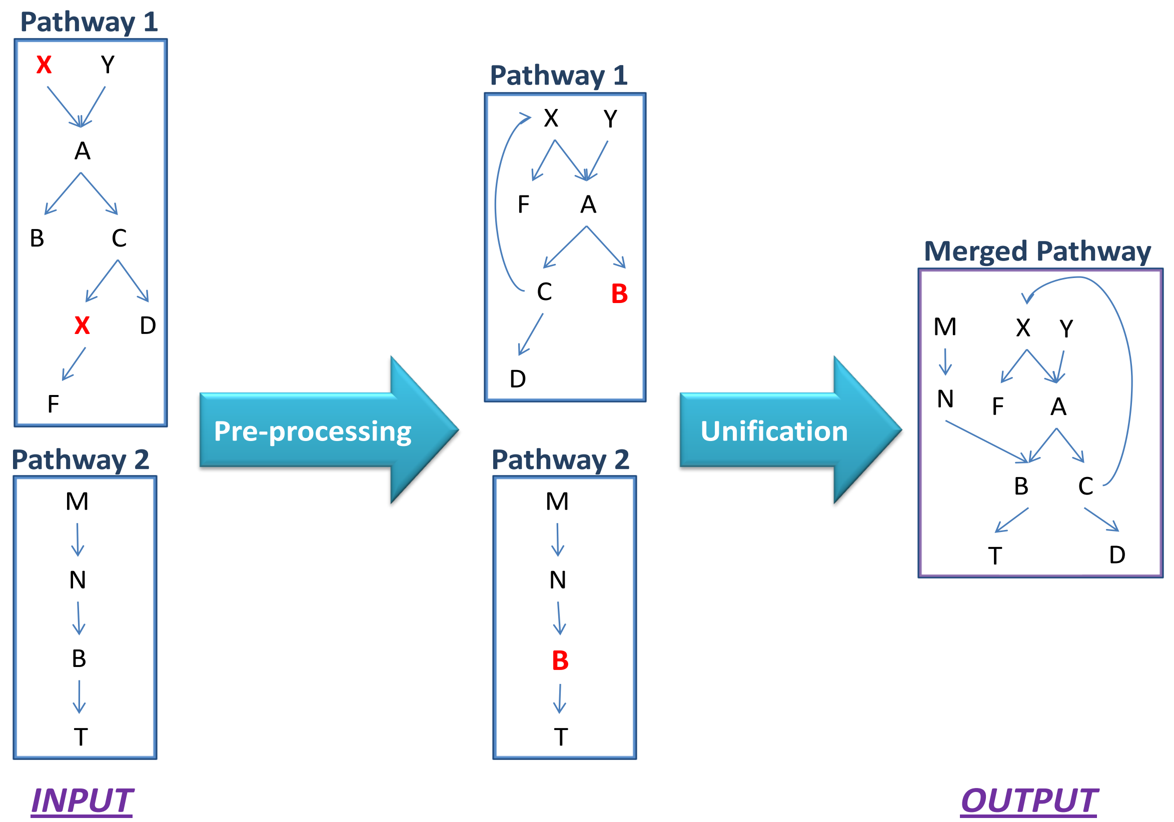


Figure 5.1: A simple example to explain merge algorithm for given pathways. In the pre-processing phase, the nodes having several clones sharing the same gene name are identified as *duplicated nodes* for both pathways. If there exists such duplicated nodes e.g., *X* in Pathway 1, only one node represents all relations of such duplications. The common nodes e.g., *B* between two input pathways are then identified. The unification phase operates the union of graph nodes and edges based on the common node(s). After performing unification, the constructed pathway is given as the output.

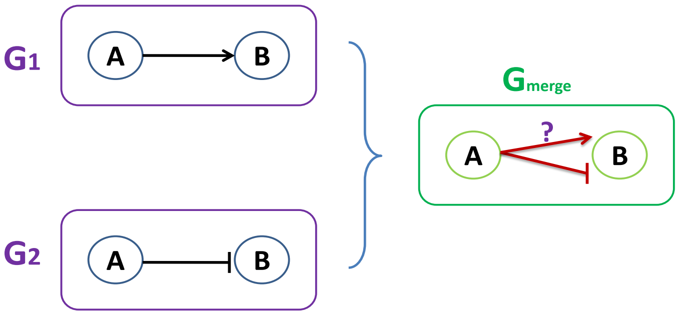


Figure 5.2: An example to explain conflicting edges problem in unification of two given graphs. Node A and B are common nodes between $\mathcal{G}_1$ and $\mathcal{G}_2$. Before performing unification, we should check the edge type between these nodes, since both nodes will appear in $\mathcal{G}_{merge}$ and type of relation between these nodes should be identical. Therefore, user decides which edge type will be assigned as the final edge relation of node A and B in $\mathcal{G}_{merge}$.

The second operation of *pre-processing phase* is the identification of common nodes between $\mathcal{G}_1$ and $\mathcal{G}_2$. Search operation is performed over $g1Hash$ and $g2Hash$ tables for $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. The nodes that share the same gene identifier are marked as *common nodes*. If search is successful, these nodes are kept in *commonNode* list. Unification two graphs is performed based on the nodes in *commonNode* list.

The last operation of *pre-processing phase* is the control of conflicting edge types between nodes in *commonNode* list. An example to explain conflicting edges problem is given in Figure 5.2. In our graphs, an edge type is set to activation or inhibition. In this example, node A and B are common nodes between $\mathcal{G}_1$ and $\mathcal{G}_2$. Before performing unification, we should check the edge type between these nodes. Both of these nodes will appear in $\mathcal{G}_{merge}$ and type of relation between these nodes should be identical. For this purpose, if the nodes in *commonNode* list are neighbors, the edges between such neighboring nodes are checked by using edge relation information taken from both $\mathcal{G}_1$ and $\mathcal{G}_2$. If there exists such a conflicting edge, the user is notified. The final decision is made by the user who assigns the final and identical edge type between node A and B in $\mathcal{G}_{merge}$.

The graph *unification phase* starts by creating an identical copy, called as $\mathcal{G}_{merge}$, of input $\mathcal{G}_1$. Unification might be considered as a variation of taking union of node and edge sets of two input graphs. In order to reduce running time of unification phase, we only add remaining nodes from $\mathcal{G}_2$ onto new $\mathcal{G}_{merge}$. For this purpose, we run for-loop in *unification phase* of Algorithm 6 that adds each node x in $\mathcal{V}_2$ and not in *commonNode* list into the new $\mathcal{G}_{merge}$. The nodes in out-adjacency of x and their edge types are also added into $\mathcal{G}_{merge}$.

Algorithm 6 : MergePathways $(\mathcal{G}_1, \mathcal{G}_2)$

```
Input:

Directed graph $\mathcal{G}_i$
$outAdj_i(x)$: out-adjacency list of node $x$ in graph $i$
$g1Hash$, $g2Hash$ : hash tables to keep node and edge lists for $\mathcal{G}_1$ and $\mathcal{G}_2$
$sign$: keeps edge types: activation (1) or inhibition (-1)
$commonNode$: keep id of common nodes between $\mathcal{G}_1$ and $\mathcal{G}_2$

Pre-processing Phase:

RemoveDuplicateNodes ($\mathcal{G}_1$)
RemoveDuplicateNodes ($\mathcal{G}_2$)

// identification of common nodes

for each vertex $x \in \mathcal{V}_1$ do
  if $isElement(ID(x), g2Hash)$ then
    $add\ (ID(x), commonNode)$   {add common node id to $commonNode$ list}

// control of conflicting edge types

for both $\mathcal{G}_1$ and $\mathcal{G}_2$ do
  Check conflicting edge types between $x \in commonNode$ and $y \in commonNode$

Unification Phase:

$\mathcal{G}_{merge} \leftarrow \mathcal{G}_1$   {Make a copy of the $\mathcal{G}_1$}
for each vertex $x \in \mathcal{V}_2$ do
  if $x \notin commonNode$ then
    $add\ (x, \mathcal{V}_{merge})$
    for each vertex $y \in outAdj_2(x)$ do
      $add\ (y, outAdj_{merge}(x))$   {edge relation of node $y$ is added to node $x$ relations in $\mathcal{G}_{merge}$}
      $add\ (sign(x, y), \mathcal{E}_{merge})$   {edge type between $x$ and $y$ is added to $\mathcal{E}_{merge}$}

Output:

$return\ \mathcal{G}_{merge}$
```

Algorithm 7 : RemoveDuplicateNodes $(\mathcal{G})$

```
Input:

Directed graph $\mathcal{G}$
$outAdj(x)$: out-adjacency list of node $x$
$ID$: gene id list of nodes in graph $\mathcal{G}$
$newHash$: hash table to keep unique node information and edge relation

for each vertex $x \in \mathcal{V}$ do
  if not $isElement(ID(x), newHash)$ then
    $add(ID(x), outAdj(x), newHash)$   {add gene id and edge relations of node $x$ in hash table}
  else
    $y = getElement(ID(x), newHash)$ {get information of duplicated node $x$ from hash table}
    for each vertex $k \in outAdj(x)$ do
      $add\ (k, outAdj(y))$   {edge relation of node $k$ is added to node $y$ relation set}
    $update(ID(y), outAdj(y), newHash)$   {update information of node $y$ in hash table}

Output:

$\mathcal{G}_{new} \leftarrow$ reconstruct node and edge relations from $newHash$ table
$return\ (\mathcal{G}_{new})$
```

5.3 Calculation of Significance and Sensitivity of Activity Scores

We designed the same permutation tests explained in Section 3.6 to evaluate significance and sensitivity of activity scores obtained by running of SiTSFlow algorithm on the new global network. After performing 50% percent shuffling in permutation procedure, the significance value i.e., α_{value} of each activity score is calculated by using Equation 3.10. The sensitivity value i.e., σ_{value} of each activity score is calculated by using Equation 3.12.

5.4 Computational Complexity

The computational complexity of Algorithm 6 involves the running time of Algorithm 7 and other pre-processing steps. Algorithm 7 performs elimination of duplicated nodes by using a hash table structure. The for-loop iterates over entire node set, so it runs $O(\mathcal{V})$ times. For each new node x, we control if it exists or not in hash table, so each control operation takes $O(1)$ time. Similarly, add and update operations in hash table also takes $O(1)$ time. The inner for-loop runs for each edge of $outAdj(x)$, in worst case, the total time spent in this loop is $O(\mathcal{E})$. Total running time for Algorithm 7 is $O(\mathcal{V} + \mathcal{E})$.

The identification of common nodes between $\mathcal{G}_1$ and $\mathcal{G}_2$ runs in the size of node set of $\mathcal{G}_1$. For each new node x, we control if there is gene with the same gene identifier or not in the hash table, so each check operation takes $O(1)$ time. The identification of common nodes totally takes $O(\mathcal{V}_1)$ time. The check of conflicting edge types runs for total number of edges that are adjacent to the nodes in *commonNode* list. In the worst case, it runs for all edges in a graph, so checking of conflicting edge types operation takes at most $O(\mathcal{E}_1)$ time.

The *unification phase* is performed by considering $\mathcal{G}_2$. The input $\mathcal{G}_1$ is identically copied into new $\mathcal{G}_{merge}$, so creation of a new graph by using hash table structure might be performed in constant time. The for-loop of *unification phase* runs for each node in $\mathcal{G}_2$, so it takes $O(\mathcal{V}_2)$ time. The inner for-loop runs for each edge of $outAdj_2(x)$, in worst case, the total time spent in this part is $O(\mathcal{E}_2)$. Add operations to new $\mathcal{V}_{merge}$ and $\mathcal{E}_{merge}$ sets takes $O(1)$ time. Total running time of *unification phase* is $O(\mathcal{V}_2 + \mathcal{E}_2)$.

Final running time for Algorithm 6 is in linear-time in the size of the pathways $\mathcal{G}_1$ and $\mathcal{G}_2$, that is $O(\mathcal{V}_1 + \mathcal{E}_1 + \mathcal{V}_2 + \mathcal{E}_2)$.

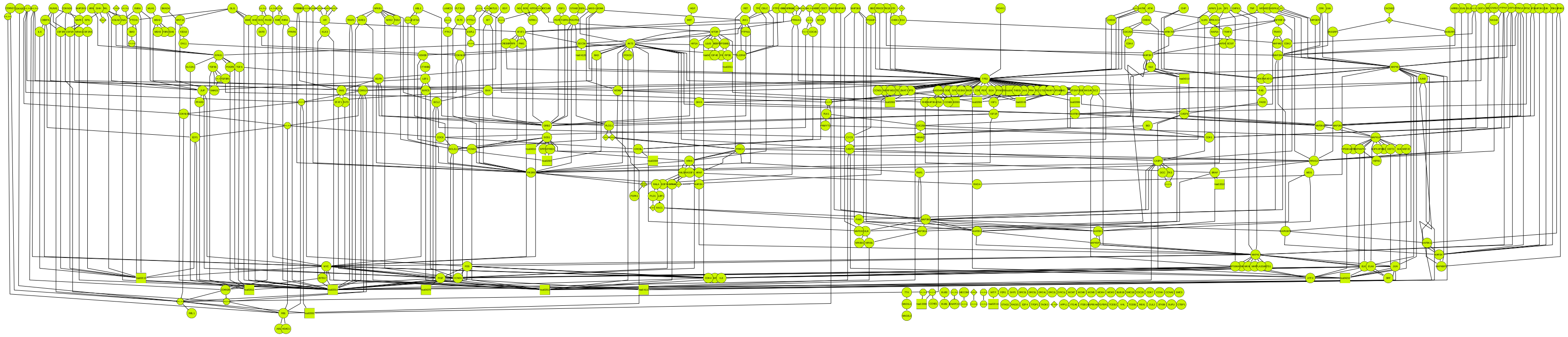


Figure 5.3: The screenshot of the global signalling network that contains 450 nodes, 650 edges, and 24 biological processes.

5.5 Experimental Results

The selected input pathways are *Cell cycle*, *Jak-STAT signalling*, *MAPK signalling*, *mTOR signalling*, *Pathways in cancer*, and *P53 signalling*. Table 4.2 summarizes the total number of nodes, genes, and processes contained in each pathway. Finally, sequential merge of six different pathways results in a larger global signalling network composed of 450 nodes, 650 edges, and 24 biological processes. The screenshot of the global signalling network is given in Figure 5.3.

The aim of construction such a large signalling network was to explore collective working mechanism of several processes and to observe divergent responses of specific processes at cell signalling level. For this purpose, new global signalling network was evaluated by SiTS-Flow algorithm based on four different data sets: HeLa cells under oxidative stress, Estradiol (E2) treated MCF7 cells, Estrogen Receptor (ER) beta treated U2OS cells, and KRas data.

Gene ranking scores obtained from microarray and ChIP-seq experiments of HeLa cells under oxidative stress were integrated to compute the self-score of each gene. SiTSFlow algorithm was applied on the global network with these scores. SiTSFlow algorithm performed 15 iterations over the entire cyclic graph until verifying the convergence threshold. Activity score of each process in global network is given in Table 5.1. When the total activity scores of target biological processes were compared, *Proliferation* process had the highest score of 5614 under the oxidative stress condition (Table 5.1). If the confidence threshold of α_{value} was set to 0.1, there was only one significant processes, *DNA biosynthesis*, out of 20 target processes. Based on σ_{value} assessment criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data. From the biological perspective, the responses of most of the processes were correlated with their activities in original pathways (see results given in Section 4.3). In other words, processes in global signalling network gave the similar biological responses with the processes of individual KEGG pathways. For example, *Anti-apoptosis* process in global network had higher activity score under oxidative stress condition (see first row of Table 5.1), similarly *Anti-apoptosis* process in *Jak-STAT* pathway has provided the same response to oxidative stress (see thirtieth row of Table 4.3). Another interesting example is related with processes having divergent behaviors under same conditions. *Anti-apoptosis* and *Evading apoptosis* processes provide replication of cells and both of these processes were dominated on oxidative stress condition in global network (see first and ninth rows of Table 5.1). However *Apoptosis* is the process of cell death, so it has divergent function in cell signalling, eventually it had domination on control sample in global network (see second row of Table 5.1). In other words, some processes represented divergent biological activities in global network and the opposite function of such processes was proved in literature.

SiTSFlow algorithm was executed on the global network by using gene scores obtained in Estradiol (E2) treated MCF7 cells. SiTSFlow algorithm performed 17 iterations over the entire cyclic graph until verifying the convergence threshold. Activity score of each process in global network is given in Table 5.2. When the total activity scores of target biological processes were compared, *Proliferation* process had the highest score of 6894 under E2 condition (Table 5.2). If the confidence threshold of α_{value} was set to 0.1, there was 5 significant processes (*Cell cycle arrest*, *DNA biosynthesis*, *Inhibition of IGF1 / mTOR pathway*, *Resistance to chemotherapy*, *Ubiquitin mediated proteolysis*) out of 23 target processes. Based on σ_{value} assessment criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data. Processes in the global signalling network gave the similar biological responses with the processes in original pathways of KEGG Database (see results given in Section 4.4). *Anti-apoptosis* and *Evading apoptosis* processes were dominated on E2 condition in global network (see first and tenth rows of Table 5.2). Although *Differentiation* process represented high activity score under control sample, *Block of differentiation* process had domination on E2 sample in global network (see seventh and third rows of Table 5.2). Similarly, *Cell cycle* process was more active in E2 sample, however, *Cell cycle arrest* process was activated in control sample. Therefore, the biological activities are divergent for some processes, *Differentiation* vs. *Block of differentiation* or *Cell cycle* vs. *Cell cycle arrest*, that have also opposite functions in cell signalling. Eventually, this fact was experimentally proved by SiTSFlow algorithm.

Another experiment was performed by using gene scores obtained in Estrogen Receptor (ER) beta treated U2OS cells. SiTSFlow algorithm performed 15 iterations over the entire cyclic graph until verifying the convergence threshold. Activity score of each process in global network is given in Table 5.3. When the total activity scores of target biological processes were compared, *Proliferation* process had the highest score of 6997 under E2 condition (Table 5.3). If the confidence threshold of α_{value} was set to 0.1, there was 4 significant processes (*DNA biosynthesis*, *DNA repair damage prevention*, *p53 signalling*, *Regulation of autophagy*) out of 23 target processes. Based on σ_{value} assessment criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data. From the biological perspective, the biological responses of processes in the global signalling network provided very similar responses with processes of original pathways (see results given in Section 4.5). Although *Anti-apoptosis* process represented high activity score under control sample, *Apoptosis* process had domination on E2 sample in global network (see first and second rows of Table 5.3). Similarly, *Cell cycle* process is more active in E2 sample, however, *Cell cycle arrest* process is activated in control sample (see fourth and fifth rows of Table 5.3). Therefore, SiTSFlow algorithm experimentally proved divergent responses of specific biological processes that have functions during the working mechanism of cell signalling.

SiTSFlow algorithm was applied on the global network by using gene scores obtained in KRas data. Activity scores of processes in global signalling network for control, Gly12Asp, and Gly12Val samples of KRas microarray data. The significance value of activity score under each sample is given by calculation of α_{value} and σ_{value}. SiTSFlow algorithm performed 16 iterations over the entire cyclic graph until verifying the convergence threshold. Activity score of each process in global network is given in Table 5.4. If the confidence threshold of α_{value} was set to 0.1, there was 8 significant processes (*Apoptosis*, *Block of differentiation*, *Cell cycle*, *Evading apoptosis*, *p53 signalling*, *Regulation of autophagy*, *Resistance to chemotherapy*, *S-phase proteins*) out of 23 target processes. Based on σ_{value} assessment criteria, almost all of the processes have remained their score consistencies even if for 50% shuffling of the input data. Biological responses of specific processes in the global signalling network provided very similar responses with processes of original pathways (see results given in Section 4.7). For example, *Cell cycle* process in original Akt pathway has been down-regulated in both Gly12Asp and Gly12Val mutations compared to control sample (see third row of Table 4.7). Similarly, *Cell cycle* process in global network was significantly down-regulated in both Gly12Asp and Gly12Val mutations (see fourth row of Table 5.4). This fact proves the hypothesis of global signalling network that was providing of easy interpretation of complex biological phenomena in a large signalling network. Although *Evading apoptosis* process was down-regulated in both Gly12Asp and Gly12Val mutations compared to control sample, *Apoptosis* process was up-regulated on control sample in global network (see ninth and second rows of Table 5.4). Similarly, *Differentiation* process was up-regulated in both Gly12Asp and Gly12Val mutations, however, *Block of differentiation* process was up-regulated in control sample (see third and sixth rows of Table 5.4). In Gly12Asp and Gly12Val mutations, an increase in *Apoptosis* and a decrease in *Evading apoptosis* were expected [94], since the results are consistent with the apoptosis-promoting role of the tumor suppressor p53. *P53 signaling pathway* was increased significantly in Gly12Asp and Gly12Val mutations. BRAF mutation has been shown to confer resistance to chemotherapy [95]. Therefore, in our analysis, where all tumors already express mutant BRAF, even control sample containing tumors have high scores for *Resistance to chemotherapy*. Mutation in KRAS oncogene has also been shown to be a predictive marker of resistance to EGFR-targeted therapy [96]. It is known that concomitant KRAS and BRAF mutations rarely occur especially in the early stages of tumors, which might explain the reduced resistance in Gly12Asp mutation compared to control sample.

5.6 Discussion

A global signalling network for human cell was constructed by running iteratively proposed pathway merge algorithm. Constructed global network was assessed by using SiTSFlow algorithm. We demonstrated the convergence of the activity scores of processes in global signalling network. Experiment specific significant processes were identified by SiTSFlow algorithm and the significant processes were also correlated with our previous results and literature. These results proved the assessment capacity of SiTSFlow algorithm even for very complex signalling networks.

Activity scores of processes in global network represented analogous biological behaviors with the individual KEGG pathways. Therefore, we can derive that, unification of several individual pathways provides an opportunity to observe how complex biological traits arise and propagate in the cell. Thus, application of the SiTSFlow algorithm on a global signalling network has been successfully performed.

Table 5.1: Activity scores of processes in global signalling network for control and oxidative stress samples in HeLa cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face.

Biological Process	**Activity Scores of Target Process**		**Significance Scores**	
	Control Sample	Oxidative Stress	α_{value}	σ_{value}
Anti-apoptosis	4581	**4798**	0.446	0.467
Apoptosis	**749**	599	0.407	0.326
Block of differentiation	**997**	653	0.382	0.250
Cell cycle	**1844**	1466	0.347	0.270
Cell growth	**223**	52	0.186	0.042
Differentiation	**2521**	1756	0.130	0.091
DNA biosynthesis	1371	**1557**	0.092	0.105
DNA repair and damage prevention	**354**	419	0.139	0.165
Evading apoptosis	1726	**1737**	0.460	0.463
Inhibition of angiogenesis and metastasis	**272**	248	0.475	0.435
Inhibition of IGF1 / mTOR pathway	**112**	75	0.454	0.303
MAPK signalling	**1540**	1154	0.136	0.102
P53 negative feedback	791	**844**	0.438	0.467
Proliferation	4572	**5614**	0.171	0.210
Regulation of autophagy	29	**97**	0.304	1.454
Resistance to chemotherapy	**84**	71	0.500	0.423
S-phase proteins	19	**51**	0.333	0.907
Sustained angiogenesis	**2259**	2005	0.364	0.323
Ubiquitin mediated proteolysis	**651**	540	0.149	0.124
VEGF signalling	1397	**1421**	0.450	0.458

Table 5.2: Activity scores of processes in global signalling network for control and E2 samples in MCF7 cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face.

Biological Process	**Activity Scores of Target Process**		**Significance Scores**	
	Control Sample	E2 Experiment	α_{value}	σ_{value}
Anti-apoptosis	4997	**6001**	0.168	0.202
Apoptosis	1080	**1286**	0.311	0.370
Block of differentiation	292	**520**	0.227	0.490
Cell cycle	2312	**2572**	0.343	0.381
Cell cycle arrest	**96**	56	0.073	0.042
Cell growth	**52**	28	0.477	0.256
Differentiation	**1001**	679	0.285	0.193
DNA biosynthesis	1685	**1900**	0.088	0.099
DNA repair and damage prevention	553	**658**	0.256	0.305
Evading apoptosis	3569	**3652**	0.461	0.472
Exosome mediated secretion	114	**115**	0.423	0.427
Inhibition of angiogenesis and metastasis	318	**398**	0.200	0.250
Inhibition of IGF1 / mTOR pathway	63	**128**	0.094	0.193
MAPK signalling	1785	**1896**	0.339	0.360
P53 negative feedback	**667**	580	0.364	0.317
p53 signalling	415	**541**	0.132	0.172
Proliferation	6329	**6894**	0.250	0.272
Regulation of autophagy	103	**147**	0.473	0.674
Resistance to chemotherapy	30	**97**	0.001	0.003
Sustained angiogenesis	1957	**2086**	0.431	0.460
Ubiquitin mediated proteolysis	**929**	407	0.008	0.004
VEGF signalling	1360	**1432**	0.471	0.496
Wnt signalling	38	**128**	0.442	1.496

Table 5.3: Activity scores of processes in global signalling network for ERb and E2 samples in U2OS cells. α_{value} is obtained by applying permutation test. σ_{value} is calculated by using variance of activity scores in permutation test. Significant activity score of each process is marked by bold face.

Biological Process	**Activity Scores of Target Process**		**Significance Scores**	
	ERb	E2	α_{value}	σ_{value}
Anti-apoptosis	**6584**	6513	0.423	0.418
Apoptosis	1152	**1740**	0.173	0.261
Block of differentiation	870	**981**	0.327	0.368
Cell cycle	2436	**2905**	0.277	0.330
Cell cycle arrest	**235**	193	0.306	0.351
Cell growth	31	**67**	0.430	0.922
Differentiation	954	**1397**	0.423	0.620
DNA biosynthesis	**2464**	2182	0.086	0.076
DNA repair and damage prevention	634	**884**	0.087	0.121
Evading apoptosis	2827	**3710**	0.103	0.135
Exosome mediated secretion	**50**	26	0.404	0.213
Inhibition of angiogenesis and metastasis	374	**418**	0.387	0.433
Inhibition of IGF1 / mTOR pathway	**127**	86	0.383	0.259
MAPK signalling	**2622**	2617	0.482	0.481
P53 negative feedback	**1163**	701	0.121	0.073
p53 signalling	184	**629**	0.099	0.338
Proliferation	6287	**6997**	0.299	0.333
Regulation of autophagy	25	**146**	0.098	0.569
Resistance to chemotherapy	55	**69**	0.500	0.626
Sustained angiogenesis	2345	**2350**	0.426	0.427
Ubiquitin mediated proteolysis	700	**828**	0.362	0.428
VEGF signalling	1421	**1516**	0.368	0.392
Wnt signalling	**174**	122	0.127	0.089

Table 5.4: Activity scores of processes in global signalling network for control (C), Gly12Asp (D), and Gly12Val (V) samples of KRas data. The significance value of each score is specified by α_{value} and σ_{value} at the right column of its score. Significant activity score of each process is marked by bold face.

Biological Process	**Activity Scores of Target Process**								
	C	α_C	σ_C	D	α_D	σ_D	V	α_V	σ_V
Anti-apoptosis	5568	0.151	0.039	4879	0.140	0.036	5850	0.123	0.042
Apoptosis	**1436**	0.073	0.462	1756	0.116	0.352	**1466**	0.012	0.388
Block of differentiation	967	0.277	0.146	**509**	0.056	0.148	727	0.277	0.146
Cell cycle	**2933**	0.096	0.205	2660	0.102	0.291	**2736**	0.018	0.195
Cell growth	228	0.178	0.292	142	0.196	0.322	89	0.164	0.210
Differentiation	1080	0.285	0.103	1352	0.144	0.127	1805	0.136	0.137
DNA biosynthesis	2583	0.138	0.058	2291	0.182	0.049	2642	0.126	0.053
DNA repair and damage prevention	421	0.194	0.016	742	0.339	0.030	742	0.216	0.028
Evading apoptosis	**3624**	0.094	0.394	**3351**	0.052	0.209	**3131**	0.012	0.178
Exosome mediated secretion	70	0.194	0.054	31	0.339	0.013	39	0.216	0.016
Inhibition of angiogenesis and metastasis	274	0.194	0.035	238	0.339	0.024	190	0.216	0.016
Inhibition of IGF1 / mTOR pathway	87	0.180	0.521	108	0.222	0.492	174	0.170	0.379
MAPK signalling	1828	0.495	0.122	1767	0.481	0.156	2269	0.379	0.118
P53 negative feedback	736	0.194	0.033	760	0.138	0.020	781	0.158	0.022
P53 signalling	332	0.106	0.483	580	0.198	0.494	**460**	0.096	0.471
Proliferation	6003	0.250	0.020	5767	0.180	0.019	6503	0.269	0.020
Regulation of autophagy	**47**	0.084	1.086	**85**	0.068	0.064	**8**	0.056	0.063
Resistance to chemotherapy	**97**	0.002	0.001	**7**	0.002	0.014	**92**	0.002	0.001
S-phase proteins	**77**	0.002	0.002	**105**	0.002	0.001	**53**	0.002	0.002
Sustained angiogenesis	2645	0.267	0.043	1880	0.256	0.058	2282	0.449	0.051
Ubiquitin mediated proteolysis	363	0.152	0.050	805	0.166	0.068	707	0.158	0.045
VEGF signalling	1917	0.361	0.064	1190	0.124	0.112	1575	0.351	0.091
Wnt signalling	98	0.345	0.318	132	0.317	0.446	84	0.259	0.386

CHAPTER 6

CONCLUSION

Recent advances in high-throughput technologies allow researchers to investigate several organisms by using genomics, transcriptome, proteomics or metabolomics large scale data. Researchers should develop new computational methods for integration, visualization, and analysis of multiple high-throughput data to answer complex biological phenomena. Computational analysis of these high-throughput technologies usually generates significant gene lists specific to experimental conditions. However, the growth of high-throughput data revealed the need for data integration during the analysis. Therefore, in order to explore a biological interpretation for such gene lists, the next step of the analysis is the association of these genes with known biological molecular or signalling networks. Thus, an enrichment process attempts to connect the significant genes with their potential biological roles through known biological pathways. Most of the methods perform pathway enrichment based on either significant gene sets or gene functional class identifications and they do not provide quantitative measure to lead assessing biological activity of a specific cellular process. Although contemplating pathway topological information and transcriptome data empowers the analysis and upgrades it to system level with both model and data, this approach has not been adequately investigated and exploited.

Machine learning research generally deals with classification or clustering of any type of data. However, the recent trend in computer science research is application of various graphical models and their corresponding solutions by spectral graph algorithms for the development of internet search engines, image segmentation, social network analysis, biological network analysis etc. For example, *PageRank* is a sophisticated algorithm used by the Google search engine that assigns a rank value for a web page to represent its relative importance within the graph created by all World Wide Web pages [97]. Image segmentation can be represented as a graph partitioning problem and *Normalized cut* approach provides a global measure for segmenting the given graph [98]. Therefore, spectral graph algorithms could easily be applied for new problems in last decades.

The described network structure based pathway enrichment system fuses and exploits transcriptome data and pathway model effectively benefiting from topological information brought in by pathway models. A score flow algorithm, *SiTSFlow* has been designed and implemented for quantitatively assessing biological activities of specific cellular processes and identifying significant paths in a pathway. The first phase of the described system is data integration in which transcriptome data is incorporated by taking the rank products of individual scores of the employed data sources. The original signalling pathway is converted into a cascaded structure by applying a linear-time graph cascading algorithm, since there might be many cyclic paths in signalling pathways. The individual gene scores are then mapped onto the nodes of cascaded graph. SiTSFlow algorithm simulates signal transduction inside the cell. Therefore, the gene scores are transferred over the nodes by traversing the path until a pre-defined target biological process is attained. Because of cyclic paths, we carry out iterations and when the scores converge, a final activity score is assigned to the pre-defined target biological process. By analysis of final activity scores of processes, user can find out related paths that would respond biological questions enquired at the design stage of transcriptome experiments. Experiment specific significant processes and paths that were identified by described system were also validated based on the information extracted from previous studies in literature. The convergence of final activity scores was also demonstrated for several cyclic pathways of KEGG PATHWAY Database. Hence, we have managed to develop a linear time score flow algorithm converging in limited number of iterations on a cyclic graph. These results proved that network structure based pathway enrichment system provides a powerful assessment tool for the user. Furthermore, SiTSFlow algorithm was implemented as Cytoscape plug-in. By using this plug-in, user can both interactively visualize pathways and apply SiTSFlow algorithm different pathways and data sources.

In the second part of the thesis, several small-size pathways have been unified and a global signalling network for human cell has been constructed. This global network was constructed by iteratively running of described pathway merge algorithm which runs in linear-time in the size of input pathways. The global signalling network was evaluated by applying SiTSFlow algorithm. Final activity scores of processes in global network converged in limited number of iterations. Activity scores of processes in global network represented analogous biological behaviors with the individual KEGG pathways and related literature studies. Thus, application of the SiTSFlow algorithm on a global signalling network has been successfully performed. Unification of several individual pathways can help to explain how complex biological traits arise and propagate in cell signalling.

The work in this thesis can be extended in several directions. SiTSFlow algorithm calculates activity score of processes by applying a deterministic score flow mechanism. On the other hand, probabilistic approaches might be adapted to assess the activities of biological processes in pathways. However, a general probabilistic graph model, *Bayesian networks*, works on directed acyclic graphs (DAG), since exact inference requires acyclic graph structure in which joint probabilities can be defined in terms of the product of conditional probabilities of nodes [99]. Due to cyclic biological pathways, we did not consider such a probabilistic model during the development of SiTSFlow algorithm. However, there exist extensions of Bayesian Networks to calculate inference of nodes in cyclic graphs. *Dynamic Bayesian Network* (DBN) might be used to estimate biological activity of processes in a pathway. DBN represents the relations between time series variables. DBNs have already been applied to represent cyclic gene networks [100, 101]. Actually, DBN is a basic type of *Hidden Markov Models* and it satisfies *first order Markov* property implying that the state of a system at time i only depends on its state at time $i-1$. By using this state dependence property, the joint probability of a cyclic graph network can be calculated by taking product of all individual conditional probabilities of each node given its parent nodes. We might adopt the original DBN approach to apply in cyclic pathways. In the modified version, time intervals can be used to represent cycles of a pathway. Experimental data employed in this thesis was not a time series one, therefore we might use time intervals of DBNs to represent the original BFS levels of a given cyclic pathway. After applying described graph cascading algorithm, each time interval can contain all of nodes of the given pathway and the length of time intervals can be equal to the highest BFS level of the given pathway. The edge relations of the given cyclic pathway might be established between the associated time intervals, i.e., BFS levels. For example, let assume that an edge e_{km} between the node k and m generates a cycle in the given graph, in modified DBN approach, the e_{km} is placed between time interval $i-1$ and i that represent BFS levels of the node k and m, respectively. The cyclic paths can be eliminated by applying this adaptation, since back edges linking to the lower time intervals are not allowed in the new representation. Convergence of conditional probability of a target process node can be achieved by passing over the time intervals with several iterations. However, convergence of conditional probabilities would not be satisfied by applying this iterative calculation, since *partitioned score transfer* method of SiTSFlow algorithm would not be represented in this adaptation of DBNs. Besides, the edge types, i.e., activation or inhibition, should be integrated in conditional probability calculation. For this purpose, a prior probability can be set to represent the weight of each edge type. By applying all these modifications, we might apply DBNs and calculate the conditional probability of each gene and process in a cyclic pathway.

Loopy Belief Propagation and *Junction Tree Algorithm* are alternative approaches to calculate approximate or exact marginal probabilities on cyclic graphs. Loopy Belief Propagation adjusts original *Belief propagation* algorithm to able to apply on cyclic graphs [102]. Belief propagation is a message passing algorithm to calculate exact inference on general graph models. The marginal probability of a variable node x is calculated by the product of all the incoming messages arriving at node x. Each of these messages is computed recursively in terms of other messages until node x has received messages from all of its neighbors. For cyclic graphs, in initialization step, all variable messages are set to unit function and all messages are passed across every edge in each direction. This message pass and update mechanism is performed at every iteration. In cyclic graphs, the algorithm converges when pending of all messages is finished. However, the algorithm may not converge in a reasonable time, so it might be terminated by user. The approximate marginal distribution of node x is then computed by using the product of last received incoming messages to node x. It is proved that graphs containing only one cycle converges to calculate the exact inference. Pathways might contain several cyclic paths, thus Loopy Belief Propagation would not provide convergence for exact inference of the marginal probabilities in such cyclic pathways. *Junction Tree Algorithm* is a method to calculate exact marginal probabilities in general graphs [103]. The algorithm can be also applicable on cyclic graphs which are transformed into DAGs. It is initiated by conversion of a directed graph into an undirected graph. In order to create a *junction tree*, each cycle in a graph is grouped into a single cluster that contains all nodes of a cycle. Finally, the algorithm performs belief propagation on the junction tree. However, representation of different edge types, i.e., activation or inhibition, might be inconvenient in such probabilistic approaches. If we modify many definitions and assumptions in SiTSFlow algorithm, Belief propagation approaches might be adopted to estimate activity of target processes in terms of marginal probabilities for cyclic pathways. Application of all these modifications still could be very troublesome and inefficient to obtain converged marginal probability of each gene and process in a cyclic pathway.

G-networks is a probabilistic queueing network having special customers, input, and service rates [104]. Each node of a pathway might be represented by a queue. We can assume that gene scores obtained with experimental data might be customers of the queue. Each queue, i.e., node, has input and service rates which represent activation and inhibition behaviors of nodes, respectively. Therefore, estimation of total input and service rates of each queue provides a measure to explain biological activity of that queue. However, the application of G-Networks on the described global network might spend too much running time due to its polynomial time complexity.

Another extension in this thesis might be performed on usage of different types of biological data. Essential transcriptome data types, microarray and ChIP-seq, are selected as the main data sources of the thesis. However, other low-throughput data sources might be integrated to calculate gene scores of the nodes in pathways. For example, MEDLINE abstracts contain text information about the genes activities in literature [105]. There are several text mining tools to search over MEDLINE abstracts. Hence, a literature profile might be constructed for each gene in a pathway by using text mining tools. This literature profile for a gene contains all terms reported in literature related with a particular gene and each term has a coefficient representing its importance [106]. However, research for some genes is excessive compared to other genes in pathway. Crucial activity of such particular genes has been so extensively studied for several decades, therefore literature data about these genes is very biased according to ordinary genes in a pathway. Another issue could be development of an algorithm to eliminate such data biases for stabilizing gene scores extracted from literature data.

The pathway merge algorithm can easily be deployed for different signalling pathways. Hence, the resulting global network would provide an universal map of cross-talk of specific pathways in cell signalling. Cross-talk refers the interactions between signalling pathways and it provides the exploration of simultaneous responses of distinct biological processes for a specific cell signal. For example, the crucial genes or process in cancer progression would be easily investigated by using such an universal map and SiTSFlow algorithm. Different methods might be implemented during merge algorithm. For instance, elimination of duplicated nodes in a pathway might be ignored, thus multiple copies of a gene in unified network would create interesting biological responses compared to current results. This possibility should be also investigated.

From biological perspective, gene knockout operations on a signalling network provides the assessment of lethality of hub-proteins for the life cycle of a cell. Gene knockout operations in global signalling network might provide very crucial information about specific proteins that have very important roles in cancer progression. Hence, application of gene knockout operations on a global network evaluated by SiTSFlow algorithm would have very effective contributions for designing of targeted drugs for these proteins.

REFERENCES

[1] H. Ji, H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, and W.H. Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nature Biotechnology*, 26(11):1293–1300, 2008.

[2] Robertson G., Hirst M., Bainbridge M., Bilenky M., Zhao Y., Zeng T., Euskirchen G., Bernier B., Varhol R., Delaney A., Thiessen N., Griffith O.L., He A., Marra M., Snyder M., and Jones S. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651, 2007.

[3] Isik Z., Atalay V., and Cetin-Atalay R. Evaluation of signaling cascades based on the weights from microarray and chip-seq data. *Journal of Machine Learning Research W&C Proceedings*, 8:44–54, 2010.

[4] Isik Z., Atalay V., Aykanat C., and Cetin-Atalay R. Data and model driven hybrid approach to activity scoring of cyclic pathway. *Lecture Notes in Electrical Engineering, Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS 2010)*, 62:91–94, 2010.

[5] Isik Z., Ersahin T., Atalay V., Aykanat C., and Cetin-Atalay R. Cyclic cellular pathway activities analyzed by a novel signal transduction score flow algorithm. *submitted to PLoS Computational Biology Journal*, 2011.

[6] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34:D354–D357, 2006.

[7] Matthews L., Gopinath G., Gillespie M., Caudy M., Croft D., de Bono B., Garapati P., Hemish J., Hermjakob H., Jassal B., Kanapin A., Lewis S., Mahajan S., May B., Schmidt E., Vastrik I., Wu G., Birney E., Stein L., and D'Eustachio P. Reactome knowledgebase of biological pathways and processes. *Nucleic Acids Res*, 37(Database issue):D619–22, 2009.

[8] Biocarta Database. *http://www.biocarta.com*. last visited date: 10.02.2011.

[9] Ingenuity Software. http://www.ingenuity.com. last visited date: 10.02.2011.

[10] Ariadne ResNet Software. *http://www.ariadnegenomics.com*. last visited date: 10.02.2011.

[11] Viswanathan G.A., Seto J., Patil S., Nudelman G., and S.C. Sealfon. Getting started in biological pathway construction and analysis. *PLoS Comput Biol*, 4(2):e16, 02 2008.

[12] Cordero F., Botta M., and Calogero R.A. Microarray data analysis and mining approaches. *Brief. in Funct. Genomics and Proteomics*, pages 1–17, 2008.

[13] Khatri P., Draghici S., Ostermeier G.C., and Krawetz S.A. Profiling gene expression using onto-express. *Genomics*, 79:266, 2002.

[14] Zeeberg B.R., Feng W., Wang G., Wang M.D., Fojo A.T., Sunshine M., Narasimhan S., Kane D.W., Reinhold W.C., Lababidi S. Bussey K.J., Riss J., Barrett J.C., and Weinstein J.N. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4:R28, 2003.

[15] Hosack D.A., Dennis G.Jr., Sherman B.T., Lane H.C., and Lempicki R.A. Identifying biological themes within lists of genes with ease. *Genome Biology*, 4:R70, 2003.

[16] Al-Shahrour F., Diaz-Uriarte R., and Dopazo J. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[17] Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., and Mesirov J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[18] Al-Shahrour F., Arbiza L., Dopazo H., Huerta-Cepas J., Minguez P., Montaner D., and Dopazo J. From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, 8:114, 2007.

[19] Kim S.Y. and Volsky D.J. Page: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6:144, 2005.

[20] Smid M. and Dorssers L.C. Go-mapper: functional analysis of gene expression data using the expression level as a score to evaluate gene ontology terms. *Bioinformatics*, 20:2618, 2004.

[21] Tarca A.L., Draghici S., Khatri P., Hassan S.S., Kim J.S. Mittal P. and, Kim C.J., Kusanovic J.P., and Romero R. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009.

[22] Efroni S., Schaefer C.F., and Buetow K.H. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*, 5:e525, 2007.

[23] Lee E., Chuang H.Y., Kim J.W., Ideker T., and Lee D. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11):e1000217, 2008.

[24] Vaske C.J., Benz S.C., Sanborn J.Z., Earl D., Szeto C., Zhu J., Haussler D., and Stuart J.M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–45, Jun 15 2010.

[25] Akutsu T. and Miyano S.and Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16:727–734, 2000.

[26] Shmulevich I., Dougherty E.R., Kim S., and Zhang W. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, 2002.

[27] Friedman N., Linial M., Nachman I., and Pe‘er D. Using bayesian network to analyze expression data. *J. Comp. Biol.*, 7:601–620, 2000.

[28] Hartemink A.J., Gifford D.K., Jaakkola T.S., and Young R.A. Combining location and expression data for principled discovery of genetic regulatory network models. pages 437–449. Pacific Symposium on Biocomputing, 2002.

[29] Chen T., He H., and Church G. Modeling gene expression with differential equations. pages 29–40. Pacific Symposium on Biocomputing, 1999.

[30] de Hoon M.J.L., Imoto S., Kobayashi K., Ogasawara N., and Miyano S. Inferring gene regulatory networks from time ordered gene expression data of bacillus subtilis using differential equations. pages 17–28. Pacific Symposium on Biocomputing, 2003.

[31] Bolstad B.M., Irizarry R.A., Astrand M., and Speed T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185, 2003.

[32] Song J.S., Johnson W.E., Zhu X., Zhang X., Li W., Manrai A.K., Liu J.S., Chen R., and Liu X.S. Model-based analysis of two-color arrays (ma2c). *Genome Biology*, 8(R178), 2007.

[33] Zhang Z.D., Rozowsky J., Lam H.Y., Du J., Snyder M., and Gerstein M. Tilescope: online analysis pipeline for high-density tiling microarray data. *Genome Biology*, 8(5):R81, 2007.

[34] Li W., Meyer C.A., and Liu X.S. A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21(Suppl. 1):i274–i282, 2005.

[35] Ji H. and Wong W.H. Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21:3629, 2005.

[36] Johnson W.E., Li W., Meyer C.A., Gottardo R., Carroll J.S., Brown M., and Liu X.S. Model-based analysis of tiling-arrays for chip-chip. *Proc. Natl. Acad. Sci.*, 103(33):12457, 2006.

[37] Kampa D., Cheng J., Kapranov P., Yamanaka M., Brubaker S., Cawley S., Drenkow J., Bekiranov S. Piccolboni A. and, Helt G., Tammana H., and Gingeras T.R. Novel rnas identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, 14:331, 2004.

[38] Keles S. Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, 63:10, 2007.

[39] Zheng M., Barrera L.O., Ren B., and Wu Y.N. Chip-chip: data, model, and analysis. *Biometrics*, 63:787, 2007.

[40] Qi Y., Rolfe A., MacIsaac K.D., Gerber G.K., Pokholok D., Zeitlinger J., Danford T., Dowell R.D., Fraenkel E., Jaakkola T.S., Young R.A., and Gifford D.K. High-resolution computational models of genome binding events. *Biotechnology*, 24:963, 2006.

[41] Reiss D.J., Facciotti M.T., and Balig N.S. Model-based deconvolution of genome-wide dna binding. *Bioinformatics*, 24:396, 2008.

[42] Toedling J., Skylar O., Krueger T., Fischer J.J., Sperling S., and Huber W. Ringo – an r/bioconductor package for analyzing chip-chip readouts. *BMC Bioinformatics*, 8:221, 2007.

[43] Bailey T.L. and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. page 2836, Menlo Park, California, USA, 1994. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press.

[44] Liu J.S., Neuwald A.F., and Lawrence C.E. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *J. Am. Stat. Assoc*, 90:1156–1170, 1995.

[45] Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., Zhang Y., Blankenberg D., Albert I., Taylor J., Miller W., Kent W.J., and Nekrutenko A. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res.*, 15:1451, 2005.

[46] Ji X., Li W., Song J., Wei L., and X.S. Li. Ceas: cis-regulatory element annotation system. *Nucleic Acids Res.*, 34:551, 2006.

[47] Cox A.J. at Illumina. Eland (efficient large-scale alignment of nucleotide databases).

[48] Jiang H. and Wong W.H. Seqmap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, 24:2395, 2008.

[49] Smith A.D., Xuan Z., and Zhang M.Q. Using quality scores and longer reads improves accuracy of solexa read mapping. *BMC Bioinformatics*, 9:128, 2008.

[50] Li R., Li Y., Kristiansen K., and Wang J. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24:713–714, 2008.

[51] Lin H., Zhang Z., Zhang M.Q., Ma B., and Li M. Zoom! zillions of oligos mapped. *Bioinformatics*, 24(21):2431–2437, 2008.

[52] Albert I., Wachi S., Jiang C., and Pugh B.F. Genetrack: a genomic data processing and visualization framework. *Bioinformatics*, 24:1305, 2008.

[53] Valouev A., Johnson D.S., Sundquist A., Medina C., Anton E., Batzoglou S., Myers R.M., and Sidow A. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat. Methods*, 5:829, 2008.

[54] Jothi R., Cuddapah S., Barski A., Cui K., and Zhao K. Genome-wide identification of in vivo protein-dna binding sites from chip-seq data. *Nucleic Acids Res*, 36:5221, 2008.

[55] Zhou Q. and Wong W.H. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci.*, page 12114, 2004.

[56] Breitling R., Armengaud P., Amtmann A., and Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573:83–92, 2004.

[57] White M.A., Nicolette C., Minden A., Polverino A., Van-Aelst L., Karin M., and Wigler M.H. Multiple ras functions can contribute to mammalian cell transformation. *Cell*, 80(4):533–41, 1995.

[58] Al-Mulla F., Milner-White E.J., Going J.J., and Birnie G.D. Structural differences between valine-12 and aspartate-12 ras proteins may modify carcinoma aggression. *J Pathol.*, 187(4):433–8, 1999.

[59] Engelman J.A. Targeting pi3k signaling in cancer: opportunities, challenges and limitations. *Nat Rev Cancer*, 9(8):550–62, 2009.

[60] Liu P., Cheng H., Roberts T.M., and Zhao J.J. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat Rev Drug Discov.*, 8(8):627–44, 2009.

[61] Tokunaga E., Oki E., Egashira A., Sadanaga N., Morita M., Kakeji Y., and Maehara Y. Deregulation of the akt pathway in human cancer. *Curr Cancer Drug Targets*, 8(1):27–36, 2008.

[62] Manning B.D. and Cantley L.C. Akt/pkb signaling: navigating downstream. *Cell*, 129(7):1261–74, 2007.

[63] Shaw R.J. and Cantley L.C. Ras, pi(3)k and mtor signaling controls tumour cell growth. *Nature*, 441(7092):424–30, 2006.

[64] Gupta S., Ramjaun A.R., Haiko P., Wang Y., Warne P.H., Nicke B., Nye E., Stamp G., Alitalo K., and Downward J. Binding of ras to phosphoinositide 3-kinase p110alpha is required for ras-driven tumorigenesis in mice. *Cell*, 129(5):957–68, 2007.

[65] Fridman J.S. and Lowe S.W. Control of apoptosis by p53. *Oncogene*, 22(56):9030–40, 2003.

[66] Vogelstein B., Lane D., and Levine A.J. Surfing the p53 network. *Nature*, 408(6810):307–10, 2000.

[67] J. Kang, M. Gemberling, M. Nakamura, F.G. Whitby, H. Handa, W.G. Fairbrother, and D. Tantin. A general mechanism for transcription regulation by oct1 and oct4 in response to genotoxic and oxidative stress. *Genes Dev.*, 23(2):208–222, 2009.

[68] J.I. Murray, M.L. Whitfield, N.D. Trinklein, R.M. Myers, P.O. Brown, and D. Botstein. Diverse and specific gene expression responses to stresses in cultured human cells. *Molecular and Cellular Biology*, 15(5):2361–2374, 2004.

[69] Hu M., Yu J., Taylor J.M., Chinnaiyan A.M., and Qin Z.S. On the detection and refinement of transcription factor binding sites using chip-seq data. *Nucleic Acids Res.*, 38(7):2154–67, 2010.

[70] Lin C.Y., Vega V.B., Thomsen J.S., Zhang T., Kong S.L., Xie M., Chiu K.P., Lipovich L., Barnett D.H., Stossi F., Yeo A., George J., Kuznetsov V.A., Lee Y.K., Charn T.H., Palanisamy N., Miller L.D., Cheung E., Katzenellenbogen B.S., Ruan Y., Bourque G., Wei C.L., and Liu E.T. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet.*, 3(6):e87, 2007.

[71] Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., and Speed T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249, 2003.

[72] Vivar O.I., Zhao X., Saunier E.F., Griffin C., Mayba O.S., Tagliaferri M., Cohen I., Speed T.P., and Leitman D.C. Estrogen receptor beta binds to and regulates three distinct classes of target genes. *J Biol Chem.*, 285(29):22059–66, 2010.

[73] Monticone M., Biollo E., Maffei M., Donadini A., Romeo F., Storlazzi C.T., Giaretti W., and Castagnola P. Gene expression deregulation by kras g12d and g12v in a braf v600e context. *Mol. Cancer*, 17(7):92, 2008.

[74] Arimoto K., Fukuda H., Imajoh-Ohmi S., Saito H., and Takekawa M. Formation of stress granules inhibits apoptosis by suppressing stress-responsive mapk pathways. *Nat Cell Biol.*, 10(11):1324–32, 2008.

[75] McMillan-Ward E. Chen Y, Kong J., Israels S.J., and Gibson S.B. Oxidative stress induces autophagic cell death independent of apoptosis in transformed and cancer cells. *Cell Death Differ.*, 15(1):171–82, 2008.

[76] Song R.X., Zhang Z., Chen Y., Bao Y., and Santen R.J. Estrogen signaling via a linear pathway involving insulin like growth factor i receptor, matrix metalloproteinases, and epidermal growth factor receptor to activate mitogen activated protein kinase in mcf7 breast cancer cells. *Endocrinology*, 148(8):4091–101, 2007.

[77] Seeger H., Wallwiener D., Kraemer E., and Mueck A.O. Comparison of possible carcinogenic estradiol metabolites: effects on proliferation, apoptosis and metastasis of human breast cancer cells. *Maturitas*, 54(1):72–7, 2006.

[78] Martinez-Campa C., Casado P., Rodriguez R., Zuazua P., Garcia-Pedrero J.M., Lazo P.S., and Ramos S. Effect of vinca alkaloids on eralpha levels and estradiol-induced responses in mcf7 cells. *Breast Cancer Res Treat.*, 98(1):81–9, 2006.

[79] Ye Y., Xiao Y., Wang W., Yearsley K., Gao J.X., Shetuni B., and Barsky S.H. Eralpha signaling through slug regulates e-cadherin and emt. *Oncogene*, 29(10):1451–62, 2010.

[80] Ropero A.B., Alonso-Magdalena P., Quesada I., and Nadal A. The role of estrogen receptors in the control of energy and glucose homeostasis. *Steroids*, 73(9-10):874–9, 2008.

[81] Foryst-Ludwig A. and Kintscher U. Metabolic impact of estrogen signaling through eralpha and erbeta. *J. Steroid Biochem. Mol. Biol.*, 122(1-3):74–81, 2010.

[82] Fan M., Nakshatri H., and Nephew K.P. Inhibiting proteasomal proteolysis sustains estrogen receptor-alpha activation. *Mol. Endocrinol.*, 18(11):2603–15, 2004.

[83] Tabuchi Y., Matsuoka J., Gunduz M., Imada T., Ono R., Ito M., Motoki T., Yamatsuji T., Shirakawa Y., Takaoka M., Haisa M., Tanaka N., Kurebayashi J., Jordan V.C., and Naomoto Y. Resistance to paclitaxel therapy is related with bcl-2 expression through an estrogen receptor mediated pathway in breast cancer. *Int. J. Oncol.*, 34(2):313–9, 2009.

[84] Stander B.A., Marais S., Vorster C.J., and Joubert A.M. In vitro effects of 2-methoxyestradiol on morphology, cell cycle progression, cell death and gene expression changes in the tumorigenic mcf-7 breast epithelial cell line. *J. Steroid Biochem. Mol. Biol.*, 119((3-5)):149–60, 2010.

[85] Oh A.S., Lorant L.A., Holloway J.N., Miller D.L., Kern F.G., and El-Ashry D. Hyperactivation of mapk induces loss of eralpha expression in breast cancer cells. *Mol. Endocrinol*, 15(8):1344–59, 2001.

[86] Brinkman J.A. and El-Ashry D. Er re-expression and re-sensitization to endocrine therapies in er-negative breast cancers. *J Mammary Gland Biol Neoplasia*, 14(1):67–78, 2009.

[87] Applanat M.P., Buteau-Lozano H., Herve M.A., and Corpet A. Vascular endothelial growth factor is a target gene for estrogen receptor and contributes to breast cancer progression. *Adv Exp Med Biol.*, 617:437–44, 2008.

[88] Hyder S.M., Liang Y., and Wu J. Estrogen regulation of thrombospondin-1 in human breast cancer cells. *Int J Cancer*, 125(5):1045–53, 2009.

[89] Cespedes M.V., Sancho F.J., Guerrero S., Parreno M., Casanova I., Pavon M.A., Marcuello E., Trias M., Cascante M., Capella G., and Mangues R. K-ras asp12 mutant neither interacts with raf, nor signals through erk and is less tumorigenic than k-ras val12. *Carcinogenesis*, 27(11):2190–200, 2006.

[90] Joneson T., White M.A., Wigler M.H., and Bar-Sagi D. Stimulation of membrane ruffling and map kinase activation by distinct effectors of ras. *Science*, 271(5250):810–2, 1996.

[91] Oliveira C., Velho S., Moutinho C., Ferreira A., Preto A., Domingo E., Capelinha A.F., Duval A., Hamelin R., Machado J.C., Schwartz S., Carneiro F., and Seruca R. Kras and braf oncogenic mutations in mss colorectal carcinoma progression. *Oncogene*, 26(1):158–63, 2007.

[92] Costa A.M., Herrero A., Fresno M.F., Heymann J., Alvarez J.A., Cameselle-Teijeiro J., and Garcia-Rostan G. Braf mutation associated with other genetic events identifies a subset of aggressive papillary thyroid carcinoma. *Clin Endocrinol*, 68(4):618–34, 2008.

[93] Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–504, 2003.

[94] Normanno N., Tejpar S., Morgillo F., De Luca A. Van Cutsem E., and Ciardiello F. Implications for kras status and egfr-targeted therapies in metastatic crc. *Nat Rev Clin Oncol.*, 6(9):519–27, Sep 2009.

[95] Loriot Y., Mordant P., Deutsch E., Olaussen K.A., and Soria J.C. Are ras mutations predictive markers of resistance to standard chemotherapy? *Nat Rev Clin Oncol.*, 6(9):528–34, Sep 2009.

[96] Heinemann V., Stintzing S., Kirchner T., Boeck S., and Jung A. Clinical relevance of egfr- and kras-status in colorectal cancer patients treated with monoclonal antibodies directed against the egfr. *Cancer Treat Rev.*, 35(3):262–71, May 2009.

[97] Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine. In: Seventh International World-Wide Web Conference, 1998.

[98] Shi J. and Malik J. Normalized cuts and image segmentation. pages 731–737. IEEE Conf. Computer Vision and Pattern Recognition, 1997.

[99] Bishop C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[100] Friedman N., Murphy K., and Russell S. Learning the structure of dynamic probabilistic networks. page 139. In: proceedings of the Conference on Uncertainty in Artificial Intelligence, 1998.

[101] Murphy K. and Mian S. Modelling gene expression data using dynamic bayesian networks. Technology report, Computer Science Division, University of California Berkeley, CA, 1999.

[102] Frey B.J. and MacKay D.J.C. A revolution: Belief propagation in graphs with cycles. Advances in Neural Information Processing Systems (NIPS), MIT Press, 1998.

[103] Lauritzen S.L. and Spiegelhalter D.J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50(2):157, 1988.

[104] Gelenbe E. Steady-state solution of probabilistic gene regulatory networks. *J Theor Biol Phys Rev E*, 76(031903), 2007.

[105] Pubmed web site. *http://www.ncbi.nlm.nih.gov/pubmed/*. last visited date: 10.02.2011.

[106] Aerts S., Lambrechts D., Maity S., Loo P.V., Coessens B., De-Smet F., Tranchevent L.C., De-Moor B., Marynen P., Hassan B., Carmeliet P., and Moreau Y. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537–544, 2006.

APPENDIX A

SCREENSHOTS OF EMPLOYED PATHWAYS

The screenshots of original KEGG pathways employed in this thesis are given in this appendix. They were created by uploading the original KGML files from KEGG PATHWAY database and displayed in the Cytoscape environment by using the developed plug-in.

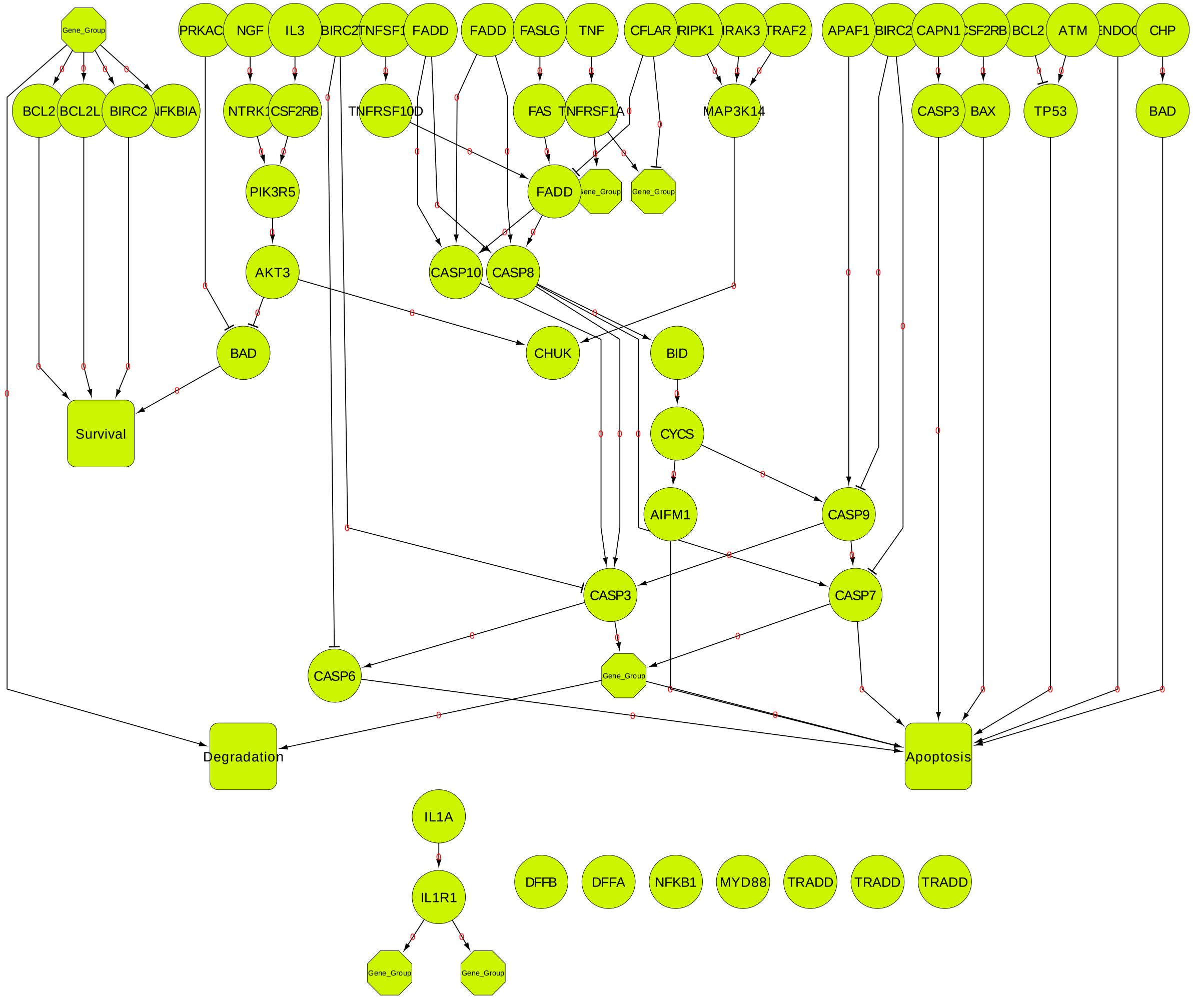


Figure A.1: Screenshot of original *Apoptosis* pathway from KEGG PATHWAY Database.

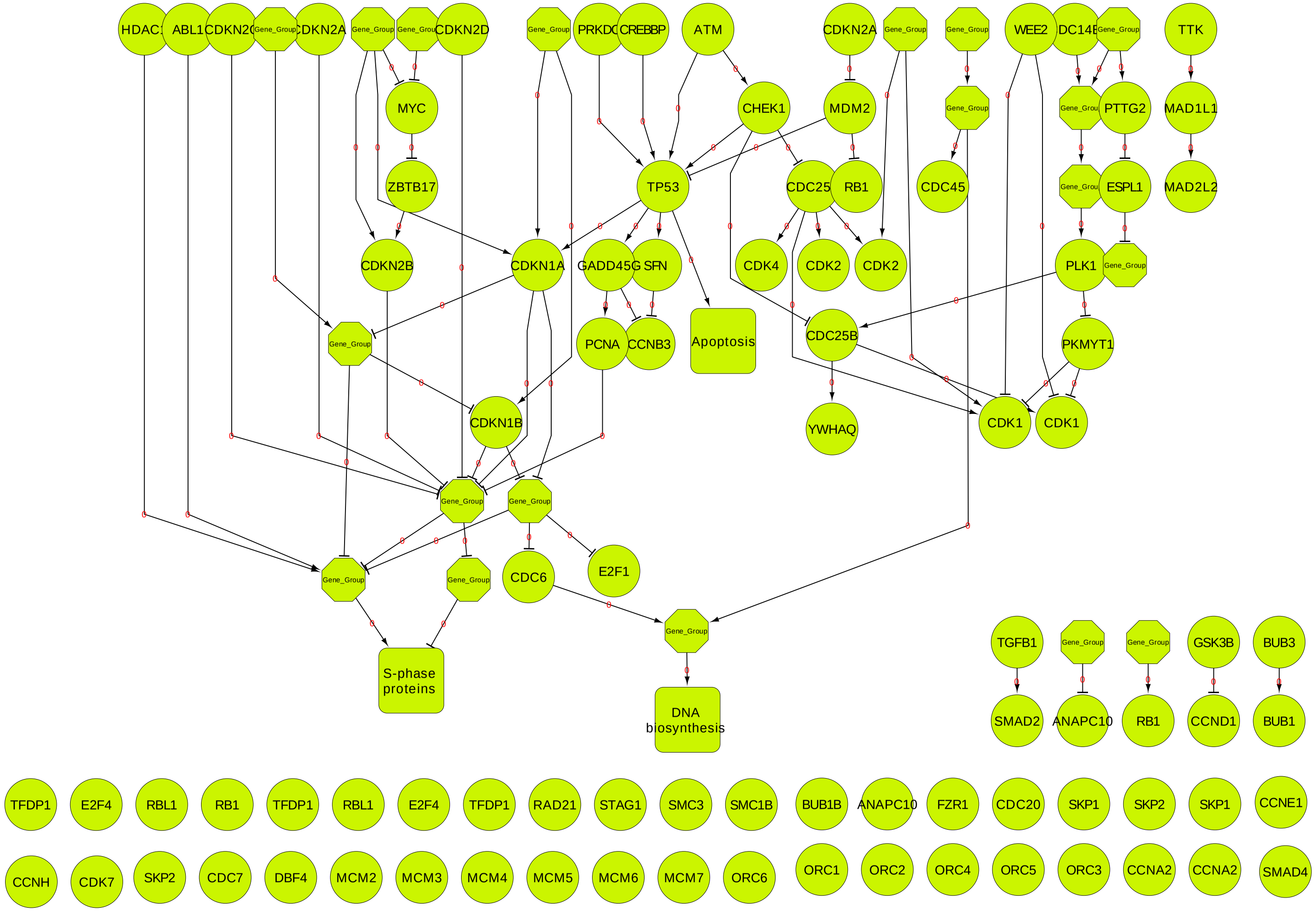


Figure A.2: Screenshot of original *Cell cycle* pathway from KEGG PATHWAY Database.

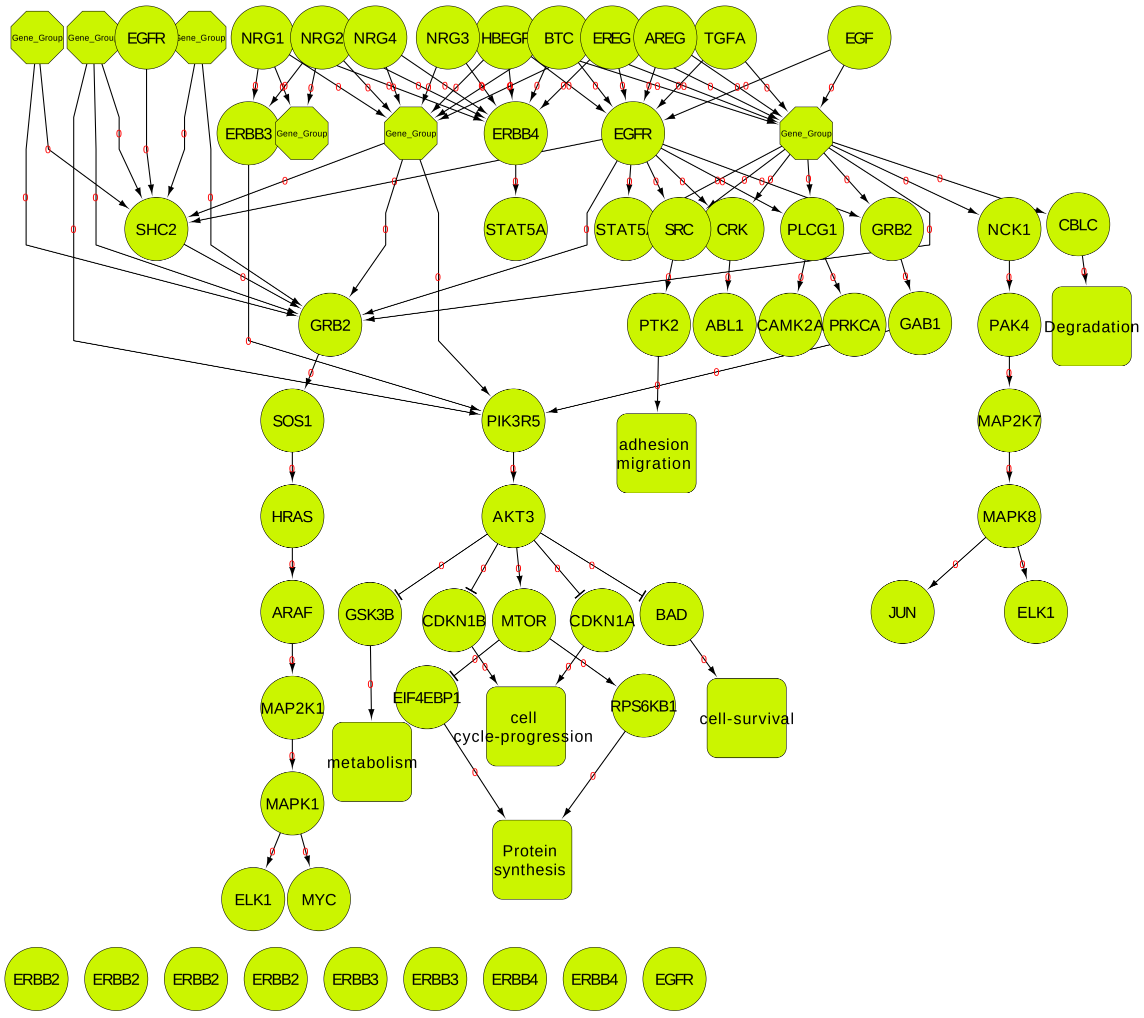


Figure A.3: Screenshot of original *ErbB signalling* pathway from KEGG PATHWAY Database.

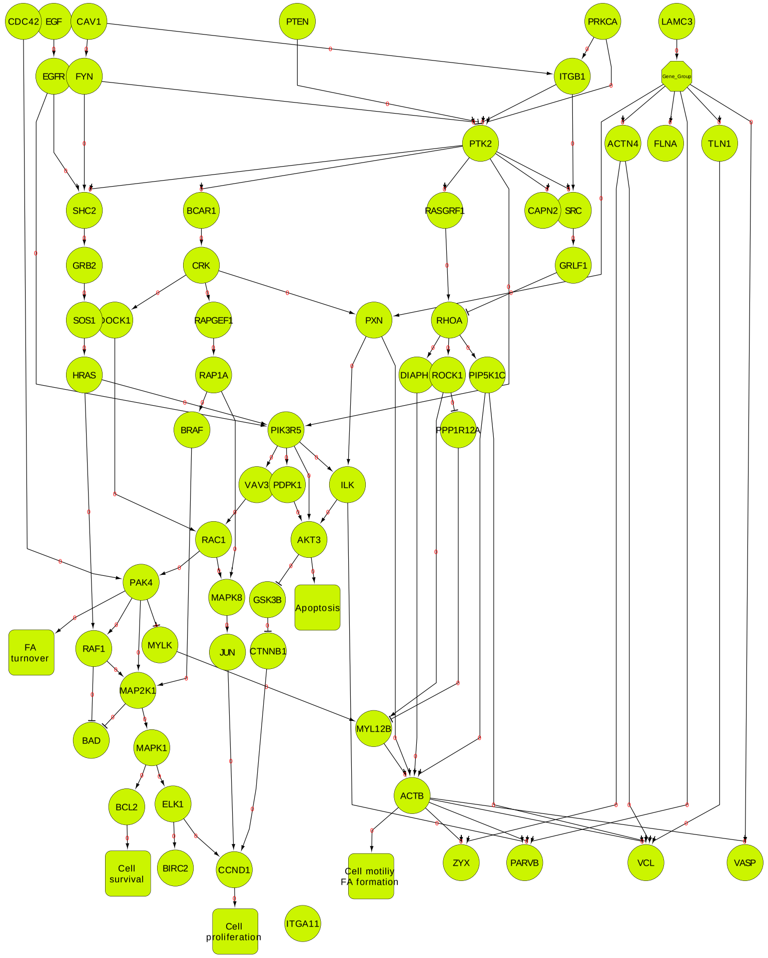


Figure A.4: Screenshot of original *Focal Adhesion* pathway from KEGG PATHWAY Database.

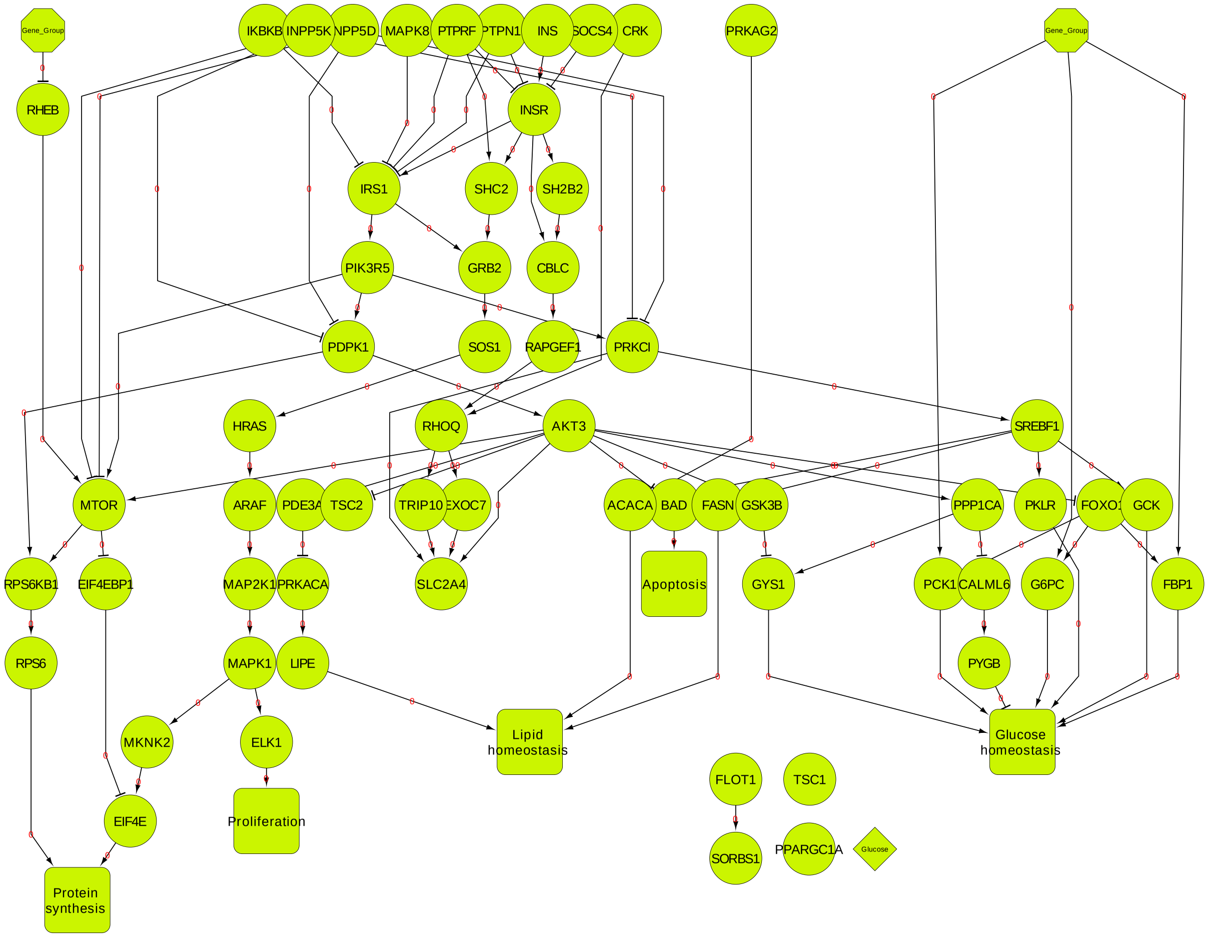


Figure A.5: Screenshot of original *Insulin signalling* pathway from KEGG PATHWAY Database.

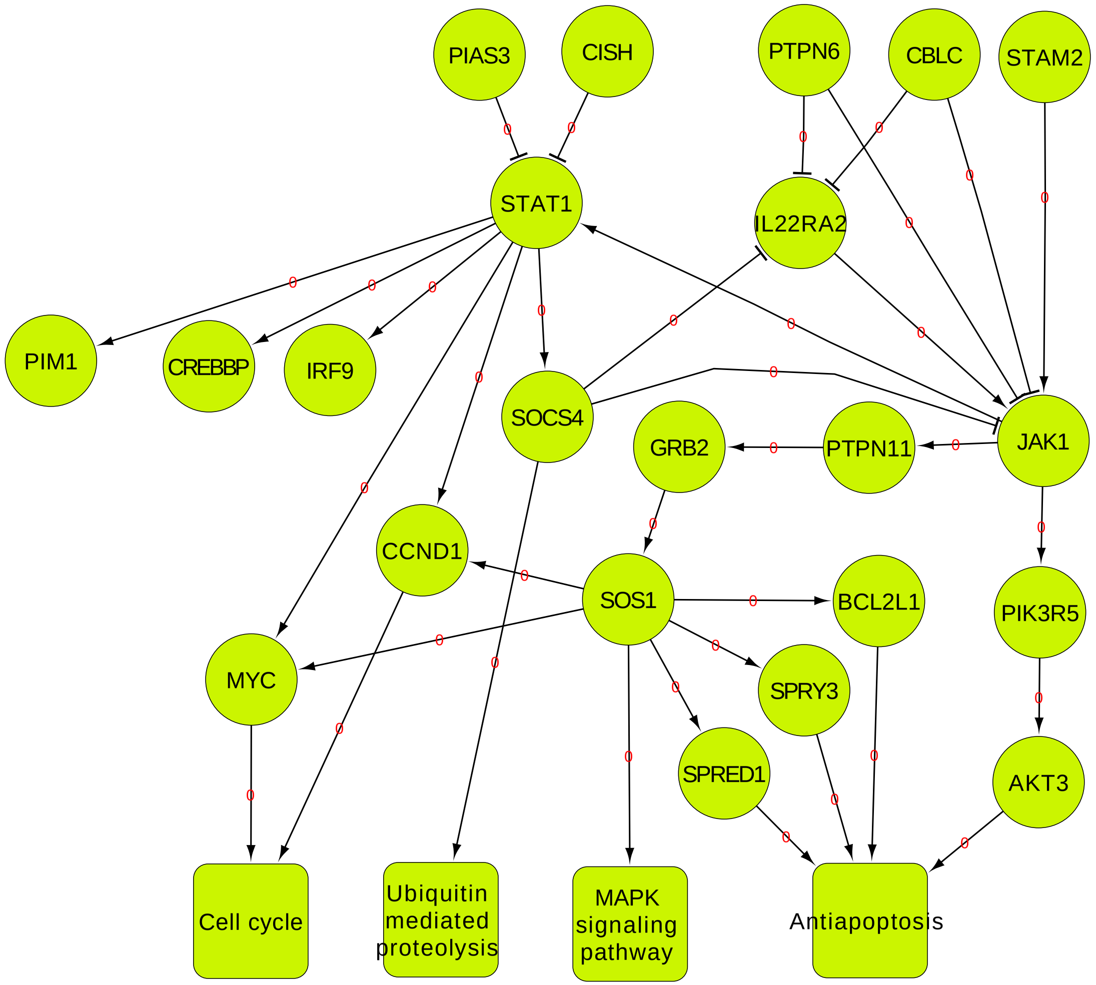


Figure A.6: Screenshot of original *Jak-STAT signalling* pathway from KEGG PATHWAY Database.

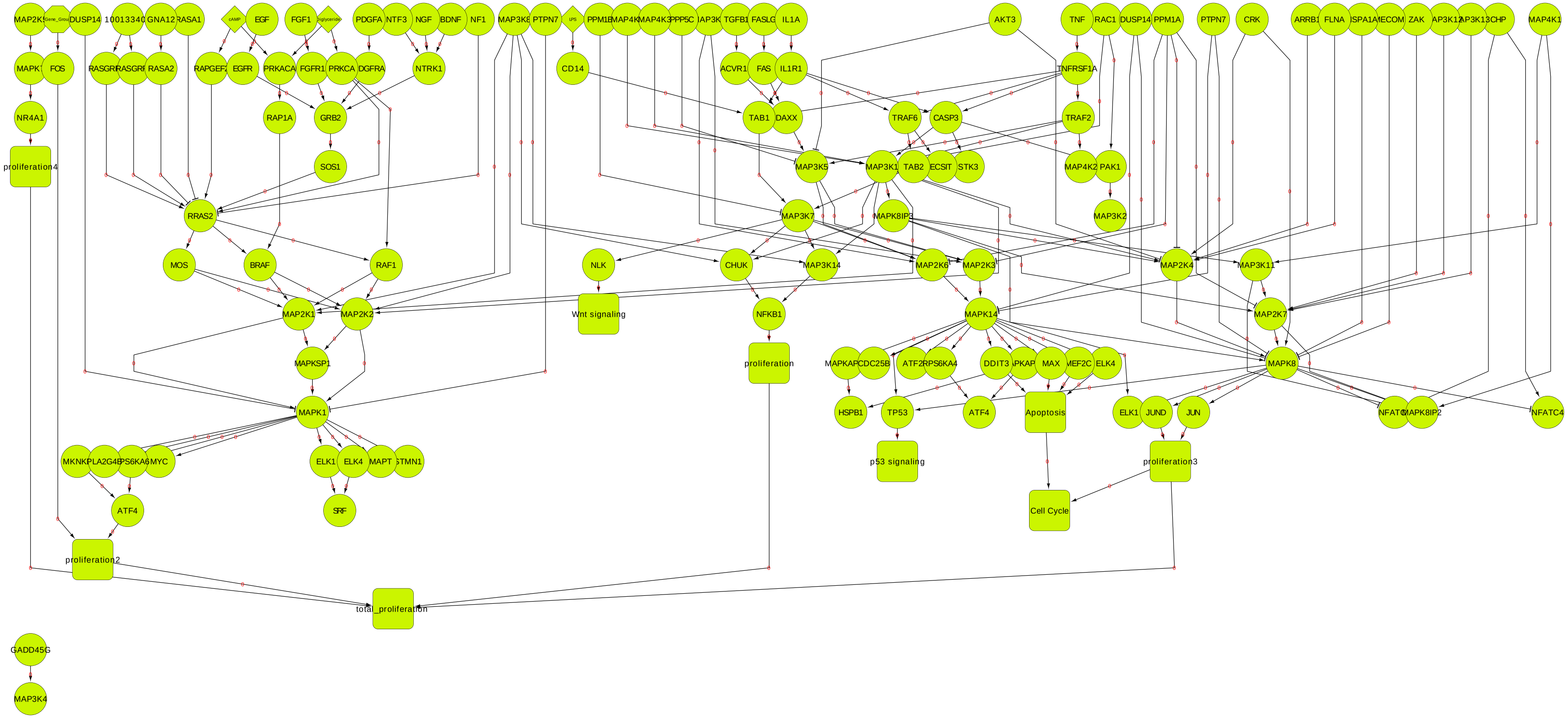


Figure A.7: Screenshot of original *MAPK signalling* pathway from KEGG PATHWAY Database.

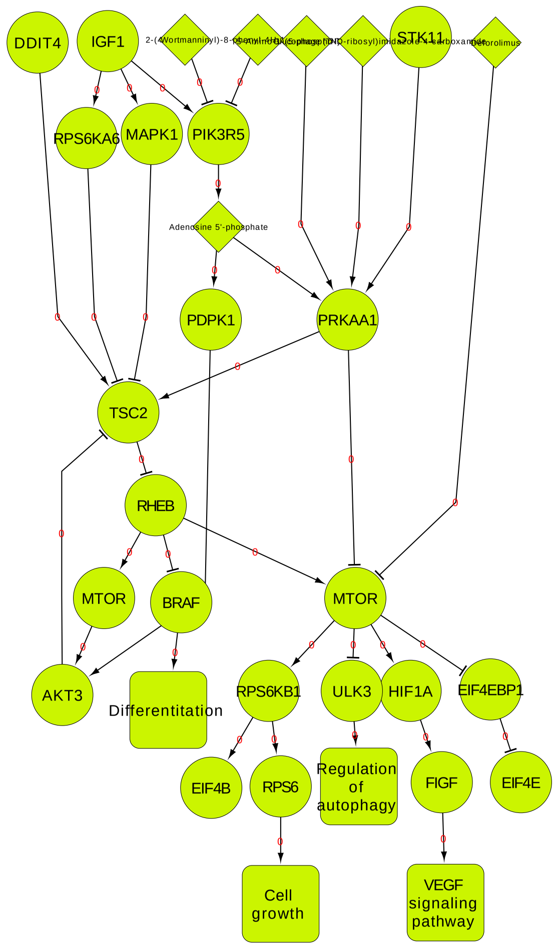


Figure A.8: Screenshot of original *mTOR signalling* pathway from KEGG PATHWAY Database.

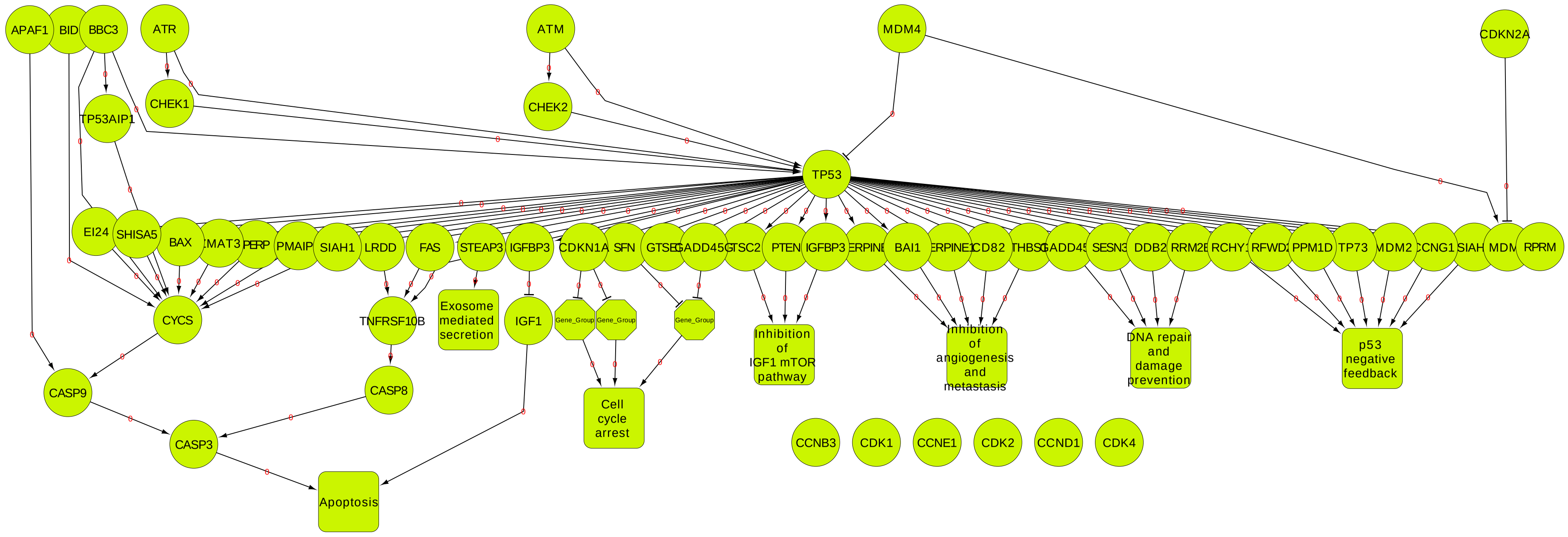


Figure A.9: Screenshot of original *P53 signalling* pathway from KEGG PATHWAY Database.

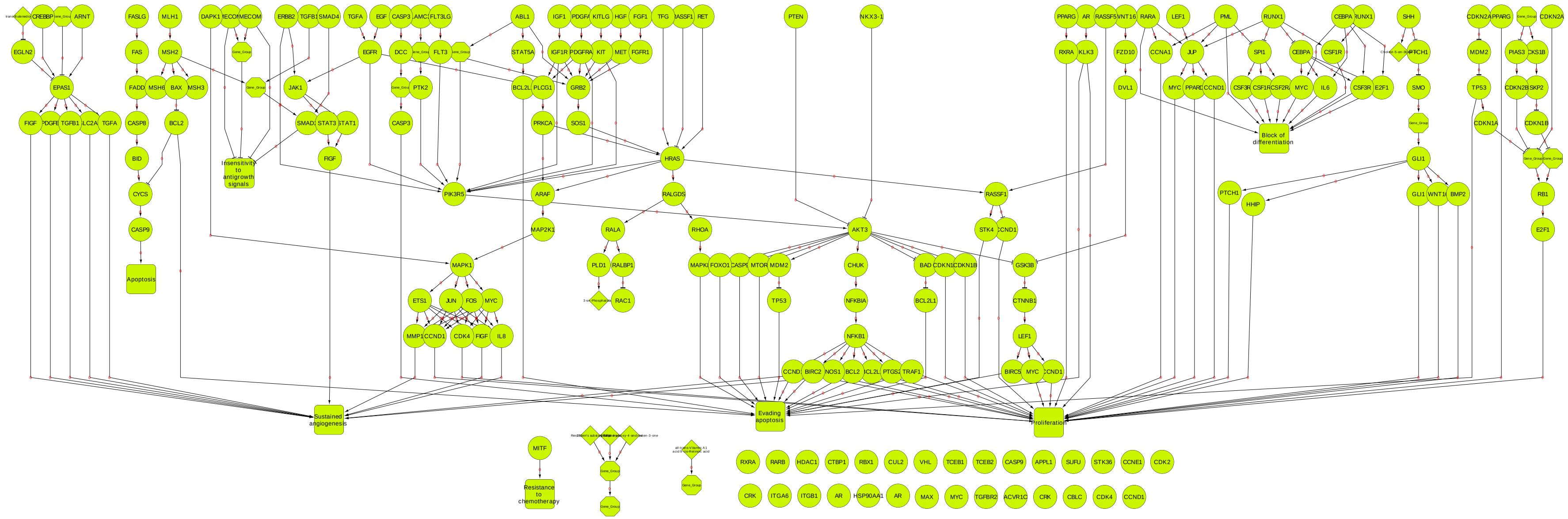


Figure A.10: Screenshot of original *Pathways in cancer* pathway from KEGG PATHWAY Database.

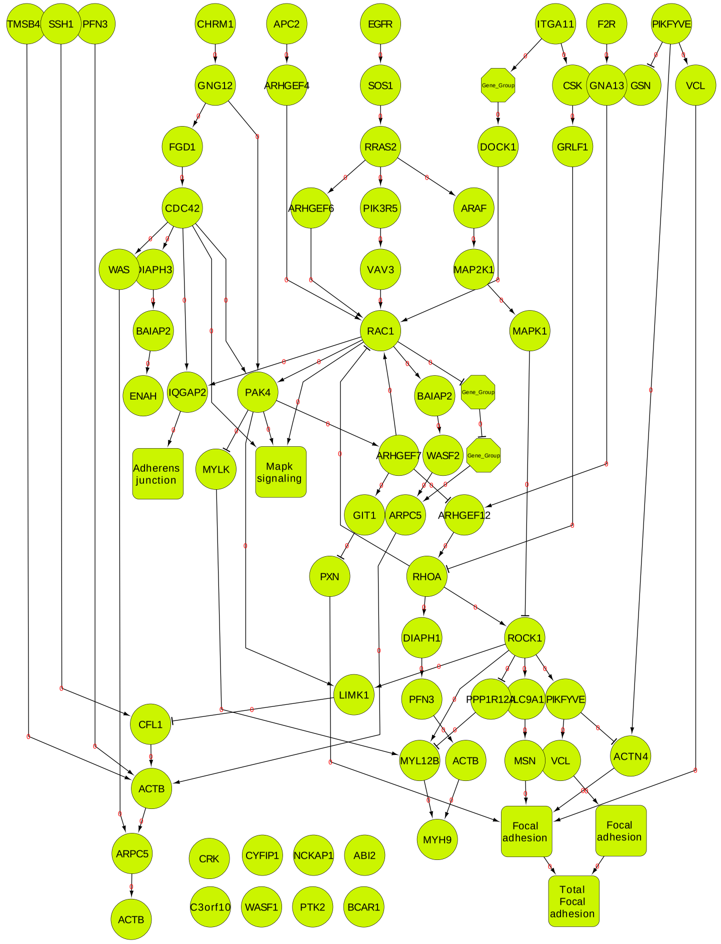


Figure A.11: Screenshot of original *Regulation of actin cytoskeleton* pathway from KEGG PATHWAY Database.

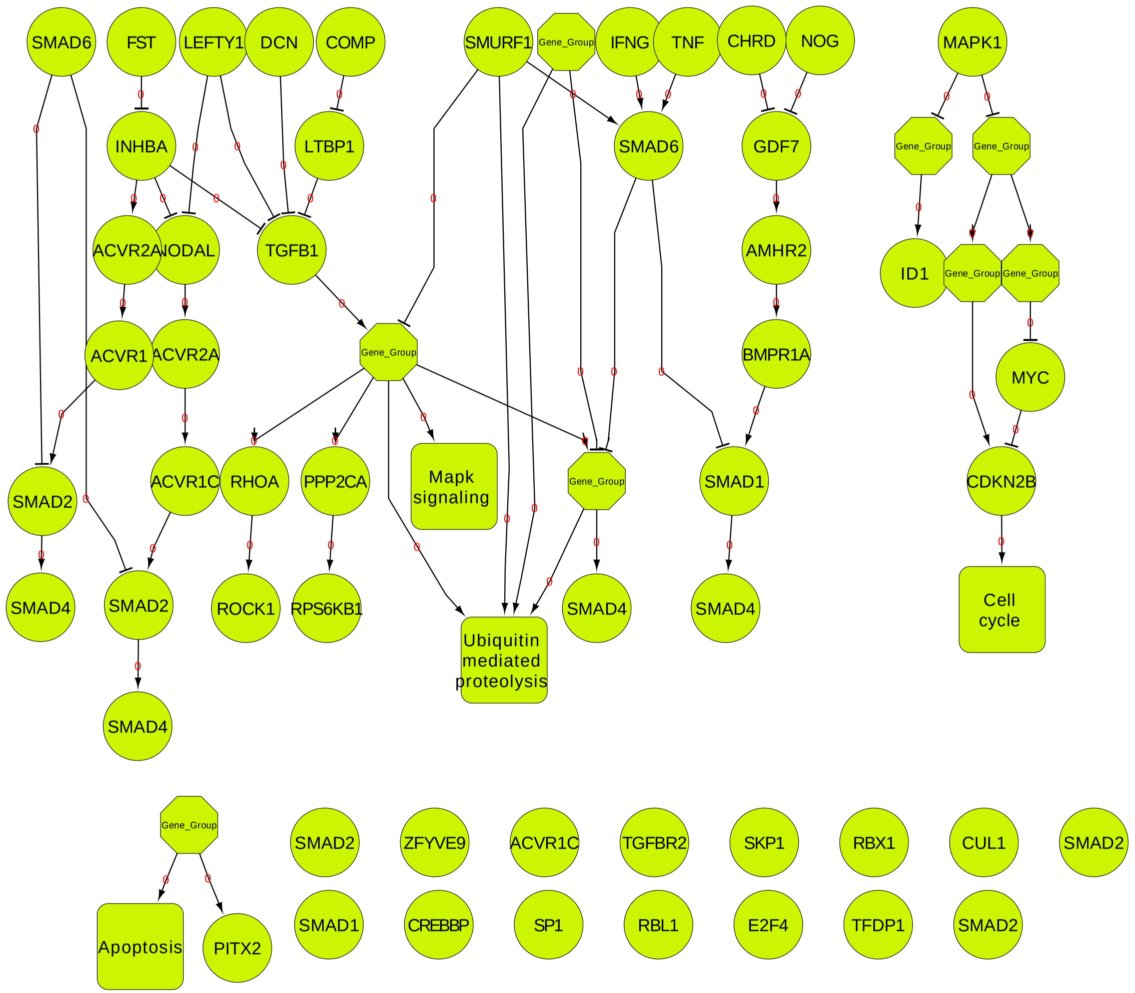


Figure A.12: Screenshot of original *TGF-β signalling* pathway from KEGG PATHWAY Database.

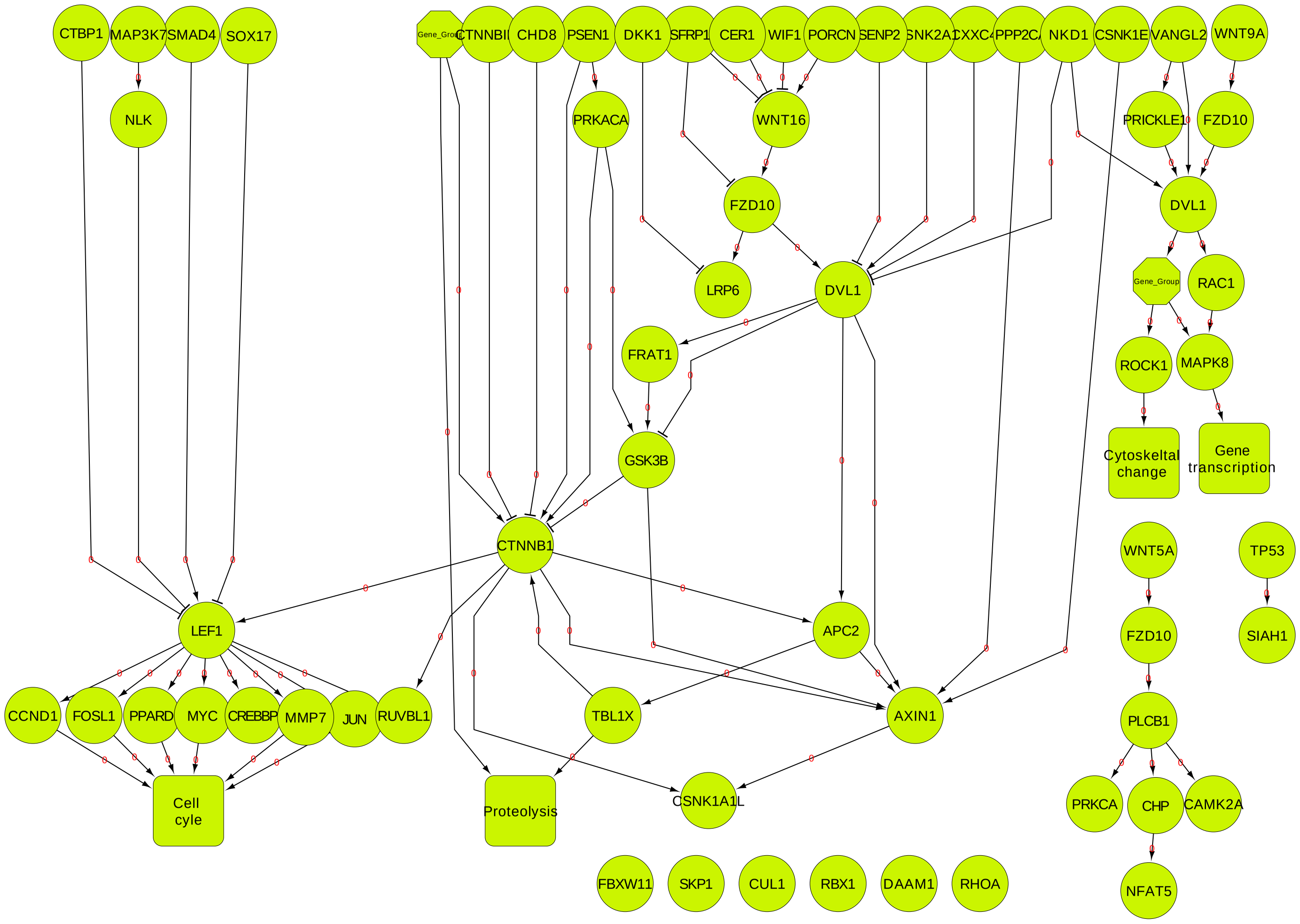


Figure A.13: Screenshot of original *Wnt signalling* pathway from KEGG PATHWAY Database.

VITA

PERSONAL INFORMATION

Surname, Name: Işık, Zerrin
Nationality: Turkish (TC)
Date and Place of Birth: 10 August 1979, İzmir
Marital Status: Single
Phone: +90 312 210 55 41
Fax: +90 312 210 55 44
email: e152115@metu.edu.tr

EDUCATION

Degree	Institution	Year of Graduation
Ph.D. in Computer Eng.	Middle East Technical University	2011
M.S. in Computer Sci. and Eng.	Sabancı University	2003
B.S. in Computer Eng.	Dokuz Eylül University	2001

WORK EXPERIENCE

Year	Place	Enrollment
2006-Present	Computer Eng. Depart. in METU	Research Assistant
2006	Computer Eng. Depart. in Çankaya University	Teaching Assistant
2004-2005	BTT Ltd. Şti.	Software Engineer
2001-2004	Computer Sci. and Eng. in Sabancı University	Teaching Assistant

PUBLICATIONS

1. Isik Z., Ersahin T., Atalay V., Aykanat C., and Cetin-Atalay R., “Cyclic Cellular Pathway Activities Analyzed by a Novel Signal Transduction Score Flow Algorithm”, submitted to PLoS Computational Biology Journal, (2011).

2. Isik Z., Atalay V., Aykanat C., and Cetin-Atalay R., “Data and Model Driven Hybrid Approach to Activity Scoring of Cyclic Pathway”, Lecture Notes in Electrical Engineering, Proceedings of the 25th International Symposium on Computer and Information Sciences (ISCIS 2010), Vol. 62, pp.91-94, (2010).

3. Isik Z., Atalay V., and Cetin-Atalay R., “Evaluation of Signaling Cascades Based on the Weights from Microarray and ChIP-seq Data”, Journal of Machine Learning Research W&C Proceedings, MIT Press, Vol.8, pp.44-54, (2010).

4. Isik Z., Atalay V., and Cetin-Atalay R., “Integrated Transcriptome Data Unified into the En Route of the Cell Signaling Pathways”, International Symposium on Health Informatics and Bioinformatics (HIBIT 2010), Turkey, (2010).

5. Sokmen Z., Atalay V., and Cetin-Atalay R., “Integration of ChIP-seq and microarray gene expression data”, International Symposium on Health Informatics and Bioinformatics (HIBIT 2009), Turkey, (2009).

6. Sokmen Z., Atalay V., and Cetin-Atalay R., “Short Time Series Microarray Data Analysis and Biological Annotation”, IEEE 16. Sinyal Isleme, Iletisim ve Uygulamalari Kurultayi (SIU 2008), Turkey, ISBN: 978-1-4244-1998-2, (2008).

7. Sokmen Z., Atalay V., and Cetin-Atalay R., “Progressive Clustering by Integration of Heterogenous Data From Multiple Sources for Target Gene Identification”, Second International Workshop on Machine Learning in Systems Biology (MLSB 2008), Brussels, (2008).

8. Sokmen Z., Yuzugullu O., Atalay V., and Cetin-Atalay R., “Short Time Series Microarray Data Analysis for Resistance to Selenium DeficiencyÓ, International Symposium on Health Informatics and Bioinformatics (HIBIT 2008), Turkey, (2008).

9. Sokmen Z., Ozturk M., Atalay V., and Cetin-Atalay R., “A Hybrid Method For The Identification of Expression Patterns From Microarray Data”, 15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 6th European Conference on Computational Biology (ECCB), Vienna - Austria, (2007).

10. Sokmen Z., Can T., Soylu R., Kocaefe C., Ozguc M., and Cetin-Atalay R., “MLC1 structure predictionÓ, International Symposium on Health Informatics and Bioinformatics (HIBIT 2007), Turkey, (2007).

11. Isik Z., Yanikoglu B., and Sezerman U., “Protein Structural Class Determination Using Support Vector Machines”, Lecture Notes in Computer Science (ISCIS 2004), Vol.3280, pp.82, (2004).

AWARD and SCHOLARSHIP

- Travel Grant by TUBITAK to Third International Workshop on Machine Learning in Systems Biology, 5-6 September, Ljubljana- Slovenia (2009)
- Graduate Courses Performance Award, Middle East Technical University (2007)
- Ph.D. Fellowship by TUBITAK (2007)
- Full Scholarship for graduate education, Sabancı University (2001)
- First honors degree, Dokuz Eylül University (2001)