REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# IMPACT ANALYSIS ALGORITHMS FOR BIOLOGICAL INTERACTION NETWORKS

OZAN ÖZIŞIK

PhD. THESIS
DEPARTMENT OF COMPUTER ENGINEERING
PROGRAM OF COMPUTER ENGINEERING

ADVISER
ASSOC. PROF. DR. BANU DİRİ

İSTANBUL, 2016

# REPUBLIC OF TURKEY
# YILDIZ TECHNICAL UNIVERSITY
# GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

## IMPACT ANALYSIS ALGORITHMS FOR BIOLOGICAL INTERACTION NETWORKS

A thesis submitted by Ozan ÖZIŞIK in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY** is approved by the committee on 13.07.2016 in Department of Computer Engineering, Computer Engineering Program.

**Thesis Adviser**

Assoc. Prof. Dr. Banu DİRİ
Yıldız Technical University

**Co- Adviser**

Assist. Prof. Dr. R. Murat DEMİRER
Üsküdar University

**Approved By the Examining Committee**

Assoc. Prof. Dr. Banu DİRİ
Yıldız Technical University ____________________

Prof. Dr. Nizamettin AYDIN, Member
Yıldız Technical University ____________________

Prof. Dr. O. Uğur SEZERMAN, Member
Acıbadem University ____________________

Assist. Prof. Dr. Mehmet AKTAŞ, Member
Yıldız Technical University ____________________

Assist. Prof. Dr. Özgür ASAR, Member
Acıbadem University ____________________

# ACKNOWLEDGEMENTS

Working on human diseases, getting the chance of making an improvement in the domain has made this study an exciting journey. I thank Assist. Prof. Dr. R. Murat DEMİRER for leading me to the bioinformatics domain. I am deeply thankful to Assoc. Prof. Dr. Banu DİRİ and Prof. Dr. Uğur SEZERMAN, who encouraged me in all the steps of the thesis, inspired me as a researcher, supported me in hard times, and made me feel really lucky as a PhD student. I thank Prof. Dr. Nizamettin Aydın for all his support as a member of thesis supervising committee and as the head of our department. I thank Assist. Prof. Dr. Burcu Bakır Güngör who has helped me a lot despite her busy schedule. I thank Assoc. Prof. Dr. Sırma Yavuz, who motivated me when I was an undergraduate student, made me love my field, and encouraged me to become a researcher.

I thank my mother and grandmother who raised me and made me who I am. Finally I want to thank my wife, who has walked with me in this tough path, who has been the biggest part of my life. This thesis is dedicated to these three beautiful ladies.

July, 2016

Ozan ÖZIŞIK

# TABLE OF CONTENTS

Page

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

# LIST OF SYMBOLS

| | |
|---|---|
| $\Phi^{-1}$ | Inverse normal cumulative distribution function |
| $J(A, B)$ | Jaccard index of sets A and B |
| $p_i$ | p-value of *i*th node |
| $z_i$ | z score of *i*th node |

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BD | Behçet's disease |
| BDNF | Brain derived neurotrophic factor |
| CAMs | Cell adhesion molecules |
| DFT | Depth first traversal |
| FAK | Focal adhesion kinase |
| GA | Genetic algorithm |
| GEP | Gene expression profiling |
| GWAS | Genome-wide association study |
| IA | Intracranial aneurysm |
| NGF | Nerve growth factor |
| NT-3 | Neurotrophin 3 |
| NT-4 | Neurotrophin 4 |
| OA | Osteoarthritis |
| PPI | Protein-protein interaction |
| RA | Rheumatoid arthritis |
| SA | Simulated annealing |
| SNP | Single nucleotide polymorphism |
| SpA | Spondyloarthritis |
| SPIA | Signaling pathway impact analysis |
| WTCCC | Wellcome Trust Case Control Consortium |

# LIST OF FIGURES

# LIST OF TABLES

## ABSTRACT

# IMPACT ANALYSIS ALGORITHMS FOR BIOLOGICAL INTERACTION NETWORKS

Ozan Özışık

Computer Engineering Department

Ph.D. Thesis

Adviser: Assoc. Prof. Dr. Banu DİRİ

Co-Adviser: Assist. Prof. Dr. R. Murat DEMİRER

Gene expression profiling (GEP) and genome-wide association studies (GWAS) are powerful tools that can provide list of genes that are related to the pathogenesis of a disease, but it is still a challenge to understand how multiple genes that have modest association with the phenotype interact and contribute to it. For this purpose, it is required to consider molecular profiles with biological interactions. In this work, we proposed two active module identification methods: an active subnetwork search method based on genetic algorithm and a network propagation method. We aimed to understand affected paths in interaction networks and reveal underlying disease mechanisms. We applied our methods to rheumatoid arthritis, intracranial aneurysm and Behçet's disease GWAS datasets. The proposed methods could successfully identify pathways that are known to be related to the diseases, and extract new mechanisms.

**Key Words:** Active subnetwork search, pathway impact analysis, genetic algorithm, network propagation

**YILDIZ TECHNICAL UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

# ÖZET

## BİYOLOJİK ETKİLEŞİM AĞLARI İÇİN ETKİ ANALİZİ ALGORİTMALARI

Ozan ÖZIŞIK

Bilgisayar Mühendisliği Bölümü
Doktora Tezi

Tez Danışmanı: Doç. Dr. Banu DİRİ
Eş Danışman: Yrd. Doç. Dr. R. Murat DEMİRER

Gen ifadesi profillemesi ve genom çapında ilişki çalışmaları bir hastalıkla ilişkili gen listeleri verseler de hastalık fenotipiyle ilişkisi zayıf genlerin bir araya gelerek fenotipe nasıl katkıda bulunabildiklerini açıklayamayabilirler. Bunun için moleküler profillerle biyolojik etkileşimlerin birlikte ele alınmaları gerekir. Bu çalışmada iki aktif modül çıkarım metodu önerilmiştir. Bunlardan ilki bir genetik algoritma tabanlı alt-ağ arama metodu, ikincisi ise bir ağ yayılım metodudur. Burada amaçlanan etkileşim ağlarında etkilenen yolları anlayabilmek ve hastalıkların altında yatan mekanizmaları ortaya çıkarmaktır. Önerilen metotları romatizmalı atardamar yangısı, kafa içi anevrizması ve Behçet hastalığı genom çapında ilişki çalışması veri kümelerine uyguladık ve önerilen metotların hastalıkla ilişkisi bilinen yolakları çıkarmanın yanı sıra yeni mekanizmalar çıkarabildiğini gördük.

**Anahtar Kelimeler:** Aktif alt-ağ arama, yolak etki analizi, genetik algoritma, ağ yayılımı

**YILDIZ TEKNİK ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ**

# CHAPTER 1

# INTRODUCTION

## 1.1 Literature Review

Gene expression profiling (GEP) and genome-wide association studies (GWAS) are powerful tools to reveal the genes that underlie genetic diseases. In GEP, activities of thousands of genes are measured and this data can be used to distinguish the differentially expressed genes in a condition. In GWAS, the statistical significance of the associations of single nucleotide polymorphisms (SNP) to a condition is examined and the roles of genetic factors can be determined. Although GEP and GWAS reveal the genes involved in complex phenotypes, it is still a challenge to understand how multiple genes that have modest association with the phenotype interact and contribute to it. In the last decade, there has been many integrative studies that overlay interaction networks with molecular profiles, e.g. differential expression data or GWAS data to identify "active modules". Active module identification is categorized into three classes by Mitra et al. [1]: i) active subnetwork search methods (also known as significant area search), ii) biclustering methods, and iii) network propagation methods. In active subnetwork search methods, it is aimed to find connected subgraphs of genes that contribute to the disease condition collectively in a biological interaction network. In biclustering it is aimed to cluster genes and expression profiles simultaneously. In network propagation methods, molecular profiles are used with directed gene networks to determine their effects on the information flow. In this study, methods in active subnetwork search and network propagation categories have been proposed and detailed literature reviews are given in section 2.1 and section 3.1, respectively.

## 1.2 Aims of the Dissertation

The aim of this work is to develop methods that can extract interaction paths in biological networks which will help us to understand disease mechanisms and enable usage of personal medication.

In this dissertation, an active subnetwork search method, a network propagation method, and a method that combines two existing methods have been proposed, and these methods are applied to Rheumatoid Arthritis (RA), Behçet's Disease (BD), and Intracranial Aneurysm (IA) GWAS datasets, respectively.

## 1.3 Hypothesis

Genome-wide association studies provide lists of genes that are related to the condition but they are incapable of explaining the underlying mechanisms. In addition, using genes with p-values less than $5x10^{-8}$ and discarding others has emerged as a standard in GWAS studies. We hypothesize that genes that have modest association with the phenotype interact and contribute to disease mechanisms, and these mechanisms can be revealed by extracting collectively affected interacting gene groups. In this study, three methods are proposed for this purpose.

Our first method, called ActiveSubnetworkGA, is a two stage genetic algorithm (GA) approach for active subnetwork search problem. In this method, we use active node list chromosome representation, branch swapping crossover operator that is similar to the crossover operator in genetic programming, multicombination of branches in crossover, mutation on duplicate individuals, pruning, and two stage genetic algorithm approach. For the calculation of subnetwork scores, the scoring metric proposed by Ideker et al. [2] is used. We applied this method to Rheumatoid Arthritis (RA) GWAS dataset and it was successful in retrieving RA related pathways.

In the second method, an active subnetwork search method and a network propagation method from the literature are combined. We applied the proposed method to Intracranial Aneurysm (IA) GWAS dataset. Our results demonstrated that applying first active subnetwork search and then the network propagation method improved the retrieved pathways.

The third method is a novel pathway impact analysis method based on network propagation that can reveal affected paths in pathways and give us the chance to observe the mechanisms in detail. By applying this method to Behçet’s disease (BD) GWAS dataset, we could obtain the disease mechanisms mentioned in the literature and new findings that will shed light on the pathogenesis of BD.

# CHAPTER 2

# ACTIVE SUBNETWORK SEARCH

## 2.1 Background and Related Work

Active subnetwork search is the search of connected subgraphs, mostly consisting of high scoring nodes in a biological interaction network. The score of a node represents the significance of a gene, as determined in a condition specific experiment; i.e. microarray study or genome-wide association study (GWAS). Proteins produced by genes interact with each other to carry out certain functions within the body and disruptions in these functions may result in a disease. Active subnetwork search in protein-protein interaction (PPI) networks would facilitate the discovery of the disease associated set of interactions and pathways. Thus, it would help understanding the disease development mechanisms. Active subnetwork search has been used to discover regulatory pathways [2], functional modules [3], [4], cancer markers [5], subnetworks consisting of dysregulated genes [6], [7], candidate subnetworks and genes for complex diseases [8], [9], to distinguish phenotype groups [10], and to predict response to treatment [11]. The seminal work in the active subnetwork search class is the work by Ideker et al. [2]. In this study a scoring method was proposed to convert significance values in differential expression data to scores for both nodes and node groups, and simulated annealing was proposed to search for the highest scoring connected components. Many other methods descend from this study and contribute by different scoring methods or search methods. Guo et al. [12] proposed edge-based scoring that combines the correlation of expression profiles of related nodes and their differential expression values. Sohler et al. [13] followed a greedy approach. In their method, a set of genes with scores exceeding a threshold are selected as seeds and in each iteration the highest scoring neighboring node is added to the subnetwork. Rajagopalan and Agarwal [14] used a heuristic graph search algorithm. In their method, nodes with positive scores that are directly connected to each other are set as initial subnetworks. Starting with the highest scoring subnetwork, depth first search with a limited depth is run to find a path to another subnetwork. These two subnetworks are merged if this operation increases the score of the former subnetwork. Chuang et al. [5] proposed a greedy search method to find the subnetwork markers for breast cancer metastasis. At each step the search considers addition of a protein within a specified network distance *d* to the current subnetwork. Jia et al. used a modified version of this method on GWAS data [8], [9]. In Ulitsky and Shamir's study [3], high-throughput gene expression data is transformed into pairwise similarity measures; and functional modules are identified by searching high similarity subnetworks in the edge-weighted network. In their method, relatively small, high scoring gene sets are found as seeds. These gene sets are later expanded by making changes on sets (adding a node, removing a node, assigning node of a set to another one, merging two sets), and sets are filtered based on significance. In [4], Ulitsky and Shamir considered interaction confidence in addition to pairwise similarities. Dittrich et al. [15] proposed an exact solution for active subnetwork search based on integer linear programming. Other studies on active subnetwork search are [16], [17], [18], [19], [20] that use greedy approach; [11], [21] that use color coding, [22], [23], [24], [25] that use mathematical programming based methods and [26], [27], [28], [29] that use genetic algorithm.

In this study, we propose a genetic algorithm (GA) method for active subnetwork search. We used active node list chromosome representation, branch swapping crossover operator that is similar to the crossover operator in genetic programming, multicombination of branches in crossover, mutation on duplicate individuals, pruning, and two stage genetic algorithm approach. Our method is tested on simulated datasets and Wellcome Trust Case Control Consortium (WTCCC) Rheumatoid Arthritis (RA) dataset. We compared our method with the simple genetic algorithm that is implemented by us and the simulated annealing method proposed in [2] as implemented in the Cytoscape [30] plugin jActive Modules. Several newly proposed works on active subnetwork search [12], [14], [15], [23], [24], [25], [27], [31], [32], [33], [34], [35] are compared to jActiveModules since it is the state of the art, and that is why we also compared our method with jActiveModules.

## 2.2 Materials and Methods

### 2.2.1 Datasets

#### 2.2.1.1 Protein-Protein Interaction (PPI) Data

In this study we used two different human protein-protein interaction (PPI) datasets. The first and smaller dataset is obtained from the supplementary material of Goh et al.'s study [36]. This dataset is composed of two systematic yeast two-hybrid experiments [37], [38] and PPIs obtained from literature by manual curation [37]. There are 61,070 interactions between 10,174 genes (22,052 non-self-interacting, non-redundant interactions) in the dataset. This PPI network is used to create the benchmark active subnetwork datasets. The second dataset is an up-to-date network that we obtained from the Biological General Repository for Interaction Datasets (BioGRID Release 3.4.127) [39]. We removed the self-interactions and redundant data. The dataset contains 174,826 interactions among 16,912 genes. This PPI network is used with real experiment dataset.

#### 2.2.1.2 Genetic Association Data of Rheumatoid Arthritis

A genome-wide association study on 1999 patients with rheumatoid arthritis and 3004 controls is obtained from WTCCC. 500,475 single nucleotide polymorphisms (SNPs) were genotyped on these 5003 samples using Affymetrix GeneChip Human Mapping 500 K Array Set. At the end of the study, 25,027 SNPs with nominal evidence of association were identified (p-values < 0.05). In [40], these SNPs were assigned to genes, genotypic p-values were weighted by the functional scores of the SNPs, and subnetworks were identified. In our study, the final PPI network with assigned significance values is used to determine active subnetwork.

#### 2.2.1.3 Simulated Datasets

Since there are no available benchmarks to test the active subnetwork search algorithms we have decided to build several sets of benchmarking networks with different score distributions based on a real protein protein interaction network. We used the PPI network from Goh et al.'s study [36]. We randomly chose a core subnetwork with selected number of nodes (e.g. 500 nodes in data 1). Then we selected nodes from the core subnetwork (250 nodes in data 1) and assigned high scores in the range of [10-9, 10-2] to them. As these high scoring nodes in the core are close to each other, they will be further called as dense high scoring nodes. In a network it is unlikely for all the high scoring nodes to be clustered in a small core subnetwork, there may be other high scoring nodes distributed around the network. We chose random nodes from the network excluding the nodes in the core (1500 nodes in data 1) and assigned high scores to them. As these are distributed around the whole network, these will be further called sparse high scoring nodes. The remaining nodes that we have not assigned a p-value yet will have low p-values in the range of [0.5, 1] (420 nodes chosen in data 1) or will not have any p-value at all (8000 nodes chosen in data 1). Genes with unknown p-values exist in the simulated datasets because we want to mimic our real dataset that contains p-values for only 4094 genes. The parameters that we have used in the creation of each dataset can be seen in Table 2.1.

Table 2.1 The parameters used for 10 different simulated datasets

| | **Core** | **Dense High Scoring** | **Sparse High Scoring** | **Low Scoring** | **Null** |
|---|---|---|---|---|---|
| Data1 | 500 | 250 | 1500 | 420 | 8000 |
| Data2 | 300 | 300 | 1500 | 370 | 8000 |
| Data3 | 500 | 250 | 1500 | 4420 | 4000 |
| Data4 | 500 | 250 | 200 | 9720 | 0 |
| Data5 | 500 | 250 | 200 | 9720 | 0 |
| Data6 | 800 | 250 | 200 | 9720 | 0 |
| Data7 | 800 | 500 | 400 | 9270 | 0 |
| Data8 | 2000 | 500 | 1000 | 8670 | 0 |
| Data9 | 8500 | 2000 | 0 | 8170 | 0 |
| Data10 | 8500 | 2000 | 0 | 8170 | 0 |

### 2.2.2 Scoring

In the presented study, we followed the scoring scheme that is proposed by Ideker et al. [2]. This scheme is also used in [14] and [8]. In this scheme, significance value $p_i$ of each gene is converted into a z-score using Eq. (2.1). In the equation, $\Phi^{-1}$ is the inverse normal cumulative distribution function. Collective z-score of a subnetwork ($z_A$) is calculated using Eq. (2.2), where $k$ denotes the number of nodes in a subnetwork and $A$ is the set of genes in the subnetwork.

$$z_i = \Phi^{-1}(1 - p_i) \tag{2.1}$$

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \tag{2.2}$$

In jActiveModules, which is the implementation of the method proposed in [2], nodes without p-values are assigned a p-value of 0.5 giving a neutral importance to these nodes. This is a reasonable approximation if there are few null-valued nodes, but it rises the problem of having too many null nodes in the final output subnetwork especially when the method is applied to datasets with more null-valued nodes, e.g., our rheumatoid arthritis dataset which contains 8147 null-valued nodes out of 10174 nodes. In our case if a node is not assigned a value, it means that it is not targeted by the SNPs in the GWAS study so giving neutral importance is misleading. Therefore we have used an alternative value, 0.9999999999999, which is the maximum value that jActiveModules allows, as 1 leads to negative infinity for a z-value.

### 2.2.3 Subnetwork Search by Genetic Algorithm

In this study, a novel genetic algorithm (GA) method is proposed for subnetwork search. To the best of our knowledge, GA is used by Klammer et al. [26], Ma et al. [27], Wu et al. [28] and Amgalan and Lee [29] in the active subnetwork search domain. In [26], GA is used to identify regulated subnetworks and individual proteins from phosphoproteomic data. In [27], a scoring function that jointly measures condition-specific changes of both nodes and edges is proposed, and search for the highest scoring subnetwork is performed by genetic algorithm. In [28], genetic algorithm is used to improve subnetworks found in [5] by a greedy search algorithm. GA runs on the neighborhood of each subnetwork instead of running on the whole network. In these three studies stated above, simple genetic algorithm implementations are used. In the simple genetic algorithm, chromosomes are represented by binary vectors, crossover is performed by swapping vector parts, and mutation is performed by flipping random bits in the vector. In [29] continuous genetic algorithm is used to obtain a weighted maximum clique that takes into account both differential expression of genes and gene-gene correlations.

#### 2.2.3.1 Chromosome Representation

In the simple genetic algorithm, binary vectors are used as chromosomes to represent the network. Cells of the vector determine whether the associated nodes are included in the subnetwork or not. An example is given in Figure 2.1. In the figure, the graph represents the biological network. Two subnetworks and their offspring are given underneath in binary vector representation form. Dark cells represent the nodes that are included in the subnetwork, which will be further referred as the nodes that are in “on” state. Arrows mark the crossover points. Vector swapping is applied as the crossover operator. Although this representation supports the ordinary mutation and crossover operations, it has two disadvantages: i) it does not guarantee connectedness of the discovered subnetworks, and ii) direct implementation is infeasible because of the huge size of the network. Alternatively, here we propose active node list representation, where each individual is a list of nodes in a subnetwork in a random order (Figure 2.2). The nodes in the list can be sorted according to the constant IDs assigned to them to make the representation analogous to the standard representation.

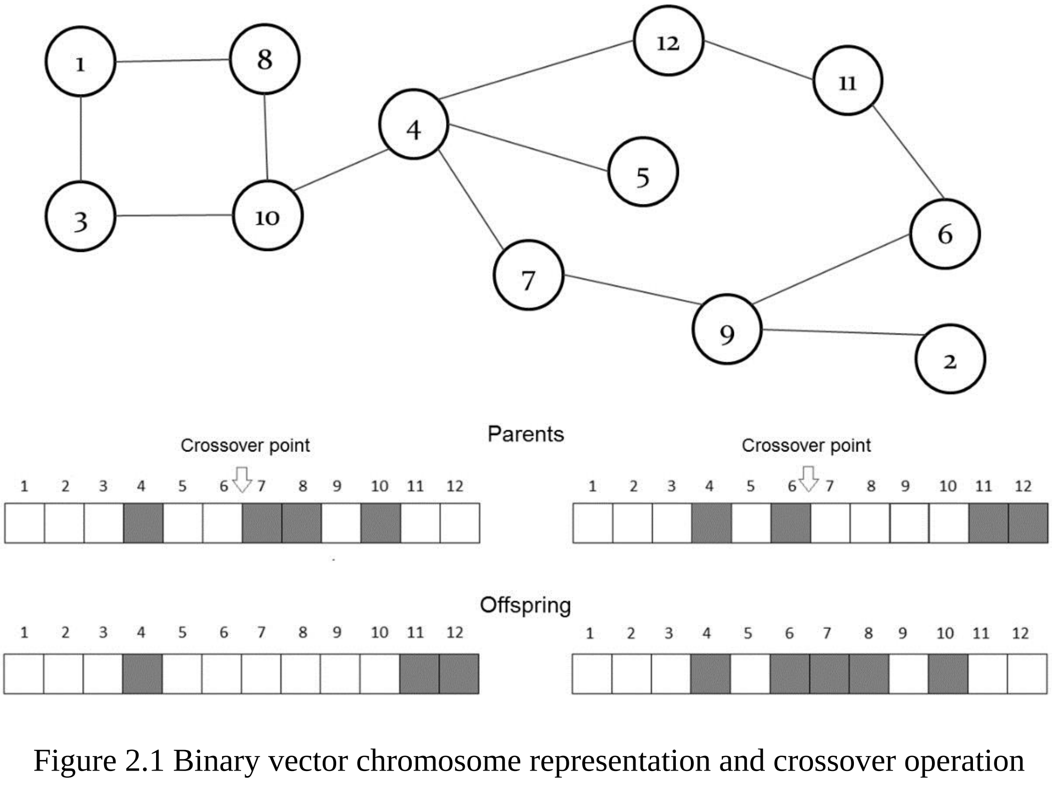

Figure 2.1 Binary vector chromosome representation and crossover operation

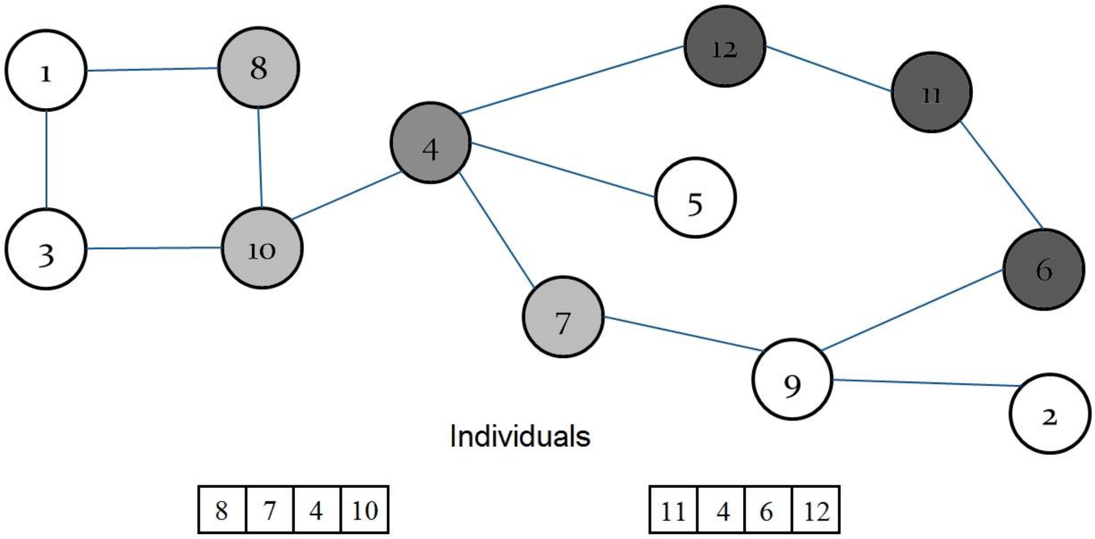

Figure 2.2 Active node list chromosome representation

### 2.2.3.2 Population

In our method, an initial population is created by selecting random nodes from the network, and expanding the subnetworks from these seed nodes using depth first traversal until the determined sizes are reached. The sizes of the initial individuals are set randomly in the [50, 200] range. In the preliminary experiments, we tried different population sizes between 400 and 3200, and observed that increasing the population size improved the obtained scores in some datasets at the cost of increasing running time. We performed our further experiments with population size of 1000 considering obtained scores and increase in time with population size.

### 2.2.3.3 Crossover

In our study, we followed two crossover approaches: i) vector swapping, ii) branch swapping. Vector swapping is the well-known crossover operator in GA, where a crossover point is determined on parents and the offspring are produced by exchanging the parts. While using vector swapping, we stored node list sorted according to node IDs in order to simulate fixed locations in vector representation. Branch swapping is similar to the operation used in genetic programming, where each parent is partitioned into two branches and these branches are swapped. Partitioning is performed by running two depth first traversal (DFT) threads, one starting from the crossover point, which is chosen randomly among the common nodes in two parents, and the other one starting from one of crossover point’s neighbors. The connected nodes are shared between two branches.

Only the node that is acting as the crossover point is common in both branches, other nodes are assigned to a single branch.

In our study, with the help of our node list representation, we use a multicombination approach in crossover. In standard one point crossover, left part of one individual is combined with the right part of the other individual. In our approach, we also check left-left and right-right combinations. We merge nodes in the combined parts, find the biggest connected component, and discard disconnected nodes. In branch swapping we do not see disconnected nodes as all nodes in a branch are connected and all branches can be combined at least by the crossover point. But in vector swapping this can occur. Once the biggest components are found in four combinations, the fittest two are chosen for the next generation. In Figure 2.3, two parents on a network, shown in light gray and dark gray are depicted. Node 1, which is a common node between these individuals, is the crossover point. Branches of the parents and four combinations of the branches are shown, respectively.

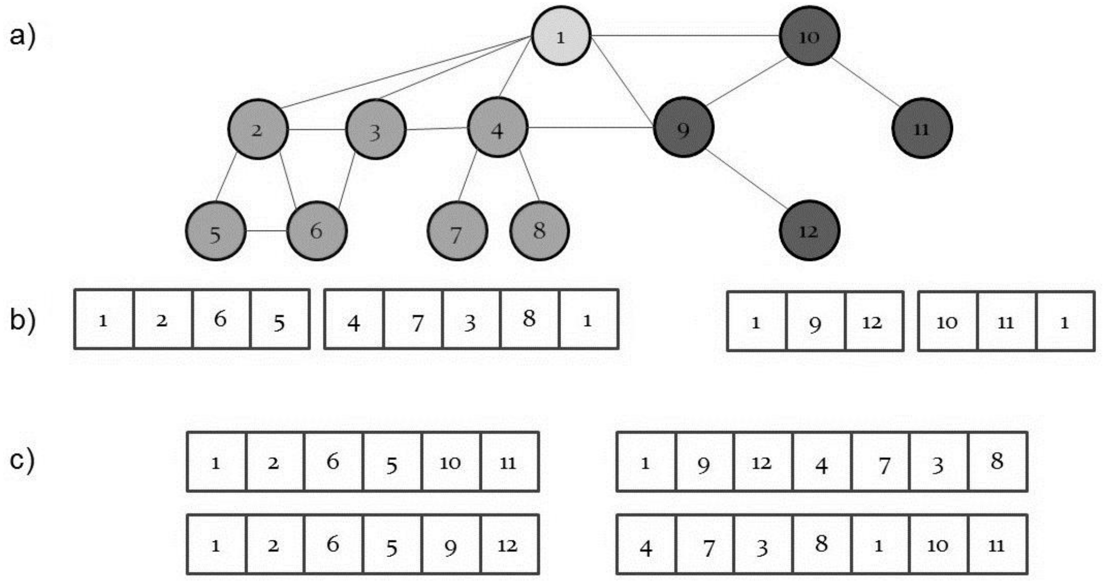

Figure 2.3 Multicombination in crossover

#### 2.2.3.4 Mutation

In the proposed method, mutation is only performed in duplicate individual addition. While adding an individual to the next generation, if the same individual exists in the next generation, instead of adding a duplicate, the individual is added after mutation. Mutation operation is performed by adding *ratioM * subnetworkSize* neighbor nodes to the individual. *ratioM* is a user defined parameter in the range [0, 1].

#### 2.2.3.5 Elitism

Elitism is the strategy of carrying the fittest individuals in one generation to the next generation directly, without any change. Considering our preliminary results on five datasets, we set elite number to 150 for further experiments with the population consisting of 1000 individuals.

#### 2.2.3.6 Stopping Condition

In our implementation, the genetic algorithm stops when an increase in the score does not reach a user defined amount in 100 iterations. Considering the score progressions in preliminary tests, we set this parameter to 0.3.

#### 2.2.3.7 Pruning

Pruning is the final operation of a GA run to remove the low scoring nodes from the best component to improve the solution. First, the nodes in the component are ordered according to their scores. Then, starting with the node with the least score, each node is eliminated successively until the score of the subnetwork starts decreasing. Integrity of the component is checked each time and if there is a disconnection, the node is added back.

#### 2.2.3.8 Two Stage Genetic Algorithm

In this study, we ran the genetic algorithm 30 times and gathered the best solutions. Then we set the best solutions as the initial population and ran the genetic algorithm once again to obtain the best individual. With the help of branch swapping crossover operation, we aimed to combine good solutions.

### 2.2.4 Functional Enrichment of the Identified Subnetwork

After detecting the highest scoring active subnetwork it is important to analyze which biological paths are altered. Functional enrichment is a widely used method to compare the set of identified genes with the set of genes that are known to be part of a biological pathway. In this study, we used ClueGO plugin [41] of Cytoscape for functional enrichment. In ClueGO we focused on KEGG pathways since KEGG database primarily categorizes genes into bona-fide biological pathways, and biological interpretation of pathways is more straightforward compared to GO terms. In order to determine the statistical significance of an enrichment of the identified subnetwork, two-sided (Enrichment/Depletion) test based on the hypergeometric distribution was chosen, and to correct the p-values for multiple testing, Bonferroni correction method was applied.

In this study we used two different human protein-protein interaction (PPI) datasets. The first and smaller dataset is obtained from the supplementary material of Goh et al.'s study [36]. This dataset is composed of two systematic yeast two-hybrid experiments [37], [38] and PPIs obtained from literature by manual curation [37]. There are 61,070 interactions between 10,174 genes (22,052 non-self-interacting, non-redundant interactions) in the dataset. This PPI network is used to create the benchmark active subnetwork datasets. The second dataset is an up-to-date network that we obtained from the Biological General Repository for Interaction Datasets (BioGRID Release 3.4.127) [39]. We removed the self-interactions and redundant data. The dataset contains 174,826 interactions among 16,912 genes. This PPI network is used with real experiment dataset.

## 2.3 Results

In this study, we propose a two stage genetic algorithm method with alternative crossover and mutation operators to the active subnetwork search problem. In order to evaluate performance of our method, we performed comparative experiments on WTCCC RA dataset and 10 simulated datasets using our genetic algorithm method, a simple genetic algorithm implementation, and jActiveModules.

On the proposed genetic algorithm, we performed our tests using both vector swapping and branch swapping crossover operators. We used the proposed node list representation in both crossover operators, but nodes were sorted while using vector swapping in order to simulate fixed locations in vector representation. Mutation of duplicate individuals and two stage genetic algorithm method are applied the same way in trials with both crossover operators. Second stage of two stage genetic algorithm method was run with branch swapping even the crossover operator was vector swapping in the first stage. In the trials we set crossover rate to 0.5, *ratioM* in mutation to 0.005, and population size to 1000.

In the simple genetic algorithm, we set crossover rate to 0.9 and mutation rate to 0.003 as these parameters gave the highest scores in our preliminary tests with five datasets. For the sake of fairness, we followed the two stage approach with simple genetic algorithm, too. In the second stage, genetic algorithm method was run with vector swapping crossover.

In jActiveModules, we set the number of modules to be found to 1 as our aim is to extract single highest scoring subnetwork. We did not choose the options of adjusting score for size and regional scoring because it was necessary to keep scoring method the same for a fair comparison. We kept starting temperature and ending temperature in default values, 1 and 0.01. In jActiveModules there is an operation that is named quenching, it runs when simulated annealing ends, and switches each node to the opposite state (“on” to “off”, “off” to “on”) to see if this change increases score. This goes on until no improvement occurs. While evaluating our method, we made comparisons with simulated annealing and simulated annealing - quenching combination. In our comparison with simulated annealing, we set iteration number to $10^6$ in jActiveModules. In the comparison with simulated annealing - quenching combination, we tried $10^6$, $10^5$, $10^4$, and 0 iterations of simulated annealing, quenching followed them, and we took the best scores for comparison.

Time is the drawback of our method. Our experiments were performed on PCs with Intel Core i3 3.10 GHz CPUs and 6 GB RAMs. Two stage GA, which requires GA to run 30 times, took approximately 15 hours on smaller PPI and 25 hours on larger PPI. In contrast, $10^6$ iterations of simulated annealing in jActiveModules took about 10 minutes and 35 minutes. Considering the difference in run times, we decided to give jActiveModules more time for search to be fair. We set iteration number to $10^8$, which is the maximum number allowed in jActiveModules. We saw that increasing iteration number did not have any effect, because the solution converged much before. Hence we also changed starting temperature and ending temperature, and set them to 10 and 0.01 respectively considering score range and acceptance probabilities. Run time of jActiveModules was about 11 hours with these settings.

Scores of the subnetworks extracted by each method is given in Table 2.2. Highest two scores are given in boldface for each dataset. In the results of Genetic Algorithms, Single Stage scores are the average scores of 30 GA runs. Two Stage GA results present the scores obtained by giving best individuals obtained in 30 runs to another GA thread. In the simulated annealing - quenching combination column (jActive SA + Quench), best results is obtained using a different combination in each dataset, the details of the combination are indicated by numbers: (1) Only quenching (2) Simulated annealing with $10^6$ iterations, no quenching (3) Simulated annealing with $10^4$ iterations and quenching (4) Simulated annealing with $10^5$ iterations and quenching (5) Simulated annealing with $10^6$ iterations and quenching.

In the table, it is clear that our method outperforms its competitors. Two stage GA found higher scoring subnetworks than the simple GA implementation in all cases. Two stage GA found higher scores than jActiveModules in 10 out of 11 datasets when branch swapping is used in the first stage, and it found higher scores in 9 out of 11 datasets when vector swapping is used. Single stage branch swapping and vector swapping GA methods both outperformed jActiveModules in 9 out of 11 datasets.

When rheumatoid arthritis dataset was used, two stage branch swapping GA extracted a subnetwork consisting of 2107 genes and jActiveModules extracted a subnetwork consisting of 840 genes. The subnetwork extracted by the proposed method contains 788 of the latter subnetwork and all but two of the remaining 1319 genes have SNPs assigned to them.

Table 2.2 Scores of active subnetworks extracted by different methods

| | Simple GA | | jActive | jActive | jActive$10^8$ | Branch Swapping | | Vector Swapping | |
|---|---|---|---|---|---|---|---|---|---|
| | SingleStage | TwoStage | SA$10^6$ | SA+Quench | SA $10^8$ | SingleStage | TwoStage | SingleStage | TwoStage |
| Data1 | 33.8 | 37.1 | 60.9 | **77.4$^{(1)}$** | 70.7 | 71.2 | **81.1** | 74.9 | 77.2 |
| Data2 | 41.5 | 44.3 | 61.6 | **71.5$^{(1)}$** | **69.3** | 63.1 | 66 | 63.4 | 65.2 |
| Data3 | 40.1 | 44.6 | 55.2 | 55.2$^{(2)}$ | 56.2 | 62.8 | **80.1** | 73.4 | **75.8** |
| Data4 | 32.9 | 36.6 | 27.5 | 30.7$^{(3)}$ | 26.6 | 42.3 | **44.6** | 42.1 | **44.4** |
| Data5 | 37.9 | 39.8 | 27.1 | 33.5$^{(4)}$ | 26.3 | 42.2 | **44.4** | 43.8 | **44.7** |
| Data6 | 31.6 | 34.8 | 27.1 | 28.8$^{(1)}$ | 25.4 | 38.6 | **42** | 39 | **39.7** |
| Data7 | 43.1 | 46.5 | 36.4 | 41.7$^{(4)}$ | 35.5 | 55.9 | **66.9** | 63.1 | **65.1** |
| Data8 | 46.8 | 49.2 | 44.4 | 45.3$^{(4)}$ | 42.3 | 59.2 | **77.1** | 65.7 | **72.6** |
| Data9 | 66 | 68.1 | 65 | 65.1$^{(5)}$ | 63.6 | 79.1 | **101** | 92 | **99.4** |
| Data10 | 67.1 | 69.8 | 65.8 | 65.9$^{(5)}$ | 62.6 | 80.5 | **102** | 93.2 | **103** |
| RA | 30.3 | 35.9 | 101 | 101$^{(2)}$ | 104 | 118.3 | **154** | 114.1 | **150** |

After the identification of the highest scoring active subnetwork in a biological network, it is necessary to assess whether this subnetwork is biologically significant. In active subnetwork search domain, biological significance of an extracted subnetwork is both related to the search algorithm and the scoring scheme, but the latter is more fundamental. As it was mentioned before, we borrowed the scoring scheme from Ideker et al. [2]. Thus, score based comparison is our main concern but we also present whether the results were biologically significant. In order to observe and compare the biological significance, we performed functional enrichment on the highest scoring active subnetwork that Two Stage GA identified on RA dataset. We used ClueGO plugin [41] of Cytoscape for functional enrichment of subnetworks.

In Table 2.3, we represent the top 20 significant pathways extracted by Two Stage GA. These pathways are mostly related to immunity and inflammation, cell adhesion and cancers. Most of these pathways (Pathways in cancer, ErbB signaling, Focal adhesion, T cell receptor signaling, MAPK signaling, Cell adhesion molecules (CAMs), Neurotrophin signaling) have been previously found to be associated with RA experimentally [42], [43], [44], [45], [46], [47], [48], [49]. A more detailed discussion regarding the relevance of the identified pathways against RA is presented in the Discussion section.

## 2.4 Discussion

In this part of the study, we proposed a two stage genetic algorithm method for active subnetwork search problem. We applied our method on 10 simulated datasets and a real dataset. We compared our results with the results of a simple genetic algorithm and the results of the simulated annealing implementation, jActiveModules. In the experiments, we saw that our two stage genetic algorithm method outperformed its competitors. Our two stage genetic algorithm with branch swapping in the first stage extracted higher scoring subnetworks than both simulated annealing and simulated annealing - quenching combination in all but one datasets. Our two stage method with vector swapping in the first stage obtained higher scores in 10 of 11 datasets compared to simulated annealing, and it obtained higher scores in 9 of 11 datasets compared to simulated annealing - quenching combination. Our method failed in Data 2, which is a simulated dataset that contains directly connected 300 high scoring nodes, and 1500 other high scoring nodes around the network. The existence of directly connected 300 high scoring nodes is an advantage for local search methods like simulated annealing and quenching. Putting all these nodes together in a solution is harder for crossover than it is for local search methods.

Table 2.3 Functional enrichment of genes in the highest scoring subnetworks found by Two Stage GA with Branch Swapping on RA dataset.

| KEGGTerm | p-value |
|---|---|
| Pathways in cancer | 2.1 E-10 |
| Rap1 signaling pathway [52], [53], [54] | 9.35 E-7 |
| cAMP signaling pathway [55], [56] | 1.46 E-6 |
| cGMP-PKG signaling pathway [55], [56] | 1.73 E-6 |
| ErbB signaling pathway [47], [48], [49] | 2.22 E-6 |
| Focal adhesion [44], [57] | 2.52 E-6 |
| Oxytocin signaling pathway | 2.67 E-6 |
| Dopaminergic synapse [58], [59], [60] | 3.23 E-6 |
| Adrenergic sign. in cardiomyocytes [61], [62] | 7.71 E-6 |
| Cell adhesion molecules (CAMs) [44], [63] | 1.01 E-5 |
| Neurotrophin signaling pathway [64], [65] | 2 E-5 |
| HTLV-I infection [42] | 2.23 E-5 |
| Ras signaling pathway [49], [52] | 2.35 E-5 |
| Insulin secretion | 2.47 E-5 |
| Glutamatergic synapse [66] | 2.63 E-5 |
| T cell receptor sign. pathway [46], [67], [68], [69] | 2.74 E-5 |
| MAPK signaling pathway [70], [71] | 7.89 E-5 |
| Hepatitis B | 2.42 E-4 |
| PI3K-Akt signaling pathway [72], [73], [74] | 2.43 E-4 |
| Prolactin signaling pathway [75], [76], [77], [78] | 2.52 E-4 |

In the trials, we saw that single stage vector swapping GA gives higher scores than single stage branch swapping GA; and two stage GA is more successful when the first stage is branch swapping. There is less node variety in different runs of vector swapping. This is good from a single stage machine learning perspective, but in a two stage method, variety in the first stage can yield better results.

The developed subnetwork identification algorithm together with functional enrichment gave us the tools to investigate the pathogenesis of RA. Here, we would like to discuss the functional relevance of the identified pathways to the pathogenesis of RA.

Pathways in cancer is the most significant term in the functional enrichment results. The association between cancer and RA is not certain and it is an important research subject. In [50], RA is reported to be positively associated with nonHodgkin’s lymphoma, Hodgkin’s disease and lung cancer, and negatively associated with colorectal cancer. In [51], a meta-analysis is performed on 16 studies. It is stated that the risk of breast cancer increased in non-Caucasians patients with RA while it decreased in Caucasian population, and some other subgroups.

Rap1 and Ras signaling pathways are significant pathways in our enrichment results. In [52], Remans et al. stated that deregulated Rap1 and Ras signaling underlied oxidative stress and consequent altered T cell function observed in RA. Role of Rap1 and Ras signaling in RA is also stated in [53]. In [54], Abreu et al. found that maintenance of T cell Rap1 signaling in murine T cells reduced disease incidence and severity in collagen-induced arthritis, and proposed that the restoration of Rap1 function in RA synovial T cells might have therapeutic benefit in RA.

cAMP and cGMP-PKG signaling pathways are the third and fourth most significant terms respectively. In [55], [56], it is stated that rheumatoid arthritis shared epitope (an HLA DRB1-encoded 5-amino acid sequence motif carried by the vast majority of RA patients) acted as a signal transduction ligand that interacted with cell surface calreticulin, triggered nitric oxide mediated signaling events in opposite cells, and this affected cGMP and cAMP.

ErbB signaling pathway is the next most significant pathway. It was found to be important for the development of RA by Menon et al. [48]. In [47] Nah et al. reported the function of epidermal growth factor in the inflammatory process of RA. The role of epidermal growth factor receptor in the pathogenesis and treatment of RA has been stated by Yuan et al. [49].

Focal adhesion pathway is experimentally shown to play a critical role in cellular processes, e.g. osteoclast pathology and angiogenesis, which are known to be important for RA [44]. Recently, Shelef et al demonstrated that the inhibition of Focal adhesion kinase (FAK) in human rheumatoid synovial fibroblasts impaired cellular invasion and migration [57].

In [58], it is found that dopamine released by dendritic cells induced IL-6-Th17 axis and caused aggravation of synovial inflammation of RA. Dopaminergic system is also associated in [59] and [60]. Dopaminergic synapse pathway is among the significant pathways.

The association of adrenergic receptors and RA have been reported in the literature [61], [62]. In [61], it is found that beta2-AR SNPs associated with RA in a population from the northern part of Sweden. In [62], it is concluded that there was a highly significant distortion in the distribution of beta2AR polymorphisms at codon 16 in RA patients which contributed to the genetic background of RA. Adrenergic signaling in cardiomyocytes term found to be significant by our method.

Role of cell adhesion molecules in RA is known [44], [63] and this pathway is among the extracted pathways.

In [64], nerve growth factor (NGF), brain derived neurotrophic factor (BDNF), neurotrophin 3 (NT-3), and neurotrophin 4 (NT-4) concentrations in the serum of spondyloarthritis (SpA), rheumatoid arthritis (RA) and osteoarthritis (OA) patients, and healthy subjects have been compared. It has been reported that NT-4 showed significantly higher concentrations in all disease groups than in healthy controls. They also detected that BDNF was significantly higher in healthy controls than all other groups. In [65], BDNF and NGF mRNA expression in synovial fluid cells has been reported. It has been detected that NGF was expressed significantly higher in RA and SpA patients than in the OA group.

In [42], it is mentioned that there is a connection between HTLV-I and RA, although the direction of causality is unknown.

Glutamatergic synapse is one of the pathways in our top 20 list. In a recent study [66], Bonnet et al. found that KA and AMPA glutamate receptors were expressed in rheumatoid arthritis, and stated that AMPA/KA glutamate receptor antagonists represented a potential treatment for arthritis. This supports the association of glutamatergic synapse with RA.

There are many studies that associate alterations in T cell receptor signaling pathway with RA [46], [67], [68]. Raychaudhuri et al. [46] showed that RA risk alleles highlight genes involved in T-cell activation. Within this pathway, CD28 gene was part of our identified subnetwork. The significance of the costimulatory molecules CD28 and CTLA-4 in the pathologic mechanism of RA has been reported by different genetic association studies [46], [69].

MAPK signaling pathway is among the top 20 significant pathways. Although the effectiveness of current methods that target this pathway is debatable [70], it is known that MAPKs have an important role in joint destruction and transducing inflammation [70], [71].

In [72], [73], [74] PI3K and PI3K-Akt signaling pathway have been pointed as a new therapeutic target for autoimmune diseases like rheumatoid arthritis.

Prolactin has been reported to promote RA pathology in [75], [76], [77]. However in [78], it is reported that high levels of circulating prolactin protected against permanent joint damage and inflammation in experimental arthritis in rats.

When the genes from GWAS dataset that has a p-value less than $5x10^{-8}$ or $5x10^{-6}$ are directly used for enrichment, almost none of the pathways in our results appear as significant. The significant pathways are enriched by 80% HLA genes, and the effects of other genes are missed.

In conclusion, the performed experiments show that the presented approach can successfully extract high scoring subnetworks in simulated datasets and identify significant Rheumatoid Arthritis associated subnetworks. This method can be easily used on the datasets of other complex diseases to detect disease-related active subnetworks.

# CHAPTER 3

# NETWORK PROPAGATION METHOD

## 3.1 Background and Related Work

In network propagation methods, molecular profiles are used with directed gene networks to determine their effects on the information flow. One of the key studies in this domain is presented by Tarca et al. [79] in which signaling pathway impact analysis (SPIA) method is proposed. In SPIA, two types of evidences are combined to determine whether a signaling pathway was impacted by observed changes: i) overrepresentation analysis of the number of differentially expressed genes, ii) the abnormal perturbation of the pathway caused by the position of the differentially expressed genes and connections in the pathway. They mention that any of the existing overrepresentation analysis or functional class scoring approaches can be used to calculate first evidence as long as it remains independent of the magnitudes of the fold-changes. For the second evidence, perturbation of each pathway is calculated. A gene's perturbation factor is the sum of its own expression change and the signed perturbation that falls to its share from its direct upstreams. The significance of pathway perturbation is calculated using a bootstrap approach in which expression changes are assigned to random genes in the pathway and perturbation score is calculated for 2000 times, and real perturbation is compared to the calculated distribution. In [80], Li et al. proposed sub-SPIA method that applies SPIA to paths under pathways. These paths are acquired by starting with significant nodes in the pathway and connecting the ones that have at most *n* non-significant genes in-between. A pathway is altered if one of its paths is significantly altered. In [81], a factor graph is created for each gene; entities in the graph represent the copy number of the genome, mRNA expression, protein level, and protein activity. Their method allows the incorporation of many types of omic data as evidence. The method predicts the degree to which a pathway's activities are altered using probabilistic inference. In PathNet proposed in [82], KEGG pathways are combined to create an interaction network. Each gene’s score based on expression data is combined with its neighborhood score using the Fisher’s method. The neighborhood score is calculated by adding the logarithms of its neighbors’ scores and then calculating the obtained score’s significance using a bootstrap approach. Related pathways are obtained using hypergeometric test. In [83], path with maximum running sum score is found for each pathway. Rotation test is used to infer the significance. The maximum scoring path is used to test the null hypothesis about the expression of the entire pathway. In [84] pathways are handled as collections of stimulus-response circuits containing alternative paths that lead an input node to an output node, each circuit’s activation probability is calculated and the number of significantly activated circuits are compared between two conditions. Gene expression measurements are transformed to node probabilities. A circuit’s activation probability is calculated by multiplying probabilities of nodes on the same path and calculating the probability of activation of at least one path.

In this section, we present an application of signaling pathway impact analysis method proposed by Tarca et al. [79] combined by the heuristic active subnetwork search method proposed by Rajagopalan and Agarwal [14]. We first apply the active subnetwork search method and then apply SPIA on KEGG pathways using the p-values of the genes in the subnetwork. SPIA is applied on the extracted subnetwork because the GWAS dataset is only a set of huge number of genes, their relation is not considered, many unrelated pathways appear in enrichment results. This is why we first apply active subnetwork search, and then apply SPIA. We applied the method on intracranial aneurysm (IA) GWAS datasets from European and Japanese populations.

## 3.2 Materials and Methods

### 3.2.1 Datasets

We evaluated the methods on intracranial aneurysm (IA) GWAS datasets obtained from

European and Japanese populations. European population IA dataset is obtained from the work of Yasuno et al. [85] in which results of a GWAS on Finnish, Dutch and Japanese cohorts are presented. In this study we use the data from European population that includes 2780 cases and 12,515 controls. Japanese population IA dataset is the result of GWAS carried out by Akiyama et al. [86] on 1069 Japanese IA patients and 904 Japanese controls. SNPs in these datasets were assigned to genes by Bakir-Gungor and Sezerman in [87].

Protein interaction data is obtained from BioGRID repository [88]. After the removal of recurring and self interactions, there are 19763 nodes and 255611 edges.

In this study we used pathway data from Kyoto Encyclopedia of Genes and Genomes (KEGG) repository [89], [90]. The download date of the pathways is February 25, 2016. Some small corrections have been made on the provided KEGG XML files which contradict with the provided graphics. In Antigen Processing and Presentation pathway, some genes that formed a group in the graphic provided by KEGG, were not connected to any gene in the XML. We connected these according to the graphic. We accepted MHCI, BiP, CANX, BRp57, CALR, β2M as a gene group; HSP70 and HSP90 as a gene group; GILT, AEP and CTSB as a gene group. In Circadian rhythm pathway, the state changes between two Per nodes and two Cry nodes were absent in the xml file in contrast to the graphic. We added these connections. In TGF-beta signaling pathway BMPR I and II, Activin I and II, NodalR I and II were not groups, Smad4, ERK and TGIF were not leading to the nodes they seemed to lead in the graphic; we fixed these.

### 3.2.2 Active Subnetwork Search Method

In [14], Rajagopalan and Agarwal proposed a heuristic active subnetwork search method. The steps of the method is as follows:

(1) Assign p-values to nodes, assign 1 if p-value is missing in the dataset for that node, and calculate node scores.

(2) Group connected nodes with positive scores as initial subnetworks.

(3) Select a previously unselected subnetwork (subnetwork A) in decreasing order of score. If all the subnetworks are used, go to step 5.

(4) Find a neighbor subnetwork with distance *d* (subnetwork B), merge A and B, if score of the new subnetwork (A’) is greater than the score of A keep the change for A and restart step 3. Otherwise reject the merge and continue the search until candidates end, and go to step 3 to select the next subnetwork.

(5) Perform pruning.

In [14] the scoring method proposed by Ideker et al. [2] is modified to decrease the extracted subnetwork size. A value is subtracted from z-scores of genes such that all nodes with p-value greater than a threshold will have a negative score. If a node is neighbor of a node with a high degree, the chance of being found increases. They also propose a score correction to make a node's score negative if its score is smaller than a p-value that is expected by chance. They use 0.01 as the score threshold.

In our application we applied the first correction method that makes scores of nodes with p-values less than a threshold negative. The expected p-value correction resulted getting subnetworks with single nodes. We also skipped step 2 because our PPI network is a densely connected subnetwork and connecting all positive scoring nodes without checking if the operation increases the subnetwork score or not, creates a huge subnetwork.

### 3.2.3 Signaling Pathway Impact Analysis

Signaling pathway impact analysis (SPIA) method is proposed by Tarca et al. [79]. In this method two evidences related to the impact on pathway are combined to obtain a global probability $P_G$: i) overrepresentation analysis of the number of differentially expressed genes ($P_{NDE}$), ii) the abnormal perturbation of the pathway caused by the position of the differentially expressed genes and connections in the pathway ($P_{PERT}$).

$P_{NDE}$ is calculated using any over-representation analysis (ORA) or functional class scoring (FCS) method in which the probability remains independent of the magnitudes of the fold-changes. We used hypergeometric test using set of genes with SNPs.

$P_{PERT}$ is the significance of the observed perturbation based on network topology and magnitudes of the fold-changes. A pathway's perturbation is based on the perturbation factors of its genes calculated using Eq. (3.1). A gene's perturbation factor is the sum of its own expression change and the signed perturbation that falls to its share from its direct upstreams.

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{n} \beta_{ij} \cdot \frac{PF(g_j)}{N_{ds}(g_j)} \qquad (3.1)$$

$\Delta E(g_i)$ represents the signed normalized measured expression change of the gene $g_i$ (log fold-change if two conditions are compared). As we are working on GWAS datasets, we used z-score calculated from the significance of the SNP instead of log fold-change.

The second term in Eq. (3.1) is the sum of perturbation factors of the genes $g_j$ directly upstream of the target gene $g_i$, normalized by the number of downstream genes of each such gene $N_{ds}(g_j)$. $\beta_{ij}$ is +1 for induction and -1 for inhibition.

Perturbation accumulation at gene level is calculated by Eq. (3.2). This subtraction is performed to ensure that a gene that is not connected to any other gene will not contribute to perturbation evidence. This gene already contributes to the first evidence.

$$Acc(g_i) = PF(g_i) - \Delta E(g_i) \tag{3.2}$$

Total accumulation of a pathway can be calculated using matrix operations in Eq. (3.3). $B$ and $\Delta E$ matrices are given in Eq. (3.4) and Eq. (3.5), respectively.

$$Acc = B.(I - B)^{-1}.\Delta E \tag{3.3}$$

$$B = \begin{pmatrix} \frac{\beta_{11}}{N_{ds}(g_1)} & \frac{\beta_{12}}{N_{ds}(g_2)} & \cdots & \frac{\beta_{1n}}{N_{ds}(g_n)} \\ \frac{\beta_{21}}{N_{ds}(g_1)} & \frac{\beta_{22}}{N_{ds}(g_2)} & \cdots & \frac{\beta_{2n}}{N_{ds}(g_n)} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\beta_{n1}}{N_{ds}(g_1)} & \frac{\beta_{n2}}{N_{ds}(g_2)} & \cdots & \frac{\beta_{nn}}{N_{ds}(g_n)} \end{pmatrix} \tag{3.4}$$

$$\Delta E = \begin{pmatrix} \Delta E(g_1) \\ \Delta E(g_2) \\ \cdots \\ \Delta E(g_n) \end{pmatrix} \tag{3.5}$$

The significance of perturbation is calculated using a bootstrap approach in which log fold changes (z-scores in our case) from the whole dataset is assigned to random genes in the pathway, total accumulation is calculated for each trial, and the number of cases that real accumulation exceeds the artificial accumulation is counted. The number of trials is 2000 in the original work and we used the same parameter. One change we applied is this, we did not use the whole dataset for bootstrap, we swapped the z-scores in the pathway. This lets us obtain the contribution of the gene's position in the pathway and pathway topology, without change in total z-score. $P_{NDE}$ and $P_{PERT}$ are combined using Eq. (3.6) and Eq. (3.7).

$$c = P_{NDE}.P_{PERT} \tag{3.6}$$

$$P_G = c - c.\ln(c) \tag{3.7}$$

## 3.3 Results

Results of signaling pathway impact analysis method combined with active subnetwork search are presented in this section.

We evaluated SPIA method in two ways: i) applying SPIA on KEGG pathways using the p-values in the whole aneurysm GWAS dataset, ii) first applying active subnetwork search proposed by Rajagopalan and Agarwal and then applying SPIA on KEGG pathways using the p-values of the genes in the extracted subnetwork.

In order to check the validity of the results we searched “aneurysm” keyword in KEGG Disease Pathways Database and compared the pathways identified by the method with the pathways in the database. The search in KEGG database returns three aneurysm related disease terms and 13 pathways. The disease terms are i) Hereditary angiopathy with nephropathy, aneurysms, and muscle cramps (HANAC), ii) Loeys-Dietz syndrome (LDS), iii) Familial thoracic aortic aneurysm and dissection (TAAD) and Aortic aneurysm familial thoracic type (AAT). The pathways are MAPK signaling pathway, Calcium signaling pathway, Cytokine-cytokine receptor interaction, Endocytosis, Vascular smooth muscle contraction, TGF-beta signaling pathway, Osteoclast differentiation, Hippo signaling pathway, Focal adhesion, ECM-receptor interaction, Adherens junction, Tight junction, and Regulation of actin cytoskeleton.

In Table 3.1 and Table 3.2, the results of SPIA alone on European and Japanese IA datasets are given. In Table 3.3 and Table 3.4 the results of applying SPIA on the subnetworks extracted from European and Japanese IA datasets using active subnetwork search are given. Known IA related pathways are given in boldface. It should be noted that the subnetwork search may result different subnetworks, so we ran the method ten times, applied SPIA on these and used the highest p-values assigned to each pathway in these trials. In active subnetwork search method, score correction threshold was set to 0.01 for European population dataset as it is the value given in the study. For the Japanese population dataset, given results are obtained with the threshold set to 0.02. We made experiments using the default threshold 0.01 and saw that the extracted subnetwork was small, and SPIA also identified a few pathways using the genes in the subnetwork. In the tables pathways names and p-values are given for both $P_{NDE}$ and $P_G$ to see if combining $P_{NDE}$ with $P_{PERT}$ to get $P_G$ improves the results.

Table 3.1 Output pathways of SPIA on European population IA GWAS dataset

| **Pathway** | **$P_{NDE}$** | | **Pathway** | **$P_G$** |
|---|---|---|---|---|
| Glutamatergic synapse | 2.64E-08 | | Glutamatergic synapse | 2.16E-07 |
| Morphine addiction | 4.72E-08 | | Morphine addiction | 3.69E-07 |
| Circadian entrainment | 6.33E-07 | | Long-term depression | 4.49E-07 |
| **Calcium signaling pathway** | 1.66E-06 | | Circadian entrainment | 4.42E-06 |
| Dilated cardiomyopathy | 4.03E-06 | | Vibrio cholerae infection | 9.27E-06 |
| Long-term depression | 8.07E-06 | | **Calcium signaling pathway** | 9.91E-06 |
| Arrhythmogenic right ventricular cardiomyopathy (ARVC) | 2.53E-05 | | Dilated cardiomyopathy | 2.99E-05 |
| Retrograde endocannabinoid signaling | 2.98E-05 | | Oxytocin signaling pathway | 5.14E-05 |
| Cholinergic synapse | 5.00E-05 | | Arrhythmogenic right ventricular cardiomyopathy (ARVC) | 1.89E-04 |
| Oxytocin signaling pathway | 7.48E-05 | | Retrograde endocannabinoid signaling | 2.32E-04 |
| **Adherens junction** | 7.59E-05 | | Thyroid hormone synthesis | 3.56E-04 |
| cGMP-PKG signaling pathway | 8.26E-05 | | **Adherens junction** | 4.98E-04 |
| **Vascular smooth muscle contraction** | 3.36E-04 | | Cholinergic synapse | 5.20E-04 |
| Gap junction | 3.80E-04 | | cGMP-PKG signaling pathway | 6.18E-04 |
| GABAergic synapse | 3.80E-04 | | Salivary secretion | 7.22E-04 |
| cAMP signaling pathway | 4.23E-04 | | cAMP signaling pathway | 1.20E-03 |
| Salivary secretion | 4.77E-04 | | Cell adhesion molecules (CAMs) | 1.60E-03 |
| Axon guidance | 5.68E-04 | | Axon guidance | 2.00E-03 |
| Thyroid hormone synthesis | 7.28E-04 | | Gap junction | 2.00E-03 |
| Cell adhesion molecules (CAMs) | 8.26E-04 | | GABAergic synapse | 2.20E-03 |
| Pancreatic secretion | 8.93E-04 | | **Vascular smooth muscle contraction** | 2.30E-03 |
| Inflammatory mediator regulation of TRP channels | 1.30E-03 | | Pancreatic secretion | 3.70E-03 |
| Serotonergic synapse | 1.70E-03 | | Rap1 signaling pathway | 3.90E-03 |
| Renin secretion | 2.30E-03 | | Transcriptional misregulation in cancer | 5.20E-03 |
| Pathways in cancer | 3.50E-03 | | Inflammatory mediator regulation of TRP channels | 5.70E-03 |

Table 3.2 Output pathways of SPIA on Japanese population IA GWAS dataset

| **Pathway** | **pNDE** | | **Pathway** | **pG** |
|---|---|---|---|---|
| Glutamatergic synapse | 5.87E-10 | | Glutamatergic synapse | 7.70E-09 |
| Axon guidance | 2.72E-07 | | Axon guidance | 2.57E-06 |
| **ECM-receptor interaction** | 2.49E-05 | | Gap junction | 5.96E-06 |
| PI3K-Akt signaling pathway | 2.56E-05 | | **ECM-receptor interaction** | 3.85E-05 |
| Arrhythmogenic right ventricular cardiomyopathy (ARVC) | 3.15E-05 | | PI3K-Akt signaling pathway | 7.53E-05 |
| **Focal adhesion** | 3.52E-05 | | **Focal adhesion** | 2.22E-04 |
| Circadian entrainment | 6.13E-05 | | Circadian entrainment | 2.36E-04 |
| Protein digestion and absorption | 1.60E-04 | | Arrhythmogenic right ventricular cardiomyopathy (ARVC) | 2.54E-04 |
| AGE-RAGE signaling pathway in diabetic complications | 2.12E-04 | | Rap1 signaling pathway | 6.07E-04 |
| Rap1 signaling pathway | 2.48E-04 | | cGMP-PKG signaling pathway | 9.06E-04 |
| Long-term potentiation | 3.93E-04 | | **Tight junction** | 1.20E-03 |
| Long-term depression | 8.38E-04 | | AGE-RAGE signaling pathway in diabetic complications | 1.50E-03 |
| cGMP-PKG signaling pathway | 9.13E-04 | | Protein digestion and absorption | 1.60E-03 |
| Insulin secretion | 1.10E-03 | | Insulin secretion | 1.60E-03 |
| **Calcium signaling pathway** | 1.20E-03 | | Long-term depression | 2.30E-03 |
| Cholinergic synapse | 1.20E-03 | | Long-term potentiation | 3.20E-03 |
| Aldosterone synthesis and secretion | 1.30E-03 | | cAMP signaling pathway | 4.90E-03 |
| cAMP signaling pathway | 1.60E-03 | | **Adherens junction** | 5.60E-03 |
| **Tight junction** | 1.60E-03 | | Aldosterone synthesis and secretion | 7.30E-03 |
| Ras signaling pathway | 2.80E-03 | | Cholinergic synapse | 7.40E-03 |
| Retrograde endocannabinoid signaling | 2.90E-03 | | **Calcium signaling pathway** | 7.40E-03 |
| Morphine addiction | 3.10E-03 | | Inositol phosphate metabolism | 8.80E-03 |
| Phospholipase D signaling pathway | 3.10E-03 | | Phospholipase D signaling pathway | 9.80E-03 |
| Pathways in cancer | 4.00E-03 | | Morphine addiction | 9.90E-03 |
| Oxytocin signaling pathway | 4.40E-03 | | Ras signaling pathway | 11.20E-03 |

Table 3.3 Output pathways of SPIA on the subnetwork extracted from European population IA GWAS dataset by active subnetwork search

| **Pathway** | **pNDE** | | **Pathway** | **pG** |
|---|---|---|---|---|
| Pathways in cancer | 1.43E-06 | | Pathways in cancer | 2.05E-05 |
| **Adherens junction** | 1.06E-05 | | **Adherens junction** | 1.29E-04 |
| **Hippo signaling pathway** | 1.96E-05 | | **Hippo signaling pathway** | 1.39E-04 |
| Chronic myeloid leukemia | 8.86E-04 | | **Vascular smooth muscle contraction** | 6.71E-04 |
| GnRH signaling pathway | 9.86E-04 | | Bacterial invasion of epithelial cells | 7.56E-04 |
| **Tight junction** | 1.10E-03 | | Chronic myeloid leukemia | 1.70E-03 |
| Bacterial invasion of epithelial cells | 1.20E-03 | | Epithelial cell sign. in Helicobacter pylori infection | 2.50E-03 |
| Inflammatory mediator regulation of TRP channels | 1.70E-03 | | GnRH signaling pathway | 3.00E-03 |
| Pancreatic cancer | 2.10E-03 | | Circadian entrainment | 5.80E-03 |
| Thyroid cancer | 2.20E-03 | | Gap junction | 6.60E-03 |
| **TGF-beta signaling pathway** | 2.20E-03 | | Pancreatic secretion | 6.70E-03 |
| Circadian entrainment | 2.80E-03 | | Inflammatory mediator regulation of TRP channels | 7.10E-03 |
| Insulin secretion | 2.80E-03 | | Dopaminergic synapse | 7.20E-03 |
| cAMP signaling pathway | 3.40E-03 | | **Tight junction** | 7.80E-03 |
| Gap junction | 3.50E-03 | | Shigellosis | 8.50E-03 |
| Proteoglycans in cancer | 3.70E-03 | | Thyroid cancer | 9.20E-03 |
| **Focal adhesion** | 3.90E-03 | | Insulin secretion | 1.23E-02 |
| Shigellosis | 7.20E-03 | | Pancreatic cancer | 1.34E-02 |
| **MAPK signaling pathway** | 7.70E-03 | | **TGF-beta signaling pathway** | 1.55E-02 |
| **Endocytosis** | 8.50E-03 | | Thyroid hormone synthesis | 1.56E-02 |
| **Vascular smooth muscle contraction** | 9.00E-03 | | ErbB signaling pathway | 1.57E-02 |
| Inflammatory bowel disease (IBD) | 9.80E-03 | | Amphetamine addiction | 1.57E-02 |
| ErbB signaling pathway | 1.01E-02 | | **Focal adhesion** | 1.61E-02 |
| Adrenergic signaling in cardiomyocytes | 1.01E-02 | | Rap1 signaling pathway | 1.88E-02 |
| Morphine addiction | 1.12E-02 | | mRNA surveillance pathway | 1.90E-02 |

Table 3.4 Output pathways of SPIA on the subnetwork extracted from Japanese population IA GWAS dataset by active subnetwork search

| **Pathway** | **pNDE** | | **Pathway** | **pG** |
|---|---|---|---|---|
| **Endocytosis** | 9.19E-06 | | **Endocytosis** | 3.50E-05 |
| Rap1 signaling pathway | 2.73E-04 | | **ECM-receptor interaction** | 1.00E-04 |
| AGE-RAGE signaling pathway in diabetic complications | 3.86E-04 | | T cell receptor signaling pathway | 4.02E-04 |
| T cell receptor signaling pathway | 5.11E-04 | | AGE-RAGE signaling pathway in diabetic complications | 1.00E-03 |
| **Tight junction** | 6.79E-04 | | Ras signaling pathway | 1.40E-03 |
| Ras signaling pathway | 7.05E-04 | | Rap1 signaling pathway | 1.70E-03 |
| HIF-1 signaling pathway | 1.00E-03 | | **Adherens junction** | 1.90E-03 |
| Protein digestion and absorption | 1.30E-03 | | **Tight junction** | 2.10E-03 |
| **ECM-receptor interaction** | 1.50E-03 | | Fatty acid biosynthesis | 6.20E-03 |
| **MAPK signaling pathway** | 2.00E-03 | | Oxytocin signaling pathway | 6.40E-03 |
| PI3K-Akt signaling pathway | 2.20E-03 | | PI3K-Akt signaling pathway | 6.80E-03 |
| Fatty acid biosynthesis | 2.20E-03 | | HIF-1 signaling pathway | 6.80E-03 |
| Pathways in cancer | 2.80E-03 | | Protein digestion and absorption | 9.10E-03 |
| Endocrine and other factor-regulated calcium reabsorption | 5.00E-03 | | **MAPK signaling pathway** | 9.40E-03 |
| Aldosterone synthesis and secretion | 5.60E-03 | | Gap junction | 1.16E-02 |
| Oxytocin signaling pathway | 7.10E-03 | | Pathways in cancer | 1.46E-02 |
| Axon guidance | 8.10E-03 | | Endocrine and other factor-regulated calcium reabsorption | 1.68E-02 |
| cAMP signaling pathway | 8.10E-03 | | Neurotrophin signaling pathway | 1.68E-02 |
| **Focal adhesion** | 1.39E-02 | | cAMP signaling pathway | 1.89E-02 |
| Long-term depression | 1.74E-02 | | Axon guidance | 2.34E-02 |
| Fc gamma R-mediated phagocytosis | 2.10E-02 | | Aldosterone synthesis and secretion | 2.61E-02 |
| Aldosterone-regulated sodium reabsorption | 2.16E-02 | | **Focal adhesion** | 3.82E-02 |
| Long-term potentiation | 2.87E-02 | | Rheumatoid arthritis | 4.10E-02 |
| Inflammatory mediator regulation of TRP channels | 2.90E-02 | | p53 signaling pathway | 4.77E-02 |
| Signaling pathways regulating pluripotency of stem cells | 2.93E-02 | | Long-term depression | 4.86E-02 |

## 3.4 Discussion

In this part of the study we combined an active subnetwork search method with a signaling pathway impact analysis (SPIA) method to identify intracranial aneurysm (IA) related KEGG pathways. Active subnetwork search is applied to the human PPI with p-values to obtain the highly affected subnetwork. Then SPIA is ran on KEGG pathways using the genes in the subnetwork.

Our results demonstrate that applying SPIA after subnetwork search improves the outputs. In Table 3.1 and Table 3.2 results of direct application of SPIA on European population and Japanese population datasets are given. Pathways that are known to be related to intracranial aneurysm are in boldface. There are 4 known pathways in Table 3.1 and 5 pathways in Table 3.2. The results of first applying active subnetwork search and then SPIA are given in Table 3.3 and Table 3.4. There are 6 known pathways in each table according to the $P_G$ results. The number of known pathways and their ranks increase by applying subnetwork search first.

In the tables $P_{NDE}$ and $P_G$ results are given side by side to analyze the effect of using perturbation evidence proposed by Tarca et al. [79]. In the results we did not see a significant contribution of perturbation, on the contrary, some known pathways in $P_{NDE}$ column of Table 3.3 became less significant and did not appear in $P_G$ list.

Our results demonstrate that common pathways are affected in European and Japanese populations, which is consistent with the results of Bakir-Gungor and Sezerman [87]. 18 of the 36 pathways identified in Japanese population according to $P_{NDE}$ are among the 46 pathways in European population, and 10 of the 26 pathways in Japanese population according to $P_G$ are among the 35 pathways in European population.

# CHAPTER 4

# PROBABILISTIC SUB-PATHWAY IMPACT ANALYSIS

## 4.1 Introduction

Sebastian-Leon et al. [84] proposed a method that handles pathways as collections of stimulus-response circuits containing alternative paths that lead an input node to an output node. Each circuit's activation probability is calculated and the number of significantly activated circuits are compared between two conditions. Gene expression measurements are transformed to node probabilities. A circuit's activation probability is calculated by multiplying probabilities of nodes on the same path and calculating the probability of activation of at least one path. Detailed literature review in this subject can be found in section 3.1.

In this study, we modified the probabilistic approach in [84] and applied our proposed method on Behçet's disease (BD) GWAS datasets from Turkish and Japanese populations. In [84], a circuit's activation probability is the probability of activation of at least one of the alternative paths it contains. Pathways are static pictures of the interactions and existence of alternative paths does not mean that the alternatives will always be available. Instead of circuits, our approach extracts affected paths. We also propose a scoring method that is insensitive to path length.

## 4.2 Materials and Methods

### 4.2.1 Datasets

BD GWAS datasets are obtained from Turkish and Japanese populations. Turkish population GWAS was conducted by Remmers et al. [91] on 1215 BD cases and 1278 unaffected controls. Japanese population GWAS was conducted by Mizuki et al. [92] on 612 Japanese individuals with BD and 740 healthy controls. SNPs in these datasets were assigned to genes by Bakir-Gungor et al. in [93].

We used pathway data from Kyoto Encyclopedia of Genes and Genomes (KEGG) repository [89], [90]. The download date of the pathways is February 25, 2016. Some small corrections have been made on the provided KEGG XML files which contradict with the provided graphics, these are explained in the previous chapter.

### 4.2.2 Proposed Probabilistic Path Impact Analysis Method

In this study we propose a network propagation method to analyze the impact on pathways by finding highly affected paths of the pathways. A path in a pathway is composed of an input node, an output node, and the intermediate nodes that constitute the path in-between. If there are alternatives in-between, each is accepted as a distinct path.

#### 4.2.2.1 Crosstalk Treatment

Donato et al. [94] showed that crosstalk can cause a biologically non-significant pathway become statistically significant and vice versa in enrichment applications. In order to prevent common genes with high scores from leading to many high scoring irrelevant paths, we decreased the significance of those genes (Eq. (4.1)). p-value of a gene is not changed if the number of pathways it belongs to ($k$) is less than or equal to a user defined crosstalk tolerance ($cT$). Otherwise, p-value of the gene is multiplied by 10 to 100 depending on the number of pathways the gene belongs to. p-value can be 0.5 at maximum.

$$pValue_{gene} = \begin{cases} pValue_{gene} & k \leq cT \\ min\left\{0.5,\ pValue_{gene} \times min\{(k - cT + 1) \times 5, 100\}\right\} & otherwise \end{cases} \tag{4.1}$$

#### 4.2.2.2 Probabilities of Nodes

In the GWAS data, we have p-values (<0.05) for SNPs assigned to genes. We assigned $1 - pValue$ as the impact probability of a gene in the pathway if the gene is in the GWAS dataset, 0.5 otherwise (Eq. (4.2)).

$$p_{gene} = \begin{cases} 1 - pValue_{gene} & if\ pValue\ exists \\ 0.5 & otherwise \end{cases} \tag{4.2}$$

In KEGG database, there are nodes that are composed of multiple alternative genes. There are also gene groups which correspond to complexes of gene products, mostly protein complexes as it is stated in KEGG Markup Language Document [95]. The probability of such a node with multiple genes is the maximum of the probabilities of the genes it contains (Eq. (4.3)).

$$p_{node} = \begin{cases} p_{gene} & if\ node\ is\ a\ single\ gene \\ \max\limits_{gene \in node} p_{gene} & if\ node\ contains\ multiple\ genes \end{cases} \tag{4.3}$$

#### 4.2.2.3 Impact Score Calculation

A path's impact probability is calculated by multiplying probabilities of nodes on it (Eq. (4.4)). In some pathways (e.g. Jak-STAT signaling pathway) one gene is represented in two consecutive nodes and connected by state change. In this condition, the gene's probability is used only once.

$$p_{impact} = \prod_{node \in path} p_{node} \tag{4.4}$$

In our method, as we do not have a reference probability for comparison, we calculated a base probability for each path by assigning 0.5 to all nodes (Eq. (4.5)). Each path's impact score is calculated using Eq. (4.6).

$$p_{base} = 0.5^{pathLength} \tag{4.5}$$

$$Score_{path} = log_{p_{impact}}(p_{base}) \tag{4.6}$$

This score is chosen over any ratio based scoring because it is not affected by the length of the path. It is close to 1 when base probability and pathway probability are close to each other, and higher when they are significantly different.

#### 4.2.2.4 Overlapping Paths in Two Populations

In Turkish and Japanese populations, if affected paths are similar, these paths are emphasized in the results. The similarity of paths are calculated using Jaccard Index. It is defined as the size of the intersection divided by the size of the union of the sample sets (Eq. (4.7)). Paths that have a Jaccard Index more than 0.5 are represented by the highest scoring path among them.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.7}$$

#### 4.2.2.5 Fusion of Pathway Impacts in Two Populations

In order to fuse the impacts on paths in Turkish and Japanese populations, we used Fisher's method (Eq. (4.8)) to combine p-values of genes in each population, obtained p-value from $\mathrm{X}^2$ test statistic for each gene, and then performed impact analysis using the new p-values. If a gene has a p-value in one population but not in the other, the missing p-value is assumed to be 0.05. If a gene has no p-value in both populations, 0.5 is assigned to the gene in both populations.

$$\mathrm{X}_{2k}^2 \sim -2\sum_{i=1}^{k} \ln(p_i) \tag{4.8}$$

### 4.3 Results

We applied our method on Behçet's disease GWAS datasets of Turkish and Japanese populations.

Crosstalk tolerance is set to 3 in these experiments. Our experiments with other crosstalk tolerance values {1, 2, 3, 4, 5} and without crosstalk treatment showed that the method is not oversensitive to the tolerance parameter. Rank of the high scoring pathways change a few steps up or down, and some new pathways are added to the bottom with the increase of the tolerance. However without crosstalk treatment many irrelevant pathways appear in the list, which showed us that crosstalk treatment is necessary.

Highly affected pathways of each population are given in Table 4.1 and Table 4.2. The pathways given here contain at least one path which has a score that is greater than the score of a hypothetical path with three genes and p-values of {0.5, 0.05, 0.05}. Pathways are sorted by the score of the highest scoring path they contain. Pathways in which very similar paths are affected in Turkish and Japanese populations are marked with *.

Pathways obtained by combining p-values in Turkish and Japanese populations are given in Table 4.3. In the tables related pathways in the literature are in boldface.

Table 4.1 Affected pathways in Behçet's disease in Turkish population

| **No** | **Pathway** | **Score** |
|---|---|---|
| 1 | **Maturity onset diabetes of the young*** | 29.14 |
| 2 | **Notch signaling pathway*** | 24.95 |
| 3 | **mTOR signaling pathway** | 21.04 |
| 4 | **Complement and coagulation cascades** | 20.66 |
| 5 | **Jak-STAT signaling pathway*** | 9.51 |
| 6 | **Hedgehog signaling pathway*** | 9.30 |
| 7 | **Wnt signaling pathway*** | 8.47 |
| 8 | Aldosterone-regulated sodium reabsorption* | 4.78 |
| 9 | Rap1 signaling pathway* | 4.51 |
| 10 | **Antigen processing and presentation*** | 4.08 |
| 11 | Fc gamma R-mediated phagocytosis* | 3.90 |
| 12 | Axon guidance* | 3.81 |
| 13 | Hippo signaling pathway* | 3.76 |
| 14 | Tuberculosis* | 3.17 |
| 15 | HIF-1 signaling pathway | 3.12 |
| 16 | Renal cell carcinoma | 3.11 |
| 17 | MAPK signaling pathway* | 3.02 |
| 18 | Pathways in cancer | 2.94 |
| 19 | Acute myeloid leukemia | 2.94 |
| 20 | Prolactin signaling pathway* | 2.92 |
| 21 | Inflammatory mediator regulation of TRP channels* | 2.92 |
| 22 | Influenza A | 2.91 |
| 23 | Staphylococcus aureus infection | 2.91 |
| 24 | Herpes simplex infection* | 2.87 |
| 25 | Insulin resistance | 2.85 |
| 26 | Signaling pathways regulating pluripotency of stem cells* | 2.84 |
| 27 | ErbB signaling pathway* | 2.82 |
| 28 | Chemokine signaling pathway | 2.82 |
| 29 | cGMP-PKG signaling pathway | 2.72 |
| 30 | Circadian entrainment | 2.72 |
| 31 | Chagas disease (American trypanosomiasis)* | 2.68 |
| 32 | Amoebiasis* | 2.68 |
| 33 | Alcoholism* | 2.68 |
| 34 | NOD-like receptor signaling pathway | 2.65 |
| 35 | Ras signaling pathway* | 2.62 |

Table 4.2 Affected pathways in Behçet’s disease in Japanese population

| No | Pathway | Score |
|---|---|---|
| 1 | **Hedgehog signaling pathway***  | 227.22 |
| 2 | **Circadian rhythm** | 104.96 |
| 3 | **Notch signaling pathway*** | 70.35 |
| 4 | **AMPK signaling pathway** | 57.49 |
| 5 | **Maturity onset diabetes of the young*** | 49.31 |
| 6 | Signaling pathways regulating pluripotency of stem cells* | 46.50 |
| 7 | MAPK signaling pathway* | 43.57 |
| 8 | Ras signaling pathway* | 19.92 |
| 9 | Axon guidance* | 11.50 |
| 10 | Amoebiasis* | 9.72 |
| 11 | **Jak-STAT signaling pathway*** | 9.40 |
| 12 | **Osteoclast differentiation** | 9.03 |
| 13 | Prolactin signaling pathway* | 8.37 |
| 14 | TNF signaling pathway | 7.01 |
| 15 | ErbB signaling pathway* | 6.99 |
| 16 | Acute myeloid leukemia | 5.98 |
| 17 | NOD-like receptor signaling pathway | 5.95 |
| 18 | Tight junction | 5.89 |
| 19 | cGMP-PKG signaling pathway | 5.50 |
| 20 | **TGF-beta signaling pathway** | 5.33 |
| 21 | Aldosterone-regulated sodium reabsorption* | 4.57 |
| 22 | Tuberculosis* | 4.52 |
| 23 | **Wnt signaling pathway*** | 4.38 |
| 24 | Fanconi anemia pathway | 3.92 |
| 25 | Shigellosis | 3.72 |
| 26 | HTLV-I infection | 3.64 |
| 27 | Neurotrophin signaling pathway | 3.57 |
| 28 | Proteoglycans in cancer | 3.50 |
| 29 | Rap1 signaling pathway* | 3.49 |
| 30 | Adherens junction | 3.31 |
| 31 | Influenza A | 3.29 |
| 32 | Inflammatory mediator regulation of TRP channels* | 3.28 |
| 33 | Hippo signaling pathway* | 3.25 |
| 34 | Basal cell carcinoma | 3.22 |
| 35 | Herpes simplex infection* | 3.18 |
| 36 | RIG-I-like receptor signaling pathway | 3.15 |
| 37 | GnRH signaling pathway | 3.11 |
| 38 | PPAR signaling pathway | 2.96 |
| 39 | **Antigen processing and presentation*** | 2.94 |
| 40 | Alcoholism* | 2.93 |

Table 4.2 (cont'd).

| | | |
|---|---|---|
| 41 | Chagas disease (American trypanosomiasis)* | 2.88 |
| 42 | Endometrial cancer | 2.88 |
| 43 | Pathways in cancer | 2.88 |
| 44 | Colorectal cancer | 2.88 |
| 45 | Serotonergic synapse | 2.86 |
| 46 | p53 signaling pathway | 2.86 |
| 47 | Inflammatory bowel disease (IBD) | 2.82 |
| 48 | Insulin secretion | 2.79 |
| 49 | AGE-RAGE signaling pathway in diabetic complications | 2.79 |
| 50 | Fc gamma R-mediated phagocytosis* | 2.77 |
| 51 | Adipocytokine signaling pathway | 2.76 |
| 52 | Progesterone-mediated oocyte maturation | 2.72 |
| 53 | cAMP signaling pathway | 2.68 |
| 54 | Non-alcoholic fatty liver disease (NAFLD) | 2.68 |
| 55 | Dopaminergic synapse | 2.67 |

Table 4.3 Pathways obtained by combining p-values of genes in Turkish and Japanese populations using Fisher's method

| **No** | **Pathway** | **Score** |
|---|---|---|
| 1 | **Hedgehog signaling pathway** | 208.81 |
| 2 | **Maturity onset diabetes of the young** | 108.70 |
| 3 | **Notch signaling pathway** | 44.20 |
| 4 | **Circadian rhythm** | 17.60 |
| 5 | **Jak-STAT signaling pathway** | 13.75 |
| 6 | NOD-like receptor signaling pathway | 12.91 |
| 7 | Tuberculosis | 11.79 |
| 8 | **AMPK signaling pathway** | 11.19 |
| 9 | MAPK signaling pathway | 10.02 |
| 10 | ErbB signaling pathway | 9.61 |
| 11 | Signaling pathways regulating pluripotency of stem cells | 9.20 |
| 12 | **Complement and coagulation cascades** | 7.74 |
| 13 | Tight junction | 6.73 |
| 14 | Ras signaling pathway | 6.40 |
| 15 | **mTOR signaling pathway** | 6.25 |
| 16 | Amoebiasis | 5.68 |
| 17 | **TGF-beta signaling pathway** | 5.30 |
| 18 | Aldosterone-regulated sodium reabsorption | 4.95 |
| 19 | **Wnt signaling pathway** | 4.39 |
| 20 | Axon guidance | 4.38 |
| 21 | Hippo signaling pathway | 3.98 |

Table 4.3 (cont’d)

| | | |
|---|---|---|
| 22 | Fanconi anemia pathway | 3.63 |
| 23 | Rap1 signaling pathway | 3.36 |
| 24 | cGMP-PKG signaling pathway | 3.30 |
| 25 | Chemokine signaling pathway | 3.30 |
| 26 | Prolactin signaling pathway | 3.30 |
| 27 | Proteoglycans in cancer | 3.30 |
| 28 | **Antigen processing and presentation** | 3.01 |
| 29 | Shigellosis | 3.01 |
| 30 | Herpes simplex infection | 2.99 |
| 31 | Inflammatory mediator regulation of TRP channels | 2.98 |
| 32 | Alcoholism | 2.98 |
| 33 | Insulin resistance | 2.93 |
| 34 | Epstein-Barr virus infection | 2.93 |
| 35 | Adherens junction | 2.87 |
| 36 | Adipocytokine signaling pathway | 2.85 |
| 37 | Neurotrophin signaling pathway | 2.80 |
| 38 | Osteoclast differentiation | 2.78 |
| 39 | PPAR signaling pathway | 2.76 |
| 40 | Colorectal cancer | 2.74 |
| 41 | Pathways in cancer | 2.74 |
| 42 | Endometrial cancer | 2.74 |
| 43 | Basal cell carcinoma | 2.73 |

## 4.4 Discussion

In Turkish population, the pathway that contains the most affected path is Maturity onset diabetes of the young (Figure 4.1). In this pathway paths starting with ONECUT1 (represented as HNF6 in the figure) are highly affected. Highly affected ONECUT1 – NEUROG3 – PAX6 path is related to pancreas development. In [96], pancreas was found to be involved in 5 of 170 BD cases in the autopsy series, but in [97] it is stated that pancreatic involvement in BD is exceptionally rare. Another highly affected path is ONECUT1 – PDX1 – NR5A2 path. In [98] anti-inflammatory role of NR5A2 (LRH-1) has been stated. In [99] it is shown that LRH-1 is a key player in the control of the hepatic acute-phase response. In [100] it is shown that LRH-1 has an important role in controlling of intestinal inflammation and the pathogenesis of inflammatory bowel disease. In KEGG the outcome of the affected paths are not clear yet.

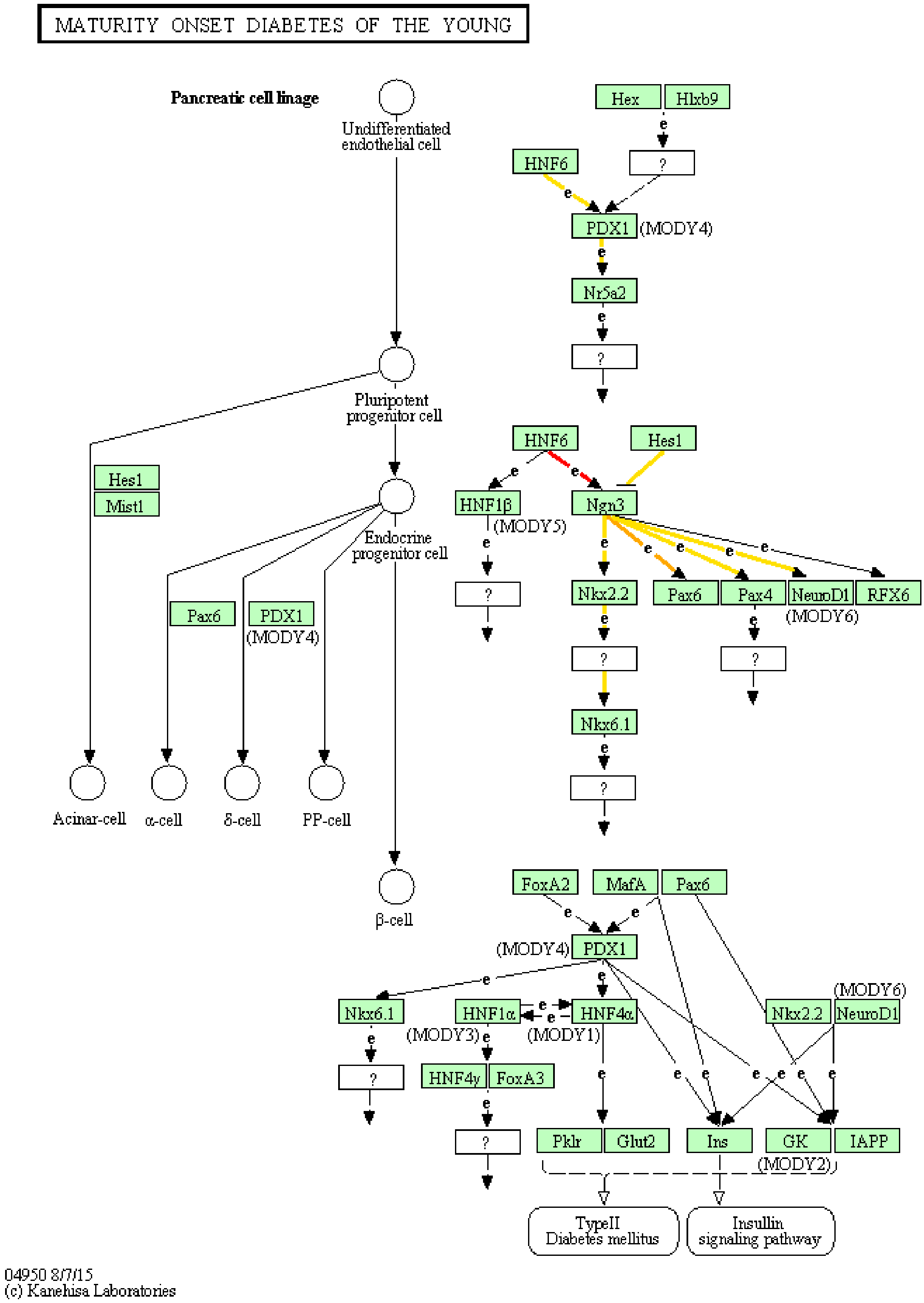

Figure 4.1 The affected paths in Maturity onset diabetes of the young pathway in Turkish population

In Notch signaling pathway, paths starting with either DLL4 (Delta) or JAG1 (Serrate) affecting DTX1 through NOTCH4 have the highest impact scores (Figure 4.2). Notch signaling pathway has been found to be highly active in BD patients and manipulation of this pathway has been offered as a therapeutic approach [101]. In [102] the role of DLL4 – Notch signaling in regulating Treg development in Experimental Autoimmune Encephalomyelitis has been shown. In [103] $CD4^+CD25^+FOXP3^+$ Treg and $CD4^+FOXP3^+$ Treg were found negatively correlated with Behçet's disease activity. In [104] $CD4^+CD25^+$ Treg cells found to be increased in the peripheral circulation of active BD patients. In [105] role of Treg cells in ocular attack in BD patients has been pointed. In [106] it is stated that DTX1 has a role in controlling Treg stability.

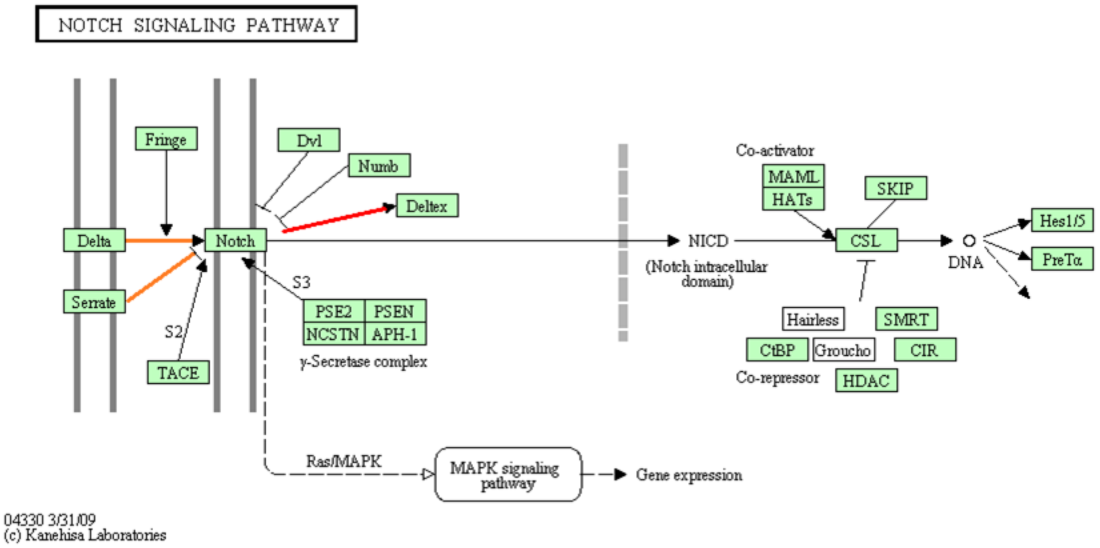

Figure 4.2 Highly affected paths in Notch signaling pathway in Turkish population

In mTOR signaling pathway RRAGC - MLST8 - ULK3 path which leads to Regulation of autophagy pathway is highly affected. The association of autophagy-related genes with autoimmune diseases is discussed in [107]. In [108] Behçet's disease systolic and diastolic blood pressure are found to be significantly higher in patients with BD. ULK3 has been associated with diastolic blood pressure in [109] and [110] in people of European descent and Japanese population respectively.

Complement and coagulation cascades pathway is affected in Turkish population. The defects in coagulation cascades in Behçet's disease has been pointed by Kiraz et al. [111]. This pathway is also found to be important in [93].

Jak-STAT signaling pathway is found to be highly affected in both populations by our method. In Figure 4.3 affected paths are shown. A red line indicates that the upstream gene is part of all the overlapping affected paths, and a yellow line indicates that the upstream gene is part of only one of the affected paths. This pathway’s activation in BD has been demonstrated by Tulunay et al. [112].

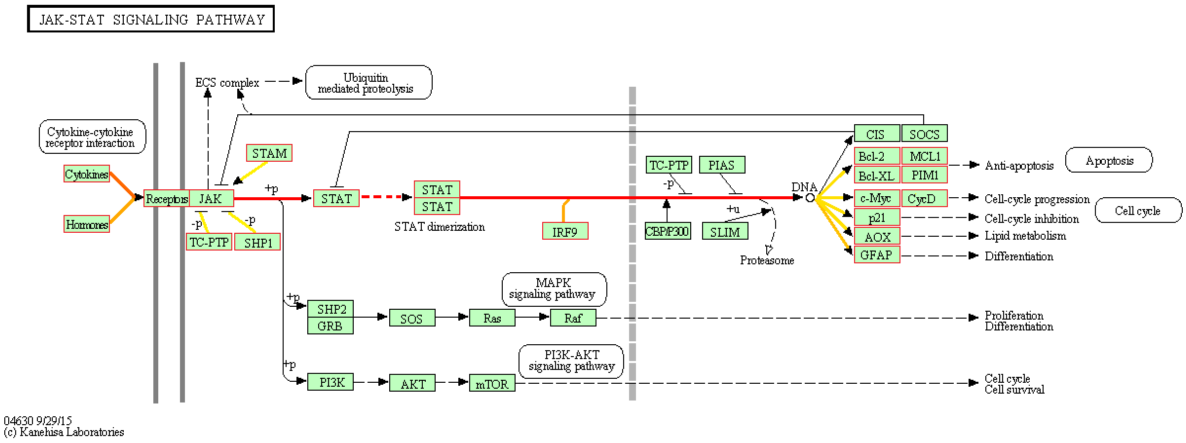

Figure 4.3 The most affected path and the thirty paths highly overlapping with it in Jak-STAT pathway in Turkish population

Hedgehog signaling pathway is affected in both populations but especially in Japanese population according to our results (Figure 4.4). In this pathway, CSNK1G3 – GLI3 – BMP2 path is highly affected. This path leads to TGF-beta signaling pathway that affects osteoblast differentiation, neurogenesis and ventral mesoderm specification through Smad1 which also has a significant SNP.

In Wnt signaling pathway CTBP2 - TCF7L1 - PPARD path is the most affected path, this path leads to cell cycle. This path is one of the discovered paths in [93].

In Japanese population Circadian rhythm pathway contains the second highest scoring path. In this pathway the most affected paths are PRKAA2 (AMPK) – CRY2 (Cry) – NPAS2 (Clock) – BHLHE40 (Dec) (Figure 4.5) and PRKAA2 (AMPK) – CRY2 (Cry) – NPAS2 (Clock) – RORA (Ror) paths. These paths control clock output / rhythmic biological processes. In [113] Miyazaki et al. showed the role of BHLHE40 (referred as Dec1 in their study) in Treg cells in mice. They reported that that BHLHE40 is required for the long-term maintenance of Treg cells after adoptive transfer to suppress effector T cell-mediated inflammation. In [114], Yu et al. showed the role of circadian clock on Th17 cell differentiation.

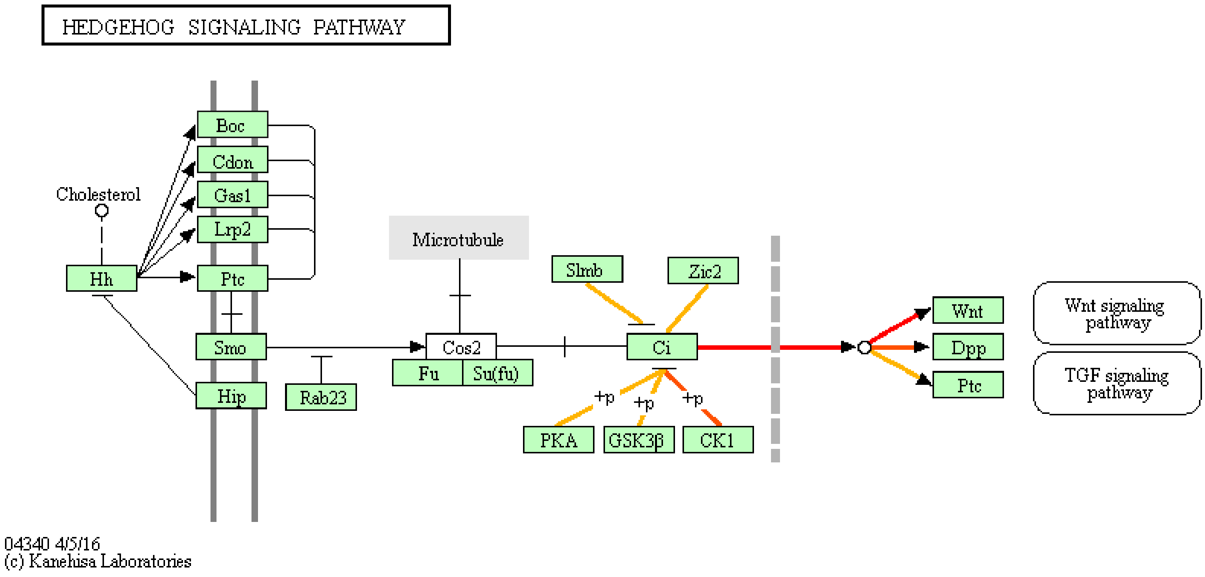

Figure 4.4 The affected paths in Hedgehog signaling pathway in Japanese population

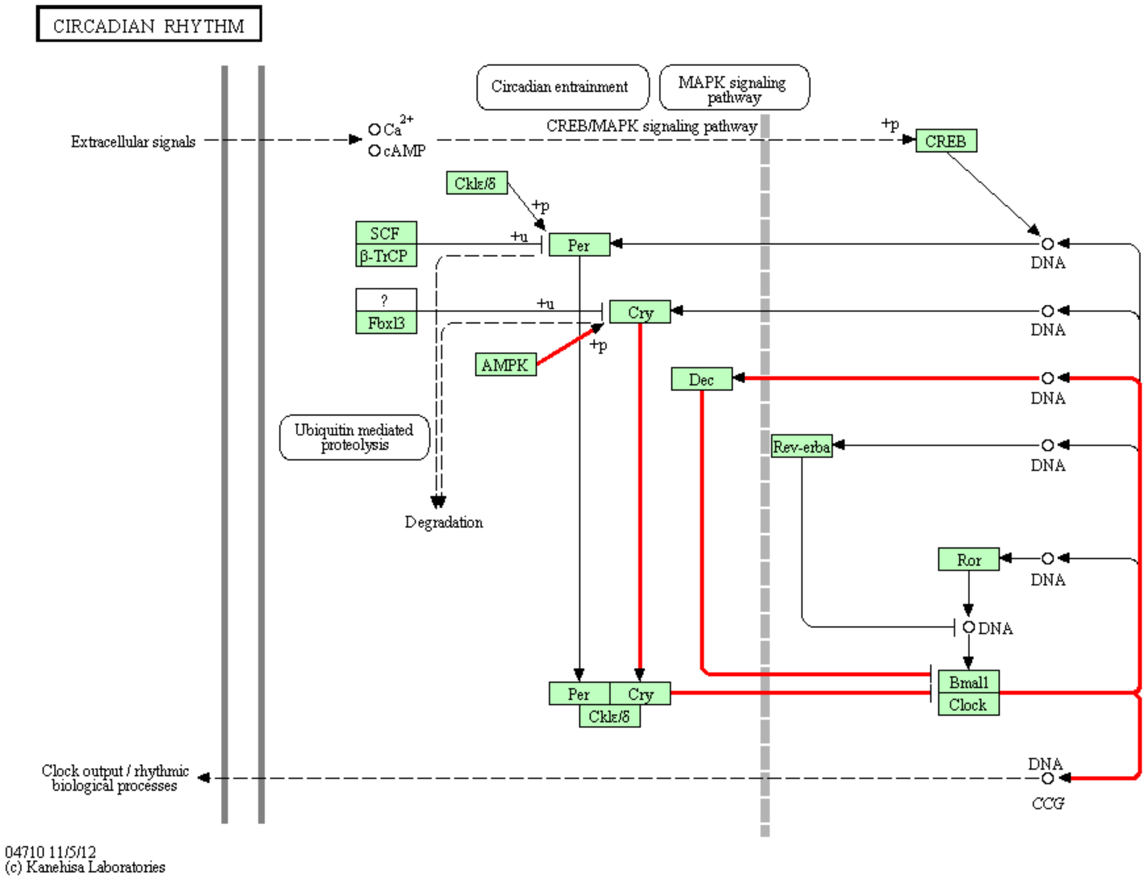

Figure 4.5 The most affected path in Circadian rhythm pathway in Japanese population

In AMPK signaling pathway MAP3K7 (Tak1) – PRKAA2 (AMPK) – PFKFB3 (PFK-2) – PFKP (PFK-1) path is the most affected path. Increased insulin resistance in Behçet’s disease patients has been reported in [115] and [116]. In [117] a SNP in PRKAA2 gene (rs2051040, which does not appear in our dataset) is reported to be associated with insulin resistance.

In Signaling pathways regulating pluripotency of stem cells pathway, NODAL – ACVR1C – SMAD2 path and BMP4 – BMPR1B – MAPK13 path are highly affected which lead to proliferation and inhibition of self-renewal in core transcriptional network respectively.

In [118] it was reported that TGFβ/Smad signalling path in T cells is overactive in patients with Behçet's disease. TGF-beta signaling pathway is one of the affected pathways in Japanese population according to our findings. The most affected path is NOG – BMP5 (BMP) – BMPR1B – SMAD1 – SMAD4 – ID3 (Id) which leads to osteoblast differentiation, neurogenesis and ventral mesoderm specification. Id proteins play a role in angiogenesis [119] [120]. Angiogenesis in vasculitides has been discussed in [121] and possible role of angiogenesis in pathogenesis of BD has been pointed in [122]

In MAPK signaling pathway, DUSP6 (MKP) – MAPK14 (p38) – MEF2C or CDC25B, PTPRR (PTP) - MAPK14 (p38) – MEF2C or CDC25B, and PPM1A (PP2CA) – MAP2K6 (MMK6) – MAPK14 (p38) – MEF2C or CDC25B paths are the affected paths.

In Osteoclast differentiation pathway, FHL2 – TRAF6 – MAP3K7 (TAK) – MAP2K6 (MKK6) – MAPK13 (p38) – FOSL2 (AP1) – CALCR (CTR) path is significantly affected. In [123] it is found that Fosl2 limits plasticity of T helper cell differentiation and plays a key role in a mouse model of autoimmune disease. In [124] it stated that T cell plasticity is an important factor in immunological diseases. In [125] contribution of plasticity of IL-17$^+$FOXP3$^+$ Treg cells towards T$_H$17 cells to the pathogenesis of rheumatoid arthritis has been shown. In [126] it is suggested that CALCR may contribute to the modulation of cytoplasmic calcium(2+) levels needed to regulate T and B cell activation and perhaps other immune functions.

# CHAPTER 5

# CONCLUSION

In this work, we have proposed two active module identification methods in order to reveal the underlying mechanisms in genetic diseases. We also combined two methods from the literature and presented the results.

The first proposed method (presented in Chapter 2) is a novel genetic algorithm approach in which, branch swapping crossover, mutation in duplicate individual addition, pruning, and two stage architecture is implemented. We applied our method on simulated datasets and rheumatoid arthritis GWAS dataset, and compared our results with the results of a simple genetic algorithm implementation and the results of the simulated annealing method that is proposed by Ideker et al. in their seminal paper [2]. The proposed method outperformed simple genetic algorithm implementation in all datasets and simulated annealing method in all but one datasets. Functional enrichment results of the extracted subnetwork present KEGG pathways whose relations to rheumatoid arthritis are supported in the literature. The performed experiments showed that the presented approach can be successfully used on the datasets of other complex diseases to detect disease-related active subnetworks. The disadvantage of this method is its long run time, this can be reduced by running GA threads in parallel.

The combined method from the literature (presented in Chapter 3) consists of the active subnetwork search proposed by Rajagopalan and Agarwal [14] and the pathway impact analysis method proposed by Tarca et al. [79]. We first applied the active subnetwork search method on intracranial aneurysm GWAS dataset and human PPI to extract the highly affected subnetwork. Then we applied the pathway impact analysis method on KEGG pathways using the genes in the subnetwork. The results demonstrate that this combined approach can find known disease related pathways. It should be noted that we did not observe a satisfying contribution of the perturbation evidence used in the pathway impact analysis method.

The second proposed method (presented in Chapter 4) is a pathway impact analysis method based on network propagation. In this method, the interacting gene sets in KEGG pathways are analyzed to discover the highly affected paths in pathways. We applied our method on Behçet’s disease GWAS dataset. Our findings were consistent with the existing literature, we also discovered new paths that could shed light on the pathogenesis of the disease.

In the future, we plan to improve the scoring method in active subnetwork search to be able to extract smaller and more focused modules. With the discovery of new interactions among proteins, human PPI gets more connected and nodes with high scores become almost totally connected in this static picture of interactions. Using network models that consider protein functions and represent the dynamic nature, more successful scoring methods can be developed.

# REFERENCES

[1] Mitra, K., Carvunis, A.R., Ramesh, S.K. and Ideker, T., (2013). "Integrative approaches for finding modular structure in biological networks", Nat. Rev. Genet., 14(10):719-732.

[2] Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F., (2002). "Discovering regulatory and signalling circuits in molecular interaction networks", Bioinformatics, 18(1):233-240.

[3] Ulitsky, I. and Shamir, R., (2007). "Identification of functional modules using network topology and high-throughput data", BMC Syst Biol, 1(1):1-17.

[4] Ulitsky, I. and Shamir, R., (2009). "Identifying functional modules using expression profiles and confidence-scored protein interactions", Bioinformatics, 25:1158-1164.

[5] Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. and Ideker, T., (2007) "Network-based classification of breast cancer metastasis", Mol Syst Biol, 3(1):1-10.

[6] Ulitsky, I., Karp, R.M. and Shamir, R., (2008). "Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles", Proceedings of Research in Computational Molecular Biology, 30 March-2 April 2008, Singapore.

[7] Chowdhury, S.A. and Koyuturk, M., (2010). "Identification of coordinately dysregulated subnetworks in complex phenotypes", Pac Symp Biocomput, 4-8 January 2010, Hawaii, 133-144.

[8] Jia, P., Zheng, S., Long, J., Zheng, W. and Zhao, Z., (2011). "dmGWAS: dense module searching for genome-wide association studies in proteinprotein interaction networks", Bioinformatics 27:95-102.

[9] Jia, P., Wang, L., Fanous, A.H., et al., (2012). "Network-assisted investigation of combined causal signals from genome-wide association studies in schizophrenia", PLoS Comput Biol, 8(7):1-11.

[10] Braun, R. and Buetow, K., (2011). "Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data", PLoS Genet, 7(6):1-13.

[11] Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A. and Sahinalp, S.C., (2011). "Optimally discriminative subnetwork markers predict response to chemotherapy", Bioinformatics, 27:205-213.

[12] Guo, Z., Wang, L., Li, Y., et al., (2007). "Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network", Bioinformatics, 23(16):2121-2128.

[13] Sohler, F., Hanisch, D. and Zimmer, R., (2004). “New methods for joint analysis of biological networks and expression data”, Bioinformatics, 20:1517-1521.

[14] Rajagopalan, D. and Agarwal, P., (2005). “Inferring pathways from gene lists using a literature-derived network of biological relationships”, Bioinformatics, 21:788-793.

[15] Dittrich, M.T., Klau, G.W., Rosenwald, A., Dandekar, T. and Muller, T., (2008). “Identifying functional modules in protein-protein interaction networks: an integrated exact approach”, Bioinformatics, 24:223-231.

[16] Breitling, R., Amtmann, A. and Herzyk, P., (2004). “Graph-based iterative Group Analysis enhances microarray interpretation”, BMC Bioinformatics, 5(1):1-10.

[17] Nacu, S., Critchley-Thorne, R., Lee, P. and Holmes, S., (2007). “Gene expression network analysis and applications to immunology”, Bioinformatics 23:850-858.

[18] Karni, S., Soreq, H. and Sharan, R., (2009). “A network-based method for predicting disease-causing genes”, J Comput Biol, 16:181-189.

[19] Fortney, K., Kotlyar, M. and Jurisica, I., (2010). “Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging”, Genome Biol, 11(2):1-15.

[20] Su, J., Yoon, B.J. and Dougherty, E.R., (2010). “Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network”, BMC Bioinformatics, 11(6):1-13.

[21] Scott, J., Ideker, T., Karp, R.M. and Sharan, R., (2006). “Efficient algorithms for detecting signaling pathways in protein interaction networks”, J Comput Biol, 13:133-144.

[22] Zhao, X.M., Wang, R.S., Chen, L. and Aihara, K., (2008). “Uncovering signal transduction networks from high-throughput data by integer linear programming”, Nucleic Acids Res, 36(9):1-12.

[23] Qiu, Y.Q., Zhang, S., Zhang, X.S. and Chen, L., (2009). “Identifying differentially expressed pathways via a mixed integer linear programming model”, IET Syst Biol, 3:475-486.

[24] Backes, C., Rurainski, A., Klau, G.W., et al., (2012). “An integer linear programming approach for finding deregulated subgraphs in regulatory networks”, Nucleic Acids Res, 40(6):1-13.

[25] Beisser, D., Brunkhorst, S., Dandekar, T., Klau, G.W., Dittrich, M.T. and Muller, T., (2012). “Robustness and accuracy of functional modules in integrated network analysis”, Bioinformatics, 28:1887-1894.

[26] Klammer, M., Godl, K., Tebbe, A. and Schaab, C., (2010). “Identifying differentially regulated subnetworks from phosphoproteomic data”, BMC Bioinformatics, 11:1-13.

[27] Ma, H., Schadt, E.E., Kaplan, L.M. and Zhao, H., (2011). “COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method”, Bioinformatics, 27:1290-1298.

[28] Wu, J., Gan, M. and Jiang, R., (2011). "A genetic algorithm for optimizing subnetwork markers for the study of breast cancer metastasis", Proceedings of Natural Computation (ICNC), 26-28 July 2011, Shanghai.

[29] Amgalan, B. and Lee, H., (2014). "WMAXC: a weighted maximum clique method for identifying condition-specific sub-network", PLoS ONE, 9(8):1-10.

[30] Shannon, P., Markiel, A., Ozier, O., et al., (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks", Genome Res, 13:2498-2504.

[31] Ulitsky, I., Krishnamurthy, A., Karp, R.M. and Shamir, R., (2010). "DEGAS: de novo discovery of dysregulated pathways in human diseases". PLoS ONE, 5:1-14.

[32] Akula, N., Baranova, A., Seto, D., et al., (2011). "A network-based approach to prioritize results from genome-wide association studies", PLoS ONE, 6:1-12.

[33] Lichtenstein, I., Charleston, M.A., Caetano, T.S., Gamble, J.R. and Vadas, M.A., (2013). "Active subnetwork recovery with a mechanism-dependent scoring function; with application to angiogenesis and organogenesis studies", BMC Bioinformatics, 14:1-22.

[34] West, J., Beck, S., Wang, X. and Teschendorff, A.E., (2013). "An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways", Sci Rep, 3:1-11.

[35] Petrochilos, D., Shojaie, A., Gennari, J. and Abernethy, N., (2013). "Using random walks to identify cancer-associated modules in expression data", BioData Min, 6:1-25.

[36] Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M. and Barabasi, A.L., (2007). "The human disease network", Proc Natl Acad Sci USA, 104:8685-8690.

[37] Rual, J.F., Venkatesan, K., Hao, T., et al., (2005). "Towards a proteome-scale map of the human protein-protein interaction network", Nature, 437:1173-1178.

[38] Stelzl, U., Worm, U., Lalowski, M., et al., (2005). "A human protein-protein interaction network: a resource for annotating the proteome", Cell, 122:957-968.

[39] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M., (2006). "BioGRID: a general repository for interaction datasets", Nucleic Acids Res, 34:535-539.

[40] Bakir-Gungor, B. and Sezerman, O.U., (2011). "A new methodology to associate SNPs with human diseases according to their pathway related context", PLoS ONE, 6:1-13.

[41] Bindea, G., Mlecnik, B., Hackl, H., et al., (2009). "ClueGO: a Cytoscape plugin to decipher functionally grouped gene ontology and pathway annotation networks", Bioinformatics, 25:1091-1093.

[42] Yakova, M., Lezin, A., Dantin, F., Lagathu, G., Olindo, S., Jean-Baptiste, G., Arfi, S. and Césaire, R., (2005). "Increased proviral load in HTLV-1-infected patients with rheumatoid arthritis or connective tissue disease", Retrovirology, 2:1-9.

[43] Plenge, R.M., Cotsapas, C., Davies, L., et al., (2007). “Two independent alleles at 6q23 associated with risk of rheumatoid arthritis”, Nat Genet, 39:1477-1482.

[44] Shahrara, S., Castro-Rueda, H.P., Haines, G.K. and Koch, A.E., (2007). “Differential expression of the FAK family kinases in rheumatoid arthritis and osteoarthritis synovial tissues”, Arthritis Res Ther, 9(5):1-10.

[45] Raychaudhuri, S., Remmers, E.F., Lee, A.T., et al., (2008). “Common variants at CD40 and other loci confer risk of rheumatoid arthritis”, Nat Genet, 40:1216-1223.

[46] Raychaudhuri, S., Thomson, B.P., Remmers, E.F., et al., (2009). “Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk”, Nat Genet, 41:1313-1318.

[47] Nah, S.S., Won, H.J., Ha, E., et al., (2010). “Epidermal growth factor increases prostaglandin E2 production via ERK1/2 MAPK and NF-kappaB pathway in fibroblast like synoviocytes from patients with rheumatoid arthritis”, Rheumatol Int, 30:443-449.

[48] Menon, R. and Farina, C., (2011). “Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes”, PLoS ONE, 6:1-9.

[49] Yuan, F.L., Li, X., Lu, W.G., Sun, J.M., Jiang, D.L. and Xu, R.S., (2013). “Epidermal growth factor receptor (EGFR) as a therapeutic target in rheumatoid arthritis”, Clin Rheumatol, 32:289-292.

[50] Mellemkjaer, L., Linet, M.S., Gridley, G., Frisch, M., Møller, H. and Olsen, J.H., (1996). “Rheumatoid arthritis and cancer risk”, Eur J Cancer, 32A(10):1753-1757.

[51] Tian, G., Liang, J.N., Wang, Z.Y. and Zhou, D., (2014). “Breast cancer risk in rheumatoid arthritis: an update meta-analysis”, Biomed Res Int, 2014:1-9.

[52] Remans, P.H., Gringhuis, S.I., Laar, J.M., et al., (2004). “Rap1 signaling is required for suppression of Ras-generated reactive oxygen species and protection against oxidative stress in T lymphocytes”, J Immunol, 173:920-931.

[53] Reedquist, K.A. and Tak, P.P., (2012). “Signal transduction pathways in chronic inflammatory autoimmune disease: small GTPases”, Open Rheumatol J, 6:259-272.

[54] Abreu, J.R., Krausz, S., Dontje, W., et al., (2010). “Sustained T cell Rap1 signaling is protective in the collagen-induced arthritis model of rheumatoid arthritis”, Arthritis Rheum, 62:3289-3299.

[55] Ling, S., Lai, A., Borschukova, O., Pumpens, P. and Holoshitz, J., (2006). “Activation of nitric oxide signaling by the rheumatoid arthritis shared epitope”, Arthritis Rheum, 54:3423-3432.

[56] Almeida, D.E., Ling, S. and Holoshitz, J., (2011). “New insights into the functional role of the rheumatoid arthritis shared epitope”, FEBS Lett, 585:3619-3626.

[57] Shelef, M.A., Bennin, D.A., Yasmin, N., et al., (2014). “Focal adhesion kinase is required for synovial fibroblast invasion, but not murine inflammatory arthritis”, Arthritis Res Ther, 16(5):1-10.

[58] Nakano, K., Yamaoka, K., Hanami, K., et al., (2011). “Dopamine induces IL-6-dependent IL-17 production via D1-like receptor on CD4 naive T cells and D1-like receptor antagonist SCH-23390 inhibits cartilage destruction in a human rheumatoid arthritis/SCID mouse chimera model”, J Immunol, 186:3745-3752.

[59] Capellino, S., Cosentino, M., Luini, A., et al., (2014). “Increased expression of dopamine receptors in synovial fibroblasts from patients with rheumatoid arthritis: inhibitory effects of dopamine on interleukin-8 and interleukin-6”, Arthritis Rheumatol, 66:2685-2693.

[60] Pacheco, R., Contreras, F. and Zouali, M., (2014). “The dopaminergic system in autoimmune diseases”, Front Immunol, 5(117):1-17.

[61] Xu, B., Arlehag, L., Rantapaa-Dahlquist, S.B. and Lefvert, A.K., (2004). “beta2-adrenergic receptor gene single-nucleotide polymorphisms are associated with rheumatoid arthritis in northern Sweden”, Scand J Rheumatol, 33:395-398.

[62] Malysheva, O., Pierer, M., Wagner, U., Wahle, M., Wagner, U. and Baerwald, C.G., (2008). “Association between beta2 adrenergic receptor polymorphisms and rheumatoid arthritis in conjunction with human leukocyte antigen (HLA)-DRB1 shared epitope”, Ann Rheum Dis, 67:1759-1764.

[63] Veale, D.J. and Maple, C., (1996). “Cell adhesion molecules in rheumatoid arthritis”, Drugs Aging, 9:87-92.

[64] Rihl, M., Kruithof, E., Barthel, C., et al., (2005). “Involvement of neurotrophins and their receptors in spondyloarthritis synovitis: relation to inflammation and response to treatment”, Ann Rheum Dis, 64:1542-1549.

[65] Barthel, C., Yeremenko, N., Jacobs, R., et al., (2009). “Nerve growth factor and receptor expression in rheumatoid arthritis and spondyloarthritis”, Arthritis Res Ther, 11(3):1-9.

[66] Bonnet, C.S., Williams, A.S., Gilbert, S.J., Harvey, A.K., Evans, B.A. and Mason, D.J., (2015). “AMPA/kainate glutamate receptors contribute to inflammation, degeneration and pain related behaviour in inflammatory stages of arthritis”, Ann Rheum Dis, 74:242-251.

[67] Cope, A.P., (2008). “T cells in rheumatoid arthritis”, Arthritis Res Ther, 10(1):1-10.

[68] Sakaguchi, S., Benham, H., Cope, A.P. and Thomas, R., (2012). “T-cell receptor signaling and the pathogenesis of autoimmune arthritis: insights from mouse and man”, Immunol Cell Biol, 90:277-287.

[69] Kormendy, D., Hoff, H., Hoff, P., Broker, B.M., Burmester, G.R. and Brunner-Weinzierl, M.C., (2013). “Impact of the CTLA-4/CD28 axis on the processes of joint inflammation in rheumatoid arthritis”, Arthritis Rheum, 65:81-87.

[70] Clark, A.R. and Dean, J.L., (2012). “The p38 MAPK Pathway in Rheumatoid Arthritis: A Sideways Look”, Open Rheumatol J, 6:209-219.

[71] Thalhamer, T., McGrath, M.A. and Harnett, M.M., (2008). “MAPKs and their relevance to arthritis and inflammation”, Rheumatology (Oxford), 47:409-414.

[72] Rommel, C., Camps, M. and Ji, H., (2007). “PI3K delta and PI3K gamma: partners in crime in inflammation in rheumatoid arthritis and beyond?”, Nat Rev Immunol, 7:191-201.

[73] Mitra, A., Raychaudhuri, S.K. and Raychaudhuri, S.P., (2012). “IL-22 induced cell proliferation is regulated by PI3K/Akt/mTOR signaling cascade”, Cytokine, 60:38-42.

[74] Bartok, B., Boyle, D.L., Liu, Y., et al., (2012). “PI3 kinase is a key regulator of synoviocyte function in rheumatoid arthritis”, Am J Pathol, 180:1906-1916.

[75] Nagafuchi, H., Suzuki, N., Kaneko, A., Asai, T. and Sakane, T., (1999). “Prolactin locally produced by synovium infiltrating T lymphocytes induces excessive synovial cell functions in patients with rheumatoid arthritis”, J Rheumatol, 26:1890-1900.

[76] Rovensky, J., Kvetnansky, R., Radikova, Z., et al., (2005). “Hormone concentrations in synovial fluid of patients with rheumatoid arthritis”, Clin Exp Rheumatol, 23:292-296.

[77] Tang, C., Li, Y., Lin, X., et al., (2014). “Prolactin increases tumor necrosis factor alpha expression in peripheral CD14 monocytes of patients with rheumatoid arthritis”, Cell Immunol, 290:164-168.

[78] Adan, N., Guzman-Morales, J., Ledesma-Colunga, M.G., et al., (2013). “Prolactin promotes cartilage survival and attenuates inflammation in inflammatory arthritis”, J Clin Invest, 123:3902-3913.

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

[80] Li, X., Shen, L., Shang, X. and Liu, W., (2015). “Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway”, PLoS ONE, 10(7):1-19.

[81] Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J.M., (2010). “Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM”, Bioinformatics, 26(12):237-245.

[82] Dutta, B., Wallqvist, A. and Reifman, J., (2012). “PathNet: a tool for pathway analysis using topological information”, Source Code Biol Med, 7(1):1-12.

[83] Haynes, W.A., Higdon, R., Stanberry, L., Collins, D. and Kolker, E., (2013). “Differential expression analysis for pathways”, PLoS Comput. Biol., 9(3):1-17.

[84] Sebastian-Leon, P., Vidal, E., Minguez, P., et al., (2014). “Understanding disease mechanisms with models of signaling pathway activities”, BMC Syst Biol, 8:1-19.

[85] Yasuno, K., Bilguvar, K., Bijlenga, P., et al., (2010). “Genome-wide association study of intracranial aneurysm identifies three new risk loci”, Nat. Genet., 42(5):420-425.

[86] Akiyama, K., Narita, A., Nakaoka, H., et al., (2010). “Genome-wide association study to identify genetic variants present in Japanese patients harboring intracranial aneurysms”, J. Hum. Genet., 55(10):656-661.

[87] Bakir-Gungor, B. and Sezerman, O.U., (2013). “The identification of pathway markers in intracranial aneurysm using genome-wide association data from two different populations”, PLoS ONE, 8(3):1-12.

[88] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M., (2006). “BioGRID: a general repository for interaction datasets”, Nucleic Acids Res., 34:535-539.

[89] Kanehisa, M. and Goto, S., (2000). “KEGG: Kyoto encyclopedia of genes and genomes”, Nucleic Acids Res., 28(1):27-30.

[90] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M., (2016). “KEGG as a reference resource for gene and protein annotation”, Nucleic Acids Res., 44(1):457-462.

[91] Remmers, E.F., Cosan, F., Kirino, Y., et al., (2010). “Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behçet's disease”, Nat. Genet., 42(8):698-702.

[92] Mizuki, N., Meguro, A., Ota, M., et al., (2010). “Genome-wide association studies identify IL23R-IL12RB2 and IL10 as Behçet's disease susceptibility loci”, Nat. Genet., 42(8):703-706.

[93] Bakir-Gungor, B., Remmers, E.F., Meguro, A., Mizuki, N., Kastner, D.L., Gul, A. and Sezerman, O.U., (2015). “Identification of possible pathogenic pathways in Behçet's disease using genome-wide association study data from two different populations”, Eur. J. Hum. Genet., 23(5):678-687.

[94] Donato, M., Xu, Z., Tomoiaga, A., et al., (2013). “Analysis and correction of crosstalk effects in pathway analysis”, Genome Res., 23(11):1885-1893.

[95] KGML Document, KEGG Markup Language, http://www.kegg.jp/kegg/xml/docs/, 1 March 2016.

[96] Lakhanpal, S., Tani, K., Lie, J.T., Katoh, K., Ishigatsubo, Y. and Ohokubo, T., (1985). “Pathologic features of Behçet's syndrome: a review of Japanese autopsy registry data”, Hum. Pathol., 16(8):790-795.

[97] Skef, W., Hamilton, M.J. and Arayssi, T., (2015). “Gastrointestinal Behçet's disease: a review”, World J. Gastroenterol., 21(13):3801-3812.

[98] Venteclef, N., Jakobsson, T., Ehrlund, A., et al., (2010). “GPS2-dependent corepressor/SUMO pathways govern anti-inflammatory actions of LRH-1 and LXRbeta in the hepatic acute phase response”, Genes Dev., 24(4):381-395.

[99] Venteclef, N., Smith, J.C., Goodwin, B. and Delerive, P., (2006). “Liver receptor homolog 1 is a negative regulator of the hepatic acute-phase response”, Mol. Cell. Biol., 26(18):6799-6807.

[100] Coste, A., Dubuquoy, L., Barnouin, R., et al., (2007). “LRH-1-mediated glucocorticoid synthesis in enterocytes protects against inflammatory bowel disease”, Proc. Natl. Acad. Sci. U.S.A., 104(32):13098-13103.

[101] Qi, J., Yang, Y., Hou, S., Qiao, Y., Wang, Q., Yu, H., Zhang, Q., Cai, T., Kijlstra, A. and Yang, P., (2014). “Increased Notch pathway activation in Behçet's disease”, Rheumatology (Oxford), 53(5):810-820.

[102] Bassil, R., Zhu, B., Lahoud, Y., Riella, L.V., Yagita, H., Elyaman, W. and Khoury, S.J., (2011). “Notch ligand delta-like 4 blockade alleviates experimental autoimmune encephalomyelitis by promoting regulatory T cell development”, J. Immunol., 187(5):2322-2328.

[103] Gündüz, E., Teke, H.U., Bilge, N.S., Cansu, D.U., Bal, C., Korkmaz, C. and Gülbaş, Z., (2013). "Regulatory T cells in Behçet's disease: is there a correlation with disease activity? Does regulatory T cell type matter?", Rheumatol. Int., 33(12):3049-3054.

[104] Hamzaoui, K., Hamzaoui, A. and Houman, H., (2006). "CD4+CD25+ regulatory T cells in patients with Behçet's disease", Clin. Exp. Rheumatol., 24(5):71-78.

[105] Nanke, Y., Kotake, S., Goto, M., Ujihara, H., Matsubara, M. and Kamatani, N., (2008). "Decreased percentages of regulatory T cells in peripheral blood of patients with Behcet's disease before ocular attack: a possible predictive marker of ocular attack", Mod Rheumatol, 18(4):354-358.

[106] Hsiao, H.W., Hsu, T.S., Liu, W.H., Hsieh, W.C., Chou, T.F., Wu, Y.J., Jiang, S.T., Lai and M.Z., (2015). "Deltex1 antagonizes HIF-1α and sustains the stability of regulatory T cells in vivo", Nat Commun, 6:1-12.

[107] Zheng, M., Yu, H., Zhang, L., Li, H., Liu, Y., Kijlstra, A. and Yang, P., (2015). "Association of ATG5 Gene Polymorphisms with Behçet's Disease and ATG10 Gene Polymorphisms with VKH Syndrome in a Chinese Han Population", Invest. Ophthalmol. Vis. Sci., 56(13):8280-8287.

[108] Aksu, T., Guler, E., Arat, N., Zorlu, A., Yılmaz, M.B., Guray, U., Tufekcioglu, O. and Kısacık, H., (2015). "Cardiovascular Involvement in Behcet's Disease", Archives of Rheumatology, 30(2):109-115.

[109] Levy, D., Ehret, G.B., Rice, K., et al., (2009). "Genome-wide association study of blood pressure and hypertension", Nat. Genet., 41(6):677-687.

[110] Takeuchi, F., Isono, M., Katsuya, T., et al., (2010). "Blood pressure and hypertension are associated with 7 loci in the Japanese population", Circulation, 121(21):2302-2309.

[111] Kiraz, S., Ertenli, I., Ozturk, M.A., Haznedaroglu, I.C., Celik, I. and Calguneri, M., (2002). "Pathological haemostasis and "prothrombotic state" in Behçet's disease", Thromb. Res., 105(2):125-133.

[112] Tulunay, A., Dozmorov, M.G., Ture-Ozdemir, F., et al., (2015). "Activation of the JAK/STAT pathway in Behcet's disease", Genes Immun., 16(2):170-175.

[113] Miyazaki, K., Miyazaki, M., Guo, Y., Yamasaki, N., Kanno, M., Honda, Z., Oda, H., Kawamoto, H. and Honda, H., (2010). "The role of the basic helix-loop-helix transcription factor Dec1 in the regulatory T cells", J. Immunol., 185(12):7330-7339.

[114] Yu, X., Rollins, D., Ruhn, K.A., Stubblefield, J.J., Green, C.B., Kashiwada, M., Rothman, P.B., Takahashi, J.S. and Hooper, L.V., (2013). "TH17 cell differentiation is regulated by the circadian clock", Science, 342(6159):727-730.

[115] Kim, S.K., Choe, J.Y., Park, S.H., Lee, S.W., Lee, G.H. and Chung, W.T., (2010). "Increased insulin resistance and serum resistin in Korean patients with Behçet's disease", Arch. Med. Res., 41(4):269-274.

[116] Erdem, H., Dinc, A., Pay, S., Simsek, I. and Turan, M., (2006). "Peripheral insulin resistance in patients with Behçet's disease", J Eur Acad Dermatol Venereol, 20(4):391-395.

[117] Horikoshi, M., Hara, K., Ohashi, J., Miyake, K., Tokunaga, K., Ito, C., Kasuga, M., Nagai, R. and Kadowaki, T., (2006). “A polymorphism in the AMPKalpha2 subunit gene is associated with insulin resistance and type 2 diabetes in the Japanese population”, Diabetes, 55(4):919-923.

[118] Shimizu, J., Izumi, T., Arimitsu, N., Fujiwara, N., Ueda, Y., Wakisaka, S., Yoshikawa, H., Kaneko, F., Suzuki, T., Takai, K. and Suzuki, N., (2012). “Skewed TGFβ/Smad signalling pathway in T cells in patients with Behçet's disease” Clin. Exp. Rheumatol., 30(3):35-39.

[119] Lyden, D., Young, A.Z., Zagzag, D., Yan, W., Gerald, W., O'Reilly, R., Bader, B.L., Hynes, R.O., Zhuang, Y., Manova, K. and Benezra, R., (1999). “Id1 and Id3 are required for neurogenesis, angiogenesis and vascularization of tumour xenografts”, Nature, 401(6754):670-677.

[120] Benezra, R., Rafii, S. and Lyden, D., (2001). “The Id proteins and angiogenesis”, Oncogene, 20(58):8334-8341.

[121] Maruotti, N., Cantatore, F.P., Nico, B., Vacca, A. and Ribatti, D., (2008). “Angiogenesis in vasculitides”, Clin. Exp. Rheumatol., 26(3):476-483.

[122] Keskin, D., Keskin, G., Inal, A. and Ozisik, L., (2014). “Serum angiostatin levels in patients with Behçet's disease: does angiogenesis play a role in the pathogenesis of Behçet's disease?”, Acta Clin Belg, 69(4):246-250.

[123] Ciofani, M., Madar, A., Galan, C., et al., (2012). “A validated regulatory network for Th17 cell specification”, Cell, 151(2):289-303.

[124] DuPage, M. and Bluestone, J.A., (2016). "Harnessing the plasticity of CD4(+) T cells to treat immune-mediated disease", Nat. Rev. Immunol., 16(3):149-163.

[125] Komatsu, N., Okamoto, K., Sawa, S., Nakashima, T., Oh-hora, M., Kodama, T., Tanaka, S., Bluestone, J.A. and Takayanagi, H., (2014). "Pathogenic conversion of Foxp3+ T cells into TH17 cells in autoimmune arthritis", Nat. Med., 20(1):62-68.

[126] Cafforio, P., De Matteo, M., Brunetti, A.E., Dammacco, F. and Silvestris, F., (2009). "Functional expression of the calcitonin receptor by human T and B cells", Hum. Immunol., 70(9):678-685.

# CURRICULUM VITAE

## PERSONAL INFORMATION

| | |
|---|---|
| **Name Surname** | : Ozan ÖZIŞIK |
| **Date of birth and place** | : 17.08.1985 / İstanbul |
| **Foreign Languages** | : English |
| **E-mail** | : ozanytu@gmail.com |

## EDUCATION

| Degree | Department | University | Date of Graduation |
|---|---|---|---|
| PhD | Computer Eng. | YTU | 2016 |
| Master | Computer Eng. | YTU | 2010 |
| Undergraduate | Computer Eng. | YTU | 2007 |
| High School | Science and Maths | Vatan AHL | 2003 |

## WORK EXPERIENCE

| Year | Corporation/Institute | Role |
|---|---|---|
| 2007-… | Yildiz Technical University | Research Assistant |

## PUBLICATIONS

### Papers

1. Ozisik, O., Bakir-Gungor, B., Diri, B. and Sezerman, O.U., (2016). “ActiveSubnetworkGA: A Two Stage Genetic Algorithm Approach to Active Subnetwork Search”, Current Bioinformatics (Accepted, DOI: 10.2174/1574893611666160527100444).
2. Ozisik, O. and Yavuz, S., (2016). “Simultaneous Localization and Mapping with Limited Sensing Using Extended Kalman Filter and Hough Transform”, Tehnički vjesnik/Technical Gazette (Accepted).

### Conference Papers

1. Arslan, A., Türk, A., Akhun, N., Kara, S., Özışık, O. and Saygılı, Ö., (2016). “Suprapatellar Fat-pad Impingement: MRG Bulguları”, Türk Manyetik Rezonans Derneği 21. Bilimsel Toplantısı (MR 2016), 26-28 May 2016, Ankara.
2. Ozisik, O., Bakir-Gungor, B., Diri, B. and Sezerman, O.U., (2013). “A Genetic Algorithm Approach to Active Subnetwork Search applied to GWAS Data”, The International Symposium on Health Informatics and Bioinformatics (HIBIT), 25-27 September 2013, Ankara.
3. Ozisik, O. and Yavuz, S., (2009). “Fuzzy-Neural Robot Controller for Unknown Environment Exploration”, 1st International Fuzzy Systems Symposium (FUZZYSS'09), 1-2 October 2009, Ankara.
4. Ozisik, O. and Yavuz, S., (2008). “An Occupancy Grid Based SLAM Method”, IEEE Computational Intelligence for Measurement Systems and Applications (CIMSA 2008), 14-16 July 2008, İstanbul.
5. Ozisik, O. and Yavuz, S., (2007). “Optical Music Recognition”, International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2007), 20-23 June 2007, İstanbul.