IDENTIFICATION OF ACTIVE DISEASE-ASSOCIATED SUBNETWORKS IN HUMAN PROTEIN-PROTEIN INTERACTION NETWORKS USING THE MCL ALGORITHM

by

KIVILCIM ÖZTÜRK

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

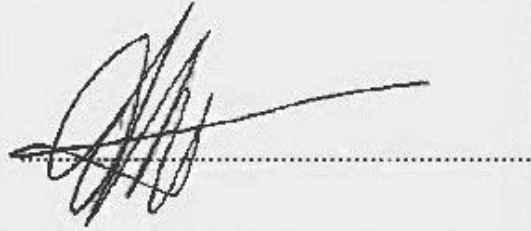Sabancı University

Spring 2015

Identification of Active Disease-Associated Subnetworks in Human Protein-Protein Interaction Networks using the MCL Algorithm
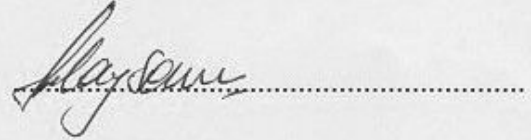
APPROVED BY:

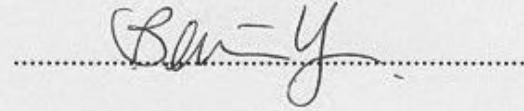Assoc. Prof. Dr. Yucel Saygin
(Thesis Supervisor)

Prof. Dr. Ugur Sezerman
(Thesis Co-Supervisor)

Prof. Dr. Erkay Savas

Assoc. Prof. Dr. Berrin Yanikoglu

Assoc. Prof. Dr. Devrim Gozuacik

DATE OF APPROVAL: 05/08/2015

ABSTRACT

IDENTIFICATION OF ACTIVE DISEASE-ASSOCIATED SUBNETWORKS IN HUMAN PROTEIN-PROTEIN INTERACTION NETWORKS USING THE MCL ALGORITHM

KIVILCIM ÖZTÜRK

Computer Science and Engineering
M.Sc. Thesis, 2015

Thesis Supervisors: Yücel Saygın and Uğur Osman Sezerman

Keywords: Active Subnetworks, Pathways, Markov Cluster Algorithm, GWAS, Rheumatoid Arthritis

An active subnetwork is a group of highly interacting genes that are associated with a particular disease in a biological interaction network. Finding these subnetworks facilitates the understanding of the molecular mechanisms of diseases and contributes to the process of devising treatment strategies, making the identification of active subnetworks an important problem. In this thesis, the use of a clustering algorithm is proposed for the detection of active subnetworks and a methodology that is based on the Markov Cluster (MCL) algorithm is implemented. The methodology uses graph representation to represent the human protein-protein interaction network, a novel scoring scheme to appoint weights to the interactions among the network, the Markov Cluster algorithm for the active subnetwork search, a scoring formula to assign scores to each found subnetwork and an elimination of subnetworks depending on those scores, followed by a functional enrichment step to discover the functionally important KEGG pathways related with found subnetworks. This methodology is applied on WTCCC Rheumatoid Arthritis (RA) dataset and identified: KEGG pathways previously found to be RA-related (e.g., NF-kappaB, Jak-STAT, Toll-like receptor, MAPK signaling pathways), and additional pathways (e.g., Serotonergic synapse) as associated with RA. The comparative study shows that the presented method outperforms state-of-the-art techniques, and functional enrichment results demonstrate that the method can successfully detect significant subnetworks that are related with RA which is a complex multifactorial disease. Therefore, it is proposed that the method can be used on the datasets of other complex diseases to identify active disease-associated subnetworks.

ÖZET

MCL ALGORİTMASI KULLANILARAK
İNSAN PROTEİN-PROTEİN İNTERAKSİYON AĞLARINDA
HASTALIK-İLİŞKİLİ AKTİF ALT-AĞLARIN SAPTANMASI

KIVILCIM ÖZTÜRK

Bilgisayar Bilimi ve Mühendisliği
Master Tezi, 2015

Tez Danışmanları: Yücel Saygın ve Uğur Osman Sezerman

Anahtar Kelimeler: Aktif Alt-Ağlar, Yolaklar, Markov Kümeleme Algoritması, GWAS, Romatoid Artrit

Biyolojik bir interaksiyon ağında, belirli bir hastalık ile alakalı ve birbiriyle yoğun etkileşim içerisinde olan genlerin bulunduğu gruplara aktif alt-ağ denilir. Bu alt-ağları bulmak hastalıkların moleküler mekanizmalarını anlamaya yardımcı olmakta ve tedavi yöntemleri tasarlamaya katkıda bulunmaktadır; bu nedenle aktif alt-ağların saptanması önemli bir problemdir. Bu tezde, aktif alt-ağların tespiti için bir kümeleme algoritmasının kullanımı önerilmektedir ve Markov Kümeleme (MCL) algoritmasına dayalı bir yöntem geliştirilmiştir. Bu yöntem, insan protein-protein etkileşim ağını temsil etmek için grafik temsili, ağdaki interaksiyonlara bir değer atamak için yeni bir skorlama tekniği, aktif alt-ağ araması için Markov Kümeleme algoritması, bulunan alt-ağlara skor atamak için yeni bir formül ve alt-ağların bazılarını elemek için de bu skorları kullanmaktadır. Bu aşama, saptanan alt-ağlarla ilişkili fonksiyonel olarak önemli olan KEGG yolaklar tespit edilerek takip edilmektedir. Tanımlanan teknik WTCCC Romatoid Artrit (RA) datası üzerinde test edilmiştir ve sıradaki yolakları RA-ilişkili yolaklar olarak saptamıştır: daha önce RA ile alakalı olduğu keşfedilmiş yolaklar (NF-kappaB, Jak-STAT, Toll-like receptor, MAPK signaling gibi) ve yeni yolaklar (Serotonergic synapse). Karşılaştırmalı bir çalışma, sunulan metodun son model tekniklerden daha iyi bir performansa sahip olduğunu göstermekte ve sonuçlar metodun başarılı bir şekilde kompleks ve multifaktoriyel bir hastalık olan RA ile alakalı alt-ağları saptayabileceğini kanıtlamaktadır. Bu nedenle, metodun başka kompleks hastalıkların dataları üzerine uygulanması durumunda o hastalıklarla ilişkili alt-ağları da tespit edebileceği önerilmektedir.

Sevgili anneme ve babama,
Sonsuz sevgileri ve destekleri için…

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisors Ugur Sezerman and Yucel Saygin for their guidance and support in completion of this project. This thesis would not have been possible without their academic and personal support.

I wish to also thank the thesis committee for their participation and recommendations which allowed this thesis to improve greatly.

Lastly, I wish to express my heartfelt thanks to my parents for their endless love, care and patience. I am grateful to them for the support they have given me during all of my educational life, especially throughout writing of this thesis. I know without a doubt that I would not have been able to complete it without their love and faith in me.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

INTRODUCTION

An active subnetwork is a group of interconnected genes in a protein-protein interaction (PPI) network and is composed of genes that are associated with a particular disease or a condition. Over the years, the problem of active subnetwork search, aiming the detection of these active subnetworks, has become increasingly important to our global understanding of the molecular mechanisms of diseases. It has been conceived that all proteins encoded by genes are responsible for the execution of specific functions which they perform by interacting with each other and destruction of these interactions may be playing a major role in the development of diseases. Therefore it is very important to identify these disease-related active subnetworks which in turn might assist in the understanding of molecular architecture of diseases and thus, hopefully, their treatment.

Due to the conceived importance of the active subnetwork detection problem, many computational methods have been proposed as a solution in the last decade. Most of these methods integrate observation data (e.g., gene expression) with the network topology to identify the potential subnetworks [1]. Frequently in these methods, the PPI network is represented as a graph where nodes denote genes and edges denote the interactions between the proteins encoded by those genes. Furthermore, the nodes are scored to reflect the significance of the genes they represent relative to the disease based on a variety of approaches including genetic variants, messenger RNA (mRNA) expression, microRNA expression, DNA methylation, protein abundance [2], with the significance being determined in a condition specific experiment such as a microarray or a genome-wide association study.

In this thesis, a clustering algorithm method is proposed for the problem of active subnetwork search in the human protein-protein interaction network. This method utilizes graph representation to represent the genes and the interactions between them, a novel edge weight calculation scheme to assign weights to those interactions, the Markov Cluster algorithm for the discovery of active subnetworks, a scoring formula to appoint scores to each found subnetwork and an elimination of subnetworks depending on those scores, followed by a functional enrichment step to discover the functionally important KEGG pathways in the found subnetworks. The method is applied on the Wellcome Trust Case Control Consortium (WTCCC) Rheumatoid Arthritis (RA) dataset.

Chapter 2

RELATED WORK and CONTRIBUTION

In literature, disease-related active subnetworks have been tried to be identified for different purposes from detecting disease-related regulatory pathways [3] and finding markers for cancer [4], to estimating response to its treatments [5]. In 2002, Ideker et al. [3] introduced a framework for active subnetwork detection from a full network of molecular interactions. This framework describes a problem which looks for the connected regions of the network that displays noticeable variations in expression on a specific set of conditions. Since then, this problem has been studied with many approaches which eventually settled to involve two parts [1]:

1. The scoring scheme: The interactions between the genes and the connected region of genes are scored so that the scores indicate the probability of the region being active.
2. The search model: The search among the connected regions is designed in a way to achieve the identification of the highest scoring regions.

The model proposed by Ideker et al. [3] acquires statistical scores of each gene based on their mRNA expression data obtained from a microarray study and assigns an overall statistical score to every subnetwork. Then the actual search for the maximal-scoring subnetworks is performed using simulated annealing.

In their study, Ideker et al. [3] demonstrated that the second part of the problem, which coincides with the active subnetwork search, is an NP-hard problem. Since then, a lot of attempts have been made to use heuristics to solve the problem, like greedy search, color coding, algorithms based on mathematical programming and again simulated

annealing. Guo et al. [6] also used simulated annealing in their study with the methodological difference being their use of edge-based scoring. The advantage of such edge-based methods which result with a list of edges (interactions) instead of a list of nodes (genes) is that they also demonstrate the active interactions in the condition rather than only displaying the active groups [1]. Ma et al. utilizes both node and edge-based approaches in their scoring scheme [7] with the F-statistic measuring gene expression, and an expected conditional F statistic (ECF) measuring correlation between genes.

To find the significant areas of the network, Sohler et al. developed a greedy approach which selects a set of seed genes according to a threshold and then performs a greedy expansion by incorporating the most significant adjacent genes based on their p-values at every iteration [8]. Chuang et al. [4] also uses a similar approach to detect the highest-scoring subnetworks in the PPI network by using gene expression profiles of tissue samples in order to find markers for breast cancer. In this search, seed proteins are chosen as the starting point for the active subnetworks, and at each step, the protein among the neighbours that are closer than a specified distance and that would yield the highest score upon being added to the current subnetwork is included. Nacu et al. [9] argues that even though the use of a greedy search reduces the amount of subnetworks being searched and thus can get stuck in a local maxima, it is still better than using a randomized algorithm by picking the neighbouring protein to be added to the current subnetwork at random which would facilitate the search of more subnetworks at a cost at speed. Since the work of Sohler et al. [8] the greedy approach has been adopted in many studies [10, 11, 12, 13]. Searching strategy in the study of Jia et al. [14] is also similar with the utilization of a greedy search algorithm, one difference being that they use GWAS data as opposed to expression data to detect a set of disease markers.

Rajagopalan and Agarwal [15] attempt a graph-based heuristic approach to detect subnetworks that maximally include all proteins of a particular biological pathway. They start by calculating corrected node scores for every gene in the network based on their p-value and then grouping nodes with positive scores into a subnetwork using a breadth-first search. Starting with the maximal-scoring subnetwork, a depth-first search detects paths to other subnetworks which are merged with the current subnetwork if the process improves the overall score. Dao et al. [5] employs a color coding technique for their network-based classification algorithm (OptDis) for the development of

subnetwork markers using expression profiles of breast cancer patients treated with combination chemotherapy. On the other hand, Qiu et al. [16] followed a mathematical programming based method where a diffusion kernel matrix describes the interaction of connected genes with the Pearson correlation based on their expression and then each gene is categorized as 'active' or 'not active' using a support vector regression approach with the tool RegMOD. In another study, Backes et al. [17] proposes a branch-and-cut based approach for the identification of deregulated subnetworks which can be performed on both directed (e.g., regulatory networks) and undirected graphs (e.g., PPI networks) for the search of maximally-connected subnetwork.

Genetic algorithms have also been used in the identification of active disease-associated subnetworks. Klammer et al. [18] presented an algorithm called SubExtractor that combines phosphoproteomic data with protein network information from STRING to identify differentially regulated subnetworks. The network created is based on a Bayesian probabilistic model that accounts for information about both differential regulation and network topology with the method being heavily constructed upon a genetic algorithm. Wu et al. [19] also uses a genetic algorithm which they argue as an improvement on the use of greedy search algorithms as though they are fast, they may not succeed in the determination of the optimal subnetwork markers and consequently reduce the performance of the successive learning machines.

Chapter 3

PRELIMINARIES

In section 3.1, background information on the genome-wide association studies is provided, followed by background information on the Rheumatoid Arthritis in section 3.2.

3.1. Background on Genome-Wide Association (GWA) Studies

Genome-wide association studies (GWAS) represent a recently developed research technique that has evolved into a powerful tool for investigating the genetic structure of human disease. GWAS aims to detect genetic risk factors for common, complex diseases (e.g., Rheumatoid Arthritis) by analyzing DNA sequence variations from across the human genome [20]. The variations that are targeted by GWAS are the single nucleotide polymorphisms (SNPs) that are common to the human genome and the purpose of the technique is to determine how these polymorphisms are distributed across different populations [21]. The ultimate aim of GWAS is to employ genetic risk factors to determine an individual's risk of developing a particular disorder and to understand the reasons of disease susceptibility in order to come up with new prevention and treatment plans [20].

Single nucleotide polymorphisms are found to be the most common type of DNA sequence variation encountered in human genome with an estimated 10 million [21]. In GWA studies, case-control setup is adopted in which two groups of individuals, one carrying the disease in question and the other being the healthy control group, are genotyped for common SNPs. It is then investigated which SNPs are encountered more

in the case group with a distinct difference which allows a statistical estimate being made about the level of heightened risk for each SNP using their odds ratio. Then with a chi-squared test, this odds ratio is converted into a p-value representing the significance of the SNP based on the frequency in which it occurs in the diseased individuals. The higher the frequency is, the lower the p-value will be.

In a notable study conducted in 2007 by the Wellcome Trust Case Control Consortium (WTCCC), 14,000 people were genotyped for seven common diseases with 2,000 people for each disease and 3,000 healthy individuals for the shared control group [22]. This study was the largest GWAS to be ever carried out at its time and it allowed many genetic markers for these common diseases to be discovered that have been helpful for the development of treatment strategies.

The GWA studies have been made more practical and less expensive by the use of the DNA microarray which is a small glass slide with a collection of microscopic DNA spots attached to it in a specific pattern [21]. The principle of microarrays is hybridization between two DNA strands. When a sample of DNA fragments is placed on the array, some of the DNA will hybridize to a probe on the surface and the rest will be washed away. Then the use of a scanning technology enables the researcher to detect in which parts of the array there has been a binding between the probe and the sample, and to what amount, which then can help with building a statistical estimation of increased risk for developing the disorder as explained above.

3.2. Background on Rheumatoid Arthritis (RA)

Rheumatoid arthritis (RA) is a chronic autoimmune disease of unknown etiology that causes joint inflammation and pain in the parts of body like feet, hands, hips and knees. The underlying mechanism involves the immune system, which is designed to protect the health of the body by attacking foreign substances (e.g., bacteria), attacking the joints instead, and consequently causing inflammation and thickening of the joint capsule. Rheumatoid arthritis occurs in 1% of the developed world's population [23] and is two to three times more prevalent in women than men with this difference being more pronounced in people of age less than 50 [24].

In the pathophysiology of Rheumatoid Arthritis, both genetic and environmental factors are implicated. While the main environmental risk to RA is thought to be smoking [23], more than half of the risk of having RA is attributed to genetic factors which are not completely discovered even though they have been researched for more than a decade. With the disease being encountered as frequently as 1 in every 100 people, it is important to continue the research to determine the genetic reasoning behind it.

Chapter 4

DATASETS

4.1. Protein-Protein Interaction (PPI) Network

In this thesis, two sets of data are used. The first dataset represents the human protein-protein interaction (PPI) network as a list of pairwise interactions between proteins and was obtained from the supplementary material of Goh et al.'s study [25]. This dataset first contains the PPIs acquired by testing binary interactions between proteins using a stringent, high-throughput yeast two-hybrid system [26, 27], and then the PPIs derived from literature by manual curation [26]. In this dataset, there are, in total, 61,070 interactions between 10,174 genes with 22,052 of them being non-self-interacting and non-redundant interactions.

4.2. Genetic Association Data of Rheumatoid Arthritis

The second dataset that was used in this thesis contains the genes that have been found, in a genome-wide association (GWA) study performed by the Wellcome Trust Case Control Consortium (WTCCC), to be significant for the disease of rheumatoid arthritis [22], which indicates these genes as being possibly involved in the development of the disease. In the mentioned GWA study, from the British population, 1999 patients with rheumatoid arthritis and 3004 healthy individuals as controls were examined. Using the Affymetrix GeneChip 500K Human Mapping Array Set, 500,475 single nucleotide polymorphisms (SNPs) were tested on these 5,003 samples. In the end, 25,027 SNPs were identified, showing nominal evidence of association with the disease, based on their genotypic p-values of association ($p < 0.05$). In a following study by Burcu-Bakir

and Sezerman [28], this SNP data and their genotypic p-values of association were used to assign these SNPs into 4,029 genes using the SPOT web server [29] by considering all known SNP/gene transcript associations. Then to take the possible associations between SNPs and their conserved transcription factor binding sites (TFBSs) into account, an additional 65 proteins (transcription factors), each protein known to bind to a TFBS a RA-associated SNP resides in, were added to the set using the SNPnexus program [30], bringing the number of genes in the dataset to a total of 4,094 genes. In order to incorporate functional information (regional score) to these genes, genotypic p-values were weighted by the functional scores of the SNPs that have been mapped to those genes, and a weighted P-value (Pw-value) was calculated for each gene which was consequently assigned to the gene as its p-value. In this thesis, this final gene set composed of 4,094 genes along with their assigned p-values [28] representing their significance to the rheumatoid arthritis is utilized to determine active subnetworks of this disease.

Chapter 5

METHOD

5.1. Scoring: Edge Weight Calculation

In the presented methodology, a novel scoring scheme is developed to assign scores to interactions between edges, called an edge weight, which would reflect the importance of said interaction. In this scheme, first, a score, $score(E)$, is assigned to the edge E that connects genes i and j, by multiplication of significance value p_u of both genes, where u represents the gene, using equation (1). Then, this score is converted into a standard score (z-score) with equation (2), where Φ^{-1} is the inverse normal cumulative distribution function and z_E denotes the z-score of the edge E. The value of z_E will be assigned to the edge as its weight.

$$score(E) = p_i * p_j \tag{1}$$

$$z_E = \Phi^{-1}\left(1 - score(E)\right) \tag{2}$$

In the method developed by Ideker et al. [3], a scoring scheme that is somewhat similar to our scheme in the way of converting p-values to z-scores is used, and a value of 0.5 is appointed to the nodes without p-values. This is equal to placing neutral significance to these nodes, which is a plausible idea when working with a PPI network that does not have many null-valued nodes. But in this case, where there are 8147 null nodes out of 10174 nodes, giving neutral significance to most of the nodes in the network will cause the final output network to have more null nodes than it is meaningful. Moreover, in a GWAS study, a node being null indicates it being insignificant for the disease, rheumatoid arthritis, as explained in section 3.1. Therefore it has been decided to assign

1 as the p-value to these genes to declare them as insignificant. However, in the case of both genes of an interaction being assigned the value of 1, instead of using equation (2) to calculate the z-score, the edge weight is set to be zero (0) directly as appointing 1 to both p-values would lead to negative infinity in the equation.

5.2. Subnetwork Search by the Markov Cluster Algorithm

In this thesis, the Markov Cluster (MCL) Algorithm that is proposed by Van Dongen is used to identify the active disease-associated subnetworks among the human protein-protein interaction network. The MCL algorithm is an unsupervised graph clustering algorithm that is based on the idea that there are more links in a cluster than between clusters and by simulating this stochastic flow in graphs, the clusters can be obtained [31].

5.2.1. Graph representation

As the MCL algorithm is a clustering algorithm for graphs, the protein-protein interaction network in this case is represented as a graph which is composed of *nodes* denoting genes and *edges*, which are the lines connecting these nodes, representing the interactions between the proteins coded by the genes denoted by said nodes. The graph is undirected, meaning that there is no distinction between the two nodes associated with each edge.

In order to be able to perform mathematical operations on the graph, it is expressed in a matrix format where each row and column denotes a gene while each matrix entry represents the edge weight between those genes. As the graph is undirected, the matrix will be symmetric at first. However, before the beginning of the MCL algorithm, it is required to perform a scaling step, in the form of normalizing each column, such that the resulting matrix will be stochastic. This means that the matrix elements on each column will correspond to probability values with each column summing up to 1 and the matrix not being symmetric anymore.

5.2.2. Clustering scheme

Natural clusters in a graph are depicted by the existence of many interactions among the nodes of a cluster, and fewer interactions between the nodes of different clusters. The MCL algorithm is based on the idea that random walks upon the graph will more likely result in staying within the natural cluster than travel between [31]. Therefore, by performing random walks on the graph, the algorithm attempts to detect where the flow tends to gather, and thus, where clusters are. The simulation of random walks is done by alternating between two processes called expansion and inflation operations.

5.2.3. Expansion operation

In expansion step, the power of Markov Chain transition matrix, edge weight matrix, is taken using the normal matrix product (e.g., matrix squaring). This allows flow to connect different regions of the graph that are not connected directly by the presence of only one edge.

Since there are only 61,070 interactions between 10,174 genes in our network, the matrix of edges will be a sparse matrix with most of the entries having zero-value. Thus in this study, while implementing the algorithm, in order to increase the speed of the matrix multiplication process and to decrease the memory demand, having a sparse matrix is taken advantage of by converting it to a sparse matrix format, which in this case is, the Compressed Row Storage (CRS) format. CRS format uses three arrays: val, which stores the values of non-zero elements of the matrix, col_ind, which stores the column indices of the elements in val array, and row_ptr, which stores the locations in the val array that start a row. In this way, the required memory cells to store an N by N matrix is reduced to 2NNZ+N+1, where NNZ denotes number of non-zero elements, from N^2 which is the number of memory cells needed to store the matrix in a standard matrix format (e.g., a 2-D vector). Then the matrix multiplication is done between the matrix and the CRS which represents the same network in a different structure, and this decreases the complexity of matrix multiplication from $O(N^3)$ to $O(NNZ \times N)$ algebraic operations including both multiplications and additions.

5.2.4. Inflation operation

Inflation coincides with raising each matrix entry to a given non-negative power, followed by a scaling step to return the matrix to a stochastic state, which is done by re-normalizing of each column. This operation is responsible for further strengthening strong currents and weakening already weak currents so that the less popular links between nodes can be demoted.

After every inflation step, edges are evaluated according to a threshold that is decided to be $1x10^{-6}$. If the weight between two nodes is less than $1x10^{-6}$, the edge between them is eliminated. In this way, inflation operation reduces the number of edges, while expansion raises them.

5.2.5. Stopping criteria

Expansion and inflation operations are iteratively used to strengthen the graph where it is strong and to weaken where it is weak. Ultimately, the iteration of these operations concludes in the segmentation of the graph into distinct components. The resulting components do not have any interactions between them anymore and the collection of these final components is understood as clustering [31].

Though global convergence is hard to prove, in practice, the process almost always converges to a doubly idempotent matrix, meaning that it does not change with further steps, and it is at a steady state [31]. In this state, every non-zero value in a single column has the same number making the column, in a sense, homogeneous.

5.2.6. Significance score calculation

After the MCL algorithm is finished, all genes in the network are separated into different clusters and every gene belongs to only one cluster. Then, in order to analyze the significance of the clusters, a scoring scheme is used to assign a score to each cluster. This *Cluster Score* (S_C) is calculated by multiplication of significance value p_i of each gene using the following equation, where n denotes the number of nodes in a cluster and C is the set of genes in the cluster.

$$S_C = \left(\prod_{i \in C} p_i \right)^{1/n}$$

Based on this formula, the lower the p-values of each gene in the cluster is, the lower the cluster score will be; which, in turn, would mean that the most significant clusters will have the lowest cluster scores.

5.2.7. Subnetwork elimination

Due to the nature of the MCL algorithm, a number of very small clusters emerge at the end; and the significance of these clusters, in terms of relation to the disease, (e.g., Rheumatoid Arthritis) should be evaluated before the other steps of the proposed method, so that the clusters that are deemed unimportant can be eliminated. Their relativity to the disease is evaluated by the usage of cluster score explained in section 5.2.6. Then, the clusters with score more than a given threshold value, and also the ones composed of less than 10 genes, simply for being too small to be significant, are eliminated.

5.3. Functional Enrichment of Identified Subnetworks

After the active subnetwork search algorithm detects the subnetworks with maximal scores, the next step is to evaluate if the genes in these subnetworks are really involved in the molecular mechanisms of the disease. Interpretation of such data is performed by finding the biological functions that are enriched in sets of genes. Functional enrichment is a technique for interpreting gene groups by statistical methods to identify functional annotations (e.g., pathways, cellular processes) the genes are associated with. It is done by comparing the group of detected genes with the genes known to be involved in a biological pathway to see if they match, which would mean that the subnetwork is related to that pathway. If the pathways found to be related to the subnetwork are also known to be a part of the development of RA, then it would be understood that the subnetwork in question is an active RA-related subnetwork.

The analysis uses the information about genes and their associated functions on biological databases (e.g., KEGG, Gene Ontology). In this thesis, for the functional enrichment of identified subnetworks, ClueGO plugin [32] of Cytoscape, which is an open-source Java program, is utilized. Even though ClueGO extract functional information about given genes utilizing KEGG, BioCarta databases and Gene Ontology [32], only the pathways obtained by using the KEGG database are used. During the functional enrichment process of ClueGO, a two-sided (enrichment/depletion) test based on the hypergeometric distribution is employed and Bonferroni correction method was used to correct the p-values for multiple testing.

Chapter 6

RESULTS and DISCUSSION

The proposed techniques were implemented in C++11; and their performance was tested on real datasets and compared with the performance of state-of-the-art techniques. The experiments were performed in a machine with 2.5Hz quad-core Intel Core i7 CPUs, 16 GB 1600MHz memory and OS X 10.10 Yosemite operating system. The complexity of the algorithm implemented is O(NNZ x N) where NNZ denotes the number of non-zero elements in the protein-protein transition matrix and N is the size of the matrix.

6.1. Parameters for Optimal Results

Starting with 4,094 genes that are found to be significant in a GWAS (WTCCC RA dataset), and a human protein-protein interaction network of 61,070 interactions between 10,174 genes, the MCL algorithm followed by a functional enrichment step was performed to identify RA-related genes and functionally important KEGG pathways. All interactions between genes were assigned an edge weight score to signify the importance of the interaction using the p-values of genes making up the interaction. Then the MCL algorithm was utilized for the search of active RA-associated subnetworks.

The MCL algorithm simulates random walks on the graph by alternating between two processes called expansion and iteration to extract potentially meaningful subnetworks by attempting to discover where the flow tends to gather in the network. After the discovery of subnetworks, functional enrichment step finds the KEGG pathways that

are associated with these subnetworks. In order to evaluate how meaningful found subnetworks are, the next step is to analyze their KEGG pathways and find how many of those pathways are related to RA. In order to do this, a detailed literature search is performed and it is seen how many of the best scoring pathways have been found to be related to Rheumatoid Arthritis in previous studies.

Since the MCL algorithm does not use a fixed expansion or inflation parameter value, different values are attempted to find the parameters that give the best results. In total, 15 combinations of parameters are used with expansion parameter taking the values of 2, 3, 4 and inflation parameter taking the values of 2, 2.5, 3, 3.5 and 4. After the use of all 15 combinations, different subnetworks are found and functional enrichment step is performed on all of these subnetworks. In order to be able to determine which parameter combination finds the best subnetworks, the KEGG pathways found to be associated with these subnetworks are evaluated and it is assessed how many of these pathways are found to be related to RA in previous studies as explained above.

Expansion parameter	Inflation parameter	Threshold	Top 20 pathways	Top 40 pathways
2	2	-	9	14
		0.30 / 0.23	12	13
	2.5	-	10	13
		0.32 / 0.22	11	11
	3	-	10	13
		0.34 / 0.26	11	11
	3.5	-	10	13
		0.32 / 0.28	11	11
	4	-	10	11
		0.44 / 0.35	9	9
3	2	-	12	25
		0.20 / 0.12	19	27
	2.5	-	12	24
		0.14 / 0.12	19	28
	3	-	14	24
		0.14 / 0.11	19	29
	3.5	-	14	25
		0.16 / 0.12	19	26
	4	-	14	25
		0.14 / 0.12	19	26
4	2	-	11	20
		0.18 / 0.13	16	23
	2.5	-	11	23
		0.22 / 0.16	16	25
	3	-	11	23
		0.18 / 0.16	15	23
	3.5	-	10	20
		0.20 / 0.19	15	24
	4	-	10	18
		0.26 / 0.22	14	22

Table 1. The number of RA-related pathways found among the top 20 and top 40 scoring pathways associated with subnetworks detected by the MCL algorithm, with the use of each parameter combination, where threshold is used to eliminate clusters with cluster score higher than it, as explained in section 5.2.7.

First deduction to be made is that, the number of RA-related pathways found among the top 20 and top 40 scoring pathways of subnetworks found by the MCL algorithm increases with the use of a cluster score threshold which supports our decision of eliminating insignificant subnetworks using this threshold. Secondly, it can be seen that the usage of expansion parameter 3 gives the best results and 4 gives acceptable results while 2 gives the worst. Though it seems that the usage of inflation parameters from 2 to 4 does not change the results a great deal, the inflation parameter 3 combined with expansion parameter 3 gives the best results by finding 19 RA-related pathways among the top 20 scoring pathways and 29 RA-related pathways among the top 40 scoring pathways. Therefore we decided to explore the results of the usage of these parameters in detail in the following sections.

Expansion parameter	Inflation parameter	Threshold
2	2	0.28
	2.5	0.28
	3	0.28
	3.5	0.28
	4	0.35
3	2	0.12
	2.5	0.12
	3	0.12
	3.5	0.12
	4	0.12
4	2	0.16
	2.5	0.16
	3	0.16
	3.5	0.20
	4	0.24

Table 2. Thresholds that give the best results for each parameter combination.

6.2. Functionally Important KEGG Pathways for RA

At the end of the implemented modified MCL algorithm, 91 subnetworks were detected. Then the functional enrichment of all of the 91 subnetworks was carried out together in order to find the subnetworks that are related to RA the most and thus the candidate active disease-associated subnetworks. As a result of the functional enrichment step, 113 KEGG pathway terms were found to be associated with only 24 of the subnetworks, reducing the number of potential active subnetworks to 24. In Table 4 and Table 5, we represent 20 maximally-scoring pathways, determined by their Term P-values, which are mostly related to immunity, inflammation, and synaptic systems. We compared our findings with previously found RA-related KEGG pathways. Most of the pathways identified by the proposed methodology have been previously found to be associated with Rheumatoid Arthritis with experimental techniques and these pathways are Notch signaling, Circadian entrainment, NF-kappa B signaling, GABAergic synapse, Axon guidance, Jak-STAT signaling, Leukocyte transendothelial migration, MAPK signaling, TGF-beta signaling and Toll-like receptor signaling pathways.

ECM-receptor interaction, which was discovered as the most significant KEGG pathway by the described methodology, is thought to be associated with RA as fibroblast-like synoviocytes (FBS) from RA synovium was detected to be binding to extracellular matrix (ECM) proteins more than the normal FBS which was concluded as the tight binding of rheumatoid FBS to the ECM proteins playing a role in ECM remodeling in the rheumatoid process in vivo [33]. The contribution of Notch signaling pathways is that macrophages are thought to play a pathogenic role in rheumatoid arthritis by secreting inflammatory mediators that contribute to joint inflammation and bone erosion and the Notch pathway has been believed to be influencing the development of macrophages for some time [34]. In following studies, Notch signaling has been demonstrated to be active in CD4$^+$ T cells during the development of RA and also to be playing a significant role in Th1 and Th17 cell differentiation which displays the role of Notch signaling pathways in the development of RA [35]. After the observation of the molecular machinery controlling the circadian rhythm being disturbed in RA patients [36], Circadian entrainment pathway is also thought to be affected by RA. NF-kappa B signaling has long been a pathway recognized with its relation to RA with the transcription factor NF-kappa B being a pivotal regulator of

inflammation; and recent studies have also supported this view by revealing a broad involvement of NF-kappa B in other aspects of RA pathology, including development of T helper 1 responses, activation, abnormal apoptosis and proliferation of RA fibroblast-like synovial cells, and differentiation and activation of bone resorbing activity of osteoclasts [37, 38, 39, 40]. Since the activation of peripheral GABA receptors were demonstrated to inhibit the development of RA in the collagen-induced arthritis (CIA) mouse model of RA [41], GABAergic synapse has thought to be involved in RA. Axon guidance is another pathway believed to be implicated in RA after recent findings of Semaphorin-3A, which is a member of a large family of conserved proteins originally implicated in axon guidance, increasing the CD4[+]NP-1[+] T cell ability to suppress alloresponses and its transient expression being altered in rheumatoid inflammation [42]. Jak-STAT signaling, Leukocyte transendothelial migration, MAPK signaling, Toll-like receptor signaling pathways are all pathways found to be significantly involved in RA [28]. Finally, TGF-beta signaling pathway have also been believed to be associated with RA via its relation to ECM with the action of transforming-growth-factor (TGF)-β following inflammatory responses is being characterized by increased production of extracellular matrix (ECM) components [43, 44].

Some of the other pathways identified by the described methodology have been previously found to be related to RA with computational techniques. These pathways are Morphine addiction, Focal adhesion, Glutamatergic synapse, Retrograde endocannabinoid signaling, Cholinergic synapse and Dopaminergic synapse pathway. All of these pathways have been shown to be associated with RA in a recent study [45] where two GWAS were carried out using RA datasets from both GAW16 (Genetic Analysis Workshop 16) and the WTCCC, and all SNPs were mapped to genome-wide autosomal genes followed by a calculation of gene-wise risk values by minimum P-value method. The KEGG pathway risk scores were determined by Fisher combination method and the significant pathways were identified by a permutation test. Focal adhesion pathway was also experimentally demonstrated to be involved in cellular processes such as osteoclast pathology and angiogenesis, which are known to be significant for RA [46].

Additionally, the result of recent experimental studies suggest PI3K-Akt signaling pathways and Complement and coagulation cascades to be in relation with RA, though the mechanisms are still not completely known. In a study, it has been shown that PI3Kɣ blockade by both genetic and pharmacological approaches reduces joint inflammation and damage in collagen-induced arthritis indicating PI3K as potentially involved in development of RA [47].

KEGG Term	Significance Score	Term P-value	Number genes	Percent genes	Number genes in pathway
ECM-receptor interaction	0.109866	2.16E-45	42	48.3%	87
Morphine addiction	0.087881	4.46E-45	31	33.7%	92
Notch signaling pathway	0.08836	2.59E-33	20	41.7%	48
Focal adhesion	0.109866	6.32E-32	46	22.2%	207
Circadian entrainment	0.087881	2.73E-30	24	25.0%	96
NF-kappa B signaling pathway	0.069851	2.09E-29	30	33.0%	91
GABAergic synapse	0.087881	2.86E-29	23	25.6%	90
Glutamatergic synapse	0.087881	3.19E-28	24	20.9%	115
Retrograde endocannabinoid signaling	0.087881	9.20E-28	23	22.5%	102
Cholinergic synapse	0.087881	9.57E-27	23	20.5%	112
Axon guidance	0.062392	6.85E-26	16	12.6%	127
Jak-STAT signaling pathway	0.074546	5.24E-25	54	34.6%	156
Serotonergic synapse	0.087881	6.46E-25	22	19.5%	113
Leukocyte transendothelial migration	0.074546	1.07E-24	47	39.8%	118
MAPK signaling pathway	0.033538	1.49E-23	28	10.9%	256
Dopaminergic synapse	0.087881	1.72E-23	22	16.9%	130
TGF-beta signaling pathway	0.073886	2.45E-23	16	20.0%	80
Complement and coagulation cascades	0.096368	4.76E-21	12	17.4%	69
Toll-like receptor signaling pathway	0.069851	6.33E-19	24	22.6%	106
PI3K-Akt signaling pathway	0.109866	3.11E-17	41	11.8%	346

Table 3. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search. Significance score is the cluster score explained in section 5.2.6 and term p-value is a score given to reflect the importance of the pathway by the functional enrichment step. 'Number genes' denotes the number of genes in the subnetwork found to be associated with the given pathway. Likewise, 'percent genes' denotes the percentage of these genes among the total number of genes of the given pathway which is denoted by 'number genes in pathway'.

Among the twenty best scoring pathways found by the proposed methodology, the only pathway found that has not been shown to be associated to RA previously, to the best of our knowledge, is Serotonergic synapse pathway. Even though this pathway has not been demonstrated to be in relation to RA by experimental or computational methods, the results of some clinical studies suggests a relation between the two. In one study, the amount of serotonin receptors in RA patients has been observed to be significantly decreased, suggesting either the reduced amounts of the receptors to cause a susceptibility to the disease or be a secondary effect of the disease [48]. Similarly, in a case study, after a SSRI uptake, which is thought to increase extracellular serotonin concentrations, a continued remission of RA in a patient has been observed which suggests serotonin receptors playing a role in mediating inflammatory processes [49].

The fact that all of the best scoring KEGG pathways identified by the described methodology have been previously found to be associated with RA experimentally, computationally or by clinical studies demonstrates the methodology as a powerful tool to detect active RA-associated subnetworks while also supporting our decision of using GWAS data as the genetic association data of RA.

All of the pathways described above along with the genes found in subnetworks to be associated with those pathways are displayed in Table 4 and 5.

KEGG Term	Significance Score	Term P-value	Num. Genes	Associated Genes Found
ECM-receptor interaction	0.109866	2.16E-45	42	ITGB1*, ITGB5*, ITGB3*, LAMA3*, TNC*, LAMC2*, LAMC1*, THBS1*, COMP*, VTN*, RELN*, ITGB8*, ITGAV*, ITGB7*, CD36*, ITGB6*, ITGA4*, LAMB3*, GP1BB*, ITGA3*, ITGA2*, ITGA1*, FN1*, GP1BA*, GP5*, HSPG2*, COL1A1*, GP9*, COL1A2*, COL2A1*, COL4A2*, COL4A1*, COL4A4*, ITGA10*, COL4A3*, ITGA11*, COL4A6*, ITGA8*, COL4A5*, ITGA6*, ITGA5*, ITGA9*,
Morphine addiction	0.087881	4.46E-45	31	PDE1C*, PDE1B*, PDE1A*, ADCY2*, PRKX*, ADCY1*, ADCY8*, GNGT1*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNG7*, PRKACG*, ADORA1*, GNG8*, PDE4A*, PRKACA*, PRKACB*, PDE4D*, PDE4C*, GNG12*, GNG11*, GNG13*, GNB2*, GNB1*, GNAS*, GNB4*, GNB3*, GNB5*,
Notch signaling pathway	0.08836	2.59E-33	20	JAG2*, NOTCH2*, PSENEN*, NOTCH3*, JAG1*, NOTCH1*, MAML2*, MAML1*, NOTCH4*, PSEN2*, DTX1*, PSEN1*, RBPJ*, DLL1*, DLL4*, LFNG*, NCSTN*, APH1B*, MFNG*, MAML3*,
Focal adhesion	0.109866	6.32E-32	46	ITGB1*, FIGF*, SHC3*, ITGB5*, FLT4*, ITGB3*, LAMA3*, TNC*, ILK*, LAMC2*, LAMC1*, ARHGAP5*, THBS1*, COMP*, VTN*, RELN*, CAPN2*, ITGB8*, FLNB*, ITGAV*, ITGB7*, ITGB6*, ITGA4*, LAMB3*, ITGA3*, HGF*, ITGA2*, ITGA1*, FN1*, PTK2*, COL1A1*, COL1A2*, COL2A1*, COL4A2*, COL4A1*, COL4A4*, ITGA10*, COL4A3*, ITGA11*, COL4A6*, ITGA8*, COL4A5*, ITGA6*, ITGA5*, TLN1*, ITGA9*,
Circadian entrainment	0.087881	2.73E-30	24	ADCY2*, PRKX*, ADCY1*, ADCY8*, GNG12*, GNG11*, GNG13*, GNGT1*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNAS*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,
NF-kappa B signaling pathway	0.069851	2.09E-29	30	TRADD*, LY96*, TNFAIP3*, TNFRSF11A*, RELA*, RELB*, IKBKB*, IRAK1*, RIPK1*, IKBKG*, MAP3K7*, TICAM2*, CHUK*, TNFSF14*, DDX58*, TRAF2*, IRAK4*, TRAF1*, NFKB1*, TIRAP*, NFKB2*, TNFRSF1A*, NFKBIA*, TRAF6*, TAB2*, TAB1*, MAP3K14*, TLR4*, MYD88*, BIRC3*,
GABAergic synapse	0.087881	2.86E-29	23	ADCY2*, PRKX*, ADCY1*, ADCY8*, GNG12*, GNG11*, GNG13*, GNGT1*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,
Glutamatergic synapse	0.087881	3.19E-28	24	ADCY2*, PRKX*, ADCY1*, ADCY8*, GNG12*, GNG11*, GNG13*, GNGT1*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNAS*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,
Retrograde endocannabinoid signaling	0.087881	9.20E-28	23	ADCY2*, PRKX*, ADCY1*, ADCY8*, GNG12*, GNG11*, GNG13*, GNGT1*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,
Cholinergic synapse	0.087881	9.57E-27	23	ADCY2*, PRKX*, ADCY1*, ADCY8*, GNG12*, GNG11*, GNG13*, GNGT1*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,

Table 4. Pathways from 1 to 10 among the 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, along with the genes associated with those pathways.

KEGG Term	Significance Score	Term P-value	Number Genes	Associated Genes Found
Axon guidance	0.062392	6.85E-26	16	EPHA5*, EPHA4*, EPHA7*, EPHA6*, EFNA5*, EFNA4*, EFNA1*, EFNB2*, EFNA3*, EFNA2*, EFNB3*, EPHB2*, EPHB1*, EPHA3*, NGEF*, EPHA2*,
Jak-STAT signaling pathway	0.074546	5.24E-25	54	IFNA5*, CSF2*, IFNA1*, IL23R*, IFNA2*, MPL*, CBLC*, IL5RA*, CBLB*, IFNA8*, GHR*, SPRED2*, SPRED1*, JAK2*, JAK1*, IFNAR2*, IL15RA*, IFNA13*, CISH*, IFNGR1*, IL15*, IFNGR2*, TYK2*, OSMR*, PRLR*, IL23A*, IL3RA*, SOS1*, SOS2*, IRF9*, IFNAR1*, CSF2RB*, PIK3R2*, PIK3R1*, CSF2RA*, SOCS3*, SOCS1*, SOCS5*, STAT5A*, STAT5B*, TSLP*, IFNB1*, STAT1*, STAT2*, STAT3*, PTPN11*, STAM*, IFNW1*, IL3*, IL5*, IL2RB*, SPRY2*, PTPN6*, IL7R*,
Serotonergic synapse	0.087881	6.46E-25	22	PRKX*, GNG12*, GNG11*, GNG13*, GNGT1*, GNG10*, GNG3*, HTR6*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNAS*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,
Leukocyte transendothelial migration	0.074546	1.07E-24	47	ITK*, ROCK1*, NCF2*, TXK*, CTNND1*, ITGB2*, GNAI3*, PIK3R2*, PIK3R1*, THY1*, CLDN2*, F11R*, CLDN1*, MLLT4*, ACTB*, ICAM1*, CDC42*, PLCG2*, PTK2B*, CTNNA3*, CTNNA2*, PLCG1*, RAC1*, JAM2*, JAM3*, PRKCG*, VASP*, VAV3*, ACTN3*, PRKCB*, CYBB*, RHOH*, CYBA*, PRKCA*, PTPN11*, ACTN4*, VAV2*, CLDN6*, CLDN5*, CLDN4*, CLDN3*, CLDN8*, CLDN7*, PECAM1*, CTNNB1*, CLDN16*, VCL*,
MAPK signaling pathway	0.033538	1.49E-23	28	ATF2*, PTPRR*, ZAK*, STK4*, DUSP16*, ELK1*, RPS6KA4*, DUSP10*, RPS6KA5*, MKNK1*, MKNK2*, MAP2K6*, MAPK3*, DUSP4*, MAP2K3*, MAP3K2*, MAP2K4*, DUSP2*, MEF2C*, MAP3K1*, DUSP1*, MAPK14*, DUSP7*, MAPK13*, MAPK11*, MAPKAPK3*, MAPKAPK5*, PTPN7*,
Dopaminergic synapse	0.087881	1.72E-23	22	PRKX*, GNG12*, GNG11*, GNG13*, GNGT1*, PPP1CB*, GNG10*, GNG3*, GNG2*, GNG5*, GNG4*, GNB2*, GNG7*, PRKACG*, GNB1*, GNAS*, GNB4*, GNB3*, GNG8*, GNB5*, PRKACA*, PRKACB*,
TGF-beta signaling pathway	0.073886	2.45E-23	16	BMPR2*, AMHR2*, NOG*, GDF6*, SMAD6*, ACVR2B*, BMP7*, GDF5*, ACVR2A*, BMP6*, GDF7*, SMAD7*, BMP4*, BMP2*, BMPR1B*, BMPR1A*,
Complement and coagulation cascades	0.096368	4.76E-21	12	F10*, VWF*, SERPINC1*, PROS1*, C4BPA*, C4BPB*, F2*, F3*, F5*, F7*, F9*, PROC*,
Toll-like receptor signaling pathway	0.069851	6.33E-19	24	TICAM2*, CHUK*, LY96*, IRAK4*, NFKB1*, RELA*, TIRAP*, IKBKB*, NFKBIA*, TLR1*, TBK1*, IRAK1*, TRAF6*, AKT3*, MAP3K8*, RIPK1*, TAB2*, IKBKG*, TAB1*, TLR5*, IKBKE*, MAP3K7*, TLR4*, MYD88*,
PI3K-Akt signaling pathway	0.109866	3.11E-17	41	ITGB1*, FIGF*, ITGB5*, FLT4*, ITGB3*, LAMA3*, TNC*, LAMC2*, LAMC1*, THBS1*, COMP*, VTN*, RELN*, ITGB8*, ITGAV*, ITGB7*, ITGB6*, ITGA4*, LAMB3*, ITGA3*, HGF*, ITGA2*, ITGA1*, FN1*, OSM*, PTK2*, COL1A1*, COL1A2*, COL2A1*, COL4A2*, COL4A1*, COL4A4*, ITGA10*, COL4A3*, ITGA11*, COL4A6*, ITGA8*, COL4A5*, ITGA6*, ITGA5*, ITGA9*,

Table 5. Pathways from 11 to 20 among the 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, along with the genes associated with those pathways.

6.3. Use of Threshold for Cluster Score

At the end of the functional enrichment step, Term P-value scores are used to order KEGG pathways found to be associated with the subnetworks identified by the methodology. The use of these scores to evaluate the significance of the pathways is very effective since during the scoring process both the number of genes in the subnetwork that are found to be associated with the particular pathway and the size of the subnetwork are taken into account. However there is one downside to using this scoring scheme, and it is that in the case of having a small subnetwork where most of the genes have a p-value of 1, meaning that they are insignificant for RA, but are found to be associated with a specific pathway; and only a small number of the genes have a p-value lower than 0.05, but are not found to be associated with the aforementioned pathway; the scheme may give very low Term P-value scores to the pathway indicating that the pathway is an important one even though it is not a pathway significant for RA since the genes found to be in relation with the pathway are not significant genes for RA (p-value = 1). In order to avoid this issue, at the end of the MCL algorithm, a cluster score is assigned to each subnetwork which reflects the significance of the subnetwork in terms of relation to RA as explained in Section 5.2.6. The lower the cluster score is, the more significant the subnetwork is for RA. For this reason, before the functional enrichment step is carried out, subnetworks that have cluster scores higher than a given threshold are thought to be very insignificant for RA and thus eliminated. One example of such elimination can be seen in Table 6 and 7. Prior to elimination, pathways that are not involved in development of RA such as Ribosome, Nucleotide excision repair, RNA transport, DNA replication, Proteasome and Mismatch repair, can be mistakenly perceived as significant based on their Term P-values, but their Cluster Scores (Significance Scores) clearly shows their insignificance for RA. Therefore, in this thesis, it is proposed that in order to evaluate the importance of found subnetworks, the use of Term P-value scores by itself is not sufficient and can lead to irrelevant pathways being classified as significant. However, the use of Term P-value scores combined with our proposed Cluster Scores (Significance Scores) is very effective and gives more accurate results.

KEGG Term	Significance Score	Term P-value
~~Ribosome~~	0.550133	9.66E-126
~~Nucleotide excision repair~~	0.148243	2.66E-61
~~RNA transport~~	0.242717	4.56E-47
ECM-receptor interaction	0.109866	2.16E-45
Morphine addiction	0.087881	4.46E-45
~~DNA replication~~	0.148243	2.15E-42
~~Proteasome~~	0.247622	4.33E-38
Notch signaling pathway	0.08836	2.59E-33
Focal adhesion	0.109866	6.32E-32
Circadian entrainment	0.087881	2.73E-30
NF-kappa B signaling pathway	0.069851	2.09E-29
GABAergic synapse	0.087881	2.86E-29
Glutamatergic synapse	0.087881	3.19E-28
Retrograde endocannabinoid signaling	0.087881	9.20E-28
Cholinergic synapse	0.087881	9.57E-27
Axon guidance	0.062392	6.85E-26
Jak-STAT signaling pathway	0.074546	5.24E-25
Serotonergic synapse	0.087881	6.46E-25
Leukocyte transendothelial migration	0.074546	1.07E-24
MAPK signaling pathway	0.033538	1.49E-23
Dopaminergic synapse	0.087881	1.72E-23
TGF-beta signaling pathway	0.073886	2.45E-23
~~Mismatch repair~~	0.148243	4.45E-23
Complement and coagulation cascades	0.096368	4.76E-21
Toll-like receptor signaling pathway	0.069851	6.33E-19
PI3K-Akt signaling pathway	0.109866	3.11E-17

Table 6. The best scoring KEGG pathways that are associated with identified subnetworks before subnetwork elimination (according to the threshold of 0.12). The pathways which are striked-through are the ones to be eliminated.

KEGG Term	Significance Score	Term P-value
ECM-receptor interaction	0.109866	2.16E-45
Morphine addiction	0.087881	4.46E-45
Notch signaling pathway	0.08836	2.59E-33
Focal adhesion	0.109866	6.32E-32
Circadian entrainment	0.087881	2.73E-30
NF-kappa B signaling pathway	0.069851	2.09E-29
GABAergic synapse	0.087881	2.86E-29
Glutamatergic synapse	0.087881	3.19E-28
Retrograde endocannabinoid signaling	0.087881	9.20E-28
Cholinergic synapse	0.087881	9.57E-27
Axon guidance	0.062392	6.85E-26
Jak-STAT signaling pathway	0.074546	5.24E-25
Serotonergic synapse	0.087881	6.46E-25
Leukocyte transendothelial migration	0.074546	1.07E-24
MAPK signaling pathway	0.033538	1.49E-23
Dopaminergic synapse	0.087881	1.72E-23
TGF-beta signaling pathway	0.073886	2.45E-23
Complement and coagulation cascades	0.096368	4.76E-21
Toll-like receptor signaling pathway	0.069851	6.33E-19
PI3K-Akt signaling pathway	0.109866	3.11E-17

Table 7. The best scoring KEGG pathways that are associated with identified subnetworks after subnetwork elimination (according to the threshold of 0.12).

6.4. Comparative Studies

In order to evaluate the performance of the proposed methodology, the results have been compared with the results of state-of-the-art techniques. One such technique is the program PANOGA [28] which uses the simulated annealing implemented in jActiveModules plugin [3] for the active subnetwork search and then the ClueGO plugin [32] of Cytoscape for the functional enrichment step. Data the method is applied upon is the PPI network from Goh et al.'s study [25] and GWAS data taken from WTCCC [22]. It is important to note that the same data is also used in this thesis, though they use both SPOT [29] and F-SNP [50] p-values to incorporate functional information into genes while we used only SPOT p-values. Since as a result of their study, they only report the 20 highest scoring pathways found by their methodology, we decided to base this comparison on those pathways even though they are not particularly the highest scoring pathways in our results. The other techniques chosen to be compared with our technique is Wu et al., Martin et al. and Zhang et al. It is important to note that, the methods they develop and the datasets they apply their techniques on differ from the ones used by this methodology to some extent. Wu et al. exploits text-mining [51], Martin et al. uses GWAS data from WTCCC and NARAC studies and performs pathway analysis to prioritize regions containing genes that are involved with RA [52] and Zhang et al. develops a multidimensional screening approach which was applied on GAW16 (Genetic Analysis Workshop) data [53].

Comparative results of the performance of the proposed methodology and these four methods are shown in Table 8, in terms of number of genes found in commonly identified KEGG pathways. Additionally, since our program and the program PANOGA utilizes the same tool, the ClueGO plugin of Cytoscape, for the functional enrichment step, leading to the Term P-value scores being used to evaluate the detected KEGG pathways in both programs, our results are further compared with the results of PANOGA by using Term P-value scores.

As can be seen in Table 8, the number of genes found by our methodology is higher, in most cases, than the genes found by the other methods. Additionally, the Term P-value given to the pathways to describe its significance in the subnetworks is almost always lower in our results than the results of PANOGA which indicates our results to be

superior since the pathways become more significant as their Term P-value gets lower. These results demonstrate that the methodology proposed in this thesis is superior to these four methods described performance-wise.

KEGG Term	Number of genes found					Term P-values	
	Martin et.al.	Wu et.al.	Zhang et.al.	PANO GA	our method	PANO GA	our method
Focal adhesion	0	36	32	30	46	9.33E-11	6.32E-32
ErbB signaling pathway	0	23	0	20	10	2.13E-10	5.79E-13
Tight junction	0	0	5	22	38	1.80E-08	1.99E-13
Chemokine signaling pathway	0	0	0	26	22	2.31E-08	3.24E-24
Adherens junction	0	0	18	17	29	1.16E-07	8.83E-15
Bacterial invasion of epithelial cells	0	0	0	16	28	1.57E-07	3.10E-13
Neurotrophin signaling pathway	0	0	0	20	15	2.36E-07	9.69E-08
Long-term potentiation	22	0	7	15	7	3.67E-07	1.61E-05
Pathways in cancer	0	0	0	32	0	1.12E-06	0
Chronic myeloid leukemia	0	21	18	14	0	1.44E-06	0
Cell adhesion molecules (CAMs)	26	0	10	18	31	1.42E-05	1.02E-07
Leukocyte transendothelial migration	24	14	0	17	47	1.72E-05	1.07E-24
T cell receptor signaling pathway	21	16	16	16	26	2.70E-05	6.90E-08
Toll-like receptor signaling pathway	0	22	6	13	24	1.97E-03	6.33E-19
Antigen processing and presentation	0	0	3	11	22	2.08E-03	1.23E-10
Allograft rejection	0	0	0	8	5	2.16E-03	4.13E-09
MAPK signaling pathway	0	43	34	20	28	6.13E-03	1.49E-23
Type I diabetes mellitus	0	0	1	8	5	6.24E-03	8.74E-09
Apoptosis	18	12	11	11	13	6.48E-03	1.38E-16
Jak-STAT signaling pathway	25	0	16	15	54	7.41E-03	5.24E-25
Prostate cancer	0	22	0	11	9	5.04E-02	9.06E-04
Calcium signaling pathway	35	0	4	16	34	1.63E-01	4.79E-07
VEGF signaling pathway	0	15	13	9	0	2.71E-01	0

Table 8. Comparison of KEGG pathways found by our method with previous studies in terms of number of genes associated within each KEGG term; and an additional comparison with method PANOGA in terms of the score, term p-values.

6.5. Best Subnetworks and Potential Gene Markers

Using term p-values and cluster score, we identified 3 significant subnetworks as the candidate active RA-associated subnetworks on the basis of their aggregate degree of genetic association with RA, in terms of the KEGG pathways found to be represented by them. These three subnetworks can be seen in Tables 9, 10 and 11.

The first active subnetwork is composed of 727 genes and 727 edges, and 26 KEGG pathways are found to be associated with this subnetwork. Most of the KEGG pathways of this subnetwork are known to be related to RA either as a result of experimental studies: Jak-STAT signaling, Leukocyte transendothelial migration, T cell receptor signaling [28], B cell receptor signaling, Ras signaling, Rap1 signaling [54], Cytokine-cytokine receptor interaction pathways; or as a result of computational studies: Adherens junction, Tight junction, Bacterial invasion of epithelial cells, Cell adhesion molecules (CAMS) [28] and Calcium signaling pathways [28, 45].

The second active subnetwork is composed of 72 genes and 71 edges, and there are 20 KEGG pathways that are represented by this subnetwork. Almost half of these pathways have been found to be associated with RA through computational means: Morphine addiction, Glutamatergic synapse, Retrograde endocannabinoid signaling, Cholinergic synapse, Dopaminergic synapse and Long-term potentiation [45]; while some of them are shown to be related to RA experimentally: Circadian entrainment [36] and GABAergic synapse [41]. The fact that almost all of the mentioned pathways (Cholinergic synapse, Glutamatergic synapse, Dopaminergic synapse and Retrograde endocannabinoid signaling pathways) are synapse-related pathways and have been discovered to be related to RA (including Morphine addiction and Long-term potentiation also) in the same previous study [45] demonstrates how closely related the genes in the subnetwork are to each other and to RA, proving the success of the MCL algorithm in clustering.

The third active subnetwork is composed of 239 genes and 239 edges, and 20 KEGG pathways have been identified to be represented by this subnetwork. Some of those pathways have been shown to be RA-related previously through experimental work: NF-kappa B signaling [37], Toll-like receptor signaling [28], TNF signaling [55] and

Neurotrophin signaling [28]; and some through computational work: Measles [56] and Prostate cancer [28]. In addition, the involvement of Epstein-Barr virus (EBV) infection and RA has been investigated for more than two decades during which EBV has been speculated to be an environmental trigger for RA and even though a definite proof is yet to be discovered, a large amount of circumstantial evidence suggest a relation between them [57, 58]. Furthermore the NOD-like receptor signaling and RIG-I-like receptor signaling pathways found in this subnetwork are also believed to be related to RA [59] even though the mechanisms relating the two are not completely understood.

Both based on their cluster score and the term p-values of the KEGG pathways associated with them, all 3 subnetworks described above are significant candidates to be recognized as active RA-associated subnetworks. The fact that most of their associated KEGG pathways have been discovered to be related to RA previously strongly supports this conclusion.

KEGG Term	Significance Score	Term P-value
Jak-STAT signaling pathway	0.074546	5.24E-25
Leukocyte transendothelial migration	0.074546	1.07E-24
Adherens junction	0.074546	8.83E-15
Tight junction	0.074546	1.99E-13
Bacterial invasion of epithelial cells	0.074546	3.10E-13
Natural killer cell mediated cytotoxicity	0.074546	4.34E-13
Fc gamma R-mediated phagocytosis	0.074546	1.01E-12
B cell receptor signaling pathway	0.074546	4.86E-11
Rap1 signaling pathway	0.074546	6.02E-11
Fc epsilon RI signaling pathway	0.074546	1.61E-09
Proteoglycans in cancer	0.074546	5.65E-09
Gap junction	0.074546	6.00E-08
T cell receptor signaling pathway	0.074546	6.90E-08
Cell adhesion molecules (CAMs)	0.074546	1.02E-07
Ras signaling pathway	0.074546	1.89E-07
Calcium signaling pathway	0.074546	4.79E-07
Platelet activation	0.074546	6.93E-07
Regulation of actin cytoskeleton	0.074546	1.38E-06
Phosphatidylinositol signaling system	0.074546	1.61E-06
Pathogenic Escherichia coli infection	0.074546	2.26E-06
Cytokine-cytokine receptor interaction	0.074546	8.14E-06
cGMP-PKG signaling pathway	0.074546	1.32E-05
cAMP signaling pathway	0.074546	2.48E-05
HIF-1 signaling pathway	0.074546	2.16E-04
Oxytocin signaling pathway	0.074546	5.93E-04
Prolactin signaling pathway	0.074546	7.93E-04

Table 9. The 26 pathways found to be related to the first active subnetwork that is composed of 727 genes and 727 edges.

KEGG Term	Significance Score	Term P-value
Morphine addiction	0.087881	4.46E-45
Circadian entrainment	0.087881	2.73E-30
GABAergic synapse	0.087881	2.86E-29
Glutamatergic synapse	0.087881	3.19E-28
Retrograde endocannabinoid signaling	0.087881	9.20E-28
Cholinergic synapse	0.087881	9.57E-27
Serotonergic synapse	0.087881	6.46E-25
Dopaminergic synapse	0.087881	1.72E-23
Alcoholism	0.087881	4.75E-16
Taste transduction	0.087881	2.84E-12
Ovarian steroidogenesis	0.087881	7.59E-11
Bile secretion	0.087881	2.82E-09
Gastric acid secretion	0.087881	9.99E-08
Insulin secretion	0.087881	3.43E-07
Salivary secretion	0.087881	5.15E-07
Thyroid hormone synthesis	0.087881	1.46E-06
Long-term potentiation	0.087881	1.61E-05
Vasopressin-regulated water reabsorption	0.087881	2.49E-05
Endocrine and other factor-regulated calcium reabsorption	0.087881	7.31E-04
Cocaine addiction	0.087881	8.93E-04

Table 10. The 20 pathways found to be related to the second active subnetwork that is composed of 72 genes and 71 edges.

KEGG Term	Significance Score	Term P-value
NF-kappa B signaling pathway	0.069851	2.09E-29
Toll-like receptor signaling pathway	0.069851	6.33E-19
Epstein-Barr virus infection	0.069851	3.44E-16
RIG-I-like receptor signaling pathway	0.069851	4.39E-15
Toxoplasmosis	0.069851	5.04E-14
Herpes simplex infection	0.069851	5.08E-14
TNF signaling pathway	0.069851	1.16E-13
Measles	0.069851	5.10E-13
NOD-like receptor signaling pathway	0.069851	1.18E-12
Osteoclast differentiation	0.069851	4.28E-11
Hepatitis C	0.069851	5.66E-11
Influenza A	0.069851	9.44E-09
Cytosolic DNA-sensing pathway	0.069851	3.42E-08
Neurotrophin signaling pathway	0.069851	9.69E-08
Hepatitis B	0.069851	1.97E-07
Chagas disease (American trypanosomiasis)	0.069851	1.16E-06
Legionellosis	0.069851	1.21E-06
Pertussis	0.069851	2.54E-06
Shigellosis	0.069851	3.73E-04
Prostate cancer	0.069851	9.06E-04

Table 11. The 20 pathways found to be related to the third active subnetwork that is composed of 239 genes and 239 edges.

It is also important to note which genes are located in the centers of these subnetworks. As the nature of the algorithm that is used in this thesis for the detection of active subnetworks, the MCL algorithm finds the subnetworks by clustering and while doing so gathers the genes that are in the cluster around a central node that is the attractor and thus is expected to be a significant gene for the condition, which in this case is RA. Following this logic, the central genes of the subnetworks are also investigated with the hope that they may be used as potential gene markers. The genes that are located in the center of the first subnetwork are EGFR and TJP1; in the center of the second subnetwork is ADCY8; and in the center of the third subnetwork are HSPA1L and MED10.

	Central Genes
Subnetwork 1	EGFR, TJP1
Subnetwork 2	ADCY8
Subnetwork 3	HSPA1L, MED10

Table 12. The genes that are found to be located in the centers of the best three subnetworks.

Significance Score	Subnetwork Size	Central Genes
0.00697	35	ERBB4, NRG1
0.013097	32	AGPAT1, PPAP2B
0.021011	28	CD247, PTPN22
0.029534	59	DSCAML1, MAGI3
0.029804	59	C3, CFB
0.033538	83	HLA-DMB, HLA-DRA
0.061454	14	CNDP2, NDRG1
0.062392	23	EFNA5, EPHA4
0.064051	62	NRXN1, SYT1
0.065805	109	HLA-DQA2, TLE1
0.065857	11	ATG10, ATG7
0.069395	18	AKAP9, KCNQ1
0.069851	239	HSPA1L, MED10
0.073886	58	BMP7, BMPR1B
0.074007	161	CALM1
0.074546	727	EGFR, TJP1
0.07594	37	RALGDS, RAP1A
0.08658	13	MDC1, MRE11A
0.087881	72	ADCY8
0.08836	88	MAML3, NOTCH4
0.094491	15	CD28, IL12A
0.096368	23	GGCX, PROS1
0.097339	58	GNAI1, OPRD1
0.103063	50	GLI1, SUFU
0.109866	267	COL4A3, FN1
0.119937	17	RAD51L1, RAD51L3

Table 13. The genes that are found to be located in the centers of the subnetworks, along with the cluster score and size of the subnetworks.

Chapter 7

CONCLUSION and FUTURE WORK

In this thesis, a clustering algorithm method is proposed in which, a novel edge weight calculation scheme to represent the interactions in the network, the Markov Cluster algorithm for the active subnetwork search, a scoring scheme to appoint scores for each found subnetwork and an elimination of subnetworks depending on those scores, is implemented, for the detection of active subnetworks in the human protein-protein interaction network. This method is applied on a real dataset (WTCCC-RA), followed by a functional enrichment step and the results are compared with the results of PANOGA [28] and methods proposed by Wu et al. [51], Martin et al. [52] and Zhang et al. [53]. The performed experiments demonstrate that the proposed method could successfully extract maximal scoring active subnetworks in human PPI networks and detect significant Rheumatoid Arthritis related subnetworks. The comparative study indicates that the presented technique outperforms the state-of-the-art active subnetwork search techniques. Therefore, it is proposed that this method can be applied upon the datasets of other complex diseases to discover active disease-associated subnetworks.

In the future studies, it is suggested to investigate the relation of the genes that are found to be located in the 3 maximal scoring subnetworks, to RA. Since the central genes of these subnetworks are especially significant due to the reason of acting as the attractor gene to all of the other genes in the cluster, a special interest should be paid to them. It is highly likely that they play an important role in the development of Rheumatoid Arthritis and thus they can be used as gene markers in the detection of the disease. Therefore, as a future experimental study, the mechanisms relating these genes to RA can be analyzed and it may be investigated if they can accurately discover the disease if

they were to be utilized as gene markers. Furthermore, they can also be studied to see if treatment strategies can be devised by targeting them through therapeutic approaches.

BIBLIOGRAPHY

1. Lichtenstein, I., Charleston, M. A., Caetano, T. S., Gamble, J. R., & Vadas, M. A. (2013). Active subnetwork recovery with a mechanism-dependent scoring function; with application to angiogenesis and organogenesis studies. *BMC Bioinformatics*, 14, 59.

2. Jiang, B., & Gribskov, M. (2014). Assessment of Subnetwork Detection Methods for Breast Cancer. *Cancer Informatics*, 13(Suppl 6), 15–23.

3. Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl 1), S233–40.

4. Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3, 140.

5. Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., & Sahinalp, S. C. (2011). Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13), i205–213.

6. Guo, Z., Wang, L., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D., Rao, S., & Wang, J. (2007). Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*, 23(16), 2121–2128.

7. Ma, H., Schadt, E., Kaplan, L., & Zhao, H. (2011). COSINE: COndition-SpecIfic sub-Network identification using a global optimization method. *Bioinformatics*, 27(9), 1290–1298.

8. Sohler, F., Hanisch, D., & Zimmer, R. (2004). New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10), 1517–1521.

9. Nacu, S., Critchley-Thorne, R., Lee, P., & Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics*, 23(7), 850–858.

10. Breitling, R., Amtmann, A., & Herzyk, P. (2004). Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics*, 5, 100.

11. Karni, S., Soreq, H., & Sharan, R. (2009). A network-based method for predicting disease-causing genes. *J. Comput. Biol.*, 16(2), 181–189.

12. Fortney, K., Kotlyar, M., & Jurisica, I. (2010). Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biol.*, 11(2), R13.

13. Su, J., Yoon, B. J., & Dougherty, E. R. (2010). Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics*, 11(Suppl 6), S8.

14. Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2011). dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1), 95–102.

15. Rajagopalan, D., & Agarwal, P. (2005). Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6), 788–793.

16. Qiu, Y., Zhang, S., Zhang, X., & Chen, L. (2010). Detecting disease associated modules and proritizing active genes based on high throuput data. *BMC Bioinformatics*, 11, 26.

17. Backes, C., Rurainski, A., Klau, G. W., Muller, O., Stockel, D., Gerasch, A., Kuntzer, J., Maisel, D., Ludwig, N., Hein, M., Keller, A., Burtscher, H., Kaufmann, M., Meese, E., & Lenhof, H. P. (2012). An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic Acids Res.*, 40(6), e43.

18. Klammer, M., Godl, K., Tebbe, A., & Schaab, C. (2010). Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC Bioinformatics*, 11, 351.

19. Wu, J., Gan, M., & Jiang, R. (2011). A genetic algorithm for optimizing subnetwork markers for the study of breast cancer metastasis. *In Natural Computation (ICNC), 2011 Seventh International Conference on*, volume 3, pages 1578–1582.

20. Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12), e1002822.

21. Norrgard, K. (2008) Genetic variation and disease: GWAS. *Nature Education*, 1(1), 87.

22. Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447(7145), 661-678.

23. Scott, D. L., Wolfe, F., & Huizinga, T. W. (2010). Rheumatoid arthritis. *Lancet*, 376(9746), 1094–108.

24. Linos, A., Worthington, J. W., O'Fallon, W. M., & Kurland, L. T. (1980). The epidemiology of rheumatoid arthritis in Rochester, Minnesota: A study of incidence, prevalence, and mortality. *Am J Epidemiol*, 111(1), 87-98.

25. Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabasi, A. L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.*, 104(21), 8685–8690.

26. Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062), 1173–1178.

27. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., & Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6), 957–968.

28. Bakir-Gungor, B., & Sezerman, O. U. (2011). A new methodology to associate SNPs with human diseases according to their pathway related context. *PLoS ONE*, 6(10), e26277.

29. Saccone, S. F., Bolze, R., Thomas, P., Quan, J. X., Mehta, G., Deelman, E., Tischfield, J. A., & Rice, J. P. (2010) SPOT: a web-based tool for using biological databases to prioritize SNPs after a genomewide association study. *Nucleic Acids Research,* 38(Web Server Issue), W201–W209.

30. Chelala, C., Khan, A., & Lemoine, N. R. (2009). SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25, 655–661.

31. Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), 1575–1584.

32. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H., Pages, F., Trajanoski, Z., & Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8), 1091–1093.

33. Rinaldi, N., Schwarz-Eywill, M., Weis, D., Leppelmann-Jansen, P., Lukoschek, M., Keilholz, U., & Barth, T. (1997). Increased expression of integrins on fibroblast-like synoviocytes from rheumatoid arthritis in vitro correlates with enhanced binding to extracellular matrix proteins. *Annals of the Rheumatic Diseases*, 56(1), 45–51.

34. Sifferlin, A. (2012, May 25). Rheumatoid Arthritis: Scientists Discover a Cell-Signaling Pathway that May Lead to New Treatment | TIME.com. Retrieved July 14, 2015.

35. Jiao, Z., Wang, W., Hua, S., Liu, M., Wang, H., Wang, X., Chen, Y., Xu, H., & Lu, L. (2014). Blockade of Notch signaling ameliorates murine collagen-induced arthritis via suppressing Th1 and Th17 cell responses. *Am J Pathol*, 184(4), 1085–93.

36. Kouri, V. P., Olkkonen, J., Kaivosoja, E., Ainola, M., Juhila, J., Hovatta, I., Konttinen, Y., & Mandelin, J. (2013). Circadian Timekeeping Is Disturbed in Rheumatoid Arthritis at Molecular Level. *PLoS ONE*, *8*(1), e54049.

37. Makarov, S. S. (2001). NF-kappaB in rheumatoid arthritis: a pivotal regulator of inflammation, hyperplasia, and tissue destruction. *Arthritis Research*, 3(4), 200–206.

38. Simmonds, R. E., & Foxwell, B. M. (2008). Signalling, inflammation and arthritis: NF-kappaB and its relevance to arthritis and inflammation. *Rheumatology (Oxford)*, 47(5), 584–90.

39. Jue, D. M., Jeon, K. I., & Jeong, J. Y. (1999). Nuclear factor kappaB (NF-kappaB) pathway as a therapeutic target in rheumatoid arthritis. *Journal of Korean Medical Science*, 14(3), 231–238.

40. Criswell, L. A. (2010). Gene discovery in rheumatoid arthritis highlights the CD40/NF-κB signaling pathway in disease pathogenesis. *Immunol Rev*, 233(1), 55–61.

41. Tian, J., Yong, J., Dang, H., & Kaufman, D. L. (2011). Oral GABA treatment downregulates inflammatory responses in a mouse model of rheumatoid arthritis. *Autoimmunity*, 44(6), 465–470.

42. Catalano, A. (2010). The neuroimmune semaphorin3A reduces inflammation and progression of experimental autoimmune arthritis. *J Immunol.*, 185(10), 6373–6383.

43. Pohlers, D., Brenmoehl, J., Löffler, I., Müller, C. K., Leipner, C., Schultze-Mosgau, S., Stallmach, A., Kinne, R. W., & Wolf, G. (2009). TGF-β and fibrosis in different organs-molecular pathway imprints. *Biochim Biophys Acta.*, 1792(8), 746–756.

44. Pohlers, D., Beyer, A., Koczan, D., Wilhelm, T., Thiesen, H.-J., & Kinne, R. W. (2007). Constitutive upregulation of the transforming growth factor-β pathway in rheumatoid arthritis synovial fibroblasts. *Arthritis Research & Therapy*, 9(3), R59.

45. Zhang, M. M., Jiang, Y. S., Lv, H. C., Mu, H. B., Li, J., Shang, Z. W., & Zhang, R. J. (2014). Pathway-based association analysis of two genome-wide screening data identifies rheumatoid arthritis-related pathways. *Genes Immun.*, 15(7), 487-94.

46. Shahrara, S., Castro-Rueda, H. P., Haines, G. K., & Koch, A. E. (2007). Differential expression of the FAK family kinases in rheumatoid arthritis and osteoarthritis synovial tissues. *Arthritis Res. Ther.*, 9(5), R112.

47. Banham-Hall, E., Clatworthy, M. R., & Okkenhaug, K. (2012). The Therapeutic Potential for PI3K Inhibitors in Autoimmune Rheumatic Diseases. *The Open Rheumatology Journal*, 6, 245–258.

48. Kling, A., Rantapää-Dahlqvist, S., Stenlund, H., & Mjörndal, T. (2006). Decreased density of serotonin 5-HT2$_A$ receptors in rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 65(6), 816–819.

49. Krishnadas, R., Krishnadas, R., & Cavanagh, J. (2011). Sustained remission of rheumatoid arthritis with a specific serotonin reuptake inhibitor antidepressant: a case report and review of the literature. *Journal of Medical Case Reports*, 5, 112.

50. Lee, P. H., & Shatkay, H. (2008). F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Research*, 36, D820–D824.

51. Wu, G., Zhu, L., Dent, J. E., & Nardini, C. (2010). A Comprehensive Molecular Interaction Map for Rheumatoid Arthritis. *PLoS ONE*, 5(4), e10137.

52. Martin, J. E., Alizadeh, B. Z., Gonzalez-Gay, M. A., Balsa, A., Pascual-Salcedo, D., Fernández-Gutiérrez, B., Raya, E., Franke, L., van't Slot, R., Coenen, M. J., van Riel, P., Radstake, T. R., Koeleman, B. P., & Martín, J. (2010). Identification of the Oxidative Stress-Related Gene MSRA as a Rheumatoid Arthritis Susceptibility Locus by Genome-Wide Pathway Analysis. *Arthritis and Rheumatism*, 62(11), 3183–3190.

53. Zhang, L., Li, W., Song, L., & Chen, L. (2010). A towards-multidimensional screening approach to predict candidate genes of rheumatoid arthritis based on SNP, structural and functional annotations. *BMC Medical Genomics*, 3, 38.

54. Remans, P. H., Gringhuis, S. I., van Laar, J. M., Sanders, M. E., Papendrecht-van der Voort, E. A., Zwartkruis, F. J., Levarht, E. W., Rosas, M., Coffer, P. J., Breedveld, F. C., Bos, J. L., Tak, P. P., Verweij, C. L., & Reedquist, K. A. (2004). Rap1 signaling is required for suppression of Ras-generated reactive oxygen species and protection against oxidative stress in T lymphocytes. *J Immunol.*, 173(2), 920–31.

55. Grossman, J. M., & Brahn, E. (1997). Rheumatoid arthritis: current clinical and research directions. *J Womens Health*, 6(6), 627–638.

56. Liu, G., Jiang, Y., Chen, X., Zhang, R., Ma, G., Feng, R., Zhang, L., Liao, M., Yingbo, M., Chen, Z., Zeng, R., & Li, K. (2013). Measles Contributes to Rheumatoid Arthritis: Evidence from Pathway and Network Analyses of Genome-Wide Association Studies. *PLoS ONE*, 8(10), e75951.

57. Ollier, W. (2000). Rheumatoid arthritis and Epstein-Barr virus: a case of living with the enemy? *Annals of the Rheumatic Diseases*, 59(7), 497–499.

58. Costenbader, K. H., & Karlson, & E. W. (2006). Epstein–Barr virus and rheumatoid arthritis: is there a link? *Arthritis Research & Therapy*, 8(1), 204.

59. Mullen, L. M., Chamberlain, G., & Sacre, S. (2015). Pattern recognition receptors as potential therapeutic targets in inflammatory rheumatic disease. *Arthritis Research & Therapy*, 17(1), 122.

KEGG Term	Significance Score	Term P-value
Nucleotide excision repair	0.228724	2.62E-60
DNA replication	0.228724	1.39E-44
Morphine addiction	0.217203	3.71E-39
Basal transcription factors	0.212178	1.99E-37
Notch signaling pathway	0.243467	2.32E-28
Glutamatergic synapse	0.217203	4.40E-26
Circadian entrainment	0.217203	9.06E-26
Retrograde endocannabinoid signaling	0.217203	3.14E-25
GABAergic synapse	0.217203	1.87E-24
Dopaminergic synapse	0.217203	4.22E-23
Cholinergic synapse	0.217203	1.57E-22
Serotonergic synapse	0.217203	1.87E-22
Mismatch repair	0.228724	6.13E-21
RNA polymerase	0.212178	2.11E-20
Pyrimidine metabolism	0.212178	2.04E-14
Cytosolic DNA-sensing pathway	0.212178	3.22E-14
Apoptosis	0.223952	3.26E-14
Cell cycle	0.228724	5.40E-14
Complement and coagulation cascades	0.020834	5.46E-13
ECM-receptor interaction	0.262625	3.30E-12

Table 14. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 2, inflation parameter 2 and threshold 0.28.

KEGG Term	Significance Score	Term P-value
DNA replication	0.21225	1.41E-46
Nucleotide excision repair	0.21225	1.20E-44
Basal transcription factors	0.208244	3.35E-31
Notch signaling pathway	0.243467	2.32E-28
GABAergic synapse	0.218292	3.67E-26
Morphine addiction	0.218292	7.01E-26
Circadian entrainment	0.218292	1.60E-25
Retrograde endocannabinoid signaling	0.218292	5.19E-25
Cholinergic synapse	0.218292	3.13E-24
Glutamatergic synapse	0.218292	5.19E-24
Mismatch repair	0.21225	7.23E-22
Serotonergic synapse	0.218292	2.58E-20
Dopaminergic synapse	0.218292	1.55E-17
Complement and coagulation cascades	0.020834	5.46E-13
Cell cycle	0.21225	2.38E-12
Staphylococcus aureus infection	0.020834	4.18E-11
Homologous recombination	0.21225	1.68E-10
mTOR signaling pathway	0.055623	9.95E-10
Dorso-ventral axis formation	0.011087	2.86E-09
Taste transduction	0.218292	2.26E-08

Table 15. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 2, inflation parameter 2.5 and threshold 0.28.

KEGG Term	Significance Score	Term P-value
DNA replication	0.253588	4.00E-47
Nucleotide excision repair	0.253588	2.49E-40
Basal transcription factors	0.205027	3.97E-29
Notch signaling pathway	0.243467	2.32E-28
GABAergic synapse	0.25613	4.22E-25
Morphine addiction	0.25613	8.06E-25
Circadian entrainment	0.25613	1.84E-24
Retrograde endocannabinoid signaling	0.25613	5.94E-24
Cholinergic synapse	0.25613	3.57E-23
Glutamatergic synapse	0.25613	5.91E-23
Mismatch repair	0.253588	4.15E-22
Serotonergic synapse	0.25613	3.11E-21
Chemokine signaling pathway	0.25613	1.24E-20
Dopaminergic synapse	0.25613	1.01E-16
Complement and coagulation cascades	0.020834	5.46E-13
Cell cycle	0.253588	1.34E-12
Staphylococcus aureus infection	0.020834	4.18E-11
Homologous recombination	0.253588	1.22E-10
Dorso-ventral axis formation	0.008507	9.55E-10
Base excision repair	0.253588	2.46E-08

Table 16. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 2, inflation parameter 3 and threshold 0.28.

KEGG Term	Significance Score	Term P-value
DNA replication	0.275742	5.32E-48
Nucleotide excision repair	0.275742	9.73E-39
Notch signaling pathway	0.218395	1.66E-29
GABAergic synapse	0.262436	8.61E-25
Morphine addiction	0.262436	1.64E-24
Circadian entrainment	0.262436	3.75E-24
Retrograde endocannabinoid signaling	0.262436	1.21E-23
Cholinergic synapse	0.262436	7.25E-23
Glutamatergic synapse	0.262436	1.20E-22
Mismatch repair	0.275742	1.73E-22
Serotonergic synapse	0.262436	5.96E-21
Chemokine signaling pathway	0.262436	2.64E-20
Dopaminergic synapse	0.262436	1.74E-16
Complement and coagulation cascades	0.020834	5.46E-13
Cell cycle	0.275742	1.26E-11
Staphylococcus aureus infection	0.020834	4.18E-11
Dorso-ventral axis formation	0.011932	9.55E-10
Homologous recombination	0.275742	4.54E-09
Base excision repair	0.275742	1.60E-08
Taste transduction	0.262436	6.28E-08

Table 17. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 2, inflation parameter 3.5 and threshold 0.28.

KEGG Term	Significance Score	Term P-value
DNA replication	0.294674	2.24E-48
Nucleotide excision repair	0.294674	4.44E-39
Basal transcription factors	0.345756	1.09E-29
Notch signaling pathway	0.218395	1.66E-29
GABAergic synapse	0.346668	3.72E-23
Morphine addiction	0.346668	6.82E-23
Mismatch repair	0.294674	1.09E-22
Circadian entrainment	0.346668	1.48E-22
Retrograde endocannabinoid signaling	0.346668	4.44E-22
Cholinergic synapse	0.346668	2.39E-21
Glutamatergic synapse	0.346668	3.83E-21
Serotonergic synapse	0.346668	1.81E-19
Dopaminergic synapse	0.346668	3.92E-15
Complement and coagulation cascades	0.020834	5.46E-13
Staphylococcus aureus infection	0.020834	4.18E-11
Dorso-ventral axis formation	0.011932	9.55E-10
Homologous recombination	0.294674	3.35E-09
Base excision repair	0.294674	1.18E-08
Taste transduction	0.346668	4.55E-08
Fanconi anemia pathway	0.294674	3.82E-07

Table 18. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 2, inflation parameter 4 and threshold 0.35.

KEGG Term	Significance Score	Term P-value
ECM-receptor interaction	0.091926	2.54E-51
Focal adhesion	0.091926	1.26E-37
Morphine addiction	0.080974	2.47E-34
NF-kappa B signaling pathway	0.081024	1.30E-32
Notch signaling pathway	0.11082	1.36E-31
Apoptosis	0.081024	1.57E-30
Axon guidance	0.070037	3.57E-30
Complement and coagulation cascades	0.091926	4.43E-29
Jak-STAT signaling pathway	0.067762	1.17E-25
Leukocyte transendothelial migration	0.067762	1.24E-24
Circadian entrainment	0.080974	4.70E-23
Natural killer cell mediated cytotoxicity	0.067762	1.11E-22
Cholinergic synapse	0.080974	1.61E-22
PI3K-Akt signaling pathway	0.091926	2.62E-22
Toll-like receptor signaling pathway	0.081024	8.28E-22
TGF-beta signaling pathway	0.10762	5.10E-21
RIG-I-like receptor signaling pathway	0.081024	3.51E-20
TNF signaling pathway	0.081024	4.60E-20
Measles	0.081024	5.88E-20
GABAergic synapse	0.080974	9.79E-20

Table 19. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 3, inflation parameter 2 and threshold 0.12.

KEGG Term	Significance Score	Term P-value
ECM-receptor interaction	0.116181	4.51E-50
Morphine addiction	0.08146	6.86E-47
Focal adhesion	0.116181	1.18E-35
Notch signaling pathway	0.086308	1.42E-33
Circadian entrainment	0.08146	4.86E-30
GABAergic synapse	0.08146	4.94E-29
Glutamatergic synapse	0.08146	5.68E-28
NF-kappa B signaling pathway	0.079797	1.16E-27
Axon guidance	0.070037	1.31E-27
Retrograde endocannabinoid signaling	0.08146	1.58E-27
Cholinergic synapse	0.08146	1.65E-26
Jak-STAT signaling pathway	0.07494	2.39E-25
Serotonergic synapse	0.08146	1.08E-24
Dopaminergic synapse	0.08146	2.86E-23
Leukocyte transendothelial migration	0.07494	3.86E-23
Complement and coagulation cascades	0.086646	1.41E-21
TGF-beta signaling pathway	0.063987	2.27E-21
Toll-like receptor signaling pathway	0.079797	4.78E-20
PI3K-Akt signaling pathway	0.116181	5.37E-20
Natural killer cell mediated cytotoxicity	0.07494	1.75E-19

Table 20. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 3, inflation parameter 2.5 and threshold 0.12.

KEGG Term	Significance Score	Term P-value
ECM-receptor interaction	0.109866	2.16E-45
Morphine addiction	0.087881	4.46E-45
Notch signaling pathway	0.08836	2.59E-33
Focal adhesion	0.109866	6.32E-32
Circadian entrainment	0.087881	2.73E-30
NF-kappa B signaling pathway	0.069851	2.09E-29
GABAergic synapse	0.087881	2.86E-29
Glutamatergic synapse	0.087881	3.19E-28
Retrograde endocannabinoid signaling	0.087881	9.20E-28
Cholinergic synapse	0.087881	9.57E-27
Axon guidance	0.062392	6.85E-26
Jak-STAT signaling pathway	0.074546	5.24E-25
Serotonergic synapse	0.087881	6.46E-25
Leukocyte transendothelial migration	0.074546	1.07E-24
MAPK signaling pathway	0.033538	1.49E-23
Dopaminergic synapse	0.087881	1.72E-23
TGF-beta signaling pathway	0.073886	2.45E-23
Complement and coagulation cascades	0.096368	4.76E-21
Toll-like receptor signaling pathway	0.069851	6.33E-19
PI3K-Akt signaling pathway	0.109866	3.11E-17

Table 21. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 3, inflation parameter 3 and threshold 0.12.

KEGG Term	Significance Score	Term P-value
ECM-receptor interaction	0.100874	3.85E-45
Morphine addiction	0.083949	1.09E-44
Notch signaling pathway	0.089066	2.43E-32
Focal adhesion	0.100874	1.16E-31
Circadian entrainment	0.083949	1.05E-29
GABAergic synapse	0.083949	1.27E-28
Glutamatergic synapse	0.083949	9.86E-28
Retrograde endocannabinoid signaling	0.083949	3.45E-27
Leukocyte transendothelial migration	0.070723	2.91E-26
Cholinergic synapse	0.083949	3.21E-26
Axon guidance	0.062392	6.85E-26
Jak-STAT signaling pathway	0.070723	3.77E-25
MAPK signaling pathway	0.030788	1.13E-24
Serotonergic synapse	0.083949	2.25E-24
TGF-beta signaling pathway	0.087756	2.45E-23
Dopaminergic synapse	0.083949	5.10E-23
Complement and coagulation cascades	0.119912	3.96E-19
NF-kappa B signaling pathway	0.057181	2.26E-18
PI3K-Akt signaling pathway	0.100874	5.04E-17
Regulation of autophagy	0.050172	8.75E-16

Table 22. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 3, inflation parameter 3.5 and threshold 0.12.

KEGG Term	Significance Score	Term P-value
ECM-receptor interaction	0.098358	2.16E-45
Morphine addiction	0.085889	6.56E-43
Notch signaling pathway	0.095922	2.59E-33
Focal adhesion	0.098358	6.32E-32
Circadian entrainment	0.085889	4.12E-28
Leukocyte transendothelial migration	0.069522	2.89E-27
GABAergic synapse	0.085889	5.20E-27
Glutamatergic synapse	0.085889	3.09E-26
MAPK signaling pathway	0.033703	4.16E-26
Jak-STAT signaling pathway	0.069522	4.94E-26
Axon guidance	0.062392	6.85E-26
Retrograde endocannabinoid signaling	0.085889	1.19E-25
Cholinergic synapse	0.085889	9.85E-25
TGF-beta signaling pathway	0.081018	4.59E-23
Serotonergic synapse	0.085889	6.45E-23
Dopaminergic synapse	0.085889	1.24E-21
NF-kappa B signaling pathway	0.060653	9.12E-20
Complement and coagulation cascades	0.031087	3.34E-19
PI3K-Akt signaling pathway	0.098358	3.11E-17
Regulation of autophagy	0.050172	8.75E-16

Table 23. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 3, inflation parameter 4 and threshold 0.12.

KEGG Term	Significance Score	Term P-value
Morphine addiction	0.121026	2.76E-46
Complement and coagulation cascades	0.076994	9.38E-39
Jak-STAT signaling pathway	0.094846	7.49E-35
Circadian entrainment	0.121026	4.53E-33
ErbB signaling pathway	0.094846	6.70E-31
GABAergic synapse	0.121026	4.68E-30
Glutamatergic synapse	0.121026	4.75E-29
Retrograde endocannabinoid signaling	0.121026	1.51E-28
ECM-receptor interaction	0.10869	1.36E-27
Cholinergic synapse	0.121026	1.58E-27
Transcriptional misregulation in cancer	0.084129	6.65E-25
Valine, leucine and isoleucine degradation	0.062028	3.17E-24
Serotonergic synapse	0.121026	6.48E-24
Chemokine signaling pathway	0.121026	4.44E-22
T cell receptor signaling pathway	0.094846	3.11E-21
Dopaminergic synapse	0.121026	6.20E-21
Natural killer cell mediated cytotoxicity	0.094846	1.14E-18
B cell receptor signaling pathway	0.094846	1.75E-18
Fc gamma R-mediated phagocytosis	0.094846	3.57E-18
Fc epsilon RI signaling pathway	0.094846	9.44E-18

Table 24. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 4, inflation parameter 2 and threshold 0.16.

KEGG Term	Significance Score	Term P-value
Morphine addiction	0.152546	1.43E-44
Circadian entrainment	0.152546	1.63E-31
GABAergic synapse	0.152546	1.66E-28
Glutamatergic synapse	0.152546	1.29E-27
Retrograde endocannabinoid signaling	0.152546	4.52E-27
Complement and coagulation cascades	0.086029	1.40E-26
Cholinergic synapse	0.152546	4.20E-26
ECM-receptor interaction	0.066727	6.54E-24
B cell receptor signaling pathway	0.13378	7.82E-24
Jak-STAT signaling pathway	0.108349	1.24E-23
Serotonergic synapse	0.152546	1.52E-22
Dopaminergic synapse	0.152546	2.92E-21
Fc gamma R-mediated phagocytosis	0.13378	3.39E-21
Cell cycle	0.157265	4.99E-21
Chemokine signaling pathway	0.152546	6.43E-21
Valine, leucine and isoleucine degradation	0.077613	3.61E-19
Apoptosis	0.123193	4.34E-18
Fc epsilon RI signaling pathway	0.13378	1.47E-16
Natural killer cell mediated cytotoxicity	0.13378	7.01E-16
T cell receptor signaling pathway	0.13378	8.91E-16

Table 25. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 4, inflation parameter 2.5 and threshold 0.16.

KEGG Term	Significance Score	Term P-value
Morphine addiction	0.153488	9.29E-43
Circadian entrainment	0.153488	8.41E-32
GABAergic synapse	0.153488	9.26E-29
Glutamatergic synapse	0.153488	6.93E-28
Retrograde endocannabinoid signaling	0.153488	2.52E-27
Cholinergic synapse	0.153488	2.34E-26
Serotonergic synapse	0.153488	9.14E-23
ECM-receptor interaction	0.143751	1.05E-22
Dopaminergic synapse	0.153488	1.75E-21
B cell receptor signaling pathway	0.144153	3.28E-21
Chemokine signaling pathway	0.153488	3.63E-21
Valine, leucine and isoleucine degradation	0.077613	3.61E-19
Fc gamma R-mediated phagocytosis	0.144153	7.60E-19
Jak-STAT signaling pathway	0.111467	7.04E-18
ErbB signaling pathway	0.144153	7.25E-18
Apoptosis	0.099167	1.04E-16
Osteoclast differentiation	0.144153	1.35E-16
Natural killer cell mediated cytotoxicity	0.144153	2.24E-16
Alcoholism	0.153488	9.52E-16
Fc epsilon RI signaling pathway	0.144153	1.66E-15

Table 26. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 4, inflation parameter 3 and threshold 0.16.

KEGG Term	Significance Score	Term P-value
Morphine addiction	0.149316	9.29E-43
Circadian entrainment	0.149316	8.41E-32
GABAergic synapse	0.149316	9.26E-29
Glutamatergic synapse	0.149316	6.93E-28
Retrograde endocannabinoid signaling	0.149316	2.52E-27
Cholinergic synapse	0.149316	2.34E-26
ECM-receptor interaction	0.190868	5.84E-23
Serotonergic synapse	0.149316	9.14E-23
Dopaminergic synapse	0.149316	1.75E-21
Chemokine signaling pathway	0.149316	3.63E-21
Valine, leucine and isoleucine degradation	0.077613	3.61E-19
Natural killer cell mediated cytotoxicity	0.186387	3.96E-17
Apoptosis	0.113795	1.79E-16
Fc gamma R-mediated phagocytosis	0.186387	1.88E-16
Alcoholism	0.149316	9.52E-16
B cell receptor signaling pathway	0.186387	2.05E-15
Complement and coagulation cascades	0.026537	6.19E-15
Osteoclast differentiation	0.186387	9.36E-15
Fc epsilon RI signaling pathway	0.186387	3.68E-14
TGF-beta signaling pathway	0.145419	3.88E-14

Table 27. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 4, inflation parameter 3.5 and threshold 0.20.

KEGG Term	Significance Score	Term P-value
Morphine addiction	0.132583	2.26E-41
Basal transcription factors	0.138179	1.43E-35
Circadian entrainment	0.132583	1.62E-28
GABAergic synapse	0.132583	2.20E-27
Glutamatergic synapse	0.132583	1.22E-26
Retrograde endocannabinoid signaling	0.132583	5.05E-26
Cholinergic synapse	0.132583	4.19E-25
ECM-receptor interaction	0.190868	5.84E-23
Serotonergic synapse	0.132583	1.55E-21
Dopaminergic synapse	0.132583	2.53E-20
Chemokine signaling pathway	0.132583	3.60E-20
Valine, leucine and isoleucine degradation	0.077613	3.61E-19
Cell cycle	0.187381	2.32E-17
RNA polymerase	0.138179	1.61E-16
Apoptosis	0.108758	1.79E-16
B cell receptor signaling pathway	0.218297	2.61E-15
Natural killer cell mediated cytotoxicity	0.218297	6.16E-15
Complement and coagulation cascades	0.026537	6.19E-15
Fc epsilon RI signaling pathway	0.218297	5.55E-14
Staphylococcus aureus infection	0.026537	4.45E-13

Table 28. The 20 most significant pathways, determined by their term p-values, found to be related to the subnetworks that are detected by the active subnetwork search, using expansion parameter 4, inflation parameter 4 and threshold 0.24.