A MULTI-LAYERED GRAPHICAL MODEL OF THE RELATION AMONG SNPS, GENES, AND PATHWAYS BASED ON SUBGRAPH SEARCH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

GÖKHAN ERSOY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

JUNE 2015

A MULTI-LAYERED GRAPHICAL MODEL OF THE RELATION BETWEEN SNPS, GENES, AND PATHWAYS BASED ON SUBGRAPH SEARCH

Submitted by **Gökhan Ersoy** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute**

Assoc. Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Health Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son _____
Supervisor, **Health Informatics**

Assoc. Prof. Dr. Tolga Can _____
Co-Supervisor, **Computer Engineering**

Examining Committee Members:

Assist. Prof. Dr. Aybar Can Acar _____
Health Informatics, METU

Assoc. Prof. Dr. Yeşim Aydın Son _____
Health Informatics, METU

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

Assist. Prof. Dr. Öznur Taştan _____
Computer Sciences,
İhsan Doğramacı Bilkent University

Dr. Nurcan Tunçbağ _____
Health Informatics, METU

Date: 17 June 2015

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

Name, Last Name : Gökhan Ersoy

Signature :

ABSTRACT

A MULTI-LAYERED GRAPHICAL MODEL OF THE RELATION AMONG SNPS, GENES, AND PATHWAYS BASED ON SUBGRAPH SEARCH

Ersoy, Gökhan

MSc., Bioinformatics Program

Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

Co-Supervisor: Assoc. Prof. Dr. Tolga Can

June 2015, 97 Pages

The analysis of Single Nucleotide Polymorphisms (SNPs) through Genome Wide Association Studies (GWAS) presents great potential for describing disease loci and gaining insight into the underlying etiology of diseases. Recently described combined p-value approach allows identification of associations at gene and pathway level. The integrated programs like METU-SNP produce simple lists of either SNP id/gene id/pathway title and their p-values and significance status or SNP id/disease id/pathway information. In this study, starting with the SNP id, we have annotated related gene ids and pathway ids consecutively. Then we have computed the intersection of these pathways, and visualized the common sub-graphs by an interactive graphical library. The tool developed in this thesis provides a visualization of the text output as graphical knowledge networks; hence, facilitates the efficient use of the information offered by the candidate SNP Biomarkers and helping discovery of SNP associated biological networks.

Keywords: gene, intersection, pathway, SNP, visualization

ÖZ

SNP, GEN VE YOLAKLAR ARASINDAKİ İLİŞKİNİN ORTAKLIKLARINA GÖRE ÇOK KATMANLI BİR GRAFİK İLE MODELLENMESİ

Ersoy, Gökhan

Yüksek Lisans, Biyoenformatik Programı

Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son

Ortak Tez Yöneticisi: Doç. Dr. Tolga Can

Haziran 2015, 97 Sayfa

Genom Boyutunda Bağlantı/İşbirliği Çalışmaları (GWAS) kapmasında yapılan "Tek Nükleotit Polimorfizm (SNP) analizi", hastalıklara özgü gen bölgelerinin tanımlanması ve hastalıkların altında yatan nedenleri kavrama açısından büyük potansiyel taşımaktadır. Son zamanlarda tanımlanmış olan "birleşik p-değeri (p-value)" yaklaşımı gen ve yolak (pathway) seviyesindeki "bağlantıları" tanımlamaya imkân sağlar. METU-SNP gibi mevcut bilgisayar programları ya "SNP numarası/gen numarası/yolak ismi, p-değeri" ya da "SNP numarası/hastalık numarası/yolak bilgisi" gibi bilgileri yazılı bir liste olarak vermektedir. Bu çalışmamızda, SNP numarasından/belirtecinden yola çıkarak sırasıyla o SNP'in yer aldığı genler ve o genlerin yer aldığı yolaklar mevcut webservisler aracılığıyla bulunmaktadır. Daha sonra bu yolaklar arası ortaklıklar bir "grafik kesişim" algoritmasıyla bulunmakta ve bir interaktif grafik kütüphanesi kullanılarak görselleştirilmektedir. Yazılı çıktının grafiksel bir bilgi ağına dönüştürülmesi; tanımlanmış muhtemel SNP biobelirteçleri (biomarker) için toplanan bilginin daha etkin kullanımına imkân sağlamakta ve SNPler ile ilişkili biyolojik ağ keşiflerine güç katmaktadır.

Anahtar Kelimeler: gen, kesişim, yolak, SNP, görselleştirme

To my family,

To myself…

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dear Assoc. Prof. Dr. Yeşim Aydın Son. I am grateful to her for an original thesis topic and her continuous support, motivation and immense knowledge. Her continuous guidance helped me in my research and writing of this thesis.

I am very glad to have Assoc. Prof. Dr. Tolga Can as my co-supervisor. He shared brilliant ideas with me and helped me through generating algorithms as part of my thesis.

I am grateful to all my teachers for their knowledgeable courses helped me to learn what I know to complete this thesis.

Examining committee members Assist. Prof. Dr. Aybar Can Acar, Dr. Nurcan Tunçbağ and Assist. Prof. Dr. Öznur Taştan are greatly acknowledged for their participation and valuable comments.

I am thankful to my former boss Temel Akgün, R&D Coordinator Ahmet Özçam and carrier manager Merve Gürkan, for giving me the chance to complete my master education.

I am thankful to my colleagues in my former company, Hamid Rustamov and Burak Karaş, for helping me to learn new technologies.

I am thankful to my managers in Social Sciences University of Ankara, Adnan Yılmaz and Abdülkadir Karabıyık, and to my colleagues for supporting me to complete my master education.

This study is dedicated to my family, my father Hasan Ersoy, my mother Zahide Ersoy, my brothers and their wives, and to my only niece Elif Vildan Ersoy. I would like to thank to them for supporting me spiritually throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API: Application Programming Interface
DB: Database
ENSG: Abbreviation for Ensembl Gene ID
GWT: Google Web Tool Kit
GWT4NB: GWT for NetBeans
GXT: Ext GWT
HGNC: HUGO Gene Nomenclature Committee
Hsa: Homo sapiens
HTML: Hyper Text Mark-up Language
HUGO: Human Genome Organization
IDE: Integrated Development Environment
Jar: Java Archive
JDK: Java Development Kit
JPA: Java Persistence API
JSON: JavaScript Object Notation
KEGG: Kyoto Encyclopedia of Genes and Genomes
KGML: KEGG Markup Language
REST: Representational State Transfer
Rs ID: Reference SNP cluster ID
SNP: Single Nucleotide Polymorphism
URL: Uniform Resource Locator
VEP: Variant Effect Predictor
XML: Extended Markup Language

CHAPTER

1. INTRODUCTION

1.1 Motivation

Single nucleotide polymorphisms, frequently called SNPs (pronounced “snips”), have high potential to identify associated loci and genes and possible molecular mechanisms of diseases and for development of new diagnostics. For better understanding the SNP effect on molecular function, the biological pathways where SNPs and their associated genes map should be investigated. For a specific list of SNPs, if we compare the common pathways and find the intersecting sub-graphs, we can better evaluate SNP effect on biological processes. To our knowledge, there isn’t any tool analyzing the SNP-gene-pathway relation in a pipeline fashion. Additionally, an integrating tool to find and visualize the common pathways of a set of associated SNPs is also required for calculation and easy interpretation of the intersections between given pathways.

1.2 Goal

This thesis study is focused on developing a new approach for the analysis of SNP-gene-pathway relations based on three very essential objectives.

The first aim of the study was finding the pathways of the genes that a given list of SNPs map. We find the union of the sets of related genes using the RS ids of some specific SNPs. Then, we convert ENSEMBL gene IDs (ENSG) to KEGG gene IDs to find the related pathways in the KEGG database. The second aim was to develop an algorithm to find the common sub-graphs of pathways. The last objective was visualizing the common sub-graphs with an interactive graphical output.

1.3 Contribution

In this thesis study, we have created a pipeline starting with a list of SNP RS ids, and finding genes and pathways consecutively, and finally calculating and visualizing the common sub-graphs of pathways. In addition, we have created a software application to serve the pipeline. The application works without any problem in three different

case studies. The pathway intersection algorithm and the pipeline -from SNPs to pathways- work as expected, for all of the three case studies.

1.4 Thesis organization

This thesis is organized in 6 chapters. Chapter I provides short introduction to the overall study. In Chapter II, background information and related works are summarized. Chapter III presents the data and materials utilized in the study and methods used during the development of the software application. Also the implementation of the software and the algorithm are explained in detail. The results from three case studies which have been run on the software application is presented as the validation of the algorithms developed within this thesis under results section in the Chapter IV. First study showed the ability of the program to compute the intersections of pathways correctly. Additionally in the first case study, we have presented the available options of the presented software where the user can start at any of the four steps. In the second case study, we run the program with a set of SNPs related with Genomic Prostate Cancer Model (Yücebaş & Aydın Son, 2014). And in the third case, network analysis with Juvenile Rheumatoid Arthritis (JRA) associated SNP set (Aydın Son *et al.*, 2015) is performed. Overall accomplishments, shortcomings and future plans of the study are summarized in Chapter VI.

2. BACKGROUND AND RELATED WORKS

2.1 Background

This study is based on the investigation of biological diseases caused by genomic variations called SNPs.

2.1.1 Human Genome

Human genome is a term for whole set of genetic information for *Homo sapiens*. This genetic information is coded as DNA sequences on chromosome pairs. The chromosome pairs are located in cell nuclei. Each cell nuclei has 23 chromosomes. 22 chromosome pairs contain the autosomal information, while the 23rd chromosome pair determines the sex (allosome). (Alberts, 2008)

DNA (Deoxyribonucleic acid) is composed of a sequence of nucleotides. A deoxyribose sugar, a phosphate group and a nucleobase builds a nucleotide. Adenine, Guanine, Cytosine and Thymine are nucleobases in the structure of the DNA. According to biochemical structure, they are divided into 2 groups. Adenine (A) and Guanine (G) are purines, Cytosine (C) and Thymine (T) are prymidines. The nucleotides in the DNA structure lay on two biopolymers and these two polymers are folded onto each other as a double stranded helix (Watson & Crick, 1953). Ribose sugar and phosphate group forms the backbone of the DNA. Nuclebases of each strand bind to each other via hydrogen bonds, and they are matched as one prymidine to one purine. Adenine binds to Thymine with 2 hydrogen bonds, while Guanine to Cytosine with 3 hydrogen bonds. (Alberts, 2008)

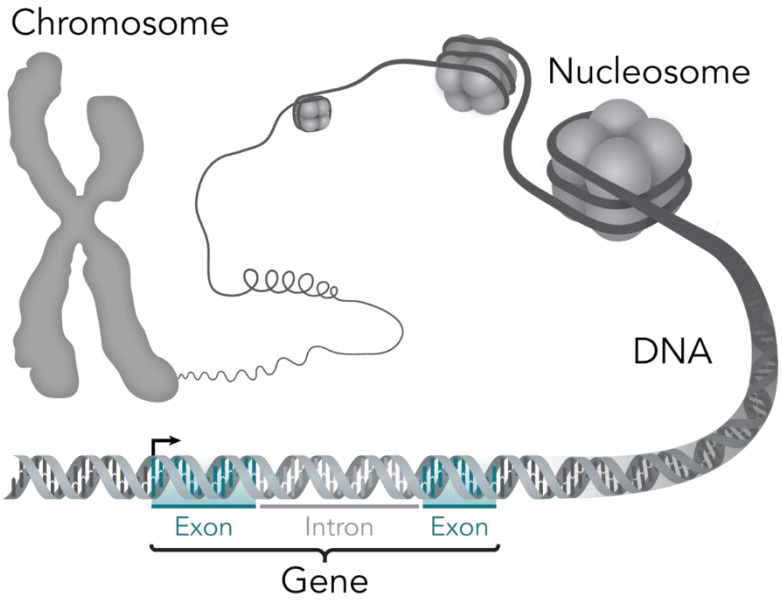

Figure 1. Human genome structure (Image is licensed under the Creative Commons Attribution-Share Alike 4.0 International license, deposited in Wikimedia Commons, submitted by Thomas Shafee)

2.1.1.1 Gene Structure

A gene is the molecular unit of heredity of a living organism. Specifically, genes are the particular regions of DNA which can encode a functional product. These functional products might be a subunit or a precursor of a protein, or an RNA molecule with important biological functions. The latest researches state that a haploid human genome contains 3 billion DNA base pairs, and 2% of which encode approximately 25,000 protein coding genes. (Pevsner, 2009) The regions of DNA that encodes proteins are called exons. The rest of the human genome (98%) consists of introns, non-coding RNAs, regulatory DNA sequences and repetitive DNA elements.

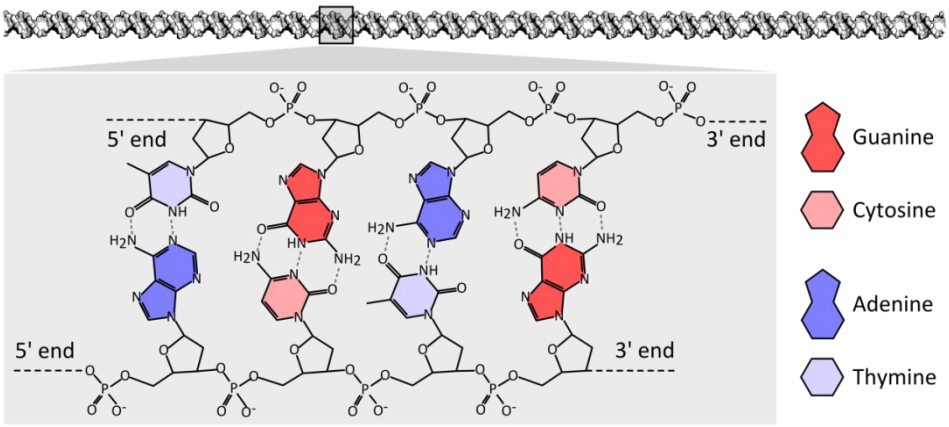

Figure 2. DNA chemical structure (Image is licensed under the Creative Commons Attribution-Share Alike 4.0 International license, deposited in Wikimedia Commons, submitted by Thomas Shafee)

2.1.2 Genomic Variations

Human genetic variation is the genetic differences both within and among populations. Genome variations are differences in the sequence of DNA from one person to another. There may be multiple variants of any given gene in the human population (genes), leading to polymorphism. Gene variants are called as alleles. Most of genes are fixed. It means that only single allele is present in the population for that gene. They are not polymorphic. On average, there is 99.5% similarity among all humans. All humans are genetically different with only 0.5% genetic variance. Even monozygotic twins have differences due to the mutations occurring during development and gene copy-number variation. Alleles occur at different frequencies in different human populations. (Alberts, 2008)

Causes of differences between individuals include the exchange of genes during meiosis and various mutational events. Genetic variation among humans occurs on many scales, from gross alterations in the human karyotype to single nucleotide changes.

Nucleotide diversity is the average proportion of nucleotides that differ between two individuals. The human nucleotide diversity is estimated to be 0.1% to 0.4% of base pairs. Approximately, there is a difference in every 1000 nucleotides. So there are 3 million nucleotide differences, because the human genome has about 3 billion nucleotides. (Alberts, 2008)

Measures of variation can be listed as

- Single Nucleotide Polymorphisms (SNPs)
- Structural Variation
- Epigenetics
- Genetic Variability
- Clines

- Haplogroups
- Variable Number Tandem Repeats

In this thesis, we focus on SNPs.

2.1.2.1 SNP

A single nucleotide polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals. Different alleles of nucleotides (A, G, C and T) cause polymorphism by a single nucleotide. If at least 1% of a population has the least frequent allele of a point mutation, then the variety among individuals of population is classified as a SNP. 90% of all genetic variations in human genome occur as SNPs.

SNPs might be bi-allelic, tri-allelic or tetra-allelic. If 2 different nucleotides might occur in the same locus in DNA sequence among individuals, the SNP is called as bi-allelic. Tri-allelic is for 3 varying nucleotides, and tetra-allelic is for 4 varying nucleotides.

SNPs might occur in two ways: *transitions* and *transversions*. Transition occurs while exchanging between purines (A and G) or between prymidines (C and T). Transversion occurs when prymidines exchange with purines.

2.1.2.2 Genetic Diseases

A SNP might be found in both coding and non-coding regions of DNA sequences as well as intergenic regions like promoter region, mRNA binding sites and splicing sites. The location of a SNP in a DNA sequence might have various effects such as phenotype differences, biological diseases, etc.

SNPs on regulatory regions cause changes in transcription and mRNA stability levels, and microRNA affectivity. (Zienolddiny & Skaug, 2012) SNPs can strongly effect the phenotype. (Wu & Jiang, 2013)

According to their effect on gene products, coding region SNPs are classified as two subtypes: Synonymous and non-synonymous SNPs.

Synonymous SNPs does not change the amino acid sequence. So the gene product doesn't change.

Non-synonymous SNPs changes the amino acid sequence which may lead a protein having different tertiary structure or no production. Any change in its tertiary structure effects the function of a protein. (Zienolddiny & Skaug, 2012) The non-synonymous SNPs are also divided into two subtypes: missense SNPs and nonsense SNPs.

Nonsense SNPs change gene sequence resulting to an early stop codon. The codons after the early stop codon cannot be read and translated to a protein. Early stop codon

causes incomplete and nonfunctional polypeptide sequences. The beta zero thalassemia disease (Chang & Kan, 1979) is caused by a nonsense SNP.

Missense SNPs change gene sequences causing a different amino acid codon. During the transcription, the new codon causes translation of a different protein product. The sickle cell anemia disease (Wang & Moult, 2001) is caused by a missense SNP.

In the course of this thesis, we will investigate the associated SNPs of Prostate Cancer and Juvenile Rheumatoid Arthritis (JRA) diseases.

2.2 Related Works

There are few existing bioinformatics tools that share common features with the proposed application. Generally, these tools perform a single function that have been integrated as a component in our application. Main tools that are available for the analysis of SNP-gene, gene-pathway relations can be listed as KEGG & KEGG PATHWAY, DAVID, SNPnexus, KEGGgraph, VisANT, KMGL-ED, PANOGA, PathVisio, Cytoscape and its plugins have been reviewed here.

KEGG database provides relations between high-level functions and utilities of the biological systems. (Ogata *et al.*, 1999) The application uses large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

A biological pathway is defined as *a series of actions among molecules in a cell that leads to a certain product or a change in a cell.*[1]

KEGG PATHWAY is officially defined as the *"collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for: 1- Metabolism 2-Genetic Information Processing 3-Environmental Information Processing 4-Cellular Processes 5-Organismal Systems 6-Human Diseases 7- Drug Development"*[2]. KEGG provides this information through a brief representation of pathway maps. KEGG only supports conversion from ncbi-gi, ncbi-geneid or uniprot ID to KEGG gene id, and matching pathways of genes can be searched only through KEGG gene ids.

These maps are manually drawn and not interactive. User can only make search of components on these maps. User does not have the ability of zooming, dragging, grouping, etc. While KEGG has REST service library for customized queries such as conversion between gene ids, finding pathways of a gene, etc. , it does not provide a graphical user interface (GUI) for easy usage of these REST services. Users should manually write and call the REST services via browsers or in application codes.

[1] http://en.wikipedia.org/wiki/Biological_pathway
[2] http://www.genome.jp/kegg/pathway.html

KEGGgraph is an R library to convert KGML files to analyzable graphs. (Zhang & Wiemann, 2009) The application can read and parse KGML files and visualize them with some limited analysis options on the graphs.

DAVID (the Database for Annotation, Visualization and Integrated Discovery) is a web based application that provides a set of functional annotation tools to understand biological significance of a submitted list of genes. (Dennis *et al.,* 2003) For any given gene list, DAVID tools provide many tasks, and among them following 3 tasks are also used in our study: 1- conversion between different types of gene identifiers 2-finding matching pathways of given genes 3-visualize genes on KEGG pathway maps. However, DAVID cannot get the intersection of pathways and cannot draw interactive graphs. The genes are mapped on KEGG generated pathways.

SNPnexus is a web server for functional annotation of novel and publicly known genetic variants. (Dayem Ullah *et al.,* 2012) SNPnexus can map given SNPs and other variants to find their genomic location and matching genes, for up-to 100,000 entries. SNPnexus accepts 3 types of input. The first is a single dbSNP id for single query. For multiple queries it accepts tab-delimited file of variants and a universal file format called VCF (Variant Call Format).

PANOGA is a web server for identification of SNP-targeted pathways from genome-wide association study data. (Bakir-Gungor *et al.*, 2014) PANOGA is developed to devise functionally important pathways through the identification of SNP-targeted genes within these pathways. The application has a multidimensional perspective like our application. It combines evidence from the following five resources:

- genetic association information obtained through GWAS,
- SNP functional information,
- protein-protein interaction network,
- linkage disequilibrium,
- biochemical pathways.

While PANOGA identifies SNP-targeted pathways similar to our application, it is quite slow; typically a run takes 7-8 hours and the user has to check the query result using the unique link, manually. PANOGA also provides outputs in following formats; pathway tables, gene list and pathway maps.

VisANT is a 3-tier enterprise system similar to the application developed in this study. (Hu *et al.*, 2008) VisANT is specialized for integrative visual data-mining of multi-scale biological networks. The VisANT platform has a local database and additionally supported by the Predictome database. VisANT is a *convenient and fast network/pathway construction tool using either update-to-date knowledge or user's data. It provides multi-scale visualization of bio-networks with functional modules, and customized node & edge annotation.* VisANT also provides exploration and navigation of KEGG pathways, and weight networks based on edge thickness, edge

color, or both. While VisANT can find matching pathways for a given gene, it doesn't accept KGML files as input, only supports BioPAX pathway file (*.owl), edge list text file (*.txt) and visML/PSI-MI file (*.xml). VisANT cannot make pathway alignment and reveal the intersection of pathways.

KGML-ED is a graphical network editor. (Klukas & Schreiber, 2007) KGML-ED enables reading KGML files, and visualizing them interactively. Users can add/delete nodes and edges to graph and save the edited pathway as a new KGML file.

PathVisio is a tool that enables displaying and editing of biological pathways, similar to drawing software. In addition, since all biological components are linked to biological data using database identifiers, PathVisio identifies the biological context of a pathway. The linkage to biological databases lets mapping and visualizing experimental data on top of the pathway drawing. PathVisio 3 allows loading of user's local dataset onto pathways, and performing statistical analyses. (Kelder *et al.*, 2012) (Kutmon *et al.*, 2015)

PathVisio enables drawing new pathway design. PathVisio can also open existing pathway files but cannot read KGML files. Users should create own pathway directory to import necessary pathway files. PathVisio allows the import of GeneMAPP pathway. On the other hand our application reads the KGML files of existing pathways and re-draw them. Like our application, PathVisio allows user to search for pathways that contain a given gene product. Application searches the local pathway directory. User can search pathways in two ways: by Gene symbol (gene name) and by Gene identifier (as defined in the identifier mapping database)

Cytoscape is an open source software specialized in biological network visualization, data integration and analysis. (Shannon *et al.*, 2003) Cytoscape integrates a network with annotation, gene expression profiles and other state data. Cytoscape provides many kinds of different interactive graph layouts and shapes. Users can easily visualize and analyze their network data which is formed in specific file formats. Cytoscape can read network/pathway files written in the following formats:

- Simple interaction file (SIF or .sif format)
- Nested network format (NNF or .nnf format)
- Graph Markup Language (GML or .gml format)
- XGMML (extensible graph markup and modelling language).
- SBML
- BioPAX
- PSI-MI Level 1 and 2.5
- Delimited text
- Excel Workbook (.xls)

Among the many plugins of Cytoscape, there are few plugins on visualization of KEGG pathways from KGML files, such as CytoKEGG[3],

[3] http://apps.cytoscape.org/apps/cytokegg

KGMLReader[4] and KEGGScape (Nishida *et al.*, 2014). CytoKegg also maps the gene expression profiles onto KEGG pathways. And CytoKegg can interactively visualize the pathway maps using Cytoscape visualization features. All these plugins CytoKegg, KGMLReader and KEGGScape only share the KGML reading and pathway visualization features with our application.

Other group of Cytoscape plugins such as CyKEGGParser (Nersisyan *et al.*, 2014) and Genoscape (Clément-Ziza *et al.*, 2009) allow manipulation with KEGG pathway maps. While they can read and visualize the pathway files in KGML format, they can also export edited and tuned pathways in KGML and BioPAX formats. CyKEGGParser provides semi-automatic correction of inconsistencies between KEGG pathway images and corresponding KGML files. Genoscape integrates the gene expression data sets (from GenoScript), a transcriptomic database and KEGG pathways using the visualization features of Cytoscape.

The integrated data is visualized as Cytoscape networks. In common with our application, Genoscape can read KEGG pathway files and re-visualize edited pathway maps.

ClueGo is specialized on creating and visualizing a functionally grouped network of terms/pathways. (Bindea *et al.*, 2009) The application can reflect the relationships between the terms based on the similarity of their associated genes. Creating and visualizing networks and reflecting relationships of them can be evaluated as similar features with our application. On the other hand, ClueGO cannot find SNP-gene mappings and cannot make conversion between gene identifiers. Like most of existing tools, ClueGo cannot align pathways and find the intersections. And ClueGo is not a web-based application. Since our application is web-based, users can use the program without any installation.

[4] http://apps.cytoscape.org/apps/kgmlreader

3. MATERIALS AND METHODS

3.1 Databases for SNP, gene and pathway information

During the development of the integrated tool for searching sub-pathways for a given SNP list, following databases are utilized as data sources to connect SNP-gene-pathway information by using SNP rsIDs, gene ids or KEGG pathway ids as reference.

The Single Nucleotide Polymorphism Database (dbSNP): is a free public archive for genetic variations within and across different species. It is developed and hosted by the National Center for Biotechnology Information (NCBI) and the National Human Genome Research Institute (NHGRI). Along with SNP entries, dbSNP also contains a variety of molecular variations such as; short deletion and insertion polymorphisms (indels/DIPs), microsatellite markers or short tandem repeats (STRs), multinucleotide polymorphisms (MNPs), heterozygous sequences and named variants. (Sherry, 2001)

Unique rsIDs are selected as the default input for the proposed application, and through rsIDs the genomic locations and associated genes are recalled from dbSNP.

Ensembl VEP: The VEP (Variant Effect Predictor) lists the known effect of any variant (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions based on the coordinates of the input variants and the nucleotide changes. The VEP can reveal information such as the genes and transcripts affected by the variants, location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions), consequence of the variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift), SIFT and PolyPhen scores for changes to protein sequence, known variants in the databases, and associated minor allele frequencies from the 1000 Genomes Project. In this study we have used the Ensembl VEP for accession to SNP-gene relations. (McLaren *et al.*, 2010)

Hyperlink Management System is defined as a tool for automatically updating and maintaining hyperlinks among major biological databases. **BioDB** is an example of "Hyperlink Management System" that can be reached from http://biodb.jp/. BioDB uses several webservices to convert unique ID and accession numbers between biological databases. Here, we have used BioDB to convert Ensembl Gene IDs (ENSG) to KEGG Gene ID.

Kyoto Encylepedia for Genes and Genomes (KEGG) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies that is served at http://www.genome.jp/kegg/. (Ogata *et al.*, 1999) KEGG PATHWAY is a collection of manually drawn pathway maps under KEGG visualizing the collective knowledge wide selection of biological and metabolical processes. The gene-pathway relations and their graphical representations are collected from the KEGG PATHWAY during the study.

3.2 Enabling Technologies

Following technologies that are listed in the **Table 1** have been used for the development of the application.

Table 1. Software development technologies.

Purpose/Type	Technology
Programming Language	Java (JDK 1.8.0 Update 25)
Application Framework for GUI	GWT 2.5.1 & GXT 3.0.1
Library for drawing graphs	yFiles for HTML GWT Overlay 1.2.1.1
Application Server	GlassFish 4.1
Database Server	MySQL 5.6
IDE	NetBeans 8.0.2
Project Type	Maven Web Application with GWT
Persistence Provider	EclipseLink (JPA 2.1)
Plugin for NetBeans	GWT4NB (for GUI Debugging and fast development)
Development Browser	Firefox 25.0 (before 26.0)
Plugin for Firefox	GWT Developer Plugin

3.3 Methods

In this study, first, a local database was designed. The database was a mirror of the KEGG database and contained all of the information fetched from KEGG as KGML files for human pathways. Other organism pathways were also added to the local database, only partially, because of the time consuming mirroring process. The structure of the local database has been shown in the **Figure 3**. The database was populated by fetching all KGML files from KEGG servers using the REST services. REST (Representational State Transfer) is a service architecture which is used for transferring data between client and server machines, in XML and JSON formats. (Fielding, 2000) KGML files are XML files structured as shown in **Figure 3**. A KGML file for a pathway has all the information about the pathway graph, such as node types, names, coordinates, colors, edges, relation types etc. But fetching and inserting all KGML files of KEGG server is not efficient and one has to update the database regularly to reflect the changes in the original KEGG database. One complete update of the database requires fetching of about 286935 pathways which takes about one week to finish; hence, making it very difficult to conduct these updates in a regular manner. Therefore, we changed our design to use only REST services and process KGML files on run-time.

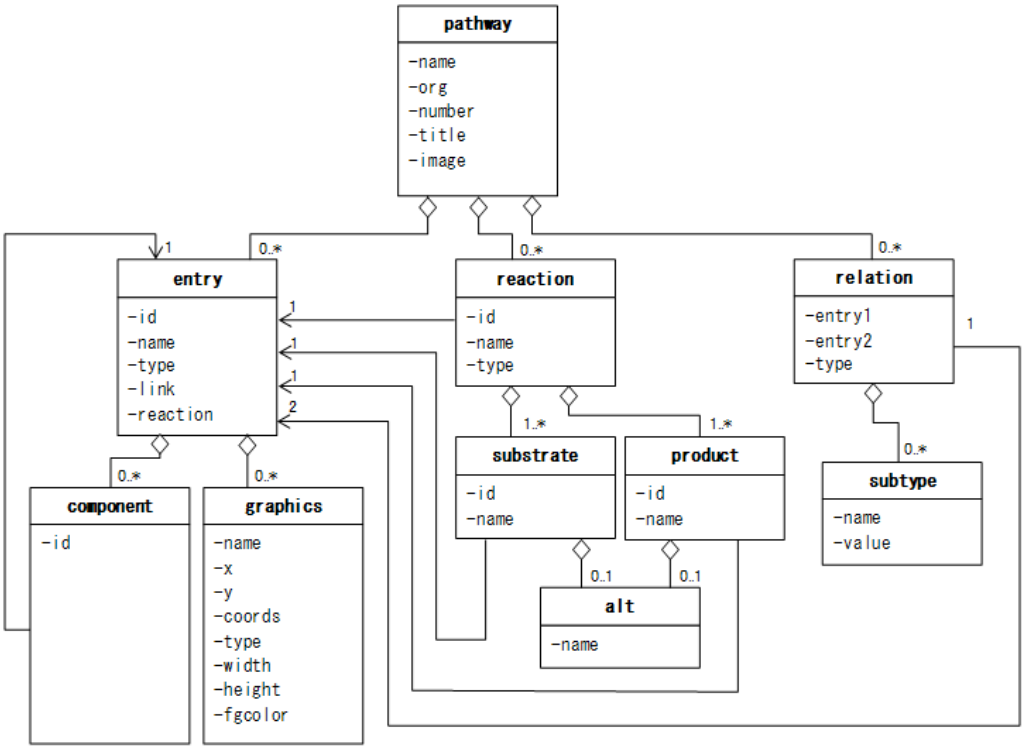

Figure 3. KGML structure, as published at http://www.kegg.jp/kegg/xml/docs/.

3.3.1 System Architecture

We have used a service-based structure that makes service calls between client side and server side. **Figure 4**, **Figure 5**, **Figure 7** and **Figure 8** show the general hierarchy of the system architecture.

In **Figure 4**, client side architecture is illustrated. Execution starts from Main.html file and it links to Main.gwt.xml which is configuration file of a GWT project. Main.gwt.xml has the linkage to MainEntryPoint class which is the starting point of a GWT project. ServiceAsync files are the interface files that are automatically generated by GWT. They are generated according to client side call of service.

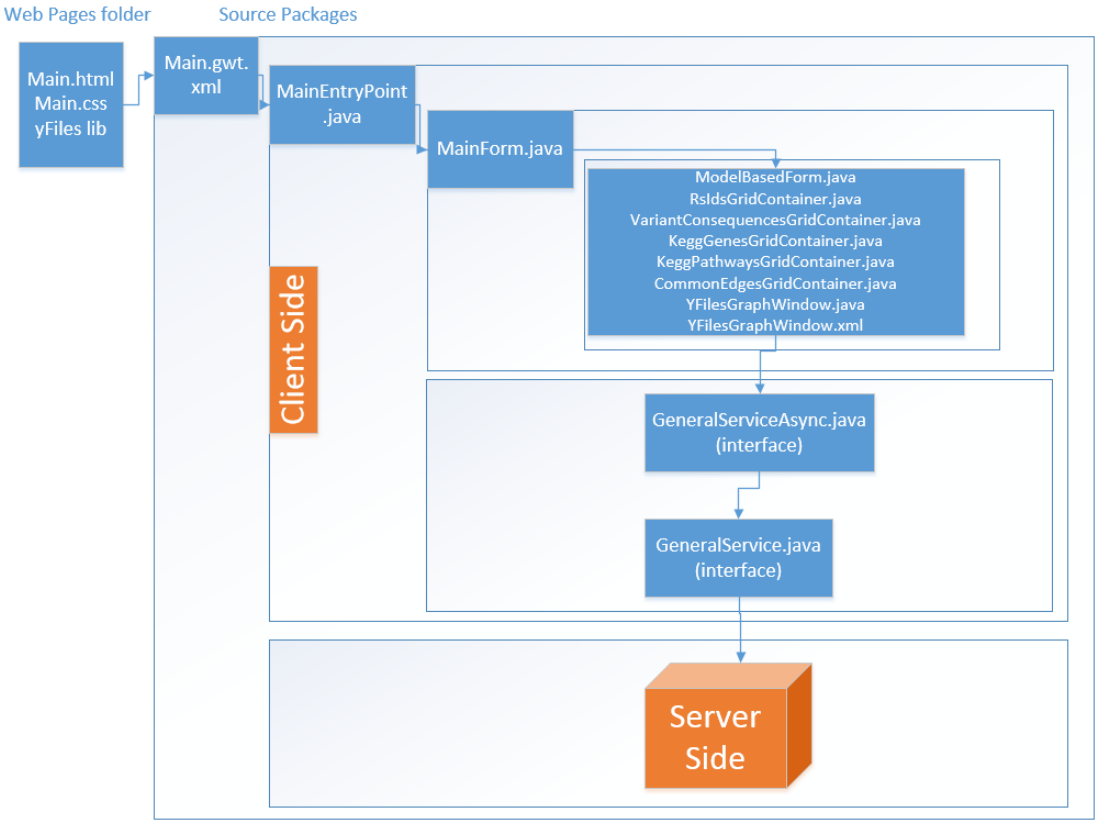

Figure 4. Client side architecture: Client side is the region where user requests are evaluated and sent to server side for data queries, and finally the query results are presented in a GUI.

In **Figure 5**, shared area classes are illustrated. These classes are only used to carry data between client side and server side. And as a result, they can be used on both sides. On the other hand, GWT can translate only "client" and "shared" packages to JavaScript code, as stated in Main.gwt.xml. Therefore, every library to be translated must be inherited in Main.gwt.xml. Otherwise GWT cannot recognize all libraries.

Figure 5. Shared area: Shared area is the region between client side and server side to carry requested data.

Main.html | Main.gwt.xml | pom.xml [Tez-web] | persistence.xml | glassfish-resources.xml

XML | General

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE module PUBLIC "-//Google Inc.//DTD Google Web Toolkit 2.0//EN"
        "http://google-web-toolkit.googlecode.com/svn/releases/2.0/distro-source/core/src/gwt-module.dtd">
<module rename-to='Main'>

    <inherits name='com.sencha.gxt.ui.GXT'/>
    <inherits name="yfiles"/>

    <!-- include a library file for browser compatibility -->
    <script src="../lib/yfiles/es5-shim.js"/>
    <!-- include the AMD module loader -->
    <script src="../lib/require.js"/>

    <set-property name="gxt.user.agent" value="gecko1_8"></set-property>

    <!-- Inherit the core Web Toolkit stuff.                        -->
    <inherits name='com.google.gwt.user.User' />

    <!-- We need the JUnit module in the main module,               -->
    <!-- otherwise eclipse complains (Google plugin bug?)           -->
    <inherits name='com.google.gwt.junit.JUnit' />

    <!-- Specify the app entry point class.                         -->
    <entry-point class='tr.com.ersoy.tez.client.MainEntryPoint' />

    <!-- Specify the paths for translatable code                    -->
    <source path='client' />
    <source path='shared' />
</module>

```

Figure 6. Preview of Main.gwt.xml: The main configuration file of the project where inherited libraries and java script files, valid browsers and the packages to be translated to JavaScript are defined.

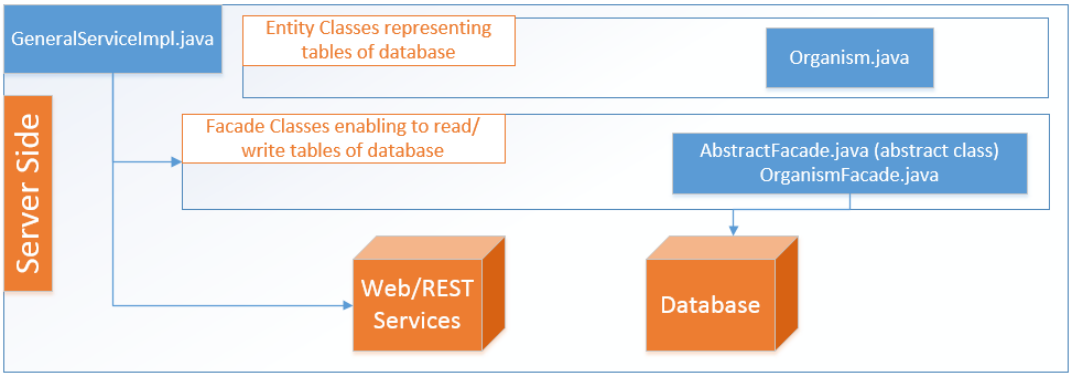

Figure 7. Server side architecture: Server side is the business layer where database queries and web service calls are implemented.

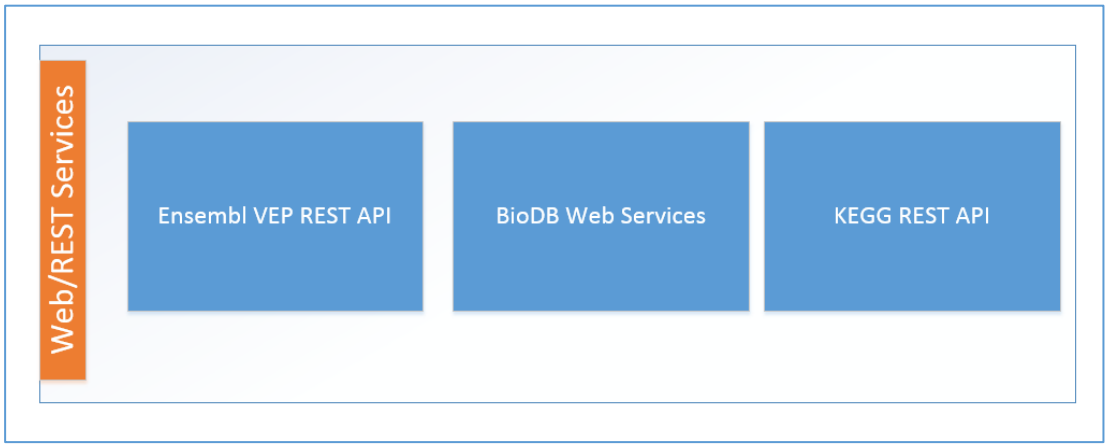

Figure 8. Web and REST services: The external database resources that are called from server side

The other important configuration files are given in **Table 2**.

Table 2. Important files and classes.

File Name	Description
pom.xml	Configuration file of Maven Project. We write all necessary library names as "dependencies" to this file. And when project is built all stated libraries are downloaded as jar files.
web.xml	Configuration file of service architecture. All Service classes and ServiceImpl classes are matched to each other in this file.
persistence.xml	Configuration file of database. Enables connection between database and facade files.
glassfish-resources.xml	Configuration file of GlassFish Application Server. Database connection properties are defined here, as

jdbc-resource and jdbc-connection-pool tags.

Previews of the configuration files are shown in Appendix H.

3.3.2 Model structure of the system

We use the model classes shown in **Table 3** to hold data retrieved from web services and to show in the tables of GUI. They are all Java classes. The details of the classes are listed in Appendix I.

Table 3. Models used as data structure.

Models
RsIdGridModel
VariantConsequencesGridModel
KeggGeneGridModel
KeggPathwayGridModel
CommonEdgeGridModel
EnrtyElementModel
GraphicsElementModel
ComponentElementModel
OrganismModel

3.3.3 Finding Gene IDs using an rsID

To find SNP associated gene ids, we have used three different REST services. These are Ensembl VEP REST API, KEGG REST API and BioDB web services.

As KEGG REST API does not offer any services to get related gene ids of a given SNP, we have used the Ensembl VEP REST API to retrieve the "variant consequences" of SNPs. Ensembl VEP has the following REST URL to get the variation information of a SNP in JSON format.

http://rest.ensembl.org/vep/human/id/rs116035550?content-type=application/json

The rsID is written into highlighted section in the URL above. This REST service returns the results as shown in Appendix A.

We have implemented two options for converting gene information to KEGG pathway. First is the conversion via gene_symbol, and the second is the conversion via gene_id. They are the tags that are used in the "transcript_consequences" array of JSON output. HGNC gene symbols are accepted in the "gene_symbol" field, and "gene_id"s are required as Ensembl gene ID (ENSG).

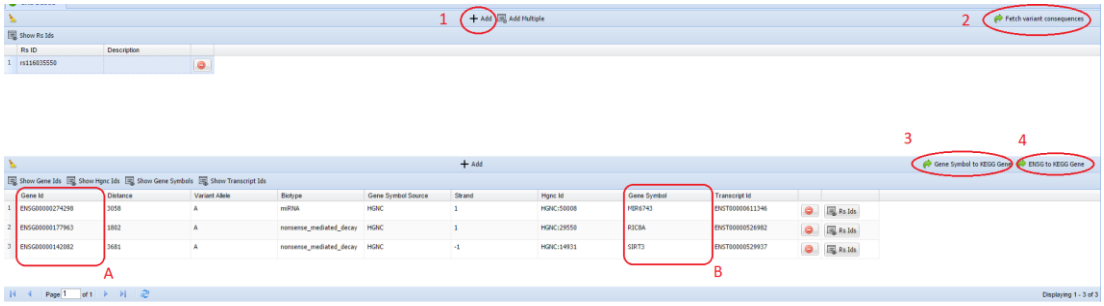

Figure 9. Variant consequences of the SNP with given RS id

When the list of variant consequences is ready, if we press the button numbered as 3 in **Figure 9**, a keyword search using "KEGG find" REST service is started. This service finds gene definitions matching with the same gene symbols in the table. Example URLs are given as following.

http://rest.kegg.jp/find/genes/"SIRT3"

http://rest.kegg.jp/find/genes/"RIC8A"

```
hsa:23410       SIRT3, SIR2L3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
ptr:450911      SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
pps:100973326   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
ggo:101126809   SIRT3; NAD-dependent protein deacetylase sirtuin-3, mitochondrial; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
pon:100453970   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
nle:100604320   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
mcc:720737      SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
mcf:102142238   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
cjc:100386098   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
mmu:64384       Sirt3, 2310003L23Rik, AI848213, Sir2l3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
rno:293615      Sirt3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
cge:100762304   Sirt3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
ngi:103751056   Sirt3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
hgl:101697494   Sirt3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
ocu:100354500   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
tup:102502361   SIRT3; sirtuin 3; K11413 NAD-dependent deacetylase sirtuin 3 [EC:3.5.1.-]
```

Figure 10. Example output of KEGG gene ids after calling

http://rest.kegg.jp/find/genes/"SIRT3"

Then, all unique gene ids for Homo sapiens (hsa) are combined and listed.

The button numbered 4 in **Figure 9** starts a one-to-one conversion from ENSG ids to KEGG gene ids through "BioDB convert" web service. An example URL for converting two ENSG ids to KEGG gene ids is given in the following link.

http://biodb.jp/convert/ensg_id/kegg/ENSG00000184005,ENSG00000080298/id.txt

Searching for gene annotations through BioDB by converting ENSG IDs directly to KEGG gene IDs is much faster, as it is using a well-structured relational database, which is updated daily.

The following two figures show the results of two different conversion methods.

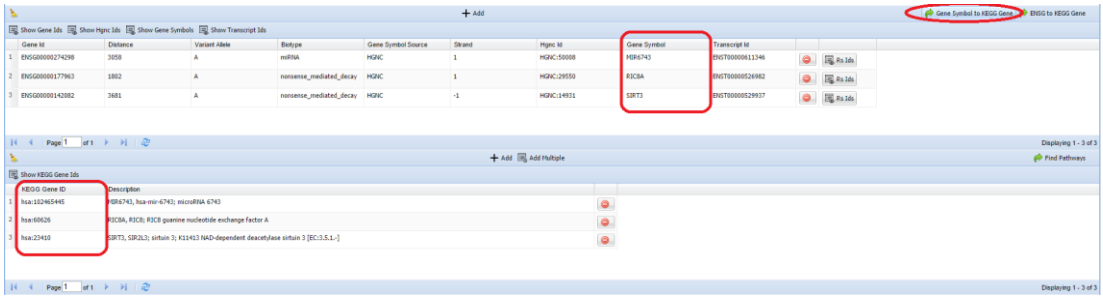

Figure 11. KEGG gene ids which are converted from gene symbols in the upper list, by "KEGG find" REST service

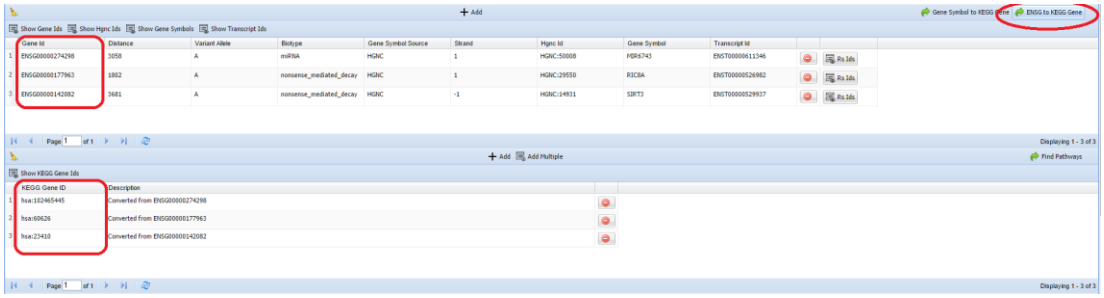

Figure 12. KEGG gene ids which are converted from ENSG ids in the upper list, by "BioDB convert" web service

We may delete some gene ids or add new ones to the list, manually. The description column helps us to get more information about the genes in the table.

3.3.4 Finding pathway ids using KEGG gene ids

Using the list of KEGG gene ids, we can run the "KEGG link" REST service to find linked (related) pathways. For example;

http://rest.kegg.jp/link/pathway/hsa:23410

The query above gives the following result as a tab separated output.

hsa:23410 path:hsa05230

Then we can fill the pathway model with these results.

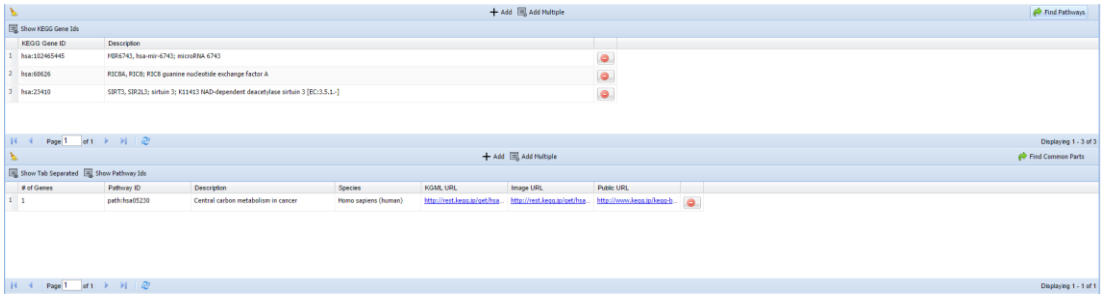

Figure 13. KEGG pathways which are related to KEGG gene ids in the upper list

Only 1 pathway was found using 3 different KEGG gene ids. We may also add pathway ids manually and proceed to the next step directly from pathways table.

"# of Genes" column shows the number of genes which are present in the related pathway. Here, only one gene is present in the pathway. If number of genes that are present in this pathway, is equal to the number of genes given in the upper list, we may say that path:hsa05230 is common for all genes.

3.3.5 Finding common sub-graphs of pathways

To find common-parts, we decided to compare unique edges of all given pathways. Edges are the smallest sub-graphs of pathways. Once we accumulate a list of unique edges, we can add-up them according to first and last nodes.

We are fetching KGML files using following sample URL. Then, we are parsing all fetched KGML files.

http://rest.kegg.jp/get/hsa05230/kgml

We used a hashmap data structure to handle unique edges and their usage counts in different pathways. A hashmap is a type of data structure that holds data in a key-value manner. Keys must be unique. **Figure 14** shows the structure of the hashmap to handle unique edges.

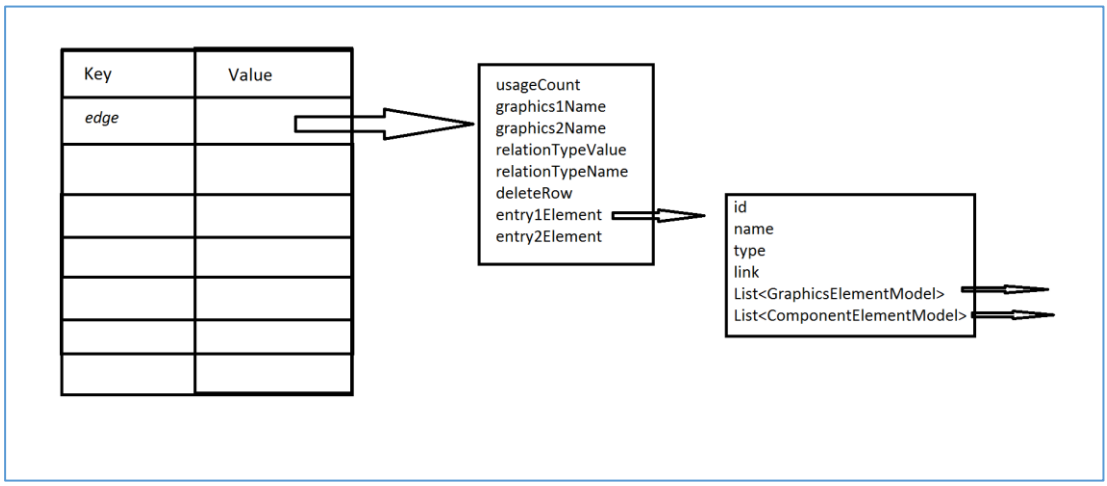

Figure 14. Proposed structure of hashmap to handle unique edges

The "edge" key structure is in the form of "Entry1Name [tab] RelationType [tab] Entry2Name". By this structure we can also distinguish edges by relation type.

If an existent edge is requested to be added to the hashmap again, we only need to increase the count value. So, the other information remains same as in the first pathway. Using the node information of only first pathway gives us the opportunity of visualizing according to only one pathway. And this feature gives us a consistent output in the manner of coordinates.

After accumulating edges of all pathways, they are listed in a table, ordered in decreasing values of usage count. If the usage count equals to the number of pathways, it means that this edge is present in all the pathways that are analyzed.

As there is only one pathway as shown in **Figure 15**, we add one more pathway manually. The new pathway is hsa05200 – Pathways in cancer. Then, we proceed to find common sub-graphs of these two pathways.

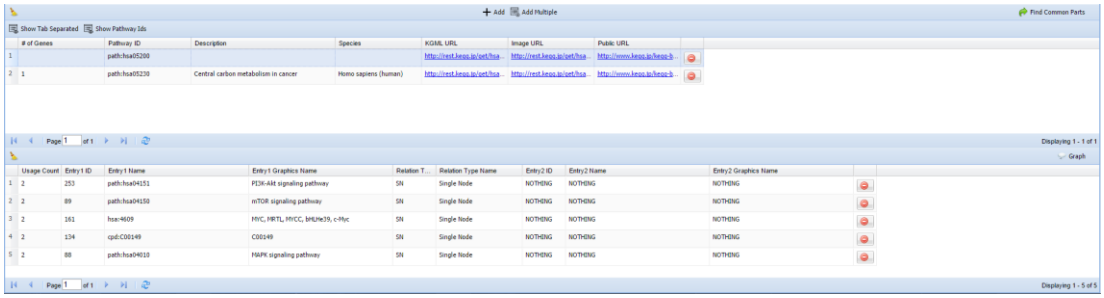

Figure 15. Common edges and single nodes of given pathways in the upper list

3.3.6 Drawing the graph of common sub-graphs

For drawing the graphical output, we add up all the edges having the same entry name as the last and first entries. It allows us to connect a chain of interactions in a linear manner.

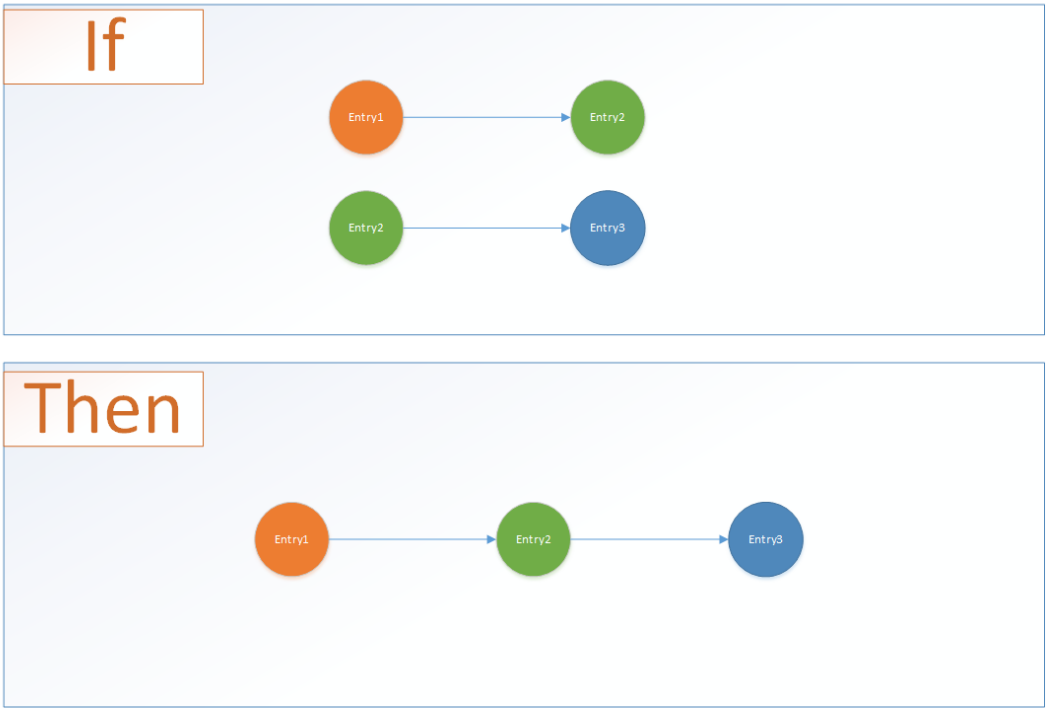

Figure 16. A simple representation of chain mechanism which is connected on same nodes.

Two layout options are added to application: default layout and tree-layout. Default layout is generated using position information given in KGML file. Every KGML file has different positions for a common sub-graph. Therefore, we are constructing the common sub-graph according to the first parsed KGML file.

Tree-layout is one of the auto-layout options served by yFiles. Other layouts supported by yFiles may also be provided as well.

4. RESULTS

In this chapter, we validate the newly developed software application with three use cases. First case study aims to show the ability of the application to get intersections of given pathways correctly. We compare the existing KEGG pathway graphs to the graphs produced by the application manually. The second case study investigates the associated SNPs of Prostate Cancer genomic model. The third case study investigates the associated SNPs of Juvenile Rheumatoid Arthritis (JRA) genomic model.

4.1 Case Study 1: Calculating intersection of two similar pathways

The software developed during this study allows the user to start a search at pathway, gene or SNP level. In our first user case, we have presented the common sub-graph search function by comparing Pathways in Cancer (hsa05200) and Pancreatic Cancer (hsa05212) pathways of human (hsa).

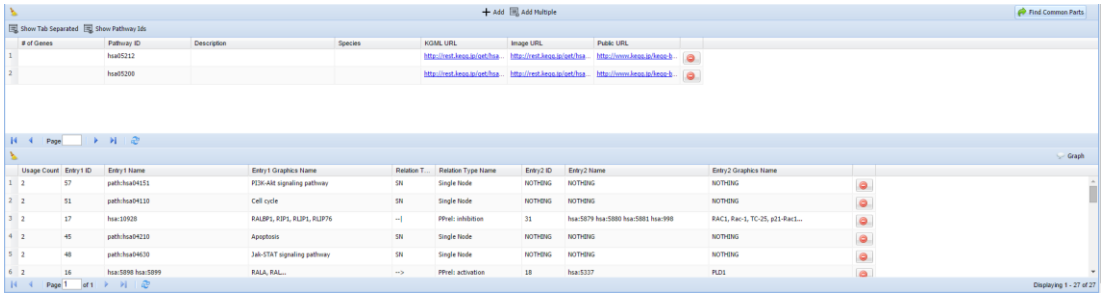

Figure 17. Common sub-graphs of hsa05200 and hsa05212: A search for common sub-graphs can be started by submitting the KEGG pathway IDs. As a result, a list of common sub-graphs identified is provided. All of 27 common sub-graphs identified in the first case are shown in **Table 12**, in Appendix B.

4.1.1 Graphical output of common sub-graphs

In KEGG Pathways, green boxes present proteins/genes, white boxes show inner maps/pathways and white tiny circles show compounds. We label the nodes with same color and shape as KEGG (**Figure 18**).

4.1.1.1 Default layout

Default layout is generated using position information given in the KGML file.

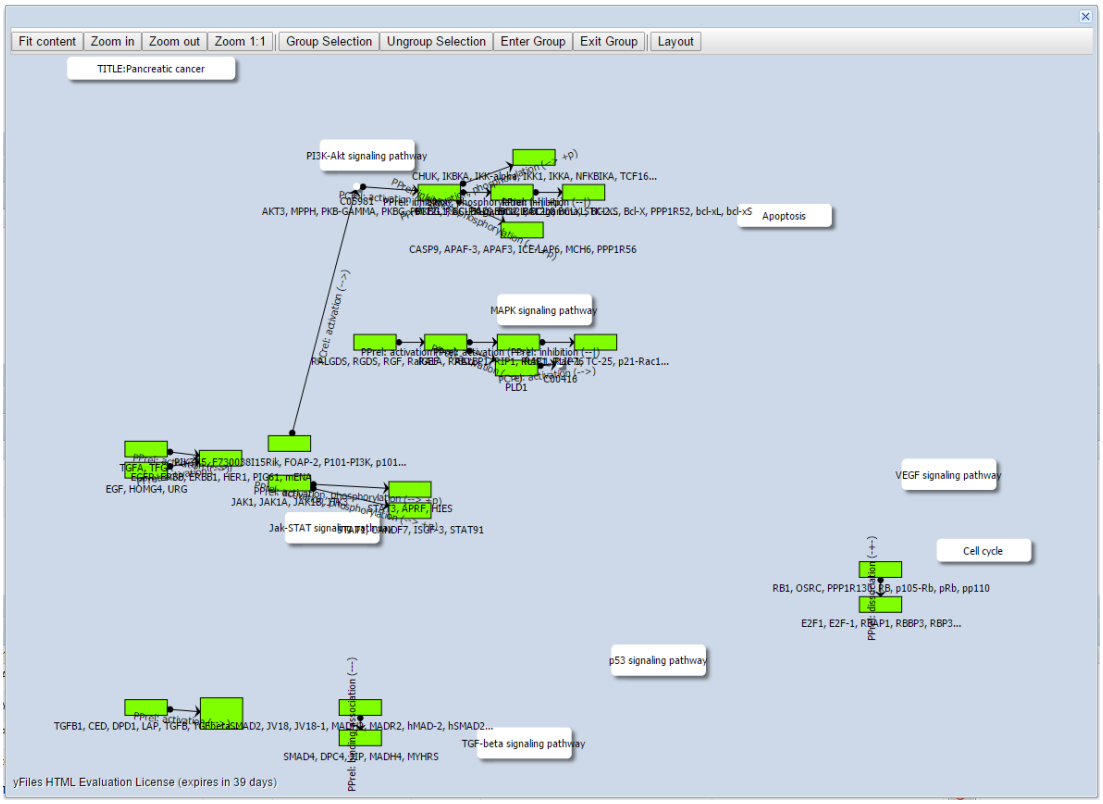

Figure 18. Graphical representation of common sub-graphs between hsa05200 and hsa05212 in default layout.

4.1.1.2 Tree layout (left-to-right)

Tree layout shows nodes and edges much more comprehensible than the default layout.

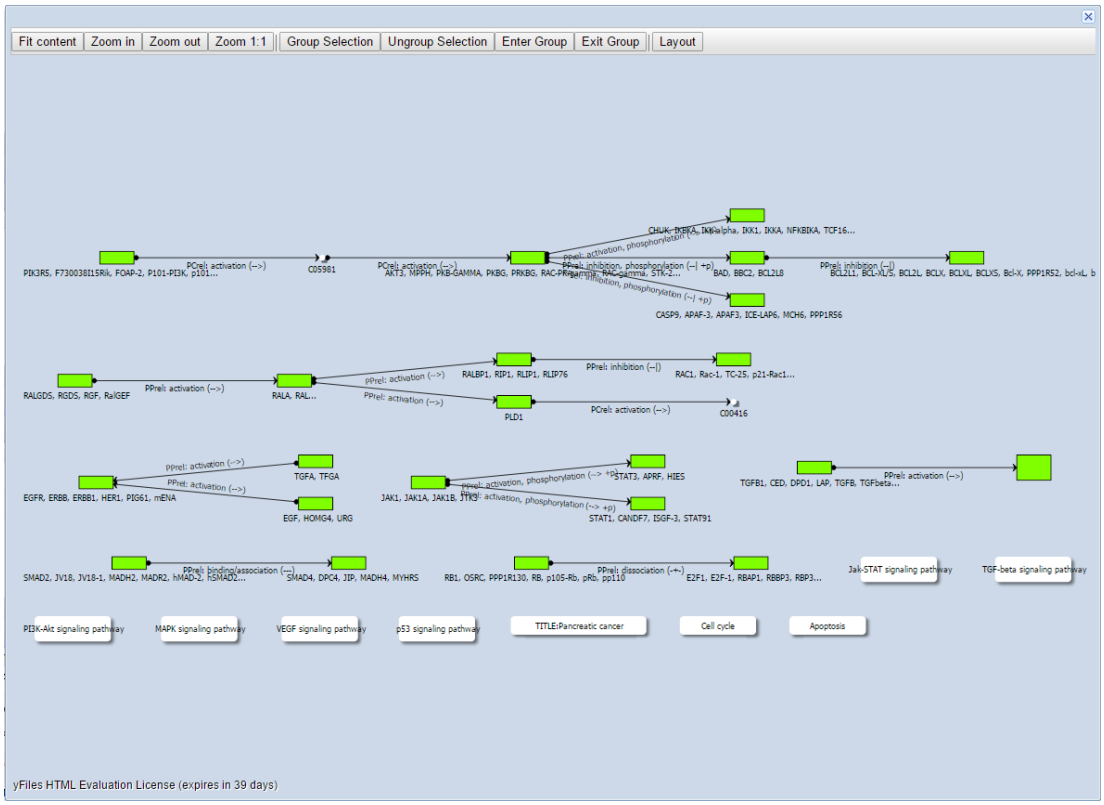

Figure 19. Graphical representation of common sub-graphs between hsa05200 and hsa05212 in tree layout (left-to-right).

4.1.2 Validation of common sub-graph of hsa05200 and hsa05212

In order to confirm the sub-graph finding algorithm we have analyzed the original KEGG pathways manually and compared it to the list of sub-graphs identified. In **Figure 20** and **Figure 21** all common sub-graphs of hsa05200 and hsa05212 pathways are marked. Red marks show common edges, and orange marks show common nodes like inner maps.

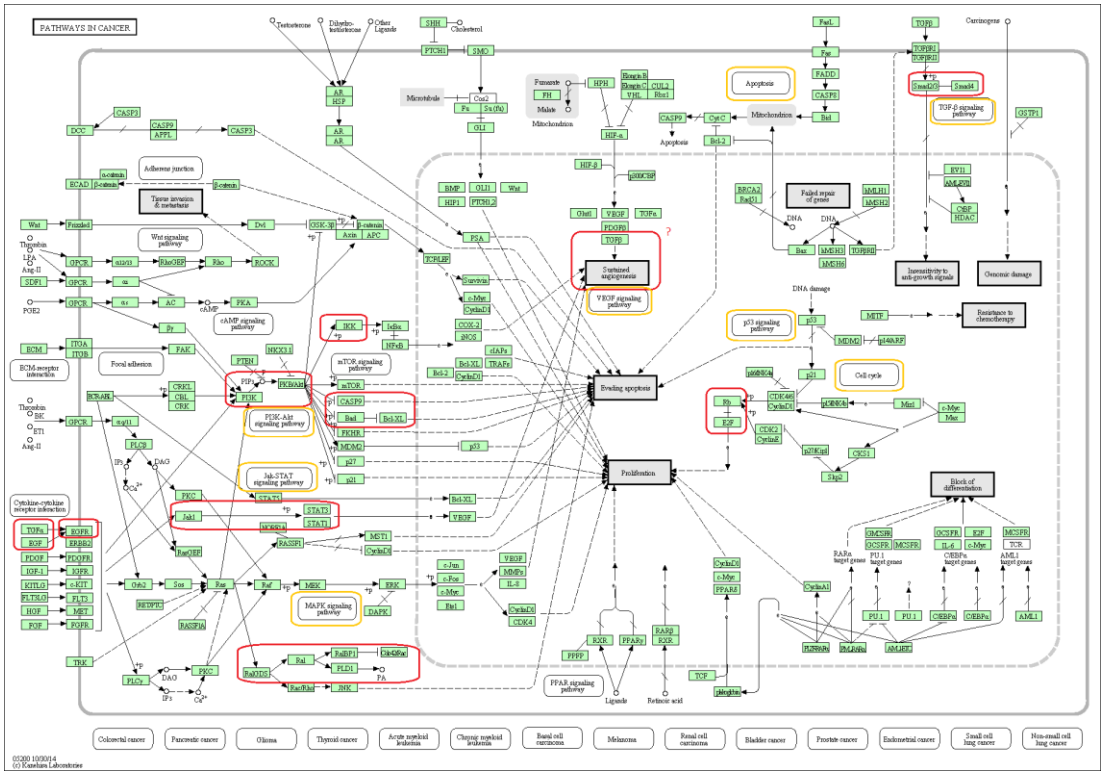

Figure 20. Common sub-graphs of hsa05200 and hsa05212 are marked on hsa05200 pathway.

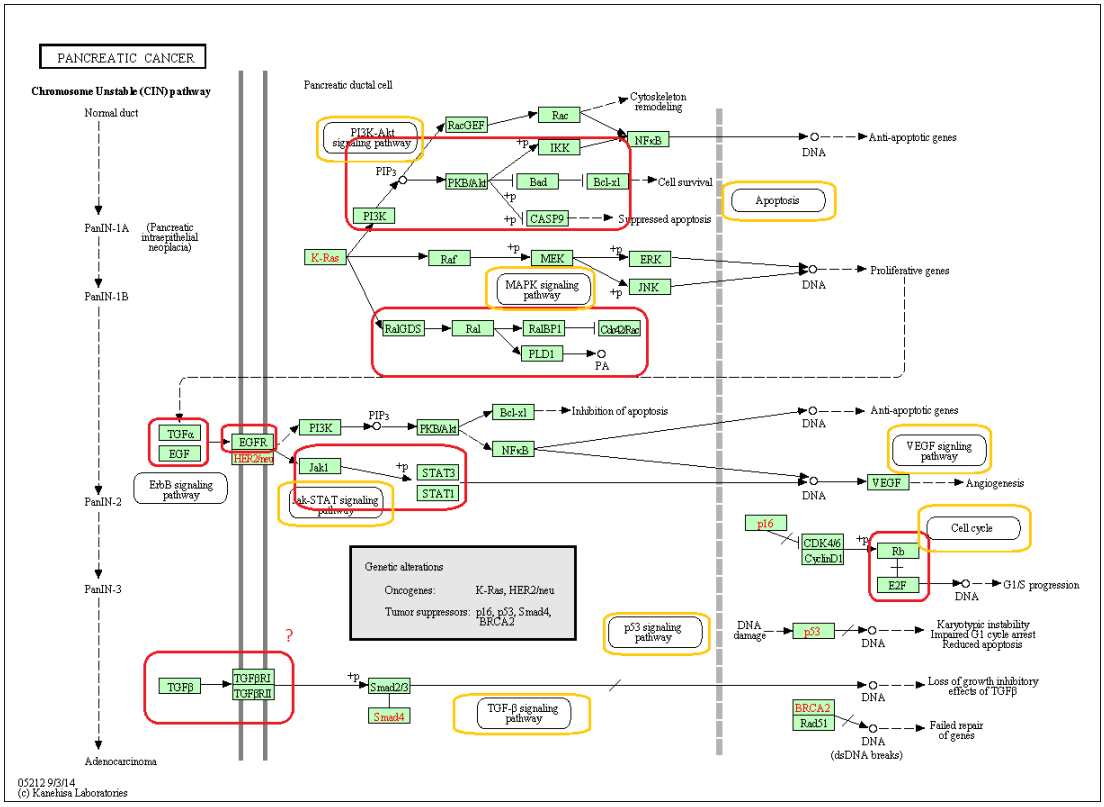

Figure 21. Common sub-graphs of hsa05200 and hsa05212 are marked on hsa05212 pathway.

4.2 Case Study 2: 107 SNPs associated with Prostate Cancer

The developed application has the capability of searching for multiple SNP-gene-pathway annotations and then identifying the common sub-graph. In the second case study we have used the list of 107 SNPs, which have been proposed to have a high classification performance in the genomic model of prostate cancer, shown in **Table 4**. (Yücebaş & Aydın Son, 2014)

Prostate is a gland in the male reproductive system and the prostate cancer is a slowly growing carcinogenic disease in the gland of prostate. However in some cases, prostate cancer grows relatively fast. The cancer cells may spread from the prostate to other parts of the body, particularly the bones and lymph nodes. Prostate cancer is also known as carcinoma of the prostate. Prostate cancer is the most common cancer among men, after skin cancer.[5] [6]

[5] http://www.cancer.org/cancer/prostatecancer/
[6] http://en.wikipedia.org/wiki/Prostate_cancer

Table 4. 107 SNPs for Prostate Cancer. (Yücebaş & Aydın Son, 2014)

rs11729739	rs2442602	rs17363393	rs7562894
rs17701543	rs3093679	rs280986	rs17595858
rs9848588	rs9347691	rs11790106	rs5972169
rs964130	rs6851444	rs11126869	rs4782945
rs10195113	rs11086671	rs7775829	rs12243805
rs1433369	rs6887293	rs9401290	rs1454186
rs12733054	rs3812906	rs17284653	rs4827384
rs17375010	rs6549458	rs1379015	rs11221701
rs766045	rs2666205	rs1965340	rs501700
rs12201462	rs7010457	rs6704731	rs17432165
rs4908656	rs10854395	rs6475584	rs1470494
rs9462806	rs12644498	rs7876199	rs744346
rs1974562	rs12247568	rs17673975	rs6774902
rs10954845	rs6686571	rs6779266	rs6747704
rs10745253	rs504207	rs16863955	rs17152800
rs12980509	rs12119983	rs9963110	rs10068915
rs2296370	rs6708126	rs960278	rs1020235
rs7843255	rs2853668	rs2115101	rs10106027
rs2194505	rs524534	rs2602296	rs17111584
rs2120806	rs17799219	rs17400029	rs17178580
rs10517581	rs7183502	rs2948268	rs3760903
rs2103869	rs13011951	rs11685549	rs11584032
rs10788555	rs12266639	rs6676372	rs4562278
rs7067548	rs2826802	rs4793790	rs11885120
rs17001078	rs7024840	rs2711134	rs7584223
rs918285	rs197265	rs4517938	rs7152946
rs7034430		rs517036	rs340542

The “Add Multiple” button is used to input multiple lines of rsIDs at once. We may paste the list into the text area. All rsIDs must be on separate lines.

There is only one option for multiple input of rsIDs. Because copy-paste action is much easier than uploading formatted files, we’ve decided to open text areas as popups for both input and output.

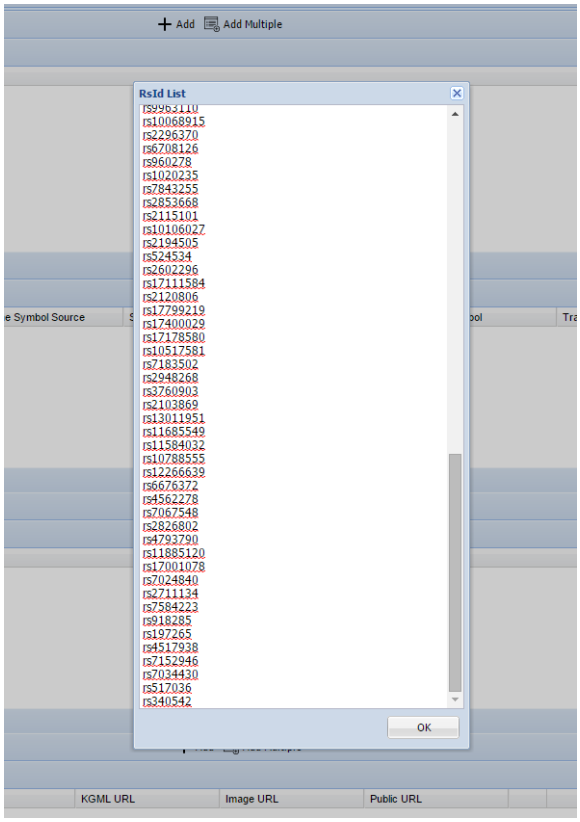

Figure 22. Adding multiple lines of rsIDs.

Using Ensembl VEP REST Service, we get the total of 88 unique gene ids for 107 SNPs. The matching SNPs of genes are shown in **Table 15**, in Appendix E.

Table 5. 88 Ensembl gene ids for the SNPs in the Prostate Cancer Model.

ENSG00000127616	ENSG00000183117	ENSG00000115306
ENSG00000170579	ENSG00000156875	ENSG00000174891
ENSG00000183023	ENSG00000118263	ENSG00000105926
ENSG00000196566	ENSG00000249699	ENSG00000213981
ENSG00000170858	ENSG00000109079	ENSG00000147316
ENSG00000258405	ENSG00000080298	ENSG00000237498
ENSG00000229298	ENSG00000157168	ENSG00000232337
ENSG00000257839	ENSG00000269509	ENSG00000231557
ENSG00000235495	ENSG00000070669	ENSG00000099810
ENSG00000143196	ENSG00000112419	ENSG00000176204
ENSG00000117501	ENSG00000172554	ENSG00000279966
ENSG00000137473	ENSG00000257453	ENSG00000198821
ENSG00000155792	ENSG00000162897	ENSG00000187231
ENSG00000213973	ENSG00000185594	ENSG00000205830
ENSG00000139289	ENSG00000105229	ENSG00000264545
ENSG00000168702	ENSG00000106078	ENSG00000253452

ENSG00000226744	ENSG00000271096	ENSG00000198633
ENSG00000255872	ENSG00000232837	ENSG00000171735
ENSG00000259282	ENSG00000162373	ENSG00000154978
ENSG00000237356	ENSG00000151693	ENSG00000196353
ENSG00000196503	ENSG00000064270	ENSG00000187391
ENSG00000043039	ENSG00000082684	ENSG00000224467
ENSG00000223761	ENSG00000171956	ENSG00000169306
ENSG00000241073	ENSG00000186094	ENSG00000184005
ENSG00000164362	ENSG00000112530	ENSG00000157087
ENSG00000107338	ENSG00000135678	ENSG00000200310
ENSG00000225913	ENSG00000222206	ENSG00000184903
ENSG00000154654	ENSG00000109083	ENSG00000091879
ENSG00000174780	ENSG00000162402	
ENSG00000230448	ENSG00000235751	

"BioDB convert" web service is used for the conversion of Ensembl gene ids (ENSG) to KEGG gene ids. It took only a few seconds to convert 88 ENSG ids to 64 KEGG gene ids. The application is web-based and all data lists are prepared on the server-side. So the performance is directly related to server machine of application and servers of external data resources (web services). And additionally, the local network bandwidth is important, too. On the client-side, the only condition that affects the performance is the amount of data to be filled in tables (grids). When the list of returned data is too long, the performance of application decreases.

Out of 88 ENSG IDs, 24 of them did not return a KEGG gene ID. And 3 of remaining 64 ENSG IDs are each converted to two different KEGG gene ids. We have excluded the incorrect ones before execution of next step. So, only 61 KEGG genes are sent as an input for next execution. The matching ENSG ids are shown in **Table 16**, in Appendix F.

We have also provided another option for gene id conversion. Gene Symbols are converted to KEGG gene ids using "KEGG find" service. But "KEGG find" service makes a keyword search in text results. So it takes almost 30 times slower than the "BioDB convert" web service.

Table 6. 64 KEGG gene ids related with Prostate Cancer.

hsa:6597	hsa:64478	hsa:6546	hsa:81552	hsa:6461
hsa:64645	hsa:6711	hsa:10200 *	hsa:132946	hsa:1368
hsa:8609	hsa:51319	hsa:7126	hsa:9914	hsa:491
hsa:51678	hsa:5991	hsa:79648	hsa:131034	hsa:493 *
hsa:147660	hsa:440	hsa:1805	hsa:8538	hsa:4685
hsa:3084	hsa:4507	hsa:80133	hsa:54437	hsa:83943
hsa:9749	hsa:80059	hsa:83894	hsa:9863	hsa:6731
hsa:54221	hsa:83953	hsa:919	hsa:27023	hsa:90410

hsa:64798	hsa:145946	hsa:221409 *	hsa:84871	hsa:285
hsa:7652	hsa:22822	hsa:51588	hsa:11141	hsa:100499171
hsa:91404	hsa:23242	hsa:147658	hsa:7015	hsa:23358
hsa:53353	hsa:79656	hsa:23261	hsa:135138	hsa:79166
hsa:100652749	hsa:9229	hsa:8853	hsa:256435	

* Incorrect conversion by BioDB web service.

The KEGG pathways of 61 KEGG genes are searched in the next step.

Totally 39 pathways, which contain the given 61 KEGG genes, were found, as shown in **Table 7**.

Table 7. 39 KEGG pathway ids related with Prostate Cancer, and their matching genes.

ID	Description	Matching KEGG Gene IDs	KEGG URL
path:hsa00250	Alanine, aspartate and glutamate metabolism	hsa:440	http://www.kegg.jp/kegg-bin/show_pathway?hsa00250
path:hsa00270	Cysteine and methionine metabolism	hsa:4507	http://www.kegg.jp/kegg-bin/show_pathway?hsa00270
path:hsa00604	Glycosphingolipid biosynthesis - ganglio series	hsa:256435	http://www.kegg.jp/kegg-bin/show_pathway?hsa00604
path:hsa01100	Metabolic pathways	hsa:440, hsa:4507, hsa:256435	http://www.kegg.jp/kegg-bin/show_pathway?hsa01100
path:hsa03018	RNA degradation	hsa:10200	http://www.kegg.jp/kegg-bin/show_pathway?hsa03018
path:hsa03060	Protein export	hsa:83943, hsa:6731	http://www.kegg.jp/kegg-bin/show_pathway?hsa03060
path:hsa04012	ErbB signaling pathway	hsa:3084	http://www.kegg.jp/kegg-bin/show_pathway?hsa04012
path:hsa04014	Ras signaling pathway	hsa:285	http://www.kegg.jp/kegg-bin/show_pathway?hsa04014
path:hsa04015	Rap1 signaling pathway	hsa:9863, hsa:285	http://www.kegg.jp/kegg-bin/show_pathway?hsa04015
path:hsa04020	Calcium signaling pathway	hsa:6546, hsa:491	http://www.kegg.jp/kegg-bin/show_pathway?hsa04020
path:hsa04022	cGMP-PKG signaling pathway	hsa:6546, hsa:491	http://www.kegg.jp/kegg-bin/show_pathway?hsa04022
path:hsa04024	cAMP signaling pathway	hsa:491	http://www.kegg.jp/kegg-bin/show_pathway?hsa04024
path:hsa04064	NF-kappa B signaling pathway	hsa:51588	http://www.kegg.jp/kegg-bin/show_pathway?hsa04064
path:hsa04066	HIF-1 signaling	hsa:285	http://www.kegg.jp/kegg-

	pathway		bin/show_pathway?hsa04066
path:hsa04120	Ubiquitin mediated proteolysis	hsa:51588	http://www.kegg.jp/kegg-bin/show_pathway?hsa04120
path:hsa04144	Endocytosis	hsa:8853	http://www.kegg.jp/kegg-bin/show_pathway?hsa04144
path:hsa04151	PI3K-Akt signaling pathway	hsa:285	http://www.kegg.jp/kegg-bin/show_pathway?hsa04151
path:hsa04260	Cardiac muscle contraction	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa04260
path:hsa04261	Adrenergic signaling in cardiomyocytes	hsa:6546, hsa:491	http://www.kegg.jp/kegg-bin/show_pathway?hsa04261
path:hsa04360	Axon guidance	hsa:54437	http://www.kegg.jp/kegg-bin/show_pathway?hsa04360
path:hsa04514	Cell adhesion molecules (CAMs)	hsa:4685	http://www.kegg.jp/kegg-bin/show_pathway?hsa04514
path:hsa04530	Tight junction	hsa:9863	http://www.kegg.jp/kegg-bin/show_pathway?hsa04530
path:hsa04630	Jak-STAT signaling pathway	hsa:51588	http://www.kegg.jp/kegg-bin/show_pathway?hsa04630
path:hsa04650	Natural killer cell mediated cytotoxicity	hsa:919	http://www.kegg.jp/kegg-bin/show_pathway?hsa04650
path:hsa04660	T cell receptor signaling pathway	hsa:919	http://www.kegg.jp/kegg-bin/show_pathway?hsa04660
path:hsa04666	Fc gamma R-mediated phagocytosis	hsa:8853	http://www.kegg.jp/kegg-bin/show_pathway?hsa04666
path:hsa04668	TNF signaling pathway	hsa:9863	http://www.kegg.jp/kegg-bin/show_pathway?hsa04668
path:hsa04724	Glutamatergic synapse	hsa:9229	http://www.kegg.jp/kegg-bin/show_pathway?hsa04724
path:hsa04961	Endocrine and other factor-regulated calcium reabsorption	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa04961
path:hsa04970	Salivary secretion	hsa:491	http://www.kegg.jp/kegg-bin/show_pathway?hsa04970
path:hsa04972	Pancreatic secretion	hsa:491	http://www.kegg.jp/kegg-bin/show_pathway?hsa04972
path:hsa04974	Protein digestion and absorption	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa04974
path:hsa04978	Mineral absorption	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa04978
path:hsa05020	Prion diseases	hsa:4685	http://www.kegg.jp/kegg-bin/show_pathway?hsa05020
path:hsa05142	Chagas disease	hsa:919	http://www.kegg.jp/kegg-

	(American trypanosomiasis)		bin/show_pathway?hsa05142
path:hsa05166	HTLV-I infection	hsa:7015	http://www.kegg.jp/kegg-bin/show_pathway?hsa05166
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa05410
path:hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa05412
path:hsa05414	Dilated cardiomyopathy	hsa:6546	http://www.kegg.jp/kegg-bin/show_pathway?hsa05414

The last step is finding the common sub-graphs of 39 KEGG pathways.

Totally 389 common sub-graphs (edges & nodes) were found. And the most common part is a compound called "cpd:C00076". It is the calcium ion. C00076 is present in 21 different pathways. The compound is a single node in all 21 pathways. The most frequent 10 common sub-graphs of all 389 are shown in

Table 13, in Appendix C.

We are drawing the graph for only common sub-graphs of all pathways. If there is any part which is common to all 39 pathways, we show it on the graph. Otherwise graph would be empty. But we can draw pathway and sub-graph matches by selection. In **Figure 23**, the parent pathways of Calcium Signaling Pathway subgraph is presented.

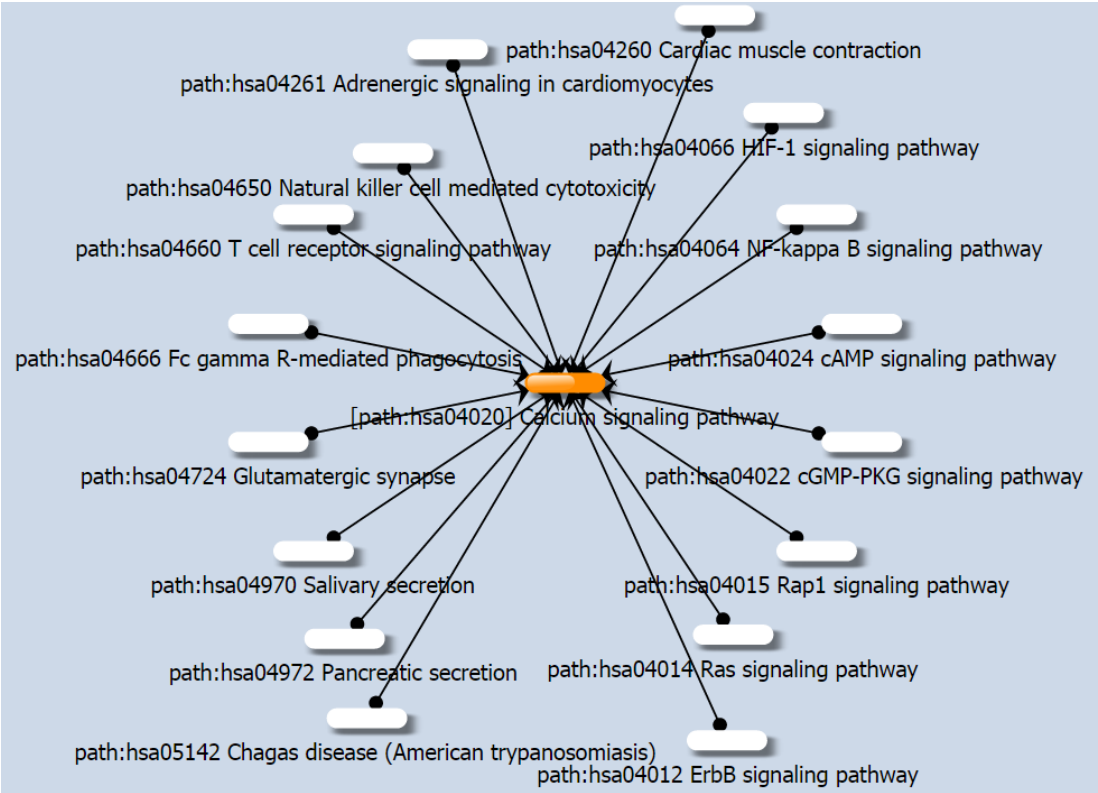

Figure 23. Parent pathways of the Calcium Signaling Pathway as a subgraph.

In the **Figure 24**, all steps of id conversions and mappings are represented with the corresponding number of inputs/outputs.

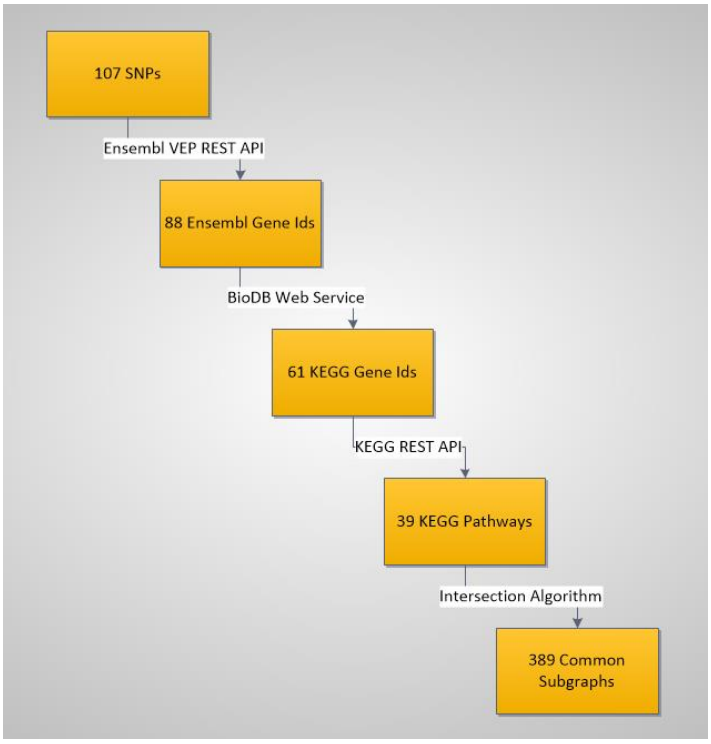

Figure 24. Conversion steps of 2nd Case Study.

4.3 Case Study 3: Network analysis of Juvenile Rheumatoid Arthritis (JRA) associated SNP set

Juvenile idiopathic arthritis (JIA) represents a group of heterogeneous diseases which are classified as arthritis of unknown origin and have onset before age of 16 years. In this study we have analyzed the SNPs that have been shown to be associated with JRA. (Aydın Son *et al.*, 2015) Total of 53 SNPs that were used to start the analysis have listed in **Table 8**.

Table 8. 53 SNPs related with JRA.

rs10244689	rs10486239	rs10789138	rs10899456
rs11208227	rs1123205	rs11602622	rs11610629
rs11611899	rs11639569	rs12597219	rs13320646
rs1340021	rs1368714	rs1369169	rs1434955
rs16904231	rs16904239	rs16904241	rs17404218
rs175310	rs17557419	rs1784354	rs1816000
rs2045093	rs216291	rs2257135	rs2367275
rs2541183	rs2886377	rs3812476	rs4274812
rs4733783	rs4733788	rs493725	rs570130
rs6105428	rs6588044	rs6711746	rs6836818
rs6844422	rs6935400	rs7115850	rs7460225

rs7462286	rs7739461	rs7740015	rs7924131
rs872674	rs9436678	rs968811	rs9874888
rs9924010			

We get 20 unique gene ids after making a search by Ensembl VEP REST API. The Ensembl gene ids and corresponding SNPs are as shown in **Table 9**. We were expecting to find following 15 RefSeq genes with at least 3 matching SNPs: ASAP1, ALK, GRID2, MAST4, ALG6, FHIT, GAB2, ITGB3BP, KCNIP4, KIRREL3, MACROD2, NXPH1, RBFOX1, UBE2CBP and VWF. 13 of 15 genes are found successfully, and 1 more gene (UBE3D) is found as the synonym of UBE2CBP. The only absent gene is MAST4. Ensembl VEP REST API cannot find the MAST4 gene for the given SNP list.

There are six extra genes:

- GRID1: protein coding gene glutamate receptor, ionotropic, delta 1
- KIRREL3-AS1: antisense gene KIRREL3 antisense RNA1
- ASAP1-IT2: sense intronic gene ASAP1 intronic transcript 2
- RP11-428L21.2: antisense gene
- RP11-420N3.3: processed transcript gene
- RP11-452H21.1: processed pseudogene gene

The four of six extra genes (KIRREL3-AS1, RP11-428L21.2, RP11-420N3.3, RP11-452H21.1) are eliminated in the next step (conversion from ENSG to KEGG gene ids).

Table 9. 20 Ensembl gene ids related with JRA with RefSeq id counterparts and matching SNPs.

ESNG ID	Matching SNPs
ENSG00000033327 (GAB2)	rs7115850, rs11602622, rs10899456
ENSG00000110799 (VWF)	rs216291, rs11611899, rs11610629
ENSG00000142856 (ITGB3BP)	rs9436678, rs6588044, rs11208227
ENSG00000153317 (ASAP1)	rs7462286, rs7460225, rs4733788, rs4733783, rs3812476, rs2045093, rs16904241, rs16904239, rs16904231, rs1340021, rs1123205
ENSG00000182771 (GRID1) *	rs7924131
ENSG00000249239 (RP11-428L21.2) *	rs1369169
ENSG00000279877 (RP11-420N3.3) *	rs12597219
ENSG00000078328 (RBFOX1)	rs9924010, rs12597219, rs11639569
ENSG00000118420 (UBE3D) **	rs7740015, rs7739461, rs6935400
ENSG00000149571 (KIRREL3)	rs570130, rs493725, rs1784354
ENSG00000171094 (ALK)	rs872674, rs6711746, rs2541183, rs2257135

ENSG00000185774 (KCNIP4)	rs6836818, rs17557419, rs1434955
ENSG00000254420 (RP11-452H21.1) *	rs7115850
ENSG00000088035 (ALG6)	rs968811, rs2367275, rs10789138
ENSG00000122584 (NXPH1)	rs17404218, rs10486239, rs10244689
ENSG00000152208 (GRID2)	rs6844422, rs4274812, rs1369169, rs1368714
ENSG00000172264 (MACROD2)	rs6105428, rs1816000, rs175310
ENSG00000189283 (FHIT)	rs9874888, rs2886377, rs13320646
ENSG00000257271 (KIRREL3-AS1) *	rs570130, rs493725
ENSG00000280543 (ASAP1-IT2) *	rs2045093

* Extra (useless) gene to be removed in the next conversion step.
** Synonym of UBE2CBP

We get 16 KEGG gene ids after conversion from Ensembl gene ids, as shown in **Table 10**.

Table 10. 16 KEGG gene ids related with JRA and matching ENSG ids.

KEGG Gene ID	Matching ENSG ID	Description
hsa:238	ENSG00000171094 (ALK)	anaplastic lymphoma receptor tyrosine kinase
hsa:2272	ENSG00000189283 (FHIT)	fragile histidine triad
hsa:2894	ENSG00000182771 (GRID1)	glutamate receptor, ionotropic, delta 1
hsa:2895	ENSG00000152208 (GRID2)	glutamate receptor, ionotropic, delta 2
hsa:7450	ENSG00000110799 (VWF)	von Willebrand factor
hsa:9846	ENSG00000033327 (GAB2)	GRB2-associated binding protein 2
hsa:23421	ENSG00000142856 (ITGB3BP)	integrin beta 3 binding protein (beta3-endonexin)
hsa:29929	ENSG00000088035 (ALG6)	ALG6, alpha-1,3-glucosyltransferase
hsa:30010	ENSG00000122584 (NXPH1)	neurexophilin 1
hsa:50807	ENSG00000153317 (ASAP1)	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1
hsa:54715	ENSG00000078328 (RBFOX1)	RNA binding protein, fox-1 homolog (C. elegans) 1
hsa:80333	ENSG00000185774 (KCNIP4)	Kv channel interacting protein 4
hsa:84623	ENSG00000149571	kin of IRRE like 3 (Drosophila)

	(KIRREL3)	
hsa:90025	ENSG00000118420 (UBE3D) **	ubiquitin protein ligase E3D
hsa:140733	ENSG00000172264 (MACROD2)	MACRO domain containing 2
hsa:100507117	ENSG00000280543 (ASAP1-IT2)	ASAP1-IT2; ASAP1 intronic transcript 2 (non-protein coding)

** Synonym of UBE2CBP

19 unique KEGG pathways are identified in which contains the 16 KEGG genes that were identified in the earlier step of the analysis. Pathway ids and descriptions are listed in the **Table 11**.

Table 11. 19 pathways related with JRA and matching KEGG gene ids.

ID	Description	Matching KEGG Gene IDs	KEGG URL
path:hsa00230	Purine metabolism	hsa:2272	Preview pathway
path:hsa00510	N-Glycan biosynthesis	hsa:29929	Preview pathway
path:hsa01100	Metabolic pathways	hsa:29929	Preview pathway
path:hsa04014	Ras signaling pathway	hsa:9846	Preview pathway
path:hsa04071	Sphingolipid signaling pathway	hsa:9846	Preview pathway
path:hsa04080	Neuroactive ligand-receptor interaction	hsa:2894, hsa:2895	Preview pathway
path:hsa04144	Endocytosis	hsa:50807	Preview pathway
path:hsa04151	PI3K-Akt signaling pathway	hsa:7450	Preview pathway
path:hsa04380	Osteoclast differentiation	hsa:9846	Preview pathway
path:hsa04510	Focal adhesion	hsa:7450	Preview pathway
path:hsa04512	ECM-receptor interaction	hsa:7450	Preview pathway
path:hsa04610	Complement and coagulation cascades	hsa:7450	Preview pathway
path:hsa04611	Platelet activation	hsa:7450	Preview pathway
path:hsa04664	Fc epsilon RI signaling pathway	hsa:9846	Preview pathway
path:hsa04666	Fc gamma R-mediated phagocytosis	hsa:9846, hsa:50807	Preview pathway
path:hsa04730	Long-term depression	hsa:2895	Preview pathway
path:hsa05220	Chronic myeloid leukemia	hsa:9846	Preview pathway
path:hsa05222	Small cell lung	hsa:2272	Preview pathway

	cancer		
path:hsa05223	Non-small cell lung cancer	hsa:2272, hsa:238	Preview pathway

In the final step of the analysis, a total of 251 common sub-graphs were revealed by integrating the KEGG pathways, where 16 of KEGG genes that the input SNPs map to have a role. Most frequently observed common sub-graphs are the inner pathways called “MAPK signaling pathway” and “PI3K-Akt signaling pathway”. They occur eight times in different pathways. The most frequent top 10 common sub-graphs out of all 251 sub-graphs returned are listed in **Table 14**, in the Appendix D.

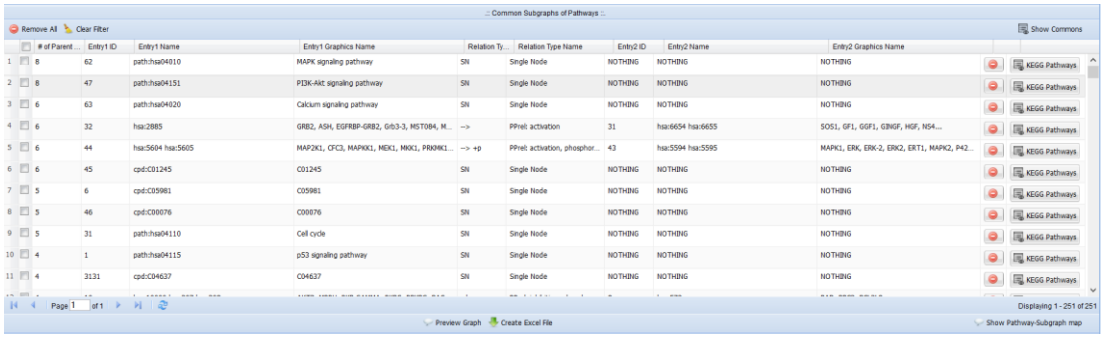

Figure 25. Common sub-graphs of 19 pathways related with JRA.

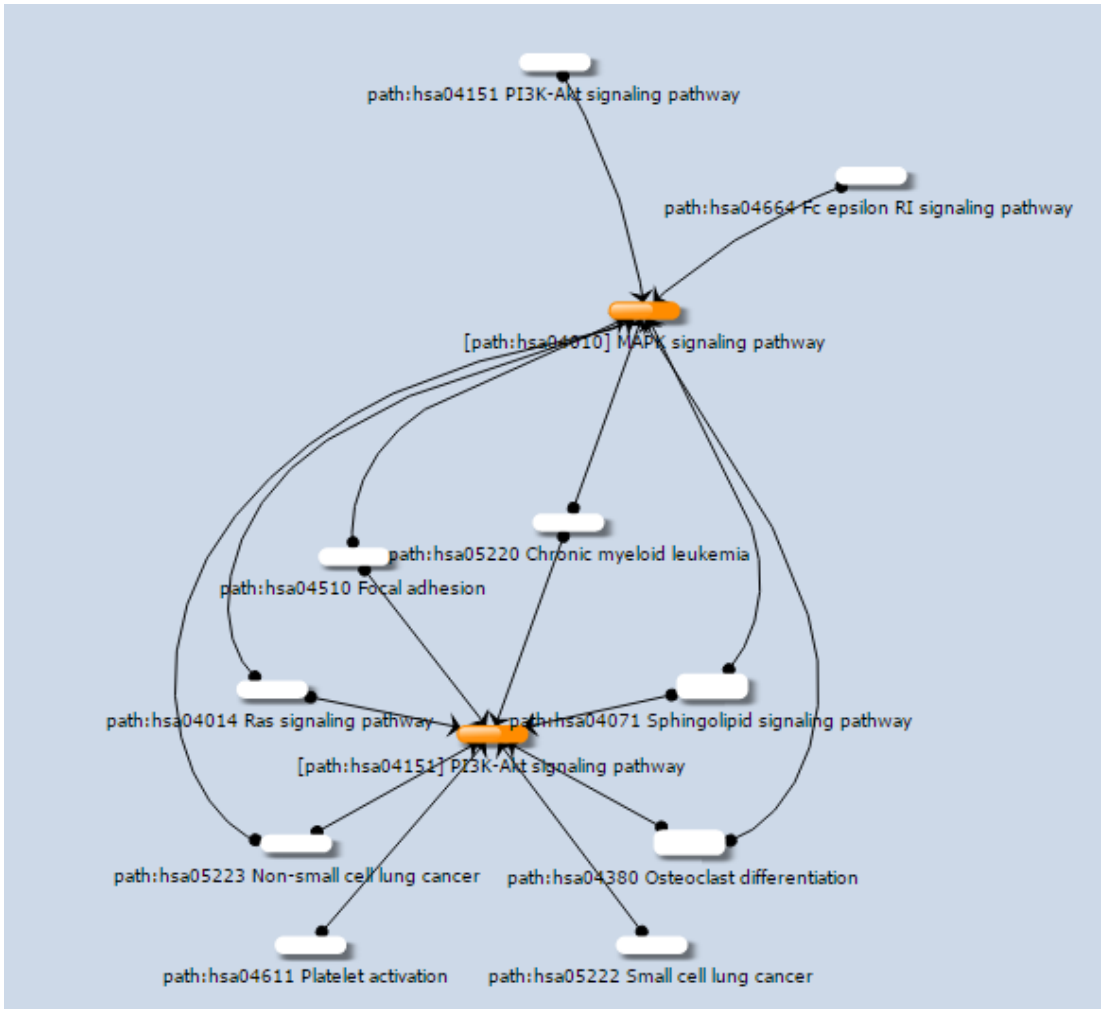

Figure 26. Parent pathways of the most common components: MAPK signaling pathway & PI3K-Akt signaling pathway.

In the **Figure 27**, all steps of id conversions and mappings are represented with the corresponding number of inputs/outputs.

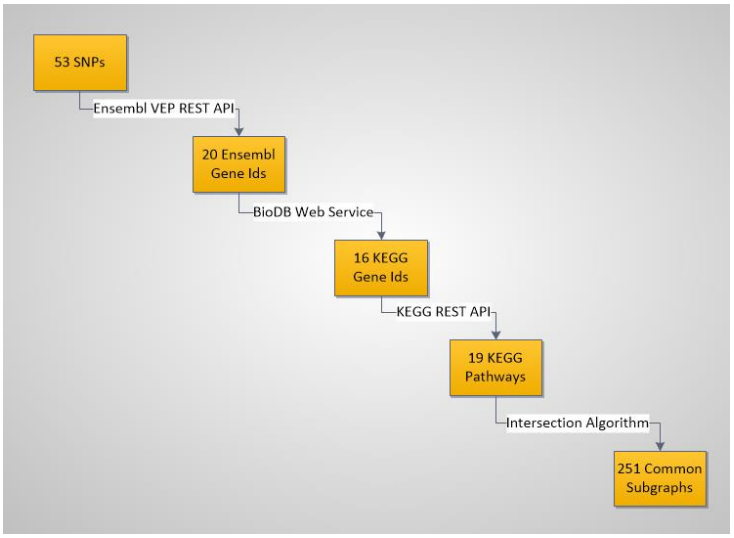

Figure 27. Conversion steps of 3rd Case Study.

5. DISCUSSION

We have developed a novel application for SNP-gene-pathway pipelining, and pathway alignment and intersection visualization. The application works as expected for all three cases.

There are similar applications already exist: SNPNexus, KEGG, KEGGgraph, KGML-ED, DAVID, Cytoscape and its plugins. We gave the details of them in the Related Works section. Most prominent advantages and disadvantages of proposed applicaiton and other are as presented below:

SNPnexus can do the first step of our software: finding matching genes of given SNPs. But it supports more variant types than our software. Our software only supports the ids of dbSNP. On the other hand, SNPnexus limits the variant number up-to 100,000. However, our software has no limitations for the number of variants.

KEGG has some disabilities compared to our application. For example, KEGG cannot find matching genes of a variant list. And, KEGG cannot find common sub-graphs of two pathways.

KEGGgraph has some features in common with our application: reading and parsing KGML files and visualizing as a graph. KGML-ED has same features with KEGGgraph, but additionaly KGML-ED can create interactive graphs where users can add/delete nodes and edges. KEGGgraph and KGML-ED cannot find the intersections of pathways.

PANOGA is a web-based project like our application. PANOGA can do SNP-gene-pathway mapping, gene id conversion, KGML file reading and visualising. However PANOGA does not support pathway alignment and cannot find the intersections of pathways.

VisANT cannot map SNPs to genes. The application can only map genes to pathways. VisANT cannot make conversion of gene identifiers. And pathway

alignment is another missing feature of application. VisANT can draw interactive pathways reading some special pathway file formats. But KGML file is not readable for the app. VisANT is not a web-based application and it has a periodically updated local database. The public VisANT implementation draws information from Predictome database. (Hu, 2004)

DAVID: Finding matching genes of a variant list, the alignment of pathways, and the interactive graph are the missing features compared to our application. The genes are mapped on KEGG generated pathways.

PathVisio enables drawing new pathway design. PathVisio can also open existing pathway files but cannot read KGML files. Users should create own pathway directory to import necessary pathway files. PathVisio allows the import of GeneMAPP pathway. On the other hand our application reads the KGML files of existing pathways and re-draw them. Like our application, PathVisio allows user to search for pathways that contain a given gene product. Application searches the local pathway directory. User can search pathways in two ways: by Gene symbol (gene name) and by Gene identifier (as defined in the identifier mapping database). On the other hand, PathVisio is a desktop application. And the application cannot make conversion between gene identifiers. Revealing pathway intersection is another missing feature of application.

Cytoscape has following features in common with our application:

- Graphical User Interface (GUI)
- Visualization
- Interactive Graph

To implement other steps, Cytoscape needs to be integrated with special databases such as dbSNP, KEGG, etc. Integration with different databases is established via plugins. Therefore, we should investigate the plugins related to SNP, gene and pathway operations: CytoKegg, KGMLReader, KEGGScape, CyKEGGParser, Genoscape and ClueGo.

Genoscape uses the visualization features of Cytoscape, but our application is using a graphics library (yFiles) for visualization. CytoKegg, KGMLReader, KEGGScape and CyKEGGParser all have same features in common with our application: reading and parsing KGML files and visualizing as a graph. For ClueGo, creating and visualizing networks and reflecting relationships of them can be evaluated as similar features with our application.

In Appendix G, **Table 17** shows the comparison between related works and our application.

In case study one, our goal was to present how the sub-graph identification algorithm working and validation of the application at the sub-graph finding level. Two pathways were selected for the analysis and common sub-graphs were identified. The application ran successfully and we were also able to manually validate the common

sub-graphs identified on the KEGG pathway images as shown in the results section. (**Figure 20** and **Figure 21**)

Second case study allowed us to present all the steps of the application starting from SNP rsID submission to final sub-graph search. Here a list of 107 SNPs that were identified in the previously published prostate cancer genomic model is investigated. The 107 SNPs were selected after a GWAS study by data mining approach, using SVM-ID3 hybrid, to best classify the prostate cases vs healthy controls.

Overall the application was able to run smoothly. Annotations of the SNP and conversion of multiple IDs are completed without an error. SNP-gene-pathway interactions for 88 SNP-gene interactions are retrieved from external databases.

The common sub-graph search revealed an empty graphic model, as there were no subgraph common to all 39 pathways identified. There were several common subgraphs, such as Calcium Signaling (common in 17 pathways), MAPK Signaling (common in 13 pathways), and PI3K-Akt signaling pathway (common in 11 pathways) and Apoptosis (common in 11 pathways) identified which have known roles in cancer biology (**Table 13**).

In the third case study statistically significant associations with JRA in the recently published GWAS are used. 53 SNPs that are mapping to 15 RefSeq genes that have at least three associated SNPs are selected for the third case study. While all the functions of the developed applications run without any problem, 251 sub-graphs are also identified within the 53 SNPs and 15 KEGG genes. The top common sub-graph revealed the enrichment of "PI3K-Akt signaling pathway" whose overactivation has been observed in autoimmune conditions (Patel & Mohan, 2005, and Xie *et al.*, 2007). Recent studies where suppression of PI3K-Akt signaling via different agents presents anti-arthritic effects shows promise for developing a new therapeutical approach for different types of autoimmune arthritis (Li *et. al.*, 2012, Han *et al.*, 2013, Yuan *et al.*, 2014). Our analysis suggests that PI3K-Akt signaling can have a role in the juvenile idiopathic arthritis, which needs further evaluation and experimental validation.

5.1 Limitations

The KEGG REST API will not support the back-end (server side) call of services, anymore. This limitation can cause re-coding of the application. We may need to move fetching operations of KGML to the front-end (client side).

Visualization of the subgraphs requires utilization of the yFiles libraries. Discontinuing the licencing of the yFiles libraries would cause failure on the graphical representation of the results.

Conversion from ENSG ids to KEGG gene ids is very lossy. Neither "KEGG find" REST service nor "BioDB convert" web service can convert ids completely.

6. CONCLUSION

6.1 Conclusion

In this thesis study, we aimed to create a pipeline starting with a list of SNP rsIDs, and finding genes and pathways consecutively, and finally calculating and visualizing the common sub-graphs of pathways. And we aimed to create a software application to serve the pipeline.

As we have presented in three different case studies, the application works without any problem. The resulting sub-graphs will decrease the amount of time and effort needed to analyze common biological pathways and will help researches reveal important information about enriched sub-pathways rapidly.

6.1.1 Accomplishments

We have managed to create a pipeline constructed on 3 different web services (Ensembl VEP, BioDB and KEGG). At the end of pipeline, we can find KEGG pathway ids which are related to given SNP RS ids.

We have managed to develop an algorithm to get the intersections of given pathways (KGML files).

We have managed to apply the yFiles graphical output to GWT working environment. This way, we solved an important problem of establishing a graphical library compatible with GWT.

And finally, we have created a web application based on Java technologies.

6.2 Future Works

As future perspectives, we can develop the application to make it more user-friendly. We can add some extra abilities such as

- Multi selection and deletion of rows

- Showing matching genes for pathways
- Showing matching pathways for common sub-graphs
- Fetching descriptions of RS ids via a web service.
- Fetching descriptions of KEGG genes via a web service.
- Adding more layout options for graphical output.
- Adding tooltips on nodes and edges of graph. So we can get a cleaner view.
- Changing arrow shapes according to relation type, like in KEGG
- Adding double-click property on inner maps to open nested pathway in a different window or to redirect to KEGG website.
- Grabbing intersections as text output, compatible with Cytoscape.
- Developing a Cytoscape plugin.
- Reporting p-values for each SNPs association if provided by the user.
- Ranking subgraphs based on features other than frequency. Parameters such as size of the subgraph can be explored.
- Reporting biggest common graph as an output.
- Integrating known SNP-disease interactions.

Current application is able to map SNPs, genes and pathways to each other. For the next version of application, disease databases can also be integrated to current entities. This integration would provide extra information about the effect of genetic variations on individuals and the relation between metabolic pathways. The new feature would have the great potential for describing disease loci and gaining insight into the underlying etiology of diseases.

REFERENCES

Alberts, B. (2008). *Molecular biology of the cell: Reference edition*. *Figure*. Retrieved from http://books.google.com/books?id=iepqmRfP3ZoC

Aydın Son Y., Batu E. D., Demirkaya E., Bilginer Y., Kasapçopur Ö., Ünsal E., Alikaşifoglu M., Özen S., "Systems-level analysis of genome wide association study results for a pilot juvenile idiopathic arthritis family study", Turkish Journal of Pediatrics, 2015 ,Acccepted for publication, Nisan-2015

Bakir-Gungor, B., Egemen, E., & Sezerman, O. U. (2014). PANOGA: A web server for identification of SNP-targeted pathways from genome-wide association study data. *Bioinformatics*, *30*(9), 1287–1289. doi:10.1093/bioinformatics/btt743

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., … Galon, J. (2009). ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091–1093. doi:10.1093/bioinformatics/btp101

Chang, J. C., & Kan, Y. W. (1979). beta 0 thalassemia, a nonsense mutation in man. Proceedings of the National Academy of Sciences of the United States of America, 76, 2886–2889. doi:10.1073/pnas.76.6.2886

Clément-Ziza, M., Malabat, C., Weber, C., Moszer, I., Aittokallio, T., Letondal, C., & Rousseau, S. (2009). Genoscape: A Cytoscape plug-in to automate the retrieval and integration of gene expression data and molecular networks. Bioinformatics, 25(19), 2617–2618. doi:10.1093/bioinformatics/btp464

Dayem Ullah, A. Z., Lemoine, N. R., & Chelala, C. (2012). SNPnexus: A web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Research, 40(W1). doi:10.1093/nar/gks364

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, *4*(5), P3. doi:10.1186/gb-2003-4-9-r60

Fielding, Roy Thomas (2000). "Chapter 5: Representational State Transfer (REST)". Architectural Styles and the Design of Network-based Software Architectures(Ph.D.). University of California, Irvine.

Han W1, Xiong Y, Li Y, Fang W, Ma Y, Liu L, Li F, Zhu X., Pharm Biol. 2013 "Anti-arthritic effects of clematichinenoside (AR-6) on PI3K/Akt signaling pathway and TNF-α associated with collagen-induced arthritis.Jan;51(1):13-22. doi: 10.3109/13880209.2012.698287. Epub 2012 Sep 21.

Hu, Z., Snitkin, E. S., & Delisi, C. (2008). VisANT: An integrative framework for networks in systems biology. Briefings in Bioinformatics, 9(4), 317–325. doi:10.1093/bib/bbn020

Kelder, T., Van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., & Pico, A. R. (2012). WikiPathways: Building research communities on biological pathways. *Nucleic Acids Research*, *40*(D1). doi:10.1093/nar/gkr1074

Klukas, C., & Schreiber, F. (2007). Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics*, *23*(3), 344–350. doi:10.1093/bioinformatics/btl611

Kutmon, M., Van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., & Evelo, C. T. (2015). PathVisio 3: An Extendable Pathway Analysis Toolbox. *PLoS Computational Biology*,11(2) doi: 10.1371/journal.pcbi.1004085

Li PP1, Liu DD, Liu YJ, Song SS, Wang QT, Chang Y, Wu YJ, Chen JY, Zhao WD, Zhang LL, Wei W. J Ethnopharmacol. 2012 May 7;141(1):290-300. doi: 10.1016/j.jep.2012.02.034. Epub 2012 Feb 27. BAFF/BAFF-R involved in antibodies production of rats with collagen-induced arthritis via PI3K-Akt-mTOR signaling and the regulation of paeoniflorin.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*, 2069–2070. doi:10.1093/bioinformatics/btq330

Nersisyan L, Samsonyan R and Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows [v1; ref status: approved 1, approved with reservations 1, http://f1000r.es/3n4] F1000Research 2014, 3:145 (doi: 10.12688/f1000research.4410.1)

Nishida K, Ono K, Kanaya S and Takahashi K. KEGGscape: a Cytoscape app for pathway data integration [v1; ref status: indexed, http://f1000r.es/3qe] F1000Research 2014, 3:144 (doi: 10.12688/f1000research.4524.1)

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. doi:10.1093/nar/27.1.29

Patel RK1, and Mohan C., Immunol Res. 2005;31(1):47-55. PI3K/AKT signaling and systemic autoimmunity.

Pevsner, J. (2009). *Bioinformatics and Functional Genomics, 2nd Ed. Tools and Applications* (p. 451). Retrieved from http://www.amazon.com/Bioinformatics-Functional-Genomics-Edition-Jonathan/dp/B004KPVA46?SubscriptionId=1V7VTJ4HA4MFT9XBJ1R2&tag=mekentosjcom-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=B004KPVA46\npapers2://publication/uuid/7BBC0F38-C8AF-4355-A446-86BB87A2081D

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504. doi:10.1101/gr.1239303

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*, 308–311. doi:10.1093/nar/29.1.308

Some organism photos are taken from http://www.arkive.org/

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. Human mutation, 17, 263–270. doi:10.1002/humu.22

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. Nature, 171, 737–738. doi:10.1097/BLO.0b013e31814b9304

Wu, J., & Jiang, R. (2013). Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. TheScientificWorldJournal, 2013, 675851. doi:10.1155/2013/675851

Xie C1, Patel R, Wu T, Zhu J, Henry T, Bhaskarabhatla M, Samudrala R, Tus K, Gong Y, Zhou H, Wakeland EK, Zhou XJ, Mohan C. Int Immunol. 2007 Apr;19(4):509-22. Epub 2007 Mar 15. PI3K/AKT/mTOR hypersignaling in autoimmune lymphoproliferative disease engendered by the epistatic interplay of Sle1b and FASlpr.

Yuan H, Yang P, Zhou D, Gao W, Qiu Z, Fang F, Ding S, Xiao W. Knockdown of sphingosine kinase 1 inhibits the migration and invasion of human rheumatoidarthritis fibroblast-like synoviocytes by down-regulating the PI3K/AKT activation and MMP-2/9 production in vitro.Mol Biol Rep. 2014 Aug;41(8):5157-65. doi: 10.1007/s11033-014-3382-4. Epub 2014 May 10.

Yücebaş, S. C., & Aydın Son, Y. (2014). A Prostate Cancer Model Build by a Novel SVM-ID3 Hybrid Feature Selection Method Using Both Genotyping and

Phenotype Data from dbGaP. *PloS One*, *9*(3), e91404. doi:10.1371/journal.pone.0091404

Zhang, J. D., & Wiemann, S. (2009). KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics, 25(11), 1470-1471. doi:10.1093/bioinformatics/btp167

Zienolddiny, S., & Skaug, V. (2012). Single nucleotide polymorphisms as susceptibility, prognostic, and therapeutic markers of nonsmall cell lung cancer. Lung Cancer: Targets and Therapy, 3, 1–14. Retrieved from http://search.proquest.com/professional/docview/1032568508?accountid=138535 \nhttp://www.dovepress.com/getfile.php?fileID=11728 LA – eng

APPENDICES

APPENDIX A: EXAMPLE JSON OUTPUT OF ENSEMBL VEP REST SERVICE TO FETCH VARIANT CONSEQUENCES BASED ON A VARIATION IDENTIFIER (SHORTENED)

http://rest.ensembl.org/vep/human/id/rs116035550?content-type=application/json

The following JSON output is the result of REST URL above. It is shortened. To see full result, you can use following URL.

http://rest.ensembl.org/documentation/info/vep_id_get


```
[
  {
    "colocated_variants": [
      {
        "aa_maf": 0,
        "ea_maf": 0.000116,
        "end": 212464,
        "seq_region_name": "11",
        "somatic": 0,
        "strand": 1,
        "aa_allele": "A",
        "id": "rs116035550",
        "ea_allele": "A",
        "allele_string": "G/A/C",
```

```
    "start": 212464
  }
],
"assembly_name": "GRCh38",
"end": 212464,
"seq_region_name": "11",
"strand": 1,
"transcript_consequences": [
  {
    "gene_id": "ENSG00000142082",
    "distance": 3681,
    "variant_allele": "A",
    "biotype": "nonsense_mediated_decay",
    "gene_symbol_source": "HGNC",
    "consequence_terms": [
      "downstream_gene_variant"
    ],
    "strand": -1,
    "hgnc_id": "HGNC:14931",
    "gene_symbol": "SIRT3",
    "transcript_id": "ENST00000529937"
  },
  {
    "gene_id": "ENSG00000142082",
    "distance": 3681,
    "variant_allele": "C",
    "biotype": "nonsense_mediated_decay",
    "gene_symbol_source": "HGNC",
```

```
      "consequence_terms": [
        "downstream_gene_variant"
      ],
      "strand": -1,
      "hgnc_id": "HGNC:14931",
      "gene_symbol": "SIRT3",
      "transcript_id": "ENST00000529937"
    },
    {
      "gene_id": "ENSG00000177963",
      "distance": 1802,
      "variant_allele": "A",
      "biotype": "nonsense_mediated_decay",
      "gene_symbol_source": "HGNC",
      "consequence_terms": [
        "downstream_gene_variant"
      ],
      "strand": 1,
      "hgnc_id": "HGNC:29550",
      "gene_symbol": "RIC8A",
      "transcript_id": "ENST00000526982"
    },
    {
      "gene_id": "ENSG00000177963",
      "distance": 1802,
      "variant_allele": "C",
      "biotype": "nonsense_mediated_decay",
      "gene_symbol_source": "HGNC",
      "consequence_terms": [
```

```
        "downstream_gene_variant"
      ],
      "strand": 1,
      "hgnc_id": "HGNC:29550",
      "gene_symbol": "RIC8A",
      "transcript_id": "ENST00000526982"
    },

     ...

    {
      "variant_allele": "A",
      "cdna_end": 1343,
      "polyphen_score": 0.986,
      "codons": "Gaa/Aaa",
      "protein_end": 340,
      "strand": 1,
      "hgnc_id": "HGNC:29550",
      "amino_acids": "E/K",
      "gene_symbol": "RIC8A",
      "cdna_start": 1343,
      "transcript_id": "ENST00000325207",
      "cds_start": 1018,
      "gene_id": "ENSG00000177963",
      "sift_prediction": "deleterious",
      "polyphen_prediction": "probably_damaging",
      "protein_start": 340,
      "biotype": "protein_coding",
      "gene_symbol_source": "HGNC",
```

```
      "cds_end": 1018,
      "sift_score": 0.01,
      "consequence_terms": [
        "missense_variant"
      ]
    },

    ...

    {
      "gene_id": "ENSG00000177963",
      "distance": 406,
      "variant_allele": "C",
      "biotype": "protein_coding",
      "gene_symbol_source": "HGNC",
      "consequence_terms": [
        "upstream_gene_variant"
      ],
      "strand": 1,
      "hgnc_id": "HGNC:29550",
      "gene_symbol": "RIC8A",
      "transcript_id": "ENST00000529275"
    }
  ],
  "id": "rs116035550",
  "allele_string": "G/A/C",
  "most_severe_consequence": "missense_variant",
  "start": 212464
}
```

```
]
```

APPENDIX B: COMMON SUB-GRAPHS OF 1st CASE STUDY

The Table 12 shows 27 common sub-graphs (nodes & edges) found at the end of Case Study 1.

Table 12. Common sub-graphs of hsa05200 and hsa05212.

Freq.	Entry1	Relation	Entry2	Type
2	[path:hsa04151] PI3K-Akt signaling pathway			Node
2	[path:hsa04110] Cell cycle			Node
2	[hsa:10928] RALBP1, RIP1, RLIP1, RLIP76	--| (PPrel: inhibition)	[hsa:5879 hsa:5880 hsa:5881 hsa:998] RAC1, Rac-1, TC-25, p21-Rac1...	Edge
2	[path:hsa04210] Apoptosis			Node
2	[path:hsa04630] Jak-STAT signaling pathway			Node
2	[hsa:5898 hsa:5899] RALA, RAL...	--> (PPrel: activation)	[hsa:5337] PLD1	Edge
2	[hsa:10000 hsa:207 hsa:208] AKT3, MPPH, PKB-GAMMA, PKBG, PRKBG, RAC-PK-gamma, RAC-gamma, STK-2...	--| +p (PPrel: inhibition, phosphorylation)	[hsa:842] CASP9, APAF-3, APAF3, ICE-LAP6, MCH6, PPP1R56	Edge
2	[hsa:1950] EGF, HOMG4, URG	--> (PPrel: activation)	[hsa:1956] EGFR, ERBB, ERBB1, HER1, PIG61, mENA	Edge
2	[path:hsa04115] p53 signaling pathway			Node

2	[hsa:572] BAD, BBC2, BCL2L8	--| (PPrel: inhibition)	[hsa:598] BCL2L1, BCL-XL/S, BCL2L, BCLX, BCLXL, BCLXS, Bcl-X, PPP1R52, bcl-xL, bcl-xS	Edge
2	[hsa:4087 hsa:4088] SMAD2, JV18, JV18-1, MADH2, MADR2, hMAD-2, hSMAD2...	--- (PPrel: binding/association)	[hsa:4089] SMAD4, DPC4, JIP, MADH4, MYHRS	Edge
2	[hsa:10000 hsa:207 hsa:208] AKT3, MPPH, PKB-GAMMA, PKBG, PRKBG, RAC-PK-gamma, RAC-gamma, STK-2...	--> +p (PPrel: activation, phosphorylation)	[hsa:1147 hsa:3551 hsa:8517] CHUK, IKBKA, IKK-alpha, IKK1, IKKA, NFKBIKA, TCF16...	Edge
2	[hsa:10000 hsa:207 hsa:208] AKT3, MPPH, PKB-GAMMA, PKBG, PRKBG, RAC-PK-gamma, RAC-gamma, STK-2...	--| +p (PPrel: inhibition, phosphorylation)	[hsa:572] BAD, BBC2, BCL2L8	Edge
2	[cpd:C05981] C05981 Phosphatidylinositol-3,4,5-trisphosphate;	--> (PCrel: activation)	[hsa:10000 hsa:207 hsa:208] AKT3, MPPH, PKB-GAMMA, PKBG, PRKBG, RAC-PK-gamma, RAC-gamma, STK-2...	Edge
2	[hsa:7040 hsa:7042 hsa:7043] TGFB1, CED, DPD1, LAP, TGFB, TGFbeta...	--> (PPrel: activation)	[Group: hsa:7046, hsa:7048]	Edge
2	[hsa:7039] TGFA, TFGA	--> (PPrel: activation)	[hsa:1956] EGFR, ERBB, ERBB1, HER1, PIG61, mENA	Edge
2	[hsa:3716] JAK1, JAK1A, JAK1B, JTK3	--> +p (PPrel: activation, phosphorylation)	[hsa:6772] STAT1, CANDF7, ISGF-3, STAT91	Edge
2	[hsa:3716] JAK1, JAK1A, JAK1B, JTK3	--> +p (PPrel: activation, phosphorylation)	[hsa:6774] STAT3, APRF, HIES	Edge
2	[hsa:5925] RB1, OSRC, PPP1R130, RB, p105-Rb, pRb, pp110	-+- (PPrel: dissociation)	[hsa:1869 hsa:1870 hsa:1871] E2F1, E2F-1, RBAP1, RBBP3, RBP3...	Edge
2	[hsa:5900] RALGDS, RGDS, RGF, RalGEF	--> (PPrel: activation)	[hsa:5898 hsa:5899] RALA, RAL...	Edge

2	[hsa:23533 hsa:5290 hsa:5291 hsa:5293 hsa:5294 hsa:5295 hsa:5296 hsa:8503] PIK3R5, F730038I15Rik, FOAP-2, P101-PI3K, p101...	--> (PCrel: activation)	[cpd:C05981] C05981 Phosphatidylinositol-3,4,5-trisphosphate;	Edge
2	[hsa:5898 hsa:5899] RALA, RAL...	--> (PPrel: activation)	[hsa:10928] RALBP1, RIP1, RLIP1, RLIP76	Edge
2	[path:hsa04010] MAPK signaling pathway			Node
2	[path:hsa04350] TGF-beta signaling pathway			Node
2	[path:hsa04370] VEGF signaling pathway			Node
2	[path:hsa05212] TITLE:Pancreatic cancer			Node
2	[hsa:5337] PLD1	--> (PCrel: activation)	[cpd:C00416] C00416 Phosphatidic acid	Edge

APPENDIX C: COMMON SUB-GRAPHS OF 2nd CASE STUDY (SHORTENED)

The Table 13 shows the most frequent 10 of 389 common sub-graphs (nodes & edges) found at the end of Case Study 2.

Table 13. The most frequent 10 of 389 common nodes and edges of 39 pathways related with Prostate Cancer, and their matching pathways.

Freq.	Entry1	Relation	Entry2	Type	Parent Pathways
21	[cpd:C00076] C00076 Calcium ion			Node	path:hsa04012 ErbB signaling pathway path:hsa04014 Ras signaling pathway path:hsa04015 Rap1 signaling pathway path:hsa04022 cGMP-PKG signaling pathway path:hsa04024 cAMP signaling pathway path:hsa04064 NF-kappa B signaling pathway path:hsa04260 Cardiac muscle contraction path:hsa04261 Adrenergic signaling in cardiomyocytes path:hsa04360 Axon guidance path:hsa04650 Natural killer cell mediated cytotoxicity path:hsa04724 Glutamatergic synapse

			path:hsa04961 Endocrine and other factor-regulated calcium reabsorption path:hsa04970 Salivary secretion path:hsa04972 Pancreatic secretion path:hsa04974 Protein digestion and absorption path:hsa04978 Mineral absorption path:hsa05142 Chagas disease (American trypanosomiasis) path:hsa05166 HTLV-I infection path:hsa05410 Hypertrophic cardiomyopathy (HCM) path:hsa05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC) path:hsa05414 Dilated cardiomyopathy
17	[path:hsa04020] Calcium signaling pathway	Node	path:hsa04012 ErbB signaling pathway path:hsa04014 Ras signaling pathway path:hsa04015 Rap1 signaling pathway path:hsa04020 Calcium signaling pathway path:hsa04022 cGMP-PKG signaling pathway path:hsa04024 cAMP signaling pathway path:hsa04064 NF-kappa B signaling pathway

			path:hsa04066 HIF-1 signaling pathway
			path:hsa04260 Cardiac muscle contraction
			path:hsa04261 Adrenergic signaling in cardiomyocytes
			path:hsa04650 Natural killer cell mediated cytotoxicity
			path:hsa04660 T cell receptor signaling pathway
			path:hsa04666 Fc gamma R-mediated phagocytosis
			path:hsa04724 Glutamatergic synapse
			path:hsa04970 Salivary secretion
			path:hsa04972 Pancreatic secretion
			path:hsa05142 Chagas disease (American trypanosomiasis)
13	[path:hsa04010] MAPK signaling pathway	Node	path:hsa04012 ErbB signaling pathway
			path:hsa04014 Ras signaling pathway
			path:hsa04015 Rap1 signaling pathway
			path:hsa04020 Calcium signaling pathway
			path:hsa04022 cGMP-PKG signaling pathway
			path:hsa04066 HIF-1 signaling pathway
			path:hsa04151 PI3K-Akt signaling pathway

			path:hsa04360 Axon guidance
			path:hsa04630 Jak-STAT signaling pathway
			path:hsa04650 Natural killer cell mediated cytotoxicity
			path:hsa04660 T cell receptor signaling pathway
			path:hsa04668 TNF signaling pathway
			path:hsa05166 HTLV-I infection
12	[cpd:C01330] C01330 Sodium ion	Node	path:hsa04024 cAMP signaling pathway path:hsa04260 Cardiac muscle contraction path:hsa04261 Adrenergic signaling in cardiomyocytes path:hsa04724 Glutamatergic synapse path:hsa04961 Endocrine and other factor-regulated calcium reabsorption path:hsa04970 Salivary secretion path:hsa04972 Pancreatic secretion path:hsa04974 Protein digestion and absorption path:hsa04978 Mineral absorption path:hsa05410 Hypertrophic cardiomyopathy (HCM) path:hsa05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC)

			path:hsa05414 Dilated cardiomyopathy
11	[path:hsa04151] PI3K-Akt signaling pathway	Node	path:hsa04012 ErbB signaling pathway
			path:hsa04014 Ras signaling pathway
			path:hsa04015 Rap1 signaling pathway
			path:hsa04022 cGMP-PKG signaling pathway
			path:hsa04024 cAMP signaling pathway
			path:hsa04066 HIF-1 signaling pathway
			path:hsa04151 PI3K-Akt signaling pathway
			path:hsa04261 Adrenergic signaling in cardiomyocytes
			path:hsa04630 Jak-STAT signaling pathway
			path:hsa04660 T cell receptor signaling pathway
			path:hsa04668 TNF signaling pathway
11	[cpd:C00238] C00238 Potassium ion	Node	path:hsa04022 cGMP-PKG signaling pathway
			path:hsa04024 cAMP signaling pathway
			path:hsa04260 Cardiac muscle contraction
			path:hsa04261 Adrenergic signaling in cardiomyocytes
			path:hsa04724 Glutamatergic synapse
			path:hsa04961 Endocrine and other factor-

				regulated calcium reabsorption
				path:hsa04970 Salivary secretion
				path:hsa04972 Pancreatic secretion
				path:hsa04974 Protein digestion and absorption
				path:hsa04978 Mineral absorption
				path:hsa05166 HTLV-I infection
11	[path:hsa04210] Apoptosis		Node	path:hsa04020 Calcium signaling pathway
				path:hsa04022 cGMP-PKG signaling pathway
				path:hsa04024 cAMP signaling pathway
				path:hsa04064 NF-kappa B signaling pathway
				path:hsa04151 PI3K-Akt signaling pathway
				path:hsa04630 Jak-STAT signaling pathway
				path:hsa04650 Natural killer cell mediated cytotoxicity
				path:hsa04668 TNF signaling pathway
				path:hsa05020 Prion diseases
				path:hsa05142 Chagas disease (American trypanosomiasis)
				path:hsa05166 HTLV-I infection

8	[hsa:5604 hsa:5605] MAP2K1, CFC3, MAPKK1, MEK1, MKK1, PRKMK1...	--> +p (PPrel: activation, phosphorylation)	[hsa:5594 hsa:5595] MAPK1, ERK, ERK-2, ERK2, ERT1, MAPK2, P42MAPK, PRKM1, PRKM2, p38, p40, p41, p41mapk, p42-MAPK...	Edge	path:hsa04012 ErbB signaling pathway path:hsa04014 Ras signaling pathway path:hsa04015 Rap1 signaling pathway path:hsa04022 cGMP-PKG signaling pathway path:hsa04024 cAMP signaling pathway path:hsa04151 PI3K-Akt signaling pathway path:hsa04650 Natural killer cell mediated cytotoxicity path:hsa04660 T cell receptor signaling pathway
8	[cpd:C05981] C05981 Phosphatidylinositol-3,4,5-trisphosphate			Node	path:hsa04012 ErbB signaling pathway path:hsa04144 Endocytosis path:hsa04151 PI3K-Akt signaling pathway path:hsa04530 Tight junction path:hsa04650 Natural killer cell mediated cytotoxicity path:hsa04660 T cell receptor signaling pathway path:hsa05142 Chagas disease (American trypanosomiasis) path:hsa05166 HTLV-I infection
8	[cpd:C01245] C01245 Inositol 1,4,5-			Node	path:hsa01100 Metabolic pathways

trisphosphate	path:hsa04012 ErbB signaling pathway path:hsa04014 Ras signaling pathway path:hsa04020 Calcium signaling pathway path:hsa04064 NF-kappa B signaling pathway path:hsa04650 Natural killer cell mediated cytotoxicity path:hsa04961 Endocrine and other factor-regulated calcium reabsorption path:hsa05142 Chagas disease (American trypanosomiasis)

APPENDIX D: COMMON SUB-GRAPHS OF 3rd CASE STUDY (SHORTENED)

The Table 14 shows the most frequent 10 of 251 common sub-graphs (nodes & edges) found at the end of Case Study 3.

Table 14. The most frequent 10 of 251 common nodes and edges of 22 pathways related with JRA, and their matching pathways.

Freq.	Entry1	Relation	Entry2	Type	Parent Pathways
8	[path:hsa04151] PI3K-Akt signaling pathway			Node	path:hsa05223 Non-small cell lung cancer path:hsa05222 Small cell lung cancer path:hsa05220 Chronic myeloid leukemia path:hsa04611 Platelet activation path:hsa04510 Focal adhesion path:hsa04380 Osteoclast differentiation path:hsa04071 Sphingolipid signaling pathway path:hsa04014 Ras signaling pathway
8	[path:hsa04010] MAPK signaling pathway			Node	path:hsa05223 Non-small cell lung cancer path:hsa05220 Chronic myeloid leukemia

			path:hsa04664 Fc epsilon RI signaling pathway path:hsa04510 Focal adhesion path:hsa04380 Osteoclast differentiation path:hsa04151 PI3K-Akt signaling pathway path:hsa04071 Sphingolipid signaling pathway path:hsa04014 Ras signaling pathway
6	[path:hsa04020] Calcium signaling pathway	Node	path:hsa05223 Non-small cell lung cancer path:hsa04730 Long-term depression path:hsa04666 Fc gamma R-mediated phagocytosis path:hsa04611 Platelet activation path:hsa04380 Osteoclast differentiation path:hsa04014 Ras signaling pathway
6	[cpd:C01245] C01245 Inositol 1,4,5-trisphosphate	Node	path:hsa05223 Non-small cell lung cancer path:hsa04730 Long-term depression path:hsa04664 Fc epsilon RI signaling

					pathway path:hsa04380 Osteoclast differentiation path:hsa04014 Ras signaling pathway path:hsa01100 Metabolic pathways
6	[hsa:2885] GRB2, ASH, EGFRBP-GRB2, Grb3-3, MST084, MSTP084, NCKAP2	--> (PPrel: activation)	[hsa:6654 hsa:6655] SOS1, GF1, GGF1, GINGF, HGF, NS4...	Edge	path:hsa05223 Non-small cell lung cancer path:hsa05220 Chronic myeloid leukemia path:hsa04664 Fc epsilon RI signaling pathway path:hsa04510 Focal adhesion path:hsa04151 PI3K-Akt signaling pathway path:hsa04014 Ras signaling pathway
6	[hsa:5604 hsa:5605] MAP2K1, CFC3, MAPKK1, MEK1, MKK1, PRKMK1...	--> +p (PPrel: activation, phosphorylation)	[hsa:5594 hsa:5595] MAPK1, ERK, ERK-2, ERK2, ERT1, MAPK2, P42MAPK, PRKM1, PRKM2, p38, p40, p41, p41mapk, p42-MAPK...	Edge	path:hsa05223 Non-small cell lung cancer path:hsa05220 Chronic myeloid leukemia path:hsa04664 Fc epsilon RI signaling pathway path:hsa04151 PI3K-Akt signaling pathway path:hsa04071 Sphingolipid signaling

			pathway path:hsa04014 Ras signaling pathway
5	[path:hsa04110] Cell cycle	Node	path:hsa05223 Non-small cell lung cancer path:hsa05222 Small cell lung cancer path:hsa05220 Chronic myeloid leukemia path:hsa04510 Focal adhesion path:hsa04151 PI3K-Akt signaling pathway
5	[cpd:C05981] C05981 Phosphatidylinositol-3,4,5-trisphosphate	Node	path:hsa05223 Non-small cell lung cancer path:hsa05220 Chronic myeloid leukemia path:hsa04664 Fc epsilon RI signaling pathway path:hsa04151 PI3K-Akt signaling pathway path:hsa04144 Endocytosis
5	[cpd:C00076] C00076 Calcium ion	Node	path:hsa05223 Non-small cell lung cancer path:hsa04664 Fc epsilon RI signaling pathway

			path:hsa04611 Platelet activation path:hsa04380 Osteoclast differentiation path:hsa04014 Ras signaling pathway
4	[path:hsa04115] p53 signaling pathway	Node	path:hsa05223 Non-small cell lung cancer path:hsa05222 Small cell lung cancer path:hsa05220 Chronic myeloid leukemia path:hsa04151 PI3K-Akt signaling pathway

APPENDIX E: MATCHING SNPS OF ENSEMBL GENE IDS IN THE 2nd CASE STUDY

The Table 15 shows the matching SNPs of 88 Ensembl gene ids in Case Study 2.

Table 15. Matching SNPs of 88 Ensembl gene ids related with Prostate Cancer.

ESNG ID	Matching SNPs
ENSG00000127616	rs17001078
ENSG00000170579	rs280986
ENSG00000183023	rs10195113
ENSG00000196566	rs10788555
ENSG00000170858	rs2296370
ENSG00000258405	rs2115101
ENSG00000229298	rs6475584
ENSG00000257839	rs1433369
ENSG00000235495	rs6708126
ENSG00000143196	rs12733054
ENSG00000117501	rs16863955
ENSG00000137473	rs964130
ENSG00000155792	rs7010457
ENSG00000213973	rs12980509
ENSG00000139289	rs1433369
ENSG00000168702	rs11885120
ENSG00000226744	rs1974562
ENSG00000255872	rs11790106
ENSG00000259282	rs3812906
ENSG00000237356	rs7152946
ENSG00000196503	rs12644498
ENSG00000043039	rs11221701
ENSG00000223761	rs10788555
ENSG00000241073	rs501700
ENSG00000164362	rs2853668
ENSG00000107338	rs11790106

ENSG00000225913	rs10788555
ENSG00000154654	rs2826802
ENSG00000174780	rs12644498
ENSG00000230448	rs1470494
ENSG00000183117	rs17432165, rs7843255, rs766045
ENSG00000156875	rs501700
ENSG00000118263	rs17284653
ENSG00000249699	rs11729739
ENSG00000109079	rs3093679
ENSG00000080298	rs7034430
ENSG00000157168	rs10954845
ENSG00000269509	rs12980509
ENSG00000070669	rs1974562
ENSG00000112419	rs7775829
ENSG00000172554	rs13011951
ENSG00000257453	rs1433369
ENSG00000162897	rs12119983
ENSG00000185594	rs3812906
ENSG00000105229	rs3760903
ENSG00000106078	rs17799219
ENSG00000271096	rs2711134
ENSG00000232837	rs10854395
ENSG00000162373	rs17375010
ENSG00000151693	rs2666205
ENSG00000064270	rs4782945
ENSG00000082684	rs2120806
ENSG00000171956	rs7183502
ENSG00000186094	rs17375010
ENSG00000112530	rs9347691
ENSG00000135678	rs1965340
ENSG00000222206	rs11729739
ENSG00000109083	rs3093679
ENSG00000162402	rs17111584
ENSG00000235751	rs918285

ENSG00000115306	rs7584223
ENSG00000174891	rs6779266
ENSG00000105926	rs2711134
ENSG00000213981	rs6704731
ENSG00000147316	rs2442602
ENSG00000237498	rs11126869
ENSG00000232337	rs197265
ENSG00000231557	rs1379015
ENSG00000099810	rs6475584
ENSG00000176204	rs6747704
ENSG00000279966	rs12980509
ENSG00000198821	rs6686571
ENSG00000187231	rs17363393
ENSG00000205830	rs11729739
ENSG00000264545	rs6475584
ENSG00000253452	rs10106027
ENSG00000198633	rs2115101
ENSG00000171735	rs4908656
ENSG00000154978	rs17673975
ENSG00000196353	rs9848588
ENSG00000187391	rs918285
ENSG00000224467	rs2194505
ENSG00000169306	rs1454186
ENSG00000184005	rs517036
ENSG00000157087	rs6774902
ENSG00000200310	rs12201462
ENSG00000184903	rs17400029
ENSG00000091879	rs2442602

APPENDIX F: MATCHING ENSEMBL GENE IDS OF KEGG GENE IDS IN 2nd CASE STUDY

The Table 16 shows the matching Ensembl gene ids of 64 KEGG gene ids in Case Study 2.

Table 16. Matching Ensembl gene ids of 64 KEGG gene ids related with Prostate Cancer.

KEGG Gene ID	Matching ENSG ID	Description
hsa:10200 *	ENSG00000105926	M-phase phosphoprotein 6
hsa:51678	ENSG00000105926	membrane protein, palmitoylated 6 (MAGUK p55 subfamily member 6)
hsa:491	ENSG00000157087	ATPase, Ca++ transporting, plasma membrane 2
hsa:493 *	ENSG00000157087	ATPase, Ca++ transporting, plasma membrane 4
hsa:145946	ENSG00000185594	spermatogenesis associated 8
hsa:221409 *	ENSG00000185594	spermatogenesis associated, serine-rich 1
hsa:100652749	ENSG00000259282	SPATA8 antisense RNA 1 (head to head)
hsa:11141	ENSG00000169306	interleukin 1 receptor accessory protein-like 1
hsa:131034	ENSG00000196353	copine IV
hsa:132946	ENSG00000196503	ADP-ribosylation factor-like 9
hsa:135138	ENSG00000112530	PARK2 co-regulated
hsa:1368	ENSG00000135678	carboxypeptidase M
hsa:147658	ENSG00000198633	zinc finger protein 534
hsa:147660	ENSG00000258405	zinc finger protein 578
hsa:1805	ENSG00000143196	Dermatopontin
hsa:22822	ENSG00000139289	pleckstrin homology-like domain, family A, member 1
hsa:23242	ENSG00000106078	cordon-bleu WH2 repeat protein
hsa:23261	ENSG00000171735	calmodulin binding transcription activator 1
hsa:23358	ENSG00000162402	ubiquitin specific peptidase 24
hsa:256435	ENSG00000184005	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 3
hsa:27023	ENSG00000171956	forkhead box B1
hsa:285	ENSG00000091879	angiopoietin 2

hsa:3084	ENSG00000157168	neuregulin 1
hsa:440	ENSG00000070669	asparagine synthetase (glutamine-hydrolyzing)
hsa:4507	ENSG00000099810	methylthioadenosine phosphorylase
hsa:4685	ENSG00000154654	neural cell adhesion molecule 2
hsa:51319	ENSG00000174891	arginine/serine-rich coiled-coil 1
hsa:51588	ENSG00000105229	protein inhibitor of activated STAT, 4
hsa:53353	ENSG00000168702	low density lipoprotein receptor-related protein 1B
hsa:54221	ENSG00000172554	syntrophin, gamma 2
hsa:54437	ENSG00000082684	sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5B
hsa:5991	ENSG00000080298	regulatory factor X, 3 (influences HLA class II expression)
hsa:64478	ENSG00000183117	CUB and Sushi multiple domains 1
hsa:6461	ENSG00000107338	Src homology 2 domain containing adaptor protein B
hsa:64645	ENSG00000156875	hippocampus abundant transcript 1
hsa:64798	ENSG00000155792	DEP domain containing MTOR-interacting protein
hsa:6546	ENSG00000183023	solute carrier family 8 (sodium/calcium exchanger), member 1
hsa:6597	ENSG00000127616	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
hsa:6711	ENSG00000115306	spectrin, beta, non-erythrocytic 1
hsa:6731	ENSG00000174780	signal recognition particle 72kDa
hsa:7015	ENSG00000164362	telomerase reverse transcriptase
hsa:7126	ENSG00000109079	tumor necrosis factor, alpha-induced protein 1 (endothelial)
hsa:7652	ENSG00000213973	zinc finger protein 99
hsa:79648	ENSG00000147316	microcephalin 1
hsa:79656	ENSG00000162373	BEN domain containing 5
hsa:80059	ENSG00000176204	leucine rich repeat transmembrane neuronal 4
hsa:80133	ENSG00000117501	maestro heat-like repeat family member 9
hsa:81552	ENSG00000154978	vesicular, overexpressed in cancer, prosurvival protein 1
hsa:83894	ENSG00000137473	tetratricopeptide repeat domain 29

hsa:83943	ENSG00000184903	IMP2 inner mitochondrial membrane peptidase-like (S. cerevisiae)
hsa:83953	ENSG00000162897	Fc receptor, IgA, IgM, high affinity
hsa:84871	ENSG00000186094	ATP/GTP binding protein-like 4
hsa:8538	ENSG00000043039	BARX homeobox 2
hsa:8609	ENSG00000118263	Kruppel-like factor 7 (ubiquitous)
hsa:8853	ENSG00000151693	ArfGAP with SH3 domain, ankyrin repeat and PH domain 2
hsa:90410	ENSG00000109083	intraflagellar transport 20
hsa:91404	ENSG00000187231	SEC14 and spectrin domains 1
hsa:919	ENSG00000198821	CD247 molecule
hsa:9229	ENSG00000170579	discs, large (Drosophila) homolog-associated protein 1
hsa:9749	ENSG00000112419	phosphatase and actin regulator 2
hsa:9863	ENSG00000187391	membrane associated guanylate kinase, WW and PDZ domain containing 2
hsa:9914	ENSG00000064270	ATPase, Ca++ transporting, type 2C, member 2

* Incorrect conversion by BioDB web service.

APPENDIX G: COMPARISON OF EXISTING APPLICATIONS

The Table 17 shows the comparison between related works and our application.

Table 17. Comparison between related works and our application.

	SNP-Gene Mapping	**Gene Identifier Conversion**	**Gene-Pathway Mapping**	**Pathway Alignment**	**Reading KGML**	**Pathway Visualization**	**Interactive Graph**	**GUI**	**Web App**	**Local Database Independency**
Our Application	+	+ [1]	+	+	+	+	+	+	+	+
KEGG	-	+ [2]	+ [3]	-	+	+	-	-	+	-
DAVID	-	+	+	-	-	+	-	+	+	-
SNPnexus	+	-	-	-	-	-	-	+	+	-

Cytoscape	-	-	-	-	-	+	+	+	-	+
CytoKegg	-	-	+ [4]	-	+	+	+	+	-	+
KEGGScape	-	-	-	-	+	+	+	+	-	+
ClueGo	-	-	+ [6]	-	+ [7]	+	+	+	-	+
CyKEGGParser	-	-	-	-	+	+	+	+	-	+
KGMLReader	-	-	-	-	+	+	+	+	-	+
Genoscape	-	-	+ [4]	-	+	+	+	+	-	+
KEGGgraph	-	-	-	- [5]	+	+	+	-	-	+
VisANT	-	-	+	-	-	+	+	+	-	-
KGML-ED	-	-	-	-	+	+	+	+	-	+

PANOGA	+	+	+	-	+	+	-	+	+	+
PathVisio	- [8]	-	+	-	-	+	+	+	-	-

1 From ENSG to KEGG Gene Id and from Gene Symbol to KEGG Gene Id
2 KEGG has appropriate REST services for conversion between KEGG gene id and ncbi-gi | ncbi-geneid | uniprot ids.
3 KEGG has appropriate REST services
4 CytoKegg and Genoscape map gene expression profiles onto KEGG pathways
5 But, KEGGgraph has a collection of tools to analyze, **dissect** and visualize graphs of KEGG pathways. (Zhang & Wiemann, 2009)
6 ClueGO visualizes the biological terms for large clusters of genes in a functionally grouped network. [7]
7 ClueGO is updated with the files from KEGG and some other databases. [8]
8 But, integration of genetic variation data can be provided by newly developed plug-ins.

[7] http://apps.cytoscape.org/apps/cluego
[8] http://apps.cytoscape.org/apps/cluego

APPENDIX H: PREVIEWS OF PROJECT CONFIGURATION FILES

The following figures show the project configuration files.

Main.html | Main.gwt.xml | pom.xml [Tez-web] | persistence.xml | glassfish-resources.xml

Source | Graph | Effective | History

```
<?xml version="1.0" encoding="UTF-8"?>
<project
    xmlns="http://maven.apache.org/POM/4.0.0"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.0.xsd">

    <modelVersion>4.0.0</modelVersion>
    <parent>
        <artifactId>Tez</artifactId>
        <groupId>tr.com.ersoy</groupId>
        <version>1.0-SNAPSHOT</version>
    </parent>

    <groupId>tr.com.ersoy</groupId>
    <artifactId>Tez-web</artifactId>
    <version>1.0-SNAPSHOT</version>
    <packaging>war</packaging>

    <name>Tez-web</name>

    <properties>
        <!-- Convenience property to set the GWT version -->
        <gwtVersion>2.5.1</gwtVersion>
        <gxt.version>3.0.1</gxt.version>
        <!-- GWT needs at least java 1.5 -->
        <webappDirectory>${project.build.directory}/${project.build.finalName}</webappDirectory>

        <endorsed.dir>${project.build.directory}/endorsed</endorsed.dir>
        <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
        <netbeans.hint.deploy.server>gfv3ee6</netbeans.hint.deploy.server>
    </properties>

    <dependencies>
        <dependency>
            <groupId>org.yworks</groupId>
            <artifactId>yfiles-gwt</artifactId>
            <version>1.2.0.5</version>
        </dependency>
        <dependency>
            <groupId>org.json</groupId>
            <artifactId>json</artifactId>
            <version>20140107</version>
        </dependency>
        <dependency>
            <groupId>com.sencha.gxt</groupId>
            <artifactId>gxt</artifactId>
            <version>${gxt.version}</version>
        </dependency>
        <dependency>
```

Figure 28. Preview of pom.xml: A configuration file of a maven project. Developers can define project dependencies like jar libraries, then all necessary libraries are download during project built.

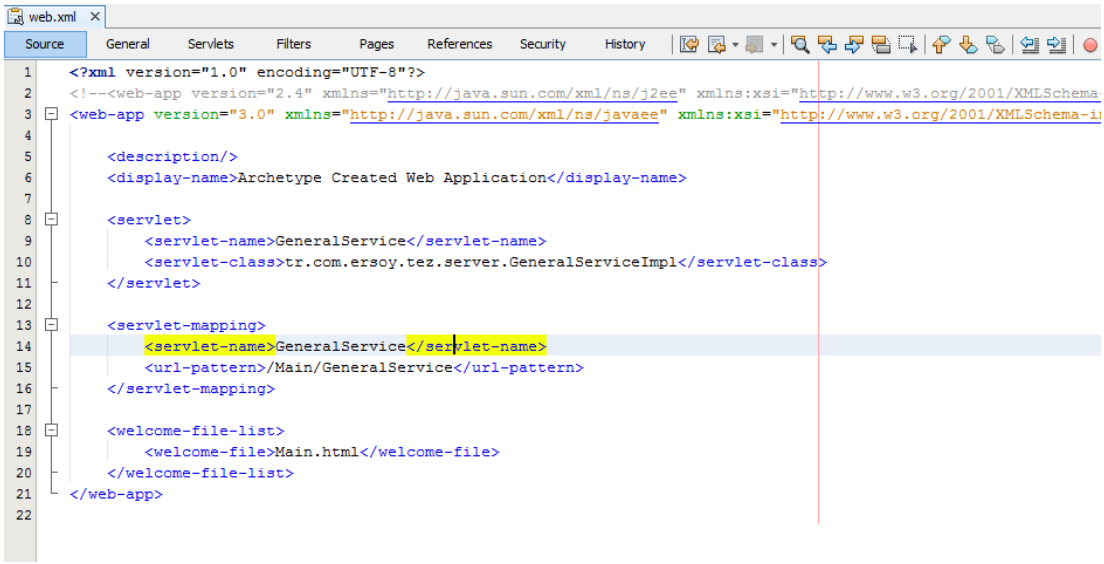

Figure 29. Preview of web.xml: A file where servlet configurations are defined.

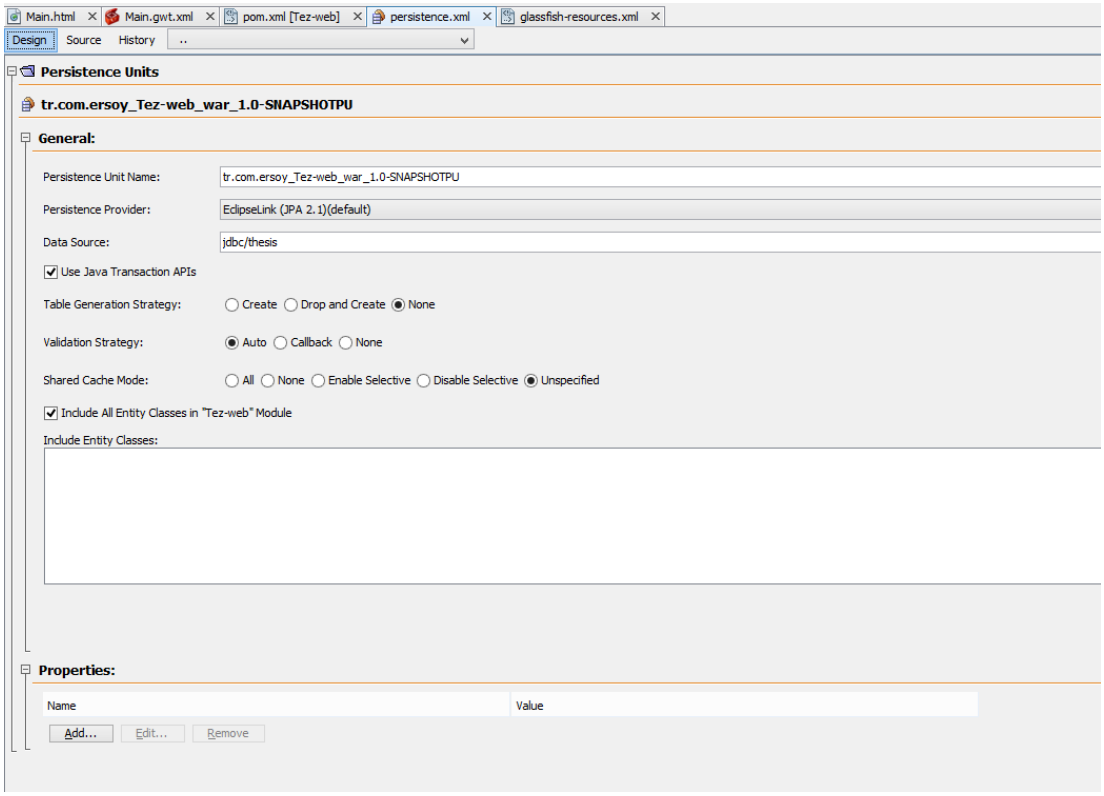

Figure 30. Preview of persistence.xml (design view): A file where database settings are configured.

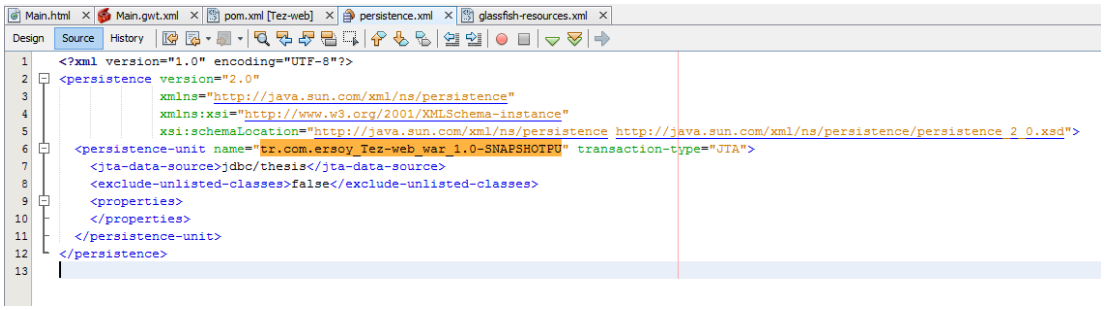

Figure 31. Preview of persistence.xml (source view)

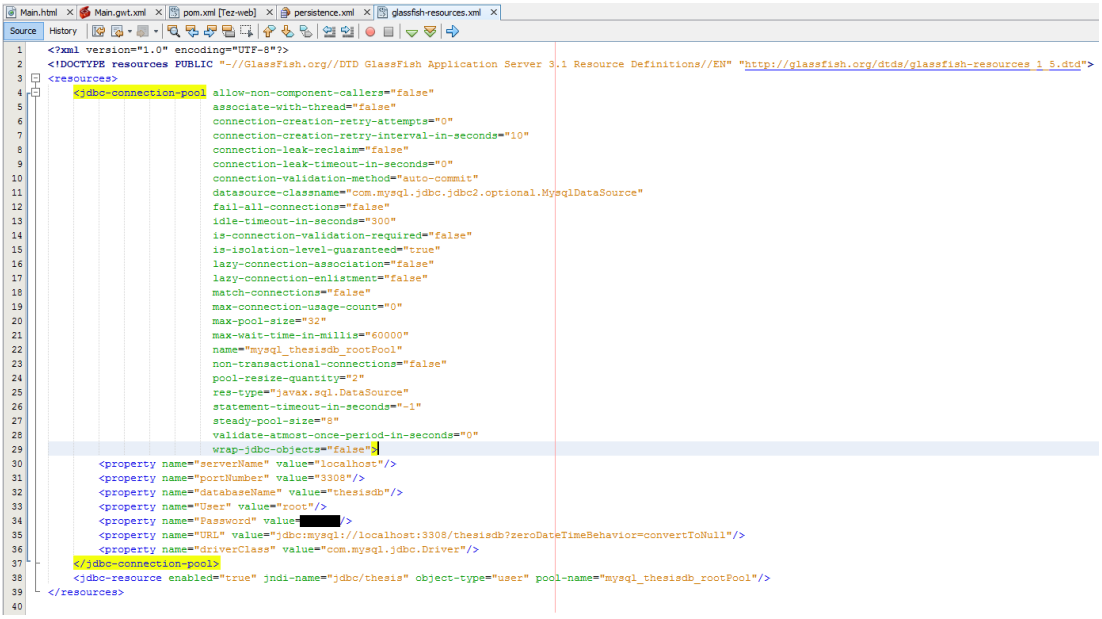

Figure 32. Preview of glassfish-resources.xml: A file where database connection info is defined for application server, GlassFish.

APPENDIX I: STRUCTURAL DETAILS OF MODEL CLASSES

The following tables show the fields of model classes to hold and transport data between client side and server side.

Table 18. RsIdGridModel to hold the data of rsId table.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
rsId	String	dbSNP id of SNPs
description	String	Detailed information

VariantConsequencesGridModel holds the data retrieved from Ensembl VEP REST API.

Table 19. VariantConsequencesGridModel to hold the data of variant consequences table.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
geneId	String	Ensembl Gene ID
distance	Integer	
variantAllele	String	
biotype	String	
geneSymbolSource	String	
strand	Integer	
hgncId	String	HGNC Gene ID

geneSymbol	String	RefSeq Gene Symbol
transcriptId	String	Transcript ID of gene

Table 20. KeggGeneGridModel to hold the data of KEGG gene ids table.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
geneId	String	KEGG Gene ID
description	String	Detailed information

Table 21. KeggPathwayGridModel to hold the data of KEGG pathways table.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
numberOfGenes	String	Number of input genes which are located on pathway.
pathwayId	String	KEGG pathway id
description	String	Detailed information.
Species	String	Species of the cell where pathway located
kgmlUrl	String	KGML file url of pathway
imageUrl	String	Image file url of pathway
publicUrl	String	Public link of pathway

Table 22. CommonEdgeGridModel to hold the data of pathways-commons table.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
usageCount	Integer	To hold the usage count of edge/node in pathways
graphics1Name	String	To hold the name of first entry of an edge
graphics2Name	String	To hold the name of second entry of an edge
relationTypeValue	String	To hold the relation as short character codes
relationTypeName	String	To hold the relation as a human-readable name
entry1Element	EnrtyElementModel	Another model to hold the details of first entry
entry2Element	EnrtyElementModel	Another model to hold the details of second entry

Table 23. EnrtyElementModel to hold the data of an "entry" element in KGML file.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
entryId	String	Unique id of entry in KGML file

Name	String	Name of entry
Type	String	Type of entry
Link	String	Hyperlink to KEGG website
graphicsList	List<GraphicsElementModel>	Graphical components of entry
componentList	List<ComponentElementModel>	The list of components if the entry has multiple sub-components

Table 24. GraphicsElementModel to hold the data of a "graphics" element in KGML file.

Field Name	**Field Type**	**Description**
pkId	Integer	To handle each model instance as a unique record
name	String	To hold the name of graphical component
fgcolor	String	To hold foreground color of graphical component
bgcolor	String	To hold the background color of graphical component
Type	String	To hold the shape type of graphical component
X	String	To hold the x coordinate of graphical component
Y	String	To hold the y coordinate of graphical component
width	String	To hold the width of graphical component
height	String	To hold the height of graphical component

Table 25. ComponentElementModel to hold the data of a "component" element in KGML file.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
componentId	String	To hold the ids of other entries. Generally used in the components of "group" type.

Table 26. OrganismModel to hold the data of Organism table in database.

Field Name	Field Type	Description
pkId	Integer	To handle each model instance as a unique record
orgId	String	Alternative and human readable unique identifier of records (O1, O2, …, O3215, etc.)
realId	String	
code	String	Short code of species (hsa, mmu, ggo, etc.)
species	String	Species of organism
taxonomy	String	Taxonomic information of organism
imageEmbedHtml	String	Image HTML of organism

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ■

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : Ersoy.....
Adı : Gökhan.....
Bölümü : Bioinformatics.....

TEZİN ADI (İngilizce) : A MULTI-LAYERED GRAPHICAL MODEL OF THE RELATION AMONG SNPS, GENES, AND PATHWAYS BASED ON..... ON SUBGRAPH SEARCH.....
.....
.....

TEZİN TÜRÜ : Yüksek Lisans ■ Doktora ☐

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ■
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullancılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....