INTEGRATION OF METU-SNP DATABASES VIA RDF FOR PI-SNP WEB SERVICE

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

BY

CEYHUN GEDİKOĞLU

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

JANUARY 2014

Approval of the thesis:

INTEGRATION OF METU-SNP DATABASES VIA RDF FOR PI-SNP WEB SERVICE

submitted by **CEYHUN GEDİKOĞLU** in partial fulfillment of the requirements for the degree of
Master of Science in Department of Health Informatics , Bioinformatics Program, Middle East Technical University by,

NAZİFE BAYKAL _____
Director, Graduate School of **Informatics Institute**

YEŞİM AYDIN SON _____
Chair, **Department of Health Informatics**

Assist. Prof. Dr. Yeşim Aydın Son _____
Supervisor, **Health Informatics Dept., METU**

Dr. Levent Çarkacıoğlu _____
Co-supervisor, **T.C. Merkez Bankası**

Examining Committee Members:

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

Assist. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU _____

Dr. Levent Çarkacıoğlu
T.C. Merkez Bankası _____

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

Assist. Prof. Dr. Didem Gökçay
Health Informatics Dept., METU _____

Date: 30.01.2014

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: CEYHUN GEDİKOĞLU

Signature :

ABSTRACT

INTEGRATION OF METU-SNP DATABASES VIA RDF FOR PI-SNP WEB SERVICE

Gedikoğlu, Ceyhun
M.S., Bioinformatics Program
Supervisor : Assist. Prof. Dr. Yeşim Aydın Son
Co-Supervisor : Dr. Levent Çarkacıoğlu
January 2014, 48 pages

Single Nucleotide Polymorphism (SNP) is a variation which occurs after a nucleotide mutates between members of a species or paired chromosomes in DNA sequence. SNP data is especially important for identifying genetic variations underlying complex diseases. The need for collection and service of this data under a standard format and globally normalized and structured metadata that houses the structured SNP data is becoming more important while recent advances in high-throughput genotyping technologies are resulting in data accumulation at large scales. This means every new research can result with new data and we need all the new data for our computations. Offline databases can only offer a collection of data but cannot provide access to updated information. So we have built an integrated (iSNP) database that is a regularly updated. This machine curated database that holds SNP and its associated metadata from publicly available databases under a structured standard format can be efficiently utilized within different applications. Also the adaptation of the METU-SNP desktop application(SNP prioritization tool for complex diseases) to the web environment, which is supported by the iSNP database is included as a part of the study. This study will help bioinformaticians from all over the world reach the METU-SNP application

with the upto date SNP information used in it via web environment.

Keywords: iSNP, METU-SNP, Semantic Web, Public Available Databases,SNP prioritization

ÖZ

PI-SNP WEB SERVİSİ İÇİN KAYNAK TANIM ÇERÇEVESİ YOLUYLA METU-SNP VERİTABANI ENTEGRASYONU

Gedikoğlu, Ceyhun
Yüksek Lisans, Biyoenformatik Programı
Tez Yöneticisi : Yard. Doç. Dr. Yeşim Aydın Son
Ortak Tez Yöneticisi : Dr. Levent Çarkacıoğlu

Ocak 2014, 48 sayfa

Tek nükleotit polimorfizmi bir türün elemanları veya bir insanın DNA sekansındaki eşlenmiş kromozomları arasındaki bir nükleotid mutasyonu sonrasındaki çeşitliliktir. SNP verisi komplike hastalıkların temelinde yatan genetik çeşitlilikleri tanımlamak için özellikle önemlidir. Bu verinin toplanması, servisi için ihtiyaç ve standart bir format altında, küresel normalize edilmiş ve biçimlendirilmiş SNP verisine evsahipliği yapan kılavuz bilgi zamanla daha önemli bir hal almaktadır. Öte yandan yüksek hızlı genotip teknolojilerindeki ilerlemeler büyük kapasiteli veri birikimlerine öncü olmaktadır. Bu demek oluyor ki; her yeni araştırma yeni veriler sağlayacaktır ve hesaplamalarımız için bu yeni verilere ihtiyacımız var. Çevrimdışı bir veritabanı güncel bilgiyi kaçırmamıza neden olabilir. Bu yüzden, entegre edilmiş, düzenli güncellenen otomatize edilmiş veritabanıyla iSNP'i oluşturduk. iSNP oluşturduğu SNP ve ilgili verileri halka açık uygun veritabanlarından standart biçimlendirilmiş bir halde farklı uygulamalar tarafından kullanılabilecek bir şekilde çekmektedir. Ayrıca METU-SNP masaüstü uygulamasının iSNP veritabanıyla da desteklenecek şekilde web ortamına entegre edilmesi çalışmalarımızın bir parçasını oluşturuyor. Bu çalışma dünyanın her tarafından araştırmacıların METU-SNP uygulamasına, güncel bilgiyi kullanarak

erişmesine olanak sağlayacaktır.

Anahtar Kelimeler: iSNP, METU-SNP, Semantik Web, Halka Açık Veritabanları, SNP önceliklendirilmesi

To My Family

ACKNOWLEDGMENTS

I am very grateful to my supervisor Assist. Prof. Dr. Yeşim Aydın Son and my co-supervisor Dr. Levent Çarkacıoğlu for many fruitful suggestions and discussions, as well as to my family and my fiance Fatma for their support and understanding at every stage of this study. I also would like to thank to my managers and colleagues for their priceless support and understanding.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

List of Algorithms

CHAPTER 1

INTRODUCTION

1.1 Motivation

Molecular biology data generated and collected every day is expanding exponentially due to the emerging high-throughput technologies, such as microarray and next generation sequencing technologies. The large scale projects completed on human genome sequencing and its annotation (Human Genome Project, 1000Genomes, HapMap, ENCODE etc.) have generated biological data, which cannot be analyzed with the traditional approaches of biology. So, the need for analysis, standardizing, keeping and properly retrieving big biological data become a challenge for bioinformaticians. As the information technologies become the key to help understand the raw biological data generated, web and database technologies has been utilized to manage the big biological data. New algorithms are developed, which requires less memory and low processing power. Today along with data management, providing access to up to date data is still one of the current research areas in bioinformatics.

Ongoing research on genomic variations aims to understand the underlying mechanisms of complex diseases, and the biological data generated on genomic variations is among the fastest expanding in the field of bioinformatics. Information related to SNPs and their associated meta-data can be found individually in many public databases. These databases serve data in a non-uniform format. In order to provide uniformity of SNP data and its associated meta-data.

In this study, we have build an integrated database structure called iSNP, in which SNP related data is collected from NCBI's dbSNP and Entrez Gene, HapMap, UCSC, Polyphen, Pathway Commons, GAD and GeneRIF-DO in RDF. In addition to development of the integrated database the METU-SNP desktop application has been redesigned as webservice and moved to the Linux environment, and connected tot he iSNP database developed for providing uptodate SNP information to support SNP prioritizaiton function of the METU-SNP application.

1.2 Genomic Variations and Genome Wide Association Studies

Single Nucleotide Polymorphisms (SNPs) are the most frequently observed genomic variations. Genome Wide Association Studies (GWAS) of SNPs and SNPs as ge-

nomic biomarkers are becoming more widespread with the potential to help for identifying genetic variations underlying complex diseases, as genotyping millions of SNPs in a short time, much lower cost is now possible with the microarray and advanced sequencing technologies [1] [2] [3]. National Center for Biotechnology Information (NCBI)'s current SNP Variation database (dbSNP build 138) holds about 45 million validated SNPs out of estimated 62 million potential SNPs within the human genome. The International HapMap Project (HapMap3) have genotyped approximately 6 million consensus SNPs between 11 different populations consistent adding the number on to the previously reported statistics from the earlier phases of the study.

Our understanding of the genetic etiology of human disease is still limited because there is an enormous number of genetic variations on the human genome, as well as the complex interplay of multiple genes and environmental factors underlie most of the diseases [4]. One of the current research areas which draw attention is the difficulty of identifying genetic variations that are the molecular basis of common diseases, such as neurodegenerative, immunological, and cardiovascular disease, diabetes and cancers. Genome Wide Association Studies (GWAS) are based on finding statistically most frequent variants (SNPs) in the individuals who have the disease compared with the healthy ones. Prioritization of the associated SNPs is the next step following a GWAS. Retrieval of up to date information for each SNP is critical during prioritization as it reveals SNPs that maps to functionally important loci and genes for the condition under study. Up to million unique SNP can be analyzed during GWAS and thousands of SNPs selected for prioritization. Integrating data for high number of SNPs requires a properly designed and easily updated database.

1.3 Semantic Web and Bio2Rdf

When first AI term was announced the world has changed. Today web serves over 10 billion pages and search engines can produce instant results to the users. In the last fifty years the results of AI research made semantic web possible. The Scientific American article made semantic web possible for the first time in 2001 [5].The Semantic Web is a linkage of actionable information, data derived through a semantic theory to interprete the symbols. The logical connection of terms establishes interoperability between systems in the the semantic theory which makes an account of "definition". [6]

The need for integrating the data on the web came from the people working on different sub fields. They are working on diverse and heterogeneous data sets but when someone needs another field's data, there was no integration among the fields. So a linkage of data within the same discipline or different discliplines was supplementary. Especially the scientists or authorities working on genomics or drug industy needed this kind of integrated web. So these communities tried to establish standarts in semantic web.

With the growing need for an integrated data web, also a standart must be produced. So organizations like the Internet Engineering Task Force and the World Wide Web Consortium(W3C) produced shared languages for interoperability purposes. The W3C defined the first Resource Description Framework(RDF) specifi-

cation in 1997. RDF is composed of triples(subject-predicate-object expression) [7]. Based on RDF, Bio2RDF creates a standart linked coherent data accross life sciences. Under Bio2RDF a large bioinformatics online database like data source created. It uses human and mouse genome to produce its over 70 million triples. General information about Bio2RDF can be found at http://bio2rdf.org [8]

1.4 Scope and Contribution of the Thesis

There are two main goals in this thesis. First one was automating the update process of the METU-SNP [9] database. Second was transforming the METU-SNP desktop application into a program which can be called with the parameters given via web environment. Before defining these two goals METU-SNP will be briefly explained. METU-SNP is a desktop application written in Java. It makes use of the techniques based on GWAS. GWAS analyzes dense maps of SNPs which includes the human genome to search allele-frequency differences among cases(individuals with specific proterties) and controls. When a significant DNA difference is seen it shows that it can be important and functional to point out disease traits among individuals [10]. Following GWAS, METU-SNP uses many public data sources to use in gene, snp, pathway and associated databases. It statistically analyzes and identifies significant SNPs, genes and pathways for the prioritization of associated SNPs.

1.4.1 Database Update Automation Process

In bioinformatics new data is always coming from various sources, so we need to use the new data to get proper and up to date results. So for METU-SNP application which had an offline database lastly updated in 2009 we needed to create an automated database update application. The program which makes the download operation is written in Java language and operating system independent. When the application is clicked it fetches the appropriate data from public databases and it downloads all the data which is needed in METU-SNP into the server database. It uses Bio2Rdf sources on the web which has roots of semantic web. The download operation is done using Sparql [11] query language(a language like sql).

1.4.2 Transforming the METU-SNP into Web Environment

The METU-SNP desktop application as its name implies can be started at a local computer by clicking on the program executable file. Besides it can run only at windows computers (All of its parts are written applicable especially for 32 bit Windows [12] computers). In order to transform the desktop application into the web environment, first we have converted the applicaiton to unix executable format, second we have modified the application. Several parts of the program used libraries which work on windows only, so libraries are changed. The METU-SNP have used few third party programs and all of them were compatible only with windows, so the third party programs working on unix are found and integrated. Some functions used in the al-

gorithm of the application didn't created the same results in unix, so modifications were made where needed.

The web site forms are created by a company called Userspots. The METU-SNP application runs at server side controlling a database called communication(The integration and communication is done via this database). When a user wants new analysis, the web site fills the necessary parts in the database with user-defined parameters. The modified METU-SNP application searches for each SNP in the database and fetches the new data when finds it. It uses the data as parameters for doing the analysis.

1.5 Thesis Outline

This dissertation is organized into five chapters. The first two introductory chapters provides general overview of the work done in this thesis and the previous work. The third chapter defines the METU-SNP application. The next chapter deals with the work in this study and the results of this work. Eventually, a concluding chapter is submitted.

- Chapter 2: Background and Literature
- Chapter 3: METU-SNP Application
- Chapter 4: Methods and Results
- Chapter 5: Conclusion and Future Work

CHAPTER 2

BACKGROUND AND LITERATURE

Here, general information about genes, SNPs are included as the METU-SNP application analyzes, prioritizes and annotates SNPs associated with diseases. Also the semantic web, bio2RDF, SPARQL is discussed, which are within the tools used for the database update of the study. Lastly a literature review is provided.

2.1 Biological Background

DNA is composed of 4 nucleic acids: Adenine (A), Cytosine (C), Thymine (T), and Guanine (G). These bases make pairs with each other (adenine with thymine and cytosine with guanine). These base combinations create DNA helix structure connected by sugar phosphate backbone.(See Figure 2.1- used with the permission of [13]). The whole human genome of an individual, build by 3 billion base pairs of DNA, provided the framework for all molecular characteristics. The DNA sequence, also called the genetic code is produced with the combination of these pairs. In almost every cell's nucleous the DNA material is found and organized into chromosomal structure, which is same for all cells in an organism.

A gene is a small portion of the genome which codes proteins. Cells use genes to synthetise proteins and this is a two step procedure. In the first step(transcription) one strand of a genome is used to create a RNA molecule. In the second step(translation) RNA molecules play the role of the guide to synthetise peptides in the ribosomes.

2.1.1 Transcription and Translation

Normally the DNA sequences are packaged and coiled by proteins. Before the transcription process, the sequence has to be opened and the strands must be spreaded. The DNA polymerase enzyme reads the related gene and produces the RNAmolecule (a single stranded chain). If there is a G in DNA strand C will be present in RNA and vice versa. But for A in DNA, RNA does not have T; instead of T Uracil will be present.(See Figure 2.2- used with the permission of [14])

After the whole gene(up to 10000 bases) is transcripted the RNA molecule detaches from the DNA strand. The messenger RNA can be modified by adding Adenins to

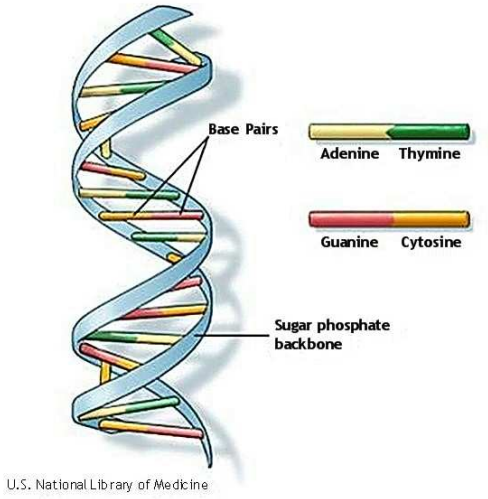

Figure 2.1: DNA Helix Structure

used with the permission of [13]

the end of the tail or excluding some parts of the introns. The removal of the introns(noncoding regions) and putting the exons together are called RNA splicing. The advantage of RNA splicing is to be able to create more than one type of protein by only one gene. It can be done due to some other factors' causing different splicing types. After the transcription the RNA can go through the pores in the nuclear membrane unlike the DNA.

In the translation process there are 3 types of RNAs: mRNA carries the needed code for translation to the ribosome from the DNA, rRNA plays structural role in the ribosome, tRNA carries amino acids for the needed codon (3 base bairs in the mRNA). Specific nucleotide sequences trigger some enzymes to support the whole translation process. The mRNA is coded in 3 to 1 fashion, 3 bases make one amino acid. There are a total of 20 amino acids in every living organism. One codon can make an amino acid or more than one codon can make one amino acid. There is a start codon(making amino acid metF) and three stop codons. Stop codons do not code for any amino acid and stops the translation process. And the genetic code for these amino acids are universal except for some organisms.(See Figure 2.3- used with the permission of [14]) When the mRNA binds among the small and big parts of the ribosome, the translation process starts. The tRNAs brings the needed amino acids one by one to form a chain of amino acids. After the stop codon comes the whole complex disassociates and the mRNA,tRNA, ribosome are directed to degradation.

The produced amino acid chain does not directly form the protein structure. First form of this amino acid chain is called the primary structure. Every amino acid has different chemical and electrical properties, so attractive and repulsive forces between amino-acids can make fold into chains, resulting in the secondary structure.

Enzymes called chaperonins play a very important role to give the final structure to

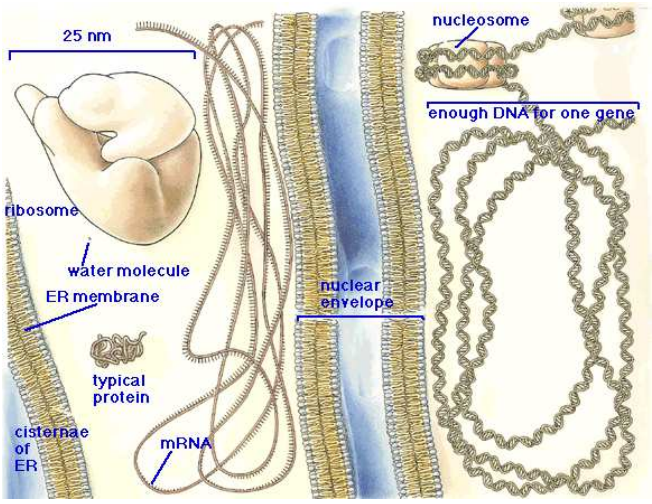

Figure 2.2: Transcription

used with the permission of [14]

the proteins by folding in specific ways. This tertiary structure can be said as the most important thing for a protein, since a protein interacts with other molecules with the help of its appropriate shape fitting the target molecule.

2.1.2 Variations

There is nearly 99.9 percent difference between genomes of all individuals. The sequences that are different are called variations and causing the unique personal charecteristics. Variations are the DNA sequence differences among individuals which can be one or more nucleotide long. The variations determine we look or which way diseases effect us. Insertions or deletions can cause these differences. In addition, extra copies of sequences can bind to anywhere. More dramatically some part of anchromosome can be transferred to somewhere else which is called translocations. These variations can have practical effects or not. According to the place of the variation(non-coding, coding, regulatory region) its biological effect changes. These can be grouped as:

- Harmless variations: Variations in non-coding regions do not make any effects, but some variations in coding regions doesn't have a known effect yet.
- Harmless change causing variations: In this group the variation occurs in the coding region and it does make a harmless change. It does not effect the protein sequence or the structure created by the variation including gene. Examples for these changes are the eye colors, human height or face looking.
- Harmful change causing variations: When a variation causes a disease it is categorised into this group. These are called mutations and occurs if the variation effects a protein that plays a role in health. For example; diabetes, cancer,

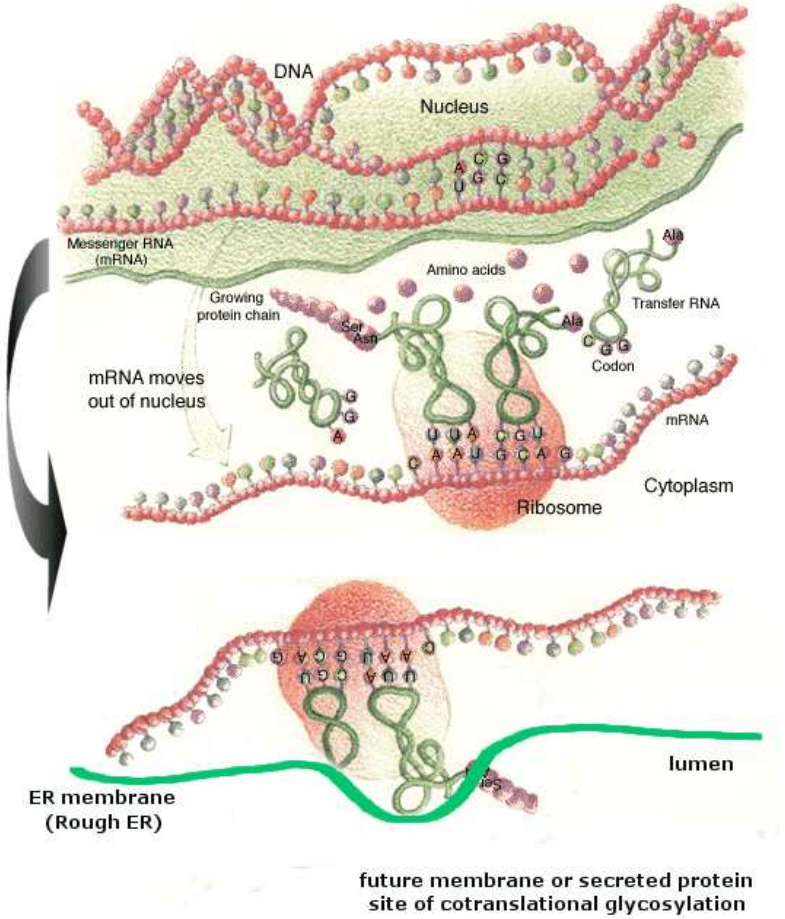

Figure 2.3: Translation

used with the permission of [14]

heart disease, Huntington's disease, and hemophilia are all the results of the mutations. But in hemophilia a mutation in only one gene is sufficient for the disease. In any cancer there must be several mutations in different genes to cause the disease.

- Latent effect causing variations: Finally there are changes that have latent effects. These changes don't have any effect on their own. After the change, the individual can see the effects according to the changes in the environment. Such changes may make one person more risky than others. For example if two persons are both smokers and drinkers one can get more effected than the other as a result of this type of variation, but if they haven't lived like that both of them wouldn't be effected.

2.1.3 Single Nucleotide Polymorphisms

The most common type of variations are single nucleotide polymorphisms(SNP). A SNP is a single base change which is seen in an important part (more than 1 percent) of the population. The single base is replaced by any of the other three bases. For example: in the sequence of TAGC, a SNP occurs when the G base changes to a C, and the sequence becomes TACC.(See Figure 2.4- used with the permission of [15])

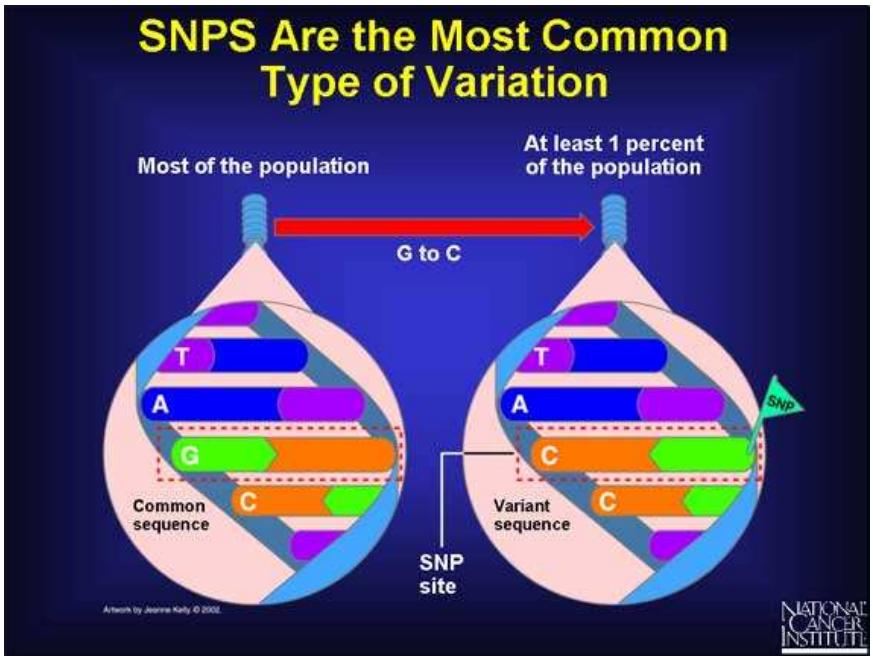

Figure 2.4: Single Nucleotide Polymorphism

used with the permission of [15]

SNPs can be found anywhere throughout the human genome, they can be on both the noncoding or coding regions. As the SNPs are also classified in to silent, harmless, harmful, or latent groups as variations. The frequency rate of finding SNPs in the

genes is very high. For ex. we can found SNPs about 1 in 1000 bases to 1 in 100 to 300 bases.

Some studies show that the SNPs are dense in the regions of recombination because the recombinations are mutagenic. [16] The high frequency of SNPs among variations and emerging techniques to rapidly identify SNPs make them significant variations. Most of the SNPs occur in the non-coding regions and do not alter the gene, but might have a regulative role. And they sometimes play a role for transcription because they can be at the region of the markers. The most important part of the SNPs are in the coding regions and they can change the structure, thus function of the proteins so they can effect the health of the individual. With the help of ongoing studies, new SNP data enters the bioinformatics literature. See Table 2.1- used with the permission of [17].

Table 2.1: The SNP Build Versions in dbSNP

Build	Release Date	Submissions	refSNP Clusters	Validated refSNP Clusters	SNPs in gene
137	June-12	192,678,553	53,567,890	38,072,522	22,450,743
135	Nov-11	178,140,935	53,327,221	41,750,143	21,247,880
134	Aug-11	179,506,198	41,365,915	6,961,883	16,880,992
132	Sep-10	143,350,315	30,442,771	19,727,605	12,212,318
131	Mar-10	114,900,250	23,653,737	14,653,228	10,375,413

2.2 Technical Background

2.2.1 Semantic Web

Today's Web is understandable by the humans, so a new approach which is easily machine-processible [18] is needed. In this new approach new techniques which utilise the backbone is suplementary and this whole structure is called the semantic web. This is not an alternative path for the existing web but can be called an improvement of the existing infrastructure.

Semantic web was firstly announced by the World Wide Web Consortium(W3C). The team leader of this propagation was Tim Berbers-Lee. In the development of the semantic web, industry and government play an important role since they have invested in this field. The US government has announced the DARPA Agent Markup Language(DAML) project [19] in the area of semantic web. Also European Union's sixth framework programme includes semantic web as a key action line.

Semantic web uses some ontology languages, important ones are:

- XML: provides surfaced syntax for structured documents.
- XML schema: a language which restricts the structure of XML based documents.

- RDF: it is a data model of resources and relations among them. These data models can be represented in XML syntax.
- RDF schema: is a vocabulary description language for defining RDF sources which generalizes hierarchies of the properties of these sources.
- OWL: is a richer vocabulary description language with respect to RDF.

Lastly semantic web has layered structure consisting of the web standarts , as presented in Figure 2.5- used with the permission of [20].

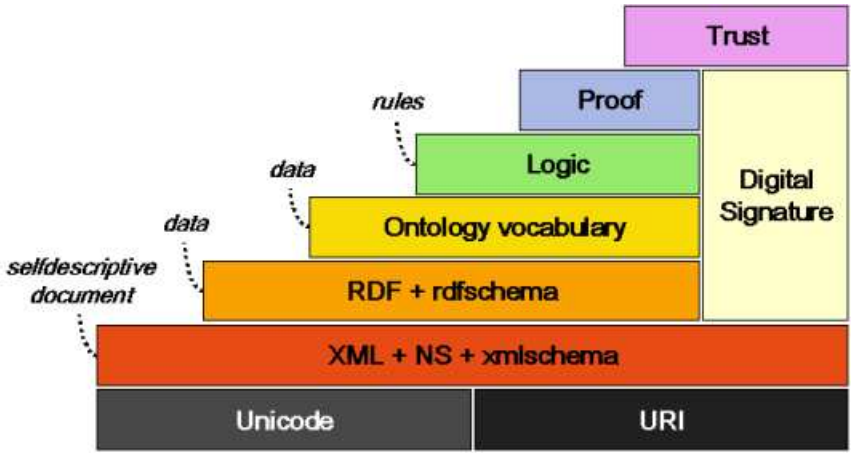

Figure 2.5: Semantic Web Layers

used with the permission of [20]

2.2.2 RDF

RDF stands for Resource Description Framework. The framework is used to represent data as an XML file in Linked Open Data cloud [21]. The things which can be present on the web have a potential to be represented as RDF. RDF is needed for situtations, where information will be used by applications rather than the humans. So, it plays the role of a bridge among applications. In RDF the data are identified using uniform resource identifiers(URIs) by the properties and their values. So simple statements can be shown as a graph which includes arcs and the properties.

RDF framework consists of the triples, which is used for each source. Subject, object and predicate create this triple. These can be defined as:

- Subject: is the URI which is described above.
- Object: can be a literal value variable according to the property. It can be a text, number or a date. Also it can be the Uri of another resource.
- Predicate: This defines the relationship type between the subject and the object.

By using these properties and the RDF standart, it is very appropriate to classify web content. After adding the RDF content to the Linked Open Data cloud it becomes part of the linked space and can link itself to another RDF content related.

In the predicate definition we saw that it is the relationship type among the object and subject. In RDF vocabulary, property term can also be used for predicates. See Table 2.2 for properties and subproperties in RDF.

Table 2.2: RDF Properties

Property name	comment	domain	range
rdf:type	The subject is an instance of a class.	rdfs:Resource	rdfs:Class
rdfs:subClassOf	The subject is a subclass of a class.	rdfs:Class	rdfs:Class
rdfs:subPropertyOf	The subject is a subproperty of a property.	rdf:Property	rdf:Property
rdfs:domain	A domain of the subject property.	rdf:Property	rdfs:Class
rdfs:range	A range of the subject property.	rdf:Property	rdfs:Class
rdfs:label	A human-readable name for the subject.	rdfs:Resource	rdfs:Literal
rdfs:comment	A description of the subject resource.	rdfs:Resource	rdfs:Literal
rdfs:member	A member of the subject resource.	rdfs:Resource	rdfs:Resource
rdf:first	The first item in the subject RDF list.	rdf:List	rdfs:Resource
rdf:rest	The rest of the subject RDF list after the first item.	rdf:List	rdf:List
rdfs:seeAlso	Further information about the subject resource.	rdfs:Resource	rdfs:Resource
rdfs:isDefinedBy	The definition of the subject resource.	rdfs:Resource	rdfs:Resource
rdf:value	Idiomatic property used for structured values	rdfs:Resource	rdfs:Resource
rdf:subject	The subject of the subject RDF statement.	rdf:Statement	rdfs:Resource
rdf:predicate	The predicate of the subject RDF statement.	rdf:Statement	rdfs:Resource
rdf:object	The object of the subject RDF statement.	rdf:Statement	rdfs:Resource

2.2.3 Bio2RDF

The Bio2RDF(http://bio2rdf.org) project creates a network of coherent linked data among the biological databases. It is created to provide knowledge integration in bioinformatics area. Bio2RDF is based on RDF documents and linked with the uri's given. It collects the public data from various popular bioinformatics databases. In other words, it can be described as a semantic web approach for knowledge integration.

Firstly an integration of bioinformatics data is made in 1995 [22]. This idea is realized by transforming the biological data to a common federated database. Then semantically equivalent data are matched. To create all the RDF data in Bio2RDF, an RDFizing process is made by the help of the Sesame open source triplestore, the Piggy Bank [23] semantic web browser plug-in for FireFox, the Protégé ontology editor and the Welkin. Ontology Web Language (OWL) is selected as the ontology to build the Bio2RDF structure, and as the implementation language [24] [25].

Bio2RDF was basically about analyzing the related html forms on the web and transforming the data into RDF form. As it was defined above, the RDF consists of the triples. After analyzing and parsing the html the problem goes to the place, where to put the read data inside the triples. The answer is putting the labels as predicates and the hyperlinks as the URIs. The semantics of these predicates can be followed in Bio2RDF ontology file [26]. RDFizer programs are produced for Bio2RDF in order to overcome the limitation on the namespaces. The external references must be edited before creating the URIs in order to properly link the triples. This procedure is called URI normalization. Here in Fig 2.6, how the URIs return documents in RDFformat from Bio2RDF.org server are presented.

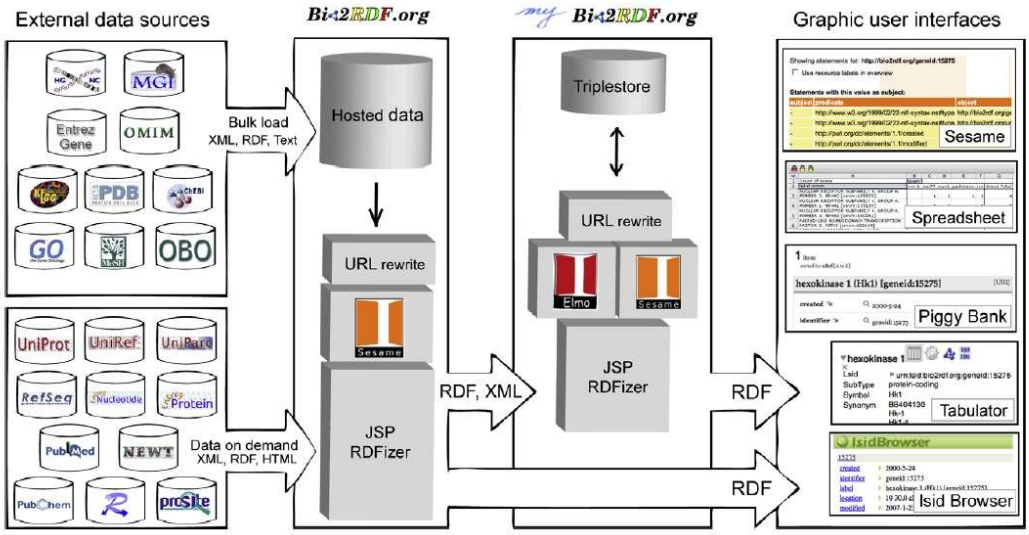

Figure 2.6: Bio2RDF Framework Architecture

-used with the permission of [8]- On the left part of the figure there are the data sources. The most important ones are Entrez Gene [27], GO [28], OBO [29], OMIM [30], Kegg, PDB [31], chEBI [32], MGI [33] and HGNC [34]. The data from these sources are transferred into the bio2RDF mysql database firstly, then the RDFization process starts. However some other data sources are directly fetched from the web sites and RDFized respectively. Also some regularly used data sources are cached for speed and availability purposes.

The Bio2RDF database is still in development with more than twenty data sources, which are represented in Figure 2.7 -used with the permission of [8]- for statistics about these sources. The database consists of 163 million well-formed RDF documents, normalised URIs which are created according to Bio2RDF ontology. Some of them are served from the local databases. By doing this, hundreds of documents can be downloaded in a couple of minutes rather than hours. Moreover it is suitable for NCBI usage restrictions [35]. Lastly the Bio2rdf database is extendable by any user who wants to contribute to this biological linked database project. Any biological or genetics data can be transformed into the rdf format and loaded into the database.

2.2.4 SPARQL Query Language

SPARQL is a RDS query language which is a recursive acronym of SPARQL Protocol and RDF Query Language. It is a query language which can retrieve and change the data stored in the databases of format Resource Description Framework [36] [37]. RDF Data Access Working Group (DAWG) has recognized the language as one of the key technologies in semantic web. This group is associated with World Wide Web Consortium. Respectively SPARQL 1.0 became official by W3C group [38] [39] on 15 January 2008, and SPARQL 1.1 in March 2013 [40].

There are similarities with the SQL since the same words like SELECT, WHERE exist, but words like OPTIONAL, FILTER are unique to SPARQL. RDF consisting

Data source	Short URI example	Number of RDF documents	Format of source data
genenames.org	hgnc:4922	27,634	Tabulated
informatics.jax.org	mgi:96103	70,172	Tabulated
ncbi.nlm.nih.gov	omim:146200	18,284	XML
ncbi.nlm.nih.gov	geneid:3098	3,315,893	XML
genome.ad.jp	path:mmu00010	68,307	KGML
genome.ad.jp	cpd:C00011	15,006	Text
genome.ad.jp	dr:D00001	6755	Text
genome.ad.jp	ec:2.7.1.1	4,958	Text
genome.ad.jp	gl:G00001	10,972	Text
genome.ad.jp	rn:R00014	7422	Text
ebi.ac.uk	chebi:16526	13,360	Tabulated
rcsb.org	pdb:1HKC	48,091	XML
geneontology.org	go:0004396	24,634	OBO/RDF
nlm.nih.gov	mesh:D006593	23,512	RDF
obofoundry.org	*obo's 54 namespaces*	108,955	OBO/RDF
beta.uniprot.org	uniparc:UPI00005AC213	30,261,843	RDF
beta.uniprot.org	uniprot:P19367	4,177,176	RDF
beta.uniprot.org	uniref:UniRef50_P19367	7,990,452	RDF
beta.uniprot.org	taxon:9606	441,422	RDF
ncbi.nlm.nih.gov	genbank:NP_277035	61,132,599	XML
ncbi.nlm.nih.gov	pubmed:3207429	17,000,000	XML
ncbi.nlm.nih.gov	pubchem:3313	38,000,000	XML
reactome.org	reactome:70326	8,332	BioPAX/RDF
expasy.org	prosite:PS00378	2,819	HTML
	Total	162,778,598	

Figure 2.7: Bio2RDF Documents from Public Databases

used with the permission of [8]

of subject, object and predicate triple as mentioned above. A SPARQL query includes these subject, object and predicate triples and the elements here can be variables. The point of a SPARQL query is making association between the triples of RDF and the triples of the SPARQL query. The SPARQL queries can be run on RDF databases, which have specific endpoints for the SPARQL. The data is retrieved via http with the help of the endpoints given as key.

A SPARQL example:

:id1 foaf:name Ceyhun
:id1 foaf: basedNear : Ankara
:id2 foaf:name Himmet
:id2 foaf:basedNear : Mersin

If we want to retrieve the names of all the people in the database we can use the query below:

SELECT ?name
WHERE }
?x foaf:name ?name
}

The variable names, intended to retrieve, start with question marks. The important part in the query is the WHERE part. The triple is subject:

?x, predicate: foaf:name, object: ?name.

There are two variables instead of subject, and predicate and one constant for the predicate. This triple is compared with the all the values in the RDF database. The constant value (foaf:name) will be matched with the same values in the RDF database. Therefore, two results will be found in this case, the name variable matches with Ceyhun and Himmet.

id1 foaf:name Ceyhun and id2 foaf:name Himmet.

2.3 METU-SNP Application

The METU-SNP application can be defined as an all in one GWAS application, which uses the sources (public databases) such as dbSNP, Entrez Gene, KEGG, Gene Ontology etc. As one of the major goals and outcomes of this thesis is updating METU-SNP database and transforming it to the web environment, the workflow of the METU-SNP will be discussed shortly.

METU-SNP, performs PLINK analysis and combined p-value calculation, and makes use of AHP based SNP prioritization and gene set enrichment analysis frameworks. It is also equipped with a machine learning algorithm called simulated annealing (SA) in order to choose representative SNPs [9].

The program is written in Java and uses java swing for user interface. It uses JDBC to connect to the databases. Unlike the common model-view-controller(MVC) it uses a pattern called model-delegate. The view in the delegate part and controller parts are combined into one layer. In the delegation part data imputation, association analysis, SNP prioritization and selection work are done. These are done with the help of third party programs. Before the application started, Java Runtime Environment [41] and mysql must be installed. The METU-SNP can be installed and run on a local computer.(See Figure 2.8- used with the permission of [9])

The program uses Plink [42], Beagle [43], Weka [44] as third party tools. Plink is a whole genome association analysis toolset, which can be used for GWAS. It includes the subjects; meta-analysis, population stratification detection, summary statistics for quality control, basic association testing, copy number variant analysis, data management, result annotation and reporting. However PLINK can only be called via the command line with some complex parameters which can be unfriendly for a typical user. So METU-SNP made PLINK more friendly by providing a graphical user interfaces.

The Beagle performs the imputation job for users trying to statistically predict the missing values with the help of already calculated ones. Many methods have been proposed to predict sporadic missing data by imputation in order to increase the effect of existing marker sets in GWAS [45]. The imputation is widely used in GWAS for meta-analysis of many diseases [46] [47]. In METU-SNP Beagle is chosen because its input data format is very similar to Plink's format and it is a stand alone jar package in order to be easily runned. User has the choice for imputation of the ungenotyped

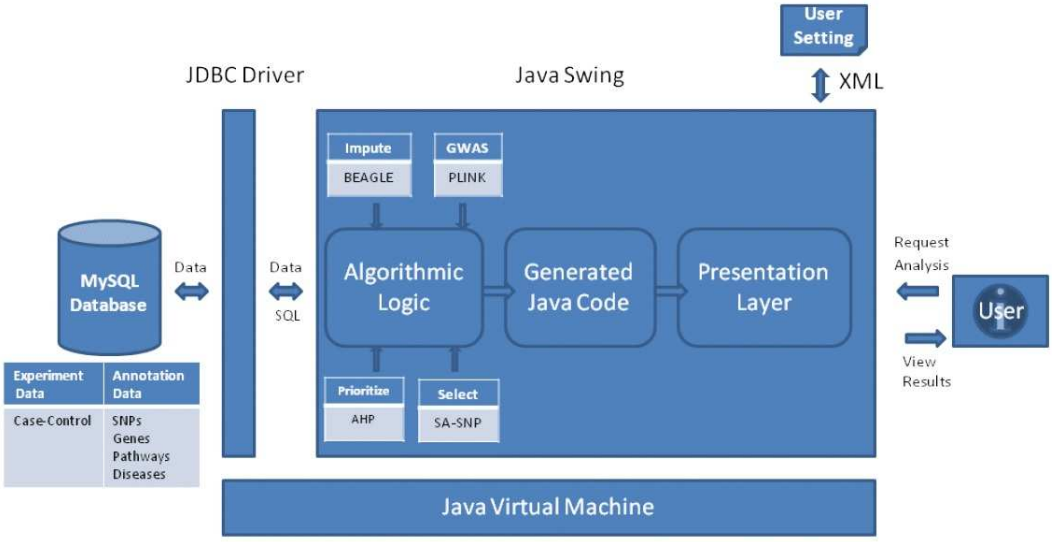

Figure 2.8: METU-SNP System Architecture

used with the permission of [9]

data, when the desired genotyping percentage isn't reached.

Lastly Weka is a machine learning and data mining tool. The aim of the program is regression, pre-processing, classification, clustering.

2.3.1 The METU-SNP Database

The METU-SNP database, which is the focus of our study, consists of the data from various public databases and based on Mysql 5. See Figure 2.9- used with the permission of [9]- for the relational database general scheme.

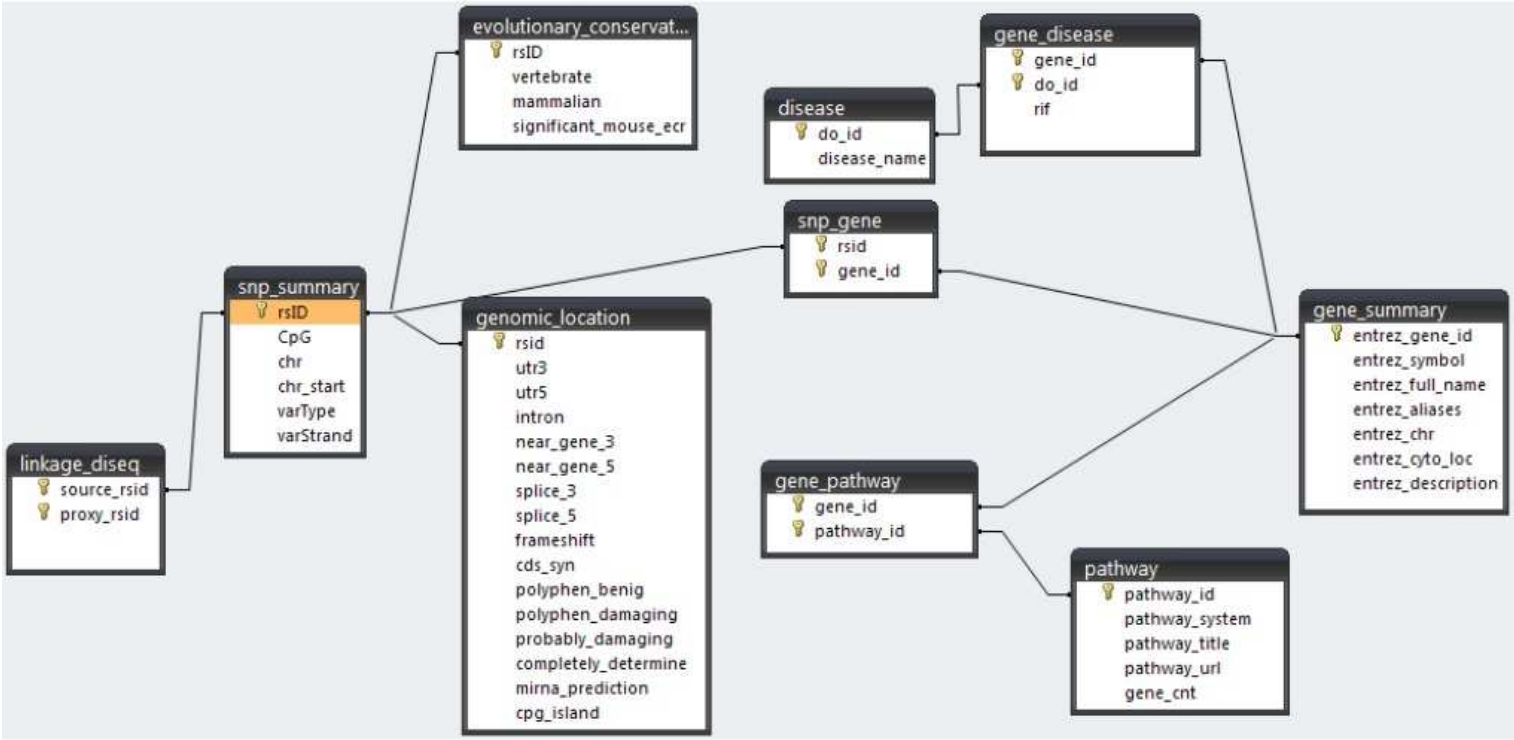

Figure 2.9: METU-SNP Database ER Diagram

used with the permission of [9]

The most important part of METU-SNP data is adapted fromSNP Logic's integrated database. [48] The SNP IDs(rsId) are collected from dbSNP, which uniquely defines individual SNPs. And other basic annotations including function class, associated gene ID and symbol are also retrieved from dbSNP. [49] Papasuite [50] provides data for scoring of SNPs based on the overlap with the specific markers. For the evolutionary conserved region data, UCSC [51] is utilized to find the conserved regions among multiple species. Lastly an online prioritization tool called SPOT [52] is used to get the linkage disequilibrium correlation values.

In METU-SNP database structure, the gene ids are downloaded from Entrez Gene. The other gene related information is provided from SPOT which is also sourced from NCBI Entrez Gene. SNP-gene associations are gotten from NCBI and dbSNP. (See Table 2.3- used with the permission of [9])

Table 2.3: Gene Based Database Annotation

Field	Description
Entrez Symbol	NCBI Entrez Gene official gene symbol
Entrez Gene ID	NCBI Entrez Gene ID
Gene type	Gene type: protein-coding, tRNA, etc.
Entrez full name	Full name from NCBI Entrez Gene
Chr	Chromosome
Start Pos (bp)	Start Position in base pairs (NCBI Mapview)
Stop Pos (bp)	Stop Position in base pairs (NCBI Mapview)
Size (kb)	Size of transcript in kb (NCBI Mapview)
Cytogenetic Pos.	Cytogenetic Position

Determining the definite molecular mechanisms underlying diseases is a popular study area. Genes and SNPs which resides in the same disease cause elements are assumed to be in the same biological pathway. METU-SNP keeps the pathway IDs and web link from Gene ontology, KEGG, Wiki Pathways and Biocarta. Lastly a mapping approach called GeneRIF-Disease Ontology (DO) [53] [54] is utilized in association of genes and diseases. DO-RIF project page of Northwestern University reflects the mapping used in METU-SNP.

2.3.2 METU-SNP Algorithms

There are many algorithms utilized in METU-SNP but two of them are selected to be included in the web service developed within this thesis: Preprocessing and GWAS analysis. During preprocessing step map and ped(pedigree) files are given as inputs(especially for plink), then quality control based filtering and imputation is done for the input.See Algorithm 1- used with the permission of [9] for the preprepreprocessing pseudocode. Second step is the GWAS analysis. This step tries to find the statistically signicant SNPs associated with the given disease. First the P values and SNP rsIDs are found then genes and pathways are identified with the help of SNP data based on the combined p-value approach [55]. See Algorithm 2- used with the permission of [9].

Algorithm 1 METU-SNP Preprocessing

Input: P: Genotype data in pedigree format, all data in the same strand.
M: Map data.
maf: Minor allele frequency threshold.
Sm: SNP missingness rate threshold.
Im: Individual missingness rate threshold.
H: Hardy Weinberg equilibrium threshold.
rsq: Allelic r^2 threshold.

Output: *Cleaned*: Quality controlling applied SNP set.

$S \leftarrow P$
if impute = true **then**
  while Chromosome i in M **do**
    $Si \Leftarrow SplitChromosome(S)$
    $Si \Leftarrow BEAGLE(Si, rsq)$
    $count++$
  end while
  $Merged = \emptyset$
  $j \Leftarrow 1$
  while $j \neq count$ **do**
    $Merged = Merged \cup Sj$
  end while
  $cleaned = PLINK(Merged, maf, H)$
else
  $cleaned = PLINK(S, Sm, Im, maf, H)$
end if

Algorithm 2 METU-SNP Two Wave Gwas Run

Input: *Cleaned* Quality controlling applied SNP set.
Type Individual SNP p-value type: Bonferroni, FDR or uncorrected.
$pSNP$ p-value threshold for significance for SNPs.
$pGENE$ p-value threshold for significance for genes.
$pPATH$ p-value threshold for significance for pathways.
Output: $signSNP$ Statistically significant SNP set.
$signGENE$ Statistically significant gene set.
$signPATH$ Statistically significant pathway set.
S List of SNPs and p-values.
$signSNP = \emptyset, signGENE = \emptyset, signPATH = \emptyset, GeneList = \emptyset, PathwayList = \emptyset$
$S \Leftarrow PLINK(Cleaned, Type)$
$countS = |S|$
$i = 1$
while $i < countS$ **do**
$GeneList = GeneList \cup AssociatedGene(Si.rsId)$
if $Si.pvalue < pSNP$ **then**
$signSNP = signSNP \cup Si.rsID$
end if
end while
$countG = |GeneList|$
$j = 1$
while $j < countG$ **do**
$PathwayList = PathwayList \cup AssociatedPathway(GeneListj.GeneId)$
if $CHIDIST(-2 \sum_{k=1}^{|} GeneListj.SNPset|log(GeneListj.SNPsetk.Pvalue), 2|GeneListj.SNPset|) <$
$Pgene$ **then**
$signGene = signGene \cup GeneListi$
end if
end while
$countP = |PathwayList|$
$l = 1$
while $l < countP$ **do**
$m = |PathwayListl.GeneSet|$
$n = |PathwayListl.GeneSet \cap signGene|$
if $1 - \sum_{p=0}^{n} \frac{\binom{signGene}{p}\binom{countG-signGene}{m-p}}{\binom{countG}{m}} < Ppath$ **then**
$signPath = signPath \cup PathwayListl$
end if
end while

CHAPTER 3

METHODS AND RESULTS

3.1 Integration of SNP Databases via RDF

The main goal of this study was updating the database of the METU-SNP application, which was defined in the previous chapter. In this section why and especially how we created a mechanism for database update will be presented.

3.1.1 Introduction

In bioinformatics studies, the genetic and biological data collected from various sources are processed with the help of computational, biological or statistical technics. First and the important step is the data collection. Access to accurate and up-to-date data is crucial for the studies. The correctness of the data must be tested and the sources of the this data must be reliable. There are many bioinformatics data sources but finding reliable and the applicable data for our studies was a critical process.

The technologies like microarray accelerated the growth rate of molecular data within the last two decades [56]. The size of the bioinformatics data collected in various databases are continuously expanding, which constrains the data sources to be up to date. If the size of the new data coming in one year is as high as the current data size and if the database is not updated, the effectiveness rate of the results decreases to nearly 50 percent. To make an illustration of how fast the genetic data is emerging, SNP data can be a good example.

In dbSNP data source there are 139 builds(updates) in the day of this thesis is written. For example in build 128, which was created in 2007 there are 34.4341.59 unique rsIds. In build 135 from 2011 there are 178.140.935 rsIds, which is approximately 5 fold of build 128 in terms of size [57]. See Figure 3.1- used with the permission of [57]- for the dbSNP growth rate in years.

3.1.2 Determining the Update Source and Technique

In this section a brief summary of the research done to access the needed data for the METU-SNP will be presented. The first step was determining the data needed for the

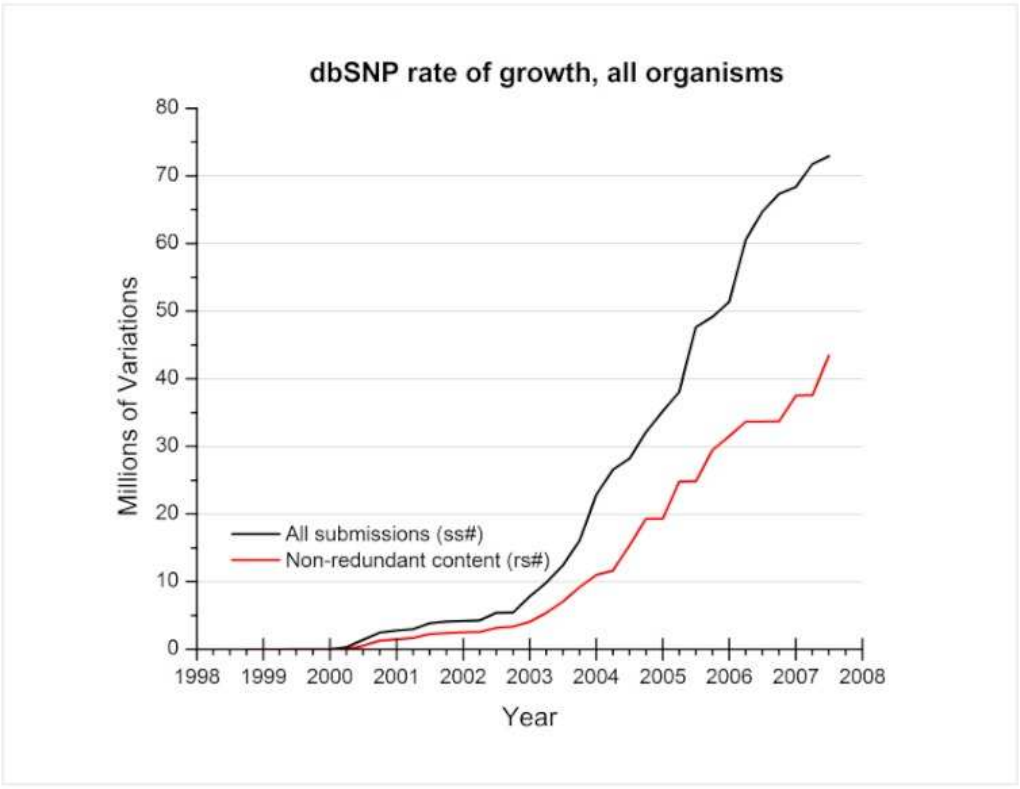

Figure 3.1: dbSNP Growth Rate

used with the permission of [57]

METU-SNP to work properly. In the Figure 2.9 explaining the METU-SNP database ER diagram , the elements in the offline database are shown. However further analysis of the application has proved that only some of the data in the diagram was utilized by the program. This analysis is done by examining the database processing layer of the METU-SNP. As a result see Figure 3.2 for the data needed to update.

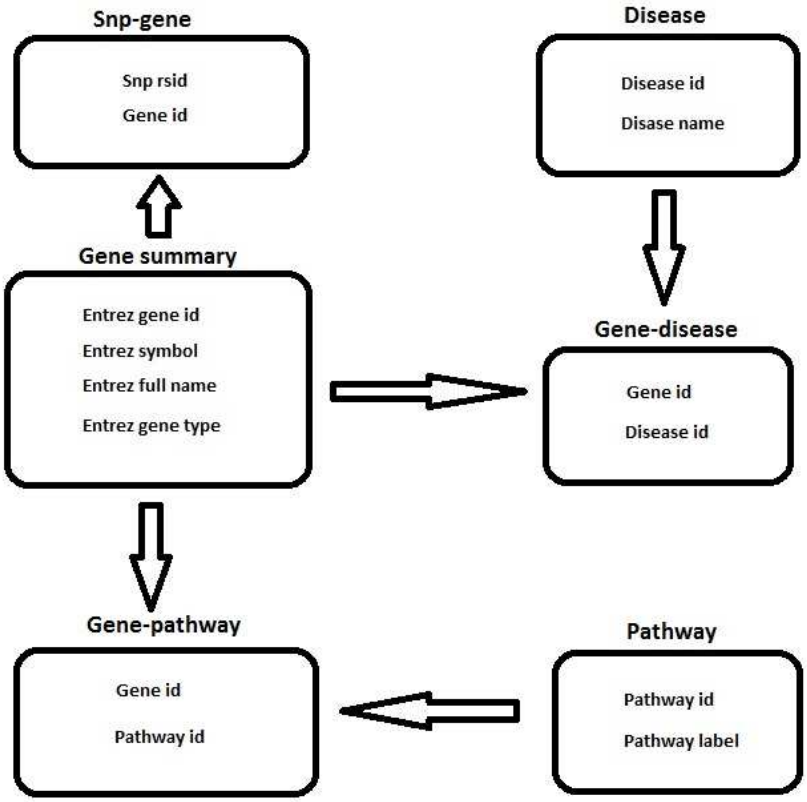

Figure 3.2: METU-SNP Data to be updated

The data is required for web service functions

After determining the data that will be updated next step was to identify the data sources available with the criteria listed below:

- Consistent with the previous data(id's or names must be suitable with the METU-SNP)
- Easily to parse
- A stabilized format(must not change how we retrieve the data)
- Always online

- Data correctness
- Lets making associations with the past data in the database

According to the conditions, investigations had been made to find the appropriate sources. At first two formats of data are found: the data in the format of ftp [58] and http [59] were the first elements.

The data embedded in the http protocol scan be easily read by humans, but computers need to know the format of the web page beforehand. The place where the data resides is an essential information for the downloader program. Besides after determining the addess, application must know the data type. The basic idea is traversing the html code and when the specific pattern which applies the needed data is found, the string is read. See Figure 3.3 for an example of Snp data html page. In this example all of the SNP results must be parsed by following the next links, then whenever an rsID pattern is found in the html code, the link must be followed for detailed snp data. This process requires complex algorithms.

Figure 3.3: Example Http SNP Data Download Page

The Integrated SNP database (iSnp) [4], the downloaded project, was first build on html parsing, but there were many disadvantages of html parsing. Whenever even a small format change is made in the target web site, the whole download operation would fail. Moreover the modularity of the program decreases, since it is very specific to the download page. If we want to change the download source we have to write a new program and this causes a fully altered source code. The time complexity is

high since even only for snp data many traverses are needed over hundreds of pages and this means high amount of time. When these disadvantages are considered we decided to continue with the ftp sourced data.

In ftp servers the data is organized like typical local computers' folder system. The problem with these, is the difficulty of finding the needed data. There are many folders in one server and many folders can also exist inside one folder. So every folder must be searched by the researcher. However every file must be opened and read, since the names of the data may not give a clue. If the data source is not eliminated by finding distinctive properties of the data (only some of them can be found, then a new search must be made with a new data source) before the related file must be downloaded.

This is another problem, since a file can be very huge and some of the data in the file can be useless. This costs both time and additional storage. Lastly we need to parse all of the downloaded file, the same problem of format sensitivity is also valid in this case. As a result, vulnerability against format changes are present. Also we need memory management techniques to read the file, since the files can be so big despite low computer memory spaces.

Finally, we have decided not to use either html or ftp based data sources. After examining all the alternatives we have found an RDF database called Bio2RDF, which meets our requirements. Bio2RDF has standarts to retrieve the needed data, some additional languages like SPARQL help us use this standart formatted data. Many public biological data sources have RDFizied forms of data in Bio2RDF, so we easily could find the information which we are interested. And lastly the probability of format change in the databases is very low.

3.1.3 The iSNP

Building the integrated SNP database, which can be updated periodically is one of the main objectives of this study. This program is designed a stand alone application and written in Java language. It runs on computers independent of the operating system with the help of the java language and exclusion of operating system specific tools. The basic idea of the program is, when clicked on the Jar file it starts to operate and downloads the required content from Bio2RDF into the local mysql database.

Some libraries help the work done by the program. The jena library is used to fetch data from Bio2RDF, it plays the role of a layer among the SPARQL and the online database. Java standart sql library is used to connect to the database. The downloaded data does not interact with the previous data, since all of the newly coming information replaces the old data. Therefore there is no consistency issues for considering the old database information.

See Algorithm 3 for iSNP pseudocode.

Algorithm 3 iSNP Pseudocode

1: $SPARQLENDPOINTGene = http://cu.gene.bio2rdf.org/sparql$
2: $SPARQLENDPOINTCtd = http://cu.ctd.bio2rdf.org/sparql$
3: $model \Leftarrow ModelFactory.createDefaultModel()$ {Create an empty Rdf model}
4: $LoadDrivers()$
5: $DriverManager.getConnection(URL, user, password)$ {Create mysql connection}
6: $EndPoint \Leftarrow SPARQLENDPOINTGene$
7: $rawGeneLabel, rawGeneSymbol, rawGeneDescription, rawGeneType \Leftarrow GetGeneInformationWithSparql()$
8: $RunSql(Deletefromgenesummary)$ {delete all the data in the genesummary table}
9: **while** GeneResults.hasNext() **do**
10: $entrezGeneId \Leftarrow parseAndGetAfter(geneId:, rawGeneLabel)$
11: $entrezGeneSymbol \Leftarrow rawGeneSymbol$
12: $entrezGeneDescription \Leftarrow rawGeneDescription$
13: $entrezGeneType \Leftarrow parseAndGetBetween(http://bio2rdf.org/geneid:vocabulary:$
14: $, http://bio2rdf.org/geneid:vocabulary:Gene, rawGeneType)$
15: **end while**
16: $InsertintogeneSummary(entrezGeneId, GeneSymbol, GeneDescription, entrezGeneType)$
17: $EndPoint \Leftarrow SPARQLENDPOINTCtd$
18: $rawGeneDiseaseLabel \Leftarrow GetGeneDiseaseInformationWithSparql()$
19: $RunSql(Deletefromgenedisease)$ {delete all the data in the genedisease table}
20: **while** GeneDiseaseResults.hasNext() **do**
21: $geneId \Leftarrow parseAndGetBetween(geneid:, /, rawGeneDiseaseLabel)$
22: $diseaseId \Leftarrow parseAndGetBetween(:, /, parseAndGetAfter(/, rawGeneDiseaseLabel))$
23: **end while**

```
24: InsertintogeneDisease(geneId, diseaseId)
25: rawDiseaseLabel ⇐ GetDiseaseInformationWithSparql()
26: RunSql(DeletefromDisease) {delete all the data in the disease table}
27: while DiseaseResults.hasNext() do
28:     diseaseName ⇐ parseAndGetBefore(/[, rawDiseaseLabel)
29:     diseaseId ⇐ parseAndGetBetween(:, /], rawDiseaseLabel)
30: end while
31: InsertintoDisease(diseaseId, diseaseName)
32: rawGenePathwayLabelrawGenePathwayAssociation ⇐
    GetGenePathwayInformationWithSparql()
33: RunSql(DeletefromGenePathway) {delete all the data in the GenePathway table}
34: while GenePathwayResults.hasNext() do
35:     if rawGenePathwayAssociation.contains(geneId:) then
36:         geneId ⇐ parseAndGetAfter(geneId :, rawGenePathwayAssociation)
37:         pathwayId ⇐ parseAndGetBetween(:, /], rawGenePathwayLabel)
38:         InsertintoGenePathway(geneId, pathwayId)
39:     end if
40: end while
41: rawPathwayLabel ⇐ GetPathwayInformationWithSparql()
42: RunSql(DeletefromPathway) {delete all the data in the Pathway table}
43: while PathwayResults.hasNext() do
44:     pathwayTitle ⇐ parseAndGetBefore(/[, rawPathwayLabel)
45:     pathwayId ⇐ parseAndGetBetween(:, /], rawGenePathwayLabel)
46:     InsertintoPathway(pathwayTitle, pathwayId)
47: end while
48: CloseModel()
```

3.1.4 SPARQL Example from iSNP

The program uses SPARQL query langugage to access the Bio2RDF data. In the pseudocode written above, the functions starting with ”parseAndGet” include SPARQL to fetch the needed data. There are many SPARQL code blocks in the program to fetch the data with the databses, where the gene information download code is illustrated in Figure 3.4 as an example. The rest of the code blocks are provided in Appendix C.

```
"PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>"+
"select distinct ?dlabel ?symbol ?description ?typer" +
"where {" +
"       " +
"       ?gener rdf:type <http://bio2rdf.org/geneid:vocabulary:Gene> ." +
"       ?gener rdf:type ?typer ." +
"       ?gener rdfs:label ?dlabel ." +
"       ?gener <http://bio2rdf.org/geneid_vocabulary:has_symbol> ?symbol ." +
"       ?gener <http://bio2rdf.org/geneid_vocabulary:has_description> ?description
"       " +
"       " +
"       }" ;
```

Figure 3.4: SPARQL Example Code in iSNP

As presented in Algorithm 3 the endpoint for gene information is selected as Gene. This determines the database to make process for the SPARQL code. This SPARQL code is like the basic sql code used in regular databases. In PREFIX there are the predefined predicate values determined by the Bio2RDF and labels are given as shown. The data required to be fetched is written after the select clause, the distinct word means eliminating the same values returned. The variables are the words with the question marks.

As it is explained in RDF and bio2RDF sections, the principle is about matching the SPARQL triples with the triples in the RDF database. So this code firstly matches all the triples in the Gene database with the predicate ”rdf:type” and object ”http://bio2rdf.org/geneid:vocabulary:Gene” . The only variable is the subject which is named ”?gener”. The other SPARQL matches belove the first comparison is applied according to the ”?gener” variable. The other triples which do not match the ”?gener” will be eliminated. Therefore the other four triples which have a specific predicate value will be returned and the objects will be the variables retrieved.

3.2 Transformation of METU-SNP into Web Environment

METU-SNP is an operating system dependent desktop application. It uses considerable amount of data from the database structure. Therefore if a user wants to use the software he must download the whole software with its dependencies. Besides, the database and the data must be installed. In order to prevent this, we have transformed the application into the web environment. We call this as prioritization through iSNP (pi-SNP).

3.2.1 Building The pi-SNP Web Service

Few alternatives has been considered for the transformation process. First was to do all the work which METU-SNP does via the web technologies. This would require much time and labor since all the coded work in METU-SNP has to be transferred as they were written in web technologies like php, java script etc..originally. Moreover transforming the code from the traditional programming language to a web based code wouldn't bring performance improvement with the complex algorithms like in the METU-SNP case.

The alternative which is chosen and called pi-SNP is explained in this section. In pi-SNP the core and algorithms of the METU-SNP is preserved as the old application. The application is still executable in the local computer. However some modifications are made to be able to get parameters from the web forms.

There are two layers in the application consisting of view-model and data layer. The modifications are made in the view-model layer. All of the needed data is injected into the application as parameters by the help of these modifications. As explained in chapter 2 we use only preprocessing and GWAS analysis algorithms of the application.

Therefore other sections of the application are not implemented. The processes of these two parts are called respectively after the parameters are read into related arrays. The parameters are firstly grouped and given with these arrays. Moreover the preprocessing and GWAS analysis processes are called respectively by the modified METU-SNP unlike the desktop application since it lets users to call these processes one by one. The modifications done on the METU-SNP was for the parameters given by a communicator program. It reads all the data from the common mysql database among the communicator program and the web site.

See Figure 3.5 for the general structure of the piSnp.

3.2.2 pi-SNP Communicator Program

This is basically a transferring application between the database and the METU-SNP application. The principle is to run the communicator every time to check the database. Therefore the application always polls new data from the database. When the web site inserts a new data group after user enters arequest for a new analysis, it searches for new data in the communication database. Afterwards the program retrieves the attributes in the database. It deletes the used data whenever it reads new one. As a result whenever the program finds data in the communicator table of the database, it calls the METU-SNP application with the read data as parameters. METU-SNP runs independent of the program after being called.

All of these are done to transfer the specific data. These data consists of disease names in order to find correlations with the SNPs. Moreover there is some statistical and biologically related data as listed in Table 3.1. Disease names are read from a separate table with the help of the disease number value. The data transferred is

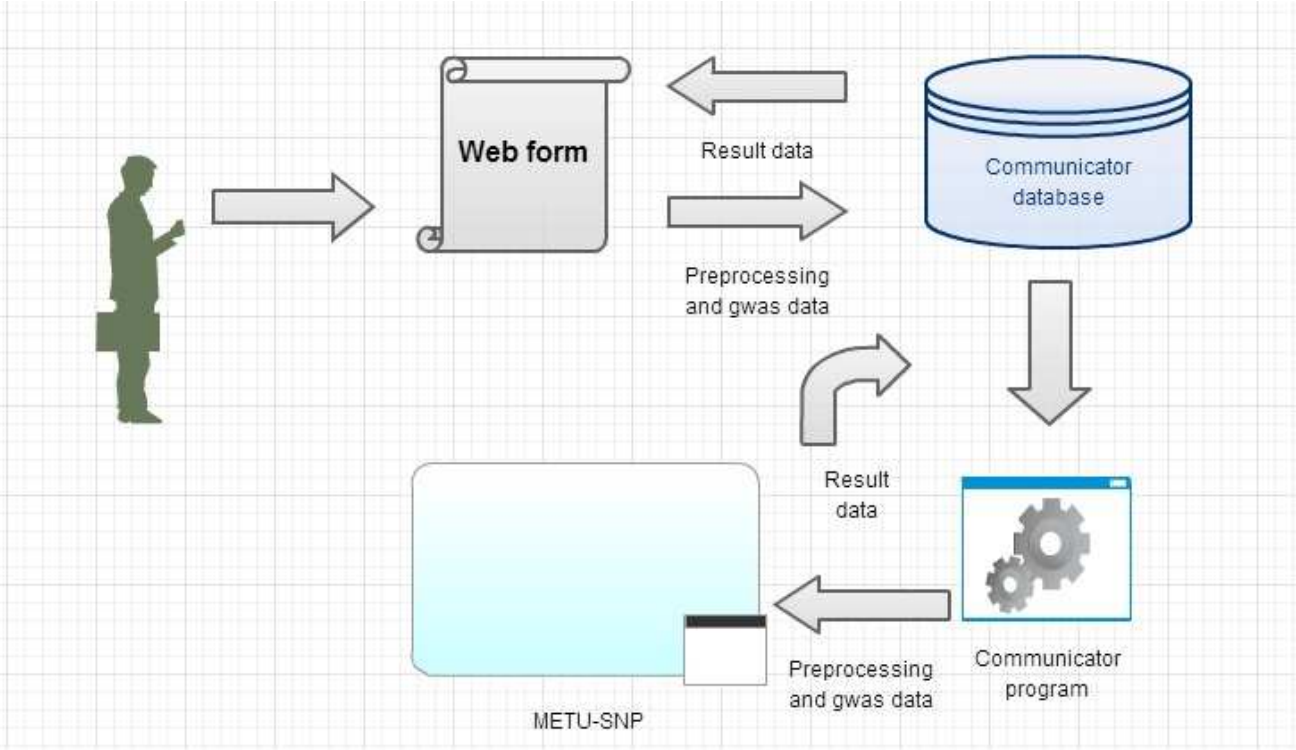

Figure 3.5: piSnp General Structure

the user input from the website. The same inputs exist in the desktop application, this means the communicator program transfers the user commands to the desktop application.

The input files may be very large (up to1 GB for a 1000 patient study) therefore user interface only accepts these files as compressed(.zip). The extraction operation is made by the communicatior program. It extracts and writes the extracted files into the parameters by the help of system commands.

Ped file and map file string values mean the paths of user designed Plink input values [60]. The web sites uploads the files and points the paths into the database. However there can be binary plink inputs which are bim, fam, bed files. If the communicator program finds these values as not null, it validates binary inputs instead of map and ped files. Since binary input files are independent of the user errors.

3.2.3 pi-SNP Web Interface

Pi-SNP web interface(forms) are designed by our group, but implemented by the company called UserSpots [61]. The forms designed as php based files. The web site has a tutorial, demo and blog parts which help users understand the aim and structure of piSnp. The web site can be accessed by getting user accounts, this lets users to see the results of their analysis and create multiple analysis projects with

Table 3.1: Communication Table

Id	integer
Ped file(path)	string
Map file(path)	string
Bim file(path)	string
Fam file(path)	string
Bed file(path)	string
Is genetic distance selected	integer(2)
Minor allele frequency value	float
Is minor allele frequency selected	integer(2)
Snp missingness rate value	float
Is SNP missingness rate selected	integer(2)
Individual missingness rate value	float
Is individual missingness rate selected	integer(2)
HWE equilibrium value	float
Is HWE equilibrium value selected	integer(2)
P value threshold value	float
Combined p value for genes	float
Is combined p value for genes selected	integer(2)
Min SNP value	integer
Is min SNP value selected	integer(2)
Max SNP value	integer
Is max SNP value selected	integer(2)
Combined p value for pathways	float
Is combined p value for pathways selected	integer(2)
Number of significant genes	integer
Is number of significant genes selected	integer(2)
Percentage of significant genes	integer
Is percentage of significant genes selected	integer(2)

different data. Figure 3.6 summarizes the workflow of the web interface, which can be access through http://pi-snp-test.ii.metu.edu.tr/ and will be moved to http://pi-snp.metu.edu.tr after the beta testing is completed.

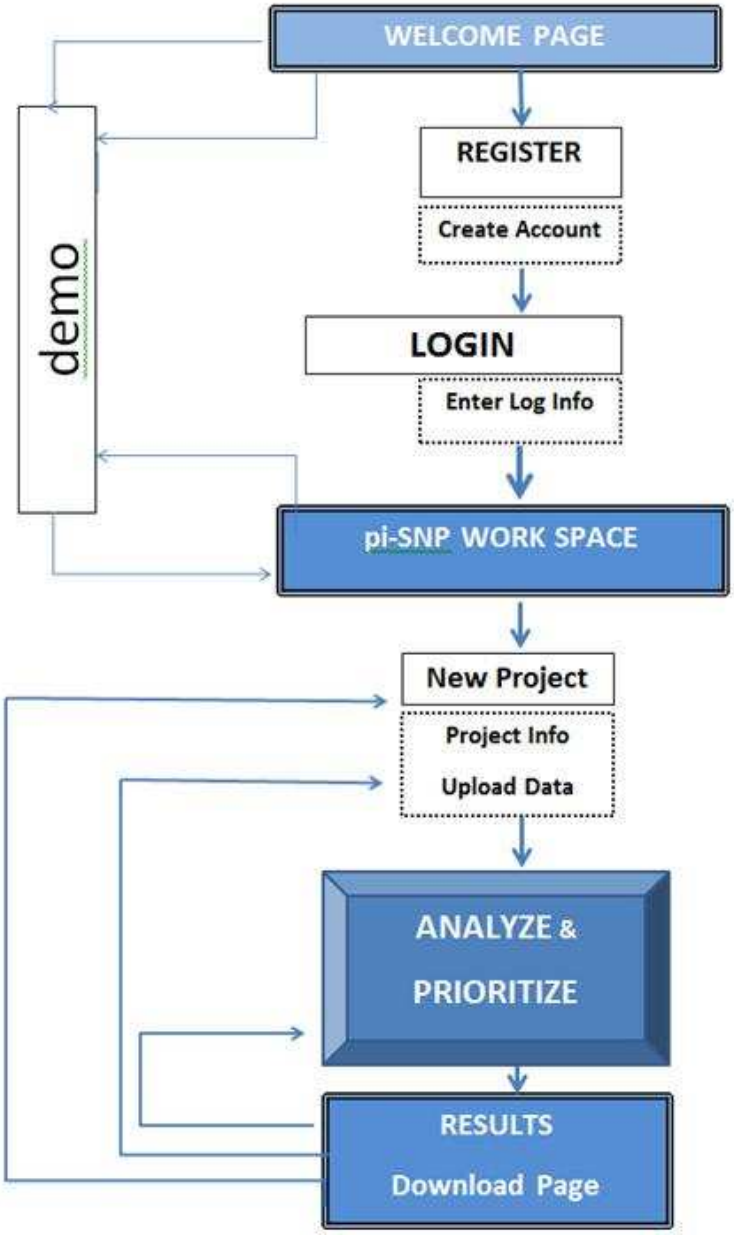

Figure 3.6: piSnp Web Site Workflow

A user can create multiple projects after logging in. A project means a single analysis, therefore users can see their past analysis as projects. Analysis is done in two steps. In the first step diseases are selected and source files are uploaded. There are two options to make the upload. User generated map and ped files can be uploaded, also machine generated binary files(bim, fam, bed) can be uploaded. All of the files in one group must be uploaded to be able to continue the analysis. Besides the file extensions

must be ”.zip”, since some files may be very large. Then in the second step quality control filtering thresholds, p-value threshold for individual SNPs, p-value threshold for genes and p value threshold for pathways for the combined p-value calculaitons are collected for the analysis. See Figure 3.7 for the analyis second web form of piSnp.

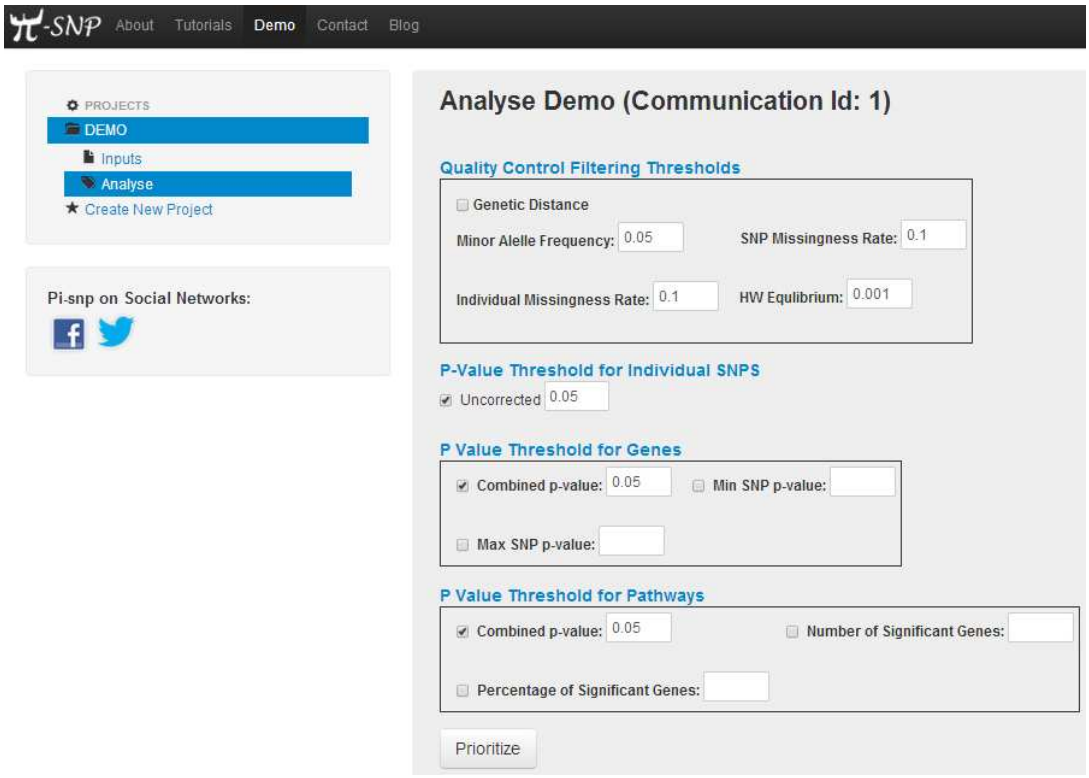

Figure 3.7: piSnp Web Site

Analysis continues at the server end, and informs the user users by mail system when the results are ready. The METU-SNP application in the server side make the analysis and get the results. It writes the results to three seperate tables (SNP, gene and pathway information with the p-values). Moreover it fills a table called ResultProgress to inform the web site about ongoing analysis and its result. The web site shows the results as three tables in seperate tabs. As the size of the results files are also large, the tables are shown inparts. Moreover users can export the results as text files.

3.3 METU-SNP Modifications for Linux

The METU-SNP desktop application is only compatible with Windows Operating System, but at the server side we prefer to have linux(ubuntu) [61] for stability and maintenance purposes. Therefore modifications for operating system adaptation was neccesary.

The essence of the compatibility problems was about the third party programs called

within METU-SNP application. All of the third party programs were in the windows 32 bit executable format [62].

As explained in METU-SNP application chapter PLINK is a widely used statistics tool in the program. The desktop version of plink was not the latest and it was only for the windows operating system. The latest version was not the same as the used version, since some parameters were changed and we were getting errors. So, the PLINK's calling parts in the application had to be changed with the new linux version. This was also necessary to call the programs in a linux environment.

In windows and linux there are differences about calling programs and system commands. For example program names are written starting with "./" in linux. So all the program names are changed with that prefix. For zipping and unzipping the files, in windows a third party program is used; but in linux it is changed into format of a command(linux has commands to make zip operations). Linux has more strict file security policies. Executable or any other readable files need extra permissions for user types. Before using any third party program permission of executing the program must be given. METU-SNP uses text files to transfer system commands among the program parts. One program part writes the commands and parameters, then another part reads and executes the command. Therefore giving permissions for these transfer text files was also necessary.

The libraries used for basic java operations are not the same as the original METUSNP application. The application was written with the windows applicable libraries dated 2008. Because of the new environment(linux) of pi-SNP we have to use linux applicable libraries. Therefore updating the libraries changed some functions. For example some string operations have different arguments in the new libraries. After getting errors for these functions minor changes are made according to the new library structure.

Some format converter third party programs used in the application are hard to found (they can be found only at its programmer's page). And when the working environment for these programs was only Windows, the source code is investigated throughout the web. From the source code a working version on linux is compiled and created.

The communication among the application and the web site was done by the communicator program. However for the result of the analysis the communication is done by the application. A modification which accesses the database by inserting information about the ongoing analysis' progress is made. The information of analysis start and finish is written into this table.

CHAPTER 4

DISCUSSIONS

METU-SNP is an operating system dependent desktop application, which requires downloading the whole software with its dependencies and its database, which build in 2008 and is outdated now. In bioinformatics access to accurate and up-to-date data is crucial, along with easy access to the analysis programs. Therefore in this study, we have built the iSNP database to be used with METU-SNP, which can be periodically updated, and transformed the METU-SNP application into a web service for wide and easy access.

There are many bioinformatics data sources but finding reliable and the applicable data for our studies was a critical process. While building the integrated SNP database, we have considered few different sources, such as html and ftp. RDF based database Bio2RDF has been selected as the source for the biological data, as it had standarts required to for the data retrival, also additional languages like SPARQL help us to use this standart formatted data.

When the Bio2RDF sources as identified a stand alone program is coded as an executable file in Java, which downloads the required content from Bio2rdf into the local mysql database. This operation can be run anytime an update to the biological databases are released, which will update the local iSNP database used by the pi-SNP web service.

There also few options for the transformation of the windows based METU-SNP desktop application into web environment. Along with transforming the need code and the third party applications into the linux environment that is used by the server a communicator program is developed to transfer the data between the web site and the METU-SNP.

The pi-SNP web service is also developed to be able to provide easy access to the METU-SNP functions and iSNP data. The goal was to develop an interface with high usability that will allow users to access their own workspaces and store their data and analysis results. The pi-SNP web server can be access through http://pi-snp-test.ii.metu.edu.tr/ . When the work on its beta testing is completed it will be moved to http://pi-snp.metu.edu.tr.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The two main goals of our study were; First, the automation of the METU-SNP database updates and second the transformation of the METU-SNP application into web environment. So, an updater program is developed that fetches data from predefined Bio2RDF sources that are required for the analysis of the SNP data. The automation of the METUSNP database updater, now provides anyone an one-click operation instead of manual database operation to access up-to-date SNP data in various biological resources. Additionally the integration of the METU-SNP application to the web environment is completed. The operations of preprocessing of SNPs and GWAS analysis can be done through a web browser. While the up to date iSNP database enables us to do more reliable annotation of the SNP and gene data, now researchers can reach our application worldwide by using the pi-SNP web site, server on METU servers.

There are few ongoing work to improve our services. The study on RDFizing process of dbSNP data is going on, which provide access to dbSNP data through iSNP in the future, and will also serve the bioinformatics community as an additional resource within the Bio2RDF project.

The current version of pi-SNP web service is not optimized for multiple users, so when there is a request for analysis it is added to a que. A multi process algorithm and parallelization of the code is within our short term plans, which will speed up the analysis and shorten the waiting time on the server.

REFERENCES

[1] D. Altshuler, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, pp. 52–58, 2010.

[2] D. Crawford and D. Nickerson, "Definition and clinical importance of haplotypes," *Annu Rev Med*, vol. 56, pp. 303–320, 2005.

[3] N. Tahri-Daizadeh and D. Nickerson, "Automated detection of informative combined effects in genetic association studies of complex traits," *Genome Research*, vol. 13, pp. 1952–1960, 2003.

[4] L. Çarkacıoğlu, C. Gedikoğlu, and Y. Aydın Son, "isnp: an integrated, automatically updated snp database," 2012.

[5] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific Am*, pp. 34–43, 2001.

[6] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic web revisited," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006.

[7] R. Çelebi, O. Gümüş, and Y. Aydın Son, "Use of open linked data in bioinformatics space: A case study," 2013.

[8] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: Towards a mashup to build bioinformatics knowledge systems," *Journal of Biomedical Informatics*, vol. In Press, Corrected Proof, pp. –, 2008.

[9] G. Üstünkar, "an integrative approach to structured snp prioritization and representative snp selection for genome-wide association studies," 2011.

[10] L. Kruglyak, "The road to genome-wide association studies," *Nature Reviews Genetics*, vol. 9, pp. 314–318, 2008.

[11] A. Chebotko and S. Lu, "Querying the semantic web: An efficient approach using relational databases," *LAP Lambert Academic Publishing*, 2009.

[12] T. Fisher, "32bit 64bit," *about.com*, 2009.

[13] "U.s. library of medicine," 2009.

[14] P. Raven, G. Johnson, and J. Losos, "Biology (7th edition)," *McGraw-Hill Co. NY*, 2008.

[15] J. Kelly, "Snps," *National Cancer Institute,USA*, 2008.

[16] M. Lercher and L. Hurst, "Human SNP variability and mutation rate are higher in regions of high recombination," *Trends in genetics*, vol. 18, no. 7, pp. 337–340, 2002.

[17] T. I. S. M. W. Group, “A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms,” *Nature*, vol. 409, 2001.

[18] G. Antoniou and F. vanHarmelen, *A Semantic Web Primer*. Cambridge, MA, USA: MIT Press, 2004.

[19] J. Hendler and D. McGuinness, “The darpa agent markup language,” *IEEE Intelligent systems*, vol. 11, no. 1, 2000.

[20] E. Miller, “The w3c’s semantic web activity: An update,” *IEEE Intelligent Systems*, vol. 19, pp. 95–96, C3, May 2004.

[21] Y. Liyang, “Linked Open Data,” *European Journal of Human Genetics*, vol. 18, pp. 409–466, July 2011.

[22] S. B. Davidson, C. Overton, and P. Buneman, “Challenges in integrating biological data sources,” *J. Comput. Biol.*, vol. 2, no. 4, pp. 557–572, 1995.

[23] D. Huynh, S. Mazzocchi, and D. R. Karger, “Piggy bank: Experience the semantic web inside your web browser.,” *J. Web Sem.*, vol. 5, no. 1, pp. 16–27, 2007.

[24] T. Gruber, “Toward principles for the design of ontologies used for knowledge sharing,” *Int J Hum Comput Stud*, vol. 43, no. 9, pp. 7–28, 1995.

[25] W. conributers, “http://www.w3.org/2004/OWL/,” 2004.

[26] B. conributers, “http://bio2rdf.org/bio2rdf-2007-02.owl,” 2007.

[27] D. R. Maglott, J. Ostell, K. D. Pruitt, and T. A. Tatusova, “Entrez gene: gene-centered information at ncbi.,” *Nucleic Acids Research*, vol. 39, no. Database-Issue, pp. 52–57, 2011.

[28] M. Ashburner, C. Ball, J. Blake, D. Botstein, and H. Butler, “Gene ontology: Tool for the unification of biology,” *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[29] “http://www.berkeleybop.org/ontologies/,” *http://www.genenames.org/*.

[30] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 52–55, 2002.

[31] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The protein data bank,” *Nucleic Acids Res*, vol. 28, pp. 235–242, Jan. 2000.

[32] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, “Chebi: a database and ontology for chemical entities of biological interest,” *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D344–D350, 2008.

[33] C. J. Bult, J. A. Blake, J. E. Richardson, J. A. Kadin, J. T. Eppig, R. M. Baldarelli, K. Barsanti, M. Baya, J. S. Beal, W. J. Boddy, D. W. Bradt, D. L. Burkart, N. E. Butler, J. Campbell, R. Corey, L. E. Corbani, S. Cousins, H. Dene, H. J. Drabkin, K. Frazer, D. M. Garippa, L. H. Glass, C. W. Goldsmith, P. L. Grant, B. L. King, M. Lennon-Pierce, J. Lewis, I. Lu, C. M. Lutz, L. J. Maltais, L. M. McKenzie, D. Miers, D. Modrusan, L. Ni, J. E. Ormsby, D. Qi, S. Ramachandran, T. B. K. Reddy, D. J. Reed, R. Sinclair, D. R. Shaw, C. L. Smith, P. Szauter, B. Taylor, P. V. Borre, M. Walker, L. Washburn, I. Witham, J. Winslow, Y. Zhu, and M. G. D. Group, "The mouse genome database (mgd): integrating biology with the genome.," *Nucleic Acids Res*, vol. 32, pp. D476–D481, Jan. 2004.

[34] B. conributers, "HUGO Gene Nomenclature Committee (HGNC), Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK,," *http://www.genenames.org/*.

[35] N. U. system requirements, "http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutilshelp.html/UserSystemRequirements,"

[36] J. Rapoza, "Sparql will make the web shine," *eWeek*, 2007.

[37] T. Segaran, C. Evans, J. Taylor, S. Toby, E. Colin, and T. Jamie, *Programming the Semantic Web*. O'Reilly Media, Inc., 1st ed., 2009.

[38] I. Herman, "W3c semantic web activity news - sparql is a recommendation," *w3.org*, 2008.

[39] Bikakis, "Xml and semantic web w3c standards timeline," *dblab*.

[40] I. Herman, "Eleven sparql 1.1 specifications are w3c recommendation," *w3.org*, 2013.

[41] O. Engineers, "Java se documentation at a glance," *Oracle.com*, 2013.

[42] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, pp. 559–575, Sept. 2007.

[43] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.," *American journal of human genetics*, vol. 81, pp. 1084–1097, Nov. 2007.

[44] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[45] M. Nothnagel, D. Ellinghaus, S. Schreiber, M. Krawczak, and A. Franke, "A comprehensive evaluation of SNP genotype imputation.," *Human genetics*, vol. 125, pp. 163–171, Mar. 2009.

[46] K. Hao, E. Chudin, J. McElwee, and E. E. Schadt, "Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies.," *BMC genetics*, vol. 10, pp. 27+, June 2009.

[47] P. I. de Bakker, M. A. Ferreira, X. Jia, B. M. Neale, S. Raychaudhuri, and B. F. Voight, "Practical aspects of imputation-driven meta-analysis of genome-wide association studies.," *Human molecular genetics*, vol. 17, pp. R122–R128, Oct. 2008.

[48] A. R. Pico, I. V. Smirnov, J. S. Chang, R.-F. Yeh, J. L. Wiemels, J. K. Wiencke, T. Tihan, B. R. Conklin, and M. Wrensch, "SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system," *Nucl. Acids Res.*, vol. 37, pp. D803–809, Jan. 2009.

[49] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, pp. 308–311, Jan. 2001.

[50] L. Conde, J. M. Vaquerizas, H. Dopazo, L. Arbiza, J. Reumers, F. Rousseau, J. Schymkowitz, and J. Dopazo, "PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes.," *Nucleic acids research*, vol. 34, pp. W621–625, July 2006.

[51] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent, "The UCSC Genome Browser Database: 2008 update.," *Nucleic Acids Res*, vol. 36, Jan. 2008.

[52] S. F. Saccone, R. Bolze, P. Thomas, J. Quan, G. Mehta, E. Deelman, J. A. Tischfield, and J. P. Rice, "SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study.," *Nucleic acids research*, vol. 38, July 2010.

[53] J. Osborne, J. Flatow, M. Holko, S. Lin, W. Kibbe, L. Zhu, M. Danila, G. Feng, and R. Chisholm, "Annotating the human genome with Disease Ontology," *BMC Genomics*, vol. 10, no. Suppl 1, pp. S6+, 2009.

[54] P. Du, G. Feng, J. Flatow, J. Song, M. Holko, W. A. Kibbe, and S. M. Lin, "From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations," *Bioinformatics*, vol. 25, pp. i63–i68, June 2009.

[55] G. Peng, L. Luo, H. Siu, Y. Zhu, P. Hu, S. Hong, J. Zhao, X. Zhou, J. D. Reveille, L. Jin, C. I. Amos, and M. Xiong, "Gene and pathway-based second-wave analysis of genome-wide association studies," *European Journal of Human Genetics*, vol. 18, pp. 111–117, July 2009.

[56] M. P. Sawicki, G. Samara, M. Hurwitz, and E. P. Jr., "Human genome project," *The American Journal of Surgery*, vol. 165, no. 2, pp. 258 – 264, 1993.

[57] http://www.ncbi.nlm.nih.gov/ contributers, "dbsnp summary @ONLINE," June 2013.

[58] J. Postel and J. K. Reynolds, "File transfer protocol," 1985.

[59] T. Berners-Lee, R. Fielding, and H. Frystyk, “Hypertext transfer protocol – http/1.0,” 1996.

[60] S. Purcell, “http://pngu.mgh.harvard.edu/ purcell/plink/,” 2009.

[61] U. contributers, “http://www.ubuntu.com/desktop,” 2013.

[62] M. Rouse, “http://whatis.techtarget.com/fileformat/EXE-Executable-file-program,”

Appendix A

Bio2Rdf

A.1 Bio2RDF sources

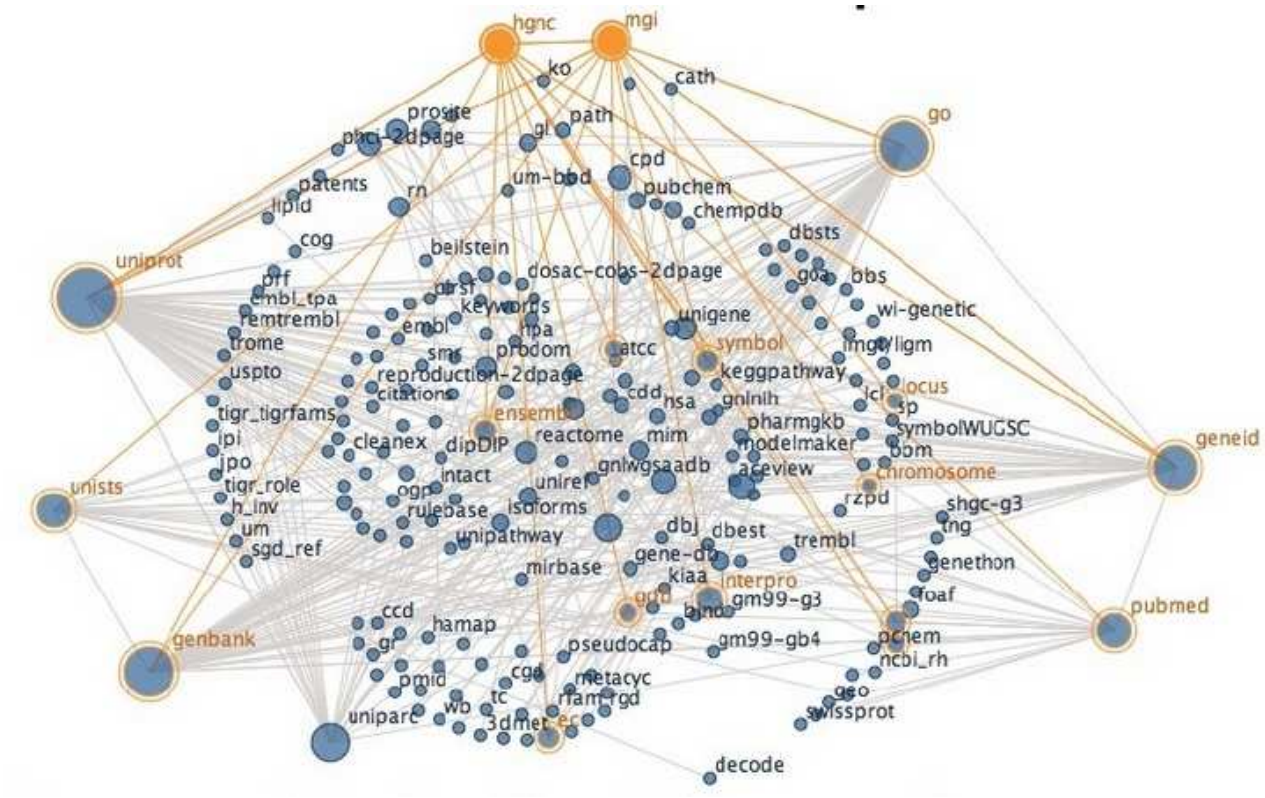

Figure A.1: Bio2RDF connections

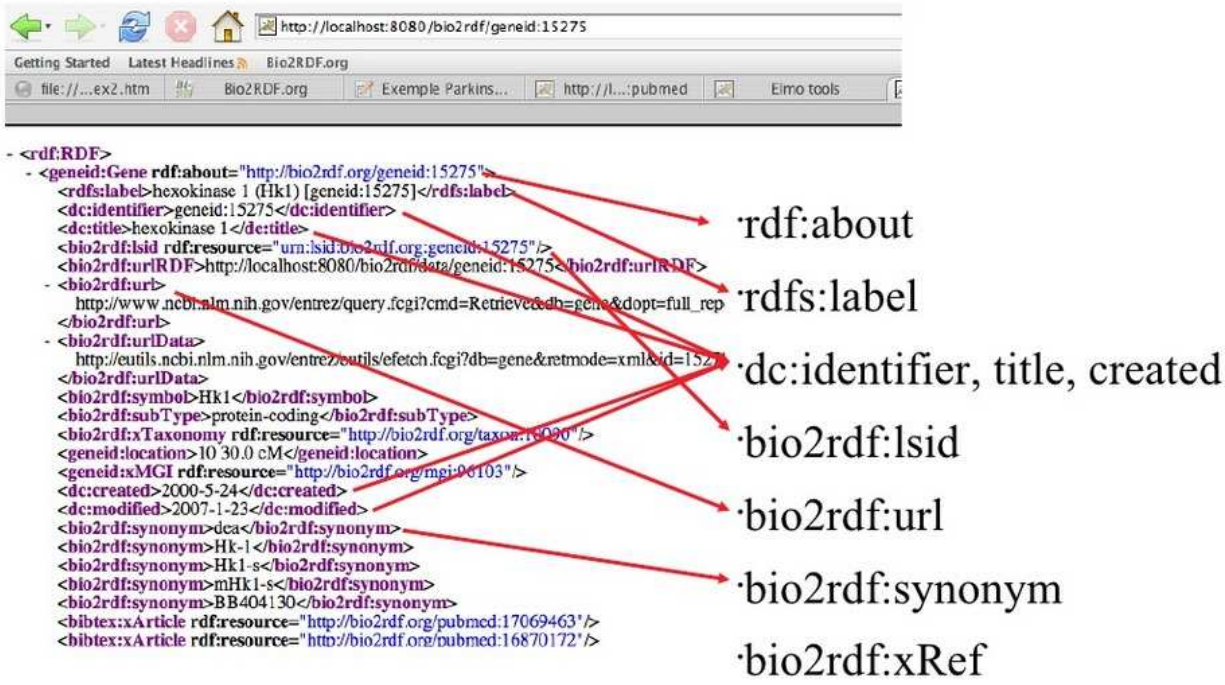

Figure A.2: Bio2RDF example of gene with id:15275

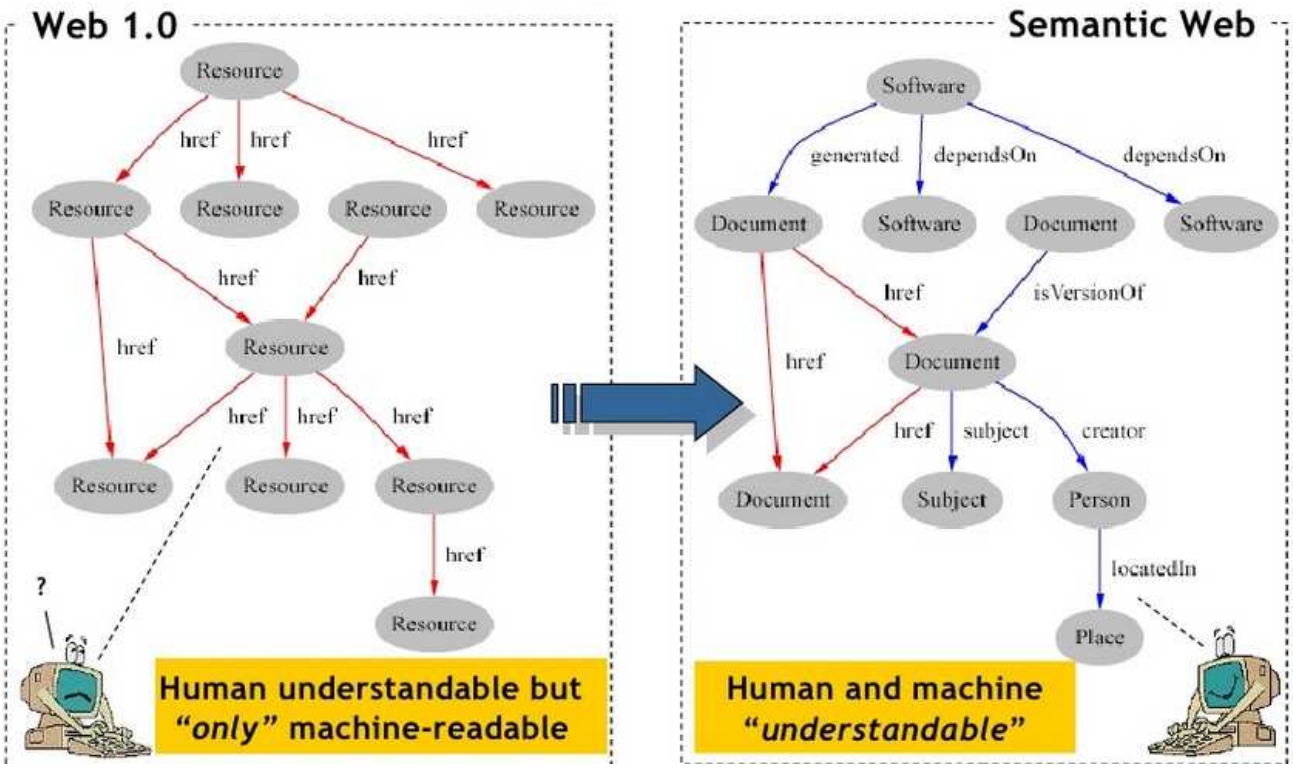

Figure A.3: Semantic Web

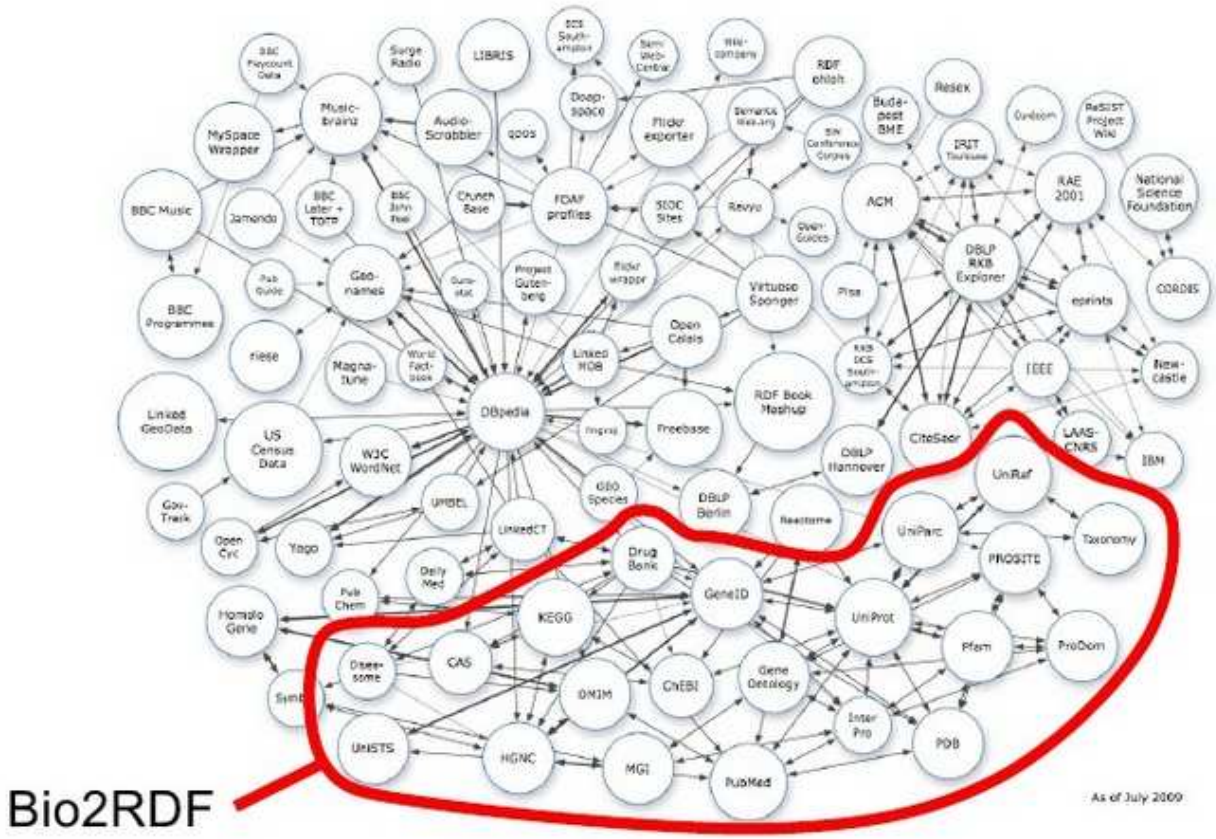

Figure A.4: Bio2RDF in Semantic Web

Appendix B

RDF

B.1 Rdfization

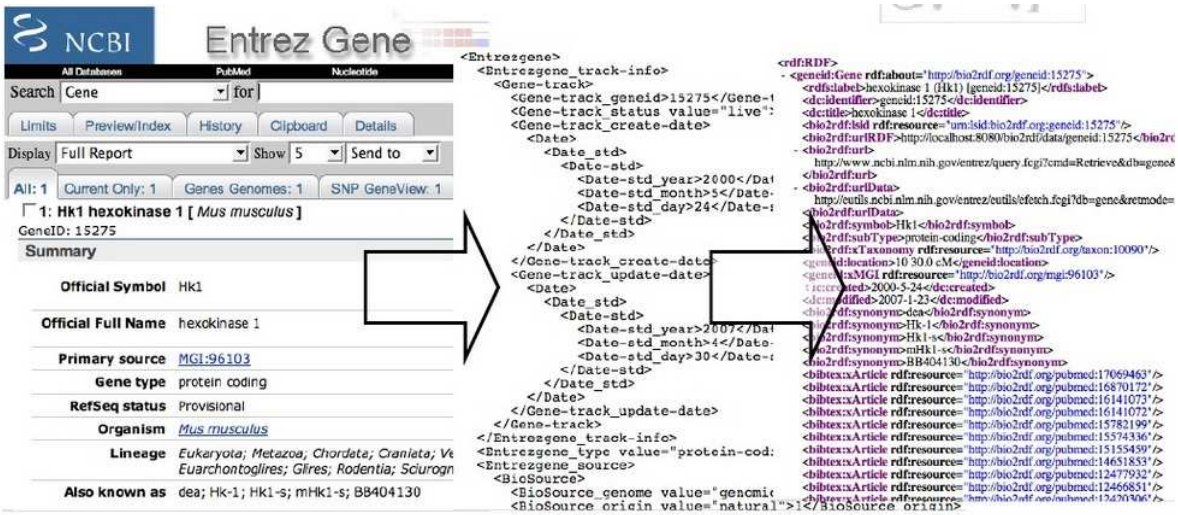

Figure B.1: Rdfization

Appendix C

SPARQL Examples

C.1 SPARQL Examples from iSNP

```
"PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>"+
        "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>"+
        "select distinct ?label" +
        "               where {" +
        "                  " +
        "                    ?assoc rdf:type <http://bio2rdf.org/ctd_vocabulary:Gene-Disease-Association> ." +
        "                    ?assoc rdfs:label ?label ." +
        "                    " +
        "                  " +
        "                  }" ;
```

Figure C.1: SPARQL Code for Gene Disease Association

```
"PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>"+
        "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>"+
        "select distinct ?label" +
        "               where {" +
        "                  " +
        "                    ?assoc rdf:type <http://bio2rdf.org/ctd_vocabulary:Disease> ." +
        "                    ?assoc rdfs:label ?label ." +
        "                    " +
        "                  " +
        "                  }" ;
```

Figure C.2: SPARQL Code for Disease Data

```
"PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>"+
        "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>"+
        "select distinct ?assoc ?label" +
        "                where {" +
        "                   " +
        "                     ?assoc <http://bio2rdf.org/ctd_vocabulary:pathway> ?kegger." +
        "                     ?kegger rdfs:label ?label ." +
        "                     " +
        "                   " +
        "                   }" ;
```

Figure C.3: SPARQL Code for Pathway Data

VITA

Bahcelievler mah.
Cankaya/Ankara

January 9th, 2014
ceyhungedikoglu@gmail.com
+90 505 751 7055

Professional Experience

2008–today	**Aselsan inc.** Expert Software Design Engineer

Education

2004–2008	**BSc in Computer Science, Bilkent University, Ankara**
1997–2004	High School, Mersin Anatolian High School, Mersin

Languages

Turkish	Mother tongue
English	**Fluent**

Publications

April 2012	Levent Carkacioglu, Ceyhun Gedikoglu, and Yesim Aydin Son. iSNP: An Integrated, Automatically Updated Snp Database. IEEE Xplore HIBIT'12, page 127-130
December 2011	Ceyhun Gedikoglu, Levent Carkacioglu, Yesim Aydin Son, iSNP: An Integrated, Automatically Updated SNP Database Server Over Web. BBC11, BeNeLux Bioinformatics Conference, Luxemburg