AUTOMATED INTEGRATION OF BIOMEDICAL INFORMATION FOR THE SUPPORT OF GENOME-WIDE ASSOCIATION STUDIES

by

ABHIJIT R TENDULKAR

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2008

ACKNOWLEDGEMENTS

I express my sincere gratitude towards my advisor Dr. Nikola Stojanovic for providing invaluable guidance throughout the project that helped me lot to conquer the problems and find out the way towards achieving the research goals. He has been a constant source of encouragement and benevolence to me at all times.

I also thank Dr. Robert Barber for his collaboration and constant support. Without him this project would not have been possible. I would also like to thank Dr. Ramez Elmasri for serving on my thesis committee.

I must mention special thanks to Dr. Bahram Khalili for advising and giving me proper direction during the course of my Master's studies. I thank all members of the Bioinformatics group for their support, guidance and help.

I appreciate constant moral and financial support from my parents. I also thank all of my friends who constantly supported and motivated me. Finally, I thank God who has showered on me the blessings and given me the strength to overcome all the troubles.

June 26, 2008

ABSTRACT

AUTOMATED INTEGRATION OF BIOMEDICAL INFORMATION
FOR THE SUPPORT OF GENOME-WIDE
ASSOCIATION STUDIES

Abhijit R. Tendulkar, M.S.

The University of Texas at Arlington, 2008

Supervising Professor: Nikola Stojanovic, PhD

Genome-wide association studies of the genetic underpinnings of complex phenotypes, and human diseases in particular, have been steadily gaining momentum over the past several years. Yet, the number of polymorphic sites in the human genome, including, but not limited to, Single Nucleotide Polymorphisms (SNPs) is so large that identifying the combination of these few which have a significant effect on the condition of interest remains an overwhelming task.

The goal of this thesis work was to identify biologically and medically relevant SNPs which could be the best possible candidates for further association studies. In this thesis we present a new networked solution, and a program GeneNAB implementing it, to the computational identification and ranking of SNPs likely to be relevant for the phenotype of interest, genome-wide. The architecture of our system is similar to that of the Distributed Annotation System (DAS), proposed by a team of prominent bioinformaticians several years ago. However, not all of the resources we use could follow the DAS protocol, so we employed a variety of methods for accessing different resources on the Internet needed for our study.

We start with a gene or a cellular pathway previously associated with the condition of interest and find SNPs in all other genes participating in the same pathway. We then rank these SNPs according to their likelihood to be biomedically relevant for the condition and report the ranked list flagging the top entries as candidates for further experimental work.

We have applied our system to the Toll-like receptor pathway which provides a mechanism for the development of inflammatory reaction in a variety of conditions, from infection to cell damage. Although many of our top-scoring SNPs still need experimental validation, we have indeed successfully located several which have been previously confirmed as medically relevant. We expect that the output of our software will be useful to guide further laboratory and clinical studies of groups of SNPs affecting any condition of interest.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Genetics

The genetic information which plays a key role in determining the structure and functions of a living organism is stored in its Deoxyribonucleic Acid, or DNA. DNA is inherited in offspring from the parents (two in the case of diploid organisms). We use the term genetic information, because the DNA, i.e. the genome does not itself perform any active role in the structural formation and functioning of the organism. Instead, DNA sequence is used to produce proteins by the complex series of processes. Proteins can either form a part of a structure of the cell (and, by extension, the organism), have an enzymatic role in enabling various chemical processes to take place or serve as signals triggering cellular processes in response to conditions in their environment.

1.1.1 DNA

DNA is a component of a cell. In the higher organisms it is present in the form of nuclear and mitochondrial DNA, and the former is divided into chromosomes, visible during the division of the cell. A schematic of the structure of a cell and chromosomes in its nucleus is shown in the Figure 1.1 below. The general structure of DNA is that of a double helix, as shown in Figure 1.2.

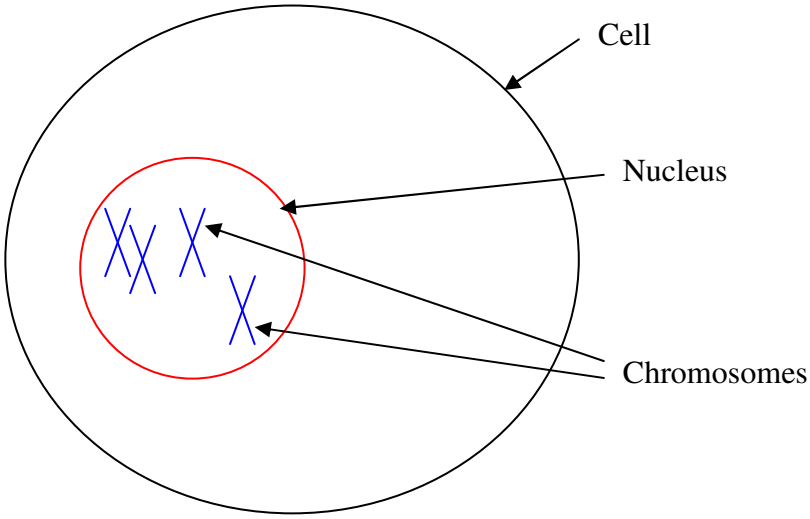

Figure 1.1 Schematic of cell structure and chromosomes

In the double helical structure of DNA two strands of polymers run anti-parallel to each other, i.e. one strand runs 5' to 3' whereas the other strand runs in 3' to 5' direction. The backbone of the polymer consists of sugars and phosphate groups which are joined by ester bonds. One of four types of molecules called nitrous bases is attached to each sugar, forming a nucleotide. These four nucleotides are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T).

Bases on the two strands pair with each other by hydrogen bonds. Adenine pairs with Thymine and Cytosine pairs with Guanine, as shown in Figure 1.3.

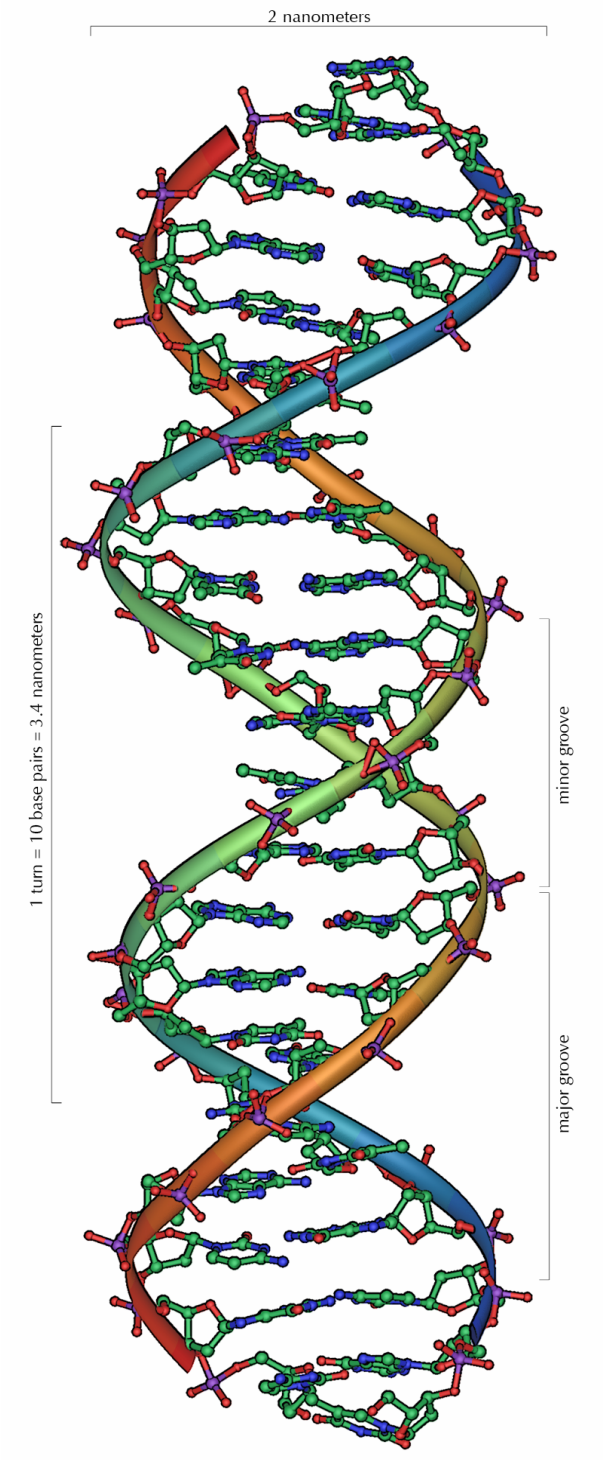

Figure 1.2 The structure of part of a DNA double helix (source: http://en.wikipedia.org)

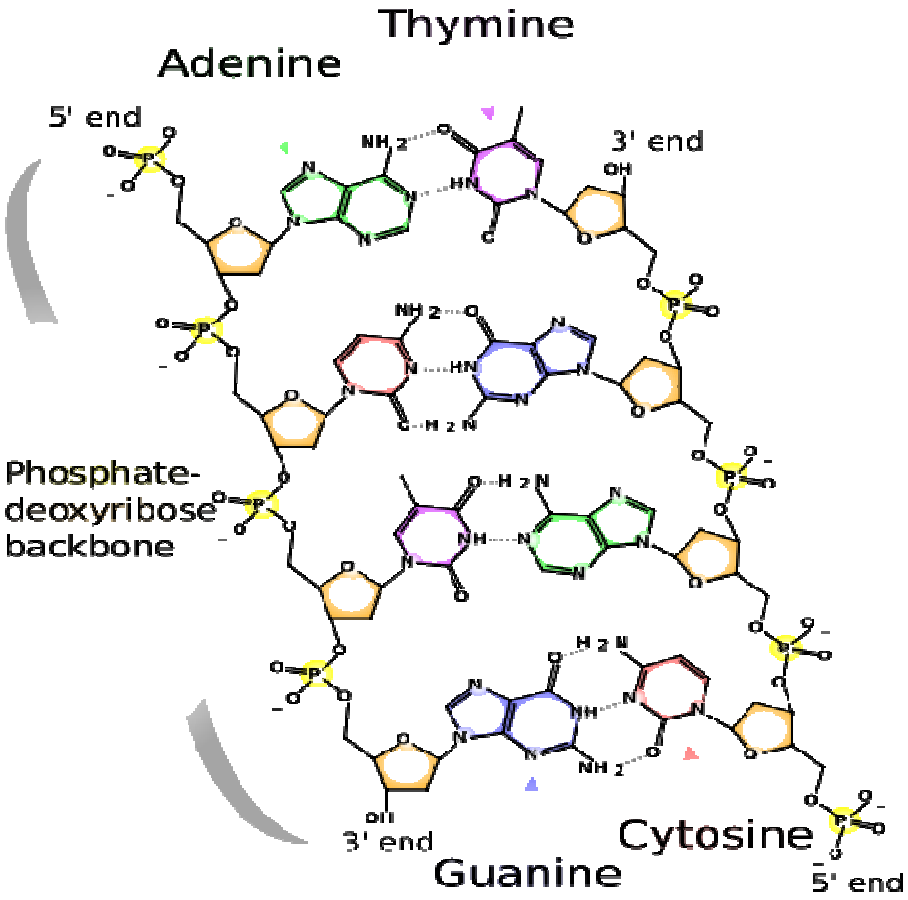

Figure 1.3 Base pairing (Source: http://en.wikipedia.org)

1.1.1.1 Genes and the genome

A gene is a discrete unit within the DNA polypeptide chain. It includes a region coding for a protein, as well as segments preceding and following the coding region (leader and trailer). In higher organisms it also contains intervening sequences (introns) between the individual coding segments (exons).

The genome of an organism is the complete DNA molecule i.e. the complete sequence of nucleotides. In higher organisms most of the DNA is non-coding, and very small part of sequence consists of protein-coding genes.

DNA information is first copied into mRNA by the process of transcription, and proteins are then synthesized using the information in mRNA by the process of translation. Proteins

consist of amino acids, and three nucleotide bases code for a single amino acid, as shown in Table 1.1 below. In mRNA, thymine (T) is replaced by uracil (U) and the deoxyribose sugar in the DNA backbone is substituted by ribose.

Table 1.1 RNA codon table

		2nd base			
		U	**C**	**A**	**G**
1st base (5' end)	**U**	UUU (Phe/F) UUC (Phe/F) UUA (Leu/L) UUG (Leu/L)	UCU (Ser/S) UCC (Ser/S) UCA (Ser/S) UCG (Ser/S)	UAU (Tyr/Y) UAC (Tyr/Y) **UAA Ochre (STOP)** **UAG Amber (STOP)**	UGU (Cys/C) UGC (Cys/C) **UGA Opal (STOP)** UGG (Trp/W)
	C	CUU (Leu/L) CUC (Leu/L) CUA (Leu/L) CUG (Leu/L)	CCU (Pro/P) CCC (Pro/P) CCA (Pro/P) CCG (Pro/P)	CAU (His/H) CAC (His/H) CAA (Gln/Q) CAG (Gln/Q)	CGU (Arg/R) CGC (Arg/R) CGA (Arg/R) CGG (Arg/R)
	A	AUU (Ile/I) AUC (Ile/I) AUA (Ile/I) **AUG (Met/M) (START)**	ACU (Thr/T) ACC (Thr/T) ACA (Thr/T) ACG (Thr/T)	AAU (Asn/N) AAC (Asn/N) AAA (Lys/K) AAG (Lys/K)	AGU (Ser/S) AGC (Ser/S) AGA (Arg/R) AGG (Arg/R)
	G	GUU (Val/V) GUC (Val/V) GUA (Val/V) GUG (Val/V)	GCU (Ala/A) GCC (Ala/A) GCA (Ala/A) GCG (Ala/A)	GAU (Asp/D) GAC (Asp/D) GAA (Glu/E) GAG (Glu/E)	GGU (Gly/G) GGC (Gly/G) GGA (Gly/G) GGG (Gly/G)

In a sequence of mRNA, the codon AUG (start codon) indicates the start of the coding region. The coding part of the gene ends with one of the stop codons UAA, UAG and UGA. Amino acids and their corresponding 1-letter and 3-letter codes are listed in Table 1.2.

Table 1.2 3-letter and 1-letter codes for amino acids

Amino Acid	**3-letter code**	**1-letter code**
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

1.1.2 Polymorphism

The human genome consists of over 3 billion base pairs. Out of these about 1.5% are protein-coding genes and the remaining sequence contains RNA genes (non-coding RNA), regulatory sequences, introns and (controversially) “junk” DNA, which is by far its largest part.

In any two individuals (humans), the genome sequence differs at approximately 3 million bases (i.e. 0.1%). And the variations are due to Single Nucleotide Polymorphism (SNP), Deletion Insertion Polymorphism (DIP), repetitive sequences (which can sometimes be highly variable from person to person) and other sources which only recently started attracting the attention of the scientific community.

SNPs are single base positions in the genome at which different sequence alternatives (alleles) exist. within the normal population. By some accounts, SNPs are the most common form of genetic variations in mammals [Brookes 1999], which exist in any part of genomic sequence including gene exons, introns, regulatory sequences or any other loci. An example of an SNP position in DNA sequence is shown in Figure 1.4.

Allele frequencies indicate the proportion of variants in a particular population or the complete human population, and alleles are classified as major or minor depending on their frequency. For example, if “C” is located at a particular base position in 60% of population and “T” in 40%, then “C” is major allele and “T” is minor allele. The frequency corresponding to “C” (60%) is major allele frequency and that corresponding to “T” (40%) is minor allele frequency.

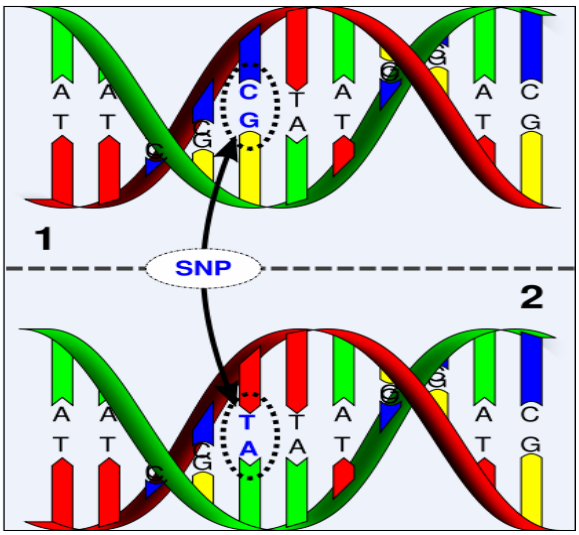

Figure 1.4 SNP (Image source: http://en.wikipedia.org)

The anticipation is that the allele information at SNP loci, referred to as a genotype, will provide a basis for assessing the individual's susceptibility to disease and perhaps the optimal choice for therapy [Hirschhorn et al. 2002; Hirschhorn et al. 2002]. A major challenge in realizing these expectations is in understanding precisely which of the millions of known SNPs affect disease. SNP sites are found approximately once in every 100 to 300 bases across the human genome [Kruglyak and Nickerson 2001] and complex traits such as most common diseases are influenced by combinations of loci which may not even lie on the same chromosome.

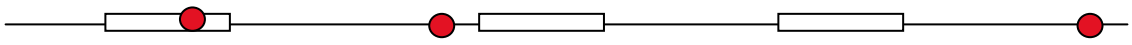

Figure 1.5 Possible layout of SNPs relative to a hypothetical 3-exon eukaryotic gene.

The SNPs may cause a disease or a condition, increase the susceptibility to disease or they may be responsible for variable response to medications by different patients. And they may lie in an exon of a gene, in a promoter region which control regulation of a gene or in any other locus, as shown in Figure 1.5.

1.1.3 Genome Projects

The Human Genome Project (HGP) [International Human Genome Sequencing Consortium, 2001] was an international effort with the primary focus to sequence the entire human genome, as it was believed that it would elucidate the molecular etiology of human disease. In the process many genes have been identified, and their number has been estimated at about 25,000.

The International HapMap Project [The International HapMap Consortium, 2005], a sequel to the Human Genome Project,was an international effort to identify and catalog genetic differences in human beings. It recorded the location and linkage information for many common human genetic variants. HapMap was a collaboration among scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria and the United States. All of the information generated by the project is available in the public domain.

The term haplotype has different meanings in different contexts. Haplotype is the genetic constitution of an individual chromosome, and it may refer to a single locus or an entire genome. In diploid organisms, haplotype is one member of the pair of alleles for each locus. In the context of the HapMap project, the term haploytpe refers to a set of SNPs on a single chromatid (one of the two copies of chromosome) that are statistically associated, i.e. the set of SNPs whose alleles can be determined by its subset, called tag SNPs.

Individuals who carry a particular SNP allele at one site often predictably carry specific alleles at other nearby variant sites. This correlation is known as linkage disequilibrium (LD). A tag SNP is representative single SNP in a region of the genome with high linkage disequilibrium. It is thus possible to identify genetic variation without genotyping every SNP in a chromosomal region. For example, consider following sequences:

……AAAGTCGTA…..

……AAGGTCCTA…..

……AAAGTCCTA…..

……AAGGTCCTA…..

……AAGGTCCTA…..

……AAAGCCCTA…..

Whenever there is “G” at 3rd position, there is always “C” at 7th position. Thus “G” at 3rd position alone determines the neighboring sequence. Hence “G” can be used as a tag SNP in this case of linkage disequilibrium.

HapMap i.e. Haplotype Map is a catalog of common genetic variants in human beings. HapMap describes what these variants are, where they occur in our DNA, and how they are distributed among people within populations, and among populations in different parts of the world.

HapMap data can be used for finding genetic variants which affect health, disease, and individual responses to medications and environmental factors. It can also help biomedical researchers find genes involved in disease and responses to therapeutic drugs.

In Phase I of the HapMap project more than one million SNPs for which genotypes have been obtained in 269 DNA samples from four populations have been recorded. The data also included ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. The four considered populations were Yoruba people of Ibadan, Nigeria, Japanese from Tokyo, Chinese from Beijing and US residents from Utah having northern & western European ancestry.

Phase II of HapMap genotyped another 4.6 million SNPs in the HapMap samples and ENCODE [The ENCODE Project Consortium, 2007] regions in additional members of each HapMap population, as well as in samples from additional populations. These results should provide an insight at the robustness and transferability of LD inferences and previously selected tag SNPs.

1.1.4 Genome-Wide Association Studies (GWAS)

Genome-Wide Association Studies (GWAS) concern the association between the SNPs or combinations of SNPs and the condition they affect. . The HapMap project, was driven towards finding genetic variants that influence health, disease and individual response to medications.

The candidate gene approach has been used frequently in allelic association studies of complex disease. This methodology employs knowledge of disease pathobiology together with results of animal, molecular, cellular and other basic science studies to select genes which are likely to be involved in disease etiology. Polymorphisms within candidate genes are then identified and evaluated for correlation with disease. Many studies over the past decade have used this strategy to establish that genetic polymorphism explains a significant portion of the individual variation in susceptibility to common illnesses such as hypertension, vascular disease, cancer, diabetes, and autoimmune disorders [Lander and Kruglyak 1995; Lander 1996]. Using a candidate gene approach, it has also been demonstrated that an individual's genetic background significantly influences susceptibility to environmental triggers [Lander and Kruglyak 1995]. For example, polymorphisms in methylene tetrahydrofolate reductase, Factor V, angiotensin converting enzyme, and CKR-5 genes are associated, respectively, with neural tube defect risk, arteriosclerosis, venous thrombosis, heart disease, and HIV resistance [Lander 1996].

In contrast, genome–wide association methods screen SNPs without prior assumptions regarding the biological role of particular genes or pathophysiology of the disease. Until recently, this approach was impractical due to the expense associated with the large number of assays required for such a study. Methods that do not focus on a particular subset of candidate genes require the determination of large numbers of genotypes ($7.5X10^8$ genotypes in a sample of 1500 to 2000 subjects). Although recent advances in high–throughput genotyping methodology have made genome–wide association scans (GWAS) more technically and

financially practicable, prioritization of SNPs that are most likely to cause disease is still of great benefit in studies subsequent to the initial GWAS. In addition, while genome wide surveys have become relatively inexpensive, they remain too costly for many researchers and require more subjects than may be available for rare diseases. Therefore, the candidate gene approach remains an important tool in the armamentarium of biologists concerned with resolving the molecular basis of human diseases.

In order to identify disease causing SNPs, we use the hypothesis that disease causing SNPs are located in highly conserved regions of the genome. To find highly conserved regions, we need to compare human genome sequences with sequences of other species using sequence alignment tools like BLAST [Altschul et al. 1990]. We also need to consider phylogenetic distance of human from species whose sequence is to be compared. Phylogenetic distance is the measure of diversity between two species. When a SNP causes change in amino acid, we evaluate significance of that change using BLOSUM62 [Henikoff and Henikoff 1992] substitution matrix which assigns substitution score for all pairs of amino acids.

1.2 Bioinformatics Databases

For GWAS we need biological information such as cellular pathways related to condition of interest, genes participating in these pathways, SNPs in these genes, information about SNPs such as population diversity data etc. All this information is not available at any single source, and hence we need to access various databases such as KEGG (Kyoto Encyclopedia of Genes and Genome), NCBI (National Center for Biotechnology Information) databases, dbSNP in particular, and Ensembl and UCSC genome browsers.

1.2.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG, The Kyoto Encyclopedia of Genes and Genomes [Kanehisa et al. 2008], is a computer database of biological networks hosted by the Bioinformatics Center at the Institute for Chemical Research, Kyoto University, and supported by grants from the Bioinformatics Research and Development of the Japan Science and Technology Agency, the Japan Ministry

of Education and the NIH/NIGMS Consortium for Functional Glycomics. It is comprised of numerous building blocks, including genes and proteins (KEGG GENES), endogenous and exogenous chemical building blocks (KEGG LIGAND), cell signaling and enzymatic networks (KEGG PATHWAY), as well as hierarchies of various biological objects (KEGG BRITE). This suite provides a robust reference base for linking genomes to biological systems, and we use it in order to identify SNPs lying within the genes participating in our target pathways.

1.2.1.1. KEGG Pathway and KGML (KEGG Markup Language)

KGML is an XML representation of the KEGG graph objects, especially KEGG pathway maps which are created and updated manually. It enables automatic drawing of KEGG pathways, as well as computational analysis and modeling of protein networks and chemical networks.

KEGG pathway maps are images representing networks of interacting molecules responsible for specific cellular functions. KEGG pathways are of two types:

1. reference pathways (manually drawn)
2. organism-specific pathways (computationally generated based on reference pathways).

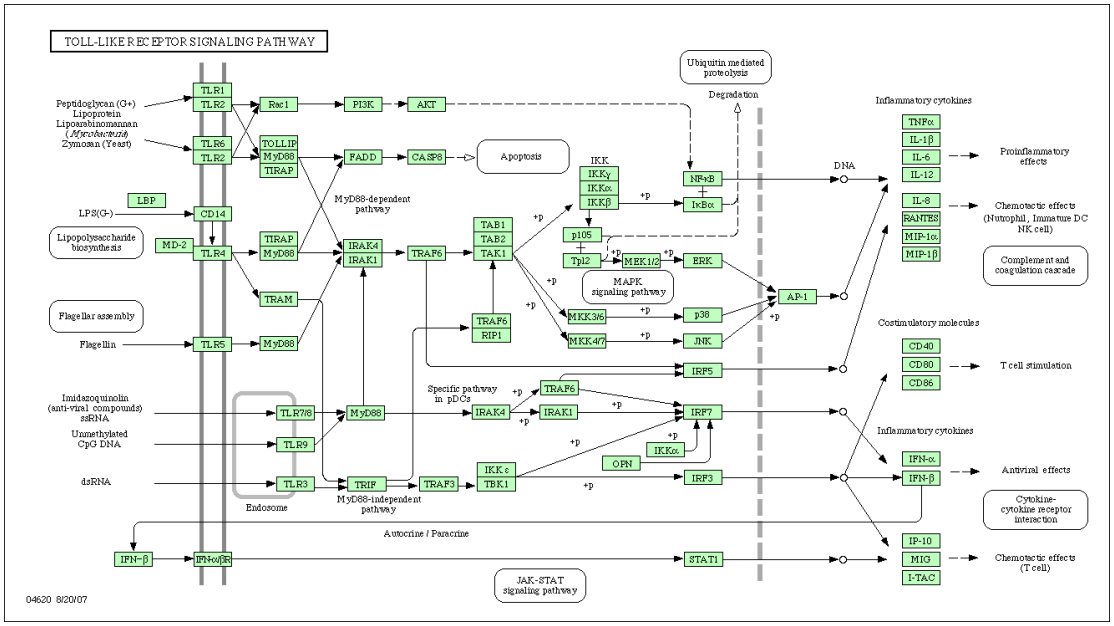

Figure 1.6 Example of a KEGG pathway (Source: www.genome.jp/kegg)

A KGML file contains the information about graphical objects and their relations in KEGG pathways. It also links the information in the KEGG GENES database.

In KGML the **pathway** element specifies one graph object with the **entry** elements as its nodes and the **relation** and **reaction** elements as its edges. The relation and reaction elements indicate the connection patterns of rectangles (gene products) and the connection patterns of circles (chemical compounds), respectively. Two types of graph objects, consisting of entry and relation elements and entry and reaction elements, are respectively called the protein network and the chemical network. Since a metabolic pathway can be viewed both as a network of proteins (enzymes) and as a network of chemical compounds, another distinction of KEGG pathways is:

- metabolic pathways can be viewed as both protein networks and chemical networks
- regulatory pathways can be viewed as protein networks only.

The pathway element is a root element, and one pathway element is specified for one pathway map in KGML. The entry, relation, and reaction elements specify the graph information, and additional elements are used to specify details about the nodes and edges of the graph.

An overview of KGML can be shown in Figure 1.7.

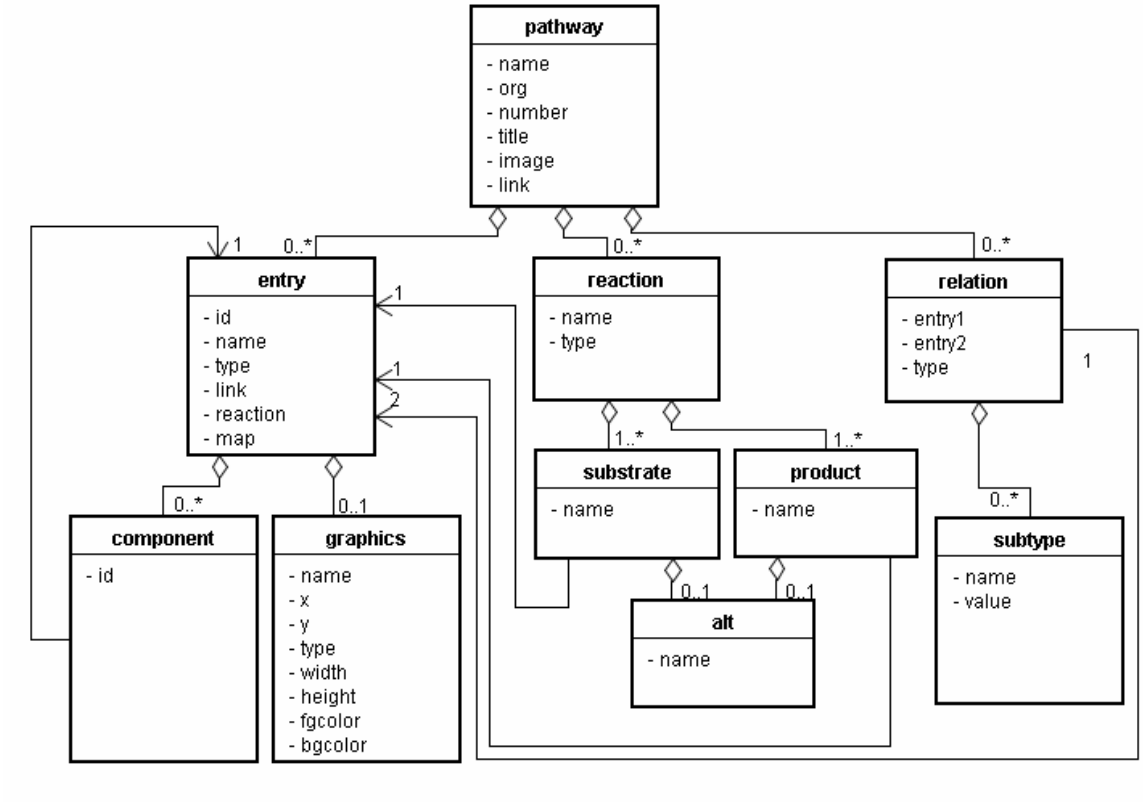

Figure 1.7 KGML overview (Source: http://www.genome.jp/kegg/docs/xml/)

Each KGML file may be acquired from: http://www.genome.jp/kegg/xml/

The KEGG pathway maps are divided into the following six categories.

1. KEGG reference metabolic pathways - enzymes are linked to the ENZYME database.
2. KEGG reference regulatory pathways - proteins are linked to the ENZYME database.
3. KEGG reference metabolic pathways linked to KO - enzymes are linked to the KO database.
4. KEGG reference regulatory pathways linked to KO - proteins are linked to the KO database.
5. KEGG organism-specific metabolic pathways - enzymes are linked to the GENES database.
6. KEGG organism-specific regulatory pathways - proteins are linked to the GENES database.

KGML files are updated daily and they are available for download at:

ftp://ftp.genome.jp/pub/kegg/xml/

1.2.1.2. Communication with KEGG

KEGG provides several mechanisms for accessing the information, and our method of choice was through the use of its Applications Programmer Interface (API). KEGG API is based on SOAP, a popular information exchange protocol based on the Extended Markup Language (XML) technology. It is implemented as a set of functions working through the Remote Procedure Call (RPC) mechanism. KEGG API enables users to develop software which allows the access to above mentioned databases and associated computational services. KEGG API has been tested with Ruby and Perl languages. However, it should work with any language that can handle SOAP/WSDL, such as Python and Java.

The WSDL file to create a SOAP client driver is available at: http://soap.genome.jp/KEGG.wsdl

Communication through the KEGG API involves use of a number of codes and identifiers, of which the following were of special interest for us:

'org' is a three-letter (or four-letter) organism code. The codes list can be found at:http://www.genome.jp/kegg/catalog/org_list.html

- 'db' is a database name used in GenomeNet service.
- 'entry_id' is a unique identifier whose format is a combination of the database name and the identifier of an entry joined by a colon sign as 'database:entry' (for instance, 'embl:J00231' means an EMBL entry 'J00231'). 'entry_id' includes 'genes_id', 'enzyme_id', 'compound_id', 'drug_id', 'glycan_id', 'reaction_id', 'pathway_id' and 'motif_id'.
- 'genes_id' is a gene identifier used in KEGG/GENES which consists of 'keggorg' and a gene name (for instance, 'eco:b0001' means an E. coli gene 'b0001').
- 'pathway_id' is a pathway identifier consisting of 'path' and a pathway number used in KEGG/PATHWAY. Pathway numbers prefixed by 'map' specify the reference pathway and

pathways prefixed by the 'keggorg' specify pathways specific to the organism (for instance, 'path:map00020' means a reference pathway for the cytrate cycle and 'path:eco00020' means the same pathway with E. coli genes marked).

KEGG API contains a large number of methods. Some of them, of special interest for us, are:

- Meta information: *list_pathways*
- PATHWAY
    - Objects in the pathway: *get_genes_by_pathway*
    - Pathways by objects: *get_pathways_by_genes*
    - Relation among pathways: *get_linked_pathways*

1.2.2 NCBI databases

NCBI (National Center for Biotechnology Information, USA) is a comprehensive resource for the biotechnology and molecular biology information. NCBI maintains public databases, conducts research in computational biology, develops software tools for analysis of genome data, and releases the biomedical information enabling a better understanding of molecular processes affecting human health and disease.

Various databases and tools maintained by NCBI include Literature Databases, Entrez Databases, Nucleotide Databases, Genome-Specific Resources, Tools for Data Mining, Tools for Sequence Analysis, Tools for 3-D Structure Display and Similarity Searching, Maps (various genetic and physical maps) and many more. Entrez is a retrieval system designed to search several linked databases. Various Nucleotide databases include GenBank, EST (Expressed Sequence Tags) Database, GSS (Genome Survey Sequences) Database, HomoloGene (gene homology tool), HTG (High-Throughput Genome Sequences) database, dbSNP (SNP database), RefSeq, STS (Sequence Tagged Sites) database, UniSTS and UniGene.

1.2.2.1 dbSNP

dbSNP is a comprehensive, and freely accessible, repository of information about both single base nucleotide variations and short insertion/deletion polymorphisms. The dbSNP database does not impose any requirement or assumption about the minimal allele frequency in its definition of a SNP. The current build of dbSNP contains over 10 million SNPs [The International HapMap Consortium, 2007], including a large number located within coding regions of genes.

1.2.2.2 Communication with dbSNP

The EFetch utility (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/efetch_help.html) from the NCBI toolkit provides a convenient method of access to SNP records. Another way of communication with dbSNP is using get() function of Perl LWP (Library for WWW in Perl, Burke 2002). The get() function takes the URL as argument, and returns the contents of URL in HTML format. We need to parse this data in HTML format to extract the information we are interested in.

1.2.3 Ensembl Genome Browser

Ensembl is a comprehensive genome information system featuring an integrated set of genome annotation, databases and other information for chordate and selected model organism and disease vector genomes [Flicek et al. 2008]. As of October 2007 Ensembl fully supports 35 species, with preliminary support for six additional species. It provides visualization for genome annotations, alignments, variation and functional genomics data and supporting additional data integration through DAS [Dowel et al. 2001] protocol.

1.3 Distributed Annotation System (DAS)

In order to integrate the information from different databases we have designed our system with an architecture similar to DAS (Distributed Annotation System) [Dowel et al. 2001], which allows the integration of sequence annotations on client side from various annotation

servers. Sequences are annotated by third party investigators at server sites, and these annotations are integrated as needed by the client-side software.

Communication between the client and server is defined by DAS XML specification. Any client or server has to adhere to this specification in order to participate in the system. The schematic of DAS is shown in Figure 1.8. In this figure, one server is the designated reference server (Washington University Genome Sequencing Center) and one or more are the annotation servers (Ensembl, Whitehead, and the Sean Eddy Laboratory) providing annotations relative to the reference sequence. The client, at Cold Spring Harbor Laboratory in this example, fetches data from multiple servers and automatically generates an integrated view.

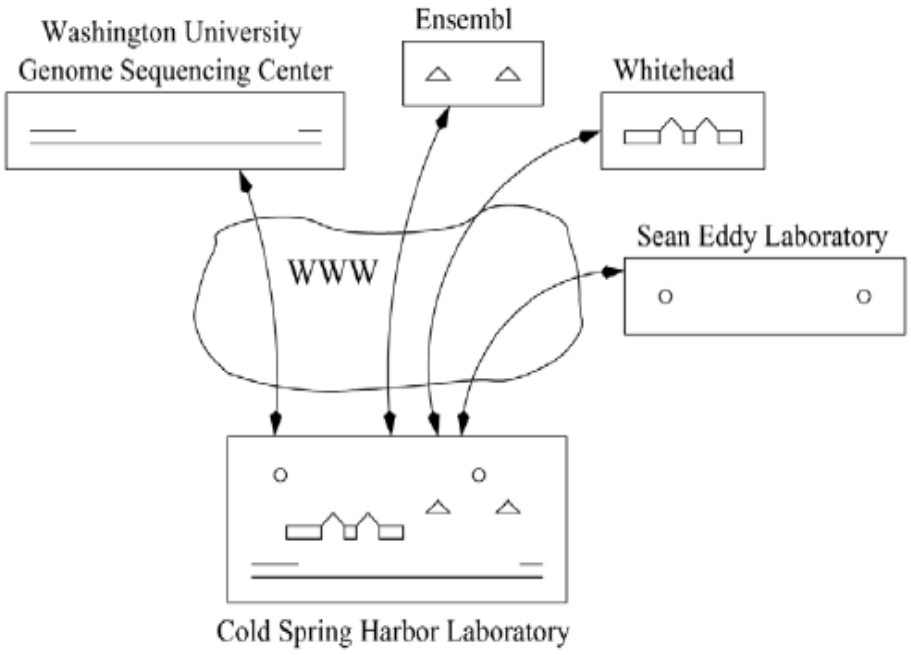

Figure 1.8 Basic Distributed Annotation System architecture. (Source: Dowel et al. 2001)

CHAPTER 2

METHODS

2.1 Architecture of GeneNAB

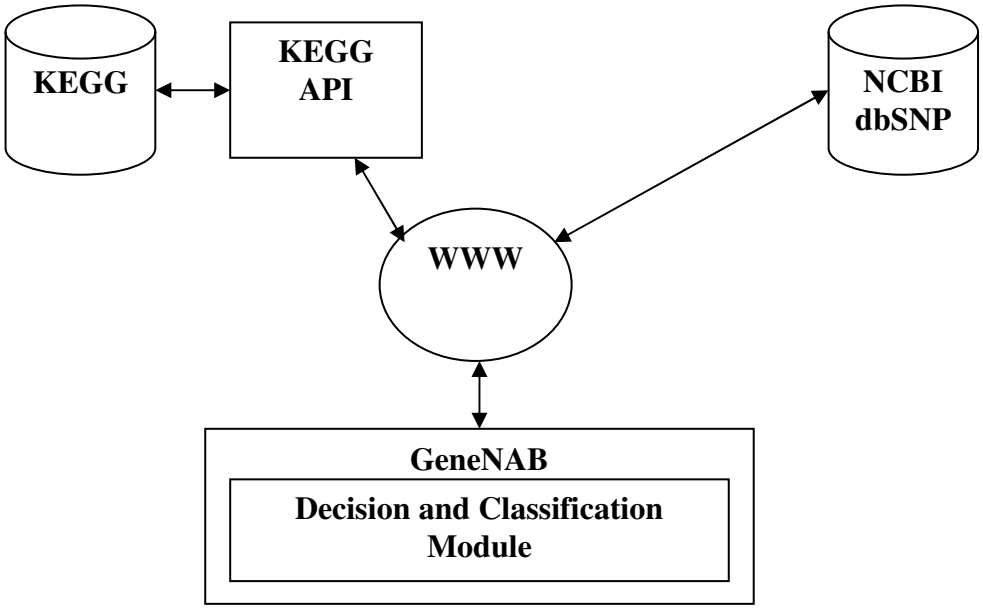

Figure 2.1 GeneNAB architecture

The architecture of our system, GeneNAB, is shown in Figure 2.1. GeneNAB uses different methods for the communication with different databases. It uses the Internet platform to communicate with public online databases, located worldwide. The GeneNAB system is implemented in Perl (ActivePerl 5.10.0) since resources like KEGG and Ensembl provide Perl API as an interface. However, in the Decision and Classification Module, methods for sorting the SNPs according to their scores are programmed in C++.

2.2 Steps in the Execution of GeneNAB

We start with a gene or pathway of interest as initial "bait". Generally, this gene or pathway is selected based on its previously determined association with the condition for which we would like to identify relevant SNPs. If we start with a gene, we find the list of pathways that gene participates in, and perform the further analysis for each of them separately. For all pathways we find the list of all genes participating in them. For this purpose we first use the RPC calls *get_pathways_by_genes* from KEGG API, and then *get_genes_by_pathway*, to locate all other genes participating in the same pathway as our gene of interest. Our postulate was that because complex traits, and diseases in particular, are influenced by sets of SNPs scattered throughout the genome (and thus subject to genome–wide association studies), one link between them is likely in that they may affect genes along the same molecular pathway.

Once we identify the genes associated with our bait, we look for all SNPs lying within the exons of these genes. These SNPs are then sorted by the Decision and Classification Module of GeneNAB, to identify SNPs which are biologically and medically more important. The flowchart of the GeneNAB system is given in Figure 2.2.

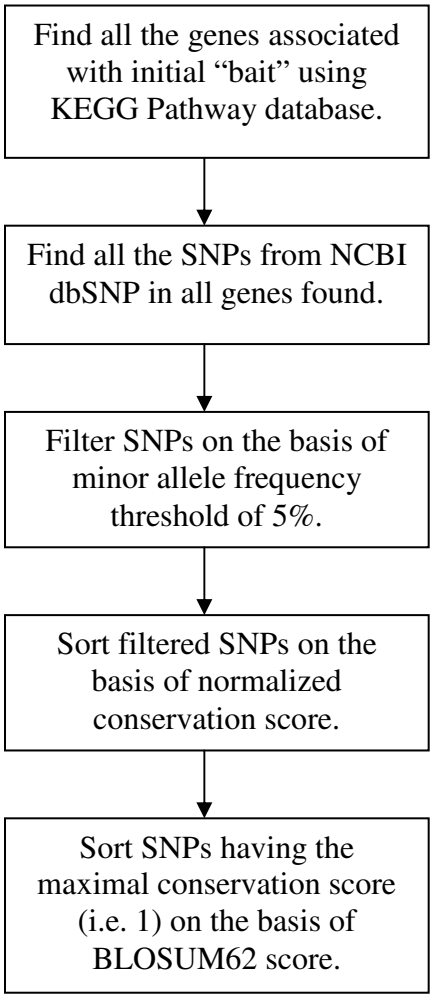

Figure 2.2 Flowchart of GeneNAB

2.3 Communication Methods used by GeneNAB

2.3.1 Communication with KEGG

GeneNAB uses KEGG API to communicate with the KEGG database. We use the RPC calls *get_pathways_by_genes*, to retrieve all molecular pathways our gene of interest participates in, and *get_genes_by_pathway*, to locate all other genes participating in the same pathway as our gene of interest. The RPC call *get_pathways_by_genes* accepts a gene or list of gene IDs as an argument. The Gene IDs are in the format <KEGG organism code> : <NCBI Gene ID> (for instance, hsa:1147, where hsa is the organism code for human). This list has to be converted into SOAP array before it can be communicated.

The RPC call *get_genes_by_pathway* accepts a pathway ID as an argument. Pathway ID is in the format <KEGG organism code> : KEGG pathway id (for instance, hsa:04620).

2.3.2 Communication with NCBI dbSNP

Even if NCBI provides the EFetch utility to access its databases, including dbSNP, it scales down the record contents and omits some fields which we are interested in. Hence, we have implemented the connection to dbSNP using the Perl LWP [Library for WWW in Perl, Burke 2002], using the gene identifiers (GeneID, rs#) we have obtained from KEGG as parameters to the URL string in the *get* function. In response, the dbSNP server returns an HTML document which we need to parse. Since this document is intended for the presentation to a human viewer, rather than a program, our parsing is somewhat heuristic, but we expect that NCBI will soon provide more automation–friendly interfaces to this resource. The major subsections of the returned document are: Submission, Fasta, Resource, GeneView, Map, Diversity, and Validation, out of which we are at this time interested in Diversity only, and in particular in the allele frequency data.

2.4 SNAPPER

SNAPPER is a program which assigns conservation scores to SNPs [Kulkarni et. al 2008]. Conservation score is assigned on the basis of the sum of phylogenetic distances within the alignment of homologous genes from several related species in the immediate neighborhood of the SNP.

SNAPPER assigns a probabilistic score to each SNP in coding regions, and this score represents the likelihood of its importance in affecting the phenotype. Its assumption is that evolutionarily conserved bases are important for proper protein function, so SNPs at these positions are more likely to be important than those in non-conserved regions. This postulate has been validated by correlating the SNAPPER results with HGMD (Human Gene Mutation Database), [Cooper et al. 1998] a database of SNPs which have been implicated in the development of diseases. The overall SNP score has also been based on several other factors known or suspected to be markers for medically relevant alleles in other published studies, such as the length of gene, type of residue change and favorability of substitution. The SNAPPER engine is based on an SVM (Support Vector Machine) classifier to predict probable medically relevant alleles, incorporating all the abovementioned factors.

Two major components of SNAPPER are:

1. Alignment of HGMD SNPs with rest of the SNPs (i.e. SNPs within the coding regions of genes which are not in HGMD).
2. Computation of a variety of metrics for each nucleotide.

The datasets resulting from these two considerations were merged, and relative importance of each metric for differentiating SNPs with low or high impact was measured using the SVM. These metrics were applied to every coding nucleotide by extrapolation. Alignments have been computed for every orthologous gene using the human genome along with eight well sequenced vertebrate genomes. The procedure for evaluating the conservation was implemented as follows:

1. For every human sequence from RefSeq (a nucleotide sequence depository in NCBI GenBank [Benson et al. 2008]), an ortholog was found in other species, using Megablast [Zhang et al. 2000].
2. Nucleotide conservation was represented by scores calculated for each human base position as a function of its conservation in other species.
3. For map back the SNPs from HGMD and dbSNP to their corresponding locations within the genes, each SNP and 50 bases upstream and downstream were BLASTed [Altschul et al. 1990] against the entire genome, and only sequences with 100% identity to the reference were retrieved. The SNPs were then positioned using the BLAST coordinates.

The SNP distribution for dbSNP SNPs was found to be most similar to that of randomly selected base positions, which indicates that most of the SNPs in dbSNP are indeed neutral. Majority of HGMD SNPs were found to lie in the most conserved regions. When only non-synonymous SNPs in dbSNP were considered the normalized nucleotide conservation score was higher, which confirmed the hypothesis that medically relevant SNPs are disproportionately distributed in most conserved regions within genes.

The pre-made list of SNP conservation scores generated by SNAPPER can be substituted by extracting the sequences of different species from the Ensembl genome browser and automatically comparing them to find the conserved regions.

Ensembl provides a Perl API to extract data from public Ensembl databases. Perl API is one of the many ways of accessing Ensembl data. It provides a level of abstraction over the Ensembl core databases. Perl API is used by the Ensembl web interface, pipeline, and gene-build systems. External users can use Perl API to automate the extraction of data, to customize Ensembl to fulfill a particular purpose, or to store additional data in Ensembl. Methods related to the communication with Ensembl are implemented in Bio::EnsEMBL module of BioPerl.

2.5 Decision and Classification Module

Once we identify the list of relevant SNPs participating in the same pathway as our starting gene, we need to rank them according to their potential relevance. First, we retain for further analysis only these whose alleles appear sufficiently often in the population. We have set a cutoff at a typically used minimum minor allele frequency (MAF) of 0.05 (5% or more within the considered population). Traditionally, this threshold is chosen based on the power considerations. For example, in a typical genome-wide association scan, a sample of 1500 to 2000 patients is sufficient to detect a clinically reasonable effect size (odds ratios of 1.5 to 2.0) associated with alleles that have a MAF greater than 5% [Dupont and Plummer, 1997]. It is possible to accrue this number of patients in a reasonable period for most traits. If the MAF is dropped to 1%, the minimum sample size increases to over 5000, which is frequently prohibitive even for multi–center studies. Furthermore, the MAF for SNPs in the Illumina [www.illumina.com] and Affymetrix [www.affymetrix.com] GWAS panels is 5%. Therefore, we used a MAF of 5% as a cutoff in GeneNAB to match the current standard for GWAS.

Consequently, we parse the Diversity record obtained from dbSNP looking for the population percentage exhibiting the minor allele. In some records the frequency data were not available, so we had to unconditionally accept these SNPs. In others there was a comprehensive record for multiple human populations, and in these cases we checked whether at least one of the represented groups features the minor allele frequency of 5% or more. All parsing was performed using basic Perl functions and pattern matching.

The last step of our pipeline involves the ranking of the identified SNPs and the selection of these likely to have a significant contributing effect on the phenotype of interest. Previous work [Kulkarni et al. 2008] has shown that the most medically relevant SNPs tend to lie in phylogenetically conserved regions, so we use the evolutionary conservation of the SNP's environment as a primary criterion for its potential significance. For this purpose we normalize the conservation score for up to 5 bases upstream and downstream of the SNP locus, as described in Kulkarni et al. 2008. At present, we are performing this ranking using locally stored

databases on our machines, but in the future we shall extend this work to automatically connect to genome browsers such as Ensembl [Flicek et al. 2008] and UCSC [Karolchik et al. 2008].

Phylogenetic conservation per se is a good criterion for selection, but it suffers from several drawbacks. The biggest problem is in that the SNP position itself may not be conserved well (indeed, it is a polymorphic site, so it may appear quite poorly conserved), and it may represent a silent change (being in third codon position of a four–fold degenerate site, for instance). We thus look at the amino acids coded for by the major and minor alleles, whether they are different, and the BLOSUM62 [Henikoff and Henikoff 1992] score for the substitution. BLOSUM62, shown in Figure 2.2, is an amino acid substitution matrix, and it specifies substitution scores for each pair of amino acids - the higher the score, the more similar amino acids are. Our assumption was that high scoring substitutions are unlikely to significantly affect the phenotype, so our top candidates were these with low BLOSUM62 values. We deliver the ranked list of most significant candidate SNPs, which should then be checked by laboratory methods for their influence on the phenotype of interest, as the output of GeneNAB.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2.3 BLOSUM62 substitution matrix

CHAPTER 3

RESULTS

3.1 Initial Bait

Given the importance of inflammation as an underlying process in complex human disease (including cancer, heart disease and the development of organ dysfunction after traumatic injury [Schwartz and Cook 2005; Engels 2008; Fairweather and Frisancho-Kiss 2008; Tsujimoto et al. 2008]), we have selected Toll–like receptor 4 (TLR4) signaling as the initial subject for the verification of our system. A receptor is a protein molecule in plasma membrane or cytoplasm, which binds to another molecule called ligand and is responsible for biochemical signaling reactions. Toll-Like receptors lie in the cell membrane, and they play an important role in initiating a signal for the immune system.

TLR4 is the cellular receptor for lipopolysaccharide (LPS), a key component of the cell wall of Gram negative bacteria. It has recently been discovered that in addition to bacterial sensing, this protein also binds to endogenous ligands that are indicative of cellular damage or stress [Bianchi 2007]. It has been demonstrated convincingly that stimulation of Toll receptor signaling results in activation nuclear factor kappa B and a subsequent innate immune inflammatory response [Tsujimoto et al. 2008]. Since the inflammation as a response varies substantially between individuals, this pathway was a good sample case for the study of the effectiveness of GeneNAB.

3.2 Pathway and Genes Identified

Our initial query to KEGG, for TLR4, returned a single pathway consisting of 102 genes, theToll-Like Receptor Signaling Pathway, as expected. Its map has been shown as an example in Figure 1.6, Chapter 1. We have used these genes for the identification of SNPs in their coding regions.

3.3 SNPs Identified and Their Scores

The subsequent communication with dbSNP resulted in a total of 25415 SNPs in the coding regions of the identified genes, out of which only 5707 exhibited minor allele frequency of more than 5% in any of the tested human populations (or for which the population data was not available). We have intersected this list with a catalog of 48362 SNPs which have been scored for phylogenetic conservation by SNAPPER. In this intersection we found 105 SNPs, out of which 39 had a maximum normalized conservation score of 1.0. We have calculated the BLOSUM62 scores for the substitutions of amino acids coded for by their major and minor alleles, and we show the top 13 SNPs identified in this analysis in Table 3.1.

Table 3.1 Top 13 SNPs in the Toll–like receptor pathway identified by the GeneNAB program. All listed SNPs scored 1.0 (maximum) for the phylogenetic conservation of their environments. The IDs of SNPs which were already clinically confirmed to be highly relevant are listed in bold.

SNP	**Gene**	**Amino Acid**	**No. of amino acid**	**Substituted amino acid**	**BLOSUM62**
rs2232613	LBP	Leu	333	Pro	-3
rs2060263	BCL2-L12	Val	47	Gly	-3
rs8177374	TIRAP	Leu	180	Ser	-2
rs2066776	MAPK12	Met	244	Thr	-1
rs10127175	IRAK1	Ser	203	Cys	-1
rs2232607	LBP	Gly	283	Asp	-1
rs11465829	IRAK1	Ile	113	Thr	-1
rs17875834	IFNAR1	Met	359	Thr	-1
rs2069830	IL-6	Ser	32	Pro	-1
rs11465830	IRAK1	His	104	Arg	0
rs5744212	LBP	Phe	339	Leu	0
rs2232619	LBP	Thr	445	Ala	0
rs17177493	CD40	Gln	78	His	0

The top scoring SNPs in our list turned out to be located within the coding regions of genes known to be critical for proper function of the innate immune response. For example, the LPS binding protein (LBP), whose coding gene contains our top scoring SNP, as well as three additional top 13 scoring SNPs, is the first protein to interact with LPS. LBP binds LPS in the blood stream and shuttles it to CD14, a cell surface protein which binds LPS and forms a signaling complex with TLR4 and MD2. Once the complex is assembled, TLR4 initiates a trans–membrane signal which triggers a phosphorylation cascade that ultimately results in ubiquitination and degradation of IkB, which allows migration of NFkB into the nucleus. Once

NFkB enters the nucleus of the cell it activates transcription of the myriad number of genes responsible for inflammation and the innate immune response. The list of the top ten scoring SNPs also contained several loci that are critical to the above–mentioned phosphorylation cascade (IRAK1, MAPK12, TIRAP), as well as a key cytokine (IL-6) and a receptor for interferon alpha. Selection of these high–scoring candidate SNPs from the 25415 SNPs within the Toll–signaling pathway would be a time consuming and frustrating task, even for an expert in Toll signaling and innate immunity.

CHAPTER 4

DISCUSSION AND FUTHER WORK

4.1 Discussion

The GeneNAB program should be helpful for the design of candidate–gene based studies, as well as the interpretation of results from a GWAS. While the utility of the program for the candidate gene approach is obvious, the role of GeneNAB in the analysis of GWAS data may not be as readily apparent. The current analysis of GWAS data involves a series of unadjusted chi–square calculations and while this method has been shown to be efficient in a number of GWAS, it does not allow consideration of the vast amount of biological knowledge regarding the genes or the disease of interest. It is our contention that GeneNAB will not only allow increased efficiency in the design of candidate gene studies, but will also aid in the analysis of genome wide association data. In addition, since our method is based on a different approach than that of currently used techniques, it can be used in a complementary way and improve the sensitivity of the computational analysis. While we feel that the limited validation we have performed so far well demonstrates the effectiveness of our method, we still need to perform the experimental check of all our top candidates.

From 25415 SNPs we couldn't fetch dbSNP records of about 20 SNPs due to connection errors. This number is small enough so that there is very low probability of missing any significant SNP. As this process is automated, it highly depends on reliable internet connection. As it continuously keeps querying dbSNP server, it might have overloaded the server with queries. The dbSNP server is not designed to accept such large number of requests within a short time interval. It provides facility of batch query for large number of queries. However it does not return the results immediately. We will address this problem in future and expect to find solution.

4.2 Future Plan

Computationally, there is still much work that can be done to extend GeneNAB. We plan to narrow the list of SNPs further from currently considered coding regions of genes, and attach more weight to these located within known active domains of their proteins. On the other hand, so far we have ignored the SNPs lying in the regulatory regions of the relevant genes, although these can substantially affect their behavior. Consequently, these SNPs should be included in the analysis, too.

Another possible extension concerns looking beyond just the pathways containing the candidate gene. By mining the Gene Ontology [The Gene Ontology Consortium, 2007] one can discover useful associations which may not be apparent from pathway information only, and we believe that these can further improve the quality of our ranking. Overall, our system is highly extensible, and even as it is producing good results in its current state we expect that in the future it can grow into one of the most powerful computational tools to aid in GWAS.

REFERENCES

1. Lewin, B. Genes VIII.
2. Bianchi, M. (2007). DAMPs, PAMPs and alarmins: all we need to know about danger. J. Leukoc. Biol. 81, 1–5.
3. Brookes, A. (1999). The essence of SNPs. Gene 234, 177–186.
4. Burke, S. M. (2002). Perl & LWP. O'Reilly.
5. Dowell, R., R. Jokerst, A. Day, S. R. Eddy, and L. Stein (2001). The distributed annotation system. BMC Bioinformatics 2, 7.
6. Dupont, W. and W. Plummer (1997). PS power and sample size program available for free on the internet. Controlled Clinical Trials 18, 274.
7. Engels, E. (2008). Inflammation in the development of lung cancer: epidemiological evidence. Expert Rev. Anticancer Ther. 8, 605–615.
8. Fairweather, D. and S. Frisancho-Kiss (2008). Mast cells and inflammatory heart disease: potential drug targets. Cardiovasc. Hematol. Disord. Drug. Targets 8, 80–90.
9. Flicek, P., B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle (2008). Ensembl 2008. Nucl. Acids Res. 36, D707–D714.

10. Henikoff, S. and J. G. Henikoff (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.
11. Hirschhorn, J., K. Lohmueller, E. Byrne, and K. Hirschhorn (2002). A comprehensive review of genetic association studies. Genet. Med. 4, 45–61.
12. Hirschhorn, J., K. Lohmueller, C. Pearce, M. Pike, and E. Lander (2002). Meta–analysis of genetic association studies supports a role for common variants in common disease risk. Am. J. Hum. Genet. 71, 360.
13. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409, 860–921.
14. Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi (2008). KEGG for linking genomes to life and the environment. Nucl. Acids Res. 36, D480–D484.
15. Karolchik, D., 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 (2008). The UCSC Genome Browser Database: 2008 update. Nucl. Acids Res. 36, D773–D779.
16. Kruglyak, L. and D. Nickerson (2001). Variation is the spice of life. Nat. Genet. 27, 234–236.
17. Kulkarni, V., M. Errami, R. Barber, and H. R. Garner (2008). Coding snps that result in disease are disproportionately distributed in the most conserved base positions. BMC Bioinformatics, In press.
18. Lander, E. (1996). The new genomics: Global views of biology. Science 274, 536–539.
19. Lander, E. and L. Kruglyak (1995). Genetic dissection of complex traits — Guidelines for interpreting and reporting linkage results. Nat. Genet. 11, 241–247.

20. Schwartz, D. and D. Cook (2005). Polymorphisms of the Toll–like receptors and human disease. Clin. Infect. Dis. 41(7), S403–407.
21. The International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437, 1299–1320.
22. The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861.
23. Tsujimoto, H., S. Ono, P. Efron, P. Scumpia, L. Moldawer, and H. Mochizuki (2008). Role of Toll–like receptors in the development of sepsis. Shock 29, 315–321.
24. Cooper, D. N., Ball, E. V., and Krawczak, M. The human gene mutation database. Nucleic Acids Res., Jan 1998; 26: 285 - 287.
25. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. J. GenBank. Nucleic Acids Research Advance Access published on December 11, 2007. Nucl. Acids Res. 2008 36: D25-D30; doi:10.1093/nar/gkm929.
26. Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.
27. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. Basic local alignment search tool. J Mol Biol 215 (3): 403–410. doi:10.1006/jmbi.1990.9999.
28. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799-816 (14 June 2007).
29. The Gene Ontology Consortium. The Gene Ontology project in 2008. Nucleic Acids Research Advance Access published on November 4, 2007. Nucl. Acids Res. 2008 36: D440-D444; doi:10.1093/nar/gkm883.

BIOGRAPHICAL INFORMATION

Abhijit R. Tendulkar received his Master of Science degree in Computer Science and Engineering from the University of Texas at Arlington in August 2008. He received the Bachelor of Engineering degree in Computer Engineering from the University of Mumbai, India, in 2003. Before starting his graduate studies in August 2006 at the University of Texas at Arlington, he worked as Research Assistant at the Indian Institute of Technology, Bombay. His primary research area is Bioinformatics.