2006

Integration of biologiccal databases

Azadeh Ahadian
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Ahadian, Azadeh, "Integration of biologiccal databases" (2006). *Master's Theses*. 2865.
DOI: <https://doi.org/10.31979/etd.ky5k-uxq2>
https://scholarworks.sjsu.edu/etd_theses/2865

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

INTEGRATION OF BIOLOGICAL DATABASES

A Thesis

Presented to

The Faculty of the Department of General Engineering

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Azadeh Ahadian

May 2006

UMI Number: 1436890

Copyright 2006 by
Ahadian, Azadeh

All rights reserved.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 1436890

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© 2006

Azadeh Ahadian

ALL RIGHTS RESEREVED

APPROVED FOR THE COLLEGE OF ENGINEERING

Dr. Michael B. Jennings

Dr. Mostafa Ronaghi, Stanford Genome Center

Dr. Frederick Thomas Sharp, IBM

Dr. Leonard P. Wesley

APPROVED FOR THE UNIVERSITY

05/19/06

ABSTRACT

INTEGRATION OF BIOLOGICAL DATABASES

By Azadeh Ahadian

With the advent of high-throughput technologies in biology that enables massive data generation, the main challenge is information management. Efficient data management requires a flexible and scalable architecture upon which data integration and interpretation is made possible. To further this effort, this research demonstrates the feasibility of integrating biological literature data with genetic variation databases. Extracting information from literature and statistical analysis of this data and integration of data with other public databases is the main purpose of this research.

A literature search program, GeneJoo, was designed in collaboration with HP laboratories. Gene/disease relation produced from this literature search program can be integrated with SNP data retrieved from public databases. The result is a graphical summary of all genes and underlying mechanism that causes a disease, helping researchers and scientists quickly find related information.

Acknowledgements

I sincerely thank Dr. Michael B. Jennings for his support and guidance during my studies at the San Jose State University and providing excellent management for the engineering department.

I owe a debt of gratitude to Dr. Mostafa Ronaghi, my supervisor, from whom I have learned much about the science of Biology. I wish to acknowledge his support and guidance throughout my studies at Stanford Genome Technology Center.

I would like to express my sincere gratitude to Dr. Thomas F. Sharp, my advisor, whose broad knowledge in computer science and English literature has been a great value to me.

I am very grateful to Dr. Leonard P. Wesley, my instructor, for guiding me during my last two semesters and for providing well informed course material that were a great help in improving my work.

I am truly grateful to Dr. Javad Dorosti who gave me the vision to see the future needs and introduced me to the field of Bioinformatics and for helping me in designing a Bioinformatics curriculum for the Engineering department so I could start as the first student in this major.

I would like to express my deepest gratitude to Dr. Ali M. Zargar for whom I have the utmost respect. I have not only enjoyed his instruction many semesters but also benefited from his advice and guidance during three years of my study at San Jose state.

I am greatly thankful to my dear friend Dr. Baback Gharizadeh for his kindness and patience in helping me get a better understanding of biomedical science and for his encouragement and guidance to help me relax and complete my writing.

I would like to extend special thanks to my dear friend Kouros T. Behzad for his support during the last three months of the program and for his constructive comments while proofreading my writings.

As always it is impossible to mention everyone who has supported me through this endeavor. I would like to express my sincerest gratitude to all those who gave me the possibility to complete this thesis, to all my friends.

Finally I would like to thank my parents and my sister for their love, complete support, and encouragement.

The work for this thesis was carried out at the Stanford Genome Technology Center in collaboration with HP Labs.

Table of Contents

1. Introduction.....	1
1.1. Terminology.....	2
1.2. Genetic Disease.....	5
1.3. SNPs.....	5
1.4. SNPs and Disease Diagnosis	6
1.5. Haplotype.....	7
1.6. Integrome	9
2. Background.....	10
2.1. Literature Text Mining.....	10
2.2. Biological Data Integration.....	16
2.2.1. Information Challenge	16
2.2.2. Problem Accessing Data in Heterogeneous Biological Databases	18
2.2.3. Database Creation	19
2.2.3.1. Flat Files.....	19
2.2.3.2. Relational Databases.....	20
3. Technical Approach.....	22
3.1. GeneJoo.....	24
3.2. SNP - Integrator	27
4. Results.....	31
4.1. Input	31
4.2. GeneJoo.....	32
4.3. SNP - Integrator	34
5. Discussion.....	35
6. Future Work	41
7. Concluding Remarks.....	43
List of Acronyms	44
References.....	45

List of Figures

Figure 1. Single Nucleotide Polymorphism.....	6
Figure 2. Haplotype	8
Figure 3. Integrome Architecture.....	22
Figure 4. GeneJoo Flow Diagram	25
Figure 5. Integrome Dataflow	30
Figure 6. User Interface for Inputting Disease Name	31
Figure 7. Medical Subject Headings for Parkinson Disease.....	32
Figure 8. Gene Relevant to Parkinson	33
Figure 9. SNP - Output of All SNPs Related to MAOB Gene	34

1. Introduction

The essence of biology is to look at the interaction between different organisms that is determined by organism's genome sequence. Proteins and metabolites mediate biological interactions. During the past decades, several technologies have been developed to produce large-scale data in different "silos" including those of genomics, genetics, proteomics, and metabolomics. A tremendous amount of research has been published providing relatively deep understanding of each field. In order to fully understand biological systems, data from each field needs to be integrated.

Database resources have become an important and invaluable tool for managing the ever growing volumes of biological data and making the information available and accessible to promote research. To date, the lack of a standard platform that could be adopted by the industry has prevented the integration of these databases for delivering a complete and comprehensive view of biology and life.

Overall, these databases contain information from all aspects of biology in disparate systems; however, each database contains information and can answer inquiries on a very specific and specialized branch of knowledge. On the other hand, to be able to address problems and questions, researchers are required to look at information across the boundaries of databases to try to correlate information. Take the following three databases for example: the Flybase database knows about the fly, the Institute for Genomic Research (TIGR) has information on orthology relationships, and the USCS database can help with map positions but each does not know about the information stored in the other databases. In support of their hypothesis, researches will need to carry

their queries across these databases to get the information that they need to make the appropriate correlations. [1]

Access to information continues to be one of the biggest challenges faced by clinical and biological researchers as the volume of published material grows. For instance there are currently about 13 million published peer-reviewed articles in the biological sciences in the Medline database. With the exponential growth in the amount and sources of biological data available, there is even a greater need to deliver some means of integrating these databases to provide researchers with accurate and efficient methods of retrieving relevant.

The objectives of this study are to simplify and enhance the usability and access to biological databases by masking the complexities of the underlying structures and applications from the user.

The following section describes the terminology used throughout this thesis. A list of acronyms has been provided at the end of the thesis.

1.1. Terminology

DNA

Deoxyribonucleic acid is the material inside a cell nucleus that is responsible for carrying the genetic information. DNA forms a few very long molecules in each cell called chromosomes. The sequenced data in a chromosome determines the functionality of a protein. [2]

Central Dogma

The central dogma of biology is the flow of genetic information in a cell.

Transcription of DNA to RNA and translation of RNA to protein establish the foundation of life. [2]

Gene Ontology (GO)

The gene ontology is a norm for explaining the biology of gene products in three ways: the function in the cell, their location in the cell and their position as a cellular element, and finally their role in the biological interactions that the gene product is involved. [3, 4]

Genome

The master blue print structure of a cell or the complete set of genetic material of an organism is called genome. [2]

Human Genome Organization (HUGO)

Since 1989 when HUGO was established as an international organization by scientists in the field of human genetics, its main goal has been to sponsor and support activities in human genetics research such as whole human genome sequencing. [5]

LocusLink

Originated at the National Center for Biotechnology Information (NCBI), LocusLink is an online repository, making available curated sequence and detailed information (such as official aliases, nomenclature, sequence accessions, phenotypes, etc.) on genetic loci. [6]

Medline

Medline is database of literature from NCBI. It contains more than 13 million published biomedical journal articles. [7]

MeSH

Medical Subject Headings (MeSH) is a thesaurus of controlled vocabulary from the National Library of Medicine that allows for searching of terms that have been grouped together in a hierarchical structure. [8]

Online Mendelian Inheritance in Man (OMIM)

This database is a list of Homo sapiens genes and consequent related gene disorders. It consists of text information, references, and numerous links to Medline database as well as the NCBI sequence database. Moreover, OMIM has further links to other information sources at NCBI and other relevant databases. [9]

PubMed

PubMed is made available through the Natural Library of Medicine (NLM) and contains over 16 million documents from databases such as Medline and scientific journals, dating back to 1950. [7]

RNA

Ribonucleic acid is found in both nucleus and cytoplasm of a cell. It can also carry genetic information and it has a very important role in protein synthesis. Unlike DNA, RNA is single-stranded. Other than nucleotide cytosine, guanine, adenine, RNA is built from Uracil instead of Thymine. [2]

1.2. Genetic Disease

A genetic disease or disorder is due to mutations, insertions, or deletions in an individual's genome. These mutations or changes in the genome can vary from minute to major differences such as a small mutation in one particular nucleotide in the DNA of one single gene to larger chromosomal aberrations, which could be the addition or deletion of one whole chromosome. Many diseases or genetic disorders are complex meaning that they are caused by irregularities or mutations in multiple genes. There are also other factors that play a significant role in disease development such as environmental and lifestyle influences. It is difficult to distinguish and measure how all these factors can affect disease development, and moreover, due to the lack of tools to identify these causative factors for disease and genetic disorders, it makes it extremely difficult for researchers and experts to develop tests to map and target these disease-causing factors.

1.3. SNPs

A Single Nucleotide Polymorphism (SNP) is a mutation that occurs in more than 1% in a given population. SNPs are genetic variations across the genome. [10]

The genetic structure consists of four building blocks or bricks referred to as nucleotides: A (adenine), C (cytosine), T (thymine), and G (guanine). SNP variation is when a single nucleotide is replaced by another nucleotide in the genome (as an example when nucleotide G is substituted with one of the nucleotides A, C, or T).

Figure 1. Single Nucleotide Polymorphism

Present estimates are that SNPs with minor allele frequencies of at least 1% are found at the rate of 1 in every 200–300 bp in the genome. [11, 12] Taking into account the entire human genome (3 billion bases), this suggests that there are 10-20 million SNPs eventually available to investigate. [13] In order to correlate particular phenotypes to genomic SNP variation, such as disease, susceptibility to disease or variable drug response, it is essential to have knowledge of how the variation is prearranged and distributed among genes, gene regions, and populations.

1.4. SNPs and Disease Diagnosis

Many genetic variations make up the distinct SNP pattern that is included in a person's genetic material. Due to dense SNP distribution across the genome, SNPs are considered to be ideal markers for large-scale genome-wide association studies to identify genes that are responsible for complex diseases. Although SNPs may sometimes be the cause of a disease, research shows that in most cases this is not true. Because

SNPs are generally found to be in the proximity of a gene that has been linked with a particular disease, by finding the SNP on the human genome map, researchers can segregate the gene. In cases where the SNP actually causes the disease, it can be used to locate the disease-causing gene.

An “association study” is a method that is used to compare results of DNA analysis between a group of people infected with a disease and a group that is not. This method can identify the SNP pattern most closely associated with the disease-causing gene. [14, 15] The first step in the process is to collect blood samples from individuals infected with a disease and those who are not. It is a requirement that the gene that is the cause of that disease already be identified. Next, researchers will analyze the DNA in the blood samples from each group for SNP patterns. Finally, the SNP patterns from both groups will be compared and the differences highlighted. These differences may be attributed to the SNP pattern possibly associated with the disease-causing gene. By assembling SNP profiles and making the information available to physicians, they will be able to predict a person’s vulnerability to a disease from DNA samples. [15]

1.5. Haplotype

Over 10 million estimated SNPs are present in humans. [14] Nucleotide variants in the gene that are in the proximity of each other have a propensity to be inherited together. For instance, an individual having the SNP nucleotide T at one specific region is likely to have similar SNP variation in the vicinity close to the location of nucleotide T. [10] These SNP variations that are associated to one another in a region of a

chromosome are considered as haplotypes. When SNPs are discovered, they are gathered together into haplotypes. [16] A chromosome part might consist of many SNPs but there are only a few tag SNPs that may significantly produce valuable information for determination of genetic variations for disease diagnosis and treatment.

Figure 2. Haplotype

Genotyping is the process of studying a person's tag SNPs. Therefore, instead of testing SNPs, researchers can use Genotyping to identify the groups of haplotypes in that person's DNA.

Investigating the large amount of SNPs in humans is a costly endeavor, making it almost impossible to genotype every individual for these SNP variations. A good approach to circumvent such cost and labor-intensive SNP genotyping is using these tag-SNPs, which are very informative and minimizes the cost significantly. It has been

roughly estimated that there are about 300,000 to 600,000 SNP tags that could be utilized for whole genome SNP screening. This is by far up to 30 times lower than the 10 million SNPs. [14]

This tag SNP information will ultimately provide valuable data for associating specific SNPs to complex genetic diseases. Such data could provide remarkable information for medicine, disease diagnosis, prophylactic measurements, personalized drugs and treatment. The tag SNPs or haplotypes may also foresee a disease in advance, such as an individual who is healthy at present, may develop diabetes in future based on the acquired genetic variation pattern.

1.6. Integrome

The main challenge in biology is managing information. Highly integrated pattern-recognition is needed to turn the processed information (generated from genomics, transcriptomics, proteomics, metabolomics, etc.) into knowledge that can be used by biologists. Integrome is a common name, which is used for a platform enabling integration of omics data.

To address part of this ongoing challenge, a new module is being introduced here to integrate literature data with SNPs in individuals in a comparison fashion to relate genes that are linked to a disease. This module will enable biologists to seamlessly go across data retrieved from literature and find associated genetic variations.

2. Background

The premise of this research is based on text mining and data integration. Some background is provided here on these topics to help readers who are not familiar with the subjects to grasp the basic concepts.

2.1. Literature Text Mining

Text mining is the process of searching through databases and analyzing text to identify patterns and relationships using a combination of statistical analysis and machine learning. Software has been developed to help users to access and examine large repositories of data for information that will aid in solving business, social, or health problems.

In the past several years text mining has become a key productivity tool in bioinformatics and has been utilized by researchers to search through biomedical literature. With its popularity, many text mining methods have been developed and are available for use in a variety of applications.

Automated literature search techniques range from simple pattern matching to full parsers and use what is known as natural language processing (NLP). A typical approach used in biomedicine is to start with the name of specific genes and then designate the text that comes before or after the gene name in the literature as relations. This approach can achieve high precision but because few relations are extracted, they get low recall. To

extract more relations, it can be assumed that phrases such as genes that occur together in a text are related. However, these co-occurrence based approaches are less precise. [17]

Information retrieval, semantics, and information extraction are the three fundamental components of NLP. Information retrieval is the part that searches a database for documents according to criteria defined in the user's query. Genes and their functions are often described with complex terminologies that make the task of information retrieval methodologies hard. That's where the semantics part comes in - semantics use classification systems from Ontology to discover the associations between the different concepts. Once an effective semantic classification is in place, ideas and concepts are extracted from text, the third component of NLP. [18]

Generally, the process of text mining of biomedical literature is performed in five stages: [19]

1. Text gathering
2. Text preprocessing
3. Data analysis
4. Visualization
5. Evaluation

This part of the research focuses on the second step, text preprocessing, and the third step, data analysis. During preprocessing, data access is optimized to increase performance and efficiency during data analysis when information is actually extracted. In the course of data analysis, text and data mining techniques are used to take advantage of NLP capabilities which go beyond keyword search in improving literature search.

Literature search can be categorized according to the following major groups:

- Ontology-based
- Word-based
- Concept-based

A weakness of the current word-based search techniques is that in order to get good results, users need to have a good understanding of the domain to choose good keywords. Also, synonyms not captured in word-based search techniques pose another problem. A way to lower the impact of the second problem is to use normalized concepts (common denominator) in place of synonymous terms. In addition, in biomedicine, Gene Ontology (GO) is used in the process of structuring large amounts of literature with the goal of enhancing methods used in keyword search. Some of the existing systems are reviewed below:

GoPubMed applies Ontology-based literature search to Gene Ontology (GO) and PubMed. The advantages of this approach are: [20]

1. It provides a high-level overview of all the search results instead of going only after the top hits.
2. It does not put incomparable articles in a one-dimensional list to represent what is inherently multi-dimensional. Instead, search results are categorized and presented in a hierarchical order from general to specific concepts facilitating quick navigation.

3. Keywords are normalized using GO, making it possible to retrieve articles using synonyms or relevant keywords as criteria.
4. Because it is a general approach, other ontologies and text bodies can also use this system.

Built upon Swanson's fundamental idea of linking literature, **LitLinker** is a system that can capture connections across biomedical literature. It combines NLP, a knowledge-based methodology, and data mining algorithms for finding new potential relationships between terms used in biomedical literature. Swanson's approach assumed one level of transitivity between correlated terms and used a combination of citation analysis and manual review. LitLinker elaborates on this idea in a different ways. LitLinker recognizes the connection between terms and using these relations performs open-ended discovery. This is in contrast to Arrowsmith and the approaches that followed it. It also raises the probability of new breakthroughs by using an intermediate literature and limits the search space for such open-ended discoveries using knowledge-based heuristics. This is in contrast to text-mining approaches used in the data-mining community. [21]

MedScan is another tool that is used for text mining. It may be used to extract information from files and HTML pages on the local computer, the Internet, and public and internal databases, in addition to MEDLINE abstracts. Following are two examples of data that MedScan can extract:

- Functional links that exist between molecules, proteins, and cell processes

- Variety of regulatory methods that are part of the process and the outcomes of the conditions set forth in regulation and experimentation

MedScan makes data amenable to data mining and computational analysis by removing redundancy and arranging them into an information library or a pathway diagram. It employs a special and sophisticated algorithm that can break down a sentence and understand the function and meaning of all words in that sentence. As a result, it goes beyond relations that are based on a common word and can extract functional associations. [22]

Genescence is also a text and data mining tool that goes beyond keyword-based searching. It uses a rule-based parser in conjunction with co-occurrence analysis that depends on corpus-based statistical techniques to extract text and find relationships, and combines them with regulatory relations that it extracts from micro array data. A term can have different meanings depending on the context and its ontology. This algorithm selects the meaning within the specified context and searches based on that. For better integration of the relationships, three ontologies are utilized: [17]

- Gene Ontology (GO)
- Human Genome (HUGO) Nomenclature
- Unified Medical Language System (UMLS)

With increasing average number of definitions that an abbreviation can have, extracting abbreviations and finding out what they stand for are becoming even a bigger challenge in biomedical literature search. To further add to the challenge, the same word or phrase may be abbreviated differently by different authors. Data mining, information

retrieval, and database integration become more complicated due to these ambiguities. In the past, different approaches have been used in different applications to map acronym definitions in hundreds and even thousands of records; however, only recently, large databases have been created that can accurately resolve definitions across millions of records. The Stanford Biomedical Abbreviation Server, SaRAD, AcroMed, and ARGH are the current four such large-scale databases. [23]

GeneJoo uses the Simple and Robust Abbreviation Dictionary (SaRAD) [24] algorithm, originally developed by Eytan Adar at the HP Laboratories. SaRAD includes three components for building a biomedical symbol dictionary:

1. The procedures to find the definition of an abbreviation,
2. The grouping of those definitions, and
3. The output of information that can lead to more effective search of Medline.

An abbreviation/definition pair has to appear more than once for it to remain in the database. [23]

GeneJoo uses an algorithm similar to that of other systems for extracting definitions. The algorithm is fast and easy to implement. It takes a window of text that precedes an abbreviation wrapped within parenthesis. Then, matching “paths” are extracted and scored based on simple heuristics (e.g., +1 for every word that matches every abbreviation character and -1 for every word that does not). The best match is the path with the highest positive score. The algorithm is suitable to computational demand of a large database such as MEDLINE because scores are added up while the paths are

being built and definitions that are likely to be incorrect or windows that are complicated, i.e., low scores, are removed early in the process, thus improving efficiency.

2.2. Biological Data Integration

Biology is the science of nature and it is as wide and deep as nature itself. It grows as the knowledge of human beings expands and deepens. Biological data contains information about different entities. Just as life and nature are complex and related, so are biological data.

2.2.1. Information Challenge

Understanding and discovering different facts and knowledge is a challenging and continuing project. New technologies have sped the process of biological discoveries to the extent that information storage has become a real challenge. For example, massive amounts of data have been generated soon after complete sequencing of human genome. A big concern is to find the best mechanism to store this information. Part of this involves deciding which format is best to store and demonstrate data. Due to the complexity of biological data, they are frequently organized in a hierarchical fashion, or alternatively, they could be defined recursively. As a result, demonstrating this kind of data in a database is difficult. Another problem that is generally encountered is updating this substantial amount of data as they appear in legacy formats. Once the information is stored, data is hard to change.

The Internet has become the most common way of retrieving data. The majority of biological data in different forms is available online and accessible to the public. The most popular techniques for storing data are web pages, flat files (literature), and relational database systems.

Not all websites of biological data are managed by database management systems. Most of these websites use different services to demonstrate integrated sets of data, which, from user points of view, look like an integrated database management system. These utility services are not able to maintain large volumes of data and cannot handle exponential growth of data in biology. [25]

Databases in bioinformatics are often designed for a specific data type such as DNA or protein sequences. Database schema is designed for that specific entity. Different programs necessitate different data types and data structures as input. Sometimes output of one program is the input of another program. Mostly, the output of one program is calculated based on input of other programs. In the process of integration there should be a flexible data transformation component to facilitate the procedure of merging and converting data. Standardization of this basic entity is going to be a very important factor in facilitating this process. This has had a huge impact on defining concepts in the bioinformatics society. HUGO and other databases have defined unique identifiers to access such data.

The nature of biological data also varies in syntax and semantics. Common terms have been used in different areas for different purposes and may cause problems in the process of integration. For instance, consider that human genome information can be

stored in different databases for different reasons and usages. As this information is being integrated, it is necessary to ensure that non-related data is not included which would ultimately render meaningless results.

2.2.2. Problem Accessing Data in Heterogeneous Biological Databases

Despite the large number of public biological databases accessible via the internet, accessing information can be a difficult, time consuming, and sometimes tedious process. The main reasons follow: [26]

- Biological databases are growing in number and volume and are being updated rapidly.
- There is a plethora of data access techniques with no standard access method.
- There are many data formats and types to store in biological databases and there is a lack of standard data storage types and formats.
- There is a lack of homogeneous databases due to the variety of elements and entities to store.
- Biological data are related to and involve two or more academic disciplines that are usually considered distinct. Coordination between these disciplines is not routine.

While searching for an effective way to integrate data from various sources, the main obstacle was the variety of bioinformatics data access techniques that exist.

Existing data retrieval/access methods have different capabilities and features; hence, there's a need for a standard and unified data access and retrieval method.

2.2.3. Database Creation

The two most common modes of storing data are flat files and Database Management Systems (DBMS). These two storage methods are described here within the context of storing and retrieving biological data.

2.2.3.1. Flat Files

Bimolecular data is generally encoded in unstructured flat files in ASCII text format. Flat files are deemed suitable for storing genetic sequence data because they are comprehensive, easy to use, and light to transfer. Biology publications are examples of research found in this format.

Many well known public databases of biological data are distributed as flat files that are internally managed by a database management system. These databases were very popular in the early days of bioinformatics database creation.

File Transfer Protocol (FTP) and CD-ROM are the two most common methods of exchanging and distributing these data files. Bimolecular data files may be searched through using text search utilities or through the Internet via different web sites and search engines. Sequence Retrieval System (SRS) is one of the most widely used tools for querying these flat files.

There are a number of constraints around using flat files for storing bimolecular data:

- Flat file repositories are not able to store and manage large-scale datasets.
- Because no industry standards have been defined and enforced around the file formats, the structure of each data file could be different from the others.

To complicate this further, file structures are generally either not described properly or only poorly explained.

The lack of consistent file and data structure contributes to difficulties in transformation of files from one form to another. It increases chances of error in parsing due to unclear object definitions, as well as retrieval of duplicate records during a search. It also makes it extremely difficult to cross reference objects thus lessening the probability of identifying relationships between data elements as there is no concept of grouping related data or data objects. Lastly, updating data is made more difficult.

2.2.3.2. Relational Databases

“Modern databases are a collection of data managed by a database management system.” [27]

DBMS is a software used for managing and manipulating (storing, retrieving, updating, and modifying) data. Data elements in a database are associated by certain relations. Relations in a database are what make the database powerful. The relationships between different groups of data have made data accessibility more efficient with little regard to the size of the database.

Like flat files, relational databases also have performance limitations around the structure and size of the database being accessed. Performance testing needs to be performed in order to determine the limits of the database for each utilized platform. Obviously, high capacity servers can handle much larger databases compared with home Personal Computers (PC).

There are many benefits and advantages to using relational databases. DBMS systems provide built-in security features for data access with the great advantage of data sharing over small and large networks. DBMS systems may be used for storing and querying sizable structured information such as sequences and graphs and semi-structured information such as manuscripts and publications. Additionally, they provide features for simplifying query of distributed data sources, and to help in data replication.

This thesis focuses on mining a subset of the biological and clinical data available from the respective databases. Given a common disease, researchers regularly need to browse through these databases to find related genes. Next, the gene-disease combination is correlated with genotyping data from databases such as dbSNP to help scientists identify the variants in the DNA sequence that may contribute to risk of disease and to understand the multi factors involved in disease development. The goal of this thesis is to automate and optimize this process by further enhancing and integrating a search engine (GeneJoo) from Stanford Genome Technology Center/HP Laboratories to create a collection of genes related to common diseases, and then integrate this database with genotyping data.

3. Technical Approach

Integrome discovers the DNA sequence variants that are part of the cause of a disease in two steps. First, important genes that cause that disease must be identified, and, second, for each gene relevant SNPs common patterns must be found.

Figure 3. Integrome Architecture

Architecturally, this prototype system is made up of:

- Many disparate public biological databases
- A local database
- Integrome

The local database is a series of flat files and a relational database that store cross-reference information generated from public biological databases. MySQL was selected as the RDBMS for this design because of the following advantages:

1. It has data security.
2. It has integrity control.
3. It is inexpensive.
4. It is open source.
5. It handles databases over 1 gigabyte in size.
6. It handles access by more than 10 concurrence users.

Examples of information maintained in the local database include gene name and alias name.

Integrome integrates a search engine (GeneJoo) from Stanford Genome Technology Center/HP Laboratories with SNP-Integrator using object-oriented methods written in Java and JSP and renders results in HTML. The main reasons for selecting Java and JSP (Java Server Pages) over CGI (Common Gateway Interface) and Perl include:

1. CGI has become less popular because of low performance and poor technical support.
2. Java is robust and reliable.
3. Java provides better code reusability and readability.
4. Java provides better maintenance and support in the long-term.

The Integrome can run on either Linux or Windows XP operating systems, uses JDBC to connect to the local database and Apache 2 as its web server.

When a client submits a request to search on a disease name through a user interface such as a web page, the first module, GeneJoo, looks at previously prepared flat files and related tables in the local database and generates a list of genes with their relevance factor causing a disease. GeneJoo then passes the list of genes to SNP-Integrator module which retrieves all SNP information for each gene.

This prototype integrates gene/disease information extracted from abstracts of published literature stored in Medline (Literature database in NCBI) with SNPs data in dbSNP.

3.1. GeneJoo

GeneJoo identifies genes that are related to a disease and summarizes its search results using a statistical method developed at the HP laboratory. [28] This method manages symbol aliasing thoroughly, computes the statistical relevance of a set of genes to a disease, and offers a new way to distinguish gene symbols from abbreviations that are not related. [28]

The main algorithm is presented in the following flowchart where each process is described in detail below:

Figure 4. GeneJoo Flow Diagram

1. Examine Medline literature and retrieve gene symbols from documents: GeneJoo uses gene symbols, both the official ones as well as the aliases from the following three online gene databases: [28]

- HUGO
- OMIM
- LocusLink

There may be instances where the same gene is associated with different symbols depending on the context or may be several genes are linked to the same symbol. The rationale behind using the databases listed above is that they are all consistent in their approach to using symbols and aliases. They all label one symbol for each gene as the official and all others as alias symbols. This approach to normalization helps solve the problem of isolating and classifying human genes when dealing with large volumes of data.

2. Replace all occurrences of symbol aliases with the official symbol: this step of the process, also called folding, involves the replacement of all symbol aliases with their respective official symbol. The algorithm used for folding is based on the count of official symbols relative to the count of their respective aliases. The level of independent occurrence, or the count of official symbols relative to the alias' count, is

the measure used in the algorithm. For each alias symbol there are four possibilities:

[28]

- There is a one to one relationship between the alias symbol and the official symbol and the official symbol is used more frequently than its alias. In this case all occurrences of the alias will be replaced with the official symbol.
- There is more than one official symbol in the set, all mapping to the same alias; however, only one official symbol has a larger count than its alias. In this case, the alias will be replaced with this official symbol.
- More than one official symbol is associated with the same alias symbol but none of the official symbols has a greater count than the alias. This is an indication that the official symbols are not yet broadly recognized by researchers and therefore the symbols are more suitable for use.
- Lastly, more than one official symbol is associated with the same alias symbol but this time two or more official symbol have a greater count than the alias. In this scenario, without further analysis, it is not possible to isolate the official symbol that was intended and therefore the alias symbols would be more suitable for use.

3. Determine how relevant the symbol is to the disease: the relevance factor is calculated by keeping a count of how many times an official symbol or its alias is repeated in all Medline articles representing a specific gene and comparing that number against the count of times when that specific gene was referenced in articles that cite the disease. [28]

4. Evaluate text to disambiguate and confirm gene symbol: identifying and extracting gene names from text remains the most difficult step of the process after isolating the symbols that are highly relevant to a disease. There are instances where abbreviations for other words in text could be taken for gene symbols thus contributing to less than accurate results. To minimize this problem, text surrounding the symbol is evaluated to confirm that the symbol is indeed representing a specific gene. When the surrounding text fails to provide any cues in this identification then statistical probability is used to make an educated guess with respect to the symbol and the gene it represents. [28]

This part of the project deals with GeneJoo and consists of two main parts:

1. Running the existing literature search program written in Perl and CGI and making it work as a stand-alone, and then running it on a web server. This version was used as a work-bench for testing the java version.
2. Rewriting in Java and adding some accuracy features based on future needs. Most biological software applications are implemented in a robust and reliable programming language such as Java or C++.

3.2. SNP - Integrator

The second part of Integrome is to develop a component for integrating gene-disease results from literature to Gene-SNP raw data. In search of publicly available SNP databases, the following were identified:

1. dbSNP database from NCBI
2. HGVbase (Human Genome Variation Database), a human gene-based polymorphism database
3. SNP consortium (TSC), established in 1999 to discover and map 300,000 common SNPs

dbSNP was selected as the source of data for this research because of its accessibility, accuracy and completeness and good documentation.

There has been a massive progress in processing SNP data in NCBI. There are many organizations working with SNP data to retrieve this information and display it in graphical format.

The criteria for selecting an application to perform SNP search were:

- It must be available as an open source project.
- It must have a method to retrieve all SNPs related to a gene.
- It must retrieve specific SNP information.

SNPHunter and BioConductor SNPper [29] were two software applications that were evaluated in detail. BioConductor SNPper was selected as the most suitable for this purpose because it met all criteria.

BioConductor

Bioconductor is an open source project that allows open participation in its development and design for the analysis and understanding of genomic data. It offers numerous statistical and graphical methods for analysis of genomic data. Bioconductor is

chiefly developed in the R programming language; however, as an open source project they accept contribution in other programming languages. [29]

R provides a wide range of statistical computation analysis and displays results graphically for easy visualization. The R system is very flexible and runs on most major operating systems, including UNIX, Windows and MAC.

BioConductor SNPper

SNPper was designed and developed for use on the web to search for known SNPs in its database, a repository that is compiled by combining data from dbSNP and UCSC Human Genome SNP databases. SNPper contains a subset of five million known SNPs. [30]

An interface to SNPper used in obtaining SNP-related data is RSNPper. [31] This interface consists of a package of functions that offer direct methods for working with an XML-RPC server. XML-RPC (Remote Procedure Call) is a web services standard that allows software running in different environments and on disparate operating systems to make procedure calls over the internet. It uses HTTP as the transport method and XML as the encoding methodology. [32]

Methods available in RSNPper enable the user to retrieve SNPs associated with a gene. There are three methods available in RSNPper function package that may be used to retrieve information about SNPs associated with a given gene: [31]

1. *GeneInfo* retrieves gene information such as gene name, gene id, chromosome number, product (what disease it produced), transcript start and stop location, coding sequence start and stop position, accession number, LocusLink id,

Omim id, UniGene id, SwissPort id, and most important of all, SNP id. This method returns information about the gene in XML format.

2. *GeneSNPs* returns all SNPs associated with a specific gene.
3. *SNPInfo* is another alternative for retrieving SNP data using SNP id as input.

In the SNP-Integrator module, the *getGeneInfo* object retrieves all information about the gene, including the gene ID. This ID is used in *getSNPForGene* object, which returns information in XML format about all SNP related to a gene. The *xmlParser* object then extracts the necessary information and the *visualizeGeneSnp* object visualizes the information.

Figure 5. Integrome Dataflow

4. Results

The Integrome is designed to integrate the information stored in published literature in NCBI Medline database with dbSNP SNP database and to return a list of genes related to a disease with their relevance factor, links to the searched literature, and detailed information on the SNPs identified as related to each gene. Following is a step-by-step of the process that the prototype system goes through and the results and reports generated at each step for the user.

4.1. Input

To test the system, “Parkinson” was input as the disease name to perform its search on:

Search for genes relevant to a disease

A screenshot of a web form titled "Search for genes relevant to a disease". It features a text input field containing the word "Parkinson" and a button labeled "Search".

Parkinson	Search
-----------	--------

Figure 6. User Interface for Inputting Disease Name

After the search completes, all medical subject headings (MeSH) are returned and displayed for the researcher. By clicking on the subject heading, a list of genes for that MeSH for Parkinson is returned:

MeSH	#docs with gene symbols	cached
<u>Antiparkinson Agents</u>	718	yes
<u>Parkinson Disease, Postencephalitic</u>	23	yes
<u>Parkinson Disease, Secondary</u>	600	yes
<u>Parkinson Disease</u>	3525	yes
<u>Parkinsonian Disorders</u>	185	yes
<u>Wolff-Parkinson-White Syndrome</u>	394	yes

Figure 7. Medical Subject Headings for Parkinson Disease

4.2. GeneJoo

The GeneJoo module is designed to find all genes related to a disease from literature stored in XML files. It was expected that GeneJoo search through published literature in Medline to return genes that were referenced in association to Parkinson's disease and for each result, display the relevance factor and links to get more information from the original document in Medline.

The top 10 results returned by GeneJoo included:

official gene symbol	relevance score	mentions with disease	SNPs	total mentions	counts from aliases			symbols not accepted
MAOB	104.39	149	Link	1183				
COMT	102.39	126	Link	882				
PARK2	99.44	28	Link	47	ARJP (24)			
MAPT	72.49	33	Link	122	PFND (7)	FTDP17 (21)	DDPAC (4)	TAU (8)
DYT3	46.20	6	Link	10				
GNPI	45.64	101	Link	2670	GPI (101)			
SNCA	44.85	19	Link	105	NACP (16)	PARK1 (1)		
PARK3	42.22	3	Link	3				
GDNF	41.65	48	Link	751				
SOD1	41.59	116	Link	4110	ALS* (104)			

Figure 8. Gene Relevant to Parkinson

4.3. SNP - Integrator

The SNP-Integrator module integrates information about gene/disease relation to gene/SNP data. For each row in the output of GeneJoo, the SNP link takes the user to the page where a list of SNPs is displayed for the selected gene. The following are the first 10 SNPs that were returned for the gene “MAOB”:

Gene Symbol	Gene ID	dbSNP ID	Chromosome	Position	Alleles	Validated
MAOB	20285	rs3027418	chrX	43372437	A/G	N
MAOB	20285	rs3027419	chrX	43372535	C/G	N
MAOB	20285	rs7890899	chrX	43373992	A/G	Y
MAOB	20285	rs5905911	chrX	43374222	A/T	Y
MAOB	20285	rs5952278	chrX	43375034	A/G	Y
MAOB	20285	rs5952659	chrX	43375035	C/T	Y
MAOB	20285	rs3027420	chrX	43375757	C/T	N
MAOB	20285	rs3027421	chrX	43375979	C/G	Y
MAOB	20285	rs3027422	chrX	43376033	A/G	N
MAOB	20285	rs5952661	chrX	43376072	C/G	N

Figure 9. SNP - Output of All SNPs Related to MAOB Gene

5. Discussion

Integration of biological data plays a key role in providing insight into complex biology. At the Stanford Genome Technology Center this problem is being tackled by using a novel approach in integrating biological data for better understanding of biological systems. The aim of this study is to describe integration of GeneJoo with SNP database, called Integrome, which is part of a larger project on integrating all sort of biological data generated from high-throughput platforms.

Integrome facilitates identification of genes related to a disease with its associated SNPs, which may have direct effect on causing a disease. This system is designed to help researchers get a summary of information from related published literature as well as delivering useful information about the relevance factor and the cause of gene mutation. In this project, GeneJoo was first developed to rapidly scan for genes associated to different diseases. The program was originally developed using Perl language, but to make it compatible with other standard software, it was converted to Java. The new re-written program was then connected to SNP databases to simultaneously retrieve associated data to every single gene.

Several important factors were taken into consideration in the design and selection of GeneJoo:

- Given a query, the identification and summarization method is used to mine only genes that are closely related, specifically, those that occur most frequently in the context.

- Assigns a higher score to a gene whose relationship to the disease has been well established and a lower score to a recently documented gene.
- Statistical techniques have been used to calculate how closely a pair of genes is related to a disease.
- The gene recognition algorithm used when searching for symbols and aliases is based on a gene name dictionary constructed from three online gene databases:
 - HUGO
 - LocusLink
 - OMIM

The main reason for the selecting these specific databases is the fact that they are consistent in assigning a unique symbol to each gene thus avoiding a common issue where one symbol is used to identify many different genes or many symbols referencing the same gene. The unique symbol is considered to be the official name while all other symbols are aliases.

- Medline currently includes more than 13 million publications dating back to 1960. GeneJoo has access to search through all the titles and abstracts in Medline.

There are a variety of algorithms that have been developed to identify genes and proteins in text. These algorithms may be based on natural language, rules, or string matching.

The most successful of these requires complex natural language processing and generates results with an accuracy of up to 98%. The algorithm used by GeneJoo could be replaced by any one of the algorithm reviewed in the background section; however, as is, the GeneJoo algorithm has the advantage of being simple and easy to implement and yields an accuracy of more than 95%, which is only slightly lower than more sophisticated algorithms mentioned above.

When searching through literature, two important criteria to be considered are performance and accuracy of results. To address the concerns around performance in GeneJoo, search engines that delved into exhaustive analysis of literature were avoided. These same search engines could also be prone to errors in results due to words or abbreviations being read in place of gene symbols. GeneJoo met both these criteria. To further improve the accuracy of the results, additional factors may be added to the GeneJoo algorithm. For example, the algorithm could also take into account the dates of the journals being searched through and restricting searches to only the top journals with high impact factors. Currently, GeneJoo searches for association of a gene to a certain disease and detail information whether the searched gene is up or down regulated in that specific disease is not determined. It will be of great value to expand the mining algorithms to search for up and down regulation if the text appears in the same sentence as the name of the gene.

Issues of various complexities were encountered in several areas, each taking hours and some times days to resolve. These included:

- Dealing with unexpected problems with the Apache server under SUSI, such as enabling CGI programs, was difficult and time consuming.
- The latest and correct versions of the source files were not available.
- Processing XML files on a laptop that was not equipped to handle the processing load was trying.
- Some scripts took a long time to run because the volume of input data was very large.
- The Linux system crashed many times after running just one program for two days. This made the testing and debugging the program very difficult. Each script had to be tested many times and would fail if run with a complete set of data. Three scripts had to be tested with half or third of the data set.
- Because scripts would not run with the complete data set, the output was not generated properly. Therefore, to test the rest of the scripts, sample data had to be created.

Open source model was used for this research project to reduce costs and to eliminate dependencies on software vendors. Open source software provides flexibility in adding new components or customizing existing features. A key factor in using open source software is code reuse which has allowed for rapid development while not compromising code quality.

Open source software used in this project was Bioconductor. There are several advantages with this software, most notably are *i*) standard source data (dbSNP), *ii*)

advanced techniques (XML-RPC), and *iii*) designed to work in both UNIX and MS Windows environments.

Although the Bioconductor project has all the advantages of the open source model, it is implemented in the newly born “R” programming language which has a learning curve to become proficient in its use. To save time and to eliminate this learning curve, the same functionality was recreated using the Java language to simulate the same methods as in RSNPper.

Integrome has the potential for usage in many areas. It may be considered for use as a biological literature search tool in stand-alone mode or as a plug-in to large systems. GeneJoo has already been used in the collaboration between the Stanford Genome Center and the Parkinson Institute to collect information about the underlying genetic factors as described in the literature. The collected data was passed through Integrome to find associated SNP data. This was previously performed manually, looking at each affected gene in Parkinson by going to different databases. The researchers in this consortium found Integrome very useful and time saving while much more additional data could be found which was almost impossible to gather through manual reading of Medline. Through Integrome researchers could search for all genes related to the disease and to connect automatically to the SNP-Integrator to get the results all in one step.

To validate the accuracy of the output of the developed tool, results were compared against Parkinson data that had been generated and tested at Stanford and where most effective genes had been identified. The obtained results by the developed automated tool contained all the genes found manually related to Parkinson Disease.

Integrome is a tool for facilitating and acquiring of information in published literature to enable the users to specifically single out related and relevant literature, and moreover, to obtain a summary list of the significant and relevant genes that are causing the disease. Furthermore, this tool will relate the attained biological literature to SNP database, and finally, helps detect the SNPs that are responsible for gene mutation and disease formation. It also gives the user the advantage of not having to know the individual building blocks of the system. It simplifies the interaction by providing a single user interface whereby the user can avoid dealing with multiple interfaces and applications to reach the results of their literature search.

6. Future Work

The current GeneJoo algorithm only looks at the title and abstract of Medline articles. Incorporating full gene names into the search algorithm and searching the whole article would improve the algorithm efficiency in terms of accuracy.

Selecting articles that have been published in the highly cited scientific journals by considering the impact factor would improve the accuracy and usefulness of the results.

The mining capability of GeneJoo could be expanded to search for up and down regulation of genes in certain diseases. Mining from other publication such as published patents could be included in the process.

Another way to improve the accuracy of the information extracted by the system is to enhance the GeneJoo algorithm to take into account the date of the journal publication to limit the searched results to only recent articles. This is similar to the process a researcher would go through during a manual search. This can easily be accomplished as the journal publication date is already part of the existing information in the Medline XML files. A disadvantage of this approach is that it may omit some original and significant papers; however, this can be addressed by providing the user with the option to specify the date range of the articles to include in the search.

Subsequent to finding SNPs for each genes related to a disease, blocks of SNP (tag SNPs) that creates haplotype could be interrogated. Currently haplotype analysis software is being evaluated to identify features matching the architectural requirements of

Integrome. Haploview is one such application. Haploview is a project hosted by SourceForge (an open source community) [33] for analyzing and viewing haplotypes. It provides key functionalities such as haplotype block analysis, haplotype frequency estimation, and Single SNP and haplotype relationship test. [16]

7. Concluding Remarks

The amount of biological data is passing exabytes level in the industry requiring development of automated tools for data mining and integration. Toward this direction and as part of the Integrome project, a literature mining tool has been developed to extract the genes and automatically find associated SNP. The developed tools could enhance the efficiency of researchers in the field by dramatically reducing the time on finding relevant disease information from literature and databases. Furthermore, this tool has demonstrated unprecedented comprehensiveness which could not be achieved by manual data mining. It is concluded that further integration of different data on biological components would increase discovery and efficiency in life science research.

List of Acronyms

ARGH	Acronym Resolving General Heuristics
CGI	Common Gateway Interface
CD-ROM	Compact Disk Read Only Memory
DBMS	Database Management System
DNA	Deoxyribonucleic Acid
FTP	File Transfer Protocol
GO	Gene Ontology
HP	Hewlett Packard, Inc.
HTML	Hyper Text Markup Language
HUGO	Human Genome Organization
JDBC	Java Database Connectivity
JSP	Java Server Pages
MAOB	A gene name: Monoamine Oxidase B
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
OMIM	Online Mendelian Inheritance in Man
PC	Personal Computers
RDBMS	Relational Database Management System
RNA	Ribonucleic Acid
RPC	Remote Procedure Call
SaRAD	Simple and Robust Abbreviation Dictionary
SNP	Single Nucleotide Polymorphism
SRS	Sequence Retrieval System
UCSC	University of California Santa Cruz
UMLS	Unified Medical Language System
XML	Extensible Markup Language

References

1. L.D. Stein, *"Integrating biological databases,"* Nat Rev Genet, **4**(5), pp. 337-45 (2003).
2. J. Meidanis and J. Setubal, *Introduction to Computational Molecular Biology*. 1997, Boston: PWS Publishing Company.
3. M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *"Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,"* Nat Genet, **25**(1), pp. 25-29 (2000).
4. W. contributors. *Gene Ontology*. (2006). [Online]. Available at: http://en.wikipedia.org/w/index.php?title=APA_style&oldid=43835099 (accessed 29 March 2006).
5. HUGO. *The Human Genome Organisation (HUGO)*. (1989). [Online]. Available at: <http://www.hugo-international.org/> (accessed 25 April 2005).
6. K.D. Pruitt and D.R. Maglott, *"RefSeq and LocusLink: NCBI gene-centered resources,"* Nucleic Acids Res, **29**(1), pp. 137-140 (2001).
7. *Medline*. [Online]. Available at: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi> (accessed 21 April 2005).
8. M.H. Coletti and H.L. Bleich, *"Medical subject headings used to search the biomedical literature,"* J Am Med Inform Assoc, **8**(4), pp. 317-23 (2001).
9. M.-N.I.f.G. Medicine, M. Johns Hopkins University (Baltimore, N.C.f.B. Information, and N.L.o. Medicine, *"Online Mendelian Inheritance in Man,"* (2000).
10. B.A. Salisbury, M. Pungliya, J.Y. Choi, R. Jiang, X.J. Sun, and J.C. Stephens, *"SNP and haplotype variation in the human genome,"* Mutat Res, **526**(1-2), pp. 53-61 (2003).
11. L. Kruglyak and D.A. Nickerson, *"Variation is the spice of life,"* Nat Genet, **27**(3), pp. 234-6 (2001).

12. J.C. Stephens, J.A. Schneider, D.A. Tanguay, J. Choi, T. Acharya, S.E. Stanley, R. Jiang, C.J. Messer, A. Chew, J.H. Han, J. Duan, J.L. Carr, M.S. Lee, B. Koshy, A.M. Kumar, G. Zhang, W.R. Newell, A. Windemuth, C. Xu, T.S. Kalbfleisch, S.L. Shaner, K. Arnold, V. Schulz, C.M. Drysdale, K. Nandabalan, R.S. Judson, G. Ruano, and G.F. Vovis, "*Haplotype variation and linkage disequilibrium in 313 human genes*," *Science*, **293**(5529), pp. 489-93 (2001).
13. M. Ronaghi and T. Langae, "*Single Nucleotide polymorphisms: discovery, detection and analysis*," *Personalized Medicine*, **2**(2), pp. 111-125 (2005).
14. T.I.H. Consortium, "*The International HapMap Project*," *Nature*, **426**(6968), pp. 789-96 (2003).
15. G.A. Thorisson, A.V. Smith, L. Krishnan, and L.D. Stein, "*The International HapMap Project Web site*," *Genome Res*, **15**(11), pp. 1592-3 (2005).
16. J.C. Barrett, B. Fry, J. Maller, and M.J. Daly, "*Haploview: analysis and visualization of LD and haplotype maps*," *Bioinformatics*, **21**(2), pp. 263-5 (2005).
17. G. Leory, H. Chen, J. Martinez, S. Eggers, R. Falsey, K. Kislin, Z. Huang, J. Li, J. Xu, D. McDonald, and G. Ng, "*Genescene: Biomedical Text And Data Mining*," *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries.*, pp. 116-118 (2003).
18. M.D. Yandell and W.H. Majoros, "*Genomics and natural language processing*," *Nat Rev Genet*, **3**(8), pp. 601-10 (2002).
19. B. Mathiak and S. Eckstein, "*Five Steps to Text Mining in Biomedical Literature*," *PKDD/ECML* pp. 43-47 (2004).
20. R. Delfs, A. Doms, A. Kozlenkov, and M. Schroeder, "*Ontology-based literature search applied to Gene Ontology and PubMed*," In *Proc. of German Bioinformatics Conference*, Bielefeld, Germany, (2004).
21. W. Pratt and M. Yetisgen-Yildiz, "*Capturing Connections across the Biomedical Literature*," *Proceedings of the 2nd international conference on Knowledge capture* pp. 105-112 (2003).
22. S. Novichkova, S. Egorov, and N. Daraselia, "*MedScan, a natural language processing engine for MEDLINE abstracts*," *Bioinformatics*, **19**(13), pp. 1699-706 (2003).

23. J.D. Wren, J.T. Chang, J. Pustejovsky, E. Adar, H.R. Garner, and R.B. Altman, "*Biomedical term mapping databases*," *Nucleic Acids Res*, **33**(Database issue), pp. D289-93 (2005).
24. E. Adar, "*SaRAD: a Simple and Robust Abbreviation Dictionary*," *Bioinformatics*, **20**(4), pp. 527-33 (2004).
25. D.P. Miranker, "*Metric-space indexes as a basis for scalable biological databases*," *Omics*, **7**(1), pp. 57-60 (2003).
26. M. Lange, A. Freier, U. Scholz, and A. Stephanik. *A computational support for the access to integrated molecular biology data*. [Online]. Available at: <http://www.bioinfo.de/isb/gcb01/poster/lange.html> (accessed 2 August 2005).
27. R. Elmasri and S.B. Navathe, *Fundamentals of Database Systems* 1994: Benjamin/Cummings.
28. L.A. Adamic, D. Wilkinson, B.A. Huberman, and E. Adar, "*A literature based method for identifying gene-disease connections*," *Proc IEEE Comput Soc Bioinform Conf*, **1** pp. 109-17 (2002).
29. R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, and J. Zhang, "*Bioconductor: open software development for computational biology and bioinformatics*," *Genome Biol*, **5**(10), pp. R80 (2004).
30. A. Riva and I.S. Kohane, "*A SNP-centric database for the investigation of the human genome*," *BMC Bioinformatics*, **5** pp. 33 (2004).
31. V. Carey. *RSNPper: utilities for SNP data*. (2005). [Online]. (accessed 10 September 2005).
32. I. UserLand Software. *XML-RPC*. (1998). [Online]. Available at: <http://www.xmlrpc.com/> (accessed 9 December 2005).
33. O.S.T. Group. *SourceForge*. (2001). [Online]. Available at: <http://sourceforge.net/> (accessed 10 July 2005).