Database Challenges in the Integration of Biomedical Data Sets

Rakesh Nagarajan
Washington University School of Medicine
Department of Pathology & Immunology
660 South Euclid Avenue, Campus Box 8118
Saint Louis, MO, 63110, USA.
rakesh@wustl.edu

Mushtaq Ahmed and Aditya Phatak
Persistent Systems Private Limited
Bhageerath, 402, Senapati Bapat Road
Pune 411016, India
{mushtaq, aditya}@persistent.co.in

Abstract

The clinical and basic science research domains present exciting and difficult data integration issues. Solving these problems is crucial as current research efforts in the field of biomedicine heavily depend upon integrated storage, querying, analysis, and visualization of clinicopathology information, genomic annotation, and large scale functional genomic research data sets. Such large scale experimental analyses are essential to decipher the pathophysiological processes occurring in most human diseases so that they may be effectively treated. In this paper, we discuss the challenges of integration of multiple biomedical data sets not only at the university level but also at the national level and present the data warehousing based solution we have employed at Washington University School of Medicine. We also describe the tools we have developed to store, query, analyze, and visualize these data sets together.

1. Introduction

It is becoming increasingly apparent that the majority of human diseases including tumorigenesis are the product of multi-step pathophysiological processes, and that each of these processes involve the complex interplay of a multitude of genes acting at different levels of the genetic program. Indeed, it is clear that genome-wide detection of genetic alterations, transcriptional profiles, and protein compositions is required to comprehensively describe the complex pathophysiology of polygenic diseases.

Fortunately, in the post-human genome sequencing era, many analyses on the genomic scale are possible. The biggest challenge in interpreting the results of these analyses lies in the data integration problem. The experimental methods employed in genomics and proteomics generate high throughput data, which is stored in different formats at multiple sources. In a university, this data is generated at various core labs and has to be shared across investigators. The data management, integration and analysis needs for this kind of heterogeneous data are enormous.

Typically, groups have utilized three major mechanisms to integrate biological databases. These include:

- Indexed data sources: This approach indexes and links a large number of data sources. Here a user begins a query with one data source, and then follows links (*e.g.* hypertext) to related information in other data sources. For example, the Sequence Retrieval System (SRS) is a popular keyword indexing and search system for biological databases [18].
- Federated databases: In this approach, the information resides in the respective source databases. Federated systems maintain a common data model and rely on schema mapping to translate heterogeneous source database schemas into the target schema for integration. For example, the Kleisli Query System provides a high-level query language, simplified SQL (sSQL), which can be used to express cross-database queries [3]. K2, a successor to Kleisli, is a view integration environment developed by the database group at the University of Pennsylvania [4], and IBM's Discovery Link is another popular integration system based on the federated approach [7].
- Data warehousing: This approach assembles data sources into a centralized system with a global data schema and indexing system for integration and navigation. This approach is dominated by relational database management systems (RDBMS).

For our university setup, the major design considerations included fast querying of data from

Proceedings of the 30th VLDB Conference,
Toronto, Canada, 2004

multiple sources, efficient handling of large amounts of data, allowing users to upload and analyze their data, and access to data via a campus-wide intranet for approximately 100 concurrent users. In addition, while the experimental data generated within the university needed to be accessed in almost real time, the annotation data coming from publicly available databases needed to be up to date only within the past few weeks. Therefore, we chose to use the data warehousing model to store these experimental and annotation data sets. In this paper, we discuss the challenges of integration of these diverse data sets not only at the university level but also at the national level and present the data warehousing based solution we have employed at Washington University School of Medicine.

The rest of the paper is organized as follows. Section 2 provides necessary background. It describes the need for data integration in detail and provides details of various bioinformatics data sets and the technologies that generate this data. In Section 3 we present our data warehouse solution and various tools we have developed to store, query, analyze, and visualize the data. In Section 4, we discuss several ongoing standardization efforts required to store and annotate such data sets in a uniform manner, and we discuss our future work and overall conclusions in Sections 5 and 6 respectively.

2. Background

Foremost among the high throughput technologies in post-genomics era is the ability to monitor the messenger RNA (mRNA) expression of all genes in a particular tissue, cell type, or pathological process [6, 14]. There is great potential in this experimental modality, termed microarray analysis, as evidenced by the recent explosion of publications using this technique to monitor genome-wide expression and to correlate expression changes to biological processes or to disease states.

Some of these modalities of molecular analysis have also been combined with clinically relevant parameters such as patient survival or the existence of metastatic disease in the study of tumorigenesis. However, it is becoming increasingly apparent that it is necessary to *simultaneously* analyze the results derived from different functional genomic experiments such as expression profiling and mutation analysis with clinical and pathological data and gene annotation. This will increase the power of the analysis and will provide complementary confirmation such that meaningful insights into the disease process may be made.

However, most end-users (biologists and clinicians) find the task of performing such integrated *in-silico* analyses daunting in the functional genomic era; the main reason for this being the variety of disparate data sources, different software applications, and varied output formats that exist today. Thus, a flexible data integration framework, which will take care of such complexities and will allow the researcher to focus on the results of analyses, is needed. In the remainder of this section, we describe the different data sets that are used in such integrated analyses.

2.1 Gene Annotation

In 2001, a draft of the entire human genome sequence or the human DNA genetic sequence was deciphered as part of the Human Genome Project [8, 10]. This enormous fund of knowledge along with the requisite annotation describing each gene is represented by a rich and diverse set of data elements. These include a total of 24 chromosomes representing approximately 3 x 10^9 base pairs of DNA sequence and the position of each of approximately 35,000 genes, 36,000 corresponding

GLOSSARY

Allele: One of the variant forms of a gene
Clinicopathology: Of or relating to clinical and/or pathology parameters
Exon: Protein coding portion of a gene
Expression/Transcription: Synthesis of mRNA from a DNA template
Functional genomics: Application of genome-wide experimental approaches to assess gene function
Homolog: Any member of a set of genes whose nucleotide sequences show a high degree of one-to-one correspondence
Intron: Intervening portion of a gene between exons, removed during the transcriptional process and not translated into protein
Metastatic disease: Cancer stage where the tumor has spread to remote tissues
Ortholog: Homologous sequences in different species that arose from a common ancestral gene during speciation
PCR: Polymerase Chain Reaction used to amplify and detect DNA
Polygenic diseases: An inherited disease controlled by several genes at once
Proteomics: The identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ, or organism that can be studied in concert to provide accurate and comprehensive data about that system.
Primer: A short synthetic piece of DNA used to initiate a PCR reaction
Transcriptional profiles: mRNA content of tissues
Tumorigenesis: Process of tumor formation

messenger RNAs and proteins, and coding region or exonic and intervening region or intronic coordinates. In addition, the following types of genomic annotation data need to be stored:

- Over 2,000,000 Single Nucleotide Polymorphisms or SNPs: These are sequence variations, which together create a unique DNA pattern in each person.
- Over 20,000 protein domains: These are independent sub-regions of proteins known to have specified functionality
- Annotation describing a gene product's molecular function, cellular compartmentalization, and/or biological process
- Protein-protein interaction and pathway information

Gene annotations reside in multiple publicly available biomedical databases, and acquiring gene annotations from various data sources involves identifying important and reliable data sources, regularly querying these sources, parsing and interpreting the results, and establishing associations between related entities. There are major difficulties at every step of this process. Each data source has custom text formats, and these formats change occasionally. Furthermore, an entire data source may be retired or completely restructured using a new schema. In addition, genomic data sources are usually updated on different schedules, and the size of such data sources may prohibit all versions of a data source from being loaded into a data warehouse. Finally, some data sources are inconsistent at the semantic level, and frequently, there is inadequate use of controlled vocabularies and common data elements to specify the metadata.

The National Center for Biotechnology Information (NCBI) is one major resource that maintains public biomedical annotation databases [17]. It includes nucleotide and protein sequence (GenBank), structure (MMDB), genome (RefSeq), and expression (GEO) databases. The OMIM (Online Mendelian Inheritance in Man) database is a catalog of human genes and genetic disorders [16]. GO (Gene Ontology) is a popular database that contains information about the cellular localization, molecular function, and the biological process in which a gene product is involved [1]. PubMed is a literature database from the National Library of Medicine, which includes over 14 million citations for biomedical articles [16].

Our warehouse fetches and stores annotations for all genes represented in humans and several other model organisms from OMIM, GO, PubMed and the following databases of NCBI: UniGene, dbSNP, RefSeq, HomoloGene, and LocusLink.

2.2 Microarray Profiling

A microarray is designed to detect the mRNA content (Expression Profiling) or the genomic DNA content (Comparative Genomic Hybridization) of thousands of genes in a particular tissue, cell type or pathological process [5, 12, 13]. It is based on the principle of hybridization between targets and probes. In this experimental modality, fluorescent-labeled nucleic acid from a sample of interest is called the target while short DNA fragments attached to a microarray are called probes. Probes on a single microarray represent most genes in the entire genome. Array experiments are typically conducted using one of two experimental formats. In the single channel system, a single sample of biological material is labeled with a fluorescent dye, hybridized to an oligonucleotide array, and the intensity value at each oligonucleotide is determined. In the two-channel system, a pair of samples is labeled with different fluorophores, hybridized to an oligonucleotide or cDNA array, and the intensity value of each fluorophore at each spot is determined.

The measured fluorescent values are meaningful only in the context of sample metadata (e.g. prostate versus breast tumor or benign versus malignant) and the associated genomic annotation of "interesting" probe sequences. Therefore, a gene expression data management system must integrate data from three different data sets: *gene expression measurements, sample metadata, and gene annotations*.

The data generated by a microarray system contains several data types. Typically, it includes

- Raw data consisting of binary image files generated by scanners
- Probe intensity data consisting of numerical values associated with each probe
- Summarized gene expression data estimates generated by combining probes representing the same gene

In a single microarray experiment, a raw image file is approximately 50 MB in size, the probe intensity data file is approximately 12MB, and the summarized gene expression data consists of between 12,000-50,000 values. Typically, a biologist would conduct between 5-100 such chip experiments and would thus have to store, query, analyze, and visualize ~100K-2500K data points.

2.3 Mutation Profiling

Microarray gene expression profiling has identified numerous genes in important pathways whose expression is altered in complex diseases. A complementary experimental modality involves the precise and comprehensive definition of the genetic changes, which are responsible for disease development or susceptibility, at the DNA level [2, 15]. This experimental methodology called mutation profiling is now possible due to the progress made in large scale DNA sequencing. Biomedical researchers may now sequence hundreds of genes in hundreds of tissue samples to identify mutations responsible for the disease phenotype.

This experimental modality also generates a rich set of data types. These include:

- Binary data: Sequencing a gene in one sample generates approximately 12-20 trace files, each of which is ~35 KB. For each gene sequenced in each sample, a binary analysis file (~6 MB) is generated. These are stored as BLOBs.
- DNA base information: The DNA sequence, quality information, and mutation probability are stored for ~500 bases in each of the trace files. This data is a combination of character, string, integer, and float data types.
- Consensus data: A consensus sequence is the overall DNA sequence derived by integrating the sequence information from all the traces of a gene. For each gene sequenced in a sample, the consensus sequence and its alignment to the reference DNA sequence are stored. This data is stored as a large string in a CLOB.

2.4 Proteomics Analysis

This analysis is aimed at high throughput separation and identification of proteins that are differentially expressed in a disease state as compared to the healthy state [9]. 2D PAGE (**2** **D**imensional **P**oly**A**crylamide **G**el **E**lectrophoresis) is by far the most commonly used method for protein separation. In this method, a complex mixture of proteins is first separated into bands based on the isoelectric point using Immobilized pH Gradient (IPG) gels. These bands of proteins are further separated into spots after being subjected to mass based separation using Sodium Dodecyl Sulfate (SDS) gels. Every spot on the gel roughly corresponds to one protein. Spot volume ratio comparison in disease state vs. normal state helps in selecting only those protein spots that are significantly different. Mass spectrometry analysis, single (MS) or tandem (MS/MS), after digestion (i.e. fragmentation) of selected spots generates corresponding spectra. These experimental spectra are compared with theoretical spectra of protein sequence digests using various software tools in order to identify the protein at each spot.

The data generated by these experiments consists of the following:

- Images (~50 MB per gel): Depending on the number of samples loaded on a gel, a gel may be scanned at multiple wavelengths to generate several image files.
- Workspaces (~50 MB per gel): Containing image analysis and comparison details
- Metadata: Describing experiments, samples gels, and spots excised from gels.
- Mass spectrometry data (~100 MB per spot)
- Spectral similarity reports

2.5 Clinical Data

Clinical data refers to any information that is contained in a patient's medical record. This information may be acquired from notes derived from a hospital admission or a doctor's visit. This data comes in various forms such as text or numbers (patient identification, demographics, history, laboratory data, etc), analog or digital signals (ECG, EEG, EMG, ENG etc), images (histological, radiological, ultrasound, etc), and videos. Furthermore, clinical studies involve specimen collection from multiple patients. The complete specimen may not be consumed at once and may be preserved in a specimen bank. Therefore, all of this patient-derived clinical and specimen-derived pathology data must be interlinked to research results derived from analyzing DNA, RNA, or protein samples. Further complicating the storage of this data is the fact that because patient identification information cannot be publicly accessible by law (HIPAA) [11], such identifiers must be removed and decoupled from other clinical parameters. Apart from humans, specimen collection and genome-wide profiling experiments may be conducted in other species such as mouse and rat. The major difficulty in storing this type of data is that each disease and species can only be adequately described using greatly different vocabularies and data elements.

3. Our Solution

In this section, we present our data warehouse solution for integrating various biomedical data sets deployed at Washington University School of Medicine. The major goals of this data integration project and the resulting Data Warehouse from the perspective of the university are:

- To develop an informatics center that will allow investigators to collect and manage large amounts of gene expression, gene sequence, proteomics, and coded clinicopathology data generated from various research studies.
- To develop data mining and analysis tools that will allow investigators to generate and validate new hypotheses based on the integration of collected functional genomic and clinicopathology data sets.
- To provide a publicly accessible venue for "publishing" experimental findings and corresponding data sets generated from investigator-based studies.
- To provide authentication, authorization and security such that investigators can give access privileges to other investigators on the data sets owned by them.

The data warehouse integrates data from the following important core facilities:

- Microarray Facility (MAF): Performs microarray experiments for investigators
- Washington University Genome Sequencing Center (GSC): Performs high throughput sequencing for mutation profiling
- Proteomics Facility (PRF): Provides access to proteomics technologies for molecular profiling
- Siteman Clinical Information Portal (SCIP): Collects and stores patient-derived clinical parameters
- Tissue Procurement Facility (TPF): Collects, stores and tracks anonymized patient identifiers, associated tissue specimens, and pathology data; generates DNA, RNA, and protein samples to be analyzed by the GSC, MAF, and PRF (Collectively called the Functional Genomic Cores or FGCs) respectively

Apart from these facilities, data is also integrated from reliable publicly available annotation data sources. The data from the above sources is integrated using the following workflow. Typically, patients are enrolled in a clinical trial, and appropriate clinical parameters depending upon the disease under question are curated from the medical record and stored in SCIP. Anonymized patient identifiers are also entered into the TPF database. As part of the clinical trial, one or more specimens are collected, tracked and stored at the TPF. The TPF processes these specimens to produce DNA, RNA, and/or protein and assigns each of these specimens and samples tracking identifiers. These biomolecular samples are then sent to the appropriate FGC to conduct microarray, mutation, and/or proteomic profiling experiments. After completion of the experiments, these FGCs load experimental data into our data warehouse. Similarly, clinical and pathology data as well as requisite inter-relationships between patient, specimen, and sample identifiers are loaded into our database from SCIP and the TPF. Genes represented in each of the experimental paradigms are annotated by importing data into our warehouse from multiple, publicly available biomedical data sources described in Section 2. Thus, clinical data (from SCIP), specimen, sample, and pathology data and their inter-relationships (from TPF), experimental data (from FGCs), and gene annotation data (from publicly available annotation databases) are loaded into our data warehouse (Figure 1).

Our data warehouse runs on Oracle 9i (version 9.2.0.4- 64 bit) database which is hosted on a Sun Enterprise 420R consisting of 4 X 450 Mhz Ultra Sparc-II processors, 4 GB of internal RAM memory, 36 GB of mirrored internal drive space, and two Sun StorEdge A1000 RAID boxes, containing 654 GB of total disk space set up to operate in RAID 5 mode. The warehouse currently has about 150 GB of data.

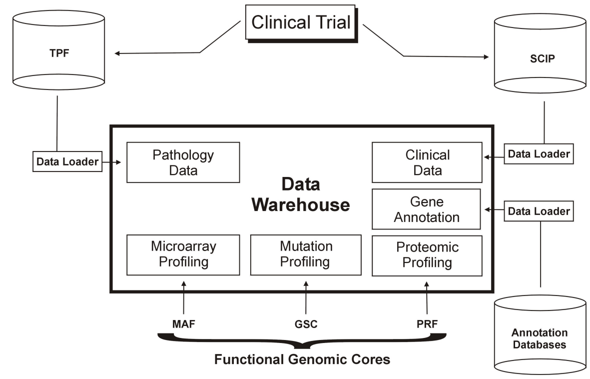

Figure 1. Clinical data (from SCIP), specimen, sample, and pathology data and their inter-relationships (from TPF), experimental data (from FGCs), and gene annotation data (from publicly available annotation databases) are loaded into our data warehouse.

We have also built analysis tools that let investigators perform integrated analysis and visualization of the various data sets stored in our data warehouse. The following subsections describe our software components that address issues of importing, querying, analyzing, and visualizing the biomedical data sets described in Section 2.

3.1 Function Express Server

Function Express Server, which is written in Java, extracts annotation data from publicly available gene annotation databases, loads it into the warehouse, and links it to genes represented in microarray, sequence, and proteomic data. Integrating these annotations into a data warehouse facilitates better representation of semantics, enhanced query performance, and superior data quality.

Function Express Server includes an Extract-Transform-Load (ETL) tool that downloads various reliable publicly available gene annotation databases, parses the data to extract relevant annotation such as the gene name and chromosomal localization, and loads them into the warehouse. Because annotation sources have custom text formats, parsers are written for each biomedical data source. These parsers are used to load data into our warehouse, and automatic updates to the warehouse are conducted at a user-specified frequency, a necessary feature as the data in the annotation databases is being updated frequently. As most annotation sources do not provide deltas, updates are detected in the ETL process.

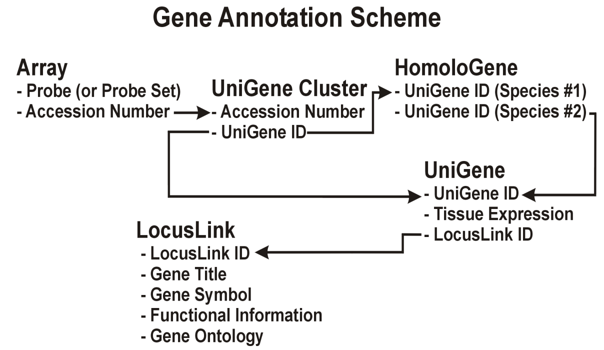

Figure 2. Gene annotation scheme portraying the links required to annotate genes represented as probes on microarrays. Genes are either directly linked via the UniGene cluster and LocusLink ID or indirectly via HomoloGene.

Function Express Server currently fetches annotations for all genes represented in humans as well as other major model organisms from UniGene, LocusLink, HomoloGene, dbSNP, OMIM, Gene Ontology, and PubMed. The linking of these annotations with individual spots or probes on a microarray is conducted as follows (Figure 2). Each probe or probeset on a microarray is linked to an accession number, a unique identifier issued by GenBank to represent a nucleotide sequence. These accession numbers are grouped together into **Uni**que **Gene** (UniGene) clusters by sequence homology. Each cluster is assigned a unique UniGene ID which in turn can be linked to a gene identified during the genome sequencing process. Each of these genes is assigned to or **link**ed to a chromosomal **Locus** (LocusLink) and is assigned a unique LocusLink ID. Individual annotations such as functional categorization (Gene Ontology), chromosomal localization (LocusLink), tissue expression (UniGene), DNA sequence variation (dbSNP), links to disease (OMIM), and gene homologs and orthologs (HomoloGene- see below) may be acquired using the accession number, UniGene ID, or LocusLink ID.

To enrich the annotation for each gene, an additional resource, HomoloGene, provided by the National Center for Biotechnology Information (NCBI) is utilized. Two gene sequences are said to be homologs of each other if they share significant sequence similarity. The HomoloGene database calculates homologs by nucleotide sequence comparisons between genes across organisms (human, mouse, rat, cow, zebrafish, frog, and fly). Using HomoloGene, we can relate functional annotation information for the same gene across species (called orthologs). Thus, while a rat or mouse gene may not be annotated with any functional information in UniGene or LocusLink, its human ortholog may be extensively annotated. The functionality of orthologous genes across species is known to be similar, and this fact is used to infer the functionality of genes that are not annotated. Because we have linked probesets on different microarray platforms (e.g. single versus two channel) to standard identifiers (Accession Number, UniGene ID, and/or LocusLink ID and HomoloGene), our database can automatically link orthologous genes from different array designs of the same or different species. Currently, we provide automated annotation for probe sets from 52 chip types representing human, mouse, and rat genes, which facilitate studies across different species.

Once the base annotation data is loaded, a set of materialized views are created in a format supportive of the queries that would run against the warehouse. In these views, not only is the annotation about each gene saved, but hierarchical trees are also generated for annotation imported from Gene Ontology (functional categorization), UniGene (tissue expression), and LocusLink (chromosomal localization).

Our data integration approach facilitates powerful queries on the annotations from multiple sources. For example, it is feasible to view all genes that are transcription factors (Gene Ontology), all genes expressed in pancreas (UniGene), or all genes located on chromosome 1p31 (LocusLink).

3.2 Chip Import Utility

The Chip Import Utility (CIU), a microarray data loader application which is written in Java, is used at the MAF. Using the CIU, the data generated from this core facility and possibly other future microarray facilities are sent to the data warehouse. The GUI allows the database curator to create and enter metadata about new experiments, investigators, samples and chips. Once this metadata is entered, the data, which includes a raw image file and one or more primary numerical fluorescence intensity files, is uploaded into the data warehouse.

However, prior to importing chip data, information about the array design or array metadata on which the experiment was performed must be provided. Since there can be multiple sources and, thus, multiple vendors for these microarrays, we have formulated a general mechanism by which the information about the arrays may be imported into the database. Namely, the species from which the probes on the array were synthesized needs to be specified (i.e. human for the Affymetrix HG-U95A, mouse for the MG-U74A, etc.). Next, the array must be given a unique name (i.e. HG-U95A, MG-U74A, etc.), and each probe must be given a unique ID (i.e. 1000_at, 1001_at, etc.). Finally, to provide automatic annotation, an accession number, UniGene ID, or LocusLink ID must be provided for each probe. Once the array metadata has been imported, the array is "registered" in the database, and data derived from experiments using this array can be imported.

The importing of chip data is complicated by the fact that the data files may be in various formats. For example, there may be a header prior to the actual data,

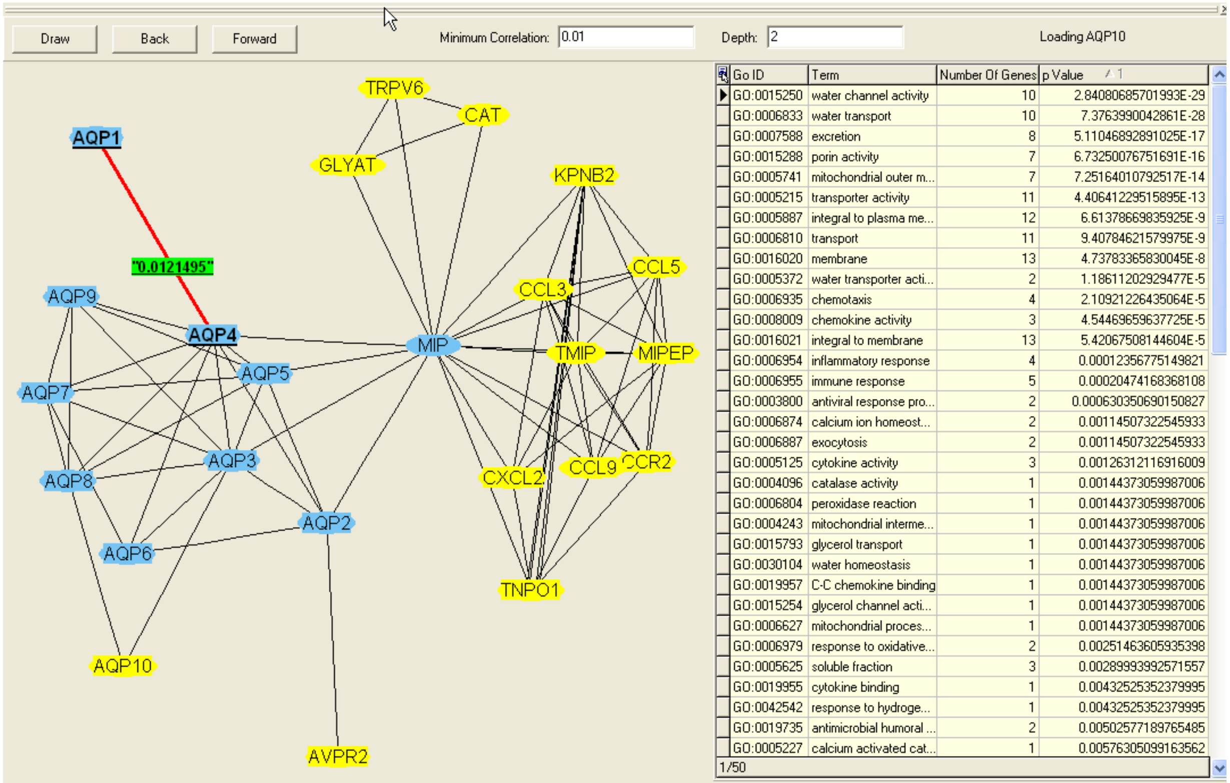

Figure 3. The Literature Gene Network window in the Function Express client displays a gene-gene literature co-citation network, centering around a selected gene (AQP4). The depth of the network (e.g. depth=2 includes all genes linked by two or fewer edges to the selected gene) and the strength of gene interaction (termed Minimum Correlation) are user configurable. The genes in this network are further grouped by Gene Ontology categories and the p value for the overrepresentation of each category in this gene set is given.

only some columns may be important to import, and different characters may delimit columns. While importing the chip data, the user will be required to provide information about these formats. To avoid entering such details again and again, an "Import Template" wizard is provided for each file format where in the user enters information about the format of data files. One can then save this template and use it for subsequent imports.

3.3 Function Express Client

The powerful features of our annotation and microarray data warehouse are leveraged by the data mining and visualization capabilities of the Function Express Client (FE). In FE, which is written in C++ using Borland Builder Enterprise 6.0, gene annotation data is accessed on demand from the database and can be coupled to gene expression data sets that are independently loaded from the MAF. Using FE it is possible to perform complex data queries using both expression and gene annotation data. For example, expression data may be filtered, normalized, and clustered; this facilitates identification of genes which are co-regulated and thus, are inferred to be involved in a particular disease process. Results of such analyses may be visualized in the context of gene annotation data. Examples of this include:

- Visualize expression of all genes that are transcription factors located on chromosome 1p and that are down-regulated in tumor samples relative to non-malignant tissue
- View expression of selected genes across different experiments conducted in same or different species on same or differing array platforms
- Display literature-based gene to gene co-citation networks

To facilitate displaying a literature-based gene network, we link over 12 million abstracts for over 500,000 gene names representing almost 200,000 distinct LocusLink IDs (or genes). The weight of a gene-gene link is calculated based on the number of abstracts where both genes are mentioned (Figure 3).

With the combined data warehouse/FE platform, the ability to access data from multiple sources for *simultaneous* meta-analysis becomes straightforward, thus increasing the analytical power of many of these studies. Through this platform, it is also possible to seamlessly

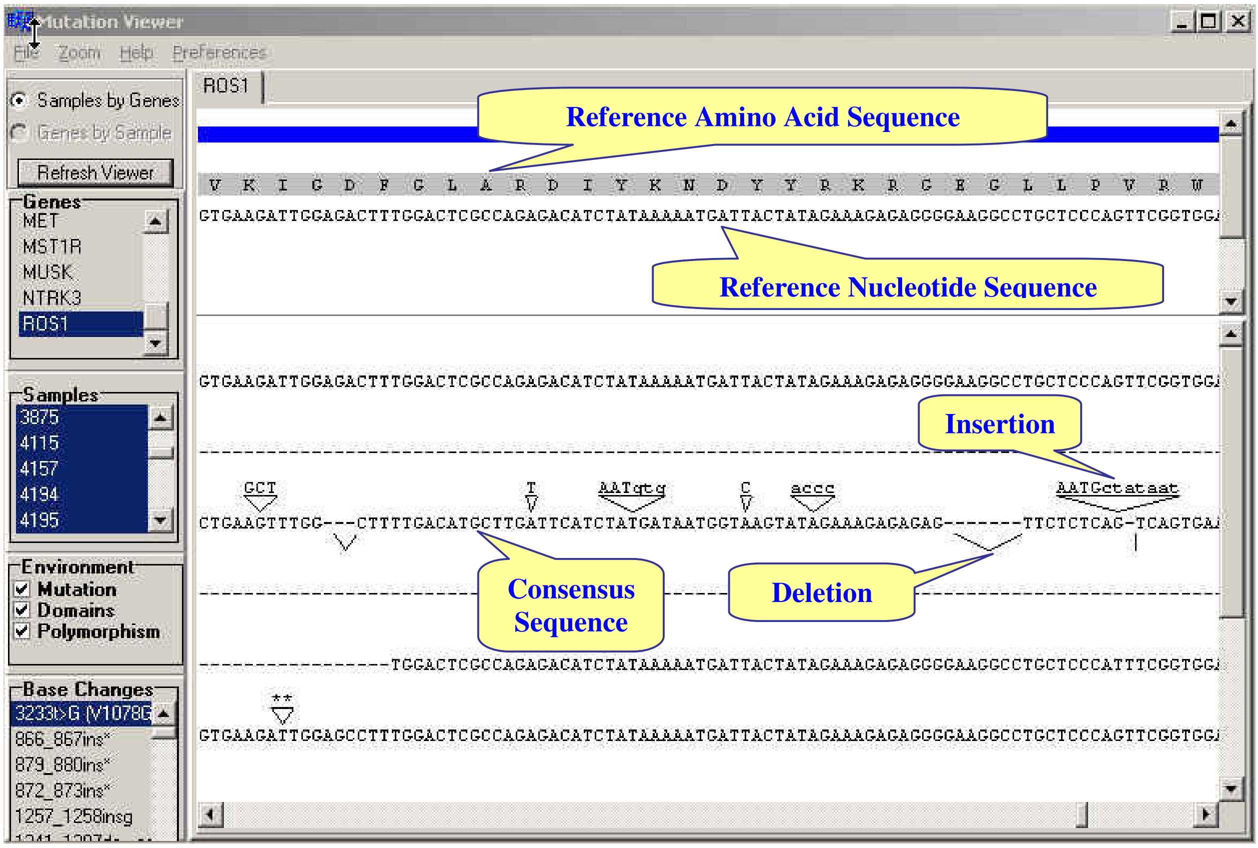

Figure 4. An example of the output from MV is shown. Selected genes, samples, and environmental features (left panel- upper two list boxes and checkbox panel) are drawn in the right two panes. The upper pane contains the reference nucleotide and amino acid sequence for ROS1. In the lower pane, each individual nucleotide sequence represents a consensus sequence from an individual patient sample, and the position of identified mutations (insertions and deletions) is indicated. Identified mutations are automatically named and are present in the lower left listbox.

integrate information from different organisms, thus clues gleaned from mouse models can be easily examined in studies involving human specimens. For instance, genes differentially regulated in a mouse model of cancer can be selected and immediately examined in human data sets to determine whether they are expressed aberrantly in corresponding human tumors. This database and software suite is constructed as an integrated set of modules so that additional genomic information, such as that derived from mutational profiling and proteomics, can also be incorporated and analyzed along with expression profiling data.

3.4 Mutation Viewer Pipeline

We have set up a process and software infrastructure where we have integrated data from the TPF and the GSC with gene annotation so that high throughput mutation profiling of large numbers of genes in hundreds of samples may be conducted. The workflow of this mutation profiling pipeline, which includes three key pieces of software we have developed, is described as follows.

DNA samples from selected specimens are bar coded, and sent to the GSC by the TPF. The design of primers for PCR amplification of selected genes is facilitated by an application, which we wrote using Borland Builder Enterprise 6.0. This software, which is a wrapper around the popular Primer3 software package, automatically designs primers for large numbers of genes in high throughput. These primers are designed using a known normal sequence called the reference sequence, which has been imported into our database by the Function Express Server from RefSeq. This primer information is then transmitted to the GSC where every primer is assigned a unique bar code, thus ensuring accurate tracking of each experiment. At the GSC, selected genes are sequenced in patient DNA samples in high throughput. The sequencing results, which may be visualized as plots for each of the bases present in DNA, need to be analyzed in an automated fashion as a single mutation profiling project may generate thousands of sequence trace files. For example, a project where 100 genes are sequenced in 100 patient samples would generate approximately 160,000 trace files.

To analyze the large number of sequence traces, we have designed and written software in Perl in collaboration with Informax, Inc. This software calculates the probability of a mutation at each base in a trace using a neural net algorithm. Traces of a single gene from an individual patient sample are analyzed together for sequence quality and are grouped together based on sequence homology to generate contigs, regions of overlapping DNA sequences. These contigs are then aligned to the reference sequence, and automated mutation/polymorphism detection is performed. The results of the initial tests with this software appear to be extremely accurate as it was necessary to manually inspect less than 1 per 1000 base calls in pilot projects. This allows us to dramatically reduce the number of traces that need to be inspected manually for potential sequence alterations. This is crucial as most mutations will occur in only a single allele and will therefore show up as 'mixed peaks' on the traces (See Figure 5). Individual trace files, consensus sequences, alignment information with respect to the reference sequence, and mutation confidence scores are imported into our warehouse for each gene-sample combination where it is then interrelated to protein domain and SNP data for each gene.

To visualize this data and to extract the salient information, we have developed a graphical user interface in C++ using Borland Builder Enterprise 6.0. In this application, called Mutation Viewer (MV, Figure 4), protein motifs (e.g. kinase domain) are shown on the DNA schematic, and mutations/polymorphisms are then "painted" onto this scaffold of protein domains, so that alterations in critical domains are easily appreciated. The presence of known SNPs (derived by scanning dbSNP) within each individual DNA are also noted on this viewer, thus commonly occurring polymorphisms can be quickly eliminated from further analysis. Furthermore, the program prioritizes mutations based on their potential functional significance (synonymous vs. non-synonymous substitutions) as well as frequency. It is also possible to zoom-in such that the amino acid and nucleotide sequence for reference and consensus sequences may be seen.

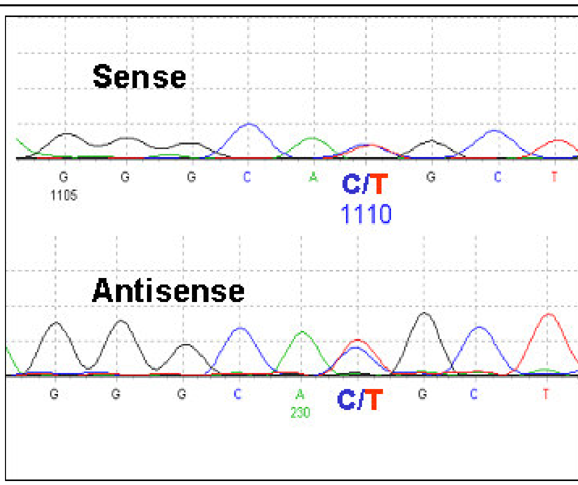

Figure 5. Example of sequence traces demonstrating a mixed peak found in a lung tumor.

Finally, the actual trace files derived from sequencing may be viewed for any consensus sequence, thus allowing the user to verify the computer-based identification of a mutation (Figure 5).

3.5 Proteomics LIMS

Experiments conducted at the PRF are aimed at identifying differentially expressed proteins in two or more patient samples (e.g. tumor versus normal) in high throughput. Protein samples, which are acquired from the TPF, are first separated into "spots" using 2D gel electrophoresis. Mass spectrometry and subsequent database searching then identify the protein at each spot. The complete workflow involves a number of laboratory steps that are sequential in nature. Users are assigned different roles based on what portion of the workflow they perform. Proteomics LIMS, which is written in Java using the Eclipse framework, is designed to automate the information flow among these roles. It also helps users in doing gel related calculations, reporting and visualization. The LIMS allows users to examine the current job queue, metadata information about each lab element, and the status of a particular lab element (tracking). The stages of the laboratory workflow where the LIMS contributes are as follows:

- Sample Procurement and Experiment Design: An investigator logs into the TPF system to request samples from already existing specimens at the TPF or to enter new specimens or samples. Metadata such as specimen source, associated pathology data, and isolation protocol are entered for each new sample. Using these samples, the investigator may design a proteomics experiment in terms of number of gels to be used and sample allocation for each gel. The completed experiment design is then sent electronically to the PRF for execution, and samples are delivered to the PRF by the TPF or the investigator.
- Sample Curation: This role uses LIMS for validating sample metadata and experiment designs.
- Sample Processing: Samples are depleted of unwanted proteins and then are subjected to a protein concentration assay. Protein concentrations are used by the LIMS to calculate the volume of each sample to be loaded on a gel.
- Gel Processing: Fluorescently labeled samples are run on IPG and SDS gels as a part of the 2D-electrophoresis. These gels are then scanned at multiple wavelengths resulting in images files that are generated for each fluorescently labeled sample. Comparative analysis of all the images from one gel or an entire experiment (consisting of multiple gels) is done using third party software to mark differentially expressed protein

spots. The LIMS stores primary image files, comparison results, and the list of differentially expressed spots. The LIMS has also implemented an image manipulation algorithm which maps coordinates of each spot from the original image file to coordinates of a second image file that the robotic picker requires for excising spots from the gel.

- Mass Spectrometry: Excised spots are trypsin (an enzyme) digested and subjected to two types of MS analysis using MALDI and ESI as ion sources. At this step, the LIMS is designed to upload spectral and peak related information for every spot into our data warehouse.

3.6 Proteomics Searching and Reporting Tools

The spectrum of a spot is searched against theoretically digested proteins from a protein sequence database to find matches. Identified proteins are automatically linked to their annotations from LocusLink, GO, UniGene and other available data sources in our data warehouse. To visualize this information, we have developed a reporting tool in Borland Builder Enterprise 6.0 that visually annotates a protein sequence with the MS and the MS/MS hits identified by multiple similarity search algorithms.

3.7 Specimen registration and banking system

Investigators use specimens and samples to carry out multi-modal (genomics, proteomics, and clinical) experiments. These specimens may be shared across experiments, modalities, or even investigators. Thus, it should be possible to analyze the results of shared specimens simultaneously. Also, it should be possible for investigators to identify specimens of interest obtained by others that have not yet been analyzed completely. This is the objective of caTIS (**ca**ncer **TIS**sue repository), the specimen registration and banking system, we are developing for the TPF. For this system to serve its purpose, every investigator must use it as a single point of entry to register their specimens. This system will be the gateway for submitting, obtaining, and querying specimens and biomolecular samples**.** This tissue banking effort will allow for the correlation of research results from a single specimen across multiple experimental modalities.

3.8 Authorization and Security

Biomedical research is frequently collaborative with researchers at a university sharing experimental data and results of analyses. Thus, we have developed extensive user authorization and security modules for our data warehouse where investigators use a single sign on system. Because different core facilities have varying data objects and differing access control requirements, our system provides role-based as well as object-based security.

In object-based security, privileges can be defined for both the data and metadata. For example, the metadata generated by the MAF is accessible to all users by default, and this allows investigators to query and search for microarray experiments of interest. An investigator can share the data of his experiment in the warehouse with a select set of investigators. Apart from the experimental data, the analysis results (e.g. a collection of 'interesting' genes) can also be shared and published.

In role-based security, roles are assigned to each user, and, each role has certain privileges (e.g. accessing or modifying a type of object). Therefore, a user having a specific role will be able to perform only the set of actions permitted for that role. For example, in the Proteomics LIMS, a user assigned to a particular role, may perform only certain workflows.

4. Emerging Trends

The data warehousing based solution we have implemented has the advantage of having all the data in one place, with the data transformed to match the desired queries. The queries are fast, and there is no dependence on individual data sources. There are, however, several further challenges we face.

First, with the tremendous diversity of data elements present in the biomedical domain, especially in the storage and representation of clinicopathology data such as that found in caTIS and SCIP, flexible and extensible data storage models must be utilized. One such medium for data storage and exchange, XML, is becoming increasingly important within the bioinformatics community. Since XML allows uniform description of data and metadata, it can be efficiently used to specify ontological descriptions of biomedical data. However, XML formats, like flat file formats, can be large, and complex, making data access difficult and inefficient. Better techniques on compression and lazy loading of XML data are required to make XML the universal medium for the storage of biological data.

Second, semantic integration is an important challenge that needs to be addressed. For example, the same protein sequence is known by different names or accession numbers in different biomedical data sources. These nomenclature differences must be resolved in order to integrate these data sources. One of the emerging trends is an effort to define semantics precisely through ontologies that attempt to capture concepts, objects, and their relationships within a biological domain. For example, Gene Ontology is a popular database that contains information about a gene product’s cellular localization, molecular function, and biological process [1]. These ontologies encapsulating controlled vocabularies may be utilized in object models with defined data elements to describe and define entities. Additionally, there is a need for data models that efficiently store the objects and data persistently.

Such new standards, vocabularies and common data elements are evolving for different biological data sets. For example, the Microarray Gene Expression Database Group (MGED) has consolidated standardization efforts for microarray data. MGED is a consortium of academic and commercial organizations with the shared goal of defining standard formats that will allow gene expression data repositories to share and exchange data. MGED has recently published the **M**inimum **I**nformation **A**bout a **M**icroarray **E**xperiment (MIAME) standard, to enable interpretation of the results of an experiment unambiguously, and, potentially to reproduce the experiment. They have also developed a data exchange format, **M**icro**A**rray **G**ene **E**xpression **M**arkup **L**anguage (MAGE-ML) and an object model, **M**icro**A**rray **G**ene **E**xpression **O**bject **M**odel (MAGE-OM). Similar to microarray data sets, HUPO (Human Proteome Organization) is developing a standard called **M**inimum **I**nformation **A**bout a **P**roteomics **E**xperiment (MIAPE). Again as with MIAME, this minimum information will be described using an ontology that not only contains vocabulary terms for describing proteomics related concepts but also defines the interrelationship between these terms.

While MIAME and MIAPE provide useful guidelines for organizing gene expression and proteomic data into a database, such adequate standards do not yet exist for the description of clinicopathology data acquired from patients afflicted with most polygenic diseases. One exception is in the field of cancer, where the National Cancer Institute Center of Bioinformatics (NCICB) has made considerable progress in developing such standards. Their Enterprise Vocabulary Service provides a controlled vocabulary for the cancer domain, and their Cancer Data Standards Repository contains a set of standardized data elements used in cancer research. While many other biomedical disease domains may be able to “borrow” essential elements and design principles from these standards, each disease research initiative will ultimately have to develop such controlled vocabularies and data elements in order to facilitate data integration and reliable representation. Third, as we move from a university to a national level, our data warehousing solution may not scale when different annotation, experimental and clinical data is gathered at multiple institutions. Again, in the field of cancer, NCICB has recently started an initiative, the **ca**ncer **B**iomedical **I**nformatics **G**rid (caBIG) that will tackle such issues. This initiative aims to deploy an integrating biomedical informatics infrastructure that will connect all the cancer centers across the United States and worldwide.

5. Future Work

We have developed and deployed a data warehousing based solution for data management, integration and analysis of the biomedical data at Washington University School of Medicine. We have also developed tools to store, query, analyze, and visualize the data sets available from core facilities and publicly available annotation data sources.

Continuing this work, we will add more extensive annotation databases and will implement common data elements and underlying controlled vocabularies. We are also currently in the process of defining XML descriptions for clinical and pathology parameters and for mutation and sequence information. To facilitate this process, Oracle9i has a dedicated XML datatype called XMLTYPE where Oracle internally shreds the XML data and puts it in separate tables. In addition, we are reconfiguring the Function Express data import and export capabilities to be MIAME/MAGEML compliant by using the MAGE-OM, so that data from different types of microarray platforms may be analyzed simultaneously. Additional future work includes making Function Express and Mutation Viewer caBIG interoperable by communicating with caBIG databases, using caBIG common data elements, ontologies and vocabularies, and supporting caBIG-compatible APIs. Finally, we will expose a web services API to deliver the linked annotation from our warehouse to the outside world.

6. Conclusions

Although difficult to achieve, database interoperability is critical to the future of biomedical research. The longer this capability is delayed, the more difficult and costly establishing interconnectivity will become. The community at large should come together and build systems that conform to standards which will support common data interchange formats, dynamic, programmatic access to local and remote data sources, and common application programming interfaces. The major issue in the integration of biomedical data is the large number of distributed, semantically disparate data sources that need to be combined into a useful and usable system for biologists. The challenges are big, but so are the rewards. For the first time many incurable illnesses may be effectively treated and even cured as integrated research becomes feasible.

Acknowledgments

We would like to thank Arvind Hulgeri for all of his effort and advice in critical editing of this paper. Our most sincere gratitude goes to Anand Deshpande for his overall guidance and assistance during the initial conception of this manuscript. Finally, we wish to thank our incredibly intelligent, hard working, and dedicated team of software engineers at Persistent Systems and at Washington University who have developed the software applications described in this paper.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, pp. 25-9, 2000.

[2] A. Bardelli, D. W. Parsons, N. Silliman, et al., "Mutational analysis of the tyrosine kinome in colorectal cancers," *Science*, vol. 300, pp. 949, 2003.

[3] S. Y. Chung and L. Wong, "Kleisli: a new tool for data integration in biology," *Trends Biotechnol*, vol. 17, pp. 351-5, 1999.

[4] S. B. Davidson, J. Crabtree, B. P. Brunk, et al., "K2/Kleisli and GUS: Experiments in integrated access to genomic data sources," *IBM Systems Journal*, vol. 40, pp. 512-531, 2001.

[5] F. Forozan, R. Karhu, J. Kononen, et al., "Genome screening by comparative genomic hybridization," *Trends Genet*, vol. 13, pp. 405-9, 1997.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-7, 1999.

[7] L. M. Haas, P. M. Schwarz, P. Kodali, et al., "DiscoveryLink: A system for integrated access to life sciences data sources," *IBM Systems Journal*, vol. 40, pp. 489-511, 2001.

[8] E. S. Lander, L. M. Linton, B. Birren, et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921, 2001.

[9] K. H. Lee, "Proteomics: a technology-driven and technology-limited discovery science," *Trends Biotechnol*, vol. 19, pp. 217-22, 2001.

[10] M. Olivier, A. Aggarwal, J. Allen, et al., "A high-resolution radiation hybrid map of the human genome draft sequence," *Science*, vol. 291, pp. 1298-302, 2001.

[11] L. H. Prince and A. Carroll-Barefield, "Management implications of the Health Insurance Portability and Accountability Act," *Health Care Manag (Frederick)*, vol. 19, pp. 44-9, 2000.

[12] G. Ramsay, "DNA chips: state-of-the art," *Nat Biotechnol*, vol. 16, pp. 40-4, 1998.

[13] M. Schena, R. A. Heller, T. P. Theriault, et al., "Microarrays: biotechnology's discovery platform for functional genomics," *Trends Biotechnol*, vol. 16, pp. 301-6, 1998.

[14] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A*, vol. 98, pp. 5116-21, 2001.

[15] Z. Wang, D. Shen, D. W. Parsons, et al., "Mutational analysis of the tyrosine phosphatome in colorectal cancers," *Science*, vol. 304, pp. 1164-6, 2004.

[16] D. L. Wheeler, C. Chappey, A. E. Lash, et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res*, vol. 28, pp. 10-4, 2000.

[17] R. M. Woodsmall and D. A. Benson, "Information resources at the National Center for Biotechnology Information," *Bull Med Libr Assoc*, vol. 81, pp. 282-4, 1993.

[18] E. M. Zdobnov, R. Lopez, R. Apweiler, et al., "The EBI SRS server-new features," *Bioinformatics*, vol. 18, pp. 1149-50, 2002.