Integration Challenges in Gene Expression Data Management

Victor M. Markowitz, John Campbell, I-Min A. Chen,
Anthony Kosky, Krishna Palaniappan, Thodoros Topaloglou

Gene Logic Inc., Data Management Systems,
2001 Center Street, Berkeley, CA 94704, U.S.A.

Introduction

DNA microarrays have emerged as the leading technology for measuring gene expression, primarily because of their high throughput. A single microarray experiment provides measurements for the mRNA transcription level for tens of thousands of genes in parallel [14]. While this technology opens new opportunities for functional genomics and drug discovery applications, it also presents new bioinformatics and data management challenges arising from the need to capture, organize, interpret and archive vast amounts of experimental data. Furthermore, in order to support meaningful biological reasoning, gene expression data need to be analyzed in the context of rich sample and gene annotations.

GeneExpress® is a data management system that contains quantitative gene expression information for thousands of normal and diseased samples, and for experimental animal model and cellular tissues, generated under a variety of treatment conditions [4]. Initially the GeneExpress system was developed with the goal of supporting effective exploration, analysis and management of gene expression data generated at Gene Logic using the Affymetrix® GeneChip® platform [13], integrated with comprehensive information on samples, clinical profiles and rich gene annotations. Building such a system required resolving various data integration problems in order to associate gene expression data with sample data and gene annotations. A subsequent goal for the GeneExpress system was to provide support for incorporating gene expression data generated outside of Gene Logic. Addressing this additional goal required the resolution of various levels of syntactic and semantic heterogeneity of sample data, gene annotations and gene expression data, often generated under different experimental conditions,. These goals have been addressed using a data warehousing methodology adapted to the special requirements of the gene expression domain [15].

In this chapter, we discuss the challenges associated with data integration in the context of a gene expression data management system, and describe how the GeneExpress system addresses these challenges. In Section 9.1, we provide an overview of the area of gene expression data management. In Section 9.2, we provide a brief description of Gene Logic's GeneExpress system. In Section 9.3, we discuss the key semantic problems associated with integrating gene expression and related data, and how they are addressed in the context of GeneExpress. In Section 9.4, we describe how third-party gene expression data can be integrated into GeneExpress. We conclude the chapter with a summary and observations in Section 9.5.

9.1. Gene Expression Data Management: Background

In this section we briefly review the gene expression data application. First, we discuss the data spaces that need to be modeled by a gene expression data management system, and then discuss initiatives to establish standards for gene expression and related data.

9.1.1. Gene Expression Data Spaces

Gene expression systems measure mRNA transcription level of protein-coding genes in a cell. The mRNA "mix" used in gene expression experiments is derived from biomaterials (samples) such as tissues and cell lines. A microarray is typically designed to detect thousands of specific target sequences associated with these genes through hybridization. The reported measurements are meaningful only when something is known about the samples and the target sequences and their associated genes. The first goal of gene expression data management is to integrate expression data with sample and gene annotations, and to allow users to use these annotations to explore, analyze and interpret expression data [3, 15]. Typically, a gene expression data management system integrates data from three different data spaces, namely sample annotations, gene annotations and gene expression measurements, which are described below.

Biological Sample Data Space

The main object in the sample data space is the *sample* representing the biological material that is involved in the experiment. Samples originate from a variety of sources with different data standards and handling protocols. Annotations associated with each such sample should address its physical features and quality, as well as the accuracy and extent of the information recorded. Ultimately sample data are recorded in the sample data space of a gene expression system. A sample can be of tissue, cell or processed RNA type, and originates from a donor organism of a given species (e.g., human, mouse, rat). Attributes associated with samples describe their nature and condition (e.g., organ site, diagnosis, disease, stage of disease), as well as donor information (e.g., demographic and clinical record for human donors, or strain, genetic modification and treatment information for animal donors). Samples are commonly organized in groups that can be further grouped into studies or projects, such as time/dose studies. Information on how samples in such groups are related to one another is therefore a necessary annotation for the sample data space.

Gene Annotations Data Space

Gene annotations help to associate the expression data reported for sequence fragments on the microarray to biological entities such as genes and proteins. The main problem here is that sequence annotations, and annotations of the function of known genes, can change over time as the availability of more sequence, better computational tools and new research lead to better gene prediction results. Further, the sources for gene annotation are usually primary or consolidated databases that are heterogeneous and may contain inconsistent data. Consequently, the effort of keeping up-to-date gene annotation data for sequence fragments on microarrays combines the complexities of database integration with the ongoing research in the field of gene identification.

The main object in the gene annotation data space is the *gene fragment*, representing an entity for which the expression level is being determined. For microarray technologies, gene fragments are associated with a specific microarray type, such as a GeneChip human probe array (e.g. HG_U95A). The annotations associated with a gene fragment describe its biological context, including its associated primary EST sequence entry in Genbank, membership in a gene-oriented sequence cluster, association with a known gene (i.e., a gene that is recorded in an official nomenclature catalogue, such as the Human Gene Nomenclature Database [8]), functional characterization, such as GO annotations, and association to known metabolic and signalling pathways.

Gene Expression Measurement Data Space

Gene expression microarray systems are broadly classified into *single channel* and *two channel* systems. A single channel system takes a single sample of biological material and provides absolute measures of gene expression for that sample, while a two channel system takes a pair of samples and provides measurements of the difference in relative gene expression between the two samples. Single channel systems are best represented by the Affymetrix GeneChip platform [2]. In this chapter we focus on the management of gene expression data generated using the GeneChip platform. Note however that most data management and integration issues discussed in this chapter apply to gene expression data in general, irrespective of the underlying technology platform.

Typically, data generated by a microarray system can be classified into three data types, each representing a different level of abstraction. This hierarchy of data types is common, with slight differences, to all microarray platforms and consists of:

1. *raw data* consisting of binary image files generated by scanners;
2. *grid* or *probe intensity* data, consisting of values associated with each probe or oligo sequence examined on a microarray; and
3. *gene expression estimates*, generated by combining data on related probes on a microarray

Each data type may have multiple data formats or representations associated with it, such as text or binary file-based formats, or database representations.

The transformation between data types is carried out by platform specific algorithms. It is not uncommon to use more than one algorithm to transform data from one data type to the next (e.g., see [11, 12]). We briefly describe below the hierarchy of data types in the context of the GeneChip platform.

Affymetrix' GeneChip microarrays (also called *probe arrays*) are tiled with oligonucleotide sequences, each of 25 base-pairs in length, known as *probes*. Each probe is designed to hybridize to a known mRNA fragment representing a target gene or EST. Probes are grouped into *probe pairs*, each of which consists of a *perfect-match* (PM) and a *mismatch* (MM) probe, with the MM probe being created from the PM probe by changing the middle (13th) base in order to measure non-specific binding. Each target gene or EST is represented by a *probe set* consisting of up to 20 probe pairs.

A GeneChip probe array experiment involves RNA sample preparation, carrying out the probe array experiment (hybridization, washing, staining) and scanning the probe array [2]. The scanning process generates a file containing an image of the probe array, which constitutes the *raw data.*

The scanned images are interpreted using methods such as the GeneChip Microarray Suite (MAS) analysis algorithms. The MAS *cell averaging* algorithm averages pixel intensities and computes cell-level intensities, where each cell represents one probe on the probe array. The output from this process is a file containing the estimated intensities for each probe on the probe array, which constitutes the *probe data*. These intensities indicate the amount of hybridization that occurred for each oligonucleotide sequence on the array.

Probe intensity files can be further analyzed with methods such as the MAS *chip analysis* algorithm, which generates *gene expression estimates* by summarizing the intensities of each probe set that corresponds to a gene or EST fragment targeted by the probe array. Alternative gene expression estimates may be based on single or multiple (e.g., replicate) experiments.

The GeneChip LIMS provides support for transforming data between the different data types, and for loading the gene expression estimates into a relational database based on the AADM model [1].

The different data types and their associated formats, result in files or data structures of different sizes. For example, for an experiment using a HG-U133 GeneChip probe array, the raw image file is around 45MB in size, the probe intensity data file is around 12MB, and the summarized gene expression data consists of roughly 22,000 values.

9.1.2. Standards: Benefits and Limitations

Effective exploration of microarray data has been hindered by the variety and heterogeneity of the data formats used. This problem has been recognized by several organizations, such as the European Bioinformatics Institute (EBI)[1], the National Center for Biotechnology Information (NCBI)[2], and the National Center for Genome Resources (NCGR)[3], in their efforts to establish public data repositories for gene expression information. Microarray manufacturers have also proposed formats, such as the Affymetrix Analysis Data Model (AADM) used for the GeneChip LIMS relational database [1], in order to facilitate data exchange between different sources of gene expression data and the development of gene expression analysis packages.

Different standardization efforts have been consolidated by the Microarray Gene Expression Database Group (MGED)[4], a consortium of academic and commercial organizations with the shared goal of defining standard formats that would allow gene expression data repositories to share and exchange data. MGED has recently published MIAME, a recommendation for the minimum information required for a microarray experiment [3], and has developed a data exchange format (MAGE-ML) and object model (MAGE-OM) for microarray experiment data.

[1] http://www.ebi.ac.uk/arrayexpress.

[2] http://www.ncbi.nlm.nih.gov/geo.

[3] http://www.ncgr.org/genex/index.html.

[4] http://www.mged.org.

Existing definitions and proposed standards for gene expression data provide useful guidelines for organizing expression data in systems such as GeneExpress. Adequate standards for the representation of sample and gene annotations, however, have not yet been established. MIAME's recommended standards for gene annotation for the fragments on a microarray are minimal in order to simplify compliance. For example, the suggested annotations for probes on a microarray consist of their identity, sequence, and the associated composite target sequence, along with gene symbol or reference to a model organism database. However, in depth gene expression data analysis requires access to functional characteristics of these target gene fragments, in order to interpret data analysis results.

Similarly, MIAME's minimum required sample annotations are not sufficient to establish the context needed for comprehensive gene expression data analysis. Clinical history, morphology and pathology for samples are needed to interpret gene expression data. For example, it is necessary to know the precise stage of a tumor or medications taken during acquisition of a cancerous sample in order to interpret expression measurements for the sample.

For sample data, standardization involves establishing controlled vocabularies of terms for specific data domains, such as SNOMED [19] for anatomy or diseases. These efforts are usually sponsored by professional organizations within a specific field, (e.g. SNOMED is supported by the College of American Pathologists) and are not easily accessible to academic organizations because of their associated cost.

For gene annotations, the most notable standardization effort is the development of the Gene Ontology by the GO Consortium [5]. The goal of GO is to provide a dynamic controlled vocabulary to describe the role of genes and gene products in terms of molecular function, biological process and cellular components.

Data exchange formats or standards emphasize the syntactic aspects of expression data, and, to a lesser degree, the meaning of the data, in cases where the representation is well documented. However, these formats do not address the semantic issues regarding the comparability (or compatibility) of gene expression data. Data comparability is a pre-requisite for analyzing expression data from multiple experiments or multiple sites together, and is discussed below in Section 9.3.

9.2. The GeneExpress System

Gene Logic's GeneExpress system provides support for managing expression data generated using the Affymetrix GeneChip platform in a high throughput production environment. Sample, gene annotation and gene expression data are collected from separate data sources: sample data are collected and managed using a sample data management system; gene annotations are acquired from a variety of public and private genome databases and integrated into a gene annotation database; and the main source for gene expression data is the Affymetrix GeneChip LIMS database. GeneExpress was built using data warehousing and on-line analytical processing (OLAP) concepts adapted to the gene expression data domain [15].

9.2.1. GeneExpress System Components

The GeneExpress data store consists of the GeneExpress Data Warehouse (GXDW). GXDW consists of component databases containing sample, gene annotation and gene expression data, and process information specific to the generation and analysis of the expression data [16].

The gene expression data in GXDW is represented by a three-dimensional array, with expression values indexed by gene fragments (identified by their target sequence and the microarray type), samples and algorithm or measurement type. This data structure is implemented by the Gene Expression Array (GXA) as a collection of matrices, each associated with a particular GeneChip probe array type (e.g. HG_U95A) and measurement type (e.g. a version of the MAS algorithm). Each matrix has axes representing samples and gene fragments. The GXA provides a basis for the GeneExpress *Analysis Engine*, which implements various analysis methods in a highly efficient manner.

The GXDW, GXA and Analysis Engine applications reside on a GeneExpress server. The server also hosts the *Workspace File System*, which allows users to store analysis results and share them throughout an organization.

Data in GXDW can be accessed using the *GeneExpress Explorer* application, which provides support for specifying gene and sample sets of interest and for analyzing gene expression data in the context of such gene and sample sets using a variety of analysis tools. GeneExpress Explorer is implemented as a client-side Java application, which runs on desktops, and accesses GXDW through JDBC and the analysis server through a CORBA layer. The main components and architecture of the GeneExpress system are illustrated in Figure 9.1. The results of gene expression analysis can be examined in the context of gene annotations, such as pathways, and can be exported to third-party tools, such as Spotifre, GeneSpring or Partek, for visualization or further analysis. The gene expression and associated data can also be accessed directly through APIs, which are available for a number of popular programming languages and platforms.

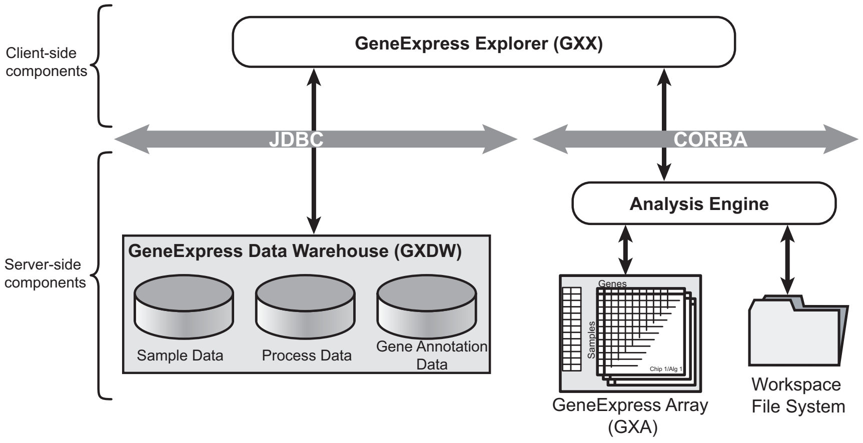

Figure 9.1: GeneExpress System Architecture

9.2.2. GeneExpress Deployment and Update Issues

In most cases, a GeneExpress system for a particular customer resides on a dedicated server. These machines are either deployed at the customer site and connected to the customer's internal network, or they are located at Gene Logic, and accessed via a VPN mechanism. The data content of each GeneExpress system, involving both GXDW and the GXA matrices, is updated on a regular schedule (e.g., bi-monthly or quarterly).

The sample, process and gene expression data components of GXDW are built by extracting the data for the relevant samples from a master *production* version of GXDW which is maintained at Gene Logic. The subset of samples provided to each GeneExpress customer is determined by the specific GeneExpress product license for the customer, and will usually contain new samples which have been processed by Gene Logic since the last content update. The sample, process and gene expression portions of the production GXDW are maintained in an incremental fashion, with new samples and experiments being added as they become available. Similarly, the set of GXA matrices for a particular customer is built by extracting the portions of the internal production GXA matrices which pertain to the samples being supplied to the customer.

The update mechanism for the gene annotation data component of GXDW is somewhat different. In order to keep abreast of current genomic data available in the public domain, it is necessary to refresh the gene annotation database periodically. The static portion of the data, such as gene fragments and array design, will not change unless new arrays are introduced. However, links to genes and all the public genomic objects may change to reflect new versions of their data sources. Because of the complex interdependencies of the various genomic data sources, and the fact that many such data sources do not provide incremental updates, it is not feasible to update the gene annotation database in an incremental fashion. Instead, it must be completely reloaded each time it is refreshed. This process is usually performed on a quarterly basis because of the high overhead involved.

9.3. Managing Gene Expression Data: Integration Challenges

In this section we present some of the key challenges that arise from the management of gene expression and related data, and briefly describe how each of these challenges are addressed in the GeneExpress system. Many of these challenges involve resolving semantic conflicts in gene expression, sample and gene annotation data in order to integrate these data in a gene expression data management system. First, we discuss the data management problems caused by differences in microarray versions, differences in algorithms and normalizations, and non-biological variability in expression data. Next, we discuss challenges regarding sample data and gene annotation data.

9.3.1. Gene Expression Data: Array Versions

Microarray platforms keep evolving with new probe array versions benefiting from technological improvements (e.g., higher density arrays and better probe selection) and advances in deciphering the genome. For example, Affymetrix recently released the HG-U133 series of the human probe arrays which replaced the previous HG-U95 series of arrays. Running the same or similar samples on two series of probe arrays doubles the amount of data generated, but, in many cases, this is necessary because the newer arrays may produce expression data for target transcript sequences that are not available on the previous versions. In addition, there may be multiple versions of a probe array within a particular array series, if problems are discovered with the a particular array (e.g. HG-U95A versions 1 and 2 within the HG-U95 series). Comparing data generated using different series of probe arrays entails addressing a very complex semantic data integration problem, with gene annotation data providing only partial support for resolving this problem.

In general, data generated using different probe array series or versions are not comparable, nor can they be transformed in order to make them comparable. This is in part due to the selection of target genes and ESTs for new probe arrays, which are often based on newly published biological information. Furthermore, representative probes for the target genes on the new probe arrays may be different due to availability of better representative sequences or improved techniques for choosing oligos within a representative sequence. New probe arrays may also be associated with improved analysis algorithms for determining summary intensity values, which will not be directly comparable with older algorithms. Consequently, in order to allow comparison of gene expression data generated for new samples using new probe arrays with data for existing samples, it is necessary to rerun the existing samples using the new probe array versions and algorithms.

On the other hand, it is often still valuable to maintain data generated using older probe arrays, since they may provide the basis for existing analyses or prediction models, which users do not wish to re-create, or because sample material may no longer be available for rerunning the experiments using new arrays, or samples may no longer be considered important enough to warrant rerunning them using new arrays. Further, older probe arrays may include fragments of interest that have been omitted or do not have a good representation on the newer arrays.

GeneExpress supports multiple probe array sets for each species, and allows users to choose a probe array set in addition to a species when performing analyses. Annotations associating homologous or related gene fragments on different versions of a probe array are provided in the gene annotation database of GXDW, and can be used to map probe sets on a given probe array to probe sets on another version of the probe array. Direct comparisons of gene expression data based on different probe arrays are not supported.

The amount of data generated with multiple probe array versions is kept manageable, in part, because GXDW and GXA contain only the estimates for expression measures and gene-level summary data. Images and probe intensity files are archived on an enterprise network accessed storage system, and are not incorporated into standard GeneExpress systems. When a new algorithm or a new probe array version needs to be supported within GeneExprees, the information describing the probe array design must be entered in the gene expression data space, and a new matrix included in the GXA.

9.3.2. Gene Expression Data: Algorithms and Normalization

Different algorithms can be applied to generate gene expression data at different levels, including image, probe level and gene expression estimate data. For example, several alternative methods have recently been developed to estimate expression measures from probe data (see [11, 12]), in addition to Affymetrix' GeneChip MAS algorithms. For GeneChip, the MAS 5.0 algorithm has recently replaced the MAS 4.0 algorithm, and is required for analyzing the data generated with the newer versions of probe arrays. In order to take advantage of a new or alternative algorithm, it is necessary to re-analyze raw or probe data and generate new estimates of gene expression. It is important to note that expression estimates generated by different algorithms are not directly comparable. Furthermore, some algorithms depend on certain parameters which may also affect the generated expression estimates.

In GeneExpress a number of factors are recorded that may determine the comparability of expression data, including:

1. the algorithms employed to generate expression estimates, namely MAS 4.0 (employed for all probe arrays through the end of 2001) or MAS 5.0 (required for the new HG-U133 probe arrays, and optional for other probe arrays)[5]; data generated using different algorithms are not comparable;
2. scaling factors used to reduce discrepancies caused by sample preparation or probe array lot variability; data generated using different scaling factors are transformed to a common factor using straightforward multiplication;
3. normalizations, which may be applied to the values generated by the MAS or other algorithms: GeneExpress provides support for several normalization methods including Standard Curve Normalization, based on using spike-ins of known concentrations for certain (bacterial) genes when preparing samples for experiments [9]; data must be generated using the same normalization in order to be comparable.

The Gene Expression analysis software, GeneExpress Explorer, ensures that data analyzed together have been generated using the same algorithms and normalization methods.

9.3.3. Gene Expression Data: Variability

Determining if gene expression data from two or more sources, such as different organizations or different sites within an organization, are comparable involves assessing non-biological differences that may affect analysis results. While gene-to-gene differences and sample-to-sample differences will be present in any set of experimental data, it is important to determine if there are other significant sources of variability. Many factors may contribute to such variability, including differences in the processes for obtaining and storing samples, differences in experimental practices and techniques, differences in adjustment of equipment, such as scanners, and so on.

[5] See http://www.affymetrix.com/products/algorithms_tech.html.

Statistical methods are used to identify the magnitude and qualitative nature of non-biological variability. Initial exploration ideally involves samples collected from the same type of tissue (that is, from the same type of organ and a similar location in the organ) and with the same pathology. In this case, one can assess data comparability using the entire set of genes involved in the experiments. If samples are from the same type of tissue but with different pathologies, then data comparability can be assessed using only genes that are not likely to be involved in the biological difference between the two groups of samples.

Exploratory statistical techniques employed for assessing the comparability of such samples include univariate (single experiment) and bivariate (pairs of experiments) analyses. One simple way to compare numerous univariate distributions is by displaying boxplots of the distributions side by side [10]. Such boxplots would allow one to determine whether there are significant effects due to, for example, scaling or saturation, which would result in a shift in the distribution of expression values. Further exploration would involve assessing the reproducibility of expression values between experiments, and the variability of expression values within each group of experiments and between groups of experiments.

Gene Logic limits non-biological sources of variability in the gene expression data it generates by following strictly controlled procedures and by monitoring the quality control measures, both for running experiments and for the collection and preparation of samples. Once data are generated from experiments, quality control procedures based on statistical methods are used to ensure that data included in GeneExpress is not unduly affected by non-biological factors.

9.3.4. Sample Data

Accurate and consistent characterization of samples is essential in dealing with gene expression data, since errors can have a substantial effect on expression analysis. It is not sufficient to base sample classification solely on annotations provided by the supplier since, first, samples may be mis-labeled (e.g., a diseased tissue being labeled as normal), and, second, there may be inconsistencies of classification due to the perspective of the pathologist or scientist who did the initial labeling. In the GeneExpress system sample classification validation involves a careful review of the micro-section images by a pathologist, as well as a thorough review of the clinical information accompanying each sample. Using the Systematized Nomenclature of Medicine (SNOMED) [19], the sample can be further characterized by topography, morphology, disease and disease stage. The use of SNOMED and other controlled vocabularies in the GeneExpress system leads to a more robust classification of samples and provides a consistent representation of the data to users. However, even with an established controlled vocabulary such as SNOMED, the choice of terms to characterize a tissue type or disease may be ambiguous, and so Gene Logic's pathologists use a consistent system of rules to determine which SNOMED terms to use.

9.3.5. Gene Annotations

Associating gene fragments with annotations from various public and private data sources provides the genomic context for interpreting gene expression data. Integrating such annotations into a data warehouse, as opposed to accessing the remote data sources through a federated database approach (see, for example, [7]), allows better representation of the semantics, powerful query expression, improved query performance and also allows the quality of the data to be checked during the integration process (a similar conclusion is reached in [6]). 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, such as the correlation of gene fragments and known genes.

Gene annotation or gene index databases are generally based on data collected from well-established and reliable public data sources. For example, gene fragments can be organized in non-redundant classes based on UniGene[6], and associated with known genes recorded in LocusLink[7]. However, such data sources may not contain genomic information for all species: some may provide very good human and mouse gene annotations, but not cover other species such as yeast or rat. In such cases, it is necessary either to find alternative data sources, or to derive gene annotations for these species by finding homologous genes on better annotated species, such as human or mouse. The choice of which approach to use may change from time to time depending on the availability of annotations.

Gene fragments are further associated with gene products (e.g., protein data from SwissProt[8]), GO ontology terms, enzymes, metabolic and signaling pahways, chromosome maps, genomic contigs and cross-species gene homologies. For genomic information such as pathways, there is no unique data source that satisfies all needs. For example, KEGG provides very good metabolic pathways, but is not complete, and other public or private pathway data sources provide valuable additional data. Integration of similar or potentially overlapping data from two or more data sources requires the potential problems of redundant and inconsistent data to be addressed.

Genomic data sources are usually updated on different schedules, and the size of such data sources usually prohibits all versions of a data source from being loaded into a data warehouse. The gene annotation component of the GeneExpress data warehouse contains over five gigabytes of data with only the most current version of data collected from various data sources. However, storing data from only one version of a data source may lead to inconsistencies, with one source referencing entities in a different version of another data source, which may have been updated or no longer exist. Further, data sources may change their data structure or schema between versions, for example adding, removing or modifying attributes or fields. In

[6] UniGene provides a partitioning of GenBank nucleotide sequences in gene-oriented non-redundant clusters. See http://www.ncbi.nlm.nih.gov/UniGene/.

[7] LocusLink provides curated sequence and descriptive information, such as official nomenclature, on genes. See http://www.ncbi.nlm.nih.gov/LocusLink/index.html.

[8] SwissProt is a curated protein sequence database that includes the description of the function of a protein. See http://www.expasy.ch/sprot/.

addition, keywords can be changed, and data files can be re-organized. Such changes necessitate revisions of data collection tools and reconciliation of data mappings.

The gene annotation component of GXDW provides an integrated view of the genomic data space, based on a unified schema that spans the various object spaces relevant to each of the public or private data sources used. One key feature of the schema is that it models the primary objects from the genomic data space in a generic way, though such objects originate from a wide variety of data sources. This minimizes the frequency of schema changes needed even as the structures of the primary data sources evolve.

In order to keep up-to-date with the evolving gene annotation data sources, the gene annotation component of GXDW is refreshed periodically. Each refresh involves extracting data from the latest versions of more than a dozen relevant public and private data sources, including UniGene, LocusLink, SwissProt, OMIM, Enzyme, GO, KEGG and proprietary pathway databases, along with model organism genome databases for organisms such as E.coli, and yeast. During the integration and the assembly process, various data transformations and data cleansing operations are performed in order to resolve conflicts and correct data errors. Due to the rapidly evolving nature of these data sources, their content may change, both syntactically and semantically, between refreshes. Consequently, establishing cross database links often requires manual curation to deal with orphans and links to retired entries. For example, LocusLink may refer to an enzyme EC number that is obsolete in the Enzyme catalog database, in which case it will be necessary to identify the correct current EC number and update the data sources

The data-warehousing strategy employed for constructing and maintaining GXDW supports various derived annotations such as cross-species homology relations between genes of different organisms and other objects. This is particularly valuable for carrying out comparative expression analysis between model organisms.

Due to the rapidly changing nature of the gene annotation data and data sources, it is important to continually search for new sources of gene annotation data, and re-evaluate existing data sources. When a new data source is considered for GeneExpress, decisions must be made regarding whether the new data source can or will replace any existing data source, whether existing curation methods must be modified, whether the data model or schema needs to be revised, and how existing data should be associated with data from the new source.

9.4. Integrating Third-Party Gene Expression Data in GeneExpress

The GeneExpress system was originally developed for the purpose of managing, exploring and analyzing gene expression data generated at Gene Logic, primarily using the Affymetrix GeneChip platform. However, as the system has been adopted by various customers, some of which have their own internal efforts to generate gene expression data, the need to integrate customer data into the GXDW in order to enable analysis of Gene Logic and customer gene expression data together, has become apparent.

In order to support the integration of customer sample and gene expression data into GeneExpress, the *GX Connect* tool has been developed at Gene Logic. GX Connect™ supports integration of gene expression data residing in an AADM-based

GeneChip LIMS database, and sample data conforming to the Gene Express Sample Data Exchange Format, into GXDW. When there is a need to integrate gene annotation data[9], gene expression data represented using alternative formats, or data which, for other reasons, cannot be integrated using GX Connect, custom data integration tools are developed.

In this section we discuss some of the challenges involved in integrating customer gene expression data with Gene Logic data, and how these challenges have been addressed in the context of GeneExpress. First, we describe data exchange formats which simplify the tasks of developing and maintaining mappings of customer data to GXDW. Next, we describe some of the structural and semantic data transformation issues that are involved in developing such mappings. Finally, we describe some of the data management issues associated with data loading and updating the Gene Logic content of a system containing both Gene Logic and customer data.

9.4.1. Data Exchange Formats

In order to avoid developing and maintaining multiple data migration and loading tools for each external data source considered for integration, *data exchange formats* serve as intermediate representations for data being transferred from various data sources to the GeneExpress data warehouse. The process of integrating external data is then divided into two phases: first, structural transformations and mappings need to be applied to the external data, in order to convert them into the data exchange formats; second, the data in the data exchange formats need to be loaded into the warehouse. Note that developing and maintaining tools that convert data from sources into a well-defined data format, such as one based on XML or similar notation, is generally easier than developing tools to transform data and populate a target data warehouse.

A number of formats have been proposed for gene expression data, as mentioned in Section 9.1.2. Since we have focused so far on integrating Affymetrix GeneChip expression data into GeneExpress, we use a subset of the Affymetrix Analysis Data Model (AADM) [1] as the data exchange format for gene expression data. In this format, expression data are associated with samples, gene fragments, analysis methods and various experimental parameters.

For sample and clinical data, standard formats such as AADM have not yet been established. Consequently, we have defined data exchange formats that satisfy our requirements.

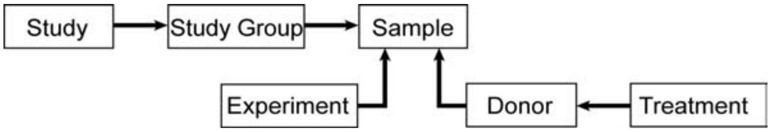

Figure 9.2: Sample Data Exchange Format

[9] Data exchange formats and integration tools for gene annotation data are planned for future versions of GX Connect.

The central object class of the Sample Data Exchange Format is *sample*, representing the biological materials (e.g. tissue or cell-line) that are investigated using probe arrays (see Figure 9.2). Attributes associated with samples may describe their structural and morphological characteristics (e.g. organ site, diagnosis, disease, stage of disease). A sample is associated with a *donor* (e.g. a human or an animal model), which may in turn be qualified by various *treatments*, and has additional attributes (e.g. clinical records and demographics for human donors, or strain and genetic modification for animal donors). Each sample may be associated with several experiments (e.g. using different chip types). Samples may be grouped into *studies*, which may be further subdivided into *study groups* based on time or treatment parameters.

Various classes in the sample data exchange format include "*catch-all*" attributes which can accommodate any data, represented as tagged-value pairs, that does not otherwise fit the format.

For data represented in the data exchange formats described above, the *GX Connect* tool can be used to control and automate the process of data transfer into the GeneExpress warehouse [4]. This tool can be deployed at customer sites and used to perform incremental (e.g. nightly) updates to GXDW. Consequently the main task associated with integrating customer data becomes defining and implementing the semantic and structural transformations necessary to convert customer data into the data exchange formats, in order to prepare it for loading into GXDW.

9.4.2. Structural Data Transformation Issues

Data from individual data sources may be supplied in a flattened or un-normalized form, such as Excel spreadsheets, so that determining their structure and how to map them to the various data exchange formats is often a complex and involved task. First, it is necessary to determine the dependencies and correlations between individual data objects, which may be provided during the data export process, or need to be determined by searching for patterns in the data. In either case, it is necessary to confirm that the correlations that have been found are consistent with the intended semantics of the data.

Data dependencies and correlations can be used to form an object model for the source data, and to define a mapping from this model to the data exchange formats. Defining such a mapping requires structural conflicts between the models to be resolved, and, in some cases, it may be necessary to choose between several possible solutions.

For example, the GeneExpress Sample Data Exchange Format classifies samples in a two-level hierarchy, with the levels represented by the classes *Study* and *Study-Group*. Sample data exported from an external data source might employ a three level hierarchy, such as *Project*, *Study* and *Treatment*. There are two possible ways to resolve such a difference in structure: either combining the exported *Study* and *Treatment* classes into the Sample Data Exchange Format *Study-Group* class and mapping the exported *Project* class to Sample Data Exchange Format *Study* class, or by mapping the exported *Project* and *Study* classes to Sample Data Exchange Format *Study* class, and the *Treatment* class to the *Study-Group* class.

In addition, it is necessary to deal with the evolution of databases and formats over time. Both the external data sources and the GeneExpress Data Warehouse may change either their structure or their controlled vocabularies or data formats, in order to reflect changes in requirements. These changes require updates to the mappings.

9.4.3. Semantic Data Mapping Issues

For gene expression data, the semantic challenges of integrating data from multiple sources are similar to those described in section 9.3. Experiment data from different platforms are generally not comparable. Even if experiments are from the same platform, expression values may have to be adjusted (e.g., to compensate for different scanner settings) before they can be compared. Moreover, expression data will not be comparable unless they are analyzed using the same version of a probe array and the same algorithm

The mappings for sample data are usually the most difficult since there is no widely accepted standard for representing clinical data (see [20], presentations at the working group on ontologies). We discuss below some of the problems of mapping sample and gene annotation data.

Sample Data Mapping: Studies

Expression data are often organized into studies. For Gene Logic data, *studies* are used to group data which address specific questions about the effects of certain variables, such as treatment conditions, disease stage, time and so on, on gene expression levels. Studies may be further divided into *study groups*, which represent samples grouped according to certain attributes, such as specific treatment conditions, time points or disease stages.

The structure and nature of a study performed outside of Gene Logic may be conceptually different from studies defined in the context of GeneExpress data. In order to group customer samples into studies or study groups, it is necessary to identify an equivalent structure in the source sample data model, which may use different terminology or organize data along different principles. In the case that there is no appropriate concept in the source data model, then rules can be incorporated into the mapping from the source data model into the sample data exchange format, allowing studies and study groups to be created based on other source data attributes, such as tissue-type or treatment. Alternatively, customer data can be organized into studies and study groups manually by editing the data once it has been converted to the sample data exchange format.

Sample Data Mapping: Nomenclature

In order to map individual sample data values to the Sample Data Exchange Format, it is first necessary to resolve differences of nomenclature, units and formatting. Differences in nomenclature are the most difficult to deal with, and often there is no single, optimal resolution for such differences. Various attributes in the data exchange formats are represented using controlled vocabularies. In particular, in the Sample Data Exchange Format, sample organ types, pathologies and disease diagnoses are represented using subsets of the SNOMED vocabulary [19].

External sample data repositories often use their own vocabularies for such concepts, and, even within a given standard such as SNOMED, different pathologists or other experts may not agree on which term should be used for a certain disease or organ type. For example, in a recent integration project, a customer included samples with the diagnosis labeled *DIABETES*. The SNOMED vocabulary includes several varieties of diabetes and related complications, so it was necessary to consult with the customer in order to determine the best choice of mapping. After some discussion it was determined that, given the differences of interpretations, the best choice was to map this to the term *OBESITY* in GeneExpress. Similarly the customer data might include abbreviations, such as *DRG*, which was mapped to *DORSAL ROOT GANGLION*, or common terms, such as *FAT*, which was mapped to *ADIPOSE TISSUE*. Moreover, a SNOMED term code is usually associated with one primary term together with one or more synonyms. Some customers may prefer a different synonym than the one chosen by Gene Logic.

Sample data may also differ in the choice of units: for example drug treatments can use units such as μMol or ng/ml, while age can be provided in days, weeks or years. A conversion table is required to map any units to comparable units in the Sample Data Exchange Format.

Formatting of individual items also needs to be resolved. For example the Sample Data Exchange Format uses the terms *Male* and *Female* to represent the sex of a donor, while a customer database may use *male* and *female*, or just *M* and *F*. Further, data may contain typographic errors, such as misspelling the name of a supplier. When vocabularies are small, or for controlled vocabularies, it may be possible to spot and correct such errors manually, but in general, these errors can go undetected.

All these different types of conflicts need to be resolved as part of the mapping from the source data format to the sample data exchange format. In some cases, it is not possible to implement rules to resolve such conflicts automatically, and so it necessary to perform manual inspection and curation of the data before mapping it. In general, if the source data is *consistent* in its use of controlled vocabularies, formatting and units, then it is possible to "hardwire" the correct mappings into the mapping implementation. However, whenever a new conflict arises, it is necessary to find a resolution and adapt the mapping implementaition.

It is worth noting that sample mapping provides consistency between Gene Logic and customer sample classifications rather than finding an optimal classification. Sample classification in GeneExpress is based on sound clinical and pathology principles in the strict framework of the SNOMED nomenclature. However, not all medical concepts map straightforwardly to SNOMED terms, and therefore there may not be a "best" classification for a concept, but rather several reasonable ones.

Gene Annotation Data Mapping

In general, gene annotations are not involved in the integration of expression data from multiple sources. In certain cases, however, it is necessary to integrate gene annotations associated with non-Gene Logic expression data, for example in order to extend the system to include custom probe arrays with proprietary gene fragments, or to support a customer's proprietary gene annotation data.

Gene annotations generally have well understood semantics, although there are ambiguities with regard to the classification of some of these annotations (see [18] for a discussion of problems associated with gene nomenclature and identification).

Since gene annotations are often stored in proprietary databases, a possible approach is to provide links to these annotations, instead of importing them into GXDW. This approach does not support the ability to query the contents of these databases directly, nor superimpose expression data on these annotations (e.g., superimpose expression levels associated with genes displayed on a pathway or chromosome map), but can make the information readily accessible from Gene Express. In such cases individual genes within GeneExpress are linked to network accessible reports or interactive services. When query access is required, custom gene annotations can be integrated into GeneExpress using a similar mechanism to that used for sample data. Defining a mapping remains non-trivial, but, as gene annotation data is often more rigorously structured than clinical information, the problem is usually less severe.

Another problem specific to gene annotations is the fact that related but different annotations are likely to reside in multiple sources. This introduces a key challenge: reconciling differences between different gene annotation sources. When different versions of a single source (e.g. UniGene) conflict, it is usually acceptable to defer to the newer version. When different sources conflict there may not be an ideal way of resolving the differences.

In addition, a customer may prefer alternative sources for gene annotation data (e.g., protein data sources other than SwissProt, or sequence clusters other than those provided in UniGene), rather than those used in GeneExpress. Even when the same data sources are used, different refresh policies may lead to the use of different versions or different builds of the same data source. Furthermore, there may be multiple ways to associate two related biological objects (e.g., links from gene fragments to known gene clusters may be based on data supplied by the probe array manufacturer, or may be based on homology searches using the fragment's target sequence). Consequently, integrating customer gene annotations with Gene Logic gene annotations requires resolving potentially complex data discrepancies.

9.4.4. Data Loading Issues

Once data from external data sources have been mapped to the data exchange formats, additional processing and curation may be required before integrating and loading them into the warehouse. First, it is necessary to detect invalid data, such as missing clinical data associated with samples or inconsistent associations of sample and gene expression data. In general, data migration tools, such as GX Connect, handle such cases by skipping the data affected by errors, and issuing warning messages in a log file. Data editing can be used in order to correct problems that have not been resolved during the mapping process.

Next, it is necessary to resolve differences between identifiers of external objects and objects already in the warehouse in order to maintain database consistency. Transformations of this type are carried out using *staging databases*, before loading data into the warehouse itself. In addition it is necessary to keep track of any identifiers created for customer data, so that if customer data objects are dropped and reloaded, for example to allow the data to be edited, they do not reappear with different identifiers.

Finally *derived data*, such as quality control data (e.g., measures of saturation for the scanners) are also computed during the final loading stage.

9.4.5. Update Issues

In Section 9.2.2 we described the process of updating a GeneExpress system containing only Gene Logic data. The content update becomes more complex if the system contains both Gene Logic data and customer data, either loaded with the GX Connect tool or with custom tools. Since both Gene Logic data and customer data change over time, content update procedures must ensure that new data from both sources are maintained correctly in the GeneExpress data warehouse.

Data in the GXDW can be classified into: (1) data *shared* by Gene Logic and customer data stores, such as controlled vocabularies; and (2) data that are not shared, that is data generated by either Gene Logic or the customer only. Examples of shared data include SNOMED terms and species information in the sample database, and probe array types and algorithm types in the expression database. When performing a content update, shared data occurring in both the Gene Logic and the customer data contents must be consolidated. Examples of data that are not shared include data pertaining to an individual sample in the sample database and experiment expression values in the expression database. Since customer sample and experiment objects are always distinct from Gene Logic sample and experiment objects, it is not necessary to merge these data, and, instead, separate spaces of object identities are maintained for customer and Gene Logic data.

Depending on the nature of the data, a variety of techniques can be used for handling updates. For example, since it is not necessary to merge data for individual experiments or samples, such as expression values, from different data sources, these data can reside in different database partitions. In this case, content update is as simple as replacing a database partition. On the other hand, controlled vocabularies and other shared data, must be consolidated, and therefore special tools are required to reconcile terms in customer and Gene Logic data, and to make sure that they are consistent in the integrated warehouse (e.g., having the same ID values). The consolidation process involves resolving the identification of objects and terms, as well as object references.

9.5. Summary

We have discussed data integration challenges involved in building a system for managing gene expression data, and how these challenges have been addressed in the GeneExpress system and in the context of several GeneExpress integration projects.

A data warehouse approach and tools were used in developing GeneExpress, and were found to provide an effective environment for developing a system that needs to support the integration and management of data from diverse sources, where data may be imprecise and may evolve over time. Other, non-warehouse (i.e., non materialized view) approaches were also briefly considered, based on our previous experience with developing genomic data management systems using the OPM tools [17], but were not adopted for reasons similar to those described in [6]. The data warehouse approach has proven to be well suited for systems such as GeneExpress, which need to integrate data from multiple data sources, with data requiring validation and cleansing, and where system performance and robustness are critical. We found, however, that the general data warehouse approach cannot be applied “as is” to the gene expression domain and therefore needs to be adapted [15]. We also learned that coping with issues of data semantics in the area of genomic applications remains complex and difficult, and often requires manual solutions.

Since good performance is a critical requirement for GeneExpress, a comprehensive set of benchmarks have been devised to continuously assess system performance as its data content grows. The benchmark involves running typical queries and expression analysis operations on a series of data sets, using various configurations of Sun SparcUltra II and III based servers and Pentium-based clients. These benchmarks first measure the single-user performance of query and analysis operations, and then measures multi-user performance with up to 300 simulated concurrent users each running analysis steps across all available array types. It was found that, given sufficient server system memory, performance for multiple users scaled linearly with the number of processors and number of concurrent users.

Though this chapter has focused on the GeneExpress system and the Affymetrix GeneChip platform, the challenges addressed by the GeneExpress system are shared by other systems for managing and analyzing gene expression data. In particular, for all gene expression platforms, the problems associated with relating the data to gene and sample annotations and issues such as compatibility of array versions and analysis algorithms are similar.

The first version of GeneExpress was released in early 2000. Through the end of 2002, the GeneExpress system has evolved through several versions and has been deployed at over twenty five biotech and pharmaceutical companies worldwide, as well as at several academic institutions. Based on the experience gained in developing tools for incorporating customer data into GeneExpress, the GX Connect tool has been developed in order to provide support for interactive extraction, transformation and loading of gene expression data generated using the Affymetrix GeneChip platform and related clinical data into GeneExpress. GeneExpress and GX Connect are deployed together as part of the Genesis Enterprise System™ [4].

Five data integration systems that provide support for integrating gene expression data from both Gene Logic and customer sources have been deployed through the end of 2002. All these systems provide support for integrating sample (clinical) data based on proprietary data formats and allow regular incremental updates of customer data; two of these systems provide support for custom Affymetrix GeneChip probe arrays; and one system also provides support for proprietary gene annotations.

Acknowledgements. We want to thank our past and present colleagues at Gene Logic who have been involved in the development of the GeneExpress system for their outstanding work. Special thanks to Mike Cariaso, Francois Collin, William Craven, Michael Elashoff, Aaron Hechmer and Dmitry Krylov for their feedback and contributions to this paper.

Trademarks. GeneExpress®, GX™ and Genesis Enterprise System™ are trademarks owned by Gene Logic Inc. Affymetrix® and GeneChip® are trademarks owned by Affymetrix, Inc.

References

1. Affymetrix, *Affymetrix Analysis Data Model*, http://www.affymetrix.com/support/.
2. Affymetrix, *GeneChip®Analysis Suite User Guide*, 2000.
3. Brazma, A. et al., *Minimum Information about a Microarray Experiment (MIAME): Towards Standards for Microarray Data*, Nature Genetics, Vol. 29, No.4, pp. 365-371, 2001.
4. Gene Logic Products. http://www.genelogic.com/products.htm. See GeneExpress® Product Line and Genesis™ Enterprise Software.
5. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium, Nature Genetics, 25, pp. 25-29, 2000. See also http://www.geneontology.org
6. Davidson, S.B., Crabtree, J., Brunk, B., Schug, J., Tannen, V, Overton, C., Stoeckert, C., *K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources"*, IBM Systems Journal, Vol. 40, No.2, pp. 512-531, 2001.
7. Eckman, B.A., Kosky, A.S., and Laroco, A.L., *Extending Traditional Query-Based Integration Approaches for Functional Characterization o Post-Genomic Data*, Bioinformatics, Vol. 17, No.7, pp. 587-601, 2001.
8. Human Gene Nomenclature Database, http://www.gene.ucl.ac.uk/nomenclature/.
9. Hill, A. A., Brown, E. L., Whitley, M.Z., Tucker-Kellogg, G., Hunter, C.P., Slonim, D.K. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. Genome Biology 2 (12), 2001, http://www.genomebiology.com/2001/2/12/research/0055/.
10. Hoaglin, D.C., Mosteller, F., Tukey, J.W., *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, Inc, 1983.
11. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P., Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data, submitted for publication.
12. Li, C. and Wong, W. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Science USA 98. pp. 31-36.
13. Lockhart D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang C., Kobayashi, M., Horton, H. and Brown, E.L., Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays, Nature Biotechnology, 14, pp. 1675-1680, 1996.
14. Lockhart D.J., and Winzeler, A.E., Genomics, Gene Expression, and DNA Arrays, Nature, 405, pp. 827-836, 2000.
15. Markowitz V.M., and Topaloglou, T., Applying Data Warehousing Concepts to Gene Expression Data Management. Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, November 2001.
16. Markowitz V.M., Chen, I.A., and Kosky, A. Gene Expression Data Management: A Case Study. Proc. of the 8th International Conference on Extending Database Technology (EDBT), Lecture Notes in Computer Science, Springer Verlag, Jensen et al (eds), 2002.
17. Markowitz V.M., Chen, I.A., Kosky, A., and Szeto E. OPM: Object-Protocol Model Data Management Tools. Bioinformatics: Databases and Systems, Letovsky S.I. (ed), Kluwer Academic, Boston, 1999, pp. 187-199.
18. Pearson, H., Biology's Name Game, Nature, 417, pp. 631-632, 2001.
19. SNOMED, Systematized Nomenclature for Medicine. http://www.snomed.org/.
20. Third International Meeting on Microarray Data Standards, Annotations, Ontologies, and Databases. Presentations. http://www.mged.org/presentations/index.html.