Gene Expression Data Management: A Case Study

Victor M. Markowitz, I-Min A. Chen, Anthony Kosky

Gene Logic Inc., Data Management Systems
2001 Center Street, Berkeley, CA 94704, U.S.A.
{victor, ichen, anthony}@genelogic.com

Abstract. One of the major challenges facing scientists dealing with gene expression data is how to integrate, explore and analyze vast quantities of related data, often residing in multiple heterogeneous data repositories. In this paper we describe the problems involved in managing gene expression data and discuss how these problems have been addressed in the context of Gene Logic's GeneExpress system. The GeneExpress system provides support for the integration of gene expression, gene annotation and sample (clinical) data with various degrees of heterogeneity, and for effective exploration of these data.

1. Introduction

DNA microarray technologies allow measurement of mRNA expression levels, that is the degree to which a gene is expressed within a cell or tissue sample, for tens of thousands of genes in parallel [7]. In order to associate biological meaning with these data, it is necessary to associate them with sample data and gene annotations. In this paper we discuss the management of gene expression, sample and gene annotation data in the context of Gene Logic's GeneExpress data management system [4]. The GeneExpress system contains quantitative gene expression information on normal and diseased tissues, and on experimental animal model and cellular tissues, subject to a variety of treated and untreated conditions. Most of Gene Logic's expression data are generated using the Affymetrix GeneChip technology [6] in a high throughput production environment. The system also contains comprehensive information on samples, clinical profiles and rich gene annotations.

This paper focuses on the data integration problems encountered with GeneExpress. Data integration for molecular biology applications has gained a great deal of attention (see, for example, the papers published in [9]), and many of the problems described in this paper apply to such applications. Solutions traditionally discussed in the literature for molecular biology data integration applications involve various degrees of resolution for semantic data heterogeneity, and different types of data access mechanisms (see, for example [3] for a description of a system that supports distributed querying across heterogeneous molecular biology data sources).

Initially the GeneExpress system was developed with the goal of supporting effective exploration and analysis of gene expression data generated at Gene Logic using the Affymetrix GeneChip platform. Achieving this goal required (1) resolution of problems related to the semantic heterogeneity of gene expression, sample and gene annotation data, and (2) development of a high performance data exploration mechanism with centralized access to the integrated data.

A subsequent goal set for the GeneExpress system was to provide support for gene

expression data generated outside of Gene Logic, possibly using different technologies. Addressing this additional goal required resolution of further semantic heterogeneity problems related to the integration of gene expression data (often generated under different experimental conditions), sample data, and gene annotations, while using the same centralized data exploration mechanism for the integrated data. Both of these goals have been addressed using a data warehousing methodology [2] adapted to the special traits of the gene expression domain [8]. GeneExpress also involves a data acquisition component designed to support data content collection for gene expression, sample, and gene annotation data.

The remainder of this paper is organized as follows. In section 2 we describe the key characteristics of the data involved in a gene expression application. In section 3 we describe the GeneExpress data management system developed at Gene Logic, and the data integration tasks performed in order to build a GeneExpress product containing native Gene Logic data. In section 4 we discuss the problems of integrating gene expression data from sources outside Gene Logic into GeneExpress and describe the mechanisms that address these problems. We conclude with a brief discussion and summary in section 5.

2. The Gene Expression Application

Gene expression applications involve exploring biological sample, gene annotation and gene expression data, each of which is sufficiently complex to warrant modeling it as a separate data space [8].

Biological samples may be collected from a variety of sources, such as hospitals or laboratories. The main object in the sample data space is the *sample* representing a biological material that is involved in an experiment. A sample can be of tissue, cell or processed RNA type, and originates from a donor of a given species (e.g., human, mouse, rat). Attributes associated with samples describe their structural and morphological characteristics (e.g., organ site, diagnosis, disease, stage of disease), and their donor data (e.g., demographic and clinical record for human donors, or strain, genetic modification and treatment information for animal donors). Samples may also be grouped into *studies* which may be further subdivided, based on their time/treatment parameters, in order to form *study groups*. The core of the sample data structure is illustrated in Figure 1.

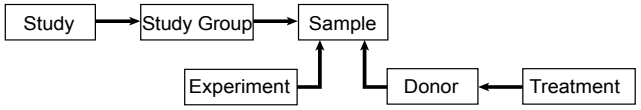

Figure 1: Sample Data Structure

Gene expression data may be generated using a variety of technologies such as different microarray platforms, and employing different algorithms for interpreting these data. For microarray data, the main object in the gene expression data space is the estimated *expression value* for a given gene or EST fragment and sample, generated using a specific microarray experiment. Data generated from microarray experiments range from raw data, such as images generated by scanners, to analyzed

data, such as quantitative gene expression measurements derived using various analysis methods.

For example, each Affymetrix GeneChip probe array contains a number of *probes* designed to bind to a particular sequence occurring within a known target mRNA fragment, which, in turn, is representative of a gene or EST of interest. Affymetrix' analysis software derives intensities for each probe by averaging the values for pixels corresponding to an individual probe in the scanned image, and then applies *analysis methods* to derive summary data for each target gene or EST fragment. The summary data generated includes expression present/absent (P/A) calls and quantitative gene expression measurements. Different analysis methods may also be applied in order to find alternative expression measurements.

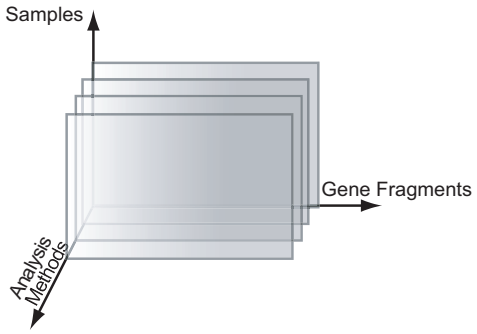

Figure 2: Gene Expression multi-dimensional array

The gene expression data may be represented by a three-dimensional matrix, with two axes representing gene fragments (identified by their target sequence and the microarray type) and samples, and a third axis representing different analysis methods (see Figure 2).

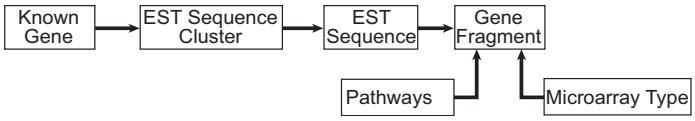

Figure 3: Gene Annotation Data Structure

Gene annotations are collected from various public and proprietary genomic resources. The main object in the gene annotation data space is the *gene fragment*, representing an entity for which the expression level is being determined as described above. 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 [5]), functional characterization, and known metabolic pathways. The core

of the gene annotation data structure is illustrated in Figure 3.

3. Primary Gene Expression Data Integration

Gene Logic's expression data are generated mainly using the Affymetrix GeneChip platform in a high throughput production environment, and are managed using the GeneExpress system. This system includes a data acquisition system for sample, gene annotation and gene expression data, a data warehouse providing support for data exploration, and data transfer and integration mechanisms for migrating data from the data acquisition system into the data warehouse (see Figure 4).

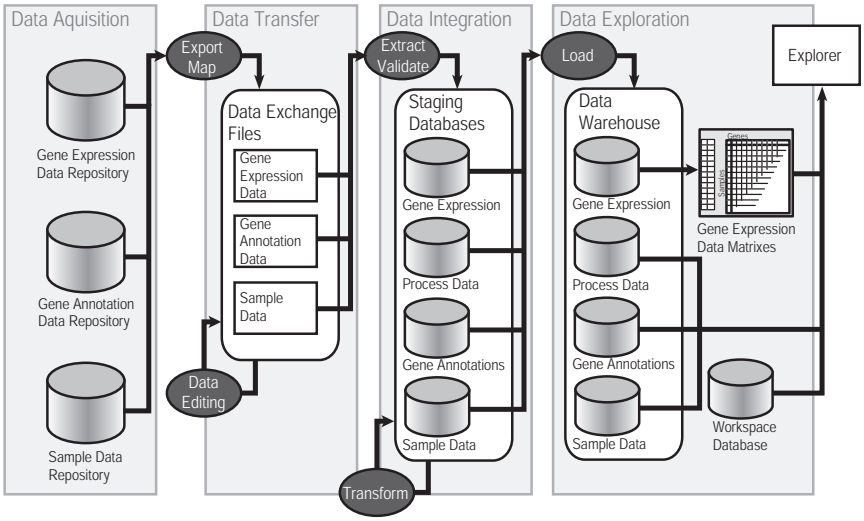

Figure 4: The main components of the GeneExpress Data Management System

The GeneExpress Data Acquisition System (GXDAS) consists of operational databases and laboratory information management system (LIMS) applications employed for data acquisition. Data management tools are used for ensuring data consistency, including checking the correct execution of the data migration and integration processes, and domain specific rules validating the sample, expression and gene annotation data.

Samples are collected from a variety of sources, such as hospitals and research centers, with sample information structured and encoded in heterogeneous formats. Format differences range from the type of data being captured to different controlled vocabularies used in order to represent anatomy, diagnoses, and medication. The *sample* component of GXDAS provides support for various sample data collection and quality control protocols, via data entry and reporting tools. This system uses domain specific vocabularies and taxonomies, such as SNOMED [11], to ensure consistency during data collection.

The *gene expression* component of GXDAS provides support for GeneChip-based production. Gene expression experiment data are recorded in GeneChip specific files

containing: (i) the binary image of a scanned probe array; (ii) average intensities for the probes on the probe array; and (iii) expression values of gene fragments tested in the probe array. The GeneChip LIMS provides support for loading the data in these files into a relational database based on Affymetrix AADM representation [1], which is the main source for gene expression data.

The gene annotation component of GXDAS provides support for assembling consistent annotations for the gene fragments underlying gene expression experiments, by acquiring, integrating, and curating data from various, mainly public, data sources. Acquiring gene annotations from public 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 fragments are organized in non-redundant classes based on UniGene, which provides a partitioning of GenBank nucleotide sequences in gene-oriented non-redundant clusters[1], and are associated with known genes recorded in LocusLink, which provides curated sequence and descriptive information, such as official nomenclature, on genes[2]. Gene fragments are further associated with gene products (e.g., from SwissProt, a curated protein sequence database that includes the description of the function of a protein, etc.[3]), enzymes, pathways (e.g., metabolic, signaling pathways), chromosome maps, genomic contigs and cross-species gene homologies.

The GeneExpress Data Warehouse (GXDW) consists of integrated component databases containing sample, gene annotation and gene expression data, and information on the analysis methods employed for generating these data.

The gene expression data in GXDW is exported to a Gene Expression Array (GXA) that implements the multi-dimensional array shown in Figure 2, as a collection of two-dimensional matrices, with axes representing samples and gene fragments. Each matrix is associated with a particular GeneChip probe array type (e.g. HG_U95), and a particular analysis method (e.g. Affymetrix MAS4 analysis). The GXA provides the basis for efficient implementation of various analysis methods.

GXDW data is explored using the *GeneExpress Explorer* application, which provides support for constructing gene and sample sets, for analyzing gene expression data in the context of gene and sample sets, and for managing individual or group analysis workspaces. The results of gene expression exploration can be further examined in the context of gene annotations, such as metabolic pathways and chromosome cytogenetic maps. The results of gene data exploration can also be exported to third-party tools for visualization or further analysis.

The GeneExpress system also contains mechanisms for migrating sample, gene annotation and gene expression data from GXDAS to GXDW (see Figure 4). These mechanism employ staging databases, and support mapping, validation, integration and loading of sample, gene annotation and gene expression data into GXDW.

4. Secondary Gene Expression Data Integration

GeneExpress is generally provided to customers with gene expression data generated

[1] See http://www.ncbi.nlm.nih.gov/UniGene/
[2] See http://www.ncbi.nlm.nih.gov/LocusLink/index.html
[3] See http://www.expasy.ch/sprot/

at Gene Logic based on samples acquired by Gene Logic. Some Gene Logic customers also have their own internal efforts to generate gene expression data. Incorporating these data into GeneExpress allows Gene Logic and customer data to be analyzed together using the same tools.

The process of incorporating customer and other external data into GeneExpress requires data migration mechanisms similar to those described in section 3 for Gene Logic's internal data sources. However, data migration from external sources needs to deal with degrees of heterogeneity that are not known ahead of time and therefore are hard to predict and support using generic tools. We describe below mechanisms that allow data from external data sources to be incorporated into the GeneExpress Data Warehouse (GXDW).

4.1 Data Export and Mapping

In order to allow data from multiple data sources to be integrated into GXDW, data exchange formats have been defined for each data space involved in the application. Data exchange formats for sample, gene expression, and gene annotation data follow the general structure described in section 3:

1. The Sample Data Exchange Format involves object classes representing samples, donors, treatments, experiments, studies and study groups (see Figure 1).
2. The Gene Expression Data Exchange Format involves object classes representing gene expression values, analysis methods, and associations of individual gene expression values with samples and gene fragments (see Figure 2), as well as related experimental parameters.
3. The Gene Annotation Data Exchange Format involves object classes representing gene fragments, known genes, EST sequences, EST sequence clusters, pathways and microarray types (see Figure 3).

All data exchange formats contain a "*catch-all*" class which can accommodate any data, represented as tagged-value pairs, that does not otherwise fit the formats.

The most significant problems in importing data into GeneExpress are involved in mapping source data to the data exchange formats. Gene expression data have, in general, well-defined semantics and usually benefit from being represented in a standard (often platform specific) format, such as AADM [1]. Gene annotations also have well understood semantics, although there are ambiguities with regard to the classification of some of these annotations (see [10] for a discussion of problems associated with gene nomenclature and identification). The mapping for sample data is usually the most difficult since there is no widely accepted standard for representing clinical data (see [12], presentations at the working group on ontologies). We describe below some of the problems of exporting sample data into data exchange formats. Note that similar problems arise with gene annotation or expression data as well.

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 [11].

Independent sample data repositories often use their own vocabularies for such concepts, and, even with 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. 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 may use 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. In some cases, when vocabularies are small, or for controlled vocabularies, it may be possible to spot and correct such errors manually, but in general, these errors are hard to spot and may go undetected.

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 best to map them to the various data exchange formats, is a complex task. First it is necessary to determine the identifiers and correlations between individual data objects, which are either provided during the data export process, or need to be determined by analyzing data patterns. In either case, it is necessary to confirm that the correlations found in the data fit the intended semantics of the data.

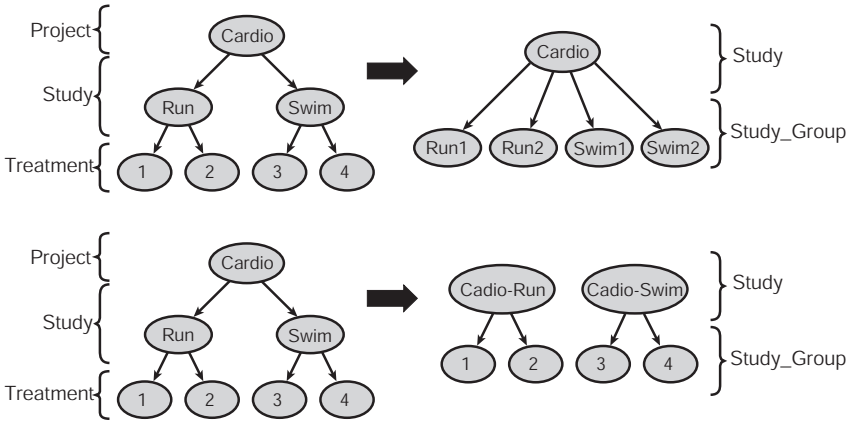

Figure 5: Alternatives for flattening data from three levels into two

Object identifiers 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 resolving structural conflicts between the models, 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 classes *Study* and *Study-Group*. Sample data exported from an external data source may employ a three level hierarchy, such as *Project*, *Study* and *Treatment*. This difference in structure can be resolved in two ways: 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. These two alternatives are illustrated in Figure 5. Note that the choice is neither obvious nor unique.

In addition to the problems described above, it is also 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 and integration tools.

4.2 Data Editing, Transformation and Loading

Once data from external data sources have been mapped to the data exchange formats, additional editing may be required before integrating and loading them into the warehouse.

First it is necessary to detect invalid data, such as missing or inconsistent data. In general the integration tools handle such cases by skipping the data effected 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, such as missing clinical data associated with samples and inconsistent associations of sample and gene expression data.

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.

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

4.3 Data Integration Applications

Data from external data sources have been incorporated into GeneExpress as part of several integration projects. The complexity of an integration project is determined by a number of factors, including whether customer data is structured using a standard format (e.g. AADM) or using a loosely defined format (e.g. a spreadsheet); whether the data involves standard commercial probe arrays or custom probe arrays (which may involve specially designed sets of gene fragments); the richness of the sample data; and the frequency of updates required for the system.

All GeneExpress integration projects have involved exporting data from external sources into data exchange formats for sample, gene annotation and gene expression data. For most of these projects, the source of gene expression data was the GeneChip LIMS, and consequently AADM was used as the gene expression data exchange format, with straightforward data export and mapping. In one project, gene expression data was stored in a custom format, and consequently the expression data was loaded directly into GXA data files, bypassing the stage of loading it into the expression database component of GXDW. Data export and mapping for gene annotations have been limited to proprietary gene annotations and, in most cases, have involved

straightforward mappings to the Gene Annotation Data Exchange Format. The most difficult problem has been presented by the sample data as discussed in section 4.1.

An additional challenge of incorporating data from external sources into GeneExpress is presented by the need to asynchronously update the Gene Logic data and data from external sources. Gene Logic provides customers with periodic updates to the GeneExpress content, on a monthly, bi-monthly, or quarterly schedule. For a standard GeneExpress installation, these updates are carried out by replacing the GXDW and GXA instances with new versions. However, if customer data is incorporated into GXDW, it is necessary to preserve these data while updating the Gene Logic data. In some cases, this problem is addressed by partitioning the GXDW databases and GXA data files into Gene Logic data and customer specific parts. The partitions are organized in such a way that the Gene Logic data can be replaced while leaving the customer data intact. The GeneExpress Explorer then accesses the union of these partitions, which are created using database or GXA union tools.

For certain databases, such as the Gene Annotations Database, the complexity of the dependencies between the data, and the overlap between the Gene Logic and the customer data, means that partitioning does not provide an adequate solution for the database update problems. It is not uncommon that customer and Gene Logic data share some controlled vocabularies, or that customer proprietary annotations refer to public genomic information such as UniGene clusters. Since controlled vocabularies, UniGene clusters and so on may change (e.g., a sequence cluster may be merged into another cluster), it is essential to re-examine customer data, and to re-establish correlations with public or Gene Logic data. For such databases it is necessary to keep track of customer data, and to re-load these data into GXDW after each periodic update, so that the consistency between public, Gene Logic and customer proprietary data is maintained.

5. Summary and Future Work

We described the challenges involved in managing and integrating gene expression data in the context of Gene Logic's GeneExpress system. The GeneExpress Warehouse is hosted on an Oracle 8i database server back-end and is supplied with a continuous stream of data from the GeneExpress Data Acquisition System. Data exploration and analysis is carried out using the GeneExpress Explorer in a three-tier (database server - analysis engine and query processor - user interface client) architecture.

Through the end of 2001, the GeneExpress system had been deployed at over twenty biotech and pharmaceutical companies, as well as at several academic institutions. At the same time, Gene Logic has completed two GeneExpress data integration projects, with the integrated systems deployed at customer sites, and has started several new integration projects. One of the completed integrated systems included support for custom Affymetrix GeneChip probe arrays, for proprietary gene annotations, and for daily incremental updates of customer data into the GXDW. All the projects have included support for integration of sample (clinical) data based on proprietary data formats, as well as gene expression data.

Based on the experience gained in developing tools for incorporating customer data into GeneExpress, we have recently developed tools that provide support for interactive extraction, transformation and loading of gene expression data generated

using the Affymetrix GeneChip LIMS into GXDW.

Acknowledgements. We want to thank our colleagues at Gene Logic who have been involved in the development of the GeneExpress system for their outstanding work. Special thanks to Mike Cariaso and Krishna Palaniappan for their contributions to this paper.

References

[1] Affymterix, "*Affymetrix Analysis Data Model*", http://www.affymetrix.com/support/-aadm/aadm.html.

[2] Chaudhuri, S., and Dayal, U., *An Overview of Data Warehousing and OLAP Technology*, SIGMOD Record, 1999.

[3] Eckman, B.A., Kosky, A.S., and Laroco, L.A., *Extending Traditional Query-Based Integration Approaches for Functional Characterization of Post-Geomic Data*, Journal of Bioinformatics, 17:587-601, 2001.

[4] *Gene Logic Products*. http://www.genelogic.com/products/ge.htm

[5] *Human Gene Nomenclature Database*, http://www.gene.ucl.ac.uk/nomenclature/.

[6] 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, 14:1675-1680, 1996.

[7] Lockhart D.J., and Winzeler, A.E., *Genomics, Gene Expression, and DNA Arrays*, Nature, 405:827-836, 2000.

[8] 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.

[9] Markowitz V.M. (ed), *Special Section on Heterogeneous Molecular Biology Databases*, Journal of Computational Biology, Vol 2, No. 4, 1995.

[10] Pearson, H., *Biology's name gam*e, Nature, 417, pp. 631-632, 2001.

[11] *SNOMED, Systematized Nomenclature for Medicine*. http://www.snomed.org/

[12] *Third International Meeting on Microarray Data Standards, Annotations, Ontologies, and Databases*. Presentations. http://www.mged.org/presentations/index.html.