Genomics & Informatics Vol. 9(1) 19-27, March 2011

Standard-based Integration of Heterogeneous Large-scale DNA Microarray Data for Improving Reusability

Yong Jung[1¶], Hwa Jeong Seo[4¶], Yu Rang Park[1], Jihun Kim[1,2,3], Sang Jay Bien[1] and Ju Han Kim[1,2,3]*

[1]Seoul National University Biomedical Informatics, [2]Systems Biomedical Informatics Research Center, and [3]Institute of Endemic Diseases, Seoul National University College of Medicine, Seoul 110-799, Korea, [4]Medical Informatics, Graduate School of Public Health, Gachon University of Medicine and Science, Incheon 405-760, Korea

Abstract

Gene Expression Omnibus (GEO) has kept the largest amount of gene-expression microarray data that have grown exponentially. Microarray data in GEO have been generated in many different formats and often lack standardized annotation and documentation. It is hard to know if preprocessing has been applied to a dataset or not and in what way. Standard-based integration of heterogeneous data formats and metadata is necessary for comprehensive data query, analysis and mining. We attempted to integrate the heterogeneous microarray data in GEO based on Minimum Information About a Microarray Experiment (MIAME) standard. We unified the data fields of GEO Data table and mapped the attributes of GEO metadata into MIAME elements. We also discriminated non-preprocessed raw datasets from others and processed ones by using a two-step classification method. Most of the procedures were developed as semi-automated algorithms with some degree of text mining techniques. We localized 2,967 Platforms, 4,867 Series and 103,590 Samples with covering 279 organisms, integrated them into a standard-based relational schema and developed a comprehensive query interface to extract. Our tool, GEOQuest is available at http://www.snubi.org/software/GEOQuest/

Keywords: gene expression data, data integration, classification

Introduction

After genome sequencing, DNA microarray analysis has become the most widely used source of genome-scale data in the life sciences (Allison *et al*., 2006; Brazma *et al*., 2001). DNA microarray is a high-throughput and data-intensive technology that provides the means of measuring the expression of thousands of genes or proteins simultaneously and brings the unprecedented development in the informatics and analysis aspect (Chaussabel and Sher, 2002; Quackenbush, 2002). Since this technique provides researchers with comprehensive understanding of biological complex features, it has been used in not only biological but also clinical field.

As many microarray experiments generate large data sets that can contain tens to hundreds of samples, however, it has needed to be managed systematically with computational tools due to their own complexity. Moreover, requirements for ensuring the scientific integrity of data and sharing the data have given rise to development of public microarray repositories like GEO (Barrett *et al*., 2007), ArrayExpress (Parkinson *et al*., 2007), Stanford Microarray Database (SMD) (Gollub *et al*., 2003), keeping pace with the standardization (Edgar and Barrett, 2006; Perou, 2001). With appearance of them, the data generated in one laboratory can be available to other researchers and various analytical methods uncover different biological insights.

Especially, NCBI GEO has kept the largest amount of gene expression data which have grown exponentially, currently holdingover 200,000 samples. Recently, researches to find clinical or biological meaning using GEO data have been actively performed to show its reusability. Butte *et al*. (2006) constructed Phenome-Genome network, Disease Nosology and Gene-Behavior-Disease through analysis of huge amounts of gene expression data in GEO using Unified Medical Language System (UMLS) (Humphreys *et al*., 1998). Yoon *et al*. (2006) constructed an application tool to provide the present large-scale approach for the analysis of GEO microarray data.

Though GEO has abundant practical possibilities, it does not support the standard format or model but its self-structured format. Moreover, it stores the microarray data without distinguishing processed data and raw data. These factors hinder providing a comprehensive biological analysis environment and future integration of

¶Both authors contributed equally.
*Corresponding author: E-mail juhan@snu.ac.kr
Tel +82-2-740-8320, Fax +82-2-747-8928
Accepted 2 March 2011

the external resources and make extracting the desired information difficult.

Currently, some tools have been developed to utilize large-scale GEO data - SeqExpress (Boyle, 2005), GEOQuery (Sean and Meltzer, 2007), ArrayQuest (Argraves *et al*., 2005). However, there is a limitation that they can handle only GEO DataSets which are curated by GEO staffs.

To solve the problem and overcome the limitation, we attempted integration of heterogeneous microarray data in GEO using microarray data standard, Minimum Information About a Microarray Experiment (MIAME). In GEO, it encourages submitters to supply MIAME standard compliant data. However, it is not related to the format, but rather to the content provided.

We performed data field unification of GEO Data table and distinguished between raw data and transformed data using classification method. To integrate the data in an efficient and accurate manner, these processes were developed in both manual and semi-automated way (Vita *et al*., 2006).

Methods

Biological research increasingly depends on computational analysis of data (Miotto *et al*., 2005). A prerequisite for computational analysis is the availability of experimental data in a formalized, structured and machine-readable format. GEO offers DataSet (GDS) which represents a curated collection of biologically and statistically comparable data for computational analysis method. However, there is a limitation that GEO data are reassembled mainly for not entire GEO data but a unit of GEO Series (GSE) only.

To overcome the limitation and provide a comprehensive biological analysis environment, we attempted integration of heterogeneous microarray data in GEO using microarray data standard, Minimum Information About a Microarray Experiment (MIAME). Workflows are as follows (Fig. 1): GEO data localization, Data field unification, Data transformation classification, Standard-based integration of heterogeneous microarray data, and GEO data update.

Gene expression omnibus

GEO is an international repository for gene expression data. It is developed and maintained by the National Library of Medicine (NLM). It serves as a public repository for a wide range of high-throughput experimental data like single and dual channel microarray-based experiments measuring mRNA, miRNA, genomic DNA and protein abundance.

In GEO, the basic entity types are Platform, Sample and Series. Platform includes a summary description of the array (Descriptive information) and a data table defining the array template (Data table). Each row in the data table corresponds to a single element, and includes sequence annotation and tracking information as

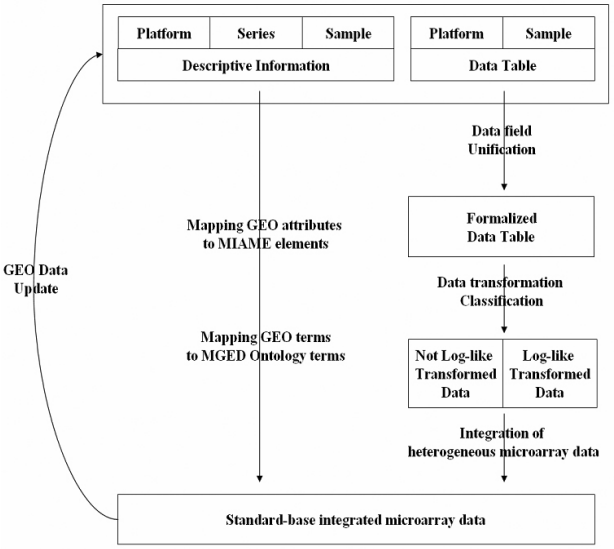

Fig. 1. The workflow of standard-based integration. As indicated, the integration of heterogeneous GEO data into standard based-format was through several processes.

Table 1. Criteria for data field unification

Entity			Property of representative fields	Number of fields
Platform			Required field of MIAME, Required field of GEO, Public biological database name (ex. GenBank, Unigene, Entrez Gene and so on (Wheeler *et al*., 2007).)	42
Sample	Dual channel		Required field of GEO, GenePix Results file format	83
	Single channel	Spotted DNA/cDNA	Required field of GEO, GEO fields used mostly and mainly	30
		Spotted oligonucleotide		53
		In situ oligonucleotide		15
		Affymetrix		37

provided by the submitter. Sample includes a description of the biological source and the experimental protocols to which it was subjected (Descriptive information), and a data table containing hybridization measurements for each element on the corresponding Platform (Data table). Series defines a set of related Samples considered to be part of a study, and describes the overall study aim and design. Each of these entities is assigned an accession number that may be used to cite and retrieve the records.

Each entity type has 'Descriptive information' for describing each entity's correspondent information. Platform and Sample have 'Data table' which consists of 'Header' and 'Matrix'. Header identifies the attributes of each column of Matrix. Matrix includes each entity's correspondent data contents.

GEO data localization

We accessed and downloaded all GEO data on January 22, 2007. Python language and MySQL DataBase Management System were used in Linux system for localizing GEO database. Since data in GEO's File Transfer Protocol were updated late at that time, we used HyperText Markup Language (HTML) documents from which users can see web page. Accession Display - accession to the data through GEO accession number - was used to automatically download all GEO data. Using specific HTML tag structure, we localized all GEO data.

To integrate increasingly large volume of GEO data, formalizing loosely-defined format of GEO is an indispensable process. First, we determined representative fields and constructed mapping tables for Data table of three technology types - 'spotted DNA/cDNA', 'spotted oligonucleotide', and 'in situ oligonucleotide'. The criteria of determining representative fields for data field unification are shown in Table 1.

To construct each mapping table, we adapted a simple text-mining method and performed through manual curation with checking a description and values of each field. Through the mapping tables, we mapped fields of all GEO Data table into each representative field. Five database tables are used to store the GEO Data table for Platform, dual channel Sample, and three technology types in single channel Sample. Ambiguous cases on mapping them are treated by EAV (Entity-Attribute-Value) table (Johnson *et al.*, 1997) to prevent loss of data caused by the mapping process.

After inserting data into each result table is completed, the four tables of Sample are integrated through the next step

Data field unification

The GEO database architecture is designed for the efficient capture and storage of heterogeneous high-throughput data sets. The structure is sufficiently flexible to accommodate evolving state of the art technologies (Barrett and Edgar, 2006). Consequently, the data have many different styles and comprise varying contents.

Due to the flexibility, GEO stores the data which include the words having the same concept but different spelling or some misspelled words without any control process. This characteristic increases heterogeneity between data in GEO and makes future integration of this resource with other biological and clinical data difficult.

Standard-based integration of heterogeneous microarray data

Data standard is an essential requirement for representation of information to ensure proper semantic integration of heterogeneous data, and also for communication standards to ensure interoperability between disparate data sources (Louie *et al.*, 2007; Martin-Sanchez *et al.*, 2004). In microarray data, the introduction of the MIAME standard has been a great success (Rayner *et al.*, 2006). However, the data format of GEO does not follow the standard. Therefore, it hinders semantic integration and interoperability between heterogeneous microarray data. For the reason, we customized the GEO data into MIAME standard-based format.

First, we analyzed attributes of Descriptive information in each GEO entity to understand that which GEO attribute corresponds to which part of MIAME. Second, we mapped GEO attributes to elements in each part of MIAME and stored them into the database. Third, we mapped values of some GEO attributes into controlled terms from the MGED Ontology (http://mged.sourceforge.net/ontologies/index.php). For example, we performed mapping from 'technology type' in Platform to *TechnologyType* class and from 'type' in Series to both *MethodologicalDesign* class and *ExperimentalFactor* class. In case of the 'type' in Series, there are the multiple values in one text, we split the text into single values. In the second and third processes, we performed both a simple text mining method and manual curation with checking a description and values of each field. Finally, microarray data in which data field unification process is done are stored into one database table according to classification result between log-like transformed data and raw data.

Data transformation classification

The hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Before the levels can be compared appropriately, a number of transformations must be carried out on the data to adjust the measured intensities to facilitate comparisons and to select genes that are significantly differentially expressed (Quackenbush, 2002).

In microarray data analysis, difference between raw data and transformed data affects analysis results importantly. Moreover, MIAME provides the conceptual structure for the representation of microarray data including raw and processed data (Brazma *et al*., 2001; Rayner *et al*., 2006). However, both raw data and transformed data are stored together in GEO with no separation. For the reasons, we must distinguish between transformed data and raw data. Even though submitters are encouraged to describe about a process of Sample 'Value' field which indicates final expression value measurements, a large part of 'Value' fields have not only an ambiguous description or no description but even wrong description (Table 2).

Table 2. Cases of transformation method treated as log-like transformation

GEO Description	Class by description	Human validation
RMA calculated Signal intensity	Log-like	Not Log
RMA Express calculated the Signal values	Log-like	Not Log
RMA-calculated Signal intensity (natural scale)	Log-like	Not Log
This is the gene expression value following quantile normalization and robust multi-array analysis.	Log-like	Not Log
Expression values represented by RMA (R/Bioconductor; http://www.bioconductor.org/)	Log-like	Not Log
Same as UNF_VALUE but with flagged values remove	Log-like	Not Log

Table 3. List in wrong description of data processing among 200 GEO Sample data sampled randomly

Representative method	List of methods
Log-like transformation value	Log transformed value UNF_VALUE Z-transformed value Robust Multichip Average (RMA) value VSN transformation

In this paper, we assumed that the log transformation are mainly used among the several transformation processes and attempted to distinguish raw data and log-like transformed data. We treated some transformation method as log-like transformation (Table 3). We propose a model for data processing classification using machine learning techniques.

Selection of a training data set

For most classification study, a training data set consisting of records whose class labels are known must be provided. It is necessary to select training data properly because the training data set is used to build a classification model. First, we performed simple random sampling and extracted 200 GEO Samples. Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed. Next, we classify data manually in compliance with three criteria for giving the correct class to data.

- Description of 'Value' field in GEO Sample
- Range of data distribution
- Symmetricity of distribution

In the classification process, guarantee of data quality is very important factor. We trimmed two-tailed 5% values of the distribution in each Sample to remove outliers before selection of training data. We extracted 188 GEO Samples for three criteria and emailed to data submitters of 12 Samples, the rest of the data set. As a result, 190 Samples are determined as training set (96 Log-like transformed data, and 94 Not-log transformed data).

Learning process for classification model

We determined the features that can explain a difference between two classes to create a classification model.

The first one is the difference of the skewness values between original distribution and its log-like transformed distribution (DSD). Skewness is a measure of the asym-

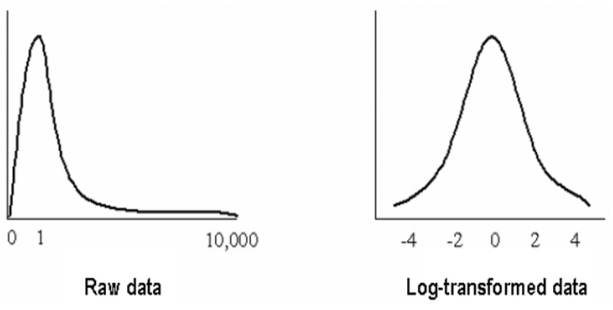

Fig. 2. Difference of distribution between raw data and log-transformed data in an identical data.

metry of the data of the probability distribution:

$$skewness = \frac{m_3}{\sigma^3}, \quad m_3 = \frac{\sum_{i=1}^{n}(x_i - X)^3}{n} \tag{1}$$

where χ_i is each value of the distribution and x represents the mean value of the distribution. *n* is the number of values in the distribution.

If expression values of raw data are log-like transformed, the skewness of original distribution is changed remarkably (Fig. 2). However, if those of log-like transformed data are log-like transformed again, the skewness is changed slightly. This characteristic makes the feature available for the classification model.

Second, a maximum value of data (MD) is concerned. Common image scanners generate typically 16-bit Tagged Information File Format (TIFF) images. Therefore, the log-like transformed data have rarely over 16. On the other hand, raw data can have values more than thousands. Since the maximum value can be a good factor for the classification process, we include this feature.

To determine the classification model, we adopted the logistic regression method with these features:

$$y = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}, \quad i = 1, \ldots, n \tag{2}$$

where *y* is defined as above and xk,i is the value of features of ith GEO Sample data. *k* is the number of features and *n* is the number of GEO Samples. The logistic regression has several strengths. 1) It is not restricted to data. 2) It can represent the model easily. 3) It can classify the data into each group with ease. 4) It finds the best explanatory variable. We can predict the class, log-like transformed or not for any GEO Sample data using the output computed by the model. If y is the positive value, the data is classified as log-like transformed data. If y is the negative value, it is classified as not log-like transformed data.

Set-wise validation of classification result

10 Sample data excluded from 200 Sample data were used to validate the classification model. We performed classification of 10 data and case-study of its result. A GEO Series is a group of related GEO Samples. Thus, we can assume that Samples in a Series should be classified as same class. However, it is found that Samples are classified with different class in a Series. To solve this case, we adopted simple voting method and classified Samples in a Series as a class with which Samples are classified more.

GEO data update

As an amount of GEO data grows exponentially, we need to update the data continuously. For incremental update, we extract a list of GEO accession numbers and release date in our database. Next, if a new data is not in the list, we download, process, and store it into the database. To trace the update status, accession numbers of updated data and the updated date are written in a log file.

Besides, authority of data in GEO changed often public to private, and vice versa. We also recorded the accession numbers of the data in the log file. Update workflow is shown in Fig. 3.

Results

Data field unification

In this section, we focused on constructing the mapping tables for each technology type and each manufacturer. Through a simple text-mining method and manual curation, we mapped the fields which are written in various strings into one representative field. Table 1 shows the example of words which have various strings for a concept.

There are around 3,000 distinct fields in Platform and Sample Data tables respectively. Since most of the data fields are recorded irregularly, we have to unify them manually. On the other hand, the fields of dual channel

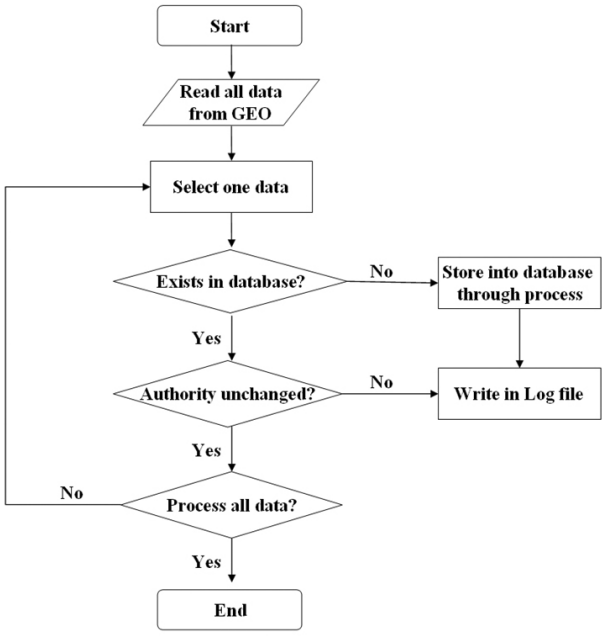

Fig. 3. General workflow of GEO data update.

Table 4. GEO field mapping result

Entity	Technology type	Number of samples	Mapped fields	Total fields
Platform	Spotted DNA/cDNA	1,437	958	1,253
	Spotted oligonucleotide	748	462	817
	In situ oligonucleotide	687	218	387
Single channel sample	Spotted DNA/cDNA	4,670	180	343
	Spotted oligonucleotide	3,217	85	97
	In situ oligonucleotide	431	16	21
	Affymetrix	49,253	193	240
Dual channel sample	All	46,019	2,705	3,858

in Sample are stored in partially regular pattern. We performed a simple text mining method (exact match) and obtained a precision of 0.8474 (544/642), a recall of 0.2011 (544/2,705), and the F-measure of 0.3248. Through its method was not reliable as the result shows, it was helpful to reduce a time-consuming process. The unification result is shown in Table 4.

Standard-based integration of heterogeneous microarray data

We designed a core relational database according to the MIAME standard-based format. In it, the GEO data is customized according to the result of mapping GEO attributes to elements in each part of MIAME. A comprehensive view of mapping result is presented in Table 5.

We mapped values of attributes in GEO ('technology type' in Platform, 'type' in Series, 'Extracted molecule' and 'Label' in Sample) into terms from MGED Ontology (*TechnologyType*, *MethodologicalDesign and ExperimentalFactor*, *and LabeledExtract* class) respectively. Like data field unification process, we performed a simple text mining method. Among 4,495 values in GEO, 2,537 values are mapped to MGED Ontology and 1,958 values are stored as they are. Other database tables have been implemented in order to collect data regarding all submitters, laboratories or organizations which take part in each experiment and to handle the case that a Sample included in multiple Series. All data regarding microarray experiment results (Data table) are stored in the additional database tables according to its transformation characteristic.

Finally, we hold 2,967 Platforms, 103,590 Samples (57,052 single channels and 46,538 dual channels) and 4,867 Series in our database. In the database, we investigated the distribution of organism, source, and extracted molecule in microarray experiments (Table 6). The view of the results reflects that the most interesting experiments in the present researches have been concerned with Homo sapiens. Moreover, it shows that the breast tissue is mostly used as a material for experiments and that breast cancer is a matter of primary concern in cancer research. In a part of Extracted Molecule, the distribution means the data stored in GEO mainly result from array-based experiments in which researches of transcriptional pattern is accomplished.

Table 5. Mapping GEO data elements into core database tables based on the MIAME standard

MIAME element	GEO Entity	GEO attribute
Experiment	Series	Title
		Type
		Summary
		Overall design
		PubMed ID
		Web link
Biological samples, preparation extraction and labeling	Sample	Organism
		Label
		Label protocol
		Extracted protocol
		Extracted molecule
		Growth protocol
		Treatment protocol
		Source
		Biomaterial provider
		Description
		Characteristic
Array	Platform	Title
		Distribution
		Technology type
		Manufacturer
		Manufacturer Protocol
		Catalog number
		Coating
		Support
		Description
Hybridization	Sample	Hybridization Protocol
		Description
		Sample type
Measurement	Sample	Scan protocol
		Data processing

Data transformation classification

Classification using Logistic Regression model

The initial classification is done by logistic regression, which is used when users have a binary dependent variable. The training data set currently consists of 190

Table 6. Distribution of top five ranked terms for organism, source, and extracted molecule

GEO feature	Term	Count
Organism	Homo sapiens	53,989
	Mus musculus	22,679
	Rattus norvegicus	8,620
	Saccharomyces cerevisiae	6,736
	Arabidopsis thaliana	4,961
Source	Breast tumor	1,691
	Pool	1,484
	DNA was obtained from NIGHS Human Genetic Cell Repository	1,152
	Breast	1,022
	Lymphoblastoid cell line	769
Extracted molecule	Total RNA	93,839
	Genomic DNA	7,936
	PolyA RNA	1,224
	Protein	385
	Other	204

Table 7. Contingency table for logistic regression learning result

Actual	Predicted		Total
	Log-like	Not Log	
Log-like	96	0	96
Not Log	0	94	94
Total	96	94	190

examples.

After learning process, we can obtain the classification model and the contingency table as a learning result (Table 7). As we can see the result, we obtained a sensitivity of 1.0 (96/96) and a specificity of 1.0 (94/94).

Set-wise validation

To test the classification model, we performed classification for the 10 Sample data excluded from 200 Sample data which were sampled randomly to make training data set. We assumed that basically, Samples in a Series have same classification results. Among the 10 classification results, however, we found a questionable result. Though two Samples are included in same Series, one was classified as Log-transformed data and the other was not. For the reason, we considered Series as a set and attempted the set-wise validation.

We applied our classification model to entire data set and extracted the data which correspond to questionable case. Entire data consist of 3,911 Series which include 97,026 Samples. The reason why the number of entire Samples is not 103,590 is that GEO Values of 6,564 Samples have only 'null' string or zero value. As a result, we found 103 Series which include 4,876 Samples. These data were classified by humans to validate the classification model (Table 8).

Table 8. Contingency table for data set in which different classification results for each sample in a series

Actual	Predicted		Total
	Log-like	Not Log	
Log-like	2,896	257	3,153
Not log	325	1,398	1,723
Total	3,221	1,655	4,876

Table 9. Comparison of logistic regression model and mixed model with validation on data subset (DCS) and entire data set

Measurement	Logistic regression		Logistic regression + voting	
	DCS	Total	DCS	Total
Accuracy	0.8806	0.9940	0.9967	0.9998
Error rate	0.1194	0.0060	0.0033	0.0002
Sensitivity	0.9185	0.9954	0.9990	0.9999
Specificity	0.8114	0.9921	0.9925	0.9997
Precision	0.8991	0.9942	0.9959	0.9998
Recall	0.9185	0.9954	0.9990	0.9999
F-measure	0.9087	0.9948	0.9974	0.9999

To solve this problem, we applied a simple voting method. For example, if the number of log-like transformed class is more than that of not-log transformed one in a Series, all Samples in the Series are assigned to the log-like transformed class. If the number of log-like transformed class is equal to that of the not-log transformed one, all Samples are assigned to what they are classified as. We tested both the original classification model and a combination of the model and voting method on data sets differently classified in a Series (DCS) and entire data set, respectively. The results are presented in Table 9 with various evaluation measurements.

As a result of classification for all GEO data, 56,012 log-like transformed data and 41,014 not-log transformed data are stored into separated database tables.

Query Interface

The integrated database can be queried on the World Wide Web at http://geo.snubi.org/~geoxperanto/html/

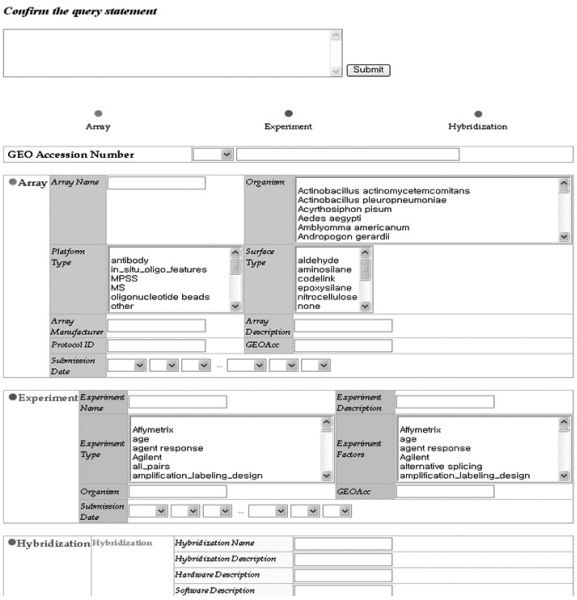

Fig. 4. Query interface of an integrated database of GEO. Some of GEO terms are mapped to MGED Ontology. It makes it possible to search a categorized term.

GEORetrieval/GEORet_Inter.html (Fig. 4) This interface enables users to search hybridization-centered data. In comparison with free text search of GEO Entrez Search System, users can search the database efficiently due to partially itemized GEO terms.

Discussion

Some main issues of management in production of microarray experiments are the large amount of information produced and their heterogeneity. Therefore, it is important to make independently collected microarray data conform to standard for sharing of data efficiently and be comparable with each other. The MIAME standard can serve both bioinformaticians and biologists to deal with the former issue. To solve the latter issue, microarray data must be classified by transformation method. In the process of data classification, we found that a combination of a classification model and other method can boost the performance of classification (Chen *et al*., 2006). We have presented a simple but effective two-step method. It consists of a Logistic Regression model as well as a simple voting method. Recently, it is clear that these datasets should be combined to generate a more comprehensive understanding of underlying biology. With appropriate integration of heterogeneous microarray data in GEO into the standard-based database, improvement of analysis results and comparison of data from different experiments can be possible. Integration strategies we proposed allow the GEO to progress remarkably toward a more standardized repository and to serve as a more uniform platform for microarray data analysis. Also, published research studies using GEO data can be expanded and improved with our database and analysis approaches (Yoon *et al*., 2006) which have been published.

Yet, we have some problem to solve further. In data field unification, we can not handle polysemy problem, which means the case that a word has many meanings. An important next step is to adapt natural language processing method to solve not only the problem but also new terms which will be input in GEO. In set-wise validation of data transformation classification, we assumed that all Samples in a Series have same classification results. This assumption may miss some case of a false positive result in a Series for entire data set. With human validation for entire data set, we can correct our validation result.

At the conclusion, we suggest to data submitters that they submit their data with the correct description in the unified format to collaborate with researchers in other fields or to provide machine-readable data for computational post-analysis. The effort may lead to improve the computational analysis results for the discovery of remarkable biological knowledge.

Acknowledgements

This study was supported by a grant from a Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028631). YRP' s educational training was supported by the Ministry of Health & Welfare, Republic of Korea (A040002).

References

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet*. 7, 55-65.

Argraves, G.L., Jani, S., Barth, J.L., and Argraves, W.S. (2005). ArrayQuest: a web resource for the analysis of DNA microarray data. *BMC Bioinformatics* 6, 287.

Ball, C.A., and Brazma, A. (2006). MGED standards: work in progress. *OMICS* 10, 138-144.

Barrett, T., and Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*. 411, 352-369.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles--database and tools update.

Nucl. Acids Res. 35, D760-765.

Boyle, J. (2005). Gene-Expression Omnibus integration and clustering tools in SeqExpress. *Bioinformatics* 21, 2550-2551.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365-371.

Burgarella, S., Cattaneo, D., Pinciroli, F., and Masseroli, M. (2005). MicroGen: a MIAME compliant web system for microarray experiment information and workflow management. *BMC Bioinformatics* 6 Suppl 4, S6.

Butte, A.J., and Chen, R. (2006). Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA. Annu. Symp. Proc.* 106-110.

Butte, A.J., and Kohane, I.S. (2006). Creation and implications of a phenome-genome network. *Nat. Biotechnol.* 24, 55-62.

Chaussabel, D., and Sher, A. (2002). Mining microarray expression data by literature profiling. *Genome Biol.* 3, RESEARCH0055.

Chen, D., Muller, H.M., and Sternberg, P.W. (2006). Automatic document classification of biological literature. *BMC Bioinformatics* 7, 370.

Edgar, R., and Barrett, T. (2006). NCBI GEO standards and services for microarray data. *Nat. Biotechnol.* 24, 1471-1472.

Gollub, J., Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J.C., Schroeder, M., Brown, P. O., Botstein, D. and Sherlock, G. (2003). The Stanford Microarray Database: data access and quality assessment tools. *Nucl. Acids Res*. 31, 94-96.

Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., and Barnett, G.O. (1998). The Unified Medical Language System: an informatics research collaboration. *J. Am. Med. Inform. Assoc*. 5, 1-11.

Johnson, S.B., Paul, T., and Khenina, A. (1997). Generic database design for patient management information. Proc. *AMIA. Annu. Fall. Symp*. 22-26.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., and Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. *J. Biomed. Inform*. 40, 5-16.

Martin-Sanchez, F., Iakovidis, I., Norager, S., Maojo, V., de Groen, P., Van der Lei, J., Jones, T., Abraham-Fuchs, K., Apweiler, R., Babic, A., Baud, R., Breton, V., Cinquin, P., Doupi, P., Dugas, M., Eils, R., Engelbrecht,R., Ghazal, P., Jehenson, P., Kulikowski, C., Lampe, K., De Moor, G., Orphanoudakis, S., Rossing, N., Sarachan, B., Sousa, A., Spekowius, G., Thireos, G., Zahlmann, G., Zvarova, J., Hermosilla, I. and Vicente, F. J. . (2004). Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *J. Biomed. Inform.* 37, 30-42.

Miotto, O., Tan, T.W., and Brusic, V. (2005). Supporting the curation of biological databases with reusable text mining. *Genome Inform.* 16, 32-44.

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U. and Brazma, A. (2007). ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucl. Acids Res.* 35, D747-750.

Perou, C.M. (2001). Show me the data! *Nat. Genet.* 29, 373.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat. Genet.* 32 Suppl, 496-501.

Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, E., Irizarry, R.A., Liu, J., Maier, D.S., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert, C. J., Jr., White, J., Whetzel, P. L., Wymore, F., Parkinson, H., Sarkans, U., Ball, C. A. and Brazma, A. (2006). A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics* 7, 489.

Sean, D., and Meltzer, P.S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846-1847.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J., Jr. and Brazma, A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 3, RESEARCH0046.

The Microarray Gene Expression Data (MGED) society. The MIAME checklist [http://www.mged.org/Workgroups/MIAME/miame_checklist.html]

Vita, R., Vaughan, K., Zarebski, L., Salimi, N., Fleri, W., Grey, H., Sathiamurthy, M., Mokili, J., Bui, H.H., Bourne, P.E., Ponomarenko, J., de Castro, R., Jr., Chan, R. K., Sidney, J., Wilson, S. S., Stewart, S., Way, S., Peters, B. and Sette, A. (2006). Curation of complex, context-dependent immunological data. *BMC Bioinformatics* 7, 341.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. and Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res*. 35, D5-12.

Yoon, S., Yang, Y., Choi, J., and Seong, J. (2006). Large scale data mining approach for gene-specific standardization of microarray gene expression data. *Bioinformatics* 22, 2898-2904.