MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Microarray Data Mining Using Gene Ontology

Songhui Li, Michael J. Becich, John Gilbertson

Center for Pathology Informatics, Department of Pathology, Benedum Oncology Informatics Center, University of Pittsburgh Cancer Institute, University of Pittsburgh Medical School Pittsburgh, PA 15261, USA

Abstract

DNA microarray technology allows scientists to study the expression of thousands of genes - potentially entire genomes - simultaneously. However the large number of genes, variety of statistical methods employed and the complexity of biologic systems complicate analysis of microarray results. We have developed a web based environment that simplifies the presentation of microarray results by combining microarray results processed for statistical significance with probe set annotation by Genbank, NCBI RefSeqs, GeneCards and the Gene Ontology. This allows rapid examination and classification of microarray experiments – annotated by NCIBI tools - by Statistical Significance and Gene Oncology Classes. By providing a simple, easily understood inierface to large microarray data sets, this tool has been particularly useful for small research groups focused on a small number of related genes and for researchers who want to ask simple questions without the overhead of complex data management and analysis.

Keywords:

Microarray, Gene Ontology, Data Mining, Gene Expression

Introduction

DNA microarray technology has been widely hailed as a powerful tool to study the global gene expression in organisms or tissues [1, 2]. Microarray can be applied to a wide range of studies including gene regulation, disease diagnosis and prognosis, cancer classification, bio-marker discovery and drug development. The microarray's capacity to compare gene expression patterns in different tissues or conditions threatens to change the way biology is practiced and understood.

However, the large amount of data that characterizes most microarray data sets has created a requirement for specialized storage, analysis, annotation and visualization methods in the management and understanding of microarray data [3]. Furthermore, microarrays are expensive, and though costs of chips and reagents have decreased, they still represent a significant investment especially when one considers that random and systemic variability in array data makes biologic replicates mandatory for any useful analysis [4]. These barriers of analysis expertise, information technology infrastructure and capital are especially difficult for small labs or research groups.

A microarray chip can hold tens of thousands of probes targeting almost the entire genome of an organism. For example, the GeneChip Human Genome U133 Set from Affymetrix contains about 45,000 probe sets targeting over 33,000 known genes [5]. A microarray experiment, performed properly, should therefore generate enough data to shed light on hundreds or thousands of scientific questions. Ideally multiple research groups should benefit from the data in a single microarray project. If microarray data could be analyzed and shared across multiple small laboratories, this would go a long way in mitigating the problems of specialized analysis, infrastructure and cost that currently limit the use of microarray data in small labs.

As custodian of microarray data sets at the University of Pittsburgh, we have fielded numerous simple queries from researchers. These queries are in the form "Is gene X regulated in data set Y?" or "Are apoptosis genes down-regulated in tumors in the prostate cancer data set?" These queries had the following characteristics: 1) they required that probe sets be mapped (annotated) to genes, 2) they required a simple, well understood measure of statistical significant gene expression differences, and 3) it was often necessary to classify genes either through a list or through a biologic process, location or function.

In response to these requests, we have developed an easy-to-use web-based application to allow researchers at the University of Pittsburgh to perform simple queries and classifications on existing microarray data sets based on the specific interests of their labs and research groups. The tools in this application allow users to search for data for genes of their interests by various combinations of statistical significance, gene and probe annotation and Gene Ontology (GO) classifications. Users can also annotate a group of genes or probes with GO terms and find out the distribution of the genes to the GO tree. This application can be a one-stop shopping for simple queries on existing microarray data sets for researchers who don't have the time or money for extensive analysis or IT infrastructure.

Materials and Methods

The system reported in this paper is a web based application that integrates microarray data sets, gene annotation from a variety of sources, and the gene ontology. The major components are listed below and are discussed in the result section.

Microarray Platforms Supported: Affymetrix GeneChip Human Genome U95 set (HG-U95) and U133 set (HG-U133) [5].

Statistical Methods Supported: Data sets were pre-analyzed using Significance Analysis of Microarrays (SAM) [7] and simple fold change prior to loading in the database. SAM analysis gives a score, representing statistical significance, for each gene, with an estimated False Discovery Rate.

Gene and Probe Set Annotation: Affymetrix HG-U95 probe set annotation was from the EnsMart database (http://www.ensembl.org/EnsMart/) of ENSEML, a joint project by the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI) and the Sanger Institute. Gene symbols, RefSeqs, Genbank Accession were from Locuslink (http://www.ncbi.nlm.nih.gov/LocusLink/) of National Center for Biotechnology Information (NCBI). The probe set annotation from EnsMart is joined with Locuslink data by its Locuslink id annotation. Links to GeneCard (http://genome-www.stanford.edu/genecards_v2.27/index.html) and Genbank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) are created dynamically

Gene Ontology Annotation: Gene Ontology data was downloaded from Gene Ontology Consortium website and is updated monthly. GO annotation of human genes was from GOA (http://www.ebi.ac.uk/GOA/) of European Bioinformatics Institute. Gene Ontology annotation is joined with Locuslink data by Locsulink id annotation at GOA.

Gene Ontology (GO) is a control vocabulary produced by Gene Ontology Consortium (http://www.geneontology.org/) to describe the function of gene products, their location in the cell and the biological process they are involved in [6]. Three structured ontologies of defined terms have been established: Molecular Function, Biological Process and Cellular Component. Gene Ontology offer two benefits to microarray study. First, the significantly differentiated genes from statistical analysis can be annotated with GO terms; second, the microarray data can be grouped according to the functions of the genes or biological processes they are involved. The first benefit can tell you where the gene products are, what they are doing and in which biological process. The second benefit is very important in a sense that the information is organized in a meaningful way. We can gain a better understanding of the data than purely statistical analysis because biological significance does not necessarily have to be statistically significant. For example, for the genes involved in a biological process, they may not be significant from statistical analysis, but if they consistently change, even though in small scales, between cancer and normal tissues, the process may be important in the understanding of the cancer.

Softwares Used: MySQL, Oracle9i, ASP, Microsoft IIS.

Results

Implementation

The application (http://bioinfo.upmc.edu) was implemented in the University of Pittsburgh Medical Center intranet with the intention to open the site to internet soon. It has a three tiered architecture. The database was created in MySQL and has been ported to Oracle 9i for administrative convenience. At the application layer, Active Server Pages is used to generate dynamic web pages for the display of data. Figure 1 illustrates the integration scheme for pre-analyzed experimental data, microarray annotation that links probe sets to genes and gene annotation including the Gene Ontology. User documentation and the technical explanation can all be found at the web site.

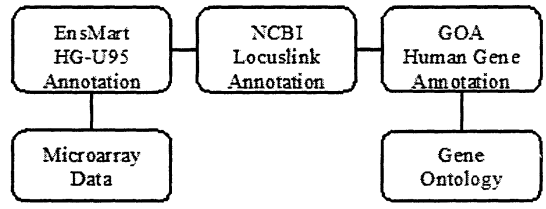


Figure 1 - Integration of data from different sources into the database

Experimental microarray data sets, after analysis, are entered into the system from an Excel spread sheet and stored in the database. We are currently supporting Significance Analysis of Microarray (SAM) analysis as well as simple fold change. SAM provides three related metrics for each probe set: the SAM score (a measure of statistical significance), the "q-value" (an estimation of the false discovery rate for a gene in the dataset), and "Rank" (a ranking of up-regulated and down-regulated genes by SAM score). In the database, Affymetrix probe sets are linked in the database to gene annotation through data downloaded from ENSEML which links Affymetrix probe sets to gene identifiers downloaded from Locuslink including HUGO official and provisional gene names, NCBI RefSeqs or Genbank Accessions. The gene ontology, updated monthly, is linked in the database to Locuslink through the data from the GOA database at the European Bioinformatics Institute. URLs are available in the Materials and Methods section.

Operations

After selecting a data set and experimental condition (ie, Prostate Cancer Data set, Tumor Tissue versus Normal Donor comparison), the system can be used in four main modes:

Search by statistical significance: One can request all genes or probes that fit a given statistical measure, such as the 40, 100 or 200 most differentially expressed genes in the data set by either SAM or Fold Change criteria. The results include gene symbol, RefSeq/Genbank Acc or probe name, Locuslink description of the corresponding gene, SAM result (score, q-value, rank) and fold change for each probe set. Links are created dynamically to Genbank and GeneCard.

Search by gene or probe set: One can enter one or more genes or probe sets. The system will accept HGNC (http://www.gene.ucl.ac.uk/nomenclature/) official symbols or Locuslink provisional names, NCBI RefSeqs or Genbank Accession numbers, or Affymetrix probe names. The system will return the statistical analysis of those genes in the data set in question (as well as the annotation discussed above). This is useful to determine if a give gene or gene list is significantly differentially ex-

pressed in the data set. Figure 2 shows part of the result to search for gene symbols TP53, IL10, and PTEN.

Gene Expression Microarray Data for: TP53 IL10 PTEN

Prostate Dataset 102802

Download Spreadsheet Format

Gene Symbol	Affy Probe Set	Description	Score	Fold Change	q-value (%)	Rank
Tumor vs. Donor						
TP53	1939_at	tumor protein p53 (Li-Fraumeni syndrome)	1.41	1.14	22.81	Up 2495
PTEN	1434_at	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	0.90	1.12	34.11	Up 4157
IL10	1548_s_at	interleukin 10	0.60	1.11	40.86	Up 5309
PTEN	31675_s_at	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	0.41	1.07	46.13	Up 6055
PTEN	39552_at	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	0.00	N/C	N/C	N/C
TP53	1974_s_at	tumor protein p53 (Li-Fraumeni syndrome)	0.00	N/C	N/C	N/C
Tumor vs. Adjacent Normal						
TP53	1939_at	tumor protein p53 (Li-Fraumeni syndrome)	2.15	1.14	8.04	Up 1185
PTEN	39552_at	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	-1.53	0.85	43.81	Down 1463
TP53	1974_s_at	tumor protein p53 (Li-Fraumeni syndrome)	1.05	1.10	27.25	Up 3910
PTEN	1434_at	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	0.75	1.06	34.41	Up 4989
IL10	1548_s_at	interleukin 10	0.00	N/C	N/C	N/C
PTEN	31675_s_at	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	0.00	N/C	N/C	N/C

Figure 2 - Extract microarray data

Search by Gene Ontology Terms: One can select a Gene Ontology term or classification. The system will return results on all genes associated with the Gene Ontology term. The system returns all genes in the node of the term in question and all the nodes under it, as well as all of the annotations and dynamic links discussed above. This search type allows a researcher to segment the microarray data into biologically significant groups, and gives a basic idea of whether a specific aspect of biology, for example apoptosis, is acts differently between the experimental groups. Figure 3 shows data for genes under GO term "extracellular matrix organization and biogenesis".

Prostate Dataset 102802

extracellular matrix organization and biogenesis (GO:0030198): Tumor vs. Donor

Download Spreadsheet Format

Gene Symbol	Affy probe set	Description	Score	Fold Change	q-value (%)	Rank
COL4A2	36659_at	collagen, type IV, alpha 2	-4.33	0.60	0.12	Down 232
COL6A2	34802_at	collagen, type VI, alpha 2	-3.94	0.57	0.12	Down 336
COL6A2	32098_at	collagen, type VI, alpha 2	-1.80	0.55	15.96	Down 1570
SPOCK2	36155_at	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2	1.10	1.83	34.11	Up 3447
ADAMTS3	36269_at	a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 3	0.65	1.12	40.86	Up 5143
COL11A2	1026_s_at	collagen, type XI, alpha 2	0.31	1.05	49.01	Up 6385
COL11A1	37892_at	collagen, type XI, alpha 1	0.00	N/C	N/C	N/C

Figure 3 - Retrieve Microarray Data by Gene Ontology Term

GO Annotation: In addition to standard searches discussed above, the system provides a mechanism to gene ontology annotation. Unlike standard GO browsers, which can be found at the Gene Ontology Consortium web site, this system allows users to annotate a list of genes. This tool has proven useful in examining a list of statistically significant probe sets or genes. By annotating the list with GO, one can get some idea of what they do and how they may be related. Figure 4 show the GO annotation for SIAH1 and SIAH2, which are homologues of Drosophila SINA.

Gene Symbols to Gene Ontology Term Mapping

Download Spreadsheet Format

Gene Symbol	Gene Name	Ontology	Gene Ontology Term
SIAH1	seven in absentia homolog 1 (Drosophila)	biological_process	GO:0006511 ubiquitin-dependent protein catabolism [GO]
			GO:0007275 development [GO]
		cellular_component	GO:0005634 nucleus [GO]
SIAH2	seven in absentia homolog 2 (Drosophila)	biological_process	GO:0006511 ubiquitin-dependent protein catabolism [GO]
			GO:0007264 small GTPase mediated signal transduction [GO]
			GO:0007275 development [GO]
		cellular_component	GO:0005634 nucleus [GO]
			GO:0005737 cytoplasm [GO]
		molecular_function	GO:0003714 transcription co-repressor activity [GO]

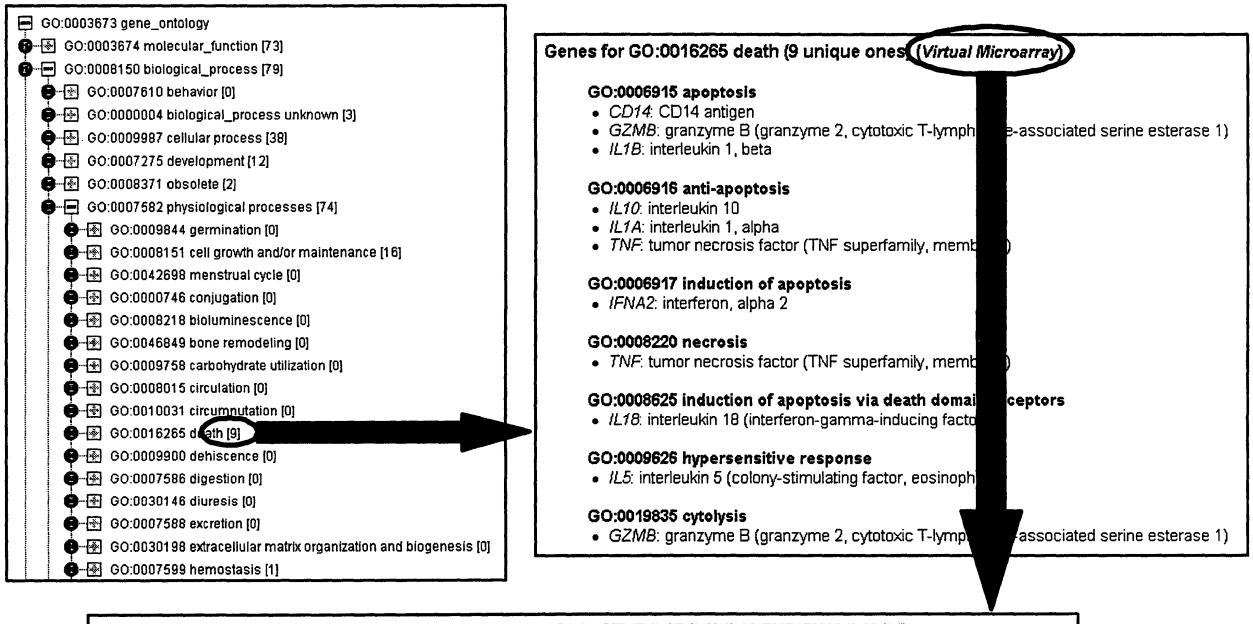
Figure 4 - GO annotation for a group of genes

Distribution of Genes or Probe Sets by GO classification and statistical significance: The system's ability to classify a list of genes or probe sets, such as the most significant genes from a statistical analysis, or simply DNA repair genes, by Gene Ontology terms allows segment the genes or probes of interest into relevant groups. From there, the microarray data can be retrieved according to the grouping.

This is demonstrated in figure 5. The Gene Ontology biologic processes tree is displayed. The numbers in brackets to the right of each branch represent the number of genes out of about 170 genes we inputted represented in the branch. For example, there are 9 genes in the "death" process branch (note, for demonstration purposes, only a subset of U95A probe sets are included in this figure). When the number is clicked, the GO terms in the "death" process branch are displayed along with genes associated with them. Terms without gene association are not displayed. When "Virtual Microarray" is clicked, it will lead you to the analyzed microarray data for the 9 genes in the experimental data set in question (In this case Prostate Data set 102802).

Discussion

Thesis and motivation: With the growing application of microarray technology an increasing number of microarray datasets have become available in public depository, such as NCI data portal generated from Director's Challenge Initiative (http://dc.nci.nih.gov), or published in scientific journals. It is likely that even more data is stored on hard discs in individual laboratories. As the main informatics groups for University of Pittsburgh Department of Pathology and University of Pittsburg Cancer Institute, we were getting increasing requests from small research groups not for complex analysis, but for simple questions such as "is a specific gene, pathway or list of genes differentially expressed in a particular dataset". There was an obvious need for an application that allowed researchers to examine large, pre-analyzed microarray data sets for the answers to simple questions without undertaking complex analysis or visualization. The microarray and gene ontology system reported here is a response to the demand. It is web- based, allowing easy access and easy-to-use. Its focus on annotation (GeneCards) and biological processes (Gene Oncology) is designed to allow res-



Gene Expression Microarray Data for: CD14 GZMB IL1B IL10 IL1A TNF IFNA2 IL18 IL5

Prostate Dataset 102802

Download Spreadsheet Format

Gene Symbol	Affy Probe Set	Description	Score	Fold Change	q-value (%)	Rank
Tumor vs. Donor						
CD14	36661_s_at	CD14 antigen	-4.85	0.45	0.12	Down 167
TNF	1852_at	tumor necrosis factor (TNF superfamily, member 2)	3.18	1.34	1.16	Up 347
TNF	259_s_at	tumor necrosis factor (TNF superfamily, member 2)	1.42	1.13	20.27	Up 2448
IL1B	39402_at	interleukin 1, beta	1.40	1.26	22.81	Up 2527
IFNA2	1791_s_at	interferon, alpha 2	0.82	1.23	40.86	Up 4435
IL5	436_at	interleukin 5 (colony-stimulating factor, eosinophil)	0.69	1.17	40.86	Up 4993
GZMB	37137_at	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	0.66	1.23	40.86	Up 5075
IL10	1548_s_at	interleukin 10	0.60	1.11	40.86	Up 5309
IL18	1165_at	interleukin 18 (interferon-gamma-inducing factor)	0.55	1.12	44.04	Up 5515
IL1B	1520_s_at	interleukin 1, beta	0.00	N/C	N/C	N/C
Tumor vs. Adjacent Normal						
TNF	259_s_at	tumor necrosis factor (TNF superfamily, member 2)	2.30	1.11	6.36	Up 987
TNF	1852_at	tumor necrosis factor (TNF superfamily, member 2)	2.06	1.12	8.70	Up 1327
IL5	436_at	interleukin 5 (colony-stimulating factor, eosinophil)	1.14	1.14	27.25	Up 3568
GZMB	37137_at	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)	1.00	1.21	34.41	Up 4040
CD14	36661_s_at	CD14 antigen	0.00	N/C	N/C	N/C
IFNA2	1791_s_at	interferon, alpha 2	0.00	N/C	N/C	N/C
IL10	1548_s_at	interleukin 10	0.00	N/C	N/C	N/C
IL18	1165_at	interleukin 18 (interferon-gamma-inducing factor)	0.00	N/C	N/C	N/C
IL1B	1520_s_at	interleukin 1, beta	0.00	N/C	N/C	N/C
IL1B	39402_at	interleukin 1, beta	0.00	N/C	N/C	N/C
Adjacent Normal vs. Donor						
CD14	36661_s_at	CD14 antigen	-4.80	0.45	0.42	Down 49

Figure 5 - Distribution of genes on GO tree and the retrieval of their data

searchers who do not normally deal with microarrays to navigate the confusion of genetic nomenclature. Even the selection of significance tests (SAM and Fold Change) was driven by the desire to give researchers both a sophisticated, respected test and a very simple ratio that is easy to understand.

Current Status: The system has been functioning at the University of Pittsburgh for about a year. It has used by 14 research groups to compare the results of traditional assays to microarray data, compare different microarray experiments and examine existing microarray data sets in the context of developing grant proposals. Perhaps its main use is for researchers to rapidly compare microarray results against the predictions of hypotheses and theory.

Future Plans: As a system that integrates statistical significance testing, gene and probe set annotation and the Gene Ontology, we expect that the system will become more and more useful as statistical testing, molecular knowledge and the biologic understanding mature. Over the next year we plan to build on the existing database to include classification by biochemical pathways. Furthermore, the systems ability to annotate microarray results will be included in a new Laboratory Information System being developed to support a clinical microarray and proteomics facility being implemented at the University of Pittsburgh.

Conclusion

We have developed a simple microarray data mining tool that integrates statistical significance testing, NCBI gene annotation and Gene Ontology categorization. The goal of the system is to allow small labs and research groups access to existing microarray data sets without the expense of specialized analysis or purchase of new chips. To this end it has been well received by researchers and it will be soon available outside the UPMC firewall. The system's focus on annotation and classification will make it an important part of a new Laboratory Information Management System (LIMS) being developed to support a Cancer Biomarkers Laboratory supporting both genomics and proteomics at the University of Pittsburgh Cancer Institute and the Department of Pathology.

Acknowledgments

We would like to thank Uma Chandran PhD and Changqing Ma MD for the microarray data analysis, and John Milnes and Gary Burdelski for technical support. The work is supported by a grant from Pennsylvania Department of Health (ME01-740 «Pennsylvania Cancer Alliance Bioinformatics Consortium (PCABC) for Cancer Research», PA Commonwealth Department of Health Tobacco Settlement) to Pennsylvania Cancer Alliance Bioinformatics Consortium (http://www.pcabc.upmc.edu) (S.L., M.J.B. and J.G.).

References

[1] Tefferi A, Bolander ME, Ansell SM, Wieben ED, Spelsberg TC. Primer on medical genomics. Part III: Microarray experiments and data analysis. *Mayo Clin Proc.* 2002: 77(9): 927-40.

[2] Young RA. *Biomedical Discovery with DNA Arrays. Cell* 2000: 102: 9-15.

[3] Butte A. The use and analysis of microarray data. *Nat Rev Drug Discov.* 2002: 1(12): 951-60.

[4] Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A.* 2000: 97(18): 9834-9.

[5] Affymetrix, Inc web site (http://www.affymetrix.com/).

[6] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000: 25: 25-29.

[7] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.* 2001: 98(9): 5116-21.

Correspondence

John Gilbertson, MD, Director, Research and Development, Center for Pathology Informatics and Benedum Oncology Informatics Center, UPMC Cancer Pavilion Suite 301, 5150 Centre Avenue, Pittsburgh, PA 15232, USA. Email: GilbertsonJR@upmc.edu

Songhui Li,

Center for Pathology Informatics,

UPMC Cancer Pavilion Suite 325D, 5150 Centre Avenue, Pittsburgh, PA 15232, USA, Email: lis@msx.upmc.edu, Telephone: (412) 623-7835.