

QUERYING MICROARRAY DATABASES

by

ZOE ALEXANDRA RAJA

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2005

ACKNOWLEDGEMENTS

I wish to thank to all those who provided help and support during the writing of this thesis. I am very grateful to my supervising professor, Dr. Ramez Elmasri for his helpful guidance, encouragement, and motivation. I would like to express my appreciation to my committee members, Dr. Leonidas Fegaras and Dr. JungHwan Oh, for their time and support of this work.

November 16, 2005

ABSTRACT

QUERYING MICROARRAY DATABASES

Publication No. _____

Zoe Alexandra Raja, M.S.

The University of Texas at Arlington, 2005

Supervising Professor: Ramez Elmasri

Microarray technology has rapidly taken a key position among bioinformatics research tools. After the completion of the Human Genome Project, microarray databases have become particularly important to the management and analysis of genomic data. These databases are ideal tools for many research areas involving gene expression patterns under different experimental conditions. This work attempts to assess the querying capabilities of current public microarray database implementations by evaluating their data management, query interfaces, and results presentation. We are

not aware of any comparative study available to date that evaluates this important class of biological databases. We examine and evaluate how several of the current existing implementations handle microarray data so that they can be queried and managed in a useful, understandable, and efficient manner. Our study identifies some of the limitations among existing microarray databases that impact querying and results presentation, leading to suggestions for areas of improvement.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
ABSTRACT.....	iii
LIST OF FIGURES	ix
LIST OF TABLES.....	xii
Chapter	
1. INTRODUCTION	1
1.1 Background	1
1.2 Contribution of this Work	2
1.3 Organization of Thesis	2
2. OVERVIEW OF MICROARRAY DATABASES	4
2.1 Reviewing Microarrays	4
2.1.1 Terminology	5
2.1.2 History of Microarray Databases	8
2.1.3 Detection Technique	9
2.1.4 Applications.....	14
2.2 Basics of Microarray Database Architecture.....	15
2.2.1 Platforms	16
2.2.2 Object Models.....	16
2.2.3 Data Storage Statistics	19

2.2.4 Security	19
2.3 Other Gene Expression Techniques	19
2.3.1 EST Technique	20
2.3.2 SAGE Technique	20
2.3.3 Comparing SAGE and Microarrays.....	21
3. DATA MANAGEMENT FOR EFFECTIVE QUERYING	22
3.1 Need for Standards in Experiment Submission (MIAME).....	23
3.2 Defining Data Types and Parameters	23
3.2.1 Raw and Processed Signal Data Types	24
3.2.2 Gene, Array, Experiment, and Sample Data Types.....	25
3.3 Metadata Structures for Microarray Data.....	26
3.3.1 XML for Microarrays: MAGE-ML	27
3.3.2 Tab-Delimited Text file	28
3.4 Example Implementations	29
3.4.1 Data Types Stored	30
3.4.2 Data Submission and Download Formats	31
3.4.3 Data Management Approaches	31
3.5 Summary Diagram	34
4. OVERVIEW OF MICROARRAY DATABASE INTERFACES	36
4.1 Web Based Interfaces	36
4.1.1 Simple Text Box or Web Form	37
4.1.2 Interactive Menus or Graphical Point-and-Click.....	37

4.1.3 Built in Tools.....	37
4.2 Software Tools for Results Visualization.....	38
4.2.1 Clustering Analysis	38
4.2.2 EMAGE for EMAP.....	40
4.2.3 Treemaps	40
4.2.4 List of Visualization Tools	42
4.3 Results Structures.....	43
5. MICROARRAY QUERIES IN RESEACH STUDIES.....	46
5.1 Role of Querying Microarrays for Research.....	46
5.2 Queries for Coexpression Studies.....	48
5.3 Queries for Gene Proximity Studies	49
5.4 Queries for Tissue Localization Studies	50
5.5 Queries for Toxicity Evaluation Studies.....	51
5.6 Queries for Data Mining Studies.....	52
6. QUERY INTERFACES AND EXAMPLE QUERIES	54
6.1 Querying the ArrayExpress Database	54
6.2 Querying the CEBS Database.....	61
6.3 Querying the GEO Database	67
6.4 Querying the EMAP Database	74
6.5 Querying the SMD Database	79
6.6 Querying the HugeIndex Database	87
7. LIMITATIONS AND SUGGESTED SOLUTIONS	94

7.1 Limitations Impacting Microarray Database Querying	94
<i>7.1.1 Limitation 1: Inconsistencies in Feature Ontology.....</i>	<i>95</i>
<i>7.1.2 Limitation 2: Lack of accommodation for free text queries.....</i>	<i>95</i>
<i>7.1.3 Limitation 3: Lack of support for time-varying image data.....</i>	<i>96</i>
<i>7.1.4 Limitation 4: Lack of consensus on gene ID numbering.....</i>	<i>97</i>
7.2 Limitations Impacting Microarray Database Implementations	99
<i>7.2.1 Limitation 1: Web forms for data retrieval and presentation</i>	<i>99</i>
<i>7.2.2 Limitation 2: Use of hyperlinks to external databases</i>	<i>100</i>
<i>7.2.3 Limitation 3: Database does not store original image.....</i>	<i>101</i>
<i>7.2.4 Limitation 4: Lack of centralized and consistent data analysis ..</i>	<i>101</i>
7.3 Inherent Limitations in Microarray Databases	103
7.4 Successfully Addressed Limitations	105
8. CONCLUSION	108
Appendix	
A. APPENDIX A LIST OF MICROARRAY DATABASES.....	112
B. APPENDIX B MIAME STANDARDS FOR MICROARRAY DATA.....	115
C. APPENDIX C MAGE-ML: XML FOR MICROARRAY DATA.....	123
D. APPENDIX D PROTEIN MICROARRAY DATABASES	128
REFERENCES	145
BIOGRAPHICAL INFORMATION.....	151

LIST OF FIGURES

Figure	Page
1. Illustration of microarray techniques	12
2. Fluorescent signals of a microarray data grid image	14
3. Roles of microarray data for medical research	15
4. Illustrative diagram for microarray data types	25
5. Data management, import, and export in ArrayExpress	35
6. MagicTool clustering analysis dendrogram of query output	39
7. Treemaps rapidly identify genes of interest	41
8. The main interface for querying ArrayExpress	55
9. Results of query for “array provider = affymetrix”	56
10. The results set of the ArrayExpress query “aging studies”	58
11. Mage-ML of a results set	59
12. ArrayExpress links to external software	60
13. Example textual data matrix in ArrayExpress	61
14. Query interface for CEBS microarray database	62
15. Query results in CEBS for “mouse and forebrain”	63
16. Navigation menus allow the user to select arrays for comparison	64
17. Expression reports in CEBS using pathway classification schemes	65
18. Expression in the 1,4-Dichlorobenzene degradation pathway	65

19. Go ontology analysis matches expression to GO categories	67
20. The GEO microarray database main query interface	68
21. The GEO Datasets query interface with query results	70
22. Data visualization and cluster analysis interface in GEO.....	71
23. Geo Profiles example query result set.....	72
24. Visualization of signal values and ranks for a condition set.....	73
25. The main query interface of EMAP with navigable tree	75
26. Tissue mapping portion of EMAP interface for basic querying	76
27. Emap interface showing first 6 of a large results set.....	77
28. Gene detail record from the example query result.....	77
29. Query form of the GDX mouse genome expression database	78
30. Main interface of the Stanford Microarray database (SMD).....	80
31. Query “caenorhabditis elegans” in SMD basic search.....	81
32. SMD advanced search interface, links to display or analyze data	83
33. Example results set for a query “experimenter = EISEN”	84
34. Selection options for data view and sort on a result set row.....	84
35. Example view showing experiment and sample details for an array.....	85
36. View showing details and signal for an individual grid position.....	86
37. Gene specific expression query interface in HugeIndex	87
38. Result set for the query “cathepsin” in tissue “lung”	88
39. Individual gene profile for a gene selected from figure 6.6.2.....	89
40. Interface for searching organ specific expression patterns	91

41. Interface to generate scatter plot for expression comparison query	92
42. Scatter plot showing Mac2 gene is a result of example query 3	93
43. Plot for an individual gene selected from the interactive scatter plot	93

LIST OF TABLES

Table	Page
1. Timeline of Recent Microarray Database Implementations	10
2. Platforms for Microarray Database Implementations.....	17
3. Data Storage Statistics for Six Example Databases.....	18
4. Example Microarray Database Implementations.....	29
5. Stored Image Formats in Example Databases	30
6. Methods of Upload and Download in Example Databases	32
7. Metadata Approaches in Example Databases.....	33
8. Customized Tools for Microarray Data Analysis and Visualization	42
9. Summary of Limitations and Solutions Affecting Querying	98
10. Design Limitations of Current Implementations.....	102
11. Minimum Data Descriptors Problem and Solution.....	105
12. Summary of Metadata Format Problem and Solution	106
13. Current status for Successfully Addressed Limitations.....	106
14. List of Microarray and Gene Expression Databases in 2005	113

CHAPTER 1

INTRODUCTION

1.1 Background

Gene expression patterns reflect which proteins are present and at what rate they are being produced (expressed) within a given cell type. Microarray technology has become the predominant method of choice for investigation and research involving gene expression. In parallel with the completion of the human genome project, the number of public gene expression databases available has risen greatly in recent years. Increasing from 12 in 2000 [B00] to 42 in 2005 [G05]. Microarray databases are ideal tools for many research areas including developmental biology, evolution theory, discovering the function of an individual gene/protein, and in the search for new pharmaceuticals. They also allow for the identification of genes that have positive impact, such as which genes are responsible for desired traits in crop plants to improve yield. Several microarray databases have been established exclusively for plant genomes. In summary, since microarray technology was first introduced in the late 1990's it has rapidly taken a key position among molecular research tools. This work attempts to describe and evaluate publicly available web-based microarray databases as tools for biological research.

1.2 Contribution of this Work

In this thesis we attempt to assess current public microarray database implementations. To our knowledge, there is no comparative study available to date that provides a comprehensive evaluation of this important class of biological databases. This work compares several of the existing microarray databases in terms of their data management, querying capabilities, and presentation of query results to the users. For these implementations, we examine and identify some limitations that impact querying or results presentation. Based on this identification, we provide suggestions for areas of improvement. As a look toward the future of these databases appendix D presents a thorough discussion of recent research efforts toward handling image data from both microarray and proteomics experiments for improving future implementations.

1.3 Organization of Thesis

This thesis is divided into 8 chapters. Chapter 1 is an overview of the project and contribution of this work. Chapter 2 provides a general background for understanding microarray databases. Three different aspects are covered. Firstly, a description of how microarray data are generated, described, and interpreted. Secondly, we provide a brief introduction to the basics of microarray database architecture. Thirdly, other gene expression techniques are reviewed and compared to microarrays. Because selecting and defining the data types and descriptors is important to effective querying, Chapter 3 reviews the microarray data type definitions, standards, and data management techniques. The metadata structures for data exchange between databases are also described. Six example microarray database implementations are contrasted

and summarized in tables. The background provided in chapter 3 helps the reader to understand the example queries of chapter 6 run on the same six example databases. Chapter 4 describes and summarizes the main components of microarray database interfaces. Software tools for results visualization provide an important extension to the utility of these databases. An introduction to some of these tools is provided with graphical illustrations. Chapter 5 explains the importance of microarray databases as a tool for biological research. To illustrate how querying these databases supports research, five different study areas of genetics relying on microarray experiments are defined and examples provided from referenced research publications. Chapter 6 presents illustrated example queries using screenshots from each of the six example databases whose data management approach was summarized previously in chapter 3. In chapter 7 we describe and assess the limitations among currently implemented microarray databases. We then suggest solutions to address these limitations. Finally, chapter 8 presents the conclusion and future work.

CHAPTER 2

OVERVIEW OF MICROARRAY DATABASES

In order to better understand the design decisions and comparative analysis of microarray database implementations, this chapter provides background and review of several important aspects in generating and managing microarray data. Section 2.1 provides terminology background and a basic review of the scientific detection technique by which the microarray data are generated. Section 2.2 provides an introduction to the basic architecture of the microarray databases by looking at both the platforms and object models. In addition, data storage statistics are provided. Section 2.3 describes two other gene expression techniques that result in data often stored in microarray databases, these are EST and SAGE techniques. Finally, use of SAGE is compared with microarrays.

2.1 Reviewing Microarrays

Microarray techniques are relatively recent, but have taken a position as one of the most useful for studying genetic information. This section provides an introduction to four aspects of microarray techniques. First, we review the important basic terminology and biological concepts for microarray data. Second, we provide a brief technical history for the basic design of microarrays. Third, we outline the scientific

process of generating and detecting data using microarrays. Fourth and finally, we provide an overview of research applications to which microarrays are particularly suitable.

2.1.1 Terminology

Microarray: also called gene array, or DNA chip; a small solid substrate to which a known set of DNA sequences is fixed for purposes of identifying unknown DNA samples based on the pattern of matches.

GeneChip: a registered trademark for a commercial microarray manufactured by Affymetrix, Inc. that is sometimes informally used to refer to DNA microarrays

Biochip: a set of microarrays, allowing higher throughput and parallel testing

Expression: when a genetic sequence on the DNA is utilized to produce the protein that it encodes two stages occur, first translation to make an mRNA template and second transcription to make the protein. These two steps result in a protein product that is the *expression* of the gene.

DNA: Deoxy-Ribonucleic Acid, a chain molecule that is the basic unit of biological information passed from one cell to another during cell division, including cells involved in reproduction for all forms of life. Two complementary strands exist for each molecule and form a *chromosome*. Where the strands are separated a short single strand sub region serves as the template to produce a temporary complement that becomes an mRNA molecule. This is the transcription step in expression.

Bases: DNA is made up of four different nucleotides or *bases*, adenine, thymadine, guanine, and cytosine commonly seen as A, T, G, and C respectively in

DNA sequence representation. When two complementary DNA strands are paired together, the A or adenine nucleotides of one strand bond most strongly to the T or thymadine of the opposite strand. Similarly the G or guanine bases bond most strongly to C or cytosine. DNA bases between two different strands naturally only pair T-A and G-C. All microarray technology is based on this rule.

mRNA: These are short single stranded subsections of DNA that are copied into molecules for temporary use as templates to assemble protein sequences in the translation step of expression. RNA is similar to DNA except that the thymadine is represented by a slight chemical variant known as uradine or U in sequences. mRNA may be an edited version of the RNA copied from the original gene, the sequence can be shortened or spliced before the protein is made so that one gene may encode several proteins.

cDNA: also called complementary DNA, is synthesized by researchers under lab conditions and does not occur in nature. cDNA was used in early microarray designs; it is now common to use a larger population of sequences shorter than cDNA. Each cDNA is the complementary strand to an mRNA, therefore at each position having a C, G, A, or U on the mRNA there will be a G, C, T, or A respectively on the cDNA. The cDNA population or *library* will be fixed to the microarray surface, and then each particular mRNA from an unknown sample adheres only to a matching cDNA sequences and forms a tightly bound double stranded complex. The closer the match, the tighter it is bound.

MGED: The definition provided on their website <http://www.mged.org/> summarizes the organization well, “The Microarray Gene Expression Data (MGED) Society is an international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. The current focus is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promoting the sharing of high quality, well-annotated data within the life sciences community. A long-term goal for the future is to extend the mission to other functional genomics and proteomics high throughput technologies”.

MAGE: Microarray Gene Expression group. A group within MGED dedicated to developing standards for microarray data and databases. These standards include MAGE-ML, an XML based metadata format derived from their Microarray Gene Expression Object Model (MAGE-OM). MAGE-OM was developed and described using the Unified Modeling Language (UML) to the specifications of OMG (Object Management Group, another international consortium that establishes standards for data modeling. In addition the MGED Ontology Working group is developing an ontology for microarray data types, details of which are provided at their websites <http://mged.sourceforge.net/ontologies/MGEDontology.php> and <http://mged.sourceforge.net/ontologies/MAGEontologies.html>. These standards began development in 2001-2002 and for many microarray databases are now either in use or planned for future implementations.

Transcriptome: The complete set of mRNA transcripts representing all expressed genes in a particular cell under a particular set of conditions. Microarray databases essentially store transcriptomes for analytical comparison. There may be hundreds of transcriptomes associated to an organism under normal conditions, and potentially thousands under different experimental conditions.

UniGene: A publicly available system for partitioning and organizing gene entries from the gene sequence repository GenBank into a nonredundant set of gene clusters. UniGene is a cDNA array based source of data for searching, mapping, and describing transcriptomes. UniGene therefore benefits microarray experiment design.

2.1.2 History of Microarray Databases

The concept of commercial array kits consisting of tiny wells for detecting a particular sequence began in the 1980's with proteins which bind to a specific antibody (immunoassays), the antibody molecule being chemically fixed to a small plate. From these techniques came the goal of a similar DNA-based assay whose main technical challenge was increasing the assay sensitivity to detect tiny quantities of DNA [EC99]. It was not until the 1990's that commercially viable DNA microarrays on small silicon chips became available. A California based company, Affymetrix lead by researcher Stephen Fodor, developed one of the first chips in 1993 (holding one million sequences), which they named GeneChip. Affymetrix are the most widely used microarrays and the company has an important role in development of standards for this technology.

Since 2000, microarrays have flourished in parallel with the success of the Human Genome Project as a tool for genetic study and source of high throughput data resulting in microarray specific searchable databases as well as standards for storing their data. Microarray databases are a natural extension of earlier gene expression databases. Because microarray data is high volume and efficient to produce, in the five years between 2000 and 2005 the majority of gene expression databases predominantly contain microarray data and many exclusively store microarray data.

Table 1 below illustrates how recently these databases have been developed. The list contains several important databases, which are the focus of the examples in chapters 3, 6, and 7 in this work. In the table, the year established is the year of first listing in the annual publication *The Molecular Biology Database Collection: Update* from the journal Nucleic Acids Research. The collection included over 700 public biological databases by 2005. Among those, the number of databases specializing in expression data has steadily risen from 12 in 2000 [B00] to 42 in 2005 [G05]. Microarray experiments have been either the major or exclusive source of data for most gene expression databases since 2002.

2.1.3 Detection Technique

The chip has known sequences of synthetically produced DNA matching those from genes fixed to its surface and indexed. Affymetrix, Inc provides 60 million probes for its commercial chip sets, these probes are the short DNA sequences, which are fixed to the chip. Available chip sets cover several species, and can detect specific subregions

Table 1: Timeline of Recent Microarray Database Implementations

Database name	Year established	Comments
SMD	2001	By 2002 only 19 gene expression entries in the Nucleic Acids Research Collection, of those only 2 are specifically listed as “microarray”; SMD and yMGV for yeast (at the time of this writing yMGV website exists, but was not updated since 2003 when it was withdrawn from the collection).
GEO	2002	One of the first microarray databases to follow formal data structures recommended by MGED.
HugeIndex	2002	HugeIndex provides data exclusively from the human genome.
ArrayExpress	2003	Although microarrays were already the main source of gene expression data in 2002, among new expression database entries for 2003 only two were specifically listed as microarray databases; ArrayExpress and TRANSPATH. (Transpath quickly evolved into a family of databases and like several other databases at this time switched from public to commercial status, charging a license fee for access)
EMAP	2004	EMAP is a graphical platform extension for querying two mouse gene expression databases, EMAGE and GXD.
CEBS	2004	This is the Phase 1 of a long-term 10-year goal of expanding capability through several other phases. Phase 1 includes microarray gene expression data, toxicology/ pathology data, and associated analysis tools.
CIBEX	2004/2005	Only a preliminary launch. Well designed, with focus on compatibility with ArrayExpress and GEO. Very few data entries stored as of 2005.

within genes. Most interest in microarray technology focuses on gene expression, but microarrays are also available which look at noncoding DNA, not only the 3% that encodes genes for proteins.

All microarray assays are based around five basic experimental steps: 1) design of the biological query, 2) preparation of the sample, 3) biochemical reaction on the chip, 4) mechanical detection of the assay, and finally 5) data visualization and modeling. In a typical experiment sample DNA from a cell culture or other source representing the query will be prepared, and typically includes chemical attachment of a small fluorescent molecule for later visualization. Each grid position contains many copies of the same known sequence and is able to collect either a few or many copies of the matching sequence if it is present in the sample, therefore important data for relative amount of a particular sequence present can be detected as the intensity of the signal. When a sample of DNA is placed in solution over the chip sequences those sequences in the sample that highly match some on the chip will chemically attract with higher affinity than those that have a lesser match. This attraction and binding is known as hybridization of the base pairs. By using precise washing conditions those sequences with the best matches will remain in place while the weaker held less close matches are washed away. The matches are then detected by a scanner as spots and the images captured.

In the following diagram 2.1 taken from MIT (Massachusetts Institute of Technology) open source internet learning modules, we see the steps of a microarray experiment. This example correlates to the 5 step definition given above of a generalized microarray experiment as follows: Step 1 experimental design is complete before RNA isolation. Step 2 preparation of the sample includes diagrams A, B, and C

below. Step 3 hybridization is in diagram D. Steps 4 and 5 detection and data analysis occur in diagram E.

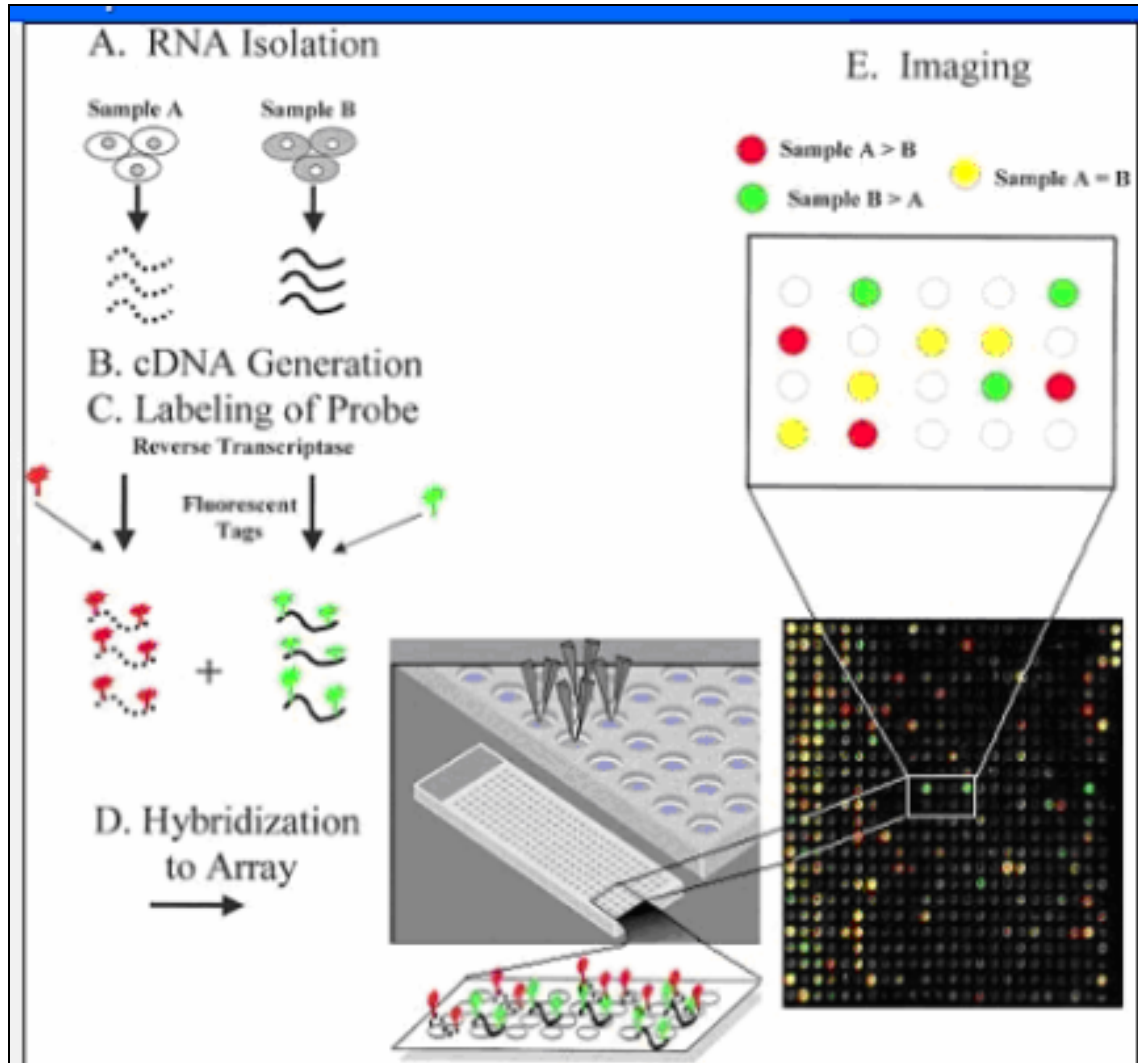


Figure 1: Illustration of microarray techniques

In order to understand the signal interpretation from figure 1, assume red represents mRNA from a normal cell of a particular type, green represents mRNA from the same cell type experimentally exposed to pesticides. Where the array shows red spots, the normal cell is expressing more of that particular gene than the pesticide-

exposed cell. Similarly, where the array shows green spots the pesticide-exposed cell is making more of that gene product than the normal cell. Where the spots are yellow both the normal and pesticide-exposed cells make the same amount of those genes. This array tells us which genes are affected (spots that are red or green instead of yellow), how they are affected (reduced or amplified expression) and how much they are affected (degree of deviation from normal expression).

Explanation of example experiment in figure 1:

- A. Isolate samples of RNA from two cell groups, the control and the experiment
- B. Generate more stable two stranded cDNA version of each RNA sequence
- C. Label 2 RNA samples with 2 different colors of fluorescent dye; for example the known control in red vs. the experimental in green
- D. Mix two labeled RNAs and hybridize to the chip
- E. Make two scans, one for each color. Combine the images to calculate ratios of amounts of both RNA samples from the control and the experimental preparations that bind to each spot.

Figure 2 below provides a more detailed example of part of a microarray data grid image showing the combinations of red and yellow fluorescent light from the signals. An additional point should be clarified regarding microarray experiments. The presence of an mRNA molecule does indicate that the protein will be synthesized but the rate of synthesis and post-synthesis chemical modification are variable. As a result, the concentration of proteins over time cannot be exactly determined by DNA microarrays.

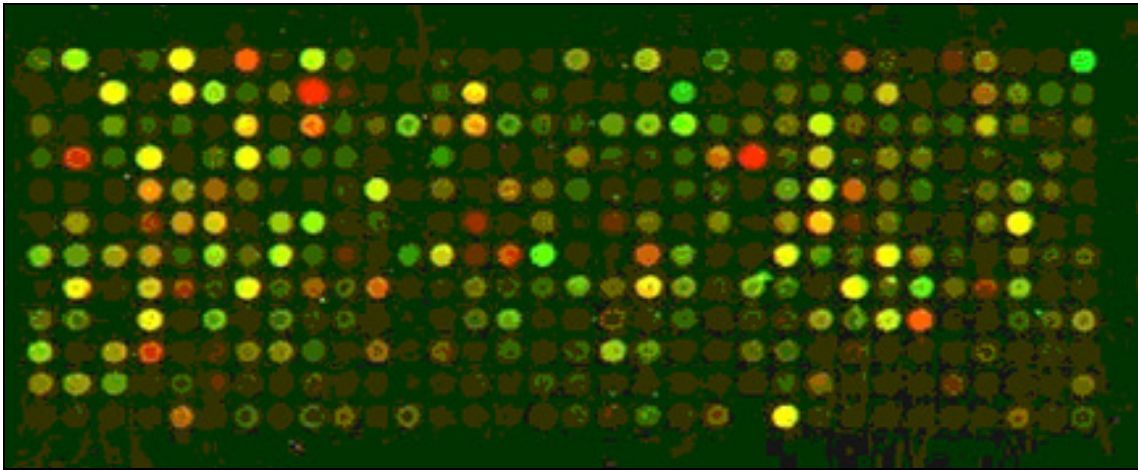


Figure 2: Fluorescent signals of a microarray data grid image

Protein microarrays have been developed instead for this purpose. This work focuses on DNA microarray data, please see appendix D for a discussion of protein analysis.

2.1.4 Applications

Microarrays are able to provide information for an entire genome, equivalent to a snapshot of all possible genetic expression by a particular cell under a given set of conditions. Microarrays are important to functional genomics (study of gene product function) because they allow visualization of how genetic expression patterns change under different cellular, physiological, and toxicological environments. For example, how cancer cells differ from normal cells, how cells respond to decreased nutrients, how cells respond to the presence of trace pesticide (toxicogenomics) or a new medical drug (pharmacogenomics). Individual genetic profiles can be tested to determine relative susceptibility to diseases, as nearly all diseases are accompanied by a change in expression profile.

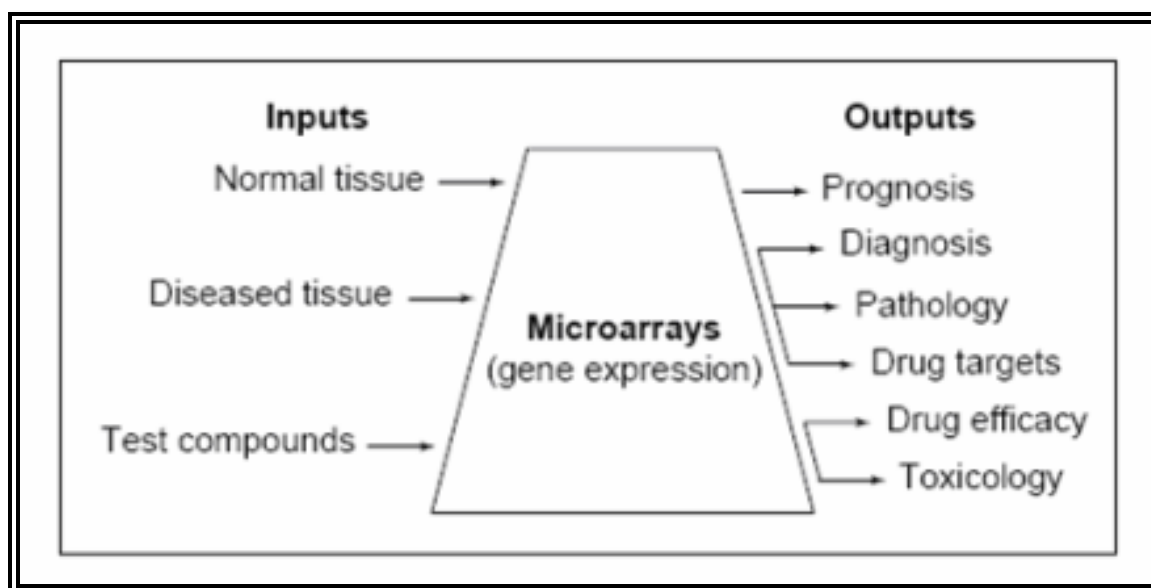


Figure 3: Roles of microarray data for medical research

In figure 3 the comparison between normal and disease genetic expression patterns is shown. Changes to those patterns after cell or tissue exposure to either potential toxins or pharmaceutical treatments are areas of intense research focus [SHTKLD98].

2.2 Basics of Microarray Database Architecture

In this section we provide a brief overview of some architectural choices that are seen among typical microarray database implementations including choice for platform, recently developed object models such as the standard recommended object model designed by the MGED group, typical scale for amount of data stored, and security issues.

2.2.1 Platforms

The platforms used for microarray database implementations have considerable variety. The implementations may be based on traditional relational database design, object oriented design, or a combination of both. Table 2 below provides a summary of the platforms used by six example implementations.

2.2.2 Object Models

Development of a single object model for unified data representation is an important goal in microarray database design. Use of a common model contributes to achieving integrated data storage and management. The model also simplifies modification and design of software to analyze and display data sets. The MAGE-OM (Microarray Gene Expression Object Model) is a data object model that attempts to define standard objects for gene expression. It was developed in 2003. MAGE-OM follows OMG (Object Management Group) specifications. It is very large and complex data driven model, much as the data types and relationships it helps to organize. It is too large to be reproduced here, but may be found at the following web link <http://www.mged.org/Workgroups/MAGE/mage-om.html>. The top level MAGE-OM view shows a broad outline of the experiment, including the name, description, associated publications, providers and BioAssays. This object model was developed by the Microarray Gene Expression Data (MGED) Society who also developed standards for data exchange and minimal information for submissions (MIAME).

Table 2: Platforms for Microarray Database Implementations

Database	Description of platform
ArrayExpress	Oracle database with query interface in Java servlets using Tomcat and Velocity. 1) assuming unix and the Oracle RDBMS, can create the DB on local computer scripts available on site. 2) J2EE application server required for software as additional query interface
CEBS	Implemented by NCT (National Center for Toxicogenomics). Server machines and database management information not available.
GEO	Implemented by NCBI; details not provided for this entry in the NCBI handbook of their databases located at: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.337 NCBI defines GEO as having a flexible and open design, acting as a centralized molecular abundance data distribution hub.
EMAP	Important software component in platform for graphics handling: MAPaint designed to executed on the UNIX operating environment (MacOSX, PC-Linux, Solaris or Irix). The program MAPaint and MA3Dview requires X Window environment with OpenGL and Motif libraries to run. CYGWIN provides an Xwindow emulator for MSWindows users.
SMD	The SMD database server is currently an eight-processor Sun V880, which has 32GB of RAM installed. Their web server is an eight-processor Sun E4500, with 8GB of RAM installed. Oracle Server Enterprise Edition version 9i. Relational database model. All source code freely available. Internet access.
HugeIndex	Object relational model implementation. Uses PostgreSQL 7.1 relational DBA and 4 tables for the schema holding data on experiments, expression levels, experimental protocols, and genes.

Although MAGE-OM is recognized as an important model for new microarray database, it has limitations. For example, it cannot include modeling for clinical data or protein data. To address the protein modeling limitation the designers of CEBS extended MAGE-OM and provide a new object model used in CEBS, the Systems Biology Object Model (SysBio-OM). SysBio-OM provides integration of proteomics and metabolomics data with microarray gene expression data. The model is open source and compatible with multiple computer platforms. The UML (universal

modeling language) description of the complete SysBio-OM is publicly available at <http://cebs.niehs.nih.gov/SysBioOM/>. MAGE-OM is implemented in ArrayExpress and CIBEX (forthcoming from the DNA Data Bank of Japan). It is a component of SysBio-OM in CEBS. It has been mapped to SMD (Stanford Microarray Database) despite the difficulties posed by the fact that SMD is not an object-oriented database, but rather a relational database [HGBDMS03]. The mapping from the MAGE object model to a relational model can be difficult. The majority of microarray databases are relational implementations, which may slow the adoption of MAGE-OM.

Table 3: Data Storage Statistics for Six Example Databases

Database	Description
ArrayExpress	As of November 2004, ArrayExpress contained data from more than 12,000 hybridizations covering 35 species. Relative to the content twelve months earlier, ArrayExpress was reported to have grown more than 10-fold. In 2003 the data stored represented 400 Gbytes or over 1 billion microarray data points.
CEBS	Statistics not found.
GEO	As of mid-2004, GEO contained data from more than 30,000 submissions covering more than 100 species. The GEO website reported that their records represent results from over 600 research groups. The data are accessed over 15,000 times each weekday. Bulk FTP downloads average 30,000 per month.
EMAP	Statistics not found.
SMD	As of mid-2005, SMD had data from over 57,000 experiments with over 35 organisms represented. SMD contributed to over 250 publications and stores over 1.6 billion microarray data points.
HugeIndex	Published statistics from 2002 state 59 experiments, and 7,000 genes represented. This number has likely increased greatly since then. Recent data not readily available.

2.2.3 Data Storage Statistics

The volume of data stored in microarray databases increases every year. The values provided in table 3 above should be regarded as a snapshot providing the scale of data storage requirements for microarray databases.

2.2.4 Security

Security is important as with any database. There are two main areas of concern, privacy for unpublished data and privacy for sensitive portions of published data. For the first concern, researchers typically wish to submit unpublished data as they gather it during an experiment so they are able to use the data storage, organization, query, and analysis features of the database. To protect this data before the research is published, it is not accessible to the public and protected by a basic login/password combination. For the second concern, clinical data may contain information about individual patients, which must be removed from the data sets before it is made publicly available.

2.3 Other Gene Expression Techniques

Before microarray technology became the predominant source of gene expression data, two other techniques were important [M01]. These began in the 1990s and still have a role in genetics research, with specialized public databases and repositories being maintained. The first is the use of expression sequence tags (ESTs) and the second is SAGE (Serial Analysis of Gene Expression). We provide descriptions of both in this section, and briefly compare SAGE to gene microarrays.

2.3.1 EST Technique

An expressed sequence tag or EST is a short sub-sequence of a transcribed DNA sequence. ESTs can represent both gene encoding and non-coding regions of DNA. The use of ESTs began as a method to identify gene transcripts, but later has an important role in gene discovery and sequence determination. Because of efficiency, microarrays have largely replaced ESTs. The basic EST technique sets up a simple single pass sequencing of the cDNA sample. This produces a sequence of low quality short fragments of between 200 to 600 nucleotides. The short fragments usually provide enough information to serve as “tags” that will uniquely hybridize to the full gene in chromosomal DNA. This allows detection of the known gene in a sample without the expense of full high quality sequencing. ESTs can be used to design probes for DNA microarrays. Some databases are dedicated to EST data including NCBI dbEST (part of GenBank).

2.3.2 SAGE Technique

Serial analysis of gene expression (SAGE) is a genetics research technique that provides a snapshot of the messenger RNA population in a particular sample. The original technique was developed circa 1995 and was important to gene expression studies in the years preceding the emergence of microarray databases. SAGE is a variant of the techniques using ESTs (described above), and also results in a set of short sequence tags. SAGE differs from standard EST techniques in that it records the abundance of each mRNA in the population. This quantitative data has been reported as slightly more precise than microarray signal intensity. In cases of very low copy

number for a particular mRNA this can be important. Additionally, SAGE results are somewhat more reproducible and serve as a quality control against which microarray data can be measured.

2.3.3 Comparing SAGE and Microarrays

There are four important distinctions between SAGE and microarrays [PMH02]. Firstly, SAGE does not require any prior knowledge of the sequences being analyzed whereas microarrays use hybridization to known sequences on the microarray chips. Secondly, in SAGE experiments each mRNA sequence undergoes a chemical processing step that increases the number of copies for that mRNA. As a result, very low levels of a particular mRNA can be accurately detected and estimated. Microarrays have no such equivalent way to amplify the quantity of low abundance mRNA sequences. Thirdly, Microarray experiments are much cheaper to perform. So large-scale studies do not typically use SAGE unless a transcriptional profile is needed for poorly characterized genes or species. Fourthly, because SAGE is a well-defined technique one can readily make direct comparisons between SAGE experiments. Microarray experiments are more difficult to directly compare because of the variation in protocols, array design, and probe design. This presents a challenge when attempting to adjust for random and systematic errors since the error sources differ widely.

CHAPTER 3

DATA MANAGEMENT FOR EFFECTIVE QUERYING

Effective querying depends to a considerable extent on the choice of data types, design of the data model, and structure of metadata files. These must be selected carefully for biological databases to provide the most relevant query results to the users. In this chapter we describe what information about microarray experiments is stored and how it is organized. We first identify the important conceptual data types in section 3.1 needed for describing the stored information. Among different implementations there is considerable variation in what data are required for a complete entry. This has resulted in a clear need for standards. One solution is the minimum set of standards MIAME, described in section 3.2. In order to facilitate exchange, suitable metadata structures are required. Section 3.3 describes the two metadata options in common use, XML files and text files. Section 3.4 provides a brief introduction to six important microarray databases as examples of actual implementations. This set of six is referenced later in chapters 5 and 6 also. Tables in this section provide an outline of the data types and data models used by these implementations. The tables illustrate the variation among different microarray databases and reflect the importance of MIAME and MAGE-ML to achieving future standardization.

3.1 Need for Standards in Experiment Submission (MIAME)

Although many significant results have been derived from microarray studies, one limitation has been the lack of standards for presenting and exchanging such data. Minimum Information About a Microarray Experiment (MIAME) describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. It concentrates on defining the content and structure of the necessary information rather than the technical format for capturing it. It is platform-independent but includes essential evidence about how the gene expression level measurements have been obtained.

MIAME is being developed by the Microarray Gene Expression Data society (<http://www.MGED.org>) and is widely regarded as the de facto standard for microarray databases. Most microarray databases either use MIAME already for submissions, and/or data export, or they have plans to do so in their published future goals. To illustrate this point, among the six example microarray databases selected for further discussion, four of them (ArrayExpress, CEBS, GEO, and EMAP) have achieved MIAME compliance and two of them (SMD and HugeIndex) have compliance as a future goal.

3.2 Defining Data Types and Parameters

In this section we identify the data types that are most central to a microarray database. These can be broadly grouped into two classes. First, the signal data is used to identify and quantify gene expression. Second, the descriptive information for the experiment and gene provide a meaningful context for interpreting that data.

3.2.1 Raw and Processed Signal Data Types

There is an important distinction to be made in characterizing the output from a microarray experiment. That distinction is between 1) the raw data files of fluorescent signal images from the microarray sample and 2) the processed signal data, which is a numeric data set from the statistical interpretation of raw data signal values. Raw data are usually generated as 16-bit TIF images one for each color of the fluorescent tags and so two TIF images per microarray. One image represents only one set of conditions and many images result from one experiment. Each image is typically 22-28 MB and as high as 200 MB in size [LGTC04]. Commonly, microarray databases do not store images because of space limitations. The processed signal data are the more valuable of the two because it provides the basis for comparisons in gene expression that answer research question.

The quality and accuracy of signal data are particularly crucial since it is the key source of information for researchers querying microarray databases. That quality and accuracy in turn depends on how carefully the raw data was processed. The best option is to provide the raw data if possible, so that users may reprocess it using newer and more accurate tools and techniques as they become available. The importance of raw image data is universally recognized for quality control even by databases that do not store them. The database ArrayExpress has chosen not to store the raw images, but instead requests that contributors keep them on local servers and provide hyperlinks to the images in their submissions. Although many databases are curated (administrators

monitor the quality of submissions) and basic standards are generally maintained, the responsibility for quality and accuracy of processed data mostly falls to the submitter.

3.2.2 Gene, Array, Experiment, and Sample Data Types

In addition to the signal data from the microarrays themselves, it is essential to provide descriptive data about both the genes that are identified and the experiment for which the microarrays were run. It is this information that creates the context for interpreting the signal data. It is also in choosing and presenting the descriptive data that we find the most variation among database implementations and the type of querying that can be done.

There are many different variations on selecting and organizing microarray experiment information in a database. Figure 4 below [WBBCIM03] shows one possible way to organize and classify data types from microarray experiments.

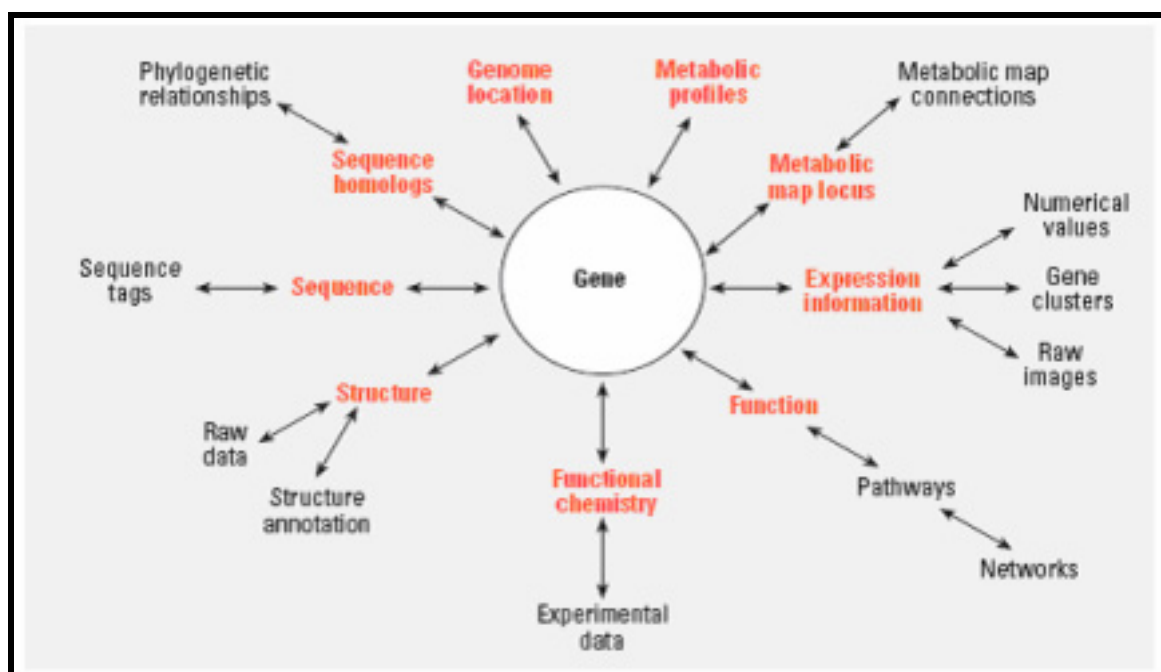


Figure 4: Illustrative diagram for microarray data types

In trying to determine a good illustrative classification, the more authoritative examples will be based on the MGED recommendations for MIAME standards (described further in section 3.2 and appendix B). The following is a simple categorization adapted from MIAME standards and reflects the types of information most important to include in a microarray database: 1) Gene annotation data (information about the gene, often done as either a hyperlink to a genetics database or an import of files from such databases for local use). 2) Sample descriptive data (information regarding the cell type, tissue, organ or species used in the experiment). 3) Experiment design data (including identification of research group conducting the experiment, brand of kit used, and physical conditions affecting the experiment such as temperature and time parameters). 4) Array design data (information regarding the choice of reporter molecules to generate the signal, whether the basic design was for spot arrays that use cDNA or Affymetrix arrays that use the short overlapping fragments of DNA termed oligonucleotides or oligos).

3.3 Metadata Structures for Microarray Data

For biological data in general and microarray data in particular the issues of data management are challenging. The complex data types and highly specialized nature of both the data and the user group make it particularly important to determine the best metadata structures. One of the most important goals for many public microarray databases is the exchange of data. To maintain the greatest distribution and the most complete data collection possible is central to the value of these databases as research tools.

Metadata structures have three important goals. The first is to represent data in a way that is scientifically meaningful and valuable to the researcher making the queries. The second is to represent as much relevant information as possible. The third is to represent the data in a consistent way so that efficient data exchange and comparisons can be made. Although these three are achieved to different degrees among implementations, many have important recognized limitations. In addition, in their documentation many databases list improved data exchange and increased breadth of data types as their future goals. At present there are two alternatives for metadata file format common to microarray databases, customized XML and tab delimited text files. We now briefly look at the characteristics of each.

3.3.1 XML for Microarrays: MAGE-ML

The nature of microarray data and the queries on that data is such that the data are easily represented as objects. Experimental results typically require contextual information in close association with each data entity. Because of this, many microarray databases are most successfully modeled based on object oriented approach. In addition to being suited to object encapsulation, XML is also the standard for web based exchange of data. Public microarray databases are nearly all web based implementations. Therefore, XML as the standard for metadata has consistently been regarded as an ideal candidate for microarray metadata storage and exchange. Published articles dating from 2001 for several microarray databases state use of XML for this purpose as one of their core future goals. As early as 2002 a strong consensus emerged to specifically use MAGE-ML rather than slightly older specifications such as

GEML (gene expression mark up language) and GeneXML. Consistent with this trend, among the six example microarray databases selected for further discussion, four of them (ArrayExpress, CEBS, SMD, and EMAP) store metadata in MAGE-ML format and two of them (GEO and HugeIndex) have storage capability for MAGE-ML as a future goal.

3.3.2 Tab-Delimited Text file

Flat file or tab-delimited text file structures are used in the earlier and more basic implementations. They may be used for submissions to the database, or for download of query results. Flat files offer the advantage of the broadest compatibility across platforms. Unfortunately, that compatibility is offset by the disadvantage that they do not provide a common organization for the data. Further, text files present a more serious limitation, image data from sample annotations and experiment annotations can not be included. Although common organization would simplify information exchange among microarray databases, the limitation regarding image data is such that text files will in future be replaced by XML where possible. Text based metadata persists because many microarray database implementations use Oracle software and are based on a standard relational database model. By contrast, MAGE-ML is based on the object model MAGE-OM from the same standards group. SMD (Stanford Microarray Database) for example, by the year 2005 was able to store data in MAGE-ML format but because it is a relational database rather than object oriented that capability was not easily achieved.

3.4 Example Implementations

In this section we review six example microarray database implementations to determine what approaches are actually in practice for selection of data types and data management. The provided tables cover the three important aspects of data management already discussed earlier in this chapter. First, we consider the experimental data types stored. These include raw image data, other image data, and notes about the stored experiment data. Second, we present the formats used for both upload of submitted data and download. As discussed earlier, the choice of formats is important for the efficiency of data exchange not only between researcher and databases but between different databases.

Table 4: Example Microarray Database Implementations

Database	Description
ArrayExpress	Very large collection. One of the first implementations of a public microarray gene expression data collection.
CEBS	Chemical Effects in Biological Systems (CEBS) knowledge base. CEBS will store data from both microarray and proteomics experiments, it is in the process of being developed. The website is operational at the first of six phase levels.
GEO	Gene expression omnibus. Provided by the NCBI (National Center for Biotechnology information). Expression data repository and online resource for expression data from any organism.
EMAP	Edinburgh mouse atlas: a digital atlas of mouse embryo development and spatially mapped gene expression.
SMD	Stanford Microarray Database. Stores both raw and normalized data from microarray experiments. Organized at Stanford University in California. Primarily stores data from research conducted at Stanford University.
HugeIndex	Human Gene Expression Index, expression levels of human genes in normal tissues.

Third, we look at the data management approaches adopted by these implementations. The tables provide a summary of two levels of data categorization. Table 4 above briefly describes the six databases selected as example implementations.

3.4.1 Data Types Stored

As has been discussed in sections 3.1 and 3.2 the choice of what data to store has some variation. All implementations at the minimum store signal data, some information about the genes identified, and basic information about the research group submitting the results. Table 5 reflects some important differences in the implementations. For example, ArrayExpress (and, until recently, also GEO) chose not to store raw image data.

Table 5: Stored Image Formats in Example Databases

DB name	Raw Images	Other Images	Notes on stored experimental data
Array Express	No	none	Includes experiments for time series responses. Stores data from different species.
GEO	Yes; recent	none	Antibody arrays, tissue arrays, comparative genomic hybridization (arrayCGH), serial analysis of gene expression (SAGE), and mass spectrometry proteomic data.
EMAP	Yes	voxel (See table 3.4.1)	Tissue types and subtypes defined using a standard ontology.
SMD	Yes	archives TIF; has normalized GIF images also	Stores data from different species. Stores both older cDNA arrays and oligonucleotide arrays such as Affymetrix.
Huge Index	Yes	none	Restricted to gene expression in normal human tissue. Serves as a data store for normal controls.
CEBS	yes	none	Includes toxicological and chemical effects data, including phenotypic profiles for chemicals. Has proteomics data.

EMAP uses the concept of a voxel to represent the domain of expression for a particular single gene at a particular stage in mouse embryo development, and saves each domain as an independent 3D (three dimensional) image.

This concept of voxel to represent a simple 3D image data point is familiar to spatial database implementations, but the variation used in EMAP is unique among microarray databases. The presence of SAGE data in GEO is relevant. As discussed in chapter 2 (section 2.3.2) SAGE provides a way to compare and validate microarray results, particularly regarding quantity when expression levels are low.

3.4.2 Data Submission and Download Formats

Data upload is usually for submission of experiment results. There is considerable variation among databases for data uploads. The most common methods to upload data are web based forms, FTP, email. For downloads we see the importance of text files (GEO, HUGOIndex) despite the limitations discussed earlier in section 3.3. MAGE-ML is used by ArrayExpress and CEBS, while EMAP and SMD also use special formats suited to custom graphical visualization tools. Table 6 below outlines the methods for data uploads and downloads among the six example databases.

3.4.3 Data Management Approaches

The two tables below summarize the data abstractions used for data management across the six example microarray database implementations. The approaches used are approximately consistent with the basic outline of MIAME standards. In some cases such as EMAP the graphical nature of the data storage and indexing requires a specialized solution using voxels as described previously in section 3.4.1.

Understanding the approach and terminology for data management will help to clarify how queries are formed when we discuss them in chapter 5.

Table 6: Methods of Upload and Download in Example Databases

DB name	Upload (submissions)	Download
Array Express	Can use MIAMExpress web based tool to generate MAGE-ML format or directly submit in MAGE-ML. May also submit spreadsheets with associated image files. Files sent via FTP or email.	The expression data can be exported as tab-delimited text, and MAGE-ML format. Electronic images available in standard .png and .svg format. Data from ArrayExpress may be exported into Expression Profiler (http://ep.ebi.ac.uk/EP/), an online tool set for gene expression analysis.
GEO	Data options are data table only, or full metadata/data table records. Format options are HTML or SOFT (Simple Omnibus Format in Text).	GEO data are available for bulk download via FTP. GEO DataSets and original records may be downloaded in a custom format (known as Simple Omnibus Format in Text (SOFT). Used to represent and exchange Gene Expression Data.) All records and raw data can be downloaded.
EMAP	Researchers may email or post submissions to their Editorial Office. Or spatially map data using MAPaint software, and submit electronically.	Provides an FTP download area from their resources weblinks: http://genex.hgu.mrc.ac.uk/Resources/intro.html . Software for EMAP / EMAGE may be installed locally for viewing 3D images and mapped gene expression patterns.
SMD	Requires FTP transfer. Accepts proprietary formats from Stanford University Shareware Scanalyze and GenePix for processed microarray data.	All data for an individual microarray can be downloaded. It can be filtered on site first. The online analysis tools provide pattern detection and clustering; their files can be downloaded and viewed in TreeView (Stanford University shareware).
Huge Index	Affymetrix data submissions are accepted by contacting the administrators for checklist of required information and sending it via email (no ftp).	Files are standard tab delimited plain text. From interface may download a list of genes. Can only download the data by performing copy and paste steps. May view and analyze the data using a spreadsheet (e.g. Excel) or comparable program.
CEBS	Set of detailed website forms request information. Raw and processed data files uploaded via webform prompts.	Once an experiment is selected, links are provided for downloading data both as text format and MAGE-ML. Other files may also be present from some experiments (microarray sample files).

Table 7: Metadata Approaches in Example Databases

DB name	Highest level of data abstraction	Second level of data abstraction
Array Express	Experiment. The ArrayExpress documentation defines the experiment as the central high level data type. It consists of one or more hybridizations, and usually a link to a publication.	1) Experiment attributes: parameters of the experiment, laboratory, experiment type, species 2) Array attributes: accession number, manufacturer or lab providing array, array design name 3) Protocol attributes: protocol accession number and type (uses MGED ontology for protocol classification).
GEO	Three upper level relational database entity types: 1) Platform 2) Sample 3) Series The platform and sample data tables are stored as text objects rather than fields in tables, to permit optimal flexibility.	1) Platform: list of elements assayed (cDNA, probe sets, tags) on an array. 2) Sample: references a platform and describes probe signal for each feature (spot or gene) in the array. 3) Series: set of related samples common to an experiment, may include summary data tables and analysis.
EMAP	EMAP uses the concept of a voxel to represent the domain of expression for a particular single gene at a particular stage in mouse embryo development, and saves each domain as an independent 3D (three dimensional) image structure in a separate file. Each submission record corresponds to a single gene.	The experimental results for mapping are usually to a single embryo. Hierarchy follows tissue classification schemas. Each node has component “child” tissues. The end or “leaf” nodes represent the smallest tissue components in the tissue ontology.
SMD	Experiment. For SMD the experiment is the central high level data type.	An experiment is represented by an image of the microarray, experiment category and subcategory, researcher, organism, and links to other databases as sources of annotation.
Huge Index	Uses an object relational model. The current schema has four tables describing the following: 1) Experiments 2) Experimental protocols 3) Expression levels 4) Genes	1) Experiment attributes: source and type of the tissue sample 2) Experimental protocols: data for standard protocols used on the stored experiments 3) Expression levels: processed signal data for each gene studied, and data values for quality of expression-level measurements 4) Gene attributes: organized in rows corresponding to each probe-set on each type of chip, includes data for the transcript targeted by each probe-set

Table 7 – Continued

DB name	Highest level of data abstraction	Second level of data abstraction
CEBS	For Phase I main data concept is the protocol. All data sets (graphs, images, numbers) within CEBS will be linked by reference to an experimental protocol number and its metadata.	Metadata will specify standard operating procedures, observations, and measurements to be recorded. Phase I will include complete sample annotation on a very large range of fields. Domain-specific metadata will introduce experimental data sets in each analytical domain: transcriptomics, toxicology, pathology, etc.
MIAME Standards	1)Image data 2)Expression data 3)Annotation data (for the gene, sample, array, and experiment)	Measurement specifications for raw and normalized signal data, Array design includes reporter probes annotation. Experiment design, Sample description, Hybridization procedures.

3.5 Summary Diagram

The database ArrayExpress has an associated publication [BPSMSV03] that describes the relationships between the databases, the role of MAGE-ML, the use of MIAME standards, and the importance of exchange with external databases. The following diagram summarizes the relationships. Figure 5 represents the ArrayExpress microarray database [BPSMSV03] and helps to illustrate the points made in this chapter. The main implementation is in Oracle. MAGE-ML is the metadata exchange format used for three purposes: 1) to receive data submissions (via MIAMExpress) 2) Exchange information with other databases and 3) export results to external data analysis software. MIAMExpress is an external web based service to help researchers create MAGE-ML files for their submissions. Expression Profiler is web based microarray data analysis software closely integrated with ArrayExpress.

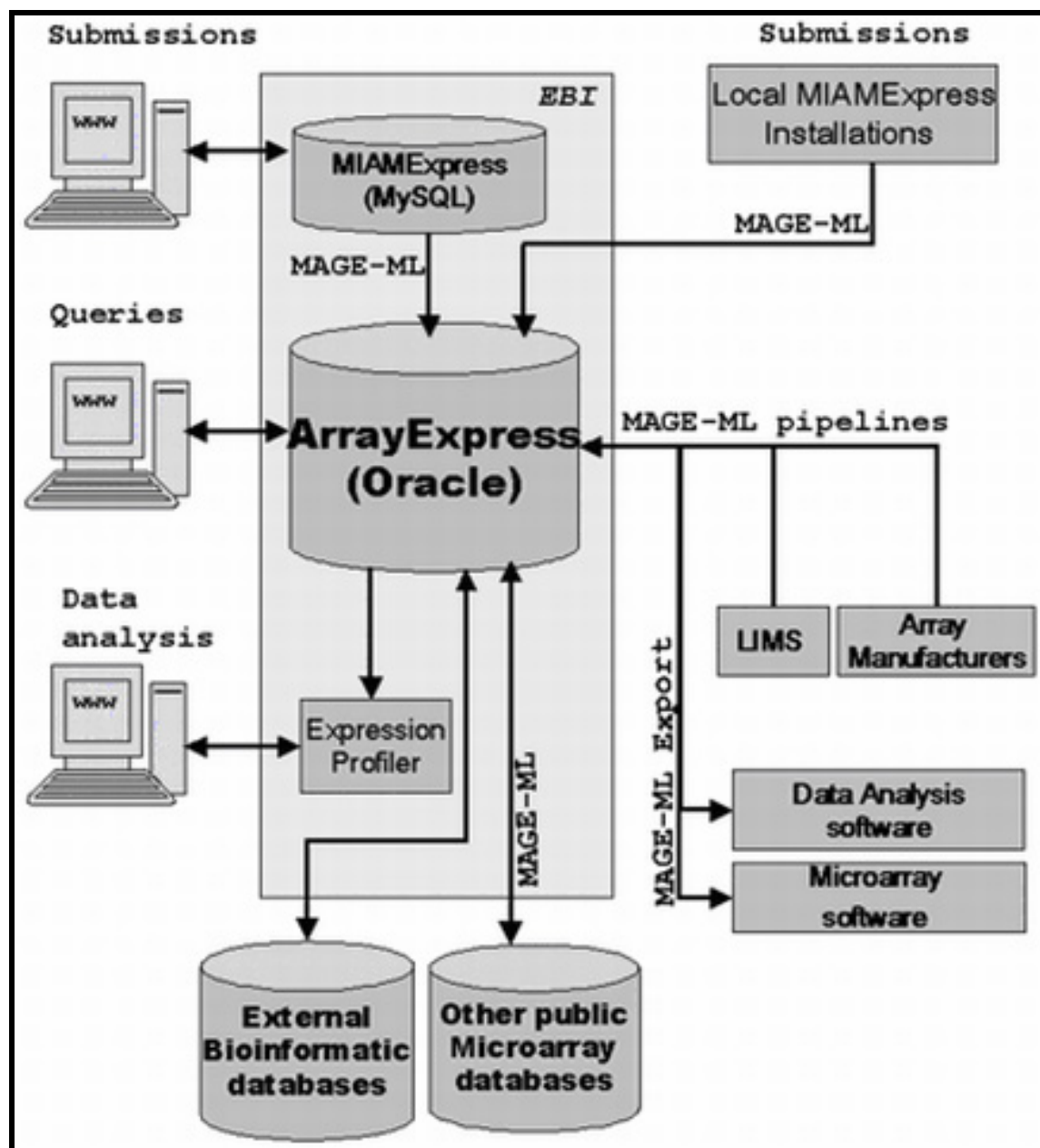


Figure 5: Data management, import, and export in ArrayExpress

CHAPTER 4

OVERVIEW OF MICROARRAY DATABASE INTERFACES

In this chapter we provide an introduction to microarray database interface design. These are general categories which we introduce here for background. Specific examples will be detailed later in chapter 6 with illustrative figures. Section 4.1 describes components used in the interfaces. These may include text boxes, graphical interactive menus or built in tools. Section 4.2 provides examples of the tools used for query results and data visualization. Because microarray data requires visualization to provide the most information from its patterns of gene expression, these tools are important to making the query results informative. Section 4.3 briefly describes the variety of results structures.

4.1 Web Based Interfaces

The web-based interface provides an essential first point of access for the users. The designs are uncluttered and relatively intuitive. Users will typically be able to type simple text entries, select from lists, or navigate through simple graphical objects. In this section we summarize the basic design approaches. There are some limitations to a web based interface which must be considered, we briefly examine these. Locally installed visualization tools provide additional capabilities to overcome them and are discussed in the next section.

4.1.1 Simple Text Box or Web Form

The GEO microarray database provides a good example; it is similar to other NCBI Entrez databases in that both simple and compound query can be achieved using simple Boolean phrases that may be either combined with, or restricted to various supported attribute fields. For example, the query “Type 1 diabetes AND apolipoprotein NOT Homo sapiens” will return all apolipoprotein related gene profiles in Type 1 diabetes-related datasets in all organisms except for human [BSTWNLR05].

4.1.2 Interactive Menus or Graphical Point-and-Click

These options are typically provided as part of navigation through an experiment result set, allowing drill down to specific details on a single data point or ‘feature’ on a microarray. The designs are intended to simplify navigation between levels of detail. The first set of selection options for queries is often on very few parameters, helping the user to narrow down their selection based on general criteria. It is after this initial filtering step that the user is then able to select the best result candidate from a summary list. Each item on the list could be a link to an individual candidate gene in answer to a query, or to annotation and visual data as noted in section 3.3.4 below.

4.1.3 Built in Tools

Many databases provide specialized tools available from within their website for both the analysis and display of microarray data. Initial queries are typically met with a list from which the user may make selections for which he or she wishes to see a detailed result. Since the result may include a large number of simple data points or values, these tools usually provide charts or graphs either embedded within the same

page or as a separate pop-up display. The data provided often summarizes microarray signal intensities, or shows other types of simple pattern distribution data such as location of gene expression in a cross section of tissue. Such tools are typically simple in their function and often implemented as Java applets, or Perl scripts.

4.2 Software Tools for Results Visualization

Beyond the tools that enhance data visualization within a browser, many microarray databases either provide software programs and packages to be installed on a local machine or recommend open source programs for visualizing results. Such programs have been developed for the purposes of microarray data analysis and display. Since visualization of the patterns is important to interpretation, these visualization tools are a research necessity and their importance in making full use of the information provided through the databases should not be under stressed.

4.2.1 Clustering Analysis

A natural basis for organizing gene expression data is to group together genes with similar patterns of expression. For any series of measurements, a number of measures of similarity in the behavior of two genes can be used, such as the Euclidean distance, angle, or dot products of the two n -dimensional vectors representing a series of n measurements. The standard correlation coefficient (i.e., the dot product of two normalized vectors) conforms well to the basic biological definition for two genes to be “coexpressed” (expressed at the same time in the same cell). There are several tools available as open source platform-independent software to perform clustering analysis on the microarray data query results including Magictool

(<http://www.bio.davidson.edu/projects/magic/magic.html>), and TreeView (<http://rana.lbl.gov/EisenSoftware.htm>). Additionally, some implementations include their own customized tools for clustering analysis and other types of data analysis.

Figure 6: MagicTool clustering analysis dendrogram of query output

4.2.2 EMAGE for EMAP

These java tools have been developed as part of the NIH funded Electronic Atlas of the Developing Human Brain project. The tools replace and extend some of the C/X11/Motif tools developed as part of the Mouse Atlas project and use the same underlying image processing libraries. Each of these has been tested under MS Windows, Solaris, Linux (Mandrake, Redhat) and Mac OSX.

- **JatlasViewer:** A 3D volume browser providing section and 3D visualization. The volume data can include segmentations, for example labeled anatomy. The viewer provides feedback and navigation through the anatomical nomenclature. Example data-sets for human and mouse embryo.
- **Jconvert:** An image format converter for generating volume data suitable for the Atlas Viewers. This allows conversion from a range of 2D and 3D image formats to the native Woolz format used in the MRC software.
- **Gene Expression Viewer:** A prototype viewer for gene-expression data that has been mapped onto a reference volume, for example as obtained by query on the EMAGE database.
- **JReconstruct:** A java version of the Reconstruct program. This implementation is incomplete and currently only allows simple re-stacking of 2D image files. Work in progress.
- **Jwarp Tools:** to support 3D warping of data by defined a series of “tie-points” or correspondences between two data volumes.

4.2.3 Treemaps

Treemaps are a relatively new data visualization technique developed in the 1990's for space-constrained display of hierarchies. It is unrelated to the branching tree menus familiar to web browsers. Treemaps are a space-filling map of differentially colored squares, clicking on a square permits drill down to the next level detail on the

data. The technique is used as a Microsoft Excel plugin and as of 2005 was being evaluated for Oracle database user interfaces. They are well suited to microarrays and allow users to view and query the data from an experiment on a single computer monitor screen.

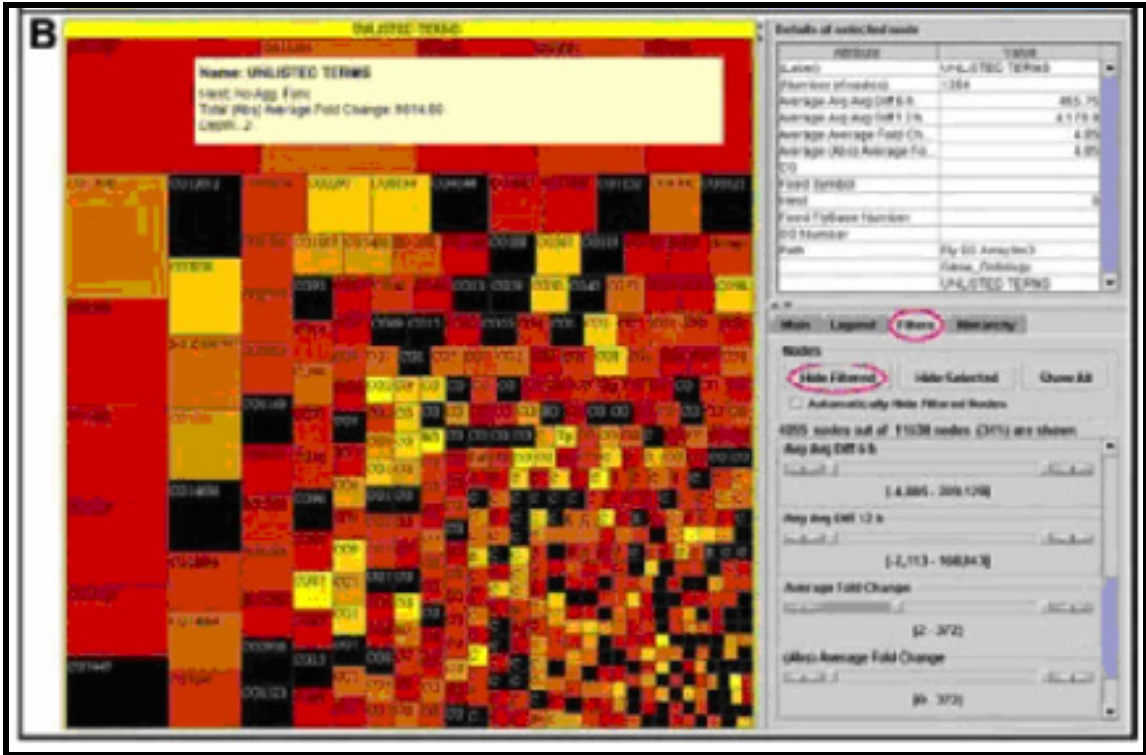


Figure 7: Treemaps rapidly identify genes of interest

In figure 7 users apply “filters” in the control panel to select genes based on specific quantitative attributes [BDBS04]. In this example, the “average fold-change” slider was moved to include values greater than two. Other filters not of interest can be turned off (image marked with circles on the filter options menu pane).

4.2.4 List of Visualization Tools

Many query results are either graphical data or large numeric sets requiring data analysis and visualization as maps, graphs, or charts. To illustrate the importance of having additional tools available to display results sets, the following table provides a summary of customized tools provided by several important microarray database implementations.

Table 8: Customized Tools for Microarray Data Analysis and Visualization

Microarray Database	Provided Tools integrated with the web-based interface	Description
ArrayExpress	<ul style="list-style-type: none">○ ArrayExpress is an implementation at the EMBL-EBI (European Bioinformatics institute)○ ArrayExpress recommends use of Expression Profiler, also developed by and available at EMBL-EBI	Expression Profiler is an open, extensible web-based collaborative platform for microarray gene expression, sequence and PPI data analysis, exposing distinct chainable components for clustering, pattern discovery, statistics (thru R), machine-learning algorithms and visualization.
GEO	GEO is an implementation within NCBI, and GEO Profiles are fully integrated with other NCBI Entrez databases such as GenBank, PubMed, Gene, UniGene, OMIM, Homologene, SNP, Taxonomy, SAGEMap and MapViewer. These databases in turn provide visualization tools and annotation information.	GEO uses the NCBI Entrez query system. Entrez has many features including cluster heat maps, query subsets profiles (for example, a user can locate gene expression levels 10-fold higher in time point 'A' than in time point 'B'). GEO BLAST database contains all GenBank sequences represented in GEO DataSets and uses NCBI's BLAST interface.
EMAP	<ul style="list-style-type: none">○ JAtlasViewer○ Jconvert○ GeneExpressionViewer○ JReconstruct○ Jwarp	Visualization tools for 2D and 3D images. Require Java environment, may be installed on MacOS, MSWindows, or Linux.

Table 8 – Continued

SMD	<ul style="list-style-type: none">○ TreeView compatible downloadable files are provided for use with a separate recommended TreeView tool.○ Clustering tools○ Pattern analysis maps○ Data quality tools	<p>Clustering and pattern analysis maps were included in the first implementation (2001).</p> <p>More recent tools have been included for assessing data quality and analysis</p>
HugeIndex	<ul style="list-style-type: none">○ Expression information display tool○ Comparison tool○ Interactive scatter plot tool	<p>The current release of the HugeIndex provides three tools with which to access and visualize the data stored within our database. These tools can display expression information about specific genes or compare multiple tissues or experiments using Boolean operators or interactive scatter plots.</p>
CEBS	<ul style="list-style-type: none">○ Data Preprocessing○ Data Comparison○ Data Visualization○ Identification of Differentially Expressed Genes○ Gene Category Analysis by BioCarta Pathways○ Gene Category Analysis by KEGG Pathways○ Gene Category Analysis by Gene Ontology (GO)	<p>Analysis tools for microarray signal data and gene expression maps; tools are available online through the web site.</p>

4.3 Results Structures

These examples are a brief introduction the results structures for microarray data queries in web-based interfaces. Separate software programs are available for more advanced types of visualization including multidimensional representations. The

software may be integrated into the website or in some cases are separate downloadable installations. For many implementations results are simply text, tables, or simple images. Chapter 6 will provide detailed illustrations of interfaces in current implementations.

Textual or Annotated: This is the most basic interface type. Data types such as author, accession number, or other simple typed keywords are used to browse records. Boxes accepting free text are often restricted by ontology such that only certain words will be recognized. Misspelled words, or alternate terminology typically yield no results as spellcheckers and prompts for similar words are not included in most implementations.

Tabular: Results may be presented as a list containing all relevant genes matching a query. Data may be provided as a tab delimited downloadable text file, correlating each feature (positive signal on the microarray) to one row.

Images: Thumbnail images may be provided for browsing, clicking on the thumbnail will then bring up the complete microarray. Cibex uses this type of navigation, as it facilitates quick comparison by simple visual inspection.

Navigable Menu: A list may be provided with either the name or other brief information that the viewer can click and navigate to drill deeper for information. Some interfaces provide a display of the microarray grid on which a spot may be clicked to retrieve details.

Graphical Statistics Chart: Visualization of the array data may be presented as simple line plots, or more detailed graphs. Intensity of expression may be provided as a

histogram and line chart for each data point, or a scatter plot analysis for a set of points as in the database Cibex.

CHAPTER 5

MICROARRAY QUERIES IN RESEACH STUDIES

In this chapter we present a general introduction to the role of gene expression queries in biological research. Section 5.1 examines in general how querying microarray databases provides information to help support or refute genetics research hypotheses. Using references to actual published work, sections 5.2 through 5.6 briefly describe each of five study areas within genetics. These are organized as general categories of research to demonstrate the range of genetics studies that benefit from querying microarray databases. Many different databases and many different queries may be used for each study area. Because the queries are implemented in part as predefined field selections to navigate through the data it would not be meaningful to classify the actual queries. Instead, the data types and data models covered in chapter 3 would serve a good basis for understanding the query constructs (chapter 6) used to search microarray data.

5.1 Role of Querying Microarrays for Research

There are several important areas of biological research that directly benefit from querying microarray databases. For a researcher studying genetics, the hypothesis under investigation is usually not answered directly by placing a single query on the microarray database. Instead, the hypothesis will be partly supported or partly refuted

based on patterns and correlations in genetic expression profiles. These patterns are found by analyzing the results of queries on specific genes and particular experiments relating to the hypothesis under investigation. Therefore, to the non-researcher a microarray database may be an inappropriate resource to answer general questions about gene expression. As with other bioinformatics databases, specialized knowledge and background is needed to effectively use microarray databases as a research tool. The brief example queries provided in this chapter are intended to be illustrative of how the interfaces have been implemented. The options chosen for query constructs and utility of the results generated may be difficult to assess for a non-researcher. We postpone evaluation and identification of limitations until chapter 7.

When a researcher begins querying a microarray database there is usually some degree of navigation through the data, and exploration of the question before arriving at a result. It is important to note that many biological research databases use simple selection menus and process the user choices through predefined queries in the system. Predefined queries are in fact the preferred and recommended method for microarray databases. The process of making the query is therefore that of an interactive session in which the researcher is guided towards the area of information that he or she is most interested in. How that data is presented becomes an extremely important component of how well the query was answered. Presentation is essential to quality of information for these types of specialized biological data. This is in contrast to traditional relational databases where results are simple clearly defined sets of information and data values typically atomic, rather than dependent on each other as part of a large subtle pattern in

need of analysis and interpretation. It is important to keep these distinctions and special properties of microarray databases in mind when viewing the example queries.

Broadly, research through microarray databases focuses queries to determine two kinds of information. First, to learn what is the normal expression pattern to use as a baseline. The correct baseline is important for meaningful comparisons to identify subtle changes. Second, what are the changes to the normal expression patterns for a given set of conditions and particular sample type (the experimental design)? The changes may be associated either with exposure to a toxin, a specific disease state, or a particular genetic variation. Genetic variations or *mutations* may offer either neutral, harmful, or beneficial effects to the organism. The term mutation in a biological research context does not imply only a negative effect. Recall that some mutations in hybrid plants improve yield and identification of those mechanisms is also an area of research.

5.2 Queries for Coexpression Studies

Coexpression is the simultaneous expression of different genes in the same cell or the same tissue site. There is a large area of research into the possible associations between coexpressed genes and common roles for those gene products, which would allow predictions of function and help in gene identification. This type of predictive research also focuses on gene proximity as noted below in 5.1.3. In the study *Coexpression Analysis of Human Genes Across Many Microarray Data Sets* [LHSQP04] examine a network of 8805 genes connected by 9.7 million coexpression links using the Stanford Microarray Database (SMD) and Gene Expression Omnibus

(GEO). From querying these databases they were able to identify 220,649 (or nearly 2.2%) of the coexpression links to be present in at least three data sets. Their findings illustrate the use of coexpression studies in the discovery functionally related groups of genes, “We show that confirmation of coexpression in multiple data sets is correlated with functional relatedness, and show how cluster analysis of the network can reveal functionally coherent groups of genes. Our findings demonstrate how the large body of accumulated microarray data can be exploited to increase the reliability of inferences about gene function.”

5.3 Queries for Gene Proximity Studies

In bacteria, clusters of genes with related function such as for the same metabolic pathway are often grouped together physically. They are encoded close together in the same subsection of a chromosome. There are ongoing studies to find similar relationships in higher organisms. This type of predictive research also uses coexpression as noted above in 5.1.2. Querying microarray databases provides data to correlate whether these grouped genes share similar expression patterns or similar function. For example, in the research article *An Abundance of Bidirectional Promoters in the Human Genome* [TAHSOM04] the authors were able to identify a class of gene pairs in which the transcription start positions are extremely close (less than 1000 base pairs) and positioned on opposite strands of the DNA (bidirectional). This discovery represents more than 10% of genes in the human genome, a surprisingly high percentage. An important next question in the article is whether the transcript levels in a bidirectional gene pair are coordinately regulated. To test that hypothesis the Stanford

Microarray Database (SMD) was queried. The results sets positively correlated 17% of the gene pairs as coordinately regulated at a statistically significant level. The results demonstrate that bidirectional arrangement is an important mechanism for expression regulation for a significant percentage of human genes.

5.4 Queries for Tissue Localization Studies

Many important questions relate to tissue localization. The expression of a particular gene may be normal in some tissues, but indicate a disease state or metabolic problem in others. It should be noted that tissue localization differs from gene proximity. Whereas gene proximity is in terms of position of the encoded gene on the chromosome, tissue localization refers to which genes are being transcribed into mRNA and protein in particular cells or tissues. The presence of certain combinations of expressed genes in a particular tissue may indicate increased risk for diseases or reactions to treatment. Being able to distinguish the patterns of gene expression specific to different tissues and also different to specific types of cells is under active research. For example, in the article *Microarray Technology: A Review of New Strategies to Discover Candidate Vulnerability Genes in Psychiatric Disorders* [BBVTLE03] the authors describe how gene location guides research in brain function, “a gene that is neuronal and is primarily expressed in the extended amygdala may lead us to hypotheses about a role in emotional reactivity, whereas a gene that is present at all synapses may lead to hypotheses relating to signaling or synaptic plasticity. The neuronal site expression pattern will then guide the choice of animal models to be pursued. For example, a gene highly expressed in emotional circuits will call for studies

using models of anxiety-like behaviors (the elevated-plus maze, light-dark box, fear conditioning). By contrast, a gene highly expressed in the hippocampus might suggest a possible role in learning and memory (e.g., the radialarm maze or Morris water maze).” The tissue and cell type localization of genes also allows precise development of animal models in which a particular cell type is genetically modified either by addition of an altered gene (transgenic models) or deletion of a gene (knockout models), both of which can provide important confirmation of the metabolic role for candidate genes indicated by microarray database queries.

5.5 Queries for Toxicity Evaluation Studies

Toxicogenomics studies evaluate the harmful effects of exposing cells or tissues to chemical compounds. This is a new approach to predicting and regulating many types of chemical exposure by examining the impact to gene expression profiles. These include evaluating the safety of new pharmaceutical drugs during the design stage, predicting drug interactions that may be harmful to a patient, and determining unsafe levels of trace contaminants in food, and water. This last purpose can be used to set federal standards and protect public health. The long term harmful effects of chemicals are also economically important to help manufacturers avoid expensive and unnecessary development of products that would later fail safety assessments. These same points are discussed in research articles such as *The Use of Toxicogenomic Data in Risk Assessment: A Regulatory Perspective* [CT05]. In addition, the authors note that although currently several governmental and commercial organizations are actively building toxicogenomics databases, “regulatory use of the toxicogenomic databases as

supportive information in the assessment procedure of new drug applications will be on a case-by-case basis until the predictive value of the databases is firmly established”. Toxicogenomics has a unique breadth of application because the information impacts commercial industries and may serve as the legal basis for mandating federal regulations. By contrast, most genomics research is limited to either improving general knowledge of molecular biology or has applications to medicine.

5.6 Queries for Data Mining Studies

Data mining is the detection of interesting patterns in large data sets. Therefore data mining techniques are commonly used with microarray databases. There are two closely associated versions of this approach used with expression data. One version is *clustering* also called cluster analysis, the automated algorithmic generation of dendrograms based on degree of similarity in expression patterns for a large set of genes. The other version is *class discovery and class prediction* which uses cluster analysis data. These are well established areas of research, and it has been noted that “cluster analysis has been a standard approach to microarray data since the beginnings of microarray technology and is the basis of most class discovery efforts.” [O03]. Members of the class may appear in different tissues but share an expression pattern, that pattern is then discovered or predicted to be associated with the same metabolic pathway or the same specific disease state. This helps to identify which genes work together. Cancer researchers use microarray cluster analysis to classify tumors and target treatments to each tumor class. Cluster analysis can help identify particular cancers in individuals. For example, in the research article *Class Discovery Analysis of*

the Lung Cancer Gene Expression Data [P04] public data is used to build a new molecular classification which they analyze with a new cluster analysis algorithm. The authors state that their analysis “reveals many additional details and subtypes of previously defined types of lung cancer. Large histological cancer types can be further divided into subclasses with different patterns of gene expression. These subtypes should be taken into account in diagnostics, drug testing, and treatment development for lung cancer patients”.

CHAPTER 6

QUERY INTERFACES AND EXAMPLE QUERIES

This chapter reviews the query interfaces of six example microarray databases. Example queries and results are shown to illustrate how users search the database for gene expression data. The documentation for these databases does not detail indexing structures, but details regarding data types have been described in chapters 3. Similarly, the query processing is typically transparent to the user and not described in documentation since it is not directly related to aspects of bioinformatics data. We therefore consider the queries from the user perspective in these examples, since that best reflects the bioinformatics nature of the queries and usability of microarray database implementations.

6.1 Querying the ArrayExpress Database

The ArrayExpress database is administered as part of the European Bioinformatics Institute. The data warehouse is based on the BioMart open source federated query architecture [DMKDDBH05]. BioMart supports queries on gene attributes and sample properties. The query interface presents a simple combination of text boxes and pull-down menu selection lists for combining parameters. The pull-down menus are provided where a limited controlled vocabulary is needed for effective searching. There are three main sections of the database to query, each with its own subsection of the interface. The three are experiments, arrays, and protocols.

Back Search Favorites

Address: <http://www.ebi.ac.uk/arrayexpress/query/entry>

Google query arrayExpress Search PageRank Popups okay ABC Check

EMBL-EBI
European Bioinformatics Institute

ArrayExpress

You are logged in as guest [Login](#) » **ArrayExpress** (960 Experiments with 27442 Hyds

Query for Experiments

Give an experiment accession number for example E-MEXP-2, [Query](#) »

or fill out some of the following fields to get a list of matching experiments:

Species	Author	Array accession number
« any species »	<input type="text"/>	<input type="text"/>
Experiment type	Laboratory	Array design name
« any type »	<input type="text"/>	<input type="text"/>
Experimental Factors	Publication	Array provider
« any factor »	« don't specify »	<input type="text"/>

Description contains the word

Query for Arrays

Give an array accession number for example A-TIGR-32, [Query](#) »

or fill out some of the following fields to get a list of matching arrays:

Array design name	Array provider
<input type="text"/>	<input type="text"/>

Query for Protocols

Give a protocol accession number for example P-SNGR-8, [Query](#) »

or fill out the following field to get a list of protocols of the given type:

Protocol type
« any type »

Figure 8: The main interface for querying ArrayExpress

The experiment section provides information about both experimental factors and the actual data. The array section provides information relating to array design. The protocol section is a simple look up of protocols based on accession number and type. Figure 8 illustrates the main query interface for ArrayExpress. It should be noted that ArrayExpress does not provide integrated tools for visualization or comparison.

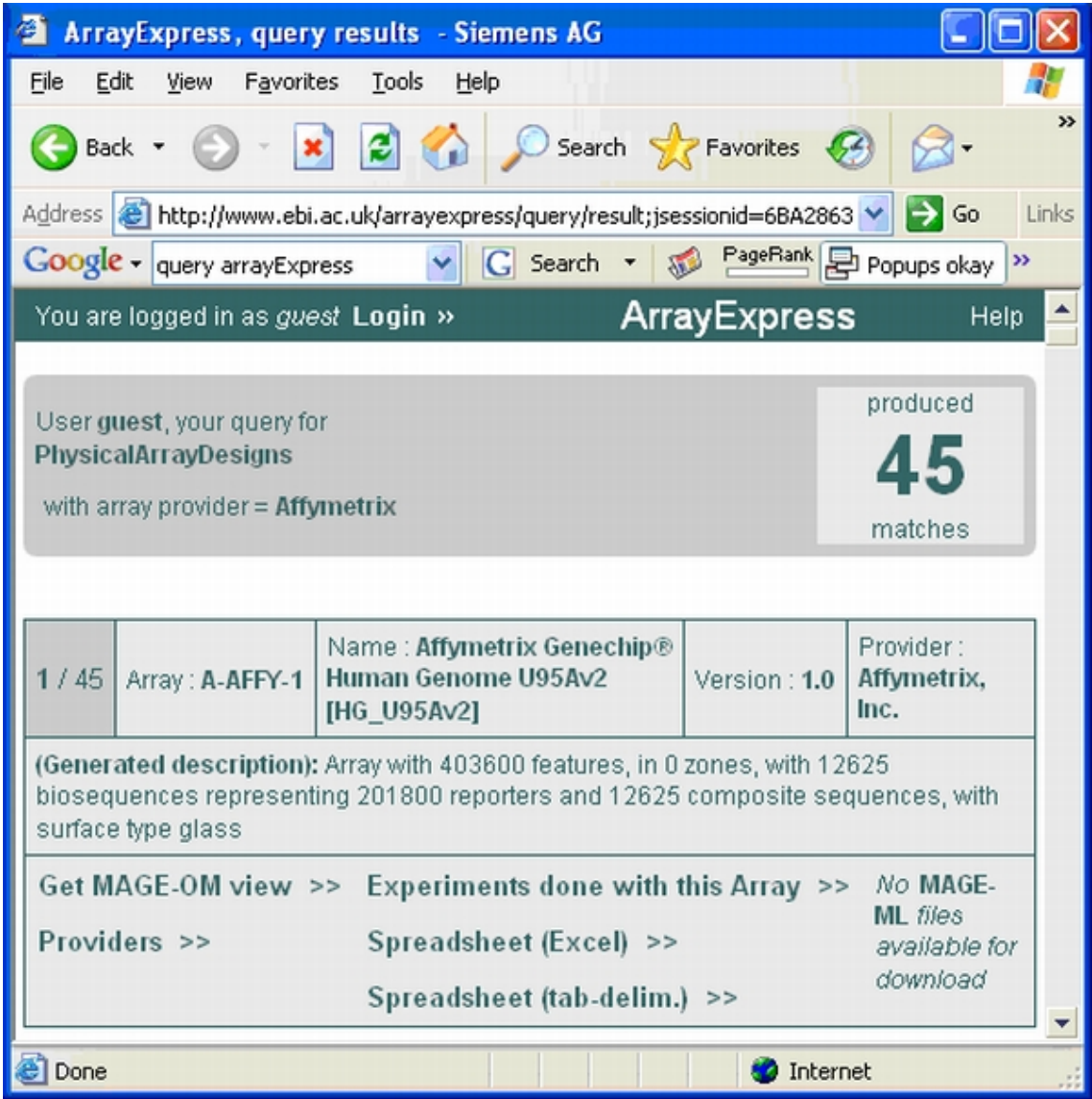


Figure 9: Results of query for “array provider = affymetrix

It is a searchable database for downloading data files of the query results. Query results from any of the sections provide links to the other two sections. The components of the database are easily accessible and navigable for the user. For example, a result set containing a particular array design will include links in the query result view to any experiment utilizing that array design.

We now present example query 1: Find all results using arrays from the provider (manufacturer) Affymetrix. Results: 45 matches in the database, with navigable linked summaries presented for each. In figure 9 we see the results set for the query “array provider = Affymetrix”. Note the first summary of all 45 matches is displayed. The summary includes navigable links to permit data download or to retrieve the list of “Experiments done with this Array”. Example query 2: Find all results for aging studies involving the species chimpanzee (select Latin name *Pan troglodytes* from pull-down menu). Results set: 1 result in database.

Selecting the MAGE-OM view from the menu in figure 10 retrieves a textual list of experimental parameters representing the experiment description as meets minimum requirements. Numeric lists of the samples used are provided as downloadable spreadsheets. Data may be downloaded in MAGE-ML (XML for microarray data) on the labeled link as a zip file. The biosamples are also provided as a graphics file for download. The ArrayExpress implementation does not allow navigation within the browser and the sample file is too large to be readable if opened in the browser as shown by figure 11 below.

ArrayExpress, query results

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Reload Print

Address <http://www.ebi.ac.uk/arrayexpress/query/result;jsessionid=8EFC6A4DEA607D> Go Links

Google query arrayExpress Search PageRank Popups okay

You are logged in as guest [Login >>](#) **ArrayExpress** [Help](#)

User **guest**, your query for Experiments
with experiment type = aging

produced
1
match

1 / 1	Experiment : E-TABM-25	Submitter(s) : Khaitovich	Lab : Evolutionary Genetics
-------	------------------------	---------------------------	-----------------------------

Experiment Design Type : organism status , aging

(Generated description): Experiment with 18 hybridizations, using 7 samples of species [Pan troglodytes], using 18 arrays of array design [Affymetrix Genechip® Human Genome U95Av2 [HG_U95Av2]], producing 18 raw data files and 0 transformed and/or normalized data files.

(Submitter's description 1): An experiment was performed to study effects of aging on gene expression in chimpanzee brains.

[Retrieve data >>](#) [Experimental protocols >>](#) [Get MAGE-OM view >>](#)

[Providers >>](#) [Array design used >>](#)

[Bibliographic references >>](#) [Samples >>](#)

- Experiment's directory in the FTP >>
- MAGE-ML : (.gz (1030 KB))
- Biosamples : (.png .svg .xls)

Internet

Figure 10: The results set of the ArrayExpress query “aging studies”

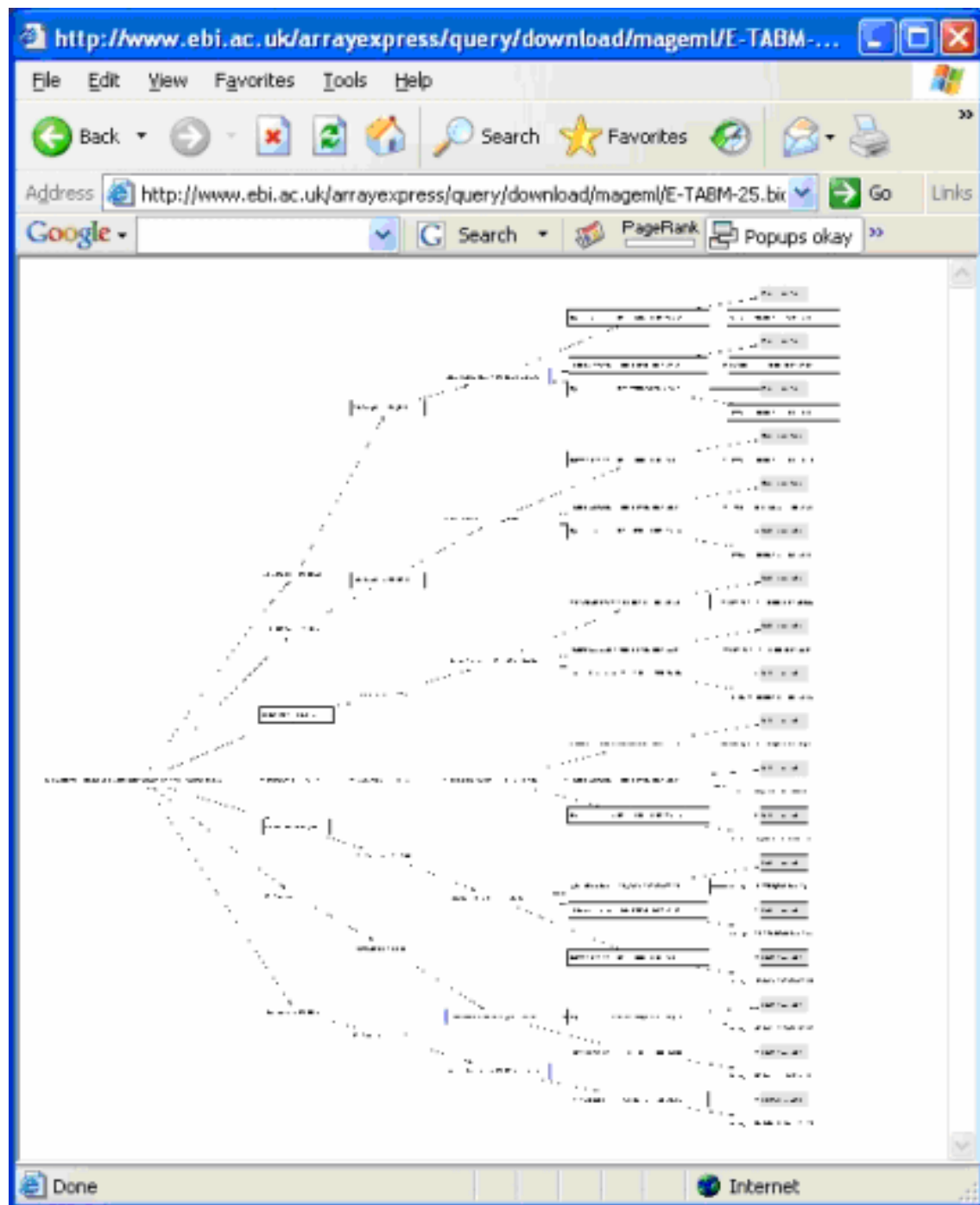


Figure 11: Mage-ML of a results set

Visualizing the data is done in external software, by clicking the Export data tab in the results summary the database presents either the option to see the data matrix as a flat

text file of numeric codes or upload to another external database Expression Profiler for visualization. In figure 12 below, The ArrayExpress database offers the option through a link to an external software package Expression Profiler for data visualization.

In figure 13 Selecting the “see data matrix” option results in a simple textual summary of data. The data requires separate software packages for meaningful patterns

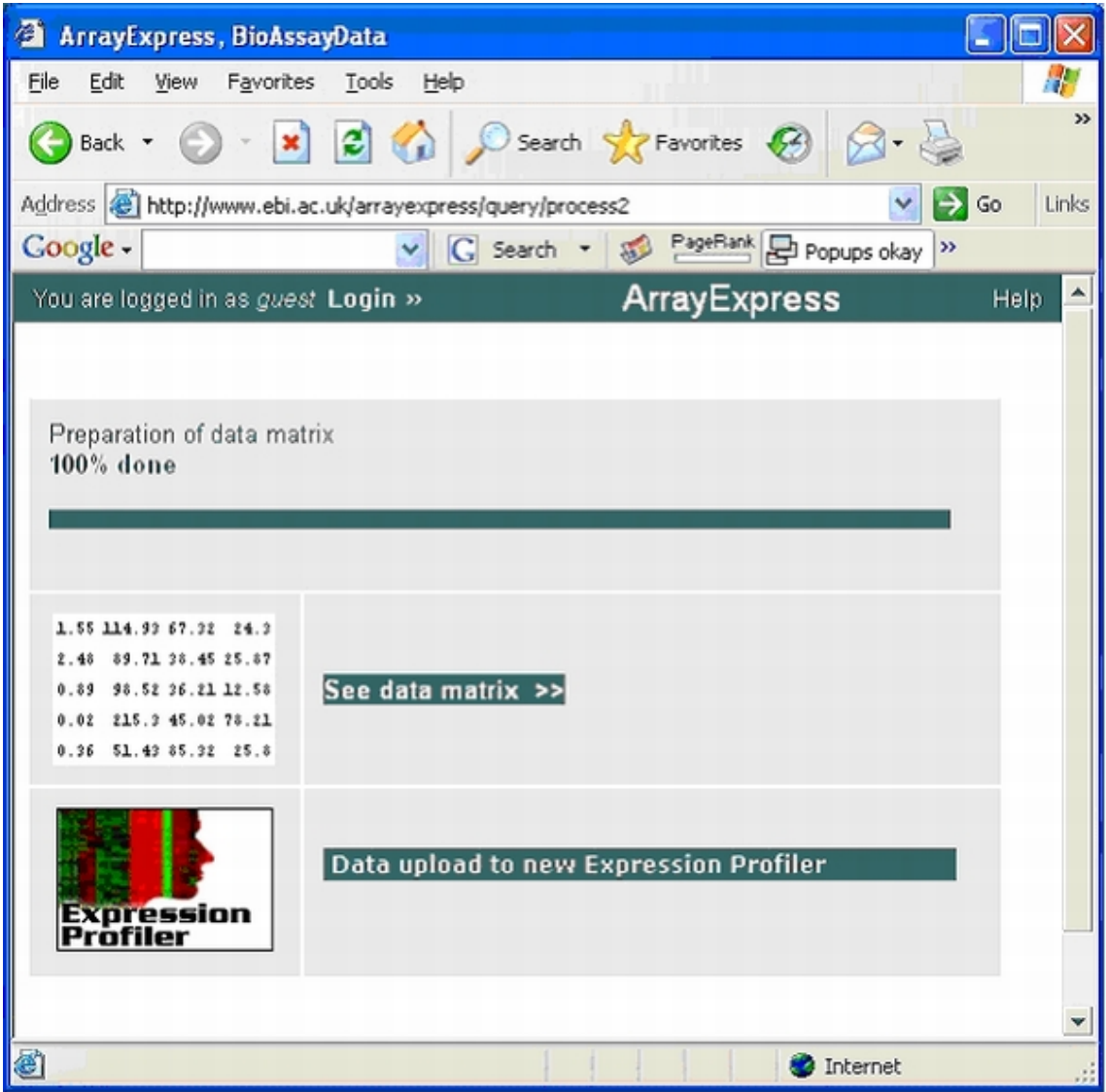


Figure 12: ArrayExpress links to external software

to be found. Efficient direct inspection of the thousands of numeric entries per result set is not possible. Certainly the significant patterns in the results set become evident only when some type of graphical representation is employed.

Affymetrix:Reporter:HG_U95Av2:1808_s_at:ProbePair13	6427735
Affymetrix:Reporter:HG_U95Av2:38598_at:ProbePair3	5947615
Affymetrix:Reporter:HG_U95Av2:1808_s_at:ProbePair13	6427737
Affymetrix:Reporter:HG_U95Av2:1849_s_at:ProbePair0	6424611
Affymetrix:Reporter:HG_U95Av2:1849_s_at:ProbePair0	6424613
Affymetrix:Reporter:HG_U95Av2:35135_at:ProbePair5	6072035
Affymetrix:Reporter:HG_U95Av2:35135_at:ProbePair5	6072037
Affymetrix:Reporter:HG_U95Av2:33865_at:ProbePair14	6232343
Affymetrix:Reporter:HG_U95Av2:33865_at:ProbePair14	6232345
Affymetrix:Reporter:HG_U95Av2:33936_at:ProbePair8	6236859
Affymetrix:Reporter:HG_U95Av2:33936_at:ProbePair8	6236861
Affymetrix:Reporter:HG_U95Av2:33189_at:ProbePair4	6407335
Affymetrix:Reporter:HG_U95Av2:38598_at:ProbePair3	5947617
Affymetrix:Reporter:HG_U95Av2:33189_at:ProbePair4	6407337
Affymetrix:Reporter:HG_U95Av2:37734_at:ProbePair10	6303439
Affymetrix:Reporter:HG_U95Av2:37734_at:ProbePair10	6303441
Affymetrix:Reporter:HG_U95Av2:41399_at:ProbePair2	6004147
Affymetrix:Reporter:HG_U95Av2:41399_at:ProbePair2	6004149
Affymetrix:Reporter:HG_U95Av2:31815_at:ProbePair8	6020221

Figure 13: Example textual data matrix in ArrayExpress

6.2 Querying the CEBS Database

The CEBS microarray database is in the first phase of a 10-year long multiphase development. The queries available at the time of this writing are relatively basic compared with the long-term goals of the designers. In its final form it will be possible to query CEBS by molecular chemical constructs so that an unknown compound can be

entered as a query and similar compounds are matched and retrieved. The genetic effects of the related compounds will then be presented based on microarray research findings and provide insight into the predicted behavior of the unknown compound. In common with the other microarray databases in this chapter, CEBS provides a basic query interface with a set of pull down menu selections for a basic guided search.

CEBS Microarray Search for Multi-Comparison - Siemens AG

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address http://cebs.niehs.nih.gov/microarray/manager

Google database search microarray Search PageRank 162 blocked ABC Check AutoLink

CEBS
CHEMICAL
EFFECTS in
BIOLOGICAL
SYSTEMS

NCT National Center for Toxicogenomics
BIOINFORMATICS to
KNOWLEDGE

Experiment Search

Microarray Home
Submit Experiment
Search and Analyze
Documentation Center

Experiment ID

Investigator's Name

Experiment Title

Tissue Name

Species

Image Processing Software

Figure 14: Query interface for CEBS microarray database

The interface is provided in figure 14 above and shows the six main types of data for use in basic Boolean query constructs. Query results are returned as a list of record summaries that can be selected and viewed in detail. Here is an example query: find experiments involving the species “mouse” and the tissue “forebrain” by selecting these two from the menu. In the result one record is returned, it is shown below in figure 15. This example record represents a brief view; a full view would fill multiple pages. In future implementations the extensive annotation data will be directly searchable.

The screenshot shows a web browser window titled "CEBS Experiment Selection - Siemens AG". The address bar displays "http://cebs.niehs.nih.gov/microarray/manager". The page features a header with the CEBS logo and the text "CHEMICAL EFFECTS in BIOLOGICAL SYSTEMS". A navigation menu on the left includes links for "Microarray Home", "Submit Experiment", "Search and Analyze", and "Documentation Center". The main content area is titled "List of Experiments for Selection" and includes a message: "The experiment search returns 1 record(s). Please use check boxes below to select experiment(s), then click on 'View Details about Selected Experiment(s)' for Analysis as well as Experiment Report." Below this message is a table with the following data:

Select	Experiment ID	Investigator	Experiment Title	Image Processing Software	Publication	Visibility	Array Design (ID)
<input type="checkbox"/>	527402005	Robert Williams	Mouse GTL strains - forebrain	Affymetrix		Williams, Public	MG_U74Av2 (11)

At the bottom of the table, there are buttons for "Reset" and "View Details about Selected".

Figure 15: Query results in CEBS for “mouse and forebrain”

Queries based on comparison between different arrays may be done through the visualization and analysis tools provided with CEBS. After an initial query to locate and select an experiment such as the one above in figure 15, a series of interfaces provides options for selection filters on the experimental data. The user is then provided with a list of all arrays from the chosen experiment and using radio buttons may select which ones to use for the comparison, as illustrated in figure 16. The comparisons provide information equivalent to new queries. A further set of menus will

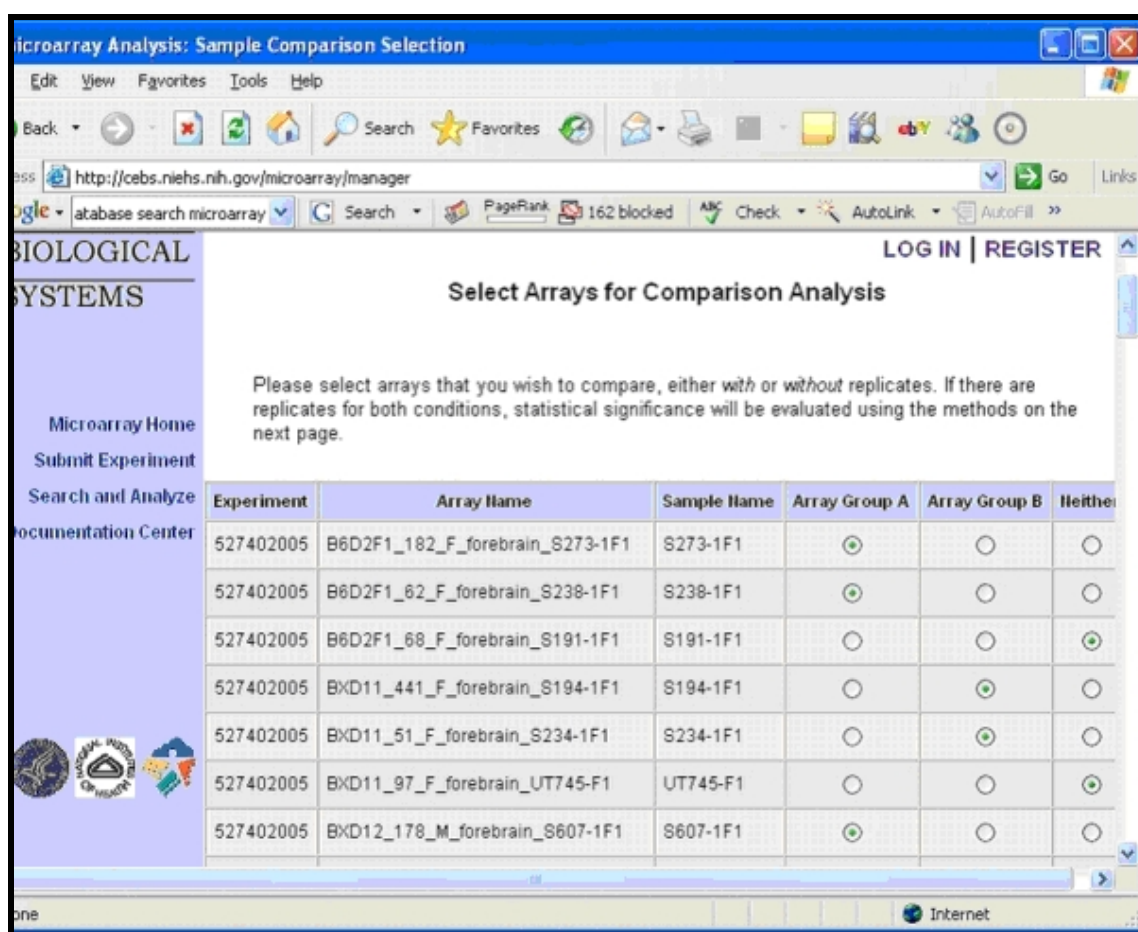


Figure 16: Navigation menus allow the user to select arrays for comparison

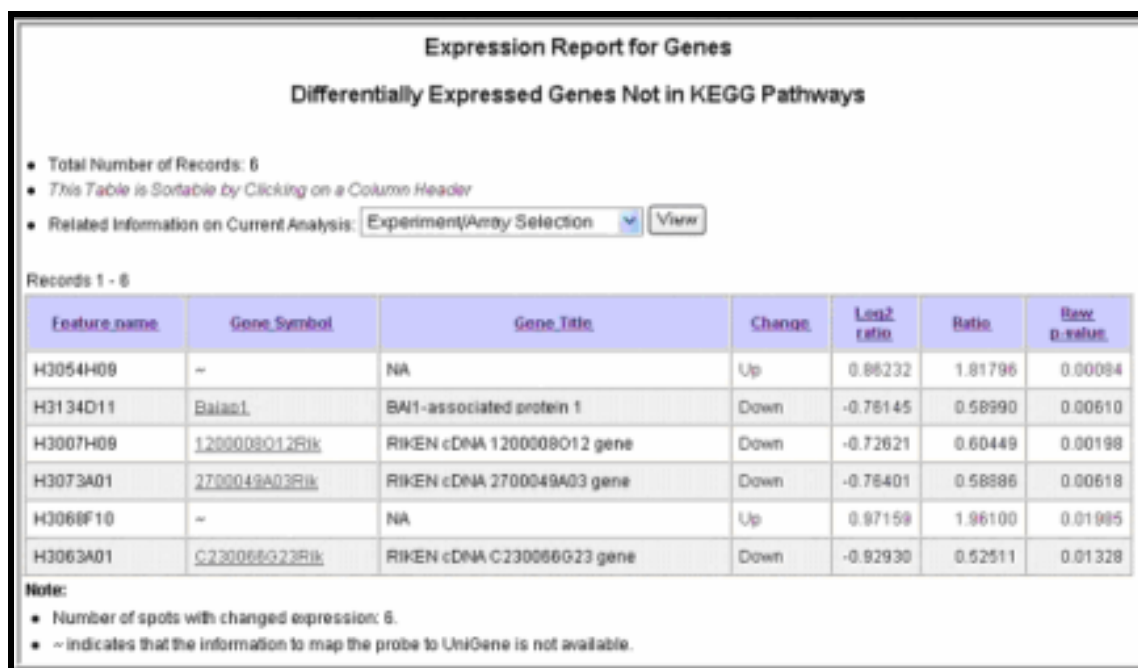


Figure 17: Expression reports in CEBS using pathway classification schemes

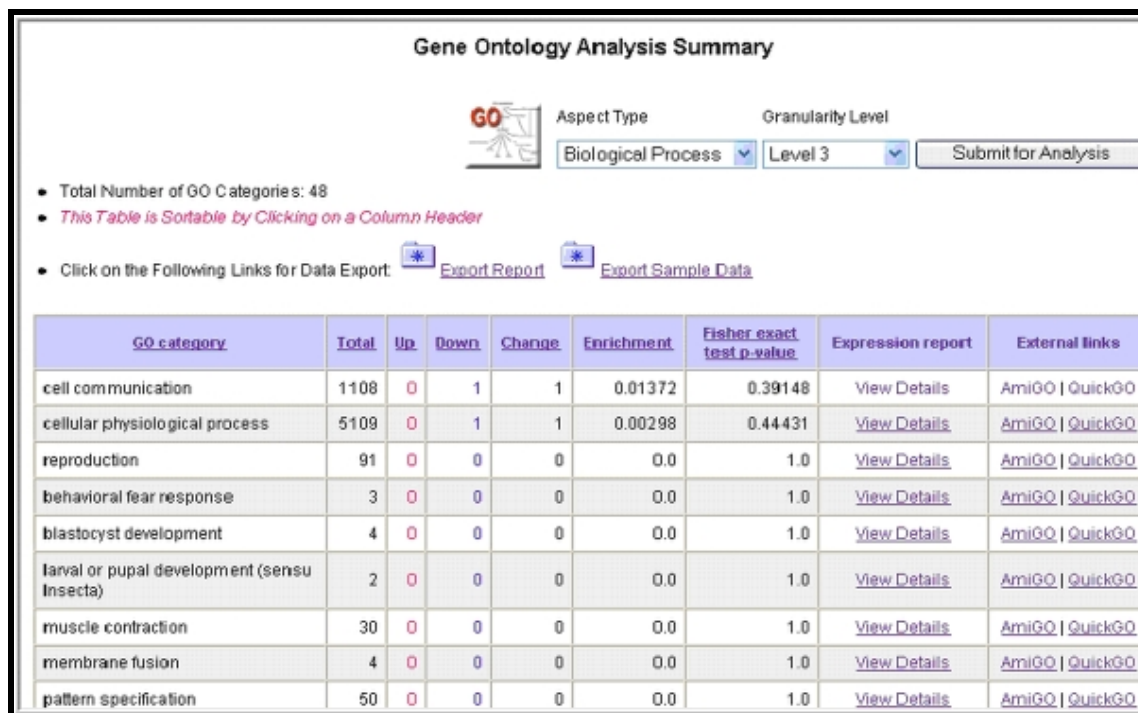


Figure 18: Expression in the 1,4-Dichlorobenzene degradation pathway

allow the user to define criteria for differentially expressed genes including statistical parameters such as t-test and p-value threshold. These menus are omitted from the examples here because they require specialized knowledge in statistical bioinformatics to be meaningful.

Comparisons among arrays can be filtered and selected based on whether or not they appear in particular important pathways. For example, there are standardized notations for biological pathways that are incorporated in the selection parameters for data analysis in CEBS. One of these notation systems is known as KEGG (used by the Kyoto Encyclopedia of Genes and Genomes suite of databases). A researcher may create a custom expression report using the CEBS database to identify the answer to the question which differentially expressed genes are not in the KEGG pathways as in figure 17, or the question which genes are in a particular KEGG pathway such as that for 1,4-dichlorobenzene degradation as shown in figure 18. CEBS is able to map microarray experiment results to gene classes based on important pathways that are affected. The analysis provides this information and is effectively serving as an indirect query engine for questions that are answered through detailed expression reports. CEBS is also able to provide classification reports for the genes of a particular experiment. For example, figure 19 below illustrates the use of GO ontology to generate a report listing to which GO category each gene in the experiment or selected subset of genes belongs.



Expression Report for Genes						
KEGG Pathway: 1,4-Dichlorobenzene degradation						
<ul style="list-style-type: none"> Total Number of Records: 7 <i>This Table is Sortable by Clicking on a Column Header</i> Related Information on Current Analysis: Experiment/Array Selection <input type="button" value="View"/> Click on the Following Links for Data Export:  Export Report  Export Sample Data 						
Feature name	Gene symbol	Gene title	Change	Log2 ratio	Ratio	Raw p-value
H3103E01	Top3a	Topoisomerase (DNA) III alpha	Unchanged	-0.20708	0.86629	0.09018
H3139A05	Top2a	Topoisomerase (DNA) II alpha	Unchanged	0.16958	1.12473	0.81249
H3122G11	Top2b	Topoisomerase (DNA) II beta	Unchanged	0.11681	1.08434	0.66708
H3098D03	Top2b	Topoisomerase (DNA) II beta	Unchanged	-0.06224	0.95777	0.63570
H3023E12	Top3b	Topoisomerase (DNA) III beta	Unchanged	-0.05130	0.96507	0.56209
H3036F09	Top2a	Topoisomerase (DNA) II alpha	Unchanged	0.40787	1.32672	0.65521
H3061C03	Top1	Topoisomerase (DNA) I	Unchanged	0.03482	1.02443	0.48257

Figure 19: Go ontology analysis matches expression to GO categories

In the example report summary both the categories “cell communication” (first row) and “cellular physiological process” (second row) show an entry of “1” under the attribute down. Therefore there are genes present in these categories that are down regulated (expression levels are reduced under the conditions of the experiment compared to normal cells). The degree to which they are down regulated is given by as numeric values from statistical formulae, the details of which are omitted here.

6.3 Querying the GEO Database

The GEO database is administered as two complementary databases, GEO Profiles and GEO Datasets. GEO is part of the NCBI (National Center for Biotechnology Information) and the GEO databases use the Entrez life sciences search engine portal common to other NCBI databases. As previously described in section 3.4.3, GEO organizes data into the three general categories: a) platform, b) sample, and



Figure 20: The GEO microarray database main query interface

c) series. These approximately correspond to a) the array design, b) signal data for each element in the array, and c) record for a group of related samples, respectively. The GEO Datasets stores collections of samples (datasets) sharing both a common platform and a common experiment type (single channel, dual channel, or SAGE if not microarray). The signal values are calculated and normalized the same way and are comparable across the set. Therefore they can be directly compared.

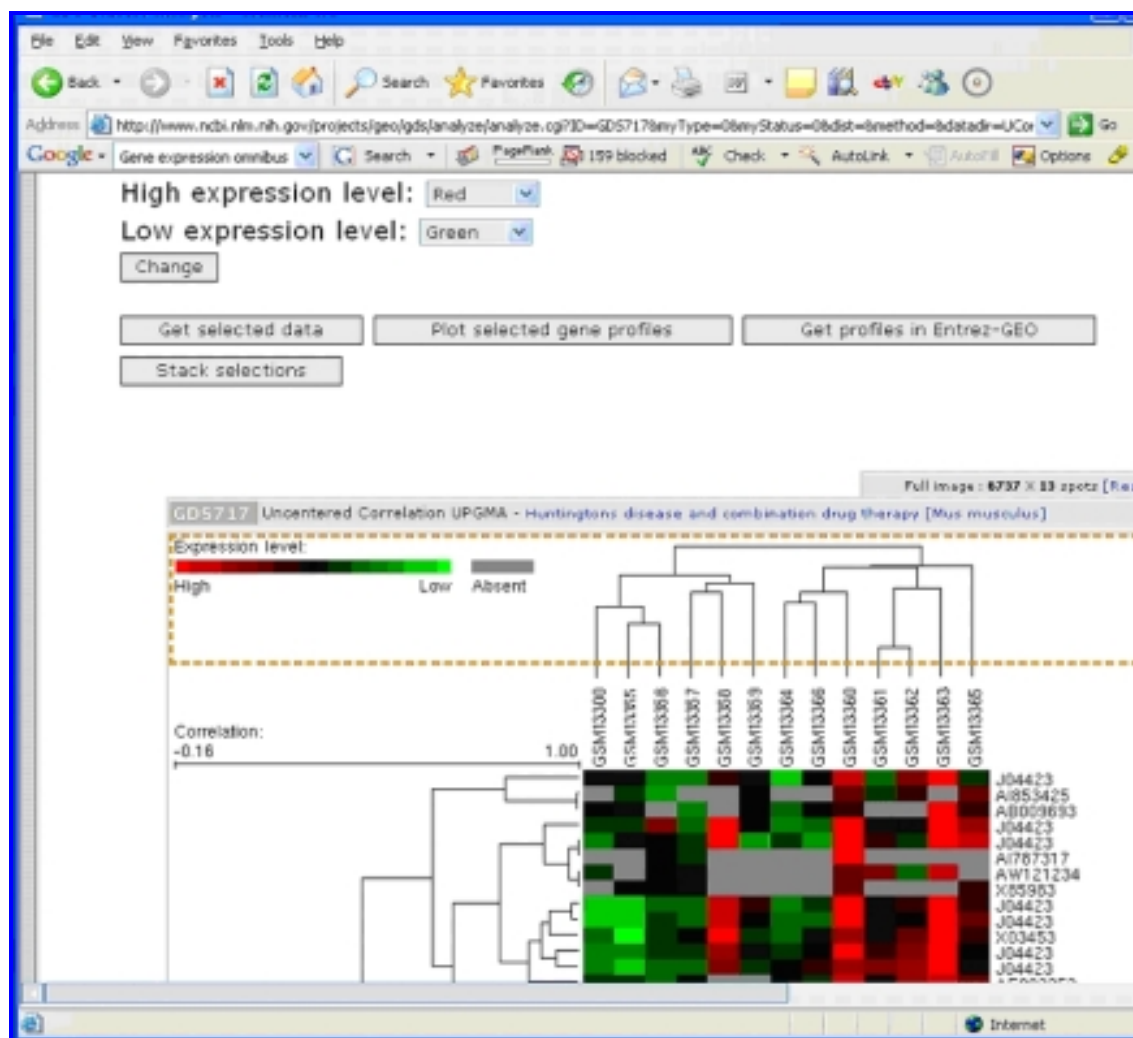
The GEO Profiles database stores individual gene expression and molecular abundance profiles. Graphical charts compare numeric data values for the same gene expressed across different experimental conditions. In brief, GEO Profiles handles the visualization of the data values while GEO Datasets stores other information. Both databases allow complex query constructs using Boolean values with optional filters on well-defined fields. The field options are described in associated documentation. The main query interface is displayed in the figure 20 below. Users may click on the tabs to browse the entries or enter search queries directly into the provided free text field. The figure 21 below shows the GEO Datasets interface with query results from the following example. GEO Datasets query 1: find all records having keywords “mouse” and “neurological”. The results retrieved 7 items, consisting of 1 dataset and 6 series. These are presented as summarized entries as seen in this screenshot.

Gene cluster analysis is provided, and linked to the record as a thumbnail icon. The analysis records are very large and often represent hundreds of genes for a given experiment. A small portion is visible in the display window. By clicking on it a navigable view with zoom, plot, and download options opens. Figure 22 below is an



Figure 21: The GEO Datasets query interface with query results

example of the data visualization interface using the thumbnail for the record in figure 21. High expression levels are in red and low expression in green. The 13 named genes are on the X-axis and the information about each sample (approximately 250) is on the Y-axis. Here only the top portion showing the first 12 samples is visible. We have now seen an example of querying GEO DataSets. Figure 23 shows the next example query uses the GEO Profiles interface. Example GEO Profiles query 1: find all experiments



studying kinase enzyme in macrophages. Enter the query “kinase AND macrophage”. The results retrieved 8297 items, each consisting of one sample. A researcher can now select a particular data set record (GDS77) such as study of kinase activity in macrophages infected with salmonella for further queries.

Example GEO Profiles query 2: find all profiles that fall into the top 5% variable molecular abundance profiles in dataset GDS77. Enter the query construct “GDS77 AND 96[Ranked Standard Deviation]”, this will return records for 241 genes.

The results are presented as summarized entries as seen in figure 23. Results may be sorted by the dimensions mean value, deviation, or outliers. Users may query based on rank or deviation directly to retrieve lists of the most abundant mRNA transcripts under selected conditions.

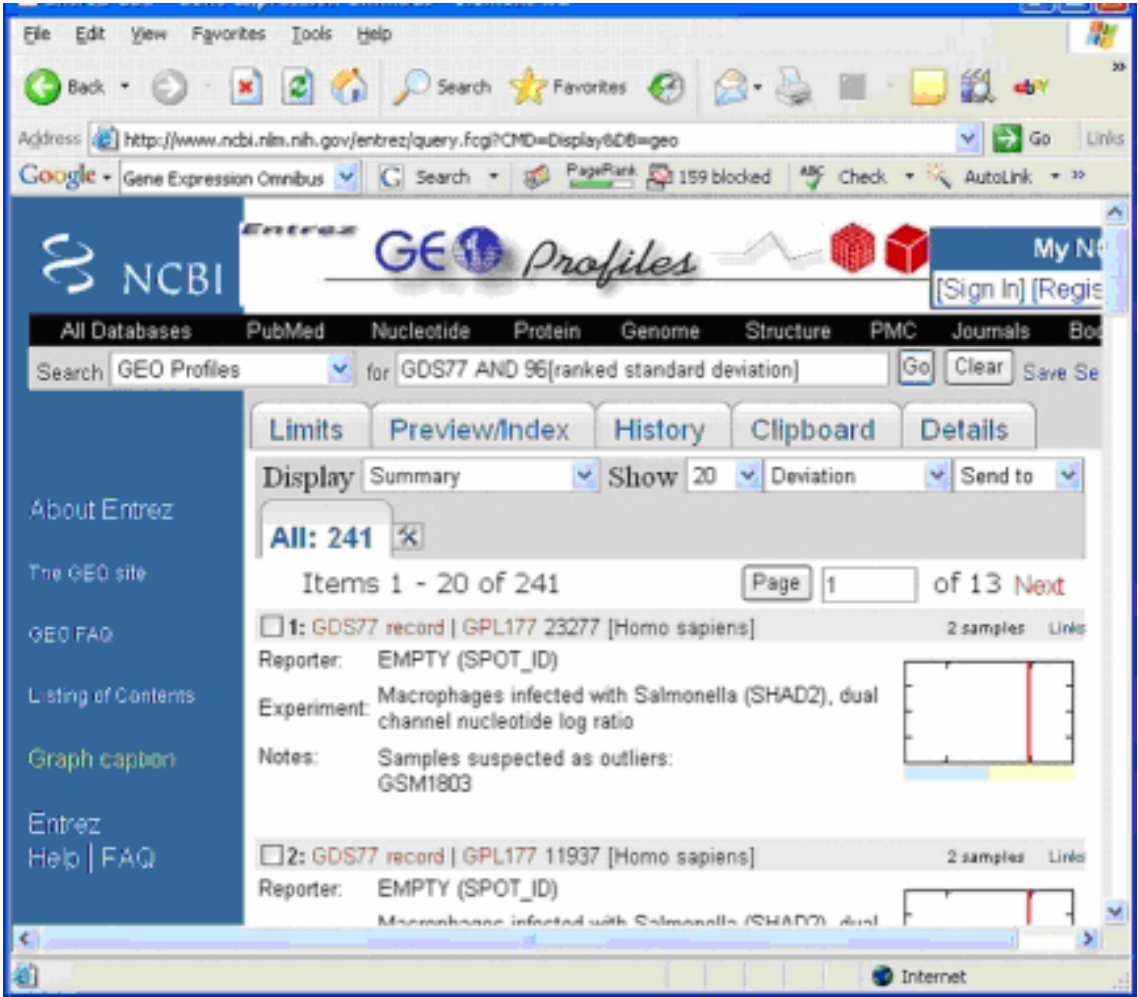


Figure 23: Geo Profiles example query result set

The thumbnail image is linked to a graphical gene signal profile in a visualization interface. The interface for an example gene is provided in figure 24 below. All the values for an each experiment are sorted and divided into 100 groups.

The blue rank bars on the left Y-axis of the chart show the approximate rank of where the expression of that gene is relative to the expression levels for all other genes on that particular array. The red bar on the right Y-axis of the chart uses arbitrary units of relative expression intensity to average, after normalization. Thirteen samples are represented in parallel. The conditions are in the light yellow and light blue blocks below the chart.

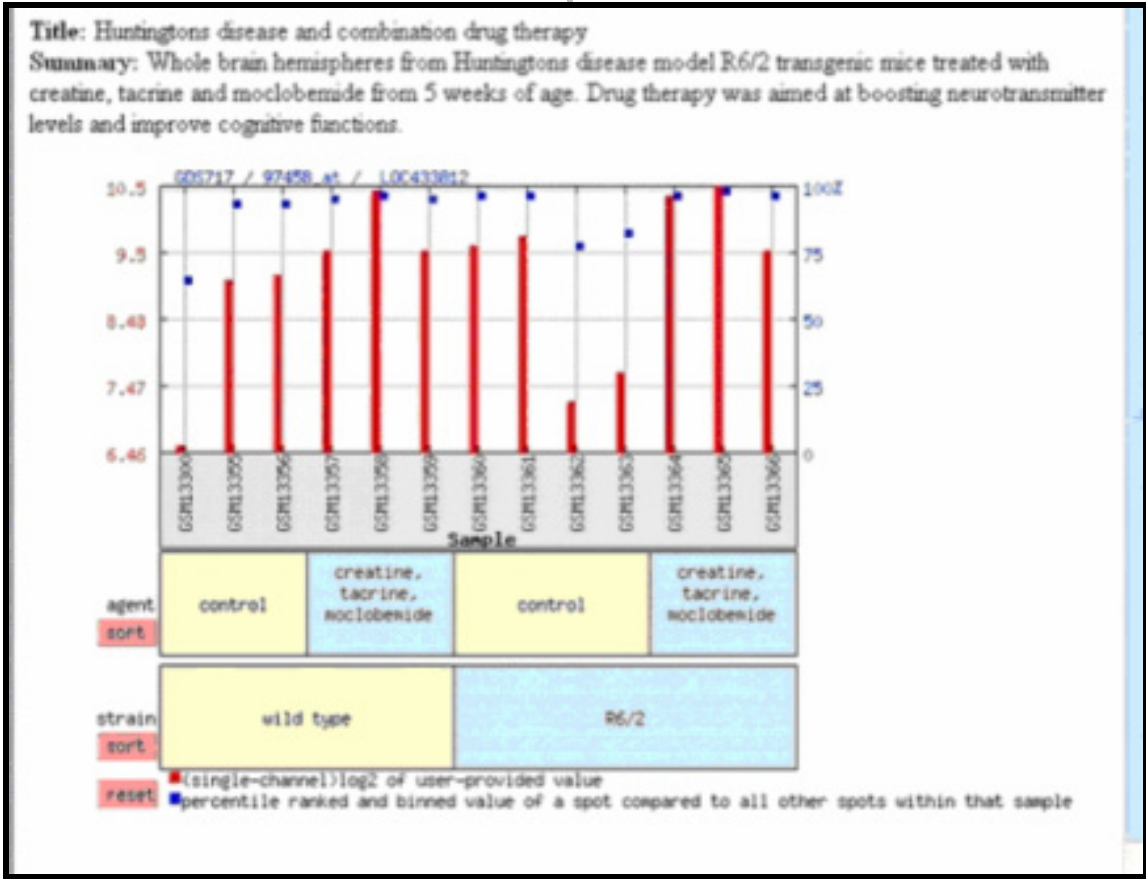


Figure 24: Visualization of signal values and ranks for a condition set

6.4 Querying the EMAP Database

The Edinburgh Mouse Atlas Project (EMAP) is a sophisticated graphical interface for mapping genes to mouse tissue, organs, and structures. EMAP provides a three dimensional spatial framework to map gene expression to any location within a virtual mouse embryo. Additionally, the stage of development for the embryo can be selected. This adds a fourth dimension of time for browsing and visualizing data, providing in effect a four dimensional atlas for gene expression. The EMAGE (the Edinburgh Mouse Atlas Gene Expression Database) is an application of the EMAP framework and provides tools for both data submission and query on stored data. Additionally, the EMAP has collaborated with the Mouse Genome Informatics (MGI) gene expression database (GXD) project. The GDX database is text based, and has been indexed through EMAP to spatially map images to the data stored in GDX.

In figure 25 below we see the query interface for EMAP. The cross sections may be selected from any of three axes (transverse, frontal, and sagittal as shown in the buttons of the lower left corner). The age of the embryo is divided into distinct Theiler stages that are selectable from a pull down menu in the upper right corner. Zoom and slider functions allow navigation in the tissue map. By simply scrolling over the map corresponding tissues are highlighted in the tree on the right. Clicking on the tissue results in the option to search either EMAGE or GDX databases for retrieval of genes corresponding to the stage and tissue selected for the query. The user may also query the tissue maps by typing the textual name of a tissue into the provided text box and clicking on the “find” button in the lower right of the frame.

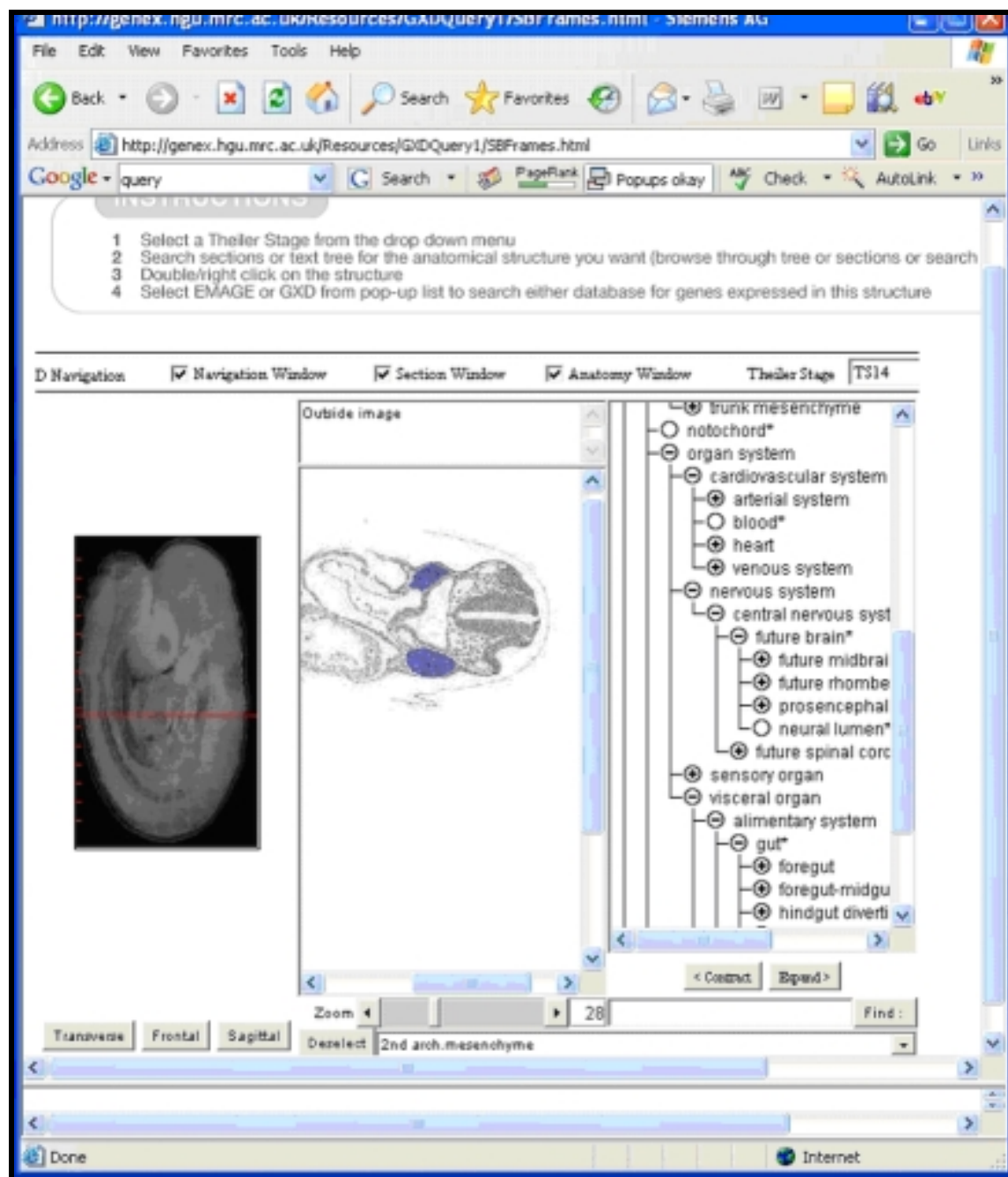


Figure 25: The main query interface of EMAP with navigable tree

Example query 1: find all genes for mesenchyme tissue in Theiler stage 14 mouse embryos. The results set shows 606 matching array results, as displayed in figures 26

and 27 below. In figure 26 the tissue type mesenchyme (blue structures) has been selected from the interactive map on the right representing a mouse embryo cross section. The tree on the left shows by marking in red font which tissue has been selected. Users may right click on the mouse embryo at the highlighted selected tissue type. The highlighted portion accesses interactive navigation that allows a search for genes for the selected tissue type. The star * marking at the end of the selected word mesenchyme refers to the option of navigating one more level in the tree and refining the search to just mesenchyme derived from head mesode or just mesenchyme from the neural crest.

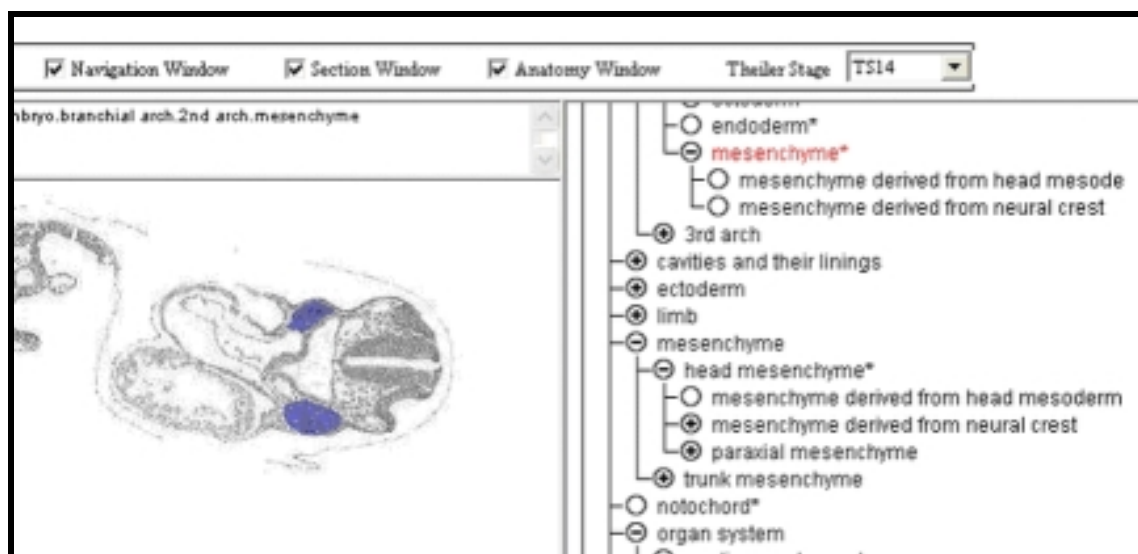


Figure 26: Tissue mapping portion of EMAP interface for basic querying

Figure 27 displays the first few results retrieved for genes matching the query “find all genes located in mesenchyme tissue from Theiler stage 14 mouse embryo”, each gene has navigable links to experiment details and data files for further analysis.

<div>?</div> <div>Gene Expression Data</div> <div>Query Results -- Summary</div>					
606 matching assay results displayed					
Gene	Assay Type	Result Details	Mutant Allele(s)	Age	Structure
1110012J17Rik	RNA in situ	MGI:2683725		E9.5	TS14: future midbrain
1110012J17Rik	RNA in situ	MGI:2683725		E9.5	TS14: future rhombencephalon
1110012J17Rik	RNA in situ	MGI:2683725		E9.5	TS14: prosencephalon
1500005K14Rik	RNA in situ	MGI:3575279		E9.0	TS14: future midbrain
1500005K14Rik	RNA in situ	MGI:3575279		E9.0	TS14: future midbrain
1500005K14Rik	RNA in situ	MGI:3575279		E9.0	TS14: prosencephalon

Figure 27: Emap interface showing first 6 of a large results set

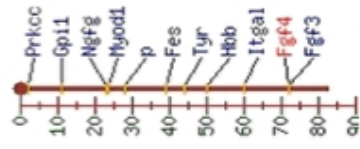
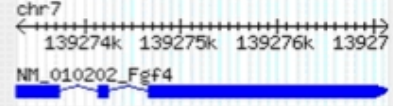
<div>?</div> <div>Gene Detail</div> <div>Your Input Welcome</div>	
Symbol Name ID	Fgf4 fibroblast growth factor 4 MGI:95518 <div>Nomenclature History</div>
Synonyms	Fgf-4, Fgfk, hst, hst-1, Hstf-1, kFGF, KS3
Genetic Map	Chromosome 7 72.4 cM Detailed Genetic Map ± 1 cM Mapping data(5) 
Sequence Map	139273295-139277152 bp, + strand (From Ensembl annotation of NCBI Build 34) Ensembl ContigView UCSC Browser NCBI Map Viewer  <div>MGI Mouse GBrowse</div>
Mammalian orthology	human; rat (Mammalian Orthology) Comparative Map (Mouse/Human Fgf4 ± 2 cM)
Sequences	Representative Sequences <div>Length Strain/Species Flank</div> <div> <input type="checkbox"/> genomic ENSMUSG00000050917 <div> Ensembl Gene Model MGI </div> </div> <div> 3858 C57BL/6J <div>± 0</div> </div>

Figure 28: Gene detail record from the example query result

Address http://www.informatics.jax.org/searches/expression_form_exp.shtml Go Links

Google GDX mouse genome Search PageRank 166 blocked ABC Check >>

Gene Expression Data Expanded Query Form

This form allows you to search for genes expressed in some anatomical structures and/or developmental stages but not in others. In addition, you can use the query parameters available on the standard Gene Expression Data Query Form.

Search Reset

Sort by: ☒ Gene symbol ☐ Age ☐ Anatomical structure ☐ Assay Type ☐ Auth

Max number of items returned: ☐ 100 ☒ 500 ☐ No limit

Return: ☐ Assays ☒ Assay Results

Gene Symbol/Name:

☐ NOT contains Search

current symbols/names & synonyms

Gene Ontology (GO) Classifications: (You can [browse the GO Classifications](#))

contains

☒ Molecular Function ☒ Biological Process ☒ Cellular Component

Chromosomal Location:

Chromosome: ANY

Restrict search to a chromosomal region? (specify one of the following)

Between and (Enter cM positions or locus symbols) Include endpoints.

Within cM of locus Include locus

Expression: (You can browse the [Anatomical Dictionary](#) or the [Stage](#))

Figure 29: Query form of the GDX mouse genome expression database

Details regarding any particular gene are retrieved within the browser and presented as a detailed record from the database. A portion of that record is shown in figure 28 above.

The EMAP query interface is a graphical mapping tool built peripherally to the mouse genome expression database GXD, and using the same data. It should be noted that the GXD query interface (the database is located at the following weblink http://www.informatics.jax.org/searches/expression_form_exp.shtml) offers additional query parameters and options. These include searches by querying with gene name, gene ontology (GO) classifications, and chromosome location. A portion of the interface is shown in figure 29 above.

6.5 Querying the SMD Database

Similar to other microarray databases, the Stanford Microarray Database (SMD) database query interface is designed for the user to narrow down the subset of results by providing selection criteria. After selecting the criteria and retrieving the results summaries, the user may look at arrays individually or combine results for final retrieval and analysis. There are three approaches to querying the database. These are the basic search, experiment list search, and advanced results search. Each is accessible from the main interface as shown in figure 30 below. Queries are on high level classification criteria and result sets contain large complete records. Unlike other microarray databases, SMD is not yet able to support query constructs to find information about a particular gene although this is a goal for their future improvements.

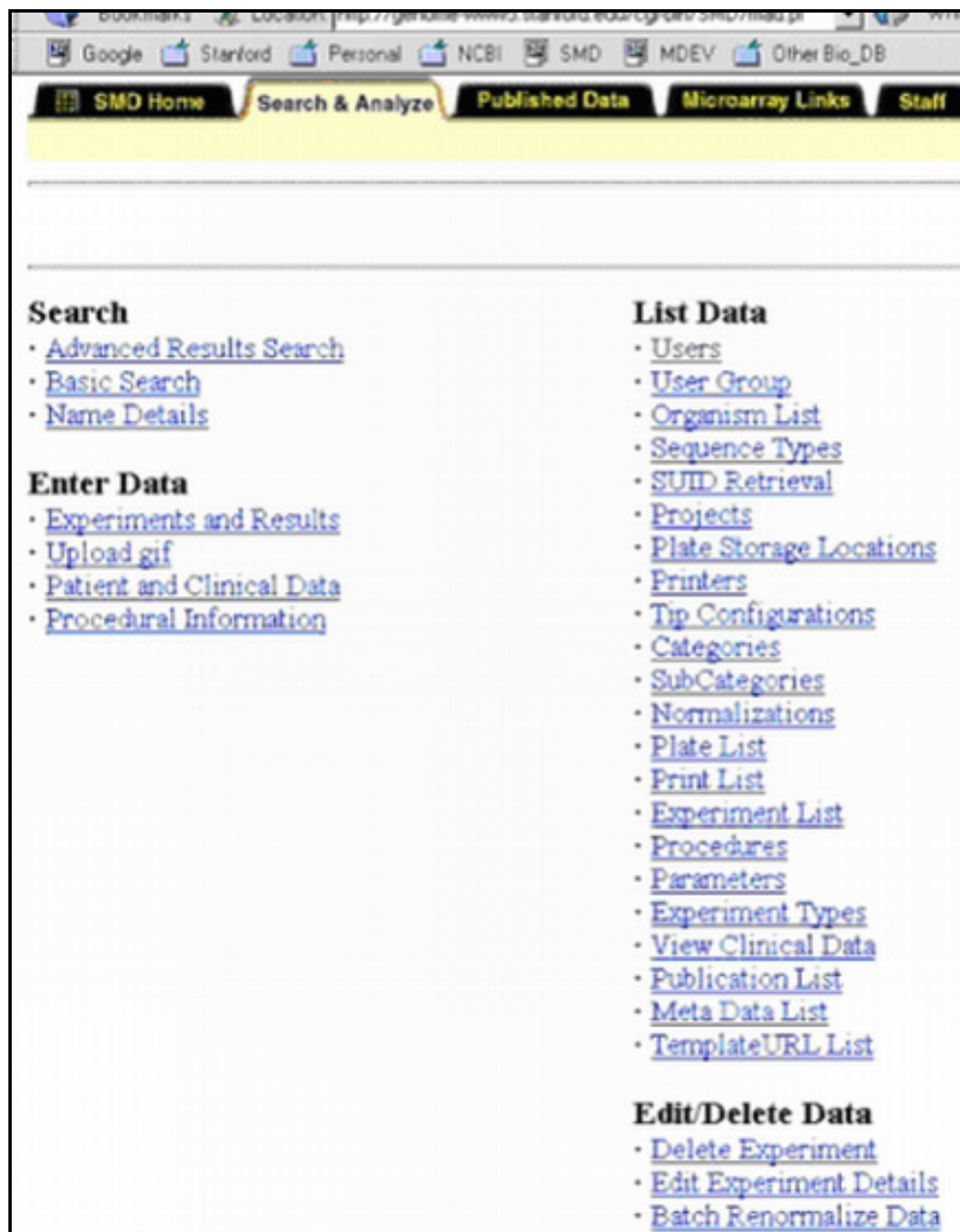


Figure 30: Main interface of the Stanford Microarray database (SMD)



Figure 31: Query “caenorhabditis elegans” in SMD basic search

In a basic search, the results may be presented in terms of related publications or in terms of an experiment category or set. An experiment set is a group of experiments that have been selected and assigned to that set. The user may customize such a group as a personal set. Figure 31 presents the basic search interface. Queries on this interface will retrieve publications, experiment sets, or experiment categories related to the organism and data identifier of interest. The interface provides selection options based on the species and experiments currently in the database. Here we see a search for microarray data on the species “*caenorhabditis elegans*”.

For the more advanced searches there are three methods for analyzing the microarray data. In the first method, Boolean operators are used to query on the three major parameters experimenter, category, and subcategory. In the second method arrays are simply selected by their print identification, and the third arrays are selected from a personal directory. The resulting datasets from any of the above methods may be displayed as array lists by clicking on the “Display Data” button or the “Data Retrieval and Analysis” button as shown below in figure 32.

Query results in figure 32 are displayed in summary lists with links to other related array sets. In the example below six arrays are returned from querying the database for results matching “experimenter = EISEN”. Each experiment is classified by both category and subcategory with links to other arrays in that same classification. A set of interactive icons summarizes options for handling the results set including a clickable image, array details, download, and plot data.

Help

Search Results

You may search for microarray experiment results using either of three methods.

- Method 1 - Allows you to select arrays by experimenter, category, subcategory and organism
- Method 2 - Allows you to select experiments by pinname
- Method 3 - Allows you to use a premade list of arrays in your personal directory

☒ **Use Method 1** to select arrays by Experimenter, Category, Subcategory and Organism.

Select Organism: All Link Lists By Organism

Experimenter	Category	SubCategory
<div style="display: flex; align-items: center;"> <div style="margin-right: 5px;">Pick:</div> <div style="border: 1px solid black; padding: 2px; flex-grow: 1;"> <div style="font-size: 0.8em; margin-bottom: 5px;">All</div> <div style="font-size: 0.8em;"> AANDRE AARNIM ABATE ACONNOLL ACUTEMI ACYC AEPSTEN AFIRE AGASCH </div> </div> </div>	<div style="display: flex; align-items: center;"> <div style="margin-right: 5px;"> <input checked="" type="radio"/> And <input type="radio"/> Or </div> <div style="border: 1px solid black; padding: 2px; flex-grow: 1;"> <div style="font-size: 0.8em; margin-bottom: 5px;">All</div> <div style="font-size: 0.8em;"> Absolute transcript levels aCOH acute lymphoblastic leukemia Adenoma aging AKT AKtest Amino Acid metabolism anaerobic growth </div> </div> </div>	<div style="display: flex; align-items: center;"> <div style="margin-right: 5px;"> <input checked="" type="radio"/> And <input type="radio"/> Or </div> <div style="border: 1px solid black; padding: 2px; flex-grow: 1;"> <div style="font-size: 0.8em; margin-bottom: 5px;">All</div> <div style="font-size: 0.8em;"> 3endseq 3-aminotriazole 3-mba 4ats 4NQO AA_starvation ace-1:ace-2 aCOH cell line aCOH tumor tissue </div> </div> </div>

☐ **Use Method 2** to select all arrays from a given point: 1.5K

☐ **Use Method 3** to use arrays from a list in your personal directory: Ovarian23K.txt

Display Data
Data Retrieval and Analysis
Reset

• 'Display Data' allows you to view and sort data, view array details or view array images and grids.
 • 'Data Retrieval and Analysis' allows you to select experiments from a list for clustering or data retrieval.

Figure 32: SMD advanced search interface, links to display or analyze data

Selection of the view and sort data icon results in a menu of options for both display and application of filters as shown in figure 34 below. Options include sorting, and display of columns for biological annotation data. Filters allow inclusion of controls and nulls in experiment results. The resulting adjustments may be saved as a downloadable file.

• Click on a Category/Subcategory to retrieve a list of all arrays using that category/subcategory
 • Click on an experimenter to view their details
 • = View and Sort Array Data
 • = Download Raw Data
 • = View Array Details
 • = View Array Image and Grids
 • = Clickable Image
 • = Plot Array Data

Your query returned 6 arrays.

ExptID	Experiment	Category	Subcategory	SlideName	Options	Exp
8280	High Molecular Weight Polysomes: Polio vs. Mock	Infection	Polio virus	nci2790	EIS	EIS
8322	Polio Polysomes	Infection	Polio virus	nci125	EIS	EIS
8293	Polio Polysomes 2 hours	Infection	Polio virus	nci430	EIS	EIS
8312	Polio Polysomes 3 hours	Infection	Polio virus	nci536	EIS	EIS
8335	Polio polysomes (3 hours) vs. total RNA	Infection	Polio virus	nci1225	EIS	EIS
8283	Total RNA Polio vs. Mock	Infection	Polio virus	nci2791	EIS	EIS

Figure 33: Example results set for a query “experimenter = EISEN”

Select data for slide : nci2790

" High Molecular Weight Polysomes: Polio vs. Mock "

Sort By:

Display:

Channel 1 Background (Mean)	Clone ID
Ch1 Background (Median)	Gene Symbol
Std Dev of Ch1 Background	Gene Name
Ch1 Net (Mean)	Cluster ID
Ch2 Intensity (Mean)	Accession

Display Rows: -

Display Control ☐ Display NULLs ☐ Make downloadable file of ALL data ☐

Filter #1: ☐

Filter #2: ☐

Filter #3: ☐

Filter #4: ☐

Filter #5: ☐

Filter #6: ☐

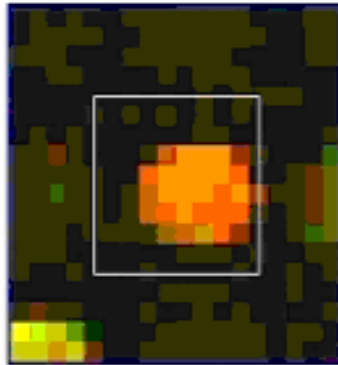
Figure 34: Selection options for data view and sort on a result set row

When the user chooses the view array details icon the database provides a record with navigable links to experimental and sample details. These include signal intensity data suited to plotting, and statistical analysis. Figure 35 below shows sample details as a record with links such as data download from the normalization data field. The detailed record in figure 36 describes one individual signal from one position on the microarray grid. This represents the hybridization of an individual gene. The record provides

Data for : nci2790	
Experimenter	EISEN
Experiment ID	8280 Compare log2(RAT2N) with experiment from the same Prin
Experiment Date	2000-08-01
Slide Name	nci2790
Experiment Name	High Molecular Weight Polysomes: Polio vs. Mock
Channel 1 Description	Mock Infected HeLa 3 hours HMW Polysomes
Channel 2 Description	Polio Infected HeLa 3 hours HMW Polysomes
Is Reverse	N
Category	Infection
Subcategory	Polio virus
Normalization	Using regression correlation normalization value is .75 Data D
Print Information	
Print ID	156
Print Name	10k_Print3
Print Configuration	Standard 4-tip
Description	
Description	This experiment was an out file converted from rana. The original fil and Subcategories were made by curators

Figure 35: Example view showing experiment and sample details for an array

"High Molecular Weight Polysomes: Polio vs. Mock"



Click spot to see in array context

Biological Information

Clone ID [IMAGE 242678](#)
Gene Name ESTs
Cluster ID Hs.269022
Accession H94252

[View expression history](#) of this entity

Spot	847
Stanford ID	237932
Ch1 Intensity (Mean)	2353
Ch1 Net (Median)	
Ch1 Intensity (Median)	
% of saturated Ch1 pixels	
Std Dev of Ch1 Intensity	
Channel 1 Background (Mean)	
Ch1 Background (Median)	916
Std Dev of Ch1 Background	
Ch1 Net (Mean)	1437
Ch2 Intensity (Mean)	3635
Ch2 Net (Mean)	3019
Ch2 Net (Median)	
Ch2 Intensity (Median)	
% of saturated Ch2 pixels	
Std Dev of Ch2 Intensity	
Channel 2 Background (Mean)	
Ch2 Background (Median)	616
Std Dev of Ch2 Background	
Ch2 Normalized Background (Median)	821
Ch2 Normalized Net (Mean)	4025
Ch2 Normalized Intensity (Mean)	4846
Normalized Ch2 Net (Median)	
Normalized Ch2 Intensity (Median)	
Regression Correlation	.96
Diameter of the spot	
Spot Flag	0
Log(base2) of R/G Normalized Ratio (Mean)	1.486
Log(base2) of R/G Normalized Ratio (Median)	
R/G Mean (per pixel)	

Figure 36: View showing details and signal for an individual grid position

numeric values for the two channels (colors) as well as normalization and other statistical information about the signal. Chapter 2 provides an overview of how signal data is interpreted.

6.6 Querying the HugeIndex Database

The HugeIndex microarray database is a repository for normal human tissues. To improve data analysis capability, each gene is cross-referenced to annotation in the LocusLink database at NCBI (National Center for Biotechnology Information). The database can be queried through menus on its web interface. In figure 37 below gene expression can be queried based on keyword and selection of an organ for the gene.

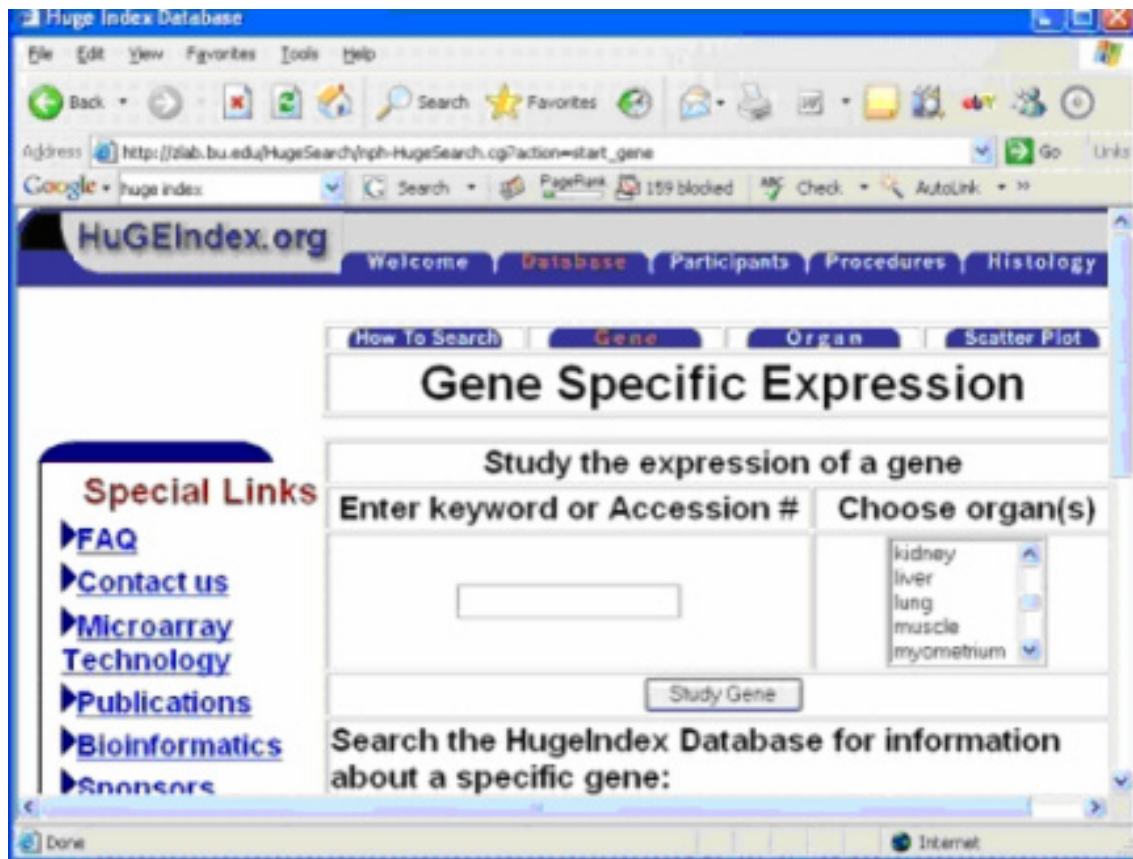


Figure 37: Gene specific expression query interface in HugeIndex



Figure 38: Result set for the query “cathepsin” in tissue “lung”

Example query 1: Find all genes encoding cathepsin proteins for lung. Enter keyword “cathepsin” and select lung. In the result set is the list of retrieved genes shown in figure 38 below. Figure 39 displays an individual gene expression profile.

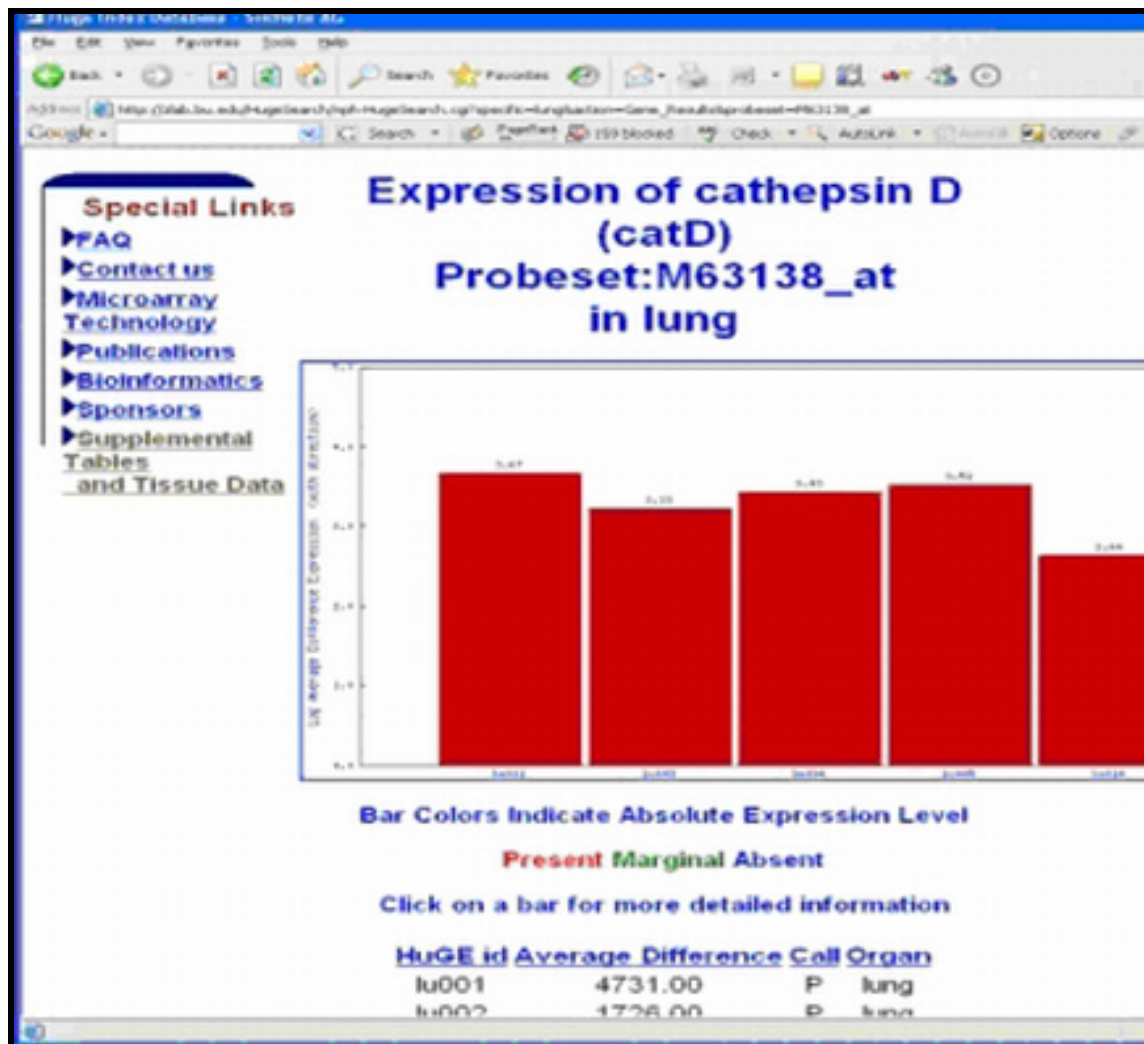


Figure 39: Individual gene profile for a gene selected from figure 6.6.2

In the above figure, the information regarding the gene is provided, including averaged signal results from multiple experiments in the same tissue. These are presented as average and standard deviation values. The Y-axis shows log base 10 average difference of expression (the difference between the sample expression and the average of expression for all genes in an experiment). Here there are six probes for the same gene (cathepsin D) providing information to allow normalization for differences

due to experimental conditions. Average differences range from 705 to 4731 based on an arbitrary signal intensity scale and corresponding log values range from 2.64 to 3.67.

The options available for composing queries can be viewed by browsing the database contents through the supplemental tables that link from the main web interface. Queries can be composed on any of three custom interfaces for each of three purposes. The first option is shown in figure 37. The query is on gene specific expression for a single organ of interest. The keyword for the gene of interest may be typed in the text box and the organ of interest selected from the menu. The second option shown in figure 40, it shows a query on expression comparison between different organ tissues. The third option is shown in figures 41, it displays the comparison of tissues or experiments through scatter plots. The design of the HugeIndex database includes query capabilities for comparison of global expression patterns among tissues. The user can query and discover gene coexpression in particular tissues, this leads to detection of the same expression patterns among related tissues as a system. The implementers describe HugeIndex in this capacity as “a reference for defining basic organ systems biology” [HWBAHJG02]. Scatter plots allow interactive selection of individual genes and separate plots of gene expression for each gene or selection of links for annotation data to external sources such as NCBI through LocusLink.

Example query 2: Find genes expressed in liver and lung, but absent from kidney, selected at organ interface shown in figure 40. In the results set the database reports that 2215 genes match these conditions, and provides links to download text file of results provided.

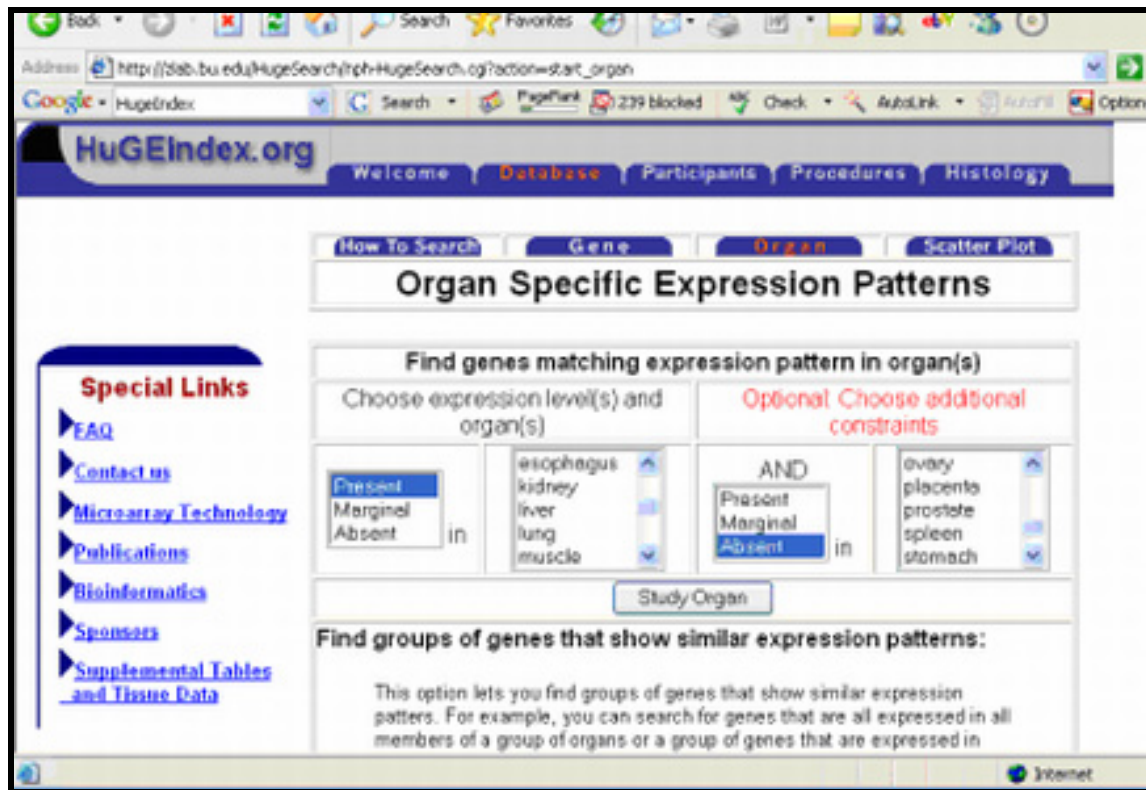


Figure 40: Interface for searching organ specific expression patterns

Example query 3: What are the differences in gene expression between brain experiment 7 from the database (pons/medulla tissue) and the average of 10 other brain tissue experiments? In the results set we are provided a scatter plot with interactive labels that appear when the cursor moves position over the data points. Each gene is identified in the label along with its status as present or absent for those selection conditions. Figures 41 through 43 below show the generated interactive scatter plot for gene expression data comparisons in brain tissue. In figure 42 the selected point in the map is for the gene Mac2 binding protein, the corresponding point can be clicked for interactive plotting options specific to each gene.

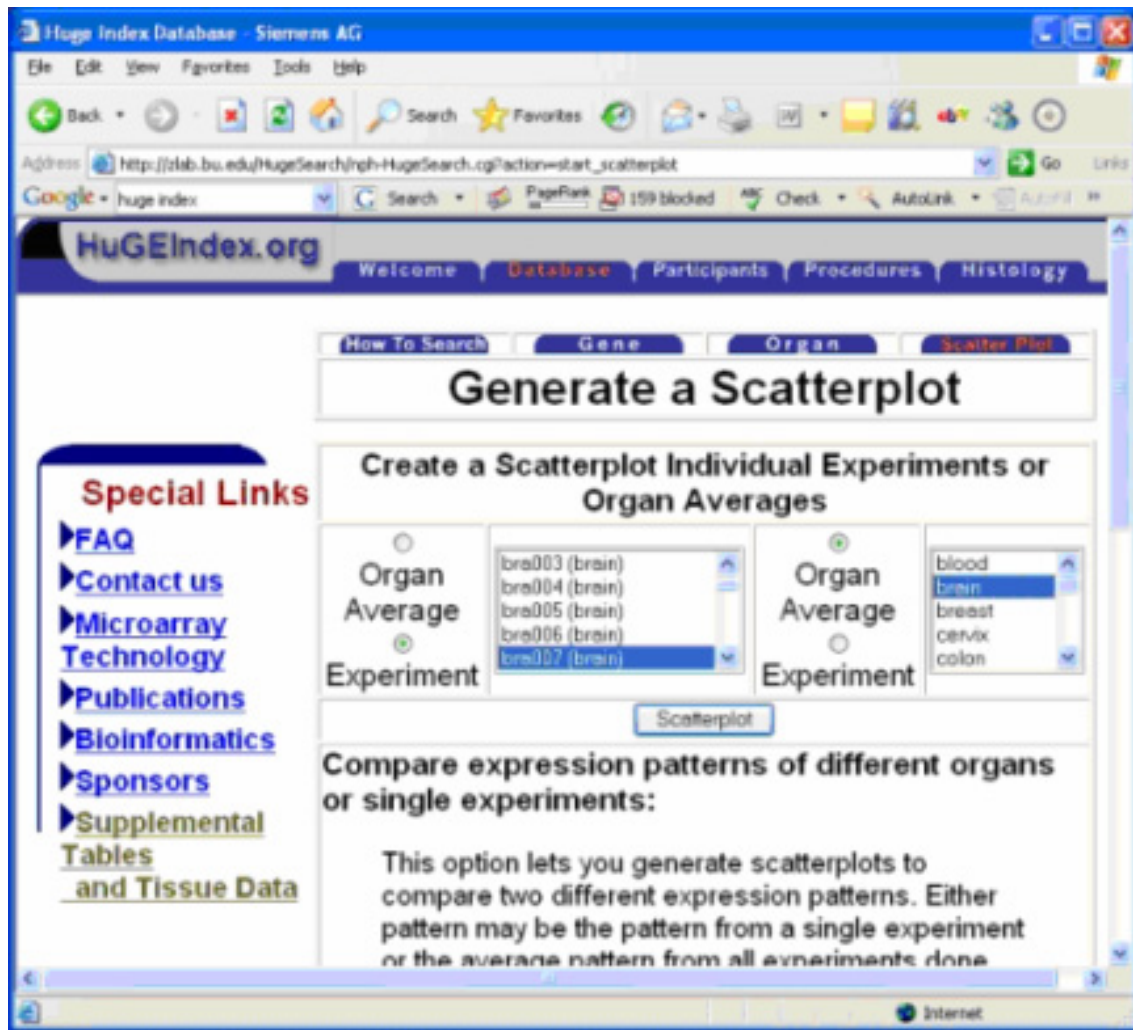


Figure 41: Interface to generate scatter plot for expression comparison query

In figure 43 which follows we will see that plot for expression of the gene for the Mac-2 binding protein in tissue sample 7. The expression is plotted as log average expression (Y axis) vs. the average expression of that gene for the set of 11 tissue samples available in the HugeIndex database. The expression is almost 30 times higher in sample 7 compared to average expression levels.

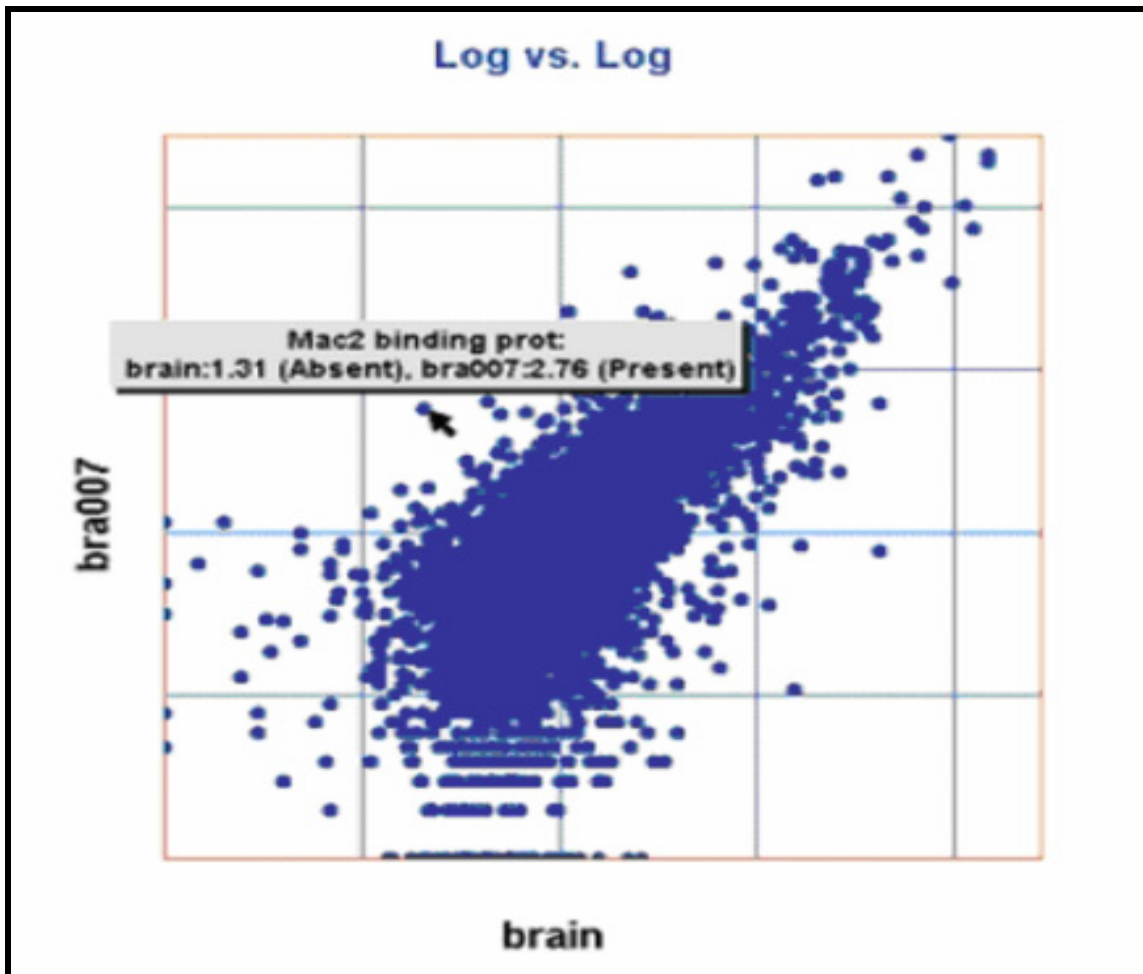


Figure 42: Scatter plot showing Mac2 gene is a result of example query 3

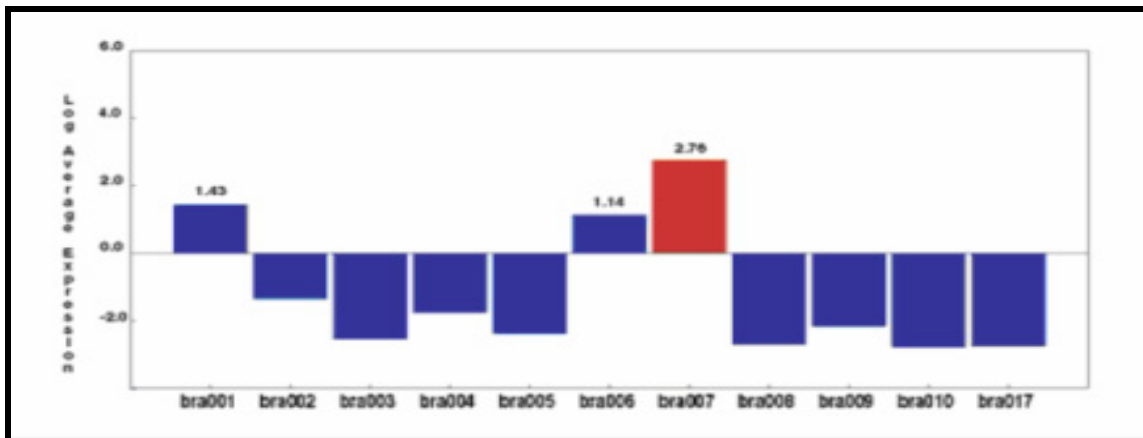


Figure 43: Plot for an individual gene selected from the interactive scatter plot

CHAPTER 7

LIMITATIONS AND SUGGESTED SOLUTIONS

In this chapter we consider some of the limitations in the microarray database implementations we have studied. Our focus is on those identified limitations for which we are able to propose reasonable solutions. The limitations have been separated into two categories based on whether they primarily impact querying and quality of results presentation or primarily impact the microarray database implementation. We define each limitation and describe its impact. We then provide a suggested solution. This chapter then considers limitations in a broader context by examining two other groupings. One group is of inherent limitations for which there are no reasonable solutions. The other group is of successfully addressed limitations for which a reasonable solution exists and is agreed upon. These are no longer true limitations unless current and new databases fail to adopt the solutions in the future.

7.1 Limitations Impacting Microarray Database Querying

In this section we will describe four limitations that impact querying, and present a suggested solution. These are inconsistencies in feature ontology, lack of accomodation for free text queries, lack of support for both time-varying image data and for a consistent system of gene ID numbering.

7.1.1 Limitation 1: Inconsistencies in Feature Ontology

Impact on querying: When different terminology is used to describe the same concept about a microarray feature (ontology). This difference poses problems when exchanging data between different databases. The lack of consistency also imposes limitations for query formation, particularly for free text searches.

Solution to Limitation 1: A potential solution would be to provide a table that cross references several ontologies permitting lookup of equivalent terms. Currently some implementations use GO consortium ontology [HCILAF04] but this is not sufficient by itself to accommodate all microarray research concepts. It should be noted that a microarray specific ontology is under development from MGED (Microarray Gene Expression Data Society at <http://www.mged.org>).

7.1.2 Limitation 2: Lack of accommodation for free text queries

Impact on querying: It is assumed that the user will not know the exact terminology or ontology particular to an individual implementation. As a simplifying solution, many implementations provide pull down menus of all possible selections for their main data types. These predefined queries are not only universal among implementations but are also recommended because they simplify the database structure. However, this lack of accommodation for more sophisticated user-constructed queries is both a restrictive and unscalable aspect of the query interface design. There is also an impact on querying by limitations on keyword search in most implementations that support it. They typically neither have a spellchecker nor the ability to prompt for a close alternate spelling when a keyword is not found. For

example, in ArrayExpress the query “affimetrix” retrieves 0 records, but the correct spelling “affymetrix” retrieves 45 records.

Solution to Limitation 2: All the implementations researched for this work do not address the need to support user-authored custom queries in free text search. Instead, they favor the use of pull down menus to view and select predefined options for combination in each query. Since the documentation and associated publications for these databases does not describe the use of formal query language, or query optimization it is difficult to propose a specific solution. By contrast the solution to supporting approximate keyword search is more obvious. This solution is to employ a spell check to determine approximate substitutions or suggestions for the query keywords, similar to what is provided in the Google.com web search tool interface. The technique known as *relaxed queries* is a more complex variation. This employs a type of synonym mapping so that terminology having closely related meanings could be checked and indexed to find information relevant to the user without imposing a single ontology.

7.1.3 Limitation 3: Lack of support for time-varying image data

Impact on querying: An important source of information in microarray experiments is the rate of change in gene expression over a window of time. Most microarray database implementations lack time interval parameters and time response data. While a set of expression snapshots at punctuated intervals could be provided relatively easily, it is not a substitute for the valuable information detailing the response in gene expression over a time period specified within the query. Such information is

important in evaluating cellular response to medication, or exposure to toxins.

Solution to Limitation 3: An interesting solution is proposed by Craig et al [CK03] in which “visual queries are supported by the combination of a traditional value against time graph representation of the data with a complementary scatter-plot representation of a specified time-period. The multiple views of the visualization are coordinated so that the user can formulate and modify queries with rapid reversible display of query results in the traditional value against time graph format. This new visualization technique allows the user to perform and combine a number of queries, including measurable change in value over a period of time queries, through an intuitive direct manipulation interface. The technique also gives the users a unique directly manipulated animated view of microarray timeseries that allows them to explore patterns over time for the entire data set and selected subsets.”

7.1.4 Limitation 4: Lack of consensus on gene ID numbering

Impact on querying: There are many gene ID (identification number) systems in use among different databases. Examples include UniGene, OMIM, EMBL, Entrez Gene, and Affymetrix Probe Set ID. Since microarray databases access other databases for annotation information about the gene it is important that the gene ID systems in place for retrieving that annotation are selected carefully. Microarray data generated from Affymetrix chips are most easily searched using the Affymetrix Probe Set ID system, however trying to use that ID to initiate annotation searches poses problems. ID systems differ among different microarray kits and experiments may submit gene IDs that only refer to a particular brand of kit creating inconsistencies within a

microarray database. Additionally, some systems such as the UniGene ID system use numbers specific to each species and do not allow cross species comparison for the same gene. [CEMKZ05]

Solution to Limitation 4: A reasonable solution would examine important gene ID systems and the correlations between these systems, then provide this information in look up tables so that appropriate gene ID codes can be used for external searches for gene annotation. Checks should be in place to account for design differences such as the use numbers specific to each species in the UniGene ID system. Table 9 below summarizes these four limitations that affect querying.

Table 9: Summary of Limitations and Solutions Affecting Querying

#	Limitation	Proposed solution
1	Inconsistencies in Feature Ontology	Provide a table that cross references several ontologies permitting lookup of equivalent terms.
2	Lack of accommodation for free text queries	Use of spell check for keyword searches. Use of synonym mapping so that terminology having closely related meanings could be checked and indexed to find information relevant to the user without imposing a single ontology.
3	Many databases are not able to provide time-varying image data	An interesting solution is proposed by Craig et al [CK03] in which “visual queries are supported by the combination of a traditional value against time graph representation of the data with a complementary scatter-plot representation of a specified time-period”.
4	Lack of consensus on gene ID numbering system	Examine important gene ID systems and the correlations between these systems, then provide this information in look up tables so that appropriate gene ID codes can be used for external searches.

They include lack of consistency in terminology, lack of flexibility for writing a free text query, inability to query how gene expression changes within a particular range of time, and the use of many different gene identification systems causing difficulties if using the gene ID to search other databases for annotation data about the gene.

7.2 Limitations Impacting Microarray Database Implementations

This section describes some important limitations in the implementations. These limitations are important to the usability of the database and therefore have an indirect impact on querying. For example, web based remote access to the data and query results is subject to slow response time or limitations in display capabilities. Many implementations do not have integrated annotation data or original unprocessed image data from the experiments. Further description of these problems and proposed solutions are provided below in table 10.

7.2.1 Limitation 1: Web forms for data retrieval and presentation

Impact on querying: In some databases which rely on Java applets or require large bandwidth to transfer image files there can be a noticeable lack of performance. In some designs hyperlinks or pop-up based retrieval of records are in conflict with common browser security settings. An additional basic restriction posed by the browser is limitation in their graphics display capability.

Solution to Limitation 1: Although web based access is important to provide efficient public open access and reach the widest community of users some implementations also offer the option of setting up the database and downloading the entire data store locally to achieve better performance for frequent querying. Several

microarray database websites also offer or recommend separate locally installed programs to provide flexible alternatives to the limitations of web browser interfaces for data visualization.

7.2.2 Limitation 2: Use of hyperlinks to external databases

Impact on querying: Use of hyperlinks to external databases such as GenBank are employed as a mechanism for providing annotation data about genes. This is a commonly implemented solution, but recognized as insufficient. Hyperlinks are used when annotation data is not compatible with the data structures in the original implementation. Hyperlinks introduce problems with presentation since the annotated information is provided only for each gene in a separate browser window as an independent search of an external database, rather than integrated in a table of entries for efficient comparison. Secondly, there are performance issues if many requests are made to external data sources.

Solution to Limitation 2: The solution is to import annotated information and integrate it into the records for more effective and efficient presentation. Adjusting the format of the data to accommodate native data structures should be done. In some cases it may be helpful to implement two or more data exchange models to accommodate a variety of annotation data. For example, the MAGE-OM data exchange model has the limitation that it can not include clinical data. Since clinical data is often very relevant to gene expression research on human disease it would be helpful to adopt a clinical data exchange model in parallel. [CMF03, BEF03]

7.2.3 Limitation 3: Database does not store original image

Impact on querying: When the original data as a large image file is discarded, it is replaced with numeric data or other representations of relative signal strengths. The original image signals will have been interpolated and processed before the values are stored. This affects how accurately the relative signal strengths are stated in the database and therefore interpretation and quality of the query results. Ideally the users should be able to access the original image so they can process the signal data according to their own filters and preferences.

Solution to Limitation 3: Store the original image in a separate database such as a BLOB (binary large object) database or similar means of accommodating the file outside of the main database implementation. Provide links to the stored original image for each experiment as a key in the relational tables of the main database. The users may then choose to download the file and process it themselves.

7.2.4 Limitation 4: Lack of centralized and consistent data analysis

Impact on querying: The overall approach in microarray database implementations assumes many independent users performing data visualization and analysis using a diverse set of tools and database implementations. Goncalves et al [GW02] identify some important limitations associated with that approach: “No central data storage unit and no data version control means that multiple different data files and versions of analysis files may be located in different directories or even on different computers. Lack of uniform experiment annotation and description of the analytical procedures employed means that collaboration and data sharing are severely limited.

Lack of integration between the various steps of the analysis makes data validation almost impossible (e.g., it is extremely difficult to go from cluster visualization to the individual spot images in order to validate the quality of the result). Last but not least, this analysis model is not scalable.”

Solution to Limitation 4: At the time of this writing the overall solution approach remained for the most part academic. The overall solution is three fold: 1) develop and enforce data format and data exchange standards to meet the core need for research exchange and collaboration 2) standardize semantics and ontology so that descriptions and meaning are readily understandable within the exchanged data sets

Table 10: Design Limitations of Current Implementations

#	Limitation	Proposed solution
1	Use of web based forms for data retrieval and presentation	Several microarray database websites also offer or recommend separate locally installed programs to provide flexible alternatives to the limitations of web browser interfaces for data visualization.
2	Use of hyperlinks to external databases such as GenBank as a mechanism for providing annotation data about genes	Adjusting the format of the data to accommodate native data structures should be done. In some cases it may be helpful to implement two or more data exchange models to accommodate a variety of annotation data.
3	Database does not store original microarray image because of its large size	Store the original image in a separate database such as a BLOB (binary large object) database. Provide links to the stored original image for each experiment as a key in the relational tables of the main database.
4	Lack of centralization and consistency in data analysis	The overall solution is three fold: 1) develop and enforce data format and data exchange standards 2) standardize semantics and ontology and 3) design of a centralized repository or public data warehouse.

and 3) design of a centralized repository or public microarray data warehouse providing integrated raw image data, annotation data, detailed records of tasks performed in data analysis, and support for research collaboration. Table 10 provides summary of the limitations and solutions covered in this section.

7.3 Inherent Limitations in Microarray Databases

Inherent limitations are those that have no simple solution. Some of these limitations are due to the nature of microarray technology and will only be solved by advances in the technology. Other limitations are due to the nature of how these microarray database implementations accommodate the needs of their users. We provide an outline of some important examples below.

- 1) mRNA is not an exact indicator of protein concentration. Gene expression is only estimated using DNA microarrays, better estimations come from emerging protein array technology but such databases are also emergent designs so microarray databases are the most important source of expression studies. Protein array databases are described in appendix D.
- 2) Lower limit on sensitivity of signal detection. Microarrays are not able to identify or detect very low copy number for some mRNA, therefore PCR verification needed to identify genes that provide only very low concentrations of mRNA.
- 3) Fundamental issues for implementations concern both diversity of platforms and lack of compatibility. Microarray database implementations can be relational databases using SQL, object-relational, or object oriented databases therefore a

query written in a particular language for one database may not be useable in many others. Because many implementations are tailored to the immediate needs of particular research groups within an institution, this diversity is inherent. It is difficult to compare data and pool data from different microarray databases. This is one of the most important concerns cited in publications discussing microarray databases. Possibly in the future a common architecture and data model will be adopted across all implementations, but it is not feasible or planned for the near future. However, improving compatibility for data formats and data exchange is being actively addressed.

- 4) Must not assume user knowledge of data structures, implementation design, or query languages. The requirement is very important and has been extensively addressed in technical literature. Making query handling transparent to the user is considered a core requirement for bioinformatics database design. Further, publications describing these implementations do not describe query handling, or optimization and rarely describe relational table structures. The consensus in the bioinformatics community is that since users are typically scientists with limited understanding of databases, so it is essential to simplify the user interface. Meeting the requirement restricts query interface design. Each database implementation examined for this work has a set of predefined queries and navigational selection of parameters to form the query. They do not permit the user to customize them.

7.4 Successfully Addressed Limitations

Our focus in this chapter is on those identified limitations for which we are able to propose reasonable solutions. Inherent limitations or constraints on using microarray databases are particularly difficult to address, these have been described in section 2.4 already. Successfully addressed limitations are those for which a reasonable

Table 11: Minimum Data Descriptors Problem and Solution

Limitation 7.4.1	Lack of standards for uploading entries and exchanging data
Impact	The lack of standards results in difficulties comparing entries and consistency in completeness of minimum information from different databases. Perhaps more important is the need for different databases to be able to exchange new data so that they are as complete and current as possible. This is one of the most important areas of concern in microarray database implementations and therefore a strong effort has gone towards the solution.
Solution	A minimum set of descriptors for a complete experiment data set has been defined. These are known as MIAME standards. MIAME compliance is now an important parameter for microarray database comparisons, as these standards are being implemented or will be implemented in most microarray databases. See section 3.2 and appendix B for more details regarding MIAME standards. [BHQSS01]

solution has been identified and made available. Adoption of the solution is either complete or planned in the future by most of the database implementations. These are no longer true limitations unless new databases do not adopt the solutions. Examples of

successfully addressed limitations appear as tables above. The first example (table 11) describes the problem of defining what minimum information to provide for a database entry record. The second example (table 12) describes the problem of determining what metadata structure to use for uploading new entries, data storage, and exchange of records between databases.

Table 12: Summary of Metadata Format Problem and Solution

Limitation 7.4.2	Use of tab delimited text files to store and transfer metadata
Impact	Although this is a commonly implemented method for data transfer, it poses two key problems: 1) it does not provide a common organization for the data which would simplify information exchange among microarray databases, and 2) neither image data from sample annotations nor experiment annotations can be included in a tab delimited text file. This second problem is the more serious limitation of the two.
Solution	The use of more descriptive metadata structures such as XML provides an effective solution to this problem. Published articles for the earliest microarray database implementations in the years 2000 and 2001 regard the use of a microarray data specific XML exchange format as one of their core future goals. Since that time MAGE-ML has emerged as the standard, and use of MAGE-ML is either currently employed or a future goal for many implementations. See section 3.3 and appendix C of this work for more details regarding MAGE-ML format. [SMSTSC02]

The third table (table 13) is a summary of current status on implementing the solutions described in tables 11 and 12.

Table 13: Current status for Successfully Addressed Limitations

Database Name	Array Express	CEBS	GEO	EMAP	SMD	HugeIndex
MIAME compliant	Yes	Yes	Yes	Yes	Future goal	Future goal
Store data as MAGE-ML	Yes	Yes	Future goal	Yes	Yes	Future goal

CHAPTER 8

CONCLUSION

In recent years biological databases have gained importance in research. The large volume of biological data generated by high throughput automated processes such as the human genome project can only be effectively managed by databases. The earliest biological databases stored DNA sequences, relatively simple data structures. Since that time the volume of more complex data such as distribution patterns of physical molecules, and structural data of biological entities has also rapidly increased in volume. Microarray experiment data is an excellent example. The more complex types of biological data require databases able to store representation of graphics and images, three dimensional structures, or time dependencies. These requirements makes complex biological data types unsuited to classical relational data models. Database technology has in recent years begun to catch up with the requirements of molecular biology data in general, and of microarray data in particular. However, no single database implementation approach is suited to all needs even within a subdiscipline such as microarray research. Every special area of study presents its own particular requirements.

As of 2005 over 700 publicly available web-based molecular databases have been implemented, representing many diverse subfields. Among these, we have selected microarray databases for the focus of this study. Microarrays represent the entire genome, all genes in an organism. A microarray database is a unique tool for the simultaneous study of hundreds of thousands of genes, and exploration of gene function and gene interaction on a huge scale. We have selected these databases for three primary reasons. Firstly, microarray databases are representative of recent technology, with most implementations having begun between 1 to 4 years ago. Since the technology is still in its infancy there is need for better understanding and improvements. Secondly, microarrays have had a particularly important role in genetics research since the completion of the Human Genome Project in 2003. Thirdly, microarray data types are complex, requiring graphical representation of expression patterns, statistical normalization of the raw data, and extensive annotation for the identified genes. Because of this complexity, microarray data poses interesting challenges for query composition, query interfaces, and results output.

In this thesis we have described and identified the important factors required for effective querying of microarray data. Ease of use is a particularly important challenge due to the highly specialized nature of the data and the queries. Making it possible for scientists to easily construct useful queries has proven a challenge for two primary reasons. Firstly, the queries placed on microarray data originate from very specialized scientific questions. And secondly, the users cannot be presumed to have any knowledge of the implementation, data structures, or understanding of a query

language. Designers and implementers have recognized that the biological researchers who use these databases should not be assumed to understand data structures, or formal query languages. Among all implementations queries are constructed based on predefined selection criteria. Defining and selecting those criteria is achieved by identifying the requirements for the meaningful characterization of a microarray experiment into data that can be searched and queried. These requirements include representing both raw and adjusted image data from fluorescent signals on the array, experimental parameter data, and gene annotation data. Biological researchers query microarray databases to find results that either support or refute a scientific hypothesis. Those result sets are not always specific answers in themselves but present valuable correlating patterns or trends. We have found that interactive software tools for graphical pattern analysis and visualization of the genetic expression pattern output are particularly important to improving the value of microarray query results to the user.

The limitations identified here have been grouped into three major categories. 1. Those limitations affecting querying capability directly, such as inconsistencies in data identification or lack of support for parameters such as querying rate of change on time-varying image data. 2. Those limitations affecting the implementation, and therefore affect querying indirectly. For example, lack of centralization and consistency in data analysis that inhibits data validation and data exchange. 3. Inherent limitations for which there are no practical solutions, such as the limits of the microarray technology itself. We have proposed solutions for each of the limitations we identified. For example, to better accommodate free text queries we suggest synonym mapping so that

terminology having closely related concepts could be matched to a likely result of interest.

In our approach we have attempted to understand the characteristics of the data to be queried, to identify challenges to effective data management, to understand the needs of intended user groups, and the purpose for the query results. In so doing we have been able to assess limitations and suggest solutions. Although this thesis describes and examines the specific case of microarray databases, the approach taken could be applied to evaluating and identifying areas of improvement in other categories of bioinformatics databases. For example, protein array databases that facilitate the study of entire protein populations (or *proteomics*, discussed further in appendix D). These are still emergent technology and present additional challenges beyond those of microarrays. In future work we would like to examine this and other categories of biological databases.

APPENDIX A
LIST OF MICROARRAY DATABASES

The following is a list of publicly available microarray databases currently available through internet access. This table is adapted from the 2005 Database Collection [G05] published by Nucleic Acids Research. The journal annually publishes a summary list of all publicly accessible internet based implementations of biological databases.

Table 14: List of Microarray and Gene Expression Databases in 2005

	NAME	DESCRIPTION	WEBSITE
1	5'SAGE	5'-end serial analysis of gene expression	http://5sage.gi.k.u-tokyo.ac.jp/
2	ArrayExpress	Public collection of microarray gene expression data	http://www.ebi.ac.uk/arrayexpress
3	Axeldb	Gene expression in <i>Xenopus laevis</i>	http://www.dkfz-heidelberg.de/abt0135/axeldb.htm
4	BodyMap	Human and mouse gene expression data	http://bodymap.ims.u-tokyo.ac.jp/
5	BGED	Brain gene expression database	http://love2.aist-nara.ac.jp/BGED
6	CleanEx	Expression reference database, linking heterogeneous expression data to facilitate cross-dataset comparisons	http://www.cleanex.isb-sib.ch/
7	dbERGEII	Database of experimental results on gene expression: genomic alignment, annotation and experimental data	http://dberge.cse.psu.edu/menu.html
8	EICO DB	Expression-based imprint candidate organiser: a database for discovery of novel imprinted genes	http://fantom2.gsc.riken.jp/EICODB/
9	emap Atlas	Edinburgh mouse atlas: a digital atlas of mouse embryo development and spatially mapped gene expression	http://genex.hgu.mrc.ac.uk/
10	EPConDB	Endocrine pancreas consortium database	http://www.cbil.upenn.edu/EPConDB
11	EpoDB	Genes expressed during human erythropoiesis	http://www.cbil.upenn.edu/EpoDB/
12	FlyView	<i>Drosophila</i> development and genetics	http://pbio07.uni-muenster.de/
13	GeneAnnot	Revised annotation of Affymetrix human gene probe sets	http://genecards.weizmann.ac.il/geneannot/
14	GeneNote	Human genes expression profiles in healthy tissues	http://genecards.weizmann.ac.il/genenote/
15	GenePaint	Gene expression patterns in the mouse	http://www.genepaint.org/Frameset.html
16	GeneTide	A transcriptome-focused member of the GeneCards suite	http://genecards.weizmann.ac.il/genetide/
17	GeneTrap	Expression patterns in an embryonic stem library of gene trap insertions	http://www.cmhd.ca/sub/genetrap.asp
18	GEO	Gene expression omnibus: gene expression profiles	http://www.ncbi.nlm.nih.gov/geo/
19	GermOnline	Gene expression in mitotic and meiotic cell cycle	http://www.germonline.org/
20	GXD	Mouse gene expression database	http://www.informatics.jax.org/menus/expression_menu.shtml
21	H-ANGEL	Human anatomic gene expression library	http://www.jbirc.aist.go.jp/hinv/index.jsp
22	HemBase	Genes expressed in differentiating human erythroid cells	http://hembase.niddk.nih.gov/

Table 14 – Continued

	NAME	DESCRIPTION	WEBSITE
23	HugeIndex	Expression levels of human genes in normal tissues	http://hugeindex.org/
24	Kidney Development Database	Kidney development and gene expression	http://golgi.ana.ed.ac.uk/kidhome.html
25	LOLA	List of lists annotated: a comparison of gene sets identified in different microarray experiments	http://www.lola.gwu.edu/
26	MAGEST	Ascidian (<i>Halocynthia roretzi</i>) gene expression patterns	http://www.genome.ad.jp/magest
27	MAMEP	Molecular anatomy of the mouse embryo project: gene expression data on mouse embryos	http://mamep.molgen.mpg.de/
28	MEPD	Medaka (freshwater fish <i>Oryzias latipes</i>) gene expression pattern database	http://www.embl.de/mepd/
29	MethDB	DNA methylation data, patterns and profiles	http://www.methdb.de/
30	Mouse SAGE	SAGE libraries from various mouse tissues and cell lines	http://mouse.biomed.cas.cz/sage
31	NASCarrays	Nottingham <i>Arabidopsis</i> Stock Centre microarray database	http://affymetrix.arabidopsis.info
32	NetAffx	Public Affymetrix probesets and annotations	http://www.affymetrix.com/
33	Osteo-Promoter Database	Genes in osteogenic proliferation and differentiation	http://www.opd.tau.ac.il
34	PEDB	Prostate expression database: ESTs from prostate tissue and cell type-specific cDNA libraries	http://www.pedb.org/
35	PEPR	Public expression profiling resource: expression profiles in a variety of diseases and conditions	http://microarray.cnmcresearch.org/pgadatatable.asp
36	RECODE	Genes using programmed translational recoding in their expression	http://recode.genetics.utah.edu/
37	RefExA	Reference database for human gene expression analysis	http://www.lsbm.org/db/index_e.html
38	rOGED	Rat ovarian gene expression database	http://web5.mccs.uky.edu/kolab/rogedendo.aspx
39	SAGEmap	NCBI's resource for SAGE data from various organisms	http://www.ncbi.nlm.nih.gov/SAGE
40	SIEGE	Smoking Induced Epithelial Gene Expression	http://pulm.bumc.bu.edu/siegeDB
41	Stanford Microarray Database	Raw and normalized data from microarray experiments	http://genome-www.stanford.edu/microarray
42	Tooth Development Database	Gene expression in dental tissue	http://bite-it.helsinki.fi/

APPENDIX B

MIAME STANDARDS FOR MICROARRAY DATA

The Microarray Gene Expression Data (MGED) Society is an international organization of molecular biology researchers, computer scientists, and data analysts whose main goal is to facilitate sharing of microarray data for the study of functional genomics and proteomics. MIAME is one of six standardization projects being pursued by researchers in the group. Provided below are the MIAME Standards for Microarray data as proposed by the EBI at their July 1999 conference meetings and released in July 2000. The standards document which follows is also publicly available from the following link: <http://www.mged.org/Workgroups/MIAME/>.

The meeting discussed draft recommendations to the microarray community proposed by the EBI and established a general consensus detailed below. These recommendations should not be regarded as an official view of the meeting, but as a starting point for wider discussions in the microarray community.

- Establishing a well-organized public repository for gene expression data will provide the bioinformatics community with a powerful tool. Establishing such a repository would be facilitated by:
- accepting a standard for the minimum information that laboratories should be encouraged to provide about microarray based experiments, to ensure reproducibility of the results;
- defining the data communication standards for such experiments;
- developing ontologies for sample description;
- developing standards for normalization, quality control, and cross-platform data comparison for microarray based experiments;

The minimum information about a published microarray based gene expression experiment should include:

1. expression level **measurement results**, in particular:
 - a. the TIFF image file from the hybridized microarray scanning;
 - b. the image analysis output (of the particular image analysis software) for each spot, for each channel;
 - c. a derived value summarizing each spot in the authors interpretation (e.g., a background subtracted intensity typically used for Stanford or Incyte technologies);
2. the following **annotations**:
 - a. array (e.g., platform type, substrate, number of spots, provider),
 - b. each element (spot) on the array (e.g., sequence or clone and relevant accession numbers),
 - c. sample source and treatment (e.g., organism, development stage, tissue, drug treatment),
 - d. controls in the sample and on the array,
 - e. hybridization extract preparation (e.g., cell rupture method, nucleic acid extraction and labeling protocol),
 - f. hybridization procedure (e.g., time, concentration, volumes, washes),
 - g. scanning procedure (e.g., hardware, output TIFF file header),
 - h. image analysis and quantification (e.g., software, version, parameters),
 - Also, MGED would like to encourage the image analysis software developers to try to design methods for standard ways of summarizing spot quality.
 - i. description of the experiment as a whole (e.g., set of related samples and hybridizations submitted together and their relationships [time series, comparative hybridizations], reference if published).

The meeting accepted the items 1a) – c) and 2a) – i) by consensus. There were two general opinions about the detailed specifications of each of the subitems. A clear majority considered the level of detail given in the "Details of the minimum

information" document is close to the minimum that has to be provided about any published experiment. Nevertheless, there were a considerable number of participants, who considered the proposed details excessive. It was agreed that the details will be specified by working groups and by e-mail discussion and proposed for discussion at a follow-up meeting.

It was agreed by consensus, that once the definition of the minimal information about a *public* experiment is accepted by the community and public repositories supporting this specification are established, journals should be encouraged to require data submissions to a public repository, where the information can be confidential until the publication.

Data storage and communication standards

1. A standard XML-based flat-file format for microarray data description and exchange, compatible with the minimum information definition discussed above, should be developed and accepted by the community. This will formalize the definition of the minimum information, as well as open a way to populate public repositories directly from laboratory databases and LIMS systems.
2. It was proposed that:
 - the flat-file format should support simultaneous submission of data from multiple experiments (i.e., unrelated hybridizations), to facilitate the uploading of data from laboratory databases into public repositories;
 - the working group for data communication standards consider ways that might allow the standards developed for data from microarray expression experiments to be extended to cover data from other kinds of microarray experiments;

- ideally, the format should support a possibility of back-referencing to items submitted to a public repository earlier.
3. A working group for developing XML standard was established at the meeting.

The standard will be reviewed and accepted in a follow-up meeting.

Ontologies for sample source and treatment description.

Ontologies should be used for sample source and treatment description (e.g., organism, development stage, tissue, cell line type, cell line, treatment type) where possible. In particular, MGED use collections of categories, each of which have their own controlled vocabularies, where the categories are themselves organized, e.g., as a tree.

1. Universally accepted ontologies or standard vocabularies currently do not exist, except for description of species (Taxonomy database). Ontologies for developmental stages and tissues are relatively well described for some organisms, mouse and fruit fly in particular.
2. A working group was established to consider where introduction of an ontology is possible, and ways achieving this. It is not feasible for the working group to develop the final ontology for any new category of sample description, but rather to:
 - identify categories which should be included in sample source and treatment description;
 - identify and review relevant ontologies developed by independent groups;

- identify the subset of required categories that can be covered by incorporating and adapting available ontologies, and identify provisional means of handling remaining categories;
 - document issues pertinent to use of other ontologies, and issues and possible approaches for fuller treatment of provisionally handled categories. The identification of high level categories and nodes where controlled vocabularies are possible will be considered for these latter categories.
3. Recommendations from the working-group will reflect on the minimum information definition and on data exchange standard.

Data normalization and cross-platform comparison

1. The microarray community should determine common controls for their arrays and experiments. In particular there may be two types of controls:
 - normalization controls
 - quality controls
2. Experiments in the public domain comparing different platforms for designing cross-platform normalization procedures should be encouraged;
3. The meeting established a working group that will develop detailed recommendations for normalization, quality control and cross-platform comparison, which will develop more detailed recommendations before the next meeting.

Database population and data submission issues.

1. XML based flat-file format will be a relatively straightforward and easy way of submission by e-mail or ftp download, enabling direct submissions from laboratory databases and LIMS systems.

2. Client side data submission tools (either Web-based or stand alone) would complement such flat-file based submissions. Ease of use and the ability to back-reference objects from the database will be essential.
3. Information about experiments and arrays may be submitted separately, with the array description being within the same or prior submission from experiments using them;
4. Use of standard protocols for hybridization extract preparation, hybridization, scanning, and image analysis should be considered. Scanning hardware and image analysis software producers should be encouraged to accept relevant standards.
5. Ideally, the database should support the reuse of objects submitted in earlier experiments (e.g., extraction and hybridization protocols), which would facilitate standardization of these categories. The XML data exchange format should support such "back-references".
6. The minimal information specified in the first section of this document should be provided by the submitter and supported by a public repository.

Data curation, quality, and ownership in a public repository

Database administrators, submitters, and users should take steps to assure the quality of data on the database.

1. Administrators of an open public database cannot police quality data, but can and should:
 - verify that data meets the minimal information requirements given above and meets obvious data consistency checks. Where possible this should be done through automated checking at the time of data submission;
 - flag database entries based on appropriately defined and accepted experimental quality assessment indicators. Possible bases for such indicators might include replication of experiments, use of recommended controls, publication of experiment in a peer-reviewed journal;
 - reserve the right to remove from the database entries that have turned out to be obviously wrong. To work out formal criteria for making such conclusions may be difficult, however;
2. Submitters of data to the database should be willing and able to update data that have proved to be in error on later analysis. For instance, if, after an experiment using an array has been loaded on the database, DNA sequencing proves a spot on the array to be unreliable, the submitter should be able to update this on the database;
3. Users of the database should be able to submit annotations. These should be identifiable as third-party annotations;
4. To ensure quality control in the early stages of the database development, administrators may at first accept data from selected collaborators. When the database reaches development stability, data submissions will be made open and public. Database submissions should be open to the whole community before they can be made obligatory prerequisite by journals.

APPENDIX C

MAGE-ML: XML FOR MICROARRAY DATA

Microarray Gene Expression Markup Language (MAGE-ML) "is a language designed to describe and communicate information about microarray based experiments. MAGE-ML is based on XML and can describe microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results. MAGE-ML has been automatically derived from Microarray Gene Expression Object Model (MAGE-OM), which is developed and described using the Unified Modelling Language (UML) -- a standard language for describing object models.

Descriptions using UML have an advantage over direct XML document type definitions (DTDs), in many respects. First they use graphical representation depicting the relationships between different entities in a way which is much easier to follow than DTDs. Second, the UML diagrams are primarily meant for humans, while DTDs are meant for computers. Therefore MAGE-OM should be considered as the primary model, and [the MGED committee report authors] will explain MAGE-ML by providing simplified fragments of MAGE-OM, rather than XML DTD or XML Schema." (from the description by Ugis Sarkans, [SMSTSC02]). The standards description that follows is an excerpt from the documentation publicly available from this link: <http://www.mged.org/Workgroups/MAGE/mage.html>.

The Minimum Information About a Microarray Experiment, also known as MIAME, was developed to specify which microarray experiment data and metadata should be reported to enable others to understand and interpret the experiment unambiguously. This is a data content standard, not a format standard.

Microarray Gene Expression Markup Language (MAGE-ML) is a formal language designed to describe and communicate information about microarray based experiments. MAGE-ML is an XML language; it can be used to describe microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results.

MAGE-ML has been automatically derived from Microarray Gene Expression Object Model (MAGE-OM), which is developed and described using the Unified Modelling Language (UML) – a standard language for describing object models. Models described using UML have advantages over pure XML technologies (DTDs or XML Schemas) in many respects, especially for didactic purposes. They use graphical representation depicting the relationships between different entities in a way which is much easier to follow for a human than DTDs. The idea behind UML diagrams is to provide a way of describe models that is both human readable and has strict semantics, while DTDs and XML Schemas are meant primarily for computers. Also, complex models (also MAGE) involve many different types of relationships between model elements, while in XML by definition information is encoded in a hierarchical manner and relationships that break the hierarchy need to be encoded in some special ways.

MIAME requires detailed annotation about experimental conditions, materials and procedures to be captured. MAGE-ML is a rich format. By using it one can encode MIAME-required information and more. The purpose of this document is to provide guidance for encoding MIAME-required information in MAGE-ML.

MIAME consists of 5 sections, [the MGED committee report authors] will follow that structure here. For each MIAME section the following is presented:

- 1) one or more UML class diagrams containing a subset of classes and associations from the corresponding MAGE-OM package(s) needed for MIAME-compliant data encoding;
- 2) a fragment of the simplified MAGE-ML DTD (MGED will call this here MAGE-ML-Lite) that is sufficient for encoding MIAME;
- 3) an informal object diagram that illustrates the structure needed for MIAME encoding in MAGE objects;
- 4) a sample MAGE-ML document template that corresponds to the object diagram.

On the class diagrams [the MGED committee report authors] have only deleted some classes and associations that are less relevant for MIAME encoding, but [the MGED committee on MAGE-ML] haven't made any structural changes. In fact, the diagram layout is the same as in the formal MAGE-OM specification.

The object identifiers for MAGE objects should have the form:

`<authority>:[<namespace>]:<object>[:<revision>]` where ":" is the field separator, "`<...>`" is a string and "`[...]`" represents an optional elements. Syntactically none of strings used in 'authority', 'namespace', 'object' and 'revision' is allowed to contain ':'. For the time being `<authority>:<object>` also is acceptable from current MAGE-ML exporters, but [the MGED committee on MAGE-ML] would recommend that submitters strive to conform to the specified format with null namespace, e.g. `<authority>::<object>`. If submitters don't have a meaningful namespace then the recommended format is `<authority>::<object>`.

<authority> is assigned by the data provider. [The MGED committee on MAGE-ML] recommend that this is done in a way that minimises the possibility of clashes, for instance following the DNS model with the providers giving names like "ebi.ac.uk", "umich.edu", "genetics.umich.edu" or "lab23.genetics.umich.edu".

[The MGED committee on MAGE-ML] recommend that the software manufacturers include the assignment of the authority during the installation of each particular copy of their software. The defaults should be set in a way that minimise the clashes. If there is a 'authority' clash during the submission to a public database (e.g., ArrayExpress), [the MGED committee on MAGE-ML] will try to resolve this via MGED. The mass software should ideally have an option which allows to change the 'authority' after the installation.

Regarding [`<namespace>`]:`<object>`[`:<revision>`], the only requirement at the moment is that objects should be guaranteed to be unique within the authority.

APPENDIX D
PROTEIN MICROARRAY DATABASES

A) INTRODUCTION

Just as genomic microarray data has a more complex and newer counterpart to protein microarray data, the graphical databases that store genomic microarray data have multimedia database counterparts to store protein microarray data. In this appendix we will review the newest type of microarray technology, for protein microarrays.

Microarray databases are among the first bioinformatics implementations to use searchable graphics. Queries are on the scale of a complete genetic profile for an entire cell or entire collection of cell types within an organism. The next step in visualization is to be able to search not just representations of images but the images themselves. In this regard multimedia databases are highly suited to bioinformatics. The next step in studies of an entire genetic expression profile is to not only look at which genes are expressed but look at the proteins those genes encode. A complete protein population or proteome is not exactly equivalent to a population of expressed genes. Protein interactions, distribution, and regulation differ from those of the expressed genes or mRNA transcripts which encode them. In appendix D we provide a detailed look at the importance of proteomics next steps in the field of genetic expression studies and the use of multimedia databases that support that research. In section B) we will first identify the relevance of multimedia databases to bioinformatics data. Then in section C) we identify some of the biological research questions that are suited to multimedia databases. From recent research efforts, we focus on location proteomics as an excellent example of how to construct an automated searchable Bioinformatic image

database. The research group lead by Robert Murphy at Carnegie-Mellon University provides the well studied case of fluorescent microscopy location proteomics. Exploring their work in sections D) and E) we examine (1) data type identification, (2) basic approach to indexing, (3) identification of query types, (4) implementation of a search tool, and (5) future goals. Section F) provides a summary.

B) ROLE OF MULTIMEDIA DATABASES FOR BIOLOGICAL DATA

Multimedia databases contain diverse data types including images, video, and audio. The Internet is an archetypical distributed multimedia database. Large variation in file size and complexity of the data is handled through the appropriate metadata structures. A typical metadata structure used for Internet files is XML. Because the data types are complex, traditional relational query techniques are not suitable. New query techniques are being developed suited to the high dimension spaces of multimedia, and multimedia content based retrieval. Two examples of recent research in querying and retrieval for multimedia are 1) developing distance metrics for nearest neighbor queries in high dimensions, and 2) the use of natural language to search for an object by name through indexing the features of an image.

Broadly, there are two areas of multimedia database technology particularly relevant to biological data, 1) static image databases (two-dimensional or *2D*) and 2) virtual reality (three dimensional or *3D*) which may include animation. There are many open questions related to indexing of complex image data, and how to construct queries for this type of data. Analysis of image data is important in biological research. Within the last few years advancements in multimedia database technology have been

accompanied by interest in its application for biological databases. Virtual Reality implementations are few and query techniques are largely confined to simple navigation, they are used as visualization tools for the content of well established relational databases. Virtual Reality graphics may have a significant role for improved usability of the most widely accessed databases which are presently limited to simple flat file and textual data retrieval. Image databases clearly have important roles and complex querying techniques have been developed for detecting patterns in both individual molecules and localization of particular proteins at the subcellular level.

Biological databases traditionally provide textual descriptions of data. A classic example are the DNA sequence databases. Even storage of images is relatively new. The concept of queries based on comparison of images is now an area of current research. Many areas of biological research depend on information from images. For image data, the manual mode of analysis has been the only option until very recently. It is limited in its efficiency and scale. Using multimedia databases with new techniques for handling image features would overcome that limitation.

C) BIOLOGICAL DATATYPES SUITED TO MULTIMEDIA DATABASES

What are some of the areas in biological research that would benefit from automated comparative image analysis, indexing, and query capability?

C1. Cell populations

As a first example, considere patterns of cell populations. Many diseases are related to protein interactions resulting in cell death. Comparing images of the cells answers whether the cells die in patches, or as a wave, and at which layers. Image

analysis also reveals what patterns of cellular migration, cell type association, and cell division occur in different disease states. Studies in biomolecular research ranging from cancer and autoimmune disorders, to physiology, and drug discovery all require these types of image pattern comparisons. Yet these are queries that conventional databases are not able to support.

C2. Location proteomics

In addition to comparison analysis, querying image patterns is important to the study of a new area in molecular biology known as location proteomics. A proteome is the complete protein profile, or the complete protein population. This will not represent all genes (the genome), since only a small subset of genes are used to make proteins at any one time. This may be in reference to either a cell, or an organism. Location proteomics is the study of the distribution properties and statistics on features of proteins in their cellular context. This type of information provides an understanding of the normal spatial and temporal patterns of protein distribution, to which a comparison can be made and therefore guides both experimental design and results interpretation. These types of information can be analyzed effectively only through image data.

C3. Whole genome analysis

A related area in image based queries, is the 3D or Virtual Reality technology for visualizing large data sets, entire existing databases, or complete systems. Again, these provide a unique and important supplement for researchers to learn from large bioinformatics databases. Through this new technology a global view of large scale patterns in comparison of DNA which textual queries are unable to accomplish.

D) OVERVIEW OF LOCATION PROTEOMICS

Murphy [MF05] makes the observation that “the focus of most biochemical research is now shifting from simply identifying gene sequences to determining the properties and functions of the proteins encoded by those genes.” From this shift we have the emerging field of proteomics which is a subfield in the wider arena of genomics. The suffix -omics refers to study of the complete system, for example the set of all proteins needed by a cell or by an organism. Location proteomics is an important area of research for mapping proteins to where they are localized in a cell. The cell is compartmentalized into different distinct subregions with specific unique tasks for cell maintenance. For example in eukaryotes (all non-bacterial cells) these structures include nucleus, mitochondria, lysosomes, Golgi apparatus, cytoskeleton, and the endoplasmic reticulum. Of fundamental importance to the adaptability and diversity of cell types is this compartmentalization.

The diversity of cell types results in specialized functions in particular tissues. Note that pancreatic cells, muscle cells, skin cells, nerve cells are all very distinct in their functions, shape, and capabilities. The ability to locate where a particular protein is found within a cell and also which cells use that protein in different types of tissue or organs is important to biologists. Additionally, researchers are interested in when the proteins are required and synthesized, which combinations are synthesized and at what concentrations. Medical problems can result from incorrect concentrations, incorrect timing or synchronization in synthesis, or incorrect location as well as from improper shape or improper function of a particular protein. Proteomics as a field is dependent

upon genomics. It is only in the last 5 years that, along with the rapid elucidation of genetic sequences, proteomics data has exploded into a huge amount of information, new technology and large volumes of data. The number of proteins exceeds the number of genes. In humans 30,000 genes are estimated to encode approximately 100,000 proteins. This is simply because the intermediate step in gene to protein (mRNA strands) can be biologically edited, with sections being deleted before it is used as a template for protein production. Similarly proteins themselves are often shortened, chemically modified, or consist of multiple independent protein strands.

Understanding the patterns of protein begins with the raw data, in contrast to other types of data on proteins such as sequence, binding partners in pathways, or metabolic activities, subcellular location has received little attention in the past partly because information and data was restricted to unstructured text in journal articles. An additional limitation of text is the lack of consistent terminology for subcellular location, despite efforts such as the GO Cellular Component Ontology vocabulary from the Gene Ontology (GO) Consortium. Query processing and query results retrieval become dependent upon qualitative terms, the table below provides an example of

Table D1: Example comparison of variation in terminology for protein location

Protein	giantin	Gpp130
Accession	Swiss-Prot Q14789	TrEMBL o00461
Comments: Subcellular location	Golgi; membrane-associated	(none)
GO Cellular component terms	0000139, Golgi membrane; 0005795, Golgi stack; 0016021, integral to membrane	0030139, endocytic vesicle; 0005801, Golgi cis-face; 0005796, Golgi lumen; 0016021, integral to membrane

textual descriptors for two proteins found in the cellular structure Golgi from which it can not readily be determined if the patterns overlap if so to what degree and what are the relative concentrations. The variation in terminology used to describe subcellular location in protein databases is illustrated by table D1 below [M04].

Other types of proteome queries can not be easily answered without image databases, for example do proteins found in the lysosome have similar distribution and concentration patterns as for endosomes? Image databases provide quantitative answers rather than qualitative. Further, by interpreting and mapping protein families at the level of a complete proteome, errors can be recognized and corrected. For example, a protein found in the mitochondria may be incorrectly assigned to another organelle and this mistake may be identifiable by its distinctive location pattern.

D.1) PROTEOMICS DATA TYPES

The most current technology typically generates proteomics data from the following five techniques:

- 1) high performance liquid chromatography (HPLC) for separation, a technique well established in chemistry
- 2) mass spectrometers for identifying individual proteins by weight and composition by each type of atom
- 3) protein microarrays (protein biochips) for measuring concentrations and interactions between proteins these were introduced commercially in 2002 as the BioPlex chip by Bio-Rad and are still actively researched.
- 4) two-hybrid systems both yeast for eukaryotes and bacterial for prokaryotes. These are genetically recombinant cell lines which enable study of protein binding and biochemical pathway elucidation in a living cell. In these systems GFP (green fluorescent protein) is synthesized attached to a protein of interest in a living cell. The protein of interest is then observed in real time within the living cell, it is assumed the fluorescent component of the hybrid protein has minimal or no impact on the biological activity of the protein.

- 5) atomic force microscopy (AFM) in which an electrode tip probes the surface of a cell or sample. The tip detects surface properties including ionic charge gradient or charge density, magnetic field, temperature, and topography [SMM04], this is one of the most recent developments in protein research.

Regarding protein microarrays it is important to note that unlike DNA, proteins are very sensitive to slight changes in their chemical surroundings which impact their ability to bind, therefore protein microarrays are still facing many challenges compared to the corresponding DNA microarray technology. This technique is based on fluorescent molecules attached to antibody proteins which are synthesized to specifically bind a protein of interest, these fluorescent antibodies are known as probes. A z-series of images taken at different positions in the z-axis of a cell allows three dimensional mapping of protein location in the entire cell, known as laser-scanning confocal microscopy. In some cases more than one protein may be observed simultaneously, each having a unique color for its fluorescent marker. Additionally, some antibodies are specific to a protein in a particular conformation or shape from a biochemical interaction.

D.2) PROTEIN IMAGE DATA INDEXING

Murphy et al [MF05, HM04, M04, MKHJC04] consider how the microscopy images of cells can form a multimedia bioinformatics database for proteomics research. Specifically, their research has developed an automated systematic analysis of the images so the image data can be classified and retrieved from an image database. Current practice in Biological Research assumes visual inspection of these images. Comparisons are made between known locations of markers and locations of the protein

being studied. Automated classification technology would replace the currently used manual system of visually inspecting images then assigning descriptions to them from a small vocabulary. The manual approach is time consuming and will not scale well. The automated system developed by the Murphy group was reported to distinguish similar images of proteins giantin and gp130 at 97% accuracy where a human observer was unable to distinguish them [M04].

An important first step to building queries on microscopy images is the data structure for such images. Murphy et al have addressed how to automatically recognize features of cell structure from the images. Previous research in this area required either special additional labeling steps for cellular structures such as the nucleus or plasma membrane or required model assumptions making them more restrictive than the approach for Murphy [HM04]. Using ten major subcellular structures they were able to achieve 92% average accuracy for 2D single cell images and 96% average accuracy for

Table D2: Descriptions of multicell morphological features

SLF Index	Multicell Morphological Feature Description
SLF1.3	The average number of pixels per object
SLF1.4	The variance of the number of pixels per object
SLF1.5	The ratio of the size of the largest object to the smallest
SLF7.9	The fraction of the non-zero pixels in a cell that are along an edge
SLF7.10	The fraction of all values in first two bins of the edge intensity histogram
SLF7.11	The ratio of the largest to the smallest value in the edge intensity histogram
SLF7.12	The ratio of the largest to next largest value in the edge intensity histogram
SLF7.13	The edge direction difference
SLF7.80	The average length of the morphological skeleton of objects
SLF7.81	The average ratio of object skeleton length to skeletal convex hull area
SLF7.82	The fraction of object pixels contained within the skeleton
SLF7.83	The fraction of object fluorescence contained within the skeleton
SLF7.84	The ratio of the number of branch points in the skeleton to skeletal length

3D single cell images. Significantly, they report better results from multicellular images than single cell partially due to a greater amount of pattern information. This eliminates the need to segment the images into individual cells. Additionally, the set of cell morphological features they selected are independent of cell rotation, eliminating the need for image orientation before processing. Figure 2 [HM04] lists thirteen of these features that describe properties of the cellular objects, these serve as examples of query parameters in a microscopy database.

D.3) QUERY TYPES FOR PROTEOMICS

Among the feature selection methods they examined, stepwise discrimination analysis (SDA) was reported as the best for subcellular pattern classification. Mapping the dataset from the images to a high dimension data structure will not be considered in detail here since this work is focused on query techniques. The automated classification techniques of Murphy et al provide data enabling queries that could include the following as output:

- 1) ranked images to provide the most representative or typical image
- 2) comparisons of sets of images whose subject matter is a specific protein under different conditions to detect changes as a result of those conditions for example in the absence of presence of a toxin or a pharmaceutical drug
- 3) grouping of proteins in a particular location within the cell and from this provide a tree hierarchy or dendrogram
- 4) content based retrieval for microscopy images from articles in online research journals or offline databases and collections

D.4) IMPLEMENTATION

The Murphy lab has created a prototype agent based service SLIF (Subcellular Location Image Finder) which can locate fluorescent micrographs by searching for articles and processing them. The SLIF service accepts text based queries, which were used to identify relevant PDF format articles from one of the largest publicly available web-based research databases NCBI PubMed Central and in a separate experiment against a single collection of 15,000 indexed articles from Proceedings of the National Academy of Sciences in XML format which can be searched using standard SQL queries.

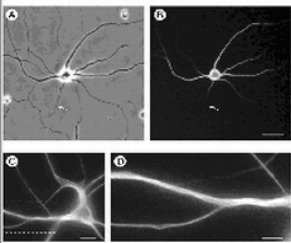
Murphy lab SLIF Service			
Home Search by caption Search by resolution Search by protein name Search FMI figure/panel <input checked="" type="radio"/> Fresh search <input type="radio"/> Search within results	Has FMI panel (s)	Figure	Caption
	Yes	 Click to view a larger image	<p>MAP2 is absent from dendritic spines. (A and B) Like the embryonic low-molecular weight variant MAP2c, high-molecular weight MAP2b is confined to the somatodendritic domain of hippocampal neurons and is absent from the axon (arrow). Shown are both a phase (A) and a fluorescence (B) image of a live neuron transfected with GFP-tagged MAP2b and kept in culture for 14 days. (Bar = 25 μm.) (C and D) This neuron transfected with MAP2b-GFP was maintained in culture for 4 weeks, by which time cells carry many dendritic spines. In the enlarged image (D) of the area outlined in C, the restriction of MAP2b-GFP fluorescence to microtubule bundles in the dendritic shaft is obvious. No fluorescence is detected in spine protuberances from the dendrite. (Bars: C = 15 μm; D = 2 μm.)</p> <p>Click to view the paper</p>
			Fluorescence video microscopy images of cells cultured on flexible polyacrylamide gels (A, B, D, and E) or a

Figure D2: Results set for a serial query using the SLIF web interface. The first query was made on the a PNAS test set [MKHJC04] and is for figures in which caption contained "microtubule," "mt," or "tubulin." A second query on the output from the first query retrieved figures containing a fluorescence microscope image (FMI). Only the first of several figures matching both queries is displayed here.

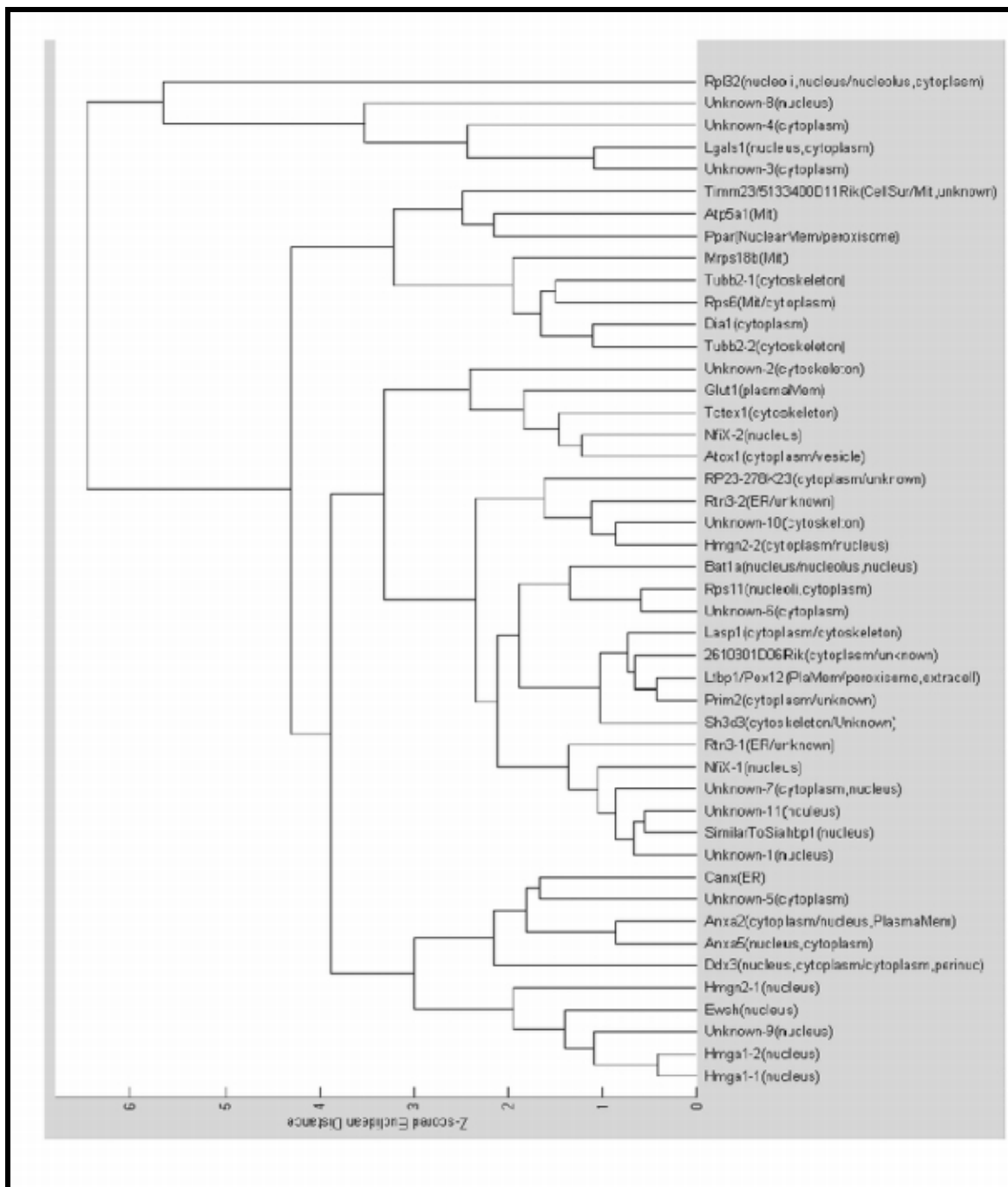


Figure D3: Dendrogram showing location patterns for 46 different proteins as arranged along the X axis by their Z-scored Euclidean Distance. Z-scores appear along the vertical axis. These distance values are determined by subtraction of mean and division by the standard deviation.

SLIF automatically retrieves the matching files, parses the image into meaningful subparts known as “panels”, identifies panels containing a fluorescent microscope image, and ranks these panels by the degree to which they match a specific query pattern. The system uses a neural network trained by using known patterns from tagged antibodies to reference proteins in images of a particular well researched cell line [MKHJC04].

D.5) OUTPUT

SQL queries utilizing Java Server Pages have been tested for the following searches: (a) particular protein name in a figure or panel, (b) specific subcellular pattern, (c) particular spatial resolution in pixel size for an individual image. The SLIF service interface is shown above in figure D2. Subcellular location trees can be generated from this type of information. Patterns vary slightly based on which cellular feature set (sample feature set described above) is used. The dendrogram in figure D3 from the research of the Murphy group appears in [CVWJM03].

E) FUTURE ISSUES

E.1) EXPANDING SLIF

Additional capabilities which are under research include providing summary reports with confidence intervals related to each query retrieval, and combining SLIF query output with other information about the proteins retrieved from annotated databases.

E.2) NEED FOR FIGURE STANDARDS

An important issue identified in this research is the lack of conventions or standards in placement and format of labels on figures for journal articles posed some difficulties in query processing which could be resolved easily by supplementing XML structures for research articles so that their content may be parsed more consistently. Examples of such standards include coordinates of each panel as pixel numbers, inclusion of a URL for an uncompressed figure, number of microns per pixel for each panel, scoping markup to match captions to panels, and database links for annotation of individual proteins.

E.3) DISTRIBUTED BIO-MOLECULAR IMAGE DATABASE

While SLIF provides an interface to collect and process images from web based journal databases, and the Murphy group have researched problems and solutions to improve image retrieval and image queries, it is restricted to fluorescent microscope images. The need is apparent for databases specifically designed to store diverse types of image data and facilitate image queries and searches. This should include time-lapsed images to show how protein activity and distribution responds to changes due to stress, pharmaceutical drugs, environmental toxins, or aging. Unfortunately, these types of data are contained within images of the figures in journal articles which, as has been previously covered this chapter, have query processing limitations. Singh et al have noted “there is currently no home for this vast amount of data, and no method readily available to discover knowledge in such a database were it available” [SMM04]. Their research toward a fully searchable distributed biomolecular image database is in its

early stages. They have defined and classified the types of queries which would be supported in the database as a set of four classes:

- 1) Metadata queries
- 2) Spatial queries
- 3) Semantic queries
- 4) Spatio-temporal queries

Feature extraction for the image database is also discussed by Singh et al [SMM04]. Of particular note is the role of a well defined distance metric to allow similarity comparisons. Equally essential is an appropriate feature set. Texture feature analysis can be adapted from aerial image processing to molecular images as both characterize region properties. At low-level resolution similarity retrieval queries can be answered. More highly structured patterns can be analyzed through statistical shape features, which would enable queries to detect specific proteins. Hierarchical techniques are needed for these high dimension datasets. Locating a specific protein may require a feature vector with hundreds of dimensions posing challenges for data clustering requiring further research both in data indexing and query retrieval.

F) SUMMARY

Challenges lie ahead for this emerging area in the field of bioinformatics. Firstly, detecting spatial and temporal patterns requires complex advanced database techniques, and the ability to handle high-dimension data indexing for the queries. The images must be analyzed and their descriptive features defined to a specified standard to allow for meaningful metadata extraction. Secondly, image databases are extremely

demanding of memory and storage resources. For example, a single image from a cell can be 4 MB, a three dimensional profile composed of 50 z-axis slices through the cell requires 200 MB, a time series of the cell to record dynamic response could be 10 GB, a single experiment contains many cells under several condition sets resulting in 100s GB. The review of location proteomics research has included some proposed solutions, but more investigation is needed.

REFERENCES

- [B00] Baxevanis, A. D. (2000). The molecular biology database collection: an online compilation of relevant database resources. Nucleic Acids Research, 28(1), 1-7.
- [BADGHH05] Ball, C. A., Ihab A. B., Awad, J. D., Gollub, J., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Matese, J. C., Nitzberg, M., Wymore, F., Zachariah Z. K., Brown, P. O., and Sherlock, G (2005). The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Research 33 (Database issue), D580-D582.
- [BBBBCF03] Baldock RA, Bard JB, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, Sharpe J, Ross A, Stevenson P, Venkataraman S, Waterhouse A, Yang Y, and D.R. Davidson. (2003): EMAP and EMAGE: a framework for understanding spatially organized data. Neuroinformatics 1(4) 309-25.
- [BBVTLE03] William E. Bunney, M.D., Blynn G. Bunney, Marquis P. Vawter, Hiroaki Tomita, , Jun Li, , Simon J. Evans, , Prabhakara V. Choudary, Richard M. Myers, Edward G. Jones, M.D., Stanley J. Watson, M.D., and Huda Akil (2003). Microarray Technology: A Review of New Strategies to Discover Candidate Vulnerability Genes in Psychiatric Disorders. American Journal of Psychiatry, 160, 657-666.
- [BBYWHBD02] Albert Burger, Richard A. Baldock, Yiya Yang, Andrew Waterhouse, Derek Houghton, Nick Burton, Duncan Davidson (2002). The Edinburgh Mouse Atlas and Gene-Expression Database: A Spatio-Temporal Database for Biological Research. Proceedings of SSDBM 2002, 239.
- [BDBS04] Eric H Baehrecke, Niem Dang, Ketan Babaria, and Ben Shneiderman (2004). Visualization and analysis of microarray and gene ontology data with treemaps. BMC Bioinformatics, 5, 84-96.
- [BEF03] Jules J Berman, Mary E Edgerton and Bruce A Friedman (2003). The tissue microarray data exchange specification: A community-based, open source tool for sharing tissue microarray data. BMC Medical Informatics and Decision Making, 3, 5-14.
- [BHQSS01] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, et al. (2001). Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. Nature Genetics, 29, 365–371.

[BPSMSV03] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Philippe Rocca-Serra and Susanna-Assunta Sansone (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Research 31(1), 68-71.

[BSTWNL05] Tanya Barrett, Tugba O. Suzek, Dennis B. Troup, Stephen E. Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E. Lash, Wataru Fujibuchi and Ron Edgar (2005). NCBI GEO: mining millions of expression profiles—database and tools Nucleic Acids Research 33 (Database issue), D562-D566.

[CEMKZ05] Scott L Carter, Aron C Eklund, Brigham H Mecham, Isaac S Kohane and Zoltan Szallasi (2005). Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. Bioinformatics 6, 107-121.

[CVWJM03] Xiang Chen, Meel Velliste, Shmuel Weinstein, Jonathan W. Jarvik, and Robert F. Murphy (2003). Location proteomics: building subcellular location trees from high-resolution 3D fluorescence microscope images of randomly tagged proteins. Proceedings of SPIE 2003, 298-307.

[CK03] Craig, P.; Kennedy, J. (2003). Coordinated graph and scatter-plot views for the visual exploration of microarray time-series data. Proceedings of INFOVIS 2003, 173-180.

[CMF03] Coyle J.F.; Mori A.R.; Huff S.M. (2003). Standards for detailed clinical models as the basis for medical data exchange and decision support. International Journal of Medical Informatics 69(18), 157-174.

[CT05] Vivian S. W Chan and Mette Due Theilade (2005). The Use of Toxicogenomic Data in Risk Assessment: A Regulatory Perspective. Clinical Toxicology 43(2), 121-126.

[CVWJM03] Xiang Chen, Meel Velliste, Shmuel Weinstein, Jonathan W. Jarvik, and Robert F. Murphy (2003). Location proteomics: building subcellular location trees from high-resolution 3D fluorescence microscope images of randomly tagged proteins. Proceedings of SPIE 2003, 298-307.

[EC99] R. Ekins and F.W. Chu (1999). Microarrays: their origins and applications Trends in Biotechnology 17, 217-218.

[EDL02] Edgar, R., Domrachev, M. and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 30, 207–210.

[G05] Michael Y. Galperin (2005). The Molecular Biology Database Collection:2005 update. Nucleic Acids Research 33 (Database issue), D5–D24.

[GBBDFH03] Jeremy Gollub, Catherine A. Ball, Gail Binkley, Janos Demeter, David B. Finkelstein, Joan M. Hebert, Tina Hernandez-Boussard, Heng Jin, Miroslava Kaloper, John C. Matese, Mark Schroeder, Patrick O. Brown, David Botstein and Gavin Sherlock (2003). The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Research, 31(1), 94-96.

[GW02] Goncalves, J., Marks, W.L. Roles and requirements for a research microarray database (2002). Engineering in Medicine and Biology Magazine, IEEE 21(6), 154-157.

[HCILAF04] Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. et al. (2004). The Gene Ontology(GO) database and informatics resource Nucleic Acids Research, 32, D258–D261.

[HGBDMS03] Hernandez-Boussard T, Gollub J, Ball CA, Demeter J, Matese JC, Sherlock G. (2003; unpublished). Mapping the MAGE-OM to data within the Stanford Microarray Database.
http://www.mged.org/Workgroups/SMD_to_MAGE_Mapping.doc
http://www.mged.org/Workgroups/MAGE-ML_to_SMD.xls

[HM04] Kai Huang, Robert F. Murphy (2004). Automated Classification of Subcellular Patterns In Multicell Images Without Segmentation Into Single Cells. Proceedings of ISBI 2004, 1139-1142

[HWBAHJG02] Peter M. Haverty, Zhiping Weng, Nathan L. Best, Kenneth R. Auerbach, Li-Li Hsiao, Roderick V. Jensen, Steven R. Gullans (2002). HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. Nucleic Acids Research 30(1), 214-217.

[IITGT03] Ikeo K, Ishii J, Tamura T, Gojobori T, Tateno Y. (2003). CIBEX: center for information biology gene expression database Comptes Rendus Biologies, 326(10-11), 1079-82.

[JPZ03] Daxin Jiang Jian Pei Aidong Zhang (2003). Towards Interactive Exploration of Gene Expression Patterns SIGKDD Explorations, 5(2), 79-90.

[KKCDIK04] Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Korner,C., Kull,M., Torrente,A., Sarkans,U., Vilo, J. et al. (2004). Expression Profiler: next generation—an online platform for analysis of microarray data. Nucleic Acids Research, 32, W465–W470.

[LGTC04] Using the Agilent Microarray Scanner Leiden Genome Technology Center, Leiden University Medical Center (2004)
http://www.lgtc.nl/home/protocols/micro-arrays/equipment_scanner_agilent_use.pdf.

[LHSQP04] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin and Paul Pavlidis (2004). Coexpression Analysis of Human Genes Across Many Microarray Data Sets Genome Research, 14, 1085-1094.

[LY04] Hautaniemi, S.; Lehmussola, A.; Yli-Harja, O. (2004) DNA microarray data preprocessing. Proceedings of ISCCSP 2004, 751–754.

[M01] D. E. Moody (2001). Genomics techniques: An overview of methods for the study of gene expression. Journal of Animal Science, 79(E. Suppl.), E128–E135.

[M04] Robert F. Murphy (2004). Automated Interpretation of Subcellular Location Patterns. Proceedings of ISBI 2004, 53-56.

[MAMSD03] Yves Moreau, Stein Aerts, Bart De Moor, Bart De Strooper and Michal Dabrowski (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. TRENDS in Genetics, 19(10), 570-577.

[MCGMFHSGCSC04] Anirban Maitra, Yoram Cohen, Susannah E.D. Gillespie, Elizabeth Mambo, Noriyoshi Fukushima, Mohammad O. Hoque, Nila Shah4, Michael Goggins, Joseph Califano, David Sidransky, and Aravinda Chakravarti. (2004). The Human MitoChip: A High-Throughput Sequencing Microarray for Mitochondrial Mutation Detection. Genome Research,14, 812-819.

[MF05] Robert F. Murphy, Christos Faloutsos (2005). Research issues in protein location image databases. Proceedings of SIGMOD 2005, 966-967.

[MKHJC04]. Robert F. Murphy, Zhenzhen Kou, Juchang Hua, Matthew Joffe, William W. Cohen (2004). Extracting and Structuring Subcellular Location Information from On-line Journal Articles: The Subcellular Location Image Finder Proceedings of IASTED: 109-114.

[MMS01] S. C. Maheshwari, Nirmala Maheshwari and S. K. Sopory (2001). Genomics, DNA chips and a revolution in plant biology. CURRENT SCIENCE, 80(2), 252-261.

[O03] Michael O'Connell (2003). Differential expression, class discovery and class prediction using S-PLUS and S+ArrayAnalyzer. SIGKDD Explorations, 5(2), 38-47.

[P04] Andrey Ptitsyn (2004). Class Discovery Analysis of the Lung Cancer Gene Expression Data. DNA and Cell Biology, 32(10), 715-721.

[PB04] Christopher J. Penkett and Jürg Bähler (2004). Navigating public microarray databases. Computational Functional Genomics, 5, 471-479.

[PMH02] Willmar D. Patino, Omar Y. Mian, Paul M. Hwang (2002). Serial Analysis of Gene Expression Technical Considerations and Applications to Cardiovascular Biology. Circulation Research, 91, 565-569.

[PSSACC05] H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. Garcia Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone and A. (2005). BraZma ArrayExpress—a public repository for microarray gene expression data at the EBI Nucleic Acids Research, 33, Database issue D553-D555.

[PT04] Peterson, D.A.; Thaut, M.H. (2004). Model and feature selection in microarray classification. Proceedings of CIBCB 2004, 56-60.

[SMM04] A. K. Singh, B. S. Manjunath, Robert F. Murphy (2004). A Distributed Database for BioMolecular Images. SIGMOD Record, 33(2), 65-71.

[SHTKLD98] Mark Schena, Renu A. Heller, Thomas P. Theriault, Ken Konrad, Eric Lachenmeier and Ronald W. Davis. (1998). Microarrays: biotechnology's discovery platform for functional genomics. Trends in Biotechnology, 16(7), 301-306.

[SMSTSC02] Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M. et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML) Genome Biology, 3, 0046.1-0046.9.

[SPLOS04] Ugis Sarkans, Helen Parkinson, Gonzalo Garcia Lara, Ahmet Oezcimen, Anjan Sharma, Niran Abeygunawardena, Sergio Contrino, Ele Holloway, Philippe Rocca-Serra, Gaurab Mukherjee, Mohammadreza Shojatalab, Misha Kapushesky, Susanna Sansone, Anna Farne, Tim Rayner, and Alvis Brazma The ArrayExpress Gene Expression Database: a Software Engineering and Implementation Perspective Bioinformatics © Oxford University Press 2004; Bioinformatics Advance Access.

[TAHSOM04] Nathan D. Trinklein, Shelley Force Aldred, Sara J. Hartman, Diane I. Schroeder, Robert P. O'tillar and Richard M. Myers (2004). An Abundance of Bidirectional Promoters in the Human Genome Genome Research, 14, 62-66.

[TLFL04] Ying Tao, Yang Liu, Carol Friedman, and Yves A. Lussier (2004). Information visualization techniques in bioinformatics during the postgenomic era. DDT: BIOSILICO, 2(6), 237-245.

[WBBCIM03] Michael Waters, Gary Boorman, Pierre Bushel, Michael Cunningham, Rick Irwin, Alex Merrick, Kenneth Olden, Richard Paules, James Selkirk, Stanley Stasiewicz, Brenda Weis, Ben Van Houten, Nigel Walker, and Raymond Tennant (2003). Systems Toxicology and the Chemical Effects in Biological Systems (CEBS) Knowledge Base. Environmental Health Perspectives, 111(6), 811-824.

BIOGRAPHICAL INFORMATION

The author Zoe Alexandra Raja was born in Vancouver, Canada and lived in Britain before moving to the United States. She received a Bachelor's Degree from the University of Wisconsin – Milwaukee in the area of Cellular and Molecular Biology with Certification in Biotechnology and two years of undergraduate research projects. She worked for two years in the University of Wisconsin System as a research assistant while taking additional seminars and courses in Biochemistry and advanced Genetics. The research frequently involved use of Biological Databases and other software tools. It was a natural step to pursue a Master's Degree in the field of Computer Science at the University of Texas at Arlington, completing graduation in 2005. Zoe Raja began working full time for Siemens Corporate Research in the department of Software Engineering as a Requirements Engineer in 2004. Her areas of interest for the future include Database Management, Software Architecture, Systems Integration, and Systems Engineering.