RI 02016 22 January 2002 Computer Science

IBM Research Report

Bioinformatics for Microarrays

Sudeshna Adak, Vishal S Batra, Deo N Bhardwaj,
P V Kamesam, Pankaj Kankar, Manish P Kurhekar,
Biplav Srivastava
IBM Research Division
IBM India Research Lab
Block I, I.I.T. Campus, Hauz Khas
New Delhi - 110016. India.

Bioinformatics for Microarrays

Sudeshna Adak, Vishal S Batra, Deo N Bhardwaj,
P V Kamesam, Pankaj Kankar, Manish P Kurhekar,
Biplav Srivastava*
IBM India Research Laboratory
Block 1, IIT Campus, Hauz Khas
New Delhi 110016, India
Email: {asudeshn, bvishal, dbhardwa, pkamesam, kpankaj,
kmanish, sbiplav}@in.ibm.com

Abstract

Microarrays (or biochips) is perhaps one of the most exciting developments in bioinformatics research. The emerging biochip technology has made it possible to simultaneously study expression (activity level) of thousands of genes or proteins in a single experiment in the laboratory. However, in order to extract relevant biological knowledge from the biochip experimental data, it is critical not only to analyze the experimental data, but also to cross-reference and correlate these large volumes of data with information available in external biological databases accessible online.

We describe a comprehensive system for knowledge management in bioinformatics called *e2e* in which data generated by the biochip experiments can be analyzed for emerging patterns among groups of genes with additional insights from related analyses like pathway scores, sequence similarity, literature text summarization, etc. To the biologist or biological applications, *e2e* exposes a common semantic view of inter-relationship among biological concepts in the form of an XML representation called eXpressML. Internally, *e2e* can use any data integration solution (like DiscoveryLink, Kleisli or natively XML-based) to retrieve data and return results corresponding to the semantic view. We have implemented an e2e prototype that demonstrates our framework by allowing a biologist to analyze her gene expression data in GEML or from a public site like Stanford, and discover knowledge through operations like querying on relevant annotated data represented in eXpressML using pathways data from KEGG, publication data from Medline and protein data from SWISS-PROT.

[*] Contact person. Author names appear in alphabetical order.

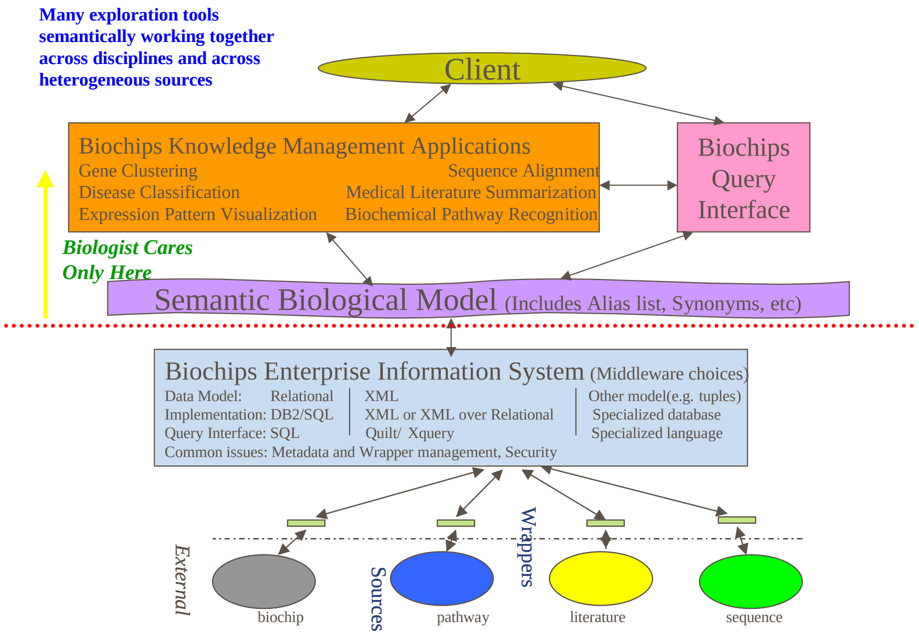

Many exploration tools semantically working together across disciplines and across heterogeneous sources
Client
Biochips Knowledge Management Applications
Gene Clustering
Sequence Alignment
Disease Classification
Medical Literature Summarization
Expression Pattern Visualization
Biochemical Pathway Recognition
Biochips Query Interface
Biologist Cares Only Here
Semantic Biological Model (Includes Alias list, Synonyms, etc)
Biochips Enterprise Information System (Middleware choices)
Data Model: Relational
Implementation: DB2/SQL
Query Interface: SQL
XML
XML or XML over Relational
Quilt/ Xquery
Other model(e.g. tuples)
Specialized database
Specialized language
Common issues: Metadata and Wrapper management, Security
Wrappers
External
Sources
biochip
pathway
literature
sequence

Figure 1: *e2e* Biochips Information System Framework Architecture.

1 Introduction

The term *biochip* is often used to refer to the variety of microarrays and miniaturized, microfluidic systems that perform massively parallel biochemical assays measuring the expression levels of thousands of genes or proteins simultaneously. The advent of high-density microarrays, introduced by [25] made it possible for the first time to measure the expression levels of thousands of genes simultaneously. Since then, DNA based microarray technologies [26, 4, 8, 19] have been used extensively in generating the expression levels of all or most of the genes of several organisms under a variety of experimental conditions. Specialized repositories and data warehousing projects are being built (NCBI's Gene Expression Omnibus[1](GEO), ArrayDB[2], ArrayExpress[3], Stanford Microarray Database[4](SMD), ChipDB[5], GeneX[21]) to store the vast quantities of data that are being generated by the biochips.

A biologist starts with analysis of the gene expression data for insightful patterns among some clusters of genes. Once a gene cluster is obtained, the main interest of a biologist lies in finding out the underlying biological mecha-

[1] http://www.ncbi.nlm.nih.gov/geo/
[2] http://genome.nhgri.nih.gov/arraydb/
[3] http://www.ebi.ac.uk/arrayexpress/
[4] http://genome-www4.Stanford.EDU/MicroArray/SMD/
[5] http://young39.wi.mit.edu/chipdb_public/front-page.html

nisms and functions causing these genes to be co-expressed and assign biological significance to this cluster. The biological relations among these genes span multidisciplinary islands of biology. Downstream annotation involves combining expression data with other sources of information to improve the range and quality of conclusions that can be drawn. Hence, high-end data management systems are required to handle the explosion in biochips data where beyond core data services, their essential capabilities to enable effective biochip usage by the life sciences industry include:

- Biochip data validation
- Comparative analytics
- Improved interpretation of these large data sets through integration of other biomedical data sources
- Scalability

However, related biomedical data[2] are numerous and hence, it will usually be available with distributed external organizations who store them in many different ways and continuously update them. The only common link is that these related data sources and annotation tools are available online. Our focus is to develop an infrastructural framework for building knowledge discovery tools for microarrays that can leverage related but continuously updated diverse online data.

To the stated end, we describe a comprehensive system for knowledge management in bioinformatics called *e2e* (see Figure 1) in which data generated by the biochip experiments can be analyzed with additional insights from analyses like pathway, sequence similarity, literature text summarization, etc. To the biologist or biological applications, *e2e* exposes a common semantic view of inter-relationship among biological concepts in the form of an XML representation called eXpressML. Internally, *e2e* can use any data integration solution (like DiscoveryLink, Kleisli or natively XML-based) to retrieve data and return results corresponding to the semantic view. We have implemented an *e2e* prototype that demonstrates our framework by allowing a biologist to analyze her gene expression data in GEML or from a public site like Stanford, and discover knowledge through operations like querying on annotated data in eXpressML, pathway scoring, text summarization, etc using pathways data from KEGG, publication data from Medline and protein data from SWISS-PROT, accordingly.

Here is a layout of the rest of the paper: we start with a background of gene expression data and microarrays/biochips, and discuss current approaches for data integration for bioinformatics. However, a biologist's focus is not just data but the semantic relationship among retrieved data. To address this, we present a desiderata for semantic integration and introduce the *e2e* framework to serve as such an infrastructure. We discuss the different components of *e2e* - the data integration middleware, eXpressML, a unified representation in XML for the complete ”annotation data” necessary to gain insight into gene expression patterns and the knowledge management (KM) applications. Next, *we give a tour of our e2e prototype that demonstrates the promise of an infrastructure capable of going beyond analysis of microarray data to intelligently cross-reference and correlate with information from external biological databases accessible online.* We conclude the discussion with our contributions and future work.

2 Background

2.1 Gene Expression and Biochips

Genes are fragments of DNA containing the fundamental code of life. The Central Dogma of biology asserts that this genetic information moves from ”DNA to RNA to protein” and this process is called gene expression. When a gene is expressed, the coded information contained in its DNA is first transcribed into messenger-RNA and then translated into the proteins present and operating in the cell. Changes in gene expression are associated with almost all biological phenomena, including aging, onset and progression of diseases, adaptive responses to the environment, and biochemical effects of drugs.

As the Human Genome Project hurtles towards completion and hundreds of novel genes are being identified in human and other organisms, DNA microarrays are helping to accelerate our understanding of the functions of these genes. The availability of this biochip data promises to have a profound impact on the understanding of basic cellular processes, the efficacy of diagnosis and treatment of disease, and improvements in our environment. Biochips, with their numerous options from DNA to protein microarrays on a wide variety of surfaces and different modes of quantification, are still in their infancy. In order to realize the full potential of biochips, the main challenges faced by the life sciences industry today are:

- Improvements in the core microarray technology to improve the accuracy of gene expression measurement
- Development of the full spectrum of specialized analytics and (bio)informatics tools required for making (biological) knowledge discoveries from biochip data.

Our focus is to develop an infrastructural framework for building knowledge discovery tools for microarrays that can leverage related but continuously updated diverse online data.

2.2 Integration of Heterogeneous Data

There are several stand-alone analysis tools today (e.g. GeneSight^{TM} from Biodiscovery, biotechnology solutions from Spotfire, etc.) that detect gene expression patterns. However, since new genomic data is continuously produced and made available online, a stand-alone tool, however sophisticated, will fail to provide the scalable, heterogeneous information integration infrastructure that is required for an up to date and comprehensive understanding of the functions of genes and their complex interactions. For the biologist, true insight is possible only by linking detected gene expression patterns with known background information on genes such as its DNA sequence and 3D structure, its role in cellular processes and disease onset/progression/prevention.

A variety of approaches have been developed for integrated access to heterogeneous data sources in genomics. Borrowing the terminology from [7], in the *link-driven federation* approach, the user can switch between sources using system-provided links in a hypermedia environment. Here, a user starts from some point of interest in a data source and then can jump to other related data sources through system created links. The user has to still interact with individual sources; only the interaction is easier through convenient links and not invoking the sources directly. SRS[10], GeneCards[23] and LinkDB[11] are examples of this approach. The link driven approach is very convenient for non-expert users because of the simple point-and-click user interface. It is also possible to perform limited keyword search on the content of a source by specifying regular expressions. The downside of link driven approaches is that it does not scale well and has no across-source capabilities.

Another approach is that of *view integration* in which a virtual global schema is created in a common data model using the descriptions of the individual sources so that the user can declaratively pose queries on the common data model that may span the content of multiple sources. The system seamlessly and automatically figures out how data from the different sources has to be retrieved [18]. A variation of view integration is the *warehousing* approach where instantiation of the global schema is created, i.e., all data of interest in remote sources is locally replicated and maintained for predictable performance. Example of general purpose database middleware for integrating heterogeneous data sources for the Life Sciences domain includes IBM's DiscoveryLink[15], Kleisli[5], and OPM[6] which provide powerful querying capabilities, but fail to provide the in-depth analysis that are provided by the ”point solutions”.

It is important to remember that the goal of a biologist is not just to get any data from different sources. Instead, she wants to access only relevant data that she can easily correlate in the pursuit of understanding the biochip assay. Hence, what is needed is *semantic integration* in which the user sees domain concepts like proteins and pathways while the infrastructural artifacts like source names (SWISS-PROT, KEGG, etc) and attribute fields (protein id, etc) are handled transparently by the user. Our goal in bioinformatics should be to provide a one-stop solution that facilitates knowledge discovery for microarrays by supporting analyses of gene expression data and cross-validation of emerging patterns with annotations of related data and applications available online. We will call such a solution to be a SIM (Sematically Integrated solution for Microarrays) system. It is clear that SIM systems will enable the biotechnology and pharmaceutical industry to realize the full potential of biochips.

3 Desiderata for a SIM system

As important biological data sources are distributed, autonomous, and heterogeneous, a biologist needs a unified view of heterogeneous data and applications that is irredundant, consistent, and semantically organized for maximum usability. The main features of a SIM system include:

1. At the core, data storage and data management of massive volumes of biochip experimental data.
2. Statistical analysis and visualization toolboxes for detecting gene expression patterns from biochip experimental data.
3. Downstream annotation/association of detected gene expression patterns with relevant biological information from heterogeneous data sources.
4. Knowledge discovery through querying, analyzing, data mining and visualization of the experimental and the downstream annotation information.

After a solution can store the large gene expression data from experiments, the data is filtered for gene expression patterns through a wide class of visualization and analysis algorithms. Next, a SIM system needs the ability to access and retrieve remote online sources so that a query and browsing interface can be built that allows the biologist to query both the biochip experimental data and the analytical results, and the annotations on related biomedical data from remote sources. Related data can be heterogeneous (e.g. sequence, pathways, literature, etc) and the user may issue queries that correlate annotations of different sources. Finally, specialized bioinformatics tools are essential to gain insight into the different functions of genes, their complex interactions and roles in disease onset/progression/ prevention.

A system related to our definition of SIM is TAMBIS[12] where a common ontology of about 1900 terms is constructed to describe the concepts and relationships in molecular biology. Users interact with TAMBIS in the ontological realm while the system internally maps them to source schemas using Kleisli[5] as its data integration middleware. However, TAMBIS is not targeted towards microarrays and does not provide the full spectrum of query/analytical capabilities (breadth) that is needed in making (biological) knowledge discoveries from biochip data.

3.1 *e2e* - An end-to-end SIM Framework

We now discuss *e2e* as a SIM framework in which semantic relationship among biological concepts is represented in eXpressML and analytical KM tools can work from this abstraction. As seen in Figure 1, *e2e* envisages a two stage approach.

The underlying infrastructure for *e2e* is a view integration middleware (called Enterprise Information System to emphasize the fact that it should be able to

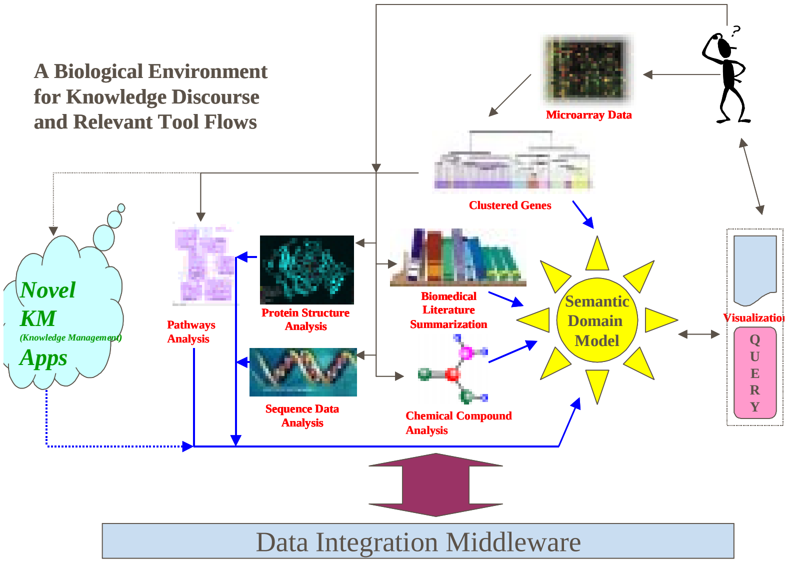


Figure 2: A Schemata of Flow between KM Analyses in *e2e*.

handle large data sizes) which can retrieve either microarray experimental data or external information from publicly available biological data sources. Recall from the background section that in a view integration system, a virtual global schema is created in a common data model using the descriptions of the individual sources and any query posed on the common data model is answered by seamlessly and automatically retrieving data from different sources. In choosing a middleware, one has to consider the issues of uniform data model, the query language to support and availability of source wrappers. For example, with a relational data model and SQL query language, DiscoveryLink[15] is a middleware solution on commercial DB2^{TM} database while Kleisli[5] uses a complex value model of data and Collection Programming Language (CPL), but either could be used within *e2e*. In the present prototype, the data management is in XML and in-memory, but will be migrated to a relational database in the next version.

To the biologist or biological applications, *e2e* exposes a common semantic view of the inter-relationship among biological concepts in the form of the XML representation of eXpressML[1]. This semantic biological model provides the user with a common biological context to view and manipulate related data and issue XML queries in Quilt[24] through a query interface.

Finally, *e2e* envisages an application layer where knowledge management tools are available for detecting gene expression patterns and downstream annotation of these patterns with heterogeneous information provided by the middleware. For example, some tools that can be used are: pathway visualization tools[17] for annotating gene clusters with pathway information, text summarization tools[27, 16] for annotating gene clusters with biological function, and sequence alignment tools[3] for annotating gene clusters with motifs/domains. Figure 2 shows a schematic flow between KM analyses tools that a biologist may take in pursuit of discovery. *Note that the input for any KM tool is a group of genes and (optionally) eXpressML while the output is some insight about the group.* By applying diverse tools, a biologist can verify her insights with analyses spanning multidisciplinary islands of biology.

e2e is a SIM system offering semantic integration of the diverse data sources to the user. Specifically, a user only needs to know about the biological domain while the system will hide the peculiarities of the sources involved to answer a domain query.

3.2 Integration in *e2e*

e2e works on two types of data - the gene expression data from microarray experiments and *annotations* of gene expression as well as relevant distributed data necessary to gain insight into gene expression patterns. The annotations are semantically arranged in the XML representation of eXpressML[1].

For gene expression, we adopted Rosetta Inpharmatics' Gene Expression Markup Language, GEML^{TM}[6], which has been accepted relatively widely by

[6] http://www.geml.org

the industry as a uniform syntax for storing and exchanging gene expression data from multiple biochip experiments. For annotations, we developed the eXpressML representation keeping following into consideration:

- The semi-structured nature of XML makes it the appropriate language for unified view of annotations as it guarantees flexibility and scalability in the data model for future extensions.
- The common view should allow querying, modeling, and browsing of complex annotations.
- The unified model should arrange the annotation information in a compact hierarchy but reflect the relationship among the biological data items and facilitate complex queries.

Though details of eXpressML have been given elsewhere[1], we indicate its organization and give a snapshot of its DTD in Figure 3 (also see Figure 7 of *e2e* tour). The clustering groups from gene expression analyses are represented under expression data analysis while annotation on related data is collated under reporter. Such data includes DNA, protein, keywords, disease, pathways, enzyme and citation information and is obtained either directly or by running KM tools on data from heterogeneous data sources relevant to the genes/proteins in the biochip experiment. Note that gene expression data itself is not part of eXpressML. A related effort is MAGE-ML[7] which represents useful annotations that describe the experimental conditions and environments (array type, number of spots, sample source, etc). However, MAGE-ML does not support annotation derived from heterogeneous external sources while eXpressML extends to this as well.

Now both GEML and eXpressML are available from *e2e* and can be queried with an expressive XML query language. The Biochips Query Interface (refer to Figure 1) select supports queries in Quilt XML query language [24] (specifically, Kweelt[8] implementation of Quilt). Quilt is the precursor to XQuery[9], which is being formalized by W3C (see a survey of XML query languages at XML Cover pages[10]). Quilt allows querying on the content and structure of XML documents - it is the latter capability that makes it more powerful in expressiveness to SQL (for relational data) or XML query languages like Xpath or XSLT. The query interface has templates for a number of pre-canned queries and the user can also pose any Quilt query which is valid (as shown later in Figure 8 of *e2e* tour).

3.3 KM Layer

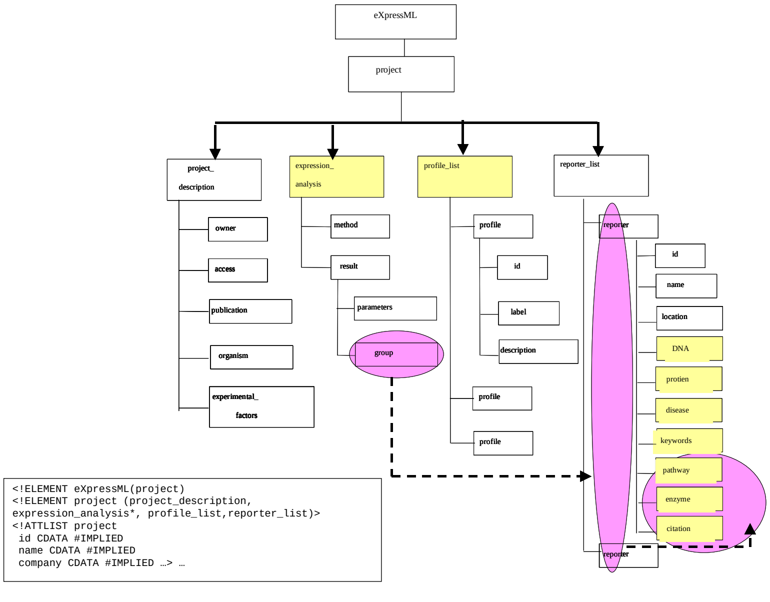
The KM layer consists of two types of applications:

[7] http://www.mged.org/Workgroups/MAGE/mage.html

[8] http://db.cis.upenn.edu/Kweelt/

[9] http://www.w3.org/TR/xquery/

[10] http://xml.coverpages.org/xmlQuery.html



```
<!ELEMENT eXpressML(project)
<!ELEMENT project (project_description,
expression_analysis*, profile_list,reporter_list)>
<!ATTLIST project
 id CDATA #IMPLIED
 name CDATA #IMPLIED
 company CDATA #IMPLIED …> …
```

Figure 3: Outline of eXpressML. A query like *'list all regulatory pathways and enzymes associated with genes that are similar in expression to gene HXK1 (Hexokinase-1)'* can be formulated against the eXpressML but it not possible with existing representations like GEML or MAGE-ML. This is because the query involves determining the genes in the same cluster (group) as gene HXK1 and finding the pathways and enzymes associated with the resulting gene list (marked by circles).

- Tools for detecting gene expression patterns by supporting clustering, classification, and visualization of biochip experimental data.

- Downstream annotation tools combining expression data with other sources of information to improve the range and quality of conclusions that can be drawn. Related data can be sequence, 3D structure, biochemical pathways, medical literature, genetic maps of diseases, etc.

Below, we describe some of the implemented front-end tools in detail but note that new tools can be built that have as input a group of genes and optionally, subset of data represented in eXpressML.

4 Knowledge Management Applications

We discuss the range of front-end tools and techniques that *e2e* incorporates to create integrated and systematic models of biological systems.

4.1 Microarray Analysis

The first tier of Microarray data analysis typically involves clustering or classification of the microarray data. In clustering (or cluster analysis), genes with similar expression patterns are grouped together. Then, it is the gene cluster rather than the individual genes that get associated with biological functions (e.g. DNA repair, galactose metabolism). For example, hierarchical clustering[9] has been used to determine the functions of gene clusters in regulating cell-cycle in yeast.

e2e provides a platform for integrating algorithms made available through third-party vendors or academic researchers seamlessly as long as they provide following basic information:

- Any initialization parameters and the format of input gene expression data (tabular or XML).

- The format of output result.

- If the algorithm supports visualization, a *handle* of the input and output panels.

Given this, the *e2e* microarray platform generates the necessary adapters for mapping GEML/eXpressML data into the inputs/outputs of the external tool. We have implemented hierarchical clustering (shown in Figure 6 of *e2e* tour) and K-means clustering in the *e2e* prototype.

4.2 Text Summarization

The biomedical literature databases are rich source of information from various disciplines of biomedical sciences. Text mining of these databases can be used to augment, confirm, or discover biologically significant information for gene clusters spanning different biological domains. The main challenges in handling biomedical citations are:

1. Querying on even a small cluster of genes retrieves tens of thousands of documents.
2. Use of multiple names and conventions in referring to genes makes it difficult to cross-reference documents with gene names.
3. Non-uniform nomenclature and language usage for same biological concepts make it difficult for text mining of the citations retrieved.
4. Highly complex and parallel interrelations among biological processes across multiple biological domains.

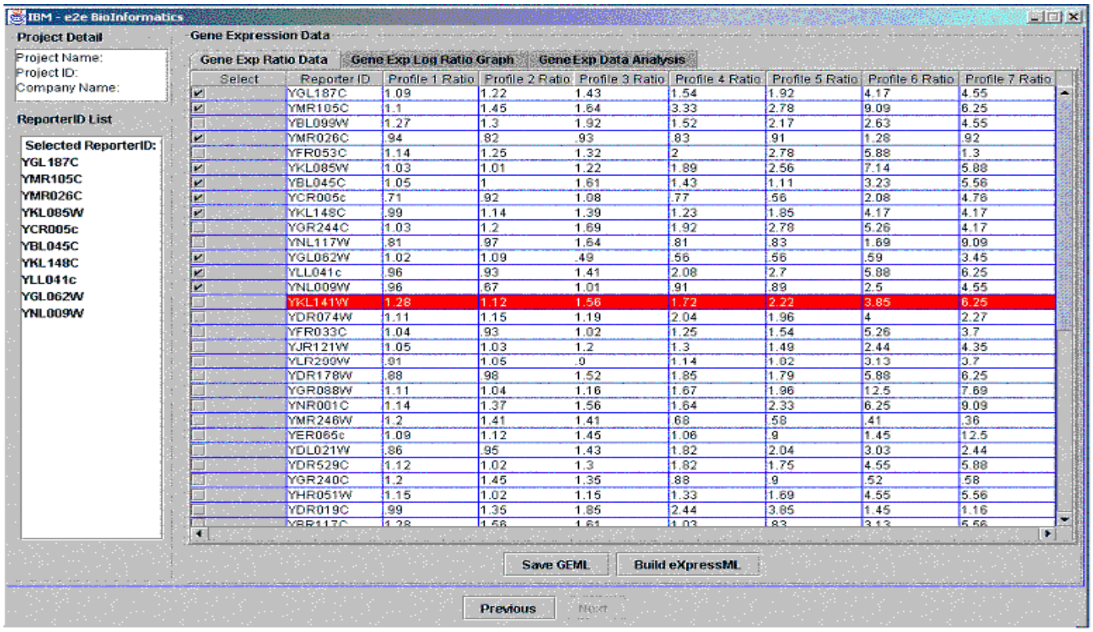
We have developed a specialized text-mining system called MedMeSH summarizer [16] that provides a summary of the citations pertaining to a group of genes in a given cluster. The MedMeSH summarizer system uses PubMed as the literature database and provides an automated document extraction and summarization solution (an output is shown in Figure 10 of *e2e* tour). PubMed, the most widely used biomedical literature database has more than 11 million citations (since 1960) and about 30,000 new citations are added each month. Key features of MedMeSH Summarizer are:

- The user is required to provide only a list of genes (gene cluster) as input.
- The output is a summary of the documents, which shows
  - The most important MeSH terms which describe the whole cluster (can be viewed as an overall list, a tree, or partitioned based on cluster-relevance).
  - Produces summaries across all biological domains, which are relevant to the cluster.

4.3 Pathways Scoring

Living organisms behave as complex systems that are flexible and adaptive to their surroundings. At the cellular level, organisms function through intricate networks of chemical reactions (metabolic pathways) and interacting molecules (regulatory pathways). These networks or biochemical pathways may be considered as the wiring diagrams for the complete biological system of an organism.

The information harnessed from microarray data can show the pathway dynamics. Genes in any organism act in concert with other genes in a pathway, and the biological functions of a gene depends on these other genes. Annotation of microarray data with pathway information can help in understanding the functions and roles of the proteins involved in various cellular processes. The pathway scoring system serves as an important tool for interpreting the



Project Detail

Project Name:
Project ID:
Company Name:

ReporterID List

Selected ReporterID:
YGL187C
YMR105C
YMR026C
YKL085W
YCR005c
YBL045C
YKL148C
YLL041c
YGL062W
YNL009W

Gene Expression Data

Select	Reporter ID	Profile 1 Ratio	Profile 2 Ratio	Profile 3 Ratio	Profile 4 Ratio	Profile 5 Ratio	Profile 6 Ratio	Profile 7 Ratio
☑	YGL187C	1.09	1.22	1.43	1.54	1.92	4.17	4.55
☑	YMR105C	1.1	1.45	1.64	3.33	2.78	9.09	6.25
□	YBL099W	1.27	1.3	1.92	1.52	2.17	2.63	4.55
☑	YMR026C	.94	.82	.93	.83	.91	1.28	.92
□	YFR053C	1.14	1.25	1.32	2	2.78	5.88	1.3
☑	YKL085W	1.03	1.01	1.22	1.89	2.56	7.14	5.88
☑	YBL045C	1.05	1	1.61	1.43	1.11	3.23	5.56
☑	YCR005c	.71	.92	1.08	.77	.56	2.08	4.76
☑	YKL148C	.99	1.14	1.39	1.23	1.85	4.17	4.17
□	YGR244C	1.03	1.2	1.69	1.92	2.78	5.26	4.17
□	YNL117W	.81	.97	1.64	.81	.83	1.69	9.09
☑	YGL062W	1.02	1.09	.49	.56	.56	.59	3.45
☑	YLL041c	.96	.93	1.41	2.08	2.7	5.88	6.25
☑	YNL009W	.96	.67	1.01	.91	.89	2.5	4.55
□	YKL141W	1.28	1.12	1.56	1.72	2.22	3.85	6.25
□	YDR074W	1.11	1.15	1.19	2.04	1.96	4	2.27
□	YFR033C	1.04	.93	1.02	1.25	1.54	5.26	3.7
□	YJR121W	1.05	1.03	1.2	1.3	1.49	2.44	4.35
□	YLR299W	.91	1.05	.9	1.14	1.02	3.13	3.7
□	YDR178W	.88	.98	1.52	1.85	1.79	5.88	6.25
□	YGR088W	1.11	1.04	1.16	1.67	1.96	12.5	7.69
□	YNR001C	1.14	1.37	1.56	1.64	2.33	6.25	9.09
□	YMR246W	1.2	1.41	1.41	.68	.58	.41	.36
□	YER065c	1.09	1.12	1.45	1.06	.9	1.45	12.5
□	YDL021W	.86	.95	1.43	1.82	2.04	3.03	2.44
□	YDR529C	1.12	1.02	1.3	1.82	1.75	4.55	5.88
□	YGR240C	1.2	1.45	1.35	.88	.9	.52	.58
□	YHR051W	1.15	1.02	1.15	1.33	1.69	4.55	5.56
□	YDR019C	.99	1.35	1.85	2.44	3.85	1.45	1.16
□	YBR117C	1.28	1.58	1.61	1.03	.83	3.13	5.56

Figure 4: Loading the gene expression data for 50 reporters.

large amount of data from microarrays, in assessing the behavior of pathways at different cell stages or the effect of stimuli on cellular processes.

We have implemented pathway scoring algorithms [17] that can be used to answer queries like:

- Which pathways are most affected during the course of the experiments?
- Which pathways are functioning differently between the two groups of patients?
- What is the nature of the effect? (details such as which genes in a pathway are most affected, are the genes over-expressed or under-expressed, which reactions are disrupted etc.)

It retrieves information using gene expression data and putative metabolic and regulatory pathways database of KEGG. The outputs are (one output is shown in Figure 11 of *e2e* tour): *pathway scores* which quantify "activity", "coregulation", and "cascade" effects in pathways as measured by the gene expression levels from the microarray experimental data, and *pathway animated visuals* which show the effects on individual pathways over the course of a microarray experiment series or between two or more groups.

4.4 Protein Sequence Analysis

An annotated model organism genome can be used as a source of reference for annotating and understanding other genomes. By comparing the complete genome of one organism to another, it is clear that certain genes have been conserved since evolutionary divergence from a common ancestor. Genes can be found in the different organisms, with identical functions and/or protein motifs. The way to do this is by sequence analysis. The sequence analyser has a host of sequence similarity tools including BLAST and FASTA and uses the SWISS-Prot database.

5 A Tour of *e2e*

As mentioned before, we have implemented an *e2e* prototype that demonstrates the promise of an end to end bioinformatics framework for microarrays. We take a tour of the system following a set of actions that a typical biologist may take for analyzing microarray results.

Figure 4 shows when the biologist has loaded the gene expression data for 50 genes (right panel) and selected 10 of them for further analysis. Now, the biologist can view the expression levels of the selected genes graphically as seen in Figure 5. In Figure 6, she has used hierarchical clustering to group the 10 genes based on their expression level.

Now, the biologist can ask the *e2e* tool to use pathways data from KEGG, publication data from Medline and protein data from SWISS-PROT to build eXpressML (which is a semantic model) for the selected genes. It is the task

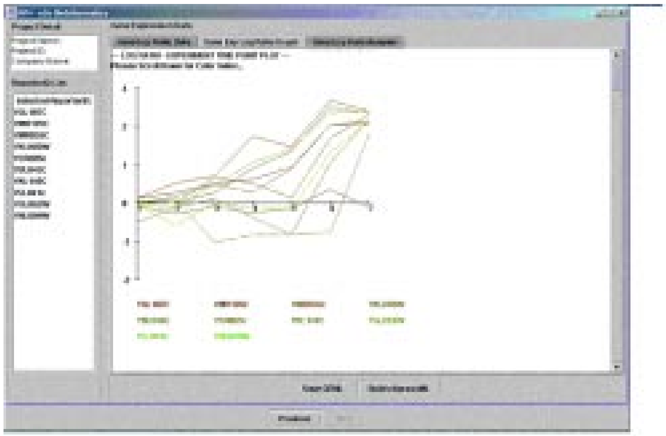


Figure 5: Viewing the expression level in graphical form.

of the tool, in conjunction with corresponding wrappers, to access the sources and *dynamically fetch* the corresponding data, and assemble it as eXpressML. Figure 7 shows the eXpressML generated for the 10 selected genes.

The *e2e* prototype allows the user to view and query on both gene expression data and eXpressML. Figure 8 shows the query interface where there are some pre-specified parameterized queries in english or the user is free to specify their own Quilt query. Queries can be diverse and cover any data represented in eXpressML and gene expression or *genes in the result of any previous query*. The latter is possible because we internally follow the convention of returning a query result with a list of applicable genes. Queries can range from asking information about a set of genes like the list of keywords, the reactions and enzymes, the expression level, or correlating information from diverse sources.

In Figure8, the biologist has posed a query using the genes from the result of a previous query (Q1), which in this case is a single gene (YHR007C), and asked for all its pathways. Figure9 shows the result containing the names of the pathways.

The user can also select a set of genes and invoke text summarization application. An example output is shown in Figure10. Additionally, she can score the pathways based on a scoring algorithm. Figure 11 gives one of the output views.

6 Conclusion and Future Work

A biologist working with microarrays needs an handle over not only gene expression data and their analyses, but also on annotations of related data like pathway scores, structural similarity, or summarization from available literature on the genes. In this paper, we presented a comprehensive bioinformatics KM

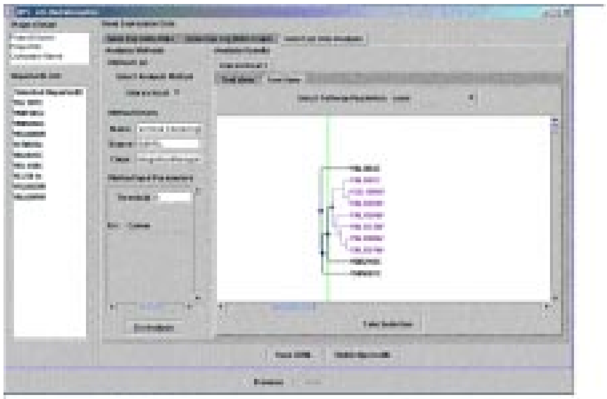
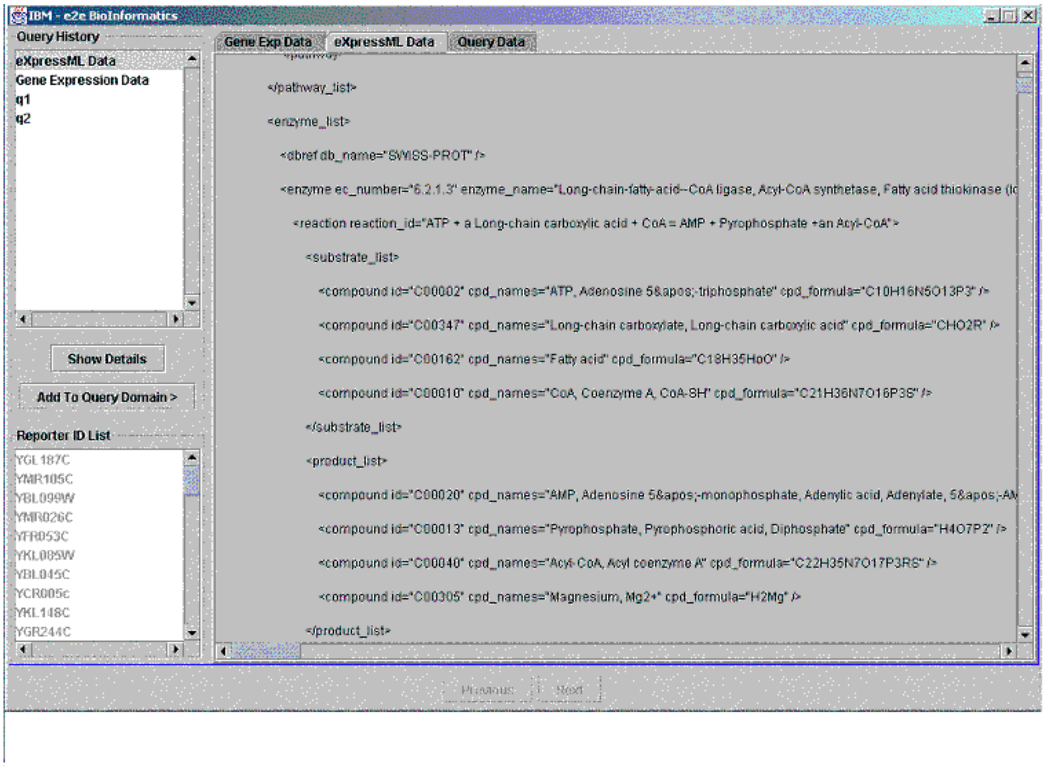
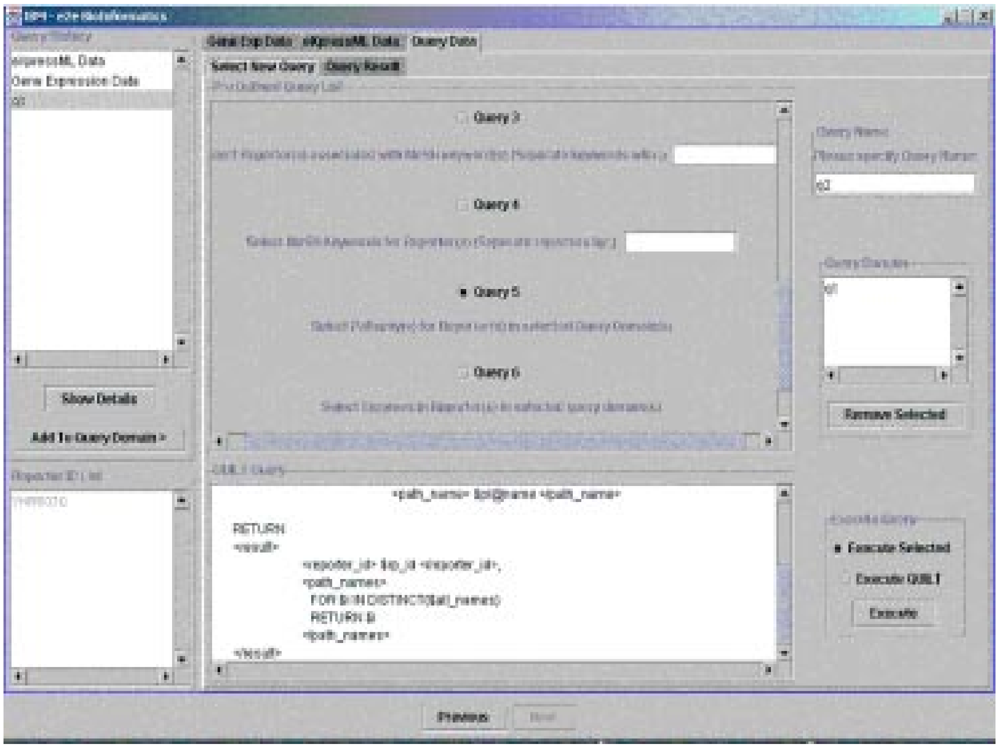


Figure 6: Clusters of genes with similar expression level.



```
</pathway_list>

<enzyme_list>

  <dbref db_name="SWISS-PROT" />

  <enzyme ec_number="6.2.1.3" enzyme_name="Long-chain-fatty-acid--CoA ligase, Acyl-CoA synthetase, Fatty acid thiokinase (lo

    <reaction reaction_id="ATP + a Long-chain carboxylic acid + CoA = AMP + Pyrophosphate +an Acyl-CoA">

      <substrate_list>

        <compound id="C00002" cpd_names="ATP, Adenosine 5&apos;-triphosphate" cpd_formula="C10H16N5O13P3" />

        <compound id="C00347" cpd_names="Long-chain carboxylate, Long-chain carboxylic acid" cpd_formula="CHO2R" />

        <compound id="C00162" cpd_names="Fatty acid" cpd_formula="C18H35HoO" />

        <compound id="C00010" cpd_names="CoA, Coenzyme A, CoA-SH" cpd_formula="C21H36N7O16P3S" />

      </substrate_list>

      <product_list>

        <compound id="C00020" cpd_names="AMP, Adenosine 5&apos;-monophosphate, Adenylic acid, Adenylate, 5&apos;-AM

        <compound id="C00013" cpd_names="Pyrophosphate, Pyrophosphoric acid, Diphosphate" cpd_formula="H4O7P2" />

        <compound id="C00040" cpd_names="Acyl-CoA, Acyl coenzyme A" cpd_formula="C22H35N7O17P3RS" />

        <compound id="C00305" cpd_names="Magnesium, Mg2+" cpd_formula="H2Mg" />

      </product_list>
```

Figure 7: The eXpressML constructed for selected genes by *dynamically* integrating data from KEGG, SWISS-Prot and PubMed.



```
                    <path_name> $pl@name <path_name>

RETURN
<result>
            <reporter_id> $rp_id </reporter_id>,
            <path_names>
              FOR $i IN DISTINCT($all_names)
              RETURN $i
            </path_names>
</result>
```

Figure 8: Query interface for posing Quilt queries on expresssion data and eXpressML.

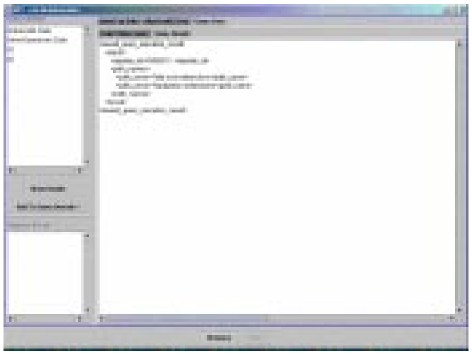


Figure 9: Result of the Quilt (XML) query posed in Figure 8.

framework called *e2e* which provides a uniform window to biochip data and related annotations. To the biologist or biological applications, *e2e* exposes a common semantic view of inter-relationship among biological concepts in the form of an XML representation called eXpressML. Internally, *e2e* can use any data integration solution (like DiscoveryLink, Kleisli or natively XML-based) to retrieve data and return results corresponding to the semantic view.

We demonstrated an *e2e* prototype that gives an early glimpse of the wide potential of an integrated KM solution for bioinformatics. Biologists who have used the *e2e* prototype value the ability it provides to cross-relate concepts and analytics from different areas. However, they want to run it with larger expression data (1000s of genes), something for which the current *e2e* prototype is slow due to the in-memory storage of XML. This limitation will be addressed in a future re-implementation of *e2e*.

We are looking at extending *e2e* along various directions:

- Improve annotation quality for different types of data. Users specifically want advanced text summarisation support that leverage known biological ontologies.

- Extend the range of annotations and the types of related data.

- Improve query interface to allow the biologist to issue natural language queries which get translated to necessary format and structure of the underlying data model.

- Improve retrieval of unstructured data along with issues like change detection and caching of results.

- Address middleware issues of effective query decomposition and scalability in the presence of large data (through available database technologies) and domain knowledge of biology[28].

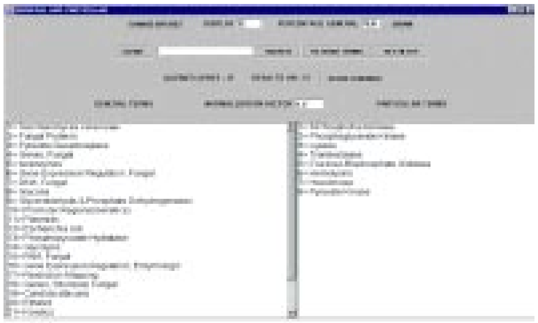


Figure 10: Summary of top keywords in medical literature that correspond to the selected genes.

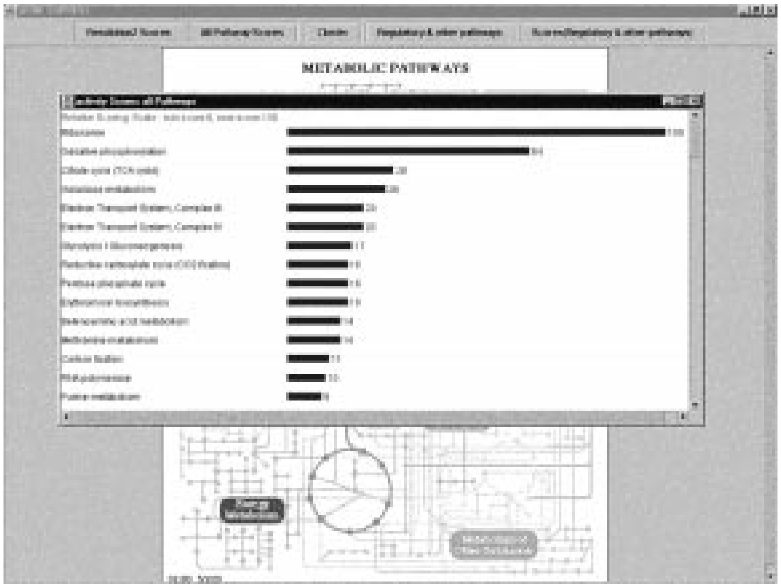

METABOLIC PATHWAYS
Ribosome 100
Oxidative phosphorylation 64
Citrate cycle (TCA cycle) 28
26
Electron Transport System, Complex III 20
Electron Transport System, Complex IV 20
Glycolysis / Gluconeogenesis 17
Reductive carboxylate cycle (CO2 fixation) 16
Pentose phosphate cycle 16
16
Selenoamino acid metabolism 14
Methionine metabolism 14
Carbon fixation 11
RNA polymerase 10
Purine metabolism 9

Figure 11: A visualization of pathway scores corresponding to the selected genes.

References

[1] Adak, S., Srivastava, B., Kankar, P., and Kurhekar, M. 2001. A Common Data Representation for Organizing and Managing Annotations of Biochip Expression Data. *Unpublished Technical Report.*

[2] Baxevanis, A. 2001. The Molecular Biology Database Collection: an updated compilation of biological database resources. *Numcleic Acids Research, Vol. 29, No. 1.*

[3] Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Research, 8:1202-1215.*

[4] Brown, P. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics, 21:33-37.*

[5] Buneman, P., Davidson, S. Hart, K., Overton, C., and Wong, L. (1995). A Data Transformation System for Biological Data Sources. *Proc. VLDB, pp 158–169.*

[6] Chen, I., Kosky, A., Markowitz, V., and Szeto, E. (1997). Constructing and maintaining scientific database views in the framework of the object-protocol model. *Proceedings SSDBM, pages 237-248.*

[7] Davidson, S., Crabtree, J., Brunk, B., Schug, J., Tannen, V., Overton, C., and Stoeckert, C. (2001). K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. *IBM Systems Journal, March 2001.*

[8] Duggan, D., Bittner, M., Chen, Y., Meltzer, P. and Trent, J. (1999). Expression profiling using cDNA microarray. two-color fluorescent probe hybridization. *Nature Genetics, 21:10-14.*

[9] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings Natl Acad Sci USA, 95:14863-14868, 1998.*

[10] Etzold, T., and Argos, P. (1993). SRS: An Indexing and Retrieval Tool for Flat File Data Libraries. *Computer Application of Biosciences, 9:49-57.*

[11] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., and Kanehisa, M. (1998). DBGET/LinkDB: an Integrated Database Retrieval System. *Pacific Sym. Biocomputing, pp 683-694.*

[12] Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., and Brass, A. (2001). Transparent Access to Multiple Bioinformatics Information Sources. *IBM Systems Journal, Vol. 40, No.2, pp 532-551.*

[13] Golub, T., Slonim, T., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., and Caligiuri M. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286:531-537.*

[14] Haab, B., Dunham, M. and Brown, P. (2001). Protein microarrays for highly parallel detection and quantification of specific proteins and antibodies in complex solutions. *Genome Biology, 2(2):research0004.1-0004.13.*

[15] Haas, L., Schwarz, P., Kodali, P., Kotlar, E., Rice, J., and Swope, W. (2001). DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal, Volume 40, Number 2, 2001.*

[16] Kankar, P., Adak, S., Sarkar, A., Murari, K. and Sharma, G. (2002). MedMeSH Summarizer: Text Mining for Gene Clusters. *To appear in Proceedings of the SIAM conference in Data Mining.*

[17] Kurhekar, M., Adak, S., Jhunjhunwala, S., and Raghupathy, K. (2002). Genome-wide pathway analysis and visualization using gene expression data. *To appear in Proceedings of the Pacific Symposium of Biocomputing.*

[18] Levy, A. 1998. Combining Artificial Intelligence and Databases for Data Integration. *At http://citeseer.nj.nec.com*

[19] Lipshutz, R., Fodor, S., Gingeras, T. and Lockhart, D. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics, 21:10-14.*

[20] Lockhart, D., Dong, H., Byrne, M. and Follettie, M. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology, 14:1675-1680.*

[21] Mangalam, H., Stewart, J., Zhou, J. and Schlauch, K. (2001). GeneX: An Open Source gene expression database and integrated tool set. *IBM Systems Journal, 40(2):552-569.*

[22] Matysiak, S., Wurtz, S., Hauser, N., Gausepohl, H. and Hoheisel, J. (1999). PNA-arrays for nucleic acid detection. *Peptide Nucleic Acids: Protocols and Applications. (P Nielsen & M Egholm, eds.), Horizon Scientific Press, Wymondham, 119-128.*

[23] Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: encyclopedia for Genes, Proteins, and Diseases. *Tech. Report, Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Rehovot, Israel.*

[24] Robie, D., Chamberlin, D. and Florescu, D. (2001). Quilt: an XML Query Language. *http://www.almaden.ibm.com/cs/people/chamberlin/quilt_euro.html*

[25] Schena, M., Shalon, D., Davis, R. and Brown, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270:467-470.*

[26] Shalon, D., Smith, S. and Brown, P. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research, 6:639-645.*

[27] Shatkay, H., Edwards, S., Wilbur, J. and Bogusk, M. (2000). Genes, themes and microarrays: Using information retrieval for large-scale gene analysis. *Proceedings of ISMB'00.*

[28] Srivastava, B. 2002. Using Planning for Query Decomposition in Bioinformatics *Sixth Intl. Conf. on AI Planning & Scheduling (AIPS-02) Workshop on "Is There Life Beyond Operator Sequencing? – Exploring Real World Planning".*