A System for Knowledge Management in Bioinformatics

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
{asudeshn, bvishal, dbhardwa, pkamesam, kpankaj, kmanish, sbiplav}@in.ibm.com

ABSTRACT

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 address this problem in a comprehensive system for knowledge management in bioinformatics called *e2e*. 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, while internally, it can use any data integration solution to retrieve data and return results corresponding to the semantic view. We have implemented an e2e prototype that enables 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.

Categories and Subject Descriptors

H.0 [**Information Systems**]: General

General Terms

Management, Performance

Keywords

Knowledge Management, Bioinformatics, Biochips

*Lead contact. Author names appear in alphabetic order.
Formerly with IBM.

CIKM'02, November 4–9, 2002, McLean, Virginia, USA.

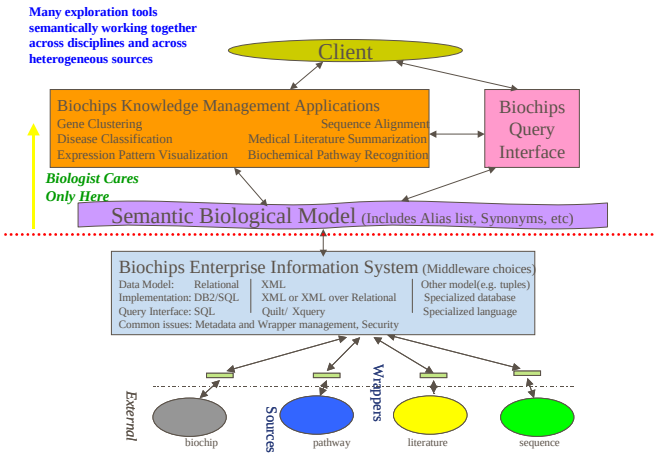

Figure 1: *e2e* Biochips Information System Framework Architecture.

1. INTRODUCTION

DNA based microarray technologies[14] 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 (e.g., NCBI's Gene Expression Omnibus[1](GEO), Stanford Microarray Database[2](SMD)) 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 mechanisms 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. However, related biomedical data[3] are numerous and 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.

[1]http://www.ncbi.nlm.nih.gov/geo/
[2]http://genome-www4.Stanford.EDU/MicroArray/SMD/

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 cross-referenced and validated with additional insights from related concepts. 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[9] , Kleisli[5] 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. More details on *e2e* can be found in the extended paper[2].

Here is the layout of the paper: we start with a background of gene expression data and discuss current approaches for data integration for bioinformatics. Next, we introduce the *e2e* framework that maintains the user's biological perspective. 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. We conclude with our contributions and future work.

2. BACKGROUND

The developments in the area of microarrays and data integration are shaping genomics today.

2.1 Gene Expression and Biochips

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.

In order to realize the full potential of biochips, the main challenges faced by the life sciences industry today are: (a) Improvements in the core microarray technology to improve the accuracy of gene expression measurement (b) Development of the full spectrum of specialized analytics and (bio)informatics tools required for making (biological) knowledge discoveries from biochip 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.

A variety of approaches have been developed for integrated access to heterogeneous data sources in genomics. In the *link-driven federation* approach, the user can switch between sources using system-provided links in a hypermedia environment. The user has to still interact with individual sources; only the interaction is easier through convenient

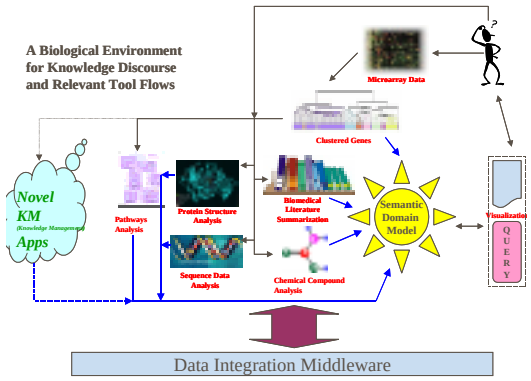

Figure 2: A Schemata of Flow between KM Analyses in *e2e*.

links and not invoking the sources directly. SRS[7] is an example of this approach. Though the link driven approach is very convenient for non-expert users and provides limited keyword search capability on the content of a source, 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 [12]. 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. Examples are IBM's DiscoveryLink[9] and Kleisli[5] which provide powerful querying capabilities, but fail to provide the in-depth analysis that are provided by the "point solutions".

A biologist 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. A system related to our definition of semantic integration is TAMBIS[8] 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. *E2E* FRAMEWORK

We now discuss *e2e* in which semantic relationship among biological concepts is represented in the XML representation of eXpressML[1] and analytical KM tools can work from this

abstraction. The biologist only needs to know about the biological domain while the system will hide the peculiarities of the sources involved to answer a domain query. 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 handle large data sizes) which can retrieve either microarray experimental data or external information from publicly available biological data 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[9] 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.

The semantic biological model provides the user with a common biological context to view and manipulate related data and issue XML queries in Quilt[13] 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. For example, some tools that can be used are: pathway visualization tools[11] for annotating gene clusters with pathway information, text summarization tools[10] for annotating gene clusters with biological function, and sequence alignment tools[4] 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.*

3.1 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}[3], which has been accepted relatively widely by 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: (a) 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. (b) The common view should allow querying, modeling, and browsing of complex annotations. (c) 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.

Related data includes DNA, protein, 3D structure, genetic maps of disease, biochemical pathways, enzyme, keywords, medical citation information, and is obtained either directly or by running KM tools on data from heterogeneous data sources. Note that gene expression data itself is not part of eXpressML. A related effort is MAGE-ML[4] 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 [13] (specifically, Kweelt[5] implementation of Quilt). The query interface has templates for a number of pre-canned queries and the user can also pose any Quilt query which is valid.

Example: 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, which are representable in eXpressML.

3.2 KM Layer

The KM layer consists of two types of applications: (a) Tools for detecting gene expression patterns by supporting clustering, classification, and visualization of biochip experimental data, and (b) Downstream annotation tools combining expression data with other sources of information to improve the range and quality of conclusions that can be drawn. 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. KM APPLICATIONS

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[6] 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: (a) Any initialization parameters and the format of input gene expression data (tabular or XML). (b) The format of output result. (c) If the algorithm supports visualization, a *handle* of the input and output panels.

We have implemented hierarchical clustering and K-means clustering in the *e2e* prototype.

4.2 Text Summarization

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

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

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

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 [10] 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 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. 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 and produces summaries across all biological domains.

4.3 Pathways Scoring

At the cellular level, organisms function through intricate networks of chemical reactions (metabolic pathways) and interacting molecules (regulatory pathways). 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 systems serve as an important tool for interpreting the 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 algorithms [11] which use gene expression data and putative metabolic and regulatory pathways database of KEGG. The outputs are: *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.

4.4 Protein Sequence Analysis

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. CONCLUSION AND FUTURE WORK

In this paper, we presented a comprehensive bioinformatics KM framework called *e2e* which provides a uniform window to biochip data and related annotations. 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.

We are looking at extending *e2e* along various directions: (a) Address middleware issues of effective query decomposition and scalability in the presence of domain knowledge of biology[15] and large data (through available database technologies). (b) Extend the range of annotations and the types of related data. (c) Improve query interface to allow the biologist to issue natural language queries. (d) Improve retrieval of unstructured data along with issues like change detection and caching of results.

6. REFERENCES

[1] Adak, S., Srivastava, B., Kankar, P., and Kurhekar, M. 2002. A Common Data Representation for Organizing and Managing Annotations of Biochip Expression Data. *IBM Research Report RI02017. Available at http://domino.watson.ibm.com/library/CyberDig.nsf/Home*

[2] Adak, S., Batra, V., Bhardwaj, D., Kamesam, P., Kankar, P., Kurhekar, and Srivastava, B. 2002. Bioinformatics for Microarrays. *IBM Research Report RI02016. Available at http://domino.watson.ibm.com/library/CyberDig.nsf/Home*

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

[4] 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.*

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

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

[8] 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.*

[9] 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.*

[10] Kankar, P., Adak, S., Sarkar, A., Murari, K. and Sharma, G. (2002). MedMeSH Summarizer: Text Mining for Gene Clusters. *In Proc. of the SIAM Conf. in Data Mining.*

[11] Kurhekar, M., Adak, S., Jhunjhunwala, S., and Raghupathy, K. (2002). Genome-wide pathway analysis and visualization using gene expression data. *In Proc. of the Pacific Symposium of Biocomputing.*

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

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

[14] 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.*

[15] 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".*