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

The BioMediator System as a Data Integration Tool to Answer Diverse Biologic Queries

Loren Donelson[a], Peter Tarczy-Hornoch[a], Peter Mork[a], Cindy Dolan[a], Joyce A. Mitchell[b], M. Barrier[a], Hao Mei[a,c]

[a]University of Washington, Seattle WA, [b]University of Missouri, Columbia MO, [c]North Carolina State University, Raleigh NC

Abstract

We present the BioMediator (www.biomediator.org) system and the process of executing queries on it. The system was designed as a tool for posing queries across semantically and syntactically heterogeneous data particularly in the biological arena. We use examples from researchers at the University of Washington, and the University of Missouri-Columbia, to discuss the BioMediator system architecture, query execution, modifications to the system to support the queries, and summarize our findings and our future directions. Finally, we discuss the system's flexibility and generalized approach and give examples of how the system can be extended for a variety of objectives.

Keywords:

BioMediator, data integration, bioinformatics

Introduction

High throughput data collection techniques developed in the last few years have led to an explosion in the number of publicly available biological databases and a new wealth of information for biologic researchers [1]. The disparate locations (from those housed at the NCBI, to those developed by consortiums and individual laboratories) and structures of the sources, however, create considerable complexity for individuals extracting related data from multiple sources.

A researcher investigating genetic information on a specific disease, for example, might be required to traverse several data sources including OMIM [2], LocusLink [3], GeneTests [4], and Swiss Prot [5]. For each source they would need source contents, organization, query syntax, and the most reliable links between sources. In a large study (e.g., one involving protein or expression array data), the process could be replicated tens, hundreds, or thousands of times and become prohibitively time consuming.

Users with novel and increasingly diverse uses for the same information make it impossible to design apriori query paths applicable to all potential queries. Consider a systems biologist interested in proteins active in a system during a particular biological process, versus a researcher using expression array data to determine highly conserved motifs within an experiment, or a clinical geneticist investigating genetic disorders caused by mutations in the same gene. All three individuals could utilize the same collection of databases to answer their queries, but they would no doubt have different questions to ask each source.

Adding to this complexity, dissimilar disciplines often have dissimilar 'views of the world' (schemata). The schema used by a molecular biologist might center on DNA and RNA structure, while that of a systems biologist centers on proteins and protein functions, and that of a clinical geneticist on genes, mutations, altered protein products and phenotypes.

Since research across domains (e.g., comparing genomes across species) has become common, a data integration system should 1) have broad application to a wide variety of potential users, 2) be flexible to accommodate diverse queries, 3) allow dissimilar centralized (mediated) schemata to support those queries, and 4) be modular enough to evolve as the data sources and uses for those sources increase.

The BioMediator (formerly GeneSeek [6]) system (www.biomediator.org) developed at the University of Washington makes querying multiple data sources with different schemata transparent to the user, while maintaining the source autonomy. We first describe the architecture of the system, providing an overview of the major software components and the communication between them (e.g. XML, XQuery). Second, we discuss the process of developing and executing queries on the system using the help of biological researchers to formulate research queries and validate their translation into system queries. Finally, we summarize our findings regarding successes and challenges of the current system, and discuss current and future uses for the BioMediator system.

Related Work

Two primary strategies exist for creating unified access to data sources—data warehousing and database federations [7]. Data warehouses consolidate the data of multiple sources into a unified location by retrieving copies of the sources. This technique allows a researcher fast access to data, but the data are only as relevant as the last time copies of the underlying sources were retrieved. Moreover, to facilitate data aggregation, warehouses contain a global schema making them less applicable to diverse groups with diverse schemata.

By comparison, database federations retrieve data only when processing a user query. These federations ameliorate the "freshness" problems of data warehousing because the data come from the live sources themselves, but often either require

that the user have in depth knowledge of the underlying sources so that they can formulate specific queries for each source, or restrict the individual querying the system to a generalized (typically fixed) central schema.

Several examples of federation projects exist in the biomedical domain, including OPM [8], Kleisli [9] and TAMBIS [10].

The Object Protocol Model (OPM) [10] is a semantic data model used to encapsulate data sources using declarative mappings and query-translation techniques to process queries across them. This presents a challenge in that each underlying source needs its own subschema developed [7] and users interested in querying all sources must understand all the subschemata to design an effective query.

The Kleisli [9] middleware solution has several incarnations, including Bio-Kleisli for molecular biology that allows a researcher to query for a variety of elements using the Collection Programming Language language. Like OPM, Kleisli's lack of a mediated schema requires that the researcher have in depth knowledge of the underlying source semantics and syntax to create the queries [7], cumbersome for queries involving a large number of sources.

The Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) [10] project uses a knowledge-based global ontology for the federation schema. [7] The global ontology (or mediated schema) represents the universe of potential concepts for queries, which a researcher augments by combining concepts. Users are able to tailor the schema to their specific needs, but the fundamental schema remains the same, making it difficult for users with widely different schemata to use the same system.

BioMediator System Overview

The BioMediator system (Figure 1) addresses the issues presented above through a modular design consisting of six components that perform data integration over multiple structured and semi-structured biological data sources.

In a broad sense, the BioMediator system defines and traverses a graph where the nodes represent data source instances for entities in the mediated schema. The edges represent instances of the relationships that connect entities in one or more sources in the schema. Using the graph, a path between two entities of interest can be constructed by concatenating several edges in the graph.

PQL (Figure 1A) [11] is a path based query language with rules permitting the user to specify query and path constraints amongst the databases in the federation. The reformulator (Figure 1B) accepts PQL input queries and enumerates all paths a) allowed by the context free grammar specified by the query, and b) supported by sources in the mediated schema, then outputs a set of queries in XQuery (a query language for semi-structured data) [12]. The source knowledge base (SKB) [6] (Figure 1C) is represented in Protégé [13] and accessed via the Protégé API. It contains a) all entities, attributes, and relationships in the mediated schema, b) the catalog of all possible sources of data and the mediated schema elements they contain, c) the mapping rules for bidirectional semantic translation of queries and source data streams [14]. The query execution engine (Qexo [15], Figure 1E) accepts XQuery input, and outputs queries as URLs in mediated schema syntax and semantics. Qexo retrieves XML documents and performs dependent joins with the results to determine subsequent documents to query. The metawrapper [14](Figure 1E) translates URLs into source specific queries using the forward mapping rules stored in the SKB. Finally, wrappers (Figure 1F) pass the queries to the specific sources. The source results are returned as an XML document through the basic components as depicted in Figure 1.

The BioMediator system addresses the problems presented by traditional database federation integration methods through three main concepts.

First, the system uses an *annotated* mediated schema to describe a 'universe' of biological data. Each user can create a completely customized mediated schema to describe their view of the 'universe' and pose queries against it. Users annotate the schema by adding additional information about data sources and their relationships. For example a user could annotate the relationship between two databases as "validated" meaning that the linkage had been curated by an external consortium. The user could restrict their query to only "validated" paths.

Second, the source knowledge base is modified through a relatively easy to use graphical user interface. Consequently, the key components (mediated schema, source knowledge, and translation rules) are easily adjusted to meet the needs of a particular researcher.

Third, the data source wrappers in the BioMediator system are generalized, such that they expose all relevant data from a data source whether or not it maps to a particular mediated schema. A source wrapper needs to be updated only in three cases: 1) when a data source extends its underlying schema and the mediated schema maps to the extensions 2) when a source changes its schema, and 3) when a source changes its interface.

Methods

The BioMediator system was developed with the intent that it could be used as a tool by a biologist without informatics training. As a final step during the development of the system, we extended our component testing to include a user beta test.

The goal of the beta testing described in this paper was to determine whether the system was functional and flexible enough to answer biologists queries, and what improvements and modifications needed to be made to make the tool useful in the research setting. Specifically, we were interested in what improvements, modifications, user interface and documentation were necessary for the system to be widely distributed and used by non-informaticist biologists.

To that end, we identified a group of four researchers with diverse questions across a variety of biological databases, and a willingness to collaborate. Each user had knowledge of the BioMediator project in a holistic sense, but had distinct interests, queries, and uses for the system. The researchers were: Hao Mei, at the time, a biomedical informatics master's candidate at

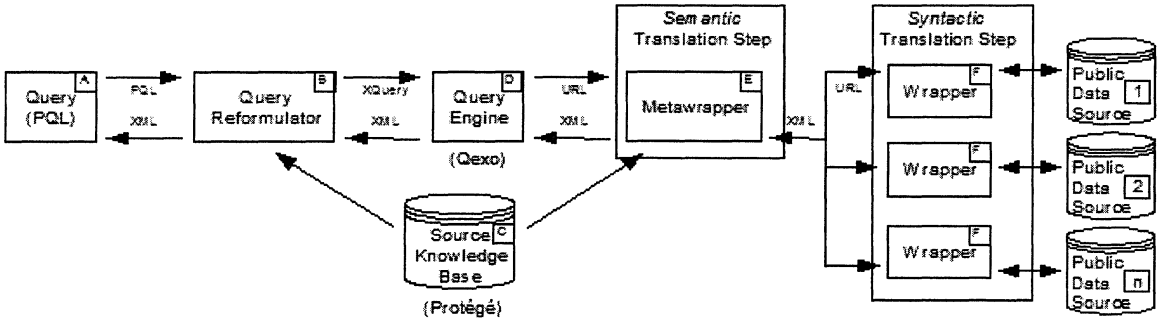

Figure 1 - BioMediator Architecture

the University of Washington developing a thesis on expression array annotation [16].

Joyce Mitchell, a geneticist and informaticist at the University of Missouri-Columbia and a Senior Scholar at the National Library of Medicine investigating online data sources for genetic researchers [17], clinicians and the public [18].

Cindy Dolan, genetic counselor with the GeneTests [4] resource, specifically interested in maintaining/updating information held in the GeneTests database.

Marianne Barrier, a post-doctoral molecular biologist at the University of Washington utilizing gene and protein expression array data.

For each group we developed system queries using the four major steps detailed below. In each step, we focus on Dolan's queries because they represent a typical example of the steps involved using BioMediator for real world data integration queries. Experiences with the other three researchers are also mentioned to provide additional insight into the query development process and end user requirements.

1. Obtain the natural language queries through user interviews. We first met with each of the four users to discuss their data integration query needs. After a brief overview of the BioMediator system (similar to that presented in the system overview above), the users were asked to articulate (in words) the underpinnings of their query or queries. For Dolan, the natural language query identified was:

> Given a disease (phenotype) name, find the associated gene(s), loci, and protein(s) in externally validated databases

This query involves several potential data sources and paths between them. Additionally, Dolan was interested in obtaining particular information from specific sources—gene symbols from HUGO [19], locus from OMIM and protein information from Swiss Prot. After further discussion, we agreed that information from other sources was acceptable provided that data from the aforementioned three sources was exhaustively retrieved.

Dolan's query also highlights annotation information attached to the mediated schema which we refer to as mediated schema metadata. Each relationship defined between sources (relevant to her query), would include an attribute specifying the validity of the relationship. This constraint was used to prune potential paths in the graph to just those that Dolan had confidence in (which she characterized as externally validated).

The four users varied in the completeness of the queries and in the breadth of desired results. Mitchell was less concerned about specific sources while Mei specifically enumerated initial paths for mediated schema entities based on his experience working with the underlying systems manually [11].

2. Translate the natural language queries into PQL queries and validate the PQL translation with the user. We began by identifying valid relationships (edges) among entities (nodes) in the mediated schema as those that were externally validated and causal. These restrictions allowed us to focus our query on relationships that Dolan felt were most useful for maintaining and updating the gene, locus, and product information stored in the GeneTests database. We then defined valid paths amongst entities as those comprised of valid edges, allowing the BioMediator system to create all joins between entities. The resultant PQL query is presented below:

```
USING
    X!validEdge == TRUE :- X!__isCausal == TRUE,
                           X!validation == "external";
    X!validPath == TRUE :- X!validEdge == TRUE;
    X.Y!validPath == TRUE :- X!validPath == TRUE,
                           Y!validPath == TRUE
WHERE
    Phenotype(Ph),
    Ph->"name"->"disease name",
    Gene(G),
    Ph->AcyclicByEdge(P)->G,
    P!validPath == TRUE
    Protein(Pr)
    Ph->AcyclicByEdge(P)->Pr
```

While each of the four users had distinct queries, paradigms emerged that allowed us to create subsequent queries faster. In general the USING and WHERE clauses (described in detail elsewhere [11]) follow the same structure as the query above, so we only needed to modify the constraints and entities of interest to create similar queries.

After creating the PQL query we contacted the researchers again to discuss the query, and confirm that we had correctly understood the data elements and path constraints of interest.

To determine the steps necessary to generate a PQL query from a researcher for beta testing purposes, we were particularly thorough in our query development. Long term we anticipate that a user would not need to perform the same analysis, but instead use a tool or set of tools to automatically create a query in PQL syntax from their research question.

3. Modify the Protégé source knowledge base (SKB) to support the researchers desired mediated schema, and confirm that the existing wrappers expose all relevant source information. Using the PQL queries as a guide for the schema entities and relationships of interest, we examined our test bed SKB and existing wrappers to confirm that the necessary elements were in place. Since our testing mediated schema contained entities for phenotype, genotype, and protein, we were able to submit the PQL query developed for Dolan without further modifications to the SKB. The gene entity contained attributes both for gene ID and locus, so query results were compiled simply by returning genes and proteins associated with a disease name.

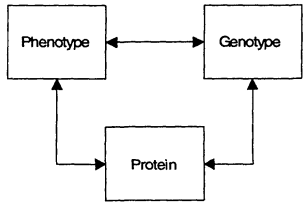

Figure 2 - High level mediated schema

In contrast, one of Mitchell's queries involved finding laboratories testing for a particular phenotype, so we modified the SKB using the Protégé GUI to include ClinicalTrial and Laboratory entities, attributes, and relationships PhenotypeAssociatedWithClinicalTrial and PhenotypeTestedByLaboratory. For microarray annotation, Mei was interested in clone information from IMAGE [20], so we modified the SKB to include a clone entity (among others).

In the future a few major mediated schemata will most likely emerge as useful for a large number of similar users. Individuals like Mei and Mitchell would probably augment an existing schema with additional entities or attributes, rather than create them de novo. Hence, the architecture allows and encourages users to have distinct schemata, but at the same time does not require the extra overhead for conceptually similar users. The use of the Protégé tool permits sharing and reuse of schemata between users in a well defined format.

The current BioMediator system includes wrappers for GO [21], HUGO, LocusLink, OMIM, Entrez [22], Swiss Prot, GEO [23], IMAGE and GeneTests, which were sufficient for answering the queries of our four collaborating researchers.

4. Execute and examine results. The last step involved submitting the query and a manual examination of the XML output provided by the reformulator. For Dolan we ran queries based on diseases with complete information in the GeneTests database to determine whether the BioMediator system found the same entities and attributes. The results were promising, with the BioMediator system returning all requested entities possible with the current list of queryable data sources.

A formal evaluation of the mature BioMediator system is needed to determine whether the results returned are useful to the researchers who pose them. This validation would include a time comparison based on the amount of time a researcher takes to find the answers to a query vs. the time spent developing and executing the PQL query, as well as an examination of the results returned vs. those returned manually (e.g. precision and recall [24]).

Findings

The BioMediator system's flexibility and generalized approach provide many benefits and accommodated all four users effectively. The system was able to answer the vast majority of the queries with minor (and straightforward) modifications.

Benefits of modular architecture: The BioMediator system's modular architecture was extremely effective. During the beta testing (methods above) we determined that using GNU Qexo would provide us benefits above and beyond our previous XQuery engine. We were able to research and swap in Qexo to process XQueries in roughly a single afternoon. Similar substitutions will allow the BioMediator system to change as new and novel technologies improve the data integration effort.

Benefits of path base query language: PQL's flexibility as a path based query language was valuable, allowing us to bind variables to not only entities within the mediated schema, but also to the paths themselves. Path constraints based on annotation information in the mediated schema pruned potentially large data sets into highly targeted, relevant, results.

Benefits of Protégé: The Protégé platform is an excellent format to represent mediated schemas within the BioMediator system. The combination of the Protégé GUI with other tools (current and future work below) eliminates the need for biologists to do programming as their needs and schemas evolve.

Need for NLP: Several of Mitchell's queries involved conducting literature searches and reviews to obtain the query answers. Since the BioMediator system does not currently perform any natural language processing, results for those queries came only from structured or semi-structured databases. Mining data from natural language unstructured documents could provide value for similar user queries.

Need for analytic tools: Barrier's queries highlighted a need by researchers to use analytic tools (e.g. BLAST) as part of the entity/relationship (node/edge) constraints. One of her questions involved using BLAST to determine similar peptide sequences from a series and then retrieving gene/protein information for each similar sequence. The current BioMediator system will assist her by simplifying the process of identifying genes and proteins once the similar sequences are determined, a valuable future incarnation would be one that performs the whole process for a given peptide sequence.

Need for better interfaces for biologists: The concepts of the PQL language are initially challenging, but once the general ideas are grasped, creating queries from examples becomes a relatively easy process. Developing explanatory diagrams, language documentation and sample queries will facilitate the process. Query generation could also be augmented using "que-

ry by example" (discussed below) techniques that exercise the Protégé API to present a GUI interface to generate PQL syntax.

Need to accommodate source evolution: A final challenge we discovered was that the interfaces to the underlying sources do change, and consequently the wrappers accessing those sources need to be updated with equal frequency. This will no doubt decrease as the sources mature, but is a challenge for all data integration projects. The BioMediator system's generalized approach minimizes the impact of such changes, wherein only the wrapper to a modified source (and perhaps its mediated schema mapping) needs updating [14].

Current and Future Work

Based on the findings above we have continued work on the BioMediator system along several avenues. Although the BioMediator system can be used as a standalone application, in practice it is even more effective when combined with other technologies. As mentioned above, we plan to implement a front-end interface designed to expose the mediated schema in Protégé, allowing researchers to select entities, add attributes to them, and specify path constraints without having to understand the specifics of the PQL language (a QBE – Query By Example – interface analogous to the one provided in Microsoft Access). Likewise, users will probably augment the system by post-processing the XML results into a human readable form.

To assist Barrier with her expression array analysis, a UW informatics researcher used the system in series with a component that parses expression array output, performs a BLAST search and then uses the results to begin annotating peptide or nucleotide sequences of interest. We plan to extend this development so that similar analytic tools can be used in much the same way that data sources are accessed. We have also begun development on a GUI tool to facilitate adding or changing sources in the database federation. This tool will assist a schema developer in defining the pair wise mappings between XML data from the sources to the XML representation of the mediated schema.

Finally, using Dolan's queries as a test bed, we are currently creating software and a methodology to test the efficacy of the system. Among other metrics, we are interested in the the BioMediator's precision and recall as a proxy for how useful it is to researchers.

Acknowledgements

We would like to thank R. Shaker and R. Pagon for their participation. Funding was provided by NHGRI (R01HG02288, Tarczy-Hornoch, PI) and NLM training grant (T15LM07442, trainees: Mork & Donelson).

References

[1] Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001;292(5518):929-34.

[2] OMIM. Online Mendelian Inheritance in Man. 2003.

[3] LocusLink. LocusLink. In; 2003.

[4] GeneTests. In; 2003.

[5] Swiss-Prot. In; 2003.

[6] Mork P, Halevy A, Tarczy-Hornoch P. A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases. *Jour Amer Med Inform Assoc, Fall Symposium Suppl* 2001:473-477.

[7] Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform* 2001;34(4):285-98.

[8] Chen IA, Markowitz VM. An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools. *Information Systems* 1995;20(5).

[9] Chung SY, Wong L. Kleisli: a new tool for data integration in biology. *Trends Biotechnol* 1999;17(9):351-5.

[10]Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton N, Goble C, Brass A. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000;16(2):184-5.

[11]Mork P, Shaker R, Halevy A, Tarczy-Hornoch P. PQL: A Declarative Query Language over Dynamic Biological Schemata. *Jour Amer Med Inform Assoc, Fall Symposium Suppl* 2002.

[12]W3C. XQuery 1.0: An XML Query Language. In; 2003.

[13]Stanford. Protege Home Page. In; 2002.

[14]Shaker R, Mork P, Barclay M, Tarczy-Hornoch P. A Rule Driven Bi-Directional Translation System Remapping Queries and Result Sets Between a Mediated Schema and Heterogeneous Data Sources. *Jour Amer Med Inform Assoc, Fall Symposium Suppl* 2002:In Press.

[15]Qexo. In; 2003.

[16]Mei H. Expression Array Anotation Using the BioMediator Biologic Data Integration System and the Bioconductor Analytic Platform. *Submitted 3/12/03 to Fall 2003 AMIA Symposium* 2003.

[17]Mitchell J. From Phenotype to Genotype: Experiences in Navigating the Available Information Resources. In: AMIA 2002 Annual Symposium; 2002; 2002. p. 1109.

[18]Mitchell J, McCray, AT, Bodenreider, O. From phenotype to genotype: issues in navigating the available information resources. *Methods of Information in Medicine, in press* 2003.

[19]HUGO. The Human Genome Organisation. In; 2003.

[20]IMAGE. The I.M.A.G.E. Consortium. In; 2003.

[21]GO. Gene Ontology Consortium. In; 2003.

[22] Entrez. Entrez Search and Retrieval System. In; 2003.

[23]GEO. Gene Expression Omnibus. In.

[24]Liu X AR. Updating a bibliography using the related articles function within PubMed. In: *Proc AMIA Symp.*; 1998; 1998. p. 750-4.

Address for correspondence

Loren J. Donelson

159 N.E. Pacific Street, HSB I-264

Box 357240 Seattle, WA 98195-7240