2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Requirement Analysis for Bio-Information Integration Systems

Sean Lee[1], PhilHyoun Lee[1], Dokyun Na[1], Doheon Lee[1], Kwanghyung Lee[1], MyungNam Bae[2]
1. { phlee, blisszen, seanlee, dhlee, kwlee } @bioif.kaist.ac.kr;
Dept. of Biosystems, Korea Advanced Institute of Science and Technology
2. mnbae@etri.re.kr; Electronics and Telecommunications Research Institute

Abstract —**Amount of biological data information has been increasing exponentially. In order to cope with this bio-information explosion, it is necessary to construct a biological data information integration system. The integration system could provide useful services for bio-application developers by answering general complex queries that require accessing information from heterogeneous bio data sources, and easily accommodate a new database into the integrated systems. In this paper, we analyze architectures and mechanisms of existing integration systems with their advantages and disadvantages. Based on this analysis and user requirement studies, we propose an integration system framework that embraces advantages of the existing systems. More specifically, we propose an integration system architecture composed of a mediator and wrappers, which can offer a service interface layer for various other applications as well as independent biologists, thus playing the role of database management system for biology applications. In other words, the system can help abstract the heterogeneous information structures and formats from the application layer. In the system, the wrappers send database-specific queries and report the result to the mediator using XML. The proposed system could facilitate *in silico* knowledge discovery by allowing combination of numerous discrete biological information databases.**

I. INTRODUCTION

CURRENTLY, there are more than 500 public biological databases[1] each containing unique information serving different goals of biologists. Moreover, with the advent of high-throughput biology research techniques, the amount of available biological data is increasing exponentially. Even though many of these biological databases are independent, they are simply specialized information describing a same biological phenomenon by each component. In other words, these disconnected biological information sources are closely related to each other in a biological sense. In order to fully utilize the available biological information, biologists must search for similar related information from each data sources and integrate the search results manually [2] before being able to see the whole picture of the concept. However, the current trend of increasing number of data sources as well as the accumulating amount of data makes it very difficult for biologists to find the useful information and deduce new knowledge from them. For example, it would be time consuming to even find an appropriate database in the midst of ever increasing pool of biological databases. In order to address this type of a problem, the meta-database [3] that shows the information about the various databases has been implemented. However, it is still up to the biologists to find the most appropriate data source for his or her needs. In addition, each of biological databases is equipped with a distinctive interface and query format. Moreover, frequently the same biological term is used to refer to different things and vice versa [4]. As described, the increasing number of databases and the individuality of database formats among them are technical road blocks to integrate the search results of different data sources. Without the integration system however, it is very time consuming for biologists to manually query each data source to integrate, analyze and manipulate the acquired data. To tackle this kind of problem, a biological integration system that enables data search, analysis, management and further, new data formulation is an essential. In this paper, we suggest a methodology and guidance in development of a biological integration system.

II. OBSTACLES IN INTEGRATION SYSTEM DEVELOPMENT

A. *Variety of stored data*

There are numerous types of biological data sources that need to be integrated. To list a few of the most popular data sources and their types, there are databases that store experimental and research results such as GenBank [5] storing genetic sequences of all publicly available DNA, Swiss-Prot containing protein sequences, ParaDB[6] containing paralogy mapping information, YMD[7] containing microarray gene expression analysis, GenMapDB[8] containing mapped artificial chromosome information, PDB[9] containing experimentally determined three-dimensional structures of biological macromolecules, InterDom[10] containing putative protein domain interaction information, and lastly there are databases containing molecular interaction networks in biological processes (PATHWAY) like KEGG[11] and EcoCyc[12]. There is another class of biological databases that contain some form of annotations based on the aforementioned experiment

data sources. For example, there are classification information databases such as ProtoMap[13], genomic annotation databases such as GeneCards[14] and FlyBase[15], nucleotide polymorphism information database such as dbSNP[16], protein motif information database like Pfam[17], journal databases like Medline[18] and PubMed[19]. There are also a variety of public tools available used to analyze and manipulate biological data. These tools perform sequence analysis, gene prediction, research data clustering, protein folding and structure prediction, fold modeling, data mining, protein interaction and pathway simulation. Due to such variety of databases as well as tools, the system should be called Bio-Information Integration System instead of Bio-Database Integration System [20].

B. Heterogeneous data formats

Currently, a wide variety of data formats of biological databases and tools exist. With such a wide variety of data formats, the primary focus of Bio- Information Integration System needs to be to integrate such heterogeneous data formats in contrast to the conventional federated databases whose primary focus is to integrate the underlying schemas [21]. More specifically, Bio-Information Integration System must parse and integrate flat file format, HTML format, XML, object-oriented database query results, as well as tool-generated analysis results [22]. Although, some biological databases have been converted to relational databases, still the most conventional format of biological data collection is the flat file format.

C. Nested data structures

Due to the nature of biological data, a lot of the data is stored in a deeply nested data structure. This property along with the complex inner relationships among databases makes it very difficult to generate a global model that can accommodate all the biological data sources. In a few cases, the data models of the data sources are not publicly available [23]. What's worse, the fast-paced development of biological concepts frequently makes it inevitable not to modify the existing data schema. Lastly, there are many cases when specific database information is difficult to be expressed in an integrated schema. For example, the complex genetic networks or inter-molecular relations can not be easily expressed in a uniform format that can be used by various biological analysis tools.

III. EXISTING BIO-INFORMATION INTEGRATION SYSTEM

Currently, the most widely used bio-information integration system is Entrez [24] of NCBI (National Center for Biotechnology Information). Entrez integrates gene sequence database, protein sequence database, biomedical literature database, genome assembly and other useful databases by 'point-and-click navigation' [25] method, in which related databases are connected using hyperlinks. EBI (European Bioinformatics Institute) developed a bio-information integration system similar to Entrez called SRS (Sequence Retrieval System), where as NCGR (National Center for Genome Resources) offers a bio-information and bio-software integration system called ISYS [27] and IBM already introduced a federated database, DB2 Life Science Data Connect [28] based on their proprietary DB2 technology. There has been other integration approaches with various technologies as TINet (Target Informatics Net) [30] based on OPM (Object-Protocol Model) [29]. Others suggested converting the data into XML format and loading it into a relational database.

Various bio-information integration systems employee unique techniques and put their focuses on unique aspects of the systems. For example, Entrez [24] integrated their various in-house bio-databases successfully, where as SRS [26] focuses on effortless expandability. TAMBIS [31], on the other hand emphasizes on transparency of source data. We further describe aforementioned systems by focusing on the existence of their global schemas, query conversion techniques to each of the underlying data sources and query results integration techniques.

A. Entrez

Entrez is an integrated retrieval system for searching approximately twenty of the linked NCBI databases such as GenBank, PubMed, Nucleotide, Structure, OMIM and Domain. In particular, it can taxonomically search for DNA sequence or protein sequence using its Taxonomy database. The system integrates the underlying databases using the simple hyperlinks instead of relying on a global schema. Users can easily download the search results of their simple Boolean based query on their hard drives in various formats. Although, it does provide relevant online resources beyond the Entrez system using "LinkOut", the lack of integration effort for external systems is considered to be one of the biggest limitations of this system.

B. SRS

SRS started as sequence retrieval system for EMBL. The system doesn't have central schemas or data models. It relies on linking algorithm that stores the linking information among bio-information sources. The system can utilize this information in search index generation as well as cross-reference link index generation. Such indexes are stored in link indices table to accelerate the search speed.
When a user selects multiple databases, the available query fields get limited to only the common existing fields of the databases. SRS however offers users with customized installation capability, along with a scripting language called ICARUS, which can be used to effortlessly develop a new wrapper for the new integrated database. As a result, the system has integrated about 200 bio-information sources and tools as of August 2003 and it can be considered as an integration system with the greatest expandability.

C. *TAMBIS*

TAMBIS was developed to provide transparent access to disparate biological data sources. The distinctive feature of the system is its biological terminology (biological Concept Model) knowledge base and usage of biological ontology as an overall schema of the integration system. Ontology is a complete and complex organization of general biological concepts that could be used to accommodate any new biological information sources. It includes the available services of each data source and source combination model that instructs how ontology information can be applied for querying underlying data sources. Lastly, the system provides a flexible means of constructing and manipulating 'knowledge-driven' queries based on its ontology. However, generating a source model for each data source is labor-intensive and generation of a complete biological ontology is still on-going. Consequently, the expandability of the system is very limited; the number of integrated data sources is only a handful.

D. *Analysis Result*

From the analysis of currently existing databases, the most important requirements of the bio-information integration system are expressive query language and expandability of additional data sources. In addition to this, optimization of the query and intuitive application interface on would enable users to fully take advantage of features offered by the integration system. Next we will discuss the methodology of implementing the integration system using the mediator-wrapper system design.

IV. BIO-INFORMATION SOURCE INTEGRATION METHODOLOGY

A. *Overall system architecture*

The integration system can be classified as data warehousing [32] or virtual integration [25] depending on whether or not the system houses the source data locally. Locally storing the source data can be accessed fast and the more complicated queries can be accommodated. However, the integration system developers must implement the searching functions for each of source data by various techniques such as creating index files on flat file data. More importantly, the integration system with warehousing technique might sometimes provide the users with stale data. Virtual integration system, on the other hand will access the data source each time the query is entered, thus always providing users with the fresh and reliable data. However, the performance needs to be sacrificed for the higher degree of reliability because the system performance and expressiveness of the query could be directly limited by underlying data sources. In other words, the system cannot accommodate any types of new queries that the data source doesn't support.

We determined that the bio-information integration system should be able to access both virtual data and locally stored data to suit the needs of the system. The developers should first examine whether a particular data source provides necessary functionalities virtually. For instance, if PDB doesn't provide necessary methods through its web interface, it is required to locally store the database in an appropriate format, parse the database accordingly and implement the necessary search functions.

In order to adapt such a flexible design, the layers of abstraction among application layer, the mediator layer and wrapper layer need to be clearly defined and distinguished. The details of the each layer will be discussed below.

1) Application layer: There could be numerous different types of applications to suit various requirements of users. Applications must access the mediator through a standard query format to a mediator regardless of whether the application is running on the web server or as a stand-alone application. Although, the applications are technically not part of the integration system, it is the method biologists can interact with the integration system. Thus, it very important for application developers to perform a thorough user-requirement analysis among biologists.

2) Mediator layer: Mediator plays a role of database systems such as DB2 or mySQL, on which other applications be built. The mediator is largely divided into three separate components: query generator, global schema and the result processor. Since the applications would be interacting with the mediator through a query language mediator understands, it must be a carefully-tuned subset of an industrial-standard query language so that it has enough expressive power in order for applications to easily access all necessary information from the source without being superfluously complicated. One of the good candidates for such a query language is a subset XQuery because it can easily access deeply nested structure of biological data. Query should be expressive enough to provide a transparent and flexible access to the system. In other words, the structures and formats of underlying data sources should be hidden to applications. Specifically, the mediator should provide an illusion that all underlying data sources are stored in XML, which can be accessed using a subset of XQuery. An application should also be able to exactly identify which databases it wants to access to answer its query instead of relying on mediator's intelligence.

There are different approaches for mediator to grasp the semantic meanings of the input query and generate the separate queries for data sources. Using the biological global ontology enables powerful query processing capability. There are several existing standard biological ontology such as GO (Gene Ontology) [34], TaO (Tambis Ontology) [31], Signal Ontology [35] and IMGT [36]. But, developing and using ontology is a daunting task and it would impede the expandability of the system. Thus, most of the existing integration systems are not employing the ontology. However, in order for the system to accommodate a powerful query, it needs to collect the semantic correspondences between underlying data sources as well as the attributes of them. We suggest the data sources be classified according to the types of data their attributes contain. Using this information, the system can generate the ontology for 'data source attributes' instead of ontology for biological concepts.

'Data source attributes' ontology service can be used in query processing in order to accurately determine which data sources as well as which fields or attributes of them should be accessed in order to answer such queries. In addition, with some additional information about the data sources and its performance, query optimization could also be performed in the mediator. Some of the other suggested standard data modeling methodologies for heterogeneous data sources include OMG [37] and STEP [38].
Once the appropriate data sources were realized and queries were generated using the semantic mapping information, mediator can distribute the queries to the specified database wrappers and result processor can collect and display the results. Result processing stage can manipulate search results from data sources as needed such as performing join operations.

3) Wrapper layer: Wrappers hide the heterogeneity of data sources from the mediator. The mediator sends a calculated query to a wrapper without having to know which data format the data is stored in such as a flat file, XML file or binary file format. In fact, the mediator does not be concerned whether the data is stored locally or need to be accessed through HTTP. In order to achieve such abstraction, the standard of wrapper-mediator communication should be defined precisely. To maximize the expandability, a wrapper development tool kit should be offered.

V. Conclusion

In order to overcome the flood of biological information and effectively fetch the targeted data from various data bases and biological tools, an integration system is a requirement for biologists. Here, we analyzed various obstacles that make integration a difficult task and properties of the existing integration systems along with their merits and demerits. Based on the analysis, we suggested the design requirements and methodology of bio-information integration system by each component. An integration system should support powerful yet not excessively complicated queries and easier expandability with the clear distinctions between three layers of abstraction, namely the application layer, mediator layer and wrapper layer. The most important component, or the core of the system is going to be the design of reliable and thorough global schema in order to produce more effective and efficient queries.

REFERENCES

[1] Andreas D. Baxevanis (2003) "The Molecular Biology Database Collection 2003 update" Nucleic Acids Research, 2003, Vol. 31, No. 1

[2] P. Baker, C.Goble et al. (1999) "An Ontology for Bioinformatics Application" Bioinformatics, Vol. 15, No. 6, 510~520

[3] C.Discala, X.Benigni et al. (2000) "DBcat : a Catalog of 500 Biological Databases" Nucleic Acids Research, Vol. 28, No. 1, 8-9

[4] Leser, U., H.Hehrach, et al. (1998) "Issues in Developing Integrated Genomic Databases and Application to the Human X Chromosome" Bioinformatics 14(7) : 583~690

[5] D. Benson et al. (2000) "GenBank" Nucleic Acids Research, Vol 28, No. 1 15-18

[6] T.K.Jenssen et al (2001) "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression" Nature Genetics, Vol. 28, No 1, 21-28

[7] Kei-Hoi Cheung et al. "YMD: A microarray database for large-scale gene expression analysis"

[8] http://genomics.med.upenn.edu/genmapdb

[9] H. M Berman et al. (2000) "The protein data bank" Nucleic Acids Research, Vol 28, 235-242

[10] http://InterDom.lit.org.sg

[11] M. Kanehisa et al. (1999) "KEGG: Kyoto encyclopedia of genes and genomes" Nucleic Acids Research, Vol. 27, 29-34

[12] P. Karp et al. (1999) "EcoCyc: Encyclopedia of the escherichia coli genes and metabolism" Nucleic Acids Research, Vol. 27(1), 55-58

[13] Yona et al. (2000) "ProtoMap: automatic classification of protein sequences and hierarchy of protein families" Nucleic Acids Research, Vol. 28, 49-55

[14] Rebhan, M et al. (1997) "GeneCards:encyclopedia for genes, proteins and disease" Weizmann Institute of Science, Bioinformatics Unit and Genome Center

[15] (1998) "FlyBase" (1998) Oxford University Press 85-88 Nucleic Acids Research, 1998, Vol. 26

[16] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp

[17] http://www.sanger.ac.uk/Software/Pfam

[18] http://medlineplus.nlm.nih.gov/medlineplus/

[19]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=PubMed

[20] Ulf Leser. (1999) "Designing a Global Information Resource for Molecular Biology" In 8th GI Fachtagung: Datenbanksysteme in Buero, Technik und Wissenschaft, Freiburg, Germany

[21] A. P. Sheth et al. (1990) "Federated Database Systems for Managing Distributed, Heterogeneous and Automated Databases" ACM Computing Surveys, Vol. 22, No. 3, 183-196

[22] Francois Bry et al. "A Molecular Biology Database Digest"

[23] Barbara A. Eckman et al (2001). "Optimized Seamless Integration of Biomolecular Data" International Conference on Bioinformatics and Biomedical Egineering, 23-32

[24] Entrez Online Documentation

http://www.ncbi.nlm.nih.gov/Database/index.html
[25] P.Karp (1995) "A Strategy for Database Interoperation" Journal of Computational Biology, Vol. 2, No. 4, 573-586
[26] Etzold, T., A.Ulyanov, et al. (1996) "SRS:Information Retrieval System for Molecular Biology Data Banks." Methods in Enzymology 266 : 114-128
[27] Siepe et al. (2001) "ISYS : A decentralized, Component based Approach to the Integration of Heterogeneous Bioinformatics Resources" Bioinformatics, Vol. 17, No. 1, 83-94
[28] L. M. Haas et al. (2001) "DiscoveryLink : A system for Integrated Access to Life Sciences Data Sources" IBM Systems Journal, Vol 40, No. 2, 532-551
[29] II-Min A. Chen et al(1995) "An overview of the object protocol model(OPM) and the OPM data management tools"
[30] Barbara A. Eckman et al (2001) "Extending traditional query-based integration approaches for functional characterization of post-genomic data" Bioinformatics Vol 17, no 7, 587-601
[31] P.Baker et al. (1999) "Tambis - Transparent Access to Multiple Bioinformatics Information Sources" ISMB' 98, pp. 25-34
[32] Y. Zhuge et al. (1995)"View maintenance in a warehousing environment" In Proc. ACM SIGMOD, 316-327
[33] Wiederhold, G. (1992). "Mediators in the Architecture of Future Information Systems" IEEE Computer 25(3) : 38~49
[34] http://www.geneontology.org/
[35] http://ontology.ims.u-tokyo.ac.jp/signalontology/
[36] http://imgt.cines.fr:8104/
[37] http://www.omg.org
[38] http://www.nist.gov/sc4/www/stepdocs.htm