Database Integration Approaches for Heterogeneous Biological Data Sources: An overview

Iskandar Ishak, Naomie Salim
Faculty of Computer Science and
Information System
University Technology of Malaysia
Email: iskandarishak@gmail.com, naomie@fsksm.utm.my

ABSTRACT

Biological data sources are known for its heterogeneous in many aspects. These aspects include data formats, physical location as well as its query capabilities. These data sources need to be integrated so that researchers can easily access it, query it and get results. Approaches like link-driven, data warehousing and mediator-based are being used to integrate these data sources. This paper tries to look at the dissimilarities of major biological data sources and some major approaches in dealing with the biological multi-database integration. Mediator based with ontologies approach is also discussed in this paper. Proposed work on merging biology ontologies in a mediated system is also presented.

KEYWORDS

Integration, biological database, mediator, ontology.

1. INTRODUCTION

Biological database like SwissProt, Genbank, EMBL, DDBJ, or Enzyme [8] store different type of biological data formats, located at different locations, and have different user interfaces. Integration is needed to put together this entire database to make it looks as a single database. There are several approaches that are being used to integrate multiple databases such as SRS[14, 4], TAMBIS[10, 2], BACIIS[9], Kleisli[1] and DiscoveryLink[3]. Despite the effort within these approaches, still, there is no single approach that complies to all the needs in bioinformatics [14].

2. BIOLOGICAL DATA SOURCES

Table 1 shows the heterogeneity of data sources available for biologist use for their research. The indifferences between the data sources are shown in their data format and user interfaces. These data sources store information about nucleotide sequences, protein sequences, 3D macromolecular structures and protein families. These data usually can be retrieved via web interface, ftp and e-mail. The underlying data model for these data sources are the flat file model, relational model and object-relational model [8].

According to Lambrix and Jakoniene (2000), most of the databanks allow for queries based on the occurrence of text within a data item (full-text search) and all the databanks support queries based on occurrence of a text string within certain predefined fields. User of these databanks is often guided by the retrieval interface of the systems, plus, it also supports command-line querying using the systems' query language. Most system supports the use of Boolean queries (using and, or, not).

There are some problems in using these different data sources. In order for users to query a data source they need to have some knowledge on the data source that they want to query. The users also need to know the query language, as well as user interfaces of the system. Users might need to query more than one data source, and to learn each one of the data source that they want to query could be a tedious job. This could also take a lot of users' time to get some knowledge on the data source that they want to use.

3. BIOLOGICAL DATABASE INTEGRATION APPROACHES

Researchers have come out with some approaches that integrate these diverse biological data sources. There are 3 integration approaches being used in addressing the issue of interoperability among biological database: navigational or link-based integration, mediator-based [4], and data warehousing [9, 1, 8, 4].

databank	data type	categories	references
Gen Bank	Genbank flat file, asn.1	Neuclotide-sequences	[8]
EMBL	Embl flat file	Neuclotide-sequences	[8]
DDBJ	Ddbj flat file	Neuclotide-sequences	[8]
Swiss-Prot	Swissprot flat file, fasta	Protein sequence	[8]
PIR	nbfr-pir, codata, fasta, xml and Oracle based-data model	Protein sequence	[13], [8]
ENZYME	enzyme flat file, asn.1	Protein sequence	[8]
PDB	pdb flat-file, mmCIF	3D macromolecular structures	[8]
MMDB	asn.1	3D macromolecular structures	[8]
PROSITE	prosite flat file	Protein families	[8]
PRINTS	Prints flat file	Protein families	[8]
BLOCKS	Blocks flat file	Protein families	[8]
Medline	Text-file	Literature	[13]

TABLE 1: Heterogeneity of biological data sources.

3.1 Navigational Approach

Navigational approach is an approach where the system will provide static links between data or records in different data sources. SRS [14, 9] provides some functionality to search across public, in house and in-licensed database.

SRS basically parses flat files or databanks that contain structured text with field names. It then creates and stores an index for each field and uses these local indexes at query-time to retrieve relevant entries. SRS parses the file or databanks and capture all the information, it uses its own parsing component known as ICARUS (a special built-in wrapper programming language [14]).

The results of this approach would be simple aggregation of records that matched the search constraint. These records can contain links that the user can follow to obtain more information about the results.

3.2 Data Warehouse Approach

Data warehousing approach is adopted by integration system like GUS [1] and DiscoveryLink [14]. This approach uses data warehouse repository that provides a single access point to a collection of data, obtained from a set of distributed, heterogeneous sources.

Data from the remote heterogeneous database are copied on a local server and the user will use a unique interface within the system to allow multi-database queries to be issued to this single interface.

3.3 Mediator approach

Another approach in biological database integration is the mediator approach. System like DiscoveryLink [14, 4, 3], K2/Kleisli [1], TAMBIS [10, 9, 2], and BACIIS [9] uses this approach. Mediator based approach does not store any data, but it provides a virtual view of the integrated sources [6].

Mediator approach basically translate query from the user, into query that is understood by the sources integrated into the system. It maps the relationship between source descriptions and the mediator and thus allows queries on the mediator to be translated to queries on the data source.

4. DISCUSSIONS

Navigational approach is used in systems like SRS [14, 9] and Entrez [8]. It is popular because of its easy to use feature and it only involves the tasks of point and click. This approach also allows the representation of cases where the page containing the desired information is only reachable through particular navigation path across other pages [4].

One of the weaknesses of information linkage approach is user must specify which data sources should be used to answer a given query. Another weakness of this approach is when a user is interested in a join between two data sources in the system, the user must manually perform the join by clicking on each entry in the first data source and following all connections to the second data source [1]. This approach does not actually integrate data sources but it only gives users a mean to retrieve information.

Data warehouse approach is somewhat different than the two other approaches. It involves large storage to copy the data from the data source involved in the system and it uses high-level query language like the SQL for querying the system.

Main advantage of data warehouse approach is, system performance tends to be much better. It is because query optimization can be performed locally, and inter-data source communication latency is eliminated. This approach is also reliable because there are fewer dependencies on network connectivity. The most important advantage of using the data warehousing approach is, since underlying data sources may contains errors, it keeps a separate copy of data called cleansed copy.

Data warehouse approach does have its tradeoffs. Results reliability and overall system maintenance are questionable as there are possibilities of returning outdated results [4]. Changes in data sources does not mean the data in the warehouse will also changed, so there is a need of detecting changes in data sources as well as automating the update of the data warehouse.

Mediator based approach is based on translating queries from the user to the one understood by the data sources. This approach does fit the description of integrating heterogeneous although it's only in a matter of database view perspective.

Mediator based approach has 2 approaches; Global-as-view(GAV) and Local-as-view(LAV) [4]. In GAV approach, mediator is based on the source relation schema and it facilitates great query reformulation While in LAV approach, source relation is based on the mediator's schema and relation. GAV approach has its drawback in adding or removing sources as it will involve modification of the mediator schema. Unlike GAV, adding or removing sources is much simpler. However LAV makes the query reformulation complicated.

Apart from the drawbacks, mediator based approach have more strong points that the other two approaches. Compared to navigational approach, mediator based is much more advanced because it involves retrieving knowledge from the data source rather than giving out static links. Data warehouse approach boast with its no network involvement and local server based, and this give advantage in terms of performance, no bottleneck or non availability of services. However, data source update in data warehouse that took so much time is such an uncompromised weakness. Mediator based would not have the update problem as the query directly goes to the original source. Mediator based can be looked as a cheaper and effective approach since it involves schema or view integration, rather than to have huge storage to store copied data from all the involved data sources.

In a general point of view, mediator based approach can overcome data source heterogeneity problem by using metadata form in a form of vocabularies or ontology to represent domain knowledge explicitly [11]. In fact ontology has already been used in a mediator based systems called Transparent Access to Multiple Biological Information Sources or TAMBIS [10].

System like TAMBIS uses ontology that addresses the semantic aspect of heterogeneous data sources. TAMBIS has its own ontology called TaO and contains nearly 2000 concepts that describe both molecular biology and bioinformatics tasks. Interface in TAMBIS helps users browsing the ontology for constructing queries. According to Wong, query in TAMBIS is formulated starting from one concept, then browsing the connected concepts and applicable bioinformatics in the ontology [14].

Ontology [12] would make an integration system much easier to be queried. It is easier in a sense of making a logic and sensible query in biological data domain. Ontology makes the mediator approach much more useful than the other integration approaches especially in ad-hoc query.

However, despite of the strength of ontology, there are also things that have to be looked into it in implementing it in mediator based. Ontology is quite subjective and there is more than one way to represent a domain using ontology. Therefore, there is a need for more ontology, so that the users can have choices of ontology for the domain that the user required, to help them query the system.

Ontology that is already stored in a system might not describe well on a specific thing. Alternative ontologies are needed to help the user to give their view from different perspective. Feature that allow user to add new ontology should be fit into the mediator based system. Through this approach, new and better ontology might come into the system and hence give further help for user to query the system.

Additional or alternative ontology also needed for extra exploratory data analysis. User should also be able to choose the ontology that they desire to in order to query the data sources.

5. SUGGESTIONS FOR FURTHER WORK

Incorporating user defined biological ontology in mediator based system is proposed in [5], in which it help users to query the data sources in different abstract level and different contexts. Users can also adapt existing ontology that is already in the system, or using existing external ontology like Gene Ontology, RiboWeb Ontology, TAMBIS Ontology, EcoCyc Ontology and Schulze-Kremer Ontology.

So, there is a need for combining ontologies in a database integration approaches. Further work will focus on the merging different ontologies to help user ask queries. Steps that will be taken in merging ontologies are:

1) Ontology pre-processing
2) Selection of Concept
3) Similarity computation
4) Reconstruction of hierarchy

The above approach is implemented in [7] using ontologies based on WordNet. However, in our situation, biological ontologies will be used. Prior to this, user defined ontology will be build using an ontology building tools.

The reconstructed ontologies will be stored in a mediator and will be use in assisting user to create query. The query then will be translated by mediator and send to the incorporated data sources. In this research, data sources incorporated will be the data sources like SWISS-PROT and Genbank. The result using the merge ontology will be compared with the result of using the ontologies individually. As a start, level of abstraction that will be focused on is the biological taxonomy level.

6. CONCLUSION

This paper begins with the description of the problem of heterogeneous biological data sources. Then it describes the approaches in integrating the data sources. 3 approaches have been presented and mediator based has the edge over other approaches in integrating biological data sources. Mediator based approach also offer the inclusion of ontology based in assisting user using integration system using this approach.

Ontology or specifically biological ontology could help user in building useful queries on mediator system. Alternative ontology should also be included, which means, user can add ontologies to the system. Combination between biological ontologies in a mediator system required further research. This will make query formulation in heterogeneous data source using the ontology much more useful.

7. REFERENCES

[1] S. B. Davidson, J. Crabtree, B. Brunk, J. Schug, V. Tannen, C. Overton and C. Toeckert, *K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources*, IBM System Journal. Deep Computing for the Life Sciences, 40 (2001).

[2] C. Goble, *Supporting Web based Biology with Ontologies*, IEEE (2000).

[3] L. M. Haas, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice and W. C. Swope, *DiscoveryLink: A system for integrated access to life science data sources*, IBM System Journal, 40 (2001).

[4] T. Hernandez and S. Kambhampati, *Integration of Biological Sources: Current Systems and Challenges Ahead*, SIGMOD Record, 33 (2004).

[5] V. Honavar, C. Andorf, D. Caragea, A. Silvescu, J. Reinoso-Castillo and D. Dobbs, *Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogenous, Distributed, Autonomous Biological Data Sources*.

[6] M. Kazemian, B. Moshiri, H. Nikhbakh and C. Lucas, *Architecture for Biological Database Integration*, Artificial Intelligence and Machine

Learning 2005 Conference (AIML 05) (2005).

[7] H. Kong, M. Hwang and P. Kim, *A New Methodology for Merging the Heterogeneous Domain Ontologies based on the WordNet*, *Proceedings of the International Conference on Next Generation Web Services Practices (NWesP'05)*, IEEE, 2005.

[8] P. Lambrix and V. Jakoniene, *Towards transparent access to multiple biological databanks*, Proceedings of the 1st Asia-Pacific Bioinformatics Conference (2000).

[9] Z. B. Miled, N. Li, G. L. Kellet, B. Sipesand and O. Bukhres, *Complex Life Science Multidatabase Queries*, (2002).

[10] N. W. Paton, R. Stevens, P. Baker, C. A. Goble, S. Bechhofer and A. Brass, *Query Processing in the TAMBIS Bioinformatics Source Integration System*, Scientific and Statistical Database Management (1999), pp. 138 – 147.

[11] K. U. Sattler, I. Geist and E. Schallehn, *Concept-based querying in mediator systems*, VLDB Journal (2005), 14 (2004), pp. 97-111.

[12] R. Stevens, C. A. Goble and S. Bechofer, *Ontology-based Knowledge Representation for Bioinformatics*, (2000).

[13] A.-P. Tsou, Y.-M. Sun, Chia-Lin, Liu, H.-D. Huang, J.-T. Horng and M.-F. Tsai, *A Biological Data Warehousing System for Identifying Transcriptional Regulatory Sites from Gene Expressions of Microarray Data*, (2005).

[14] L. Wong, *Technologies for Integrating Biological*, Briefing in Bioinformatics, 3 (2002), pp. 389 - 404.