

Customized Biological Database Integration System for cDNA microarray

Myungguen Chung¹, Myungeun Lim¹, Myungnam Bae¹, Sunhee Park¹

Keywords: biological database integration, wrapper, mediator, wrapper description language

1 Introduction

The integration of biological data is just one phase of the entire molecular biology research and genomic hypothesis discovery process. The researcher will find it difficult to image conducting their work without access to one or more of biological database. Then, biologists need to access an integrated view of remote or local heterogeneous data sources with advanced data accessing, analyzing, and visualization tools. For decade, several integration methods were proposed and integration products were developed already. However there are many challenges of the usage in the future

2 Method

We developed the database integration of biological and genomic source, which has recently become a major focus of the customized data integration. The most flexible data integration designs adopt a mediator approach that introduces an intermediate processing layer to decouple the underlying heterogeneous distributed data source and the client layer of end user and application. The mediator is a collection of software component performing the task of data integration. The mediator layer performs the core function of data transformation and integration and communicates with the wrappers and the user application layer. In other words, a mediator in the information integration context is a control system that is responsible for reformulating at runtime a query given by user on a single mediated schema into a local schema of the underlying data sources. Most database mediator system use a wrapper layer to handle the task of data access, data retrieval, and data translation. The Wrapper access specific data sources, extract selected data, and translate source data formats into a comma data model designated for the integration system. We also present an approach to wrapping web data sources, databases, flat files, or data generated by tools. Generally, a wrapper has two tasks: it sends a query to source to retrieve data and, builds the expected output with respect to the virtual structure. In system level, each of wrappers is made from Wrapper Description Language (WDL) which is developed to easily design for biologist.

We also made a demonstration system for our collaboration with DNA-chip. Consider this scenario; one molecular biologist wants to know annotation and other additional information of his/him DNA-chip on the laboratory. First, the researcher goes to the 'Genbank' web site at NCBI to get an annotation and reference information of spot on the DNA-chip. Next, 'Entrez Gene' is visited to get whole genomic information in review. Then, the researcher goes to the GO website to look up the GO term that is related to his/her experimental purpose. Finally, the researcher downloads the entire list of protein that is related to his/her purpose from the GO web site. This scenario needs 4 wrappers (Genbank, Gene, GO, Uniprot). User query are broke down into local query of the underlying data sources on runtime by mediator. The above fragment query is passing into the wrapper which will query to various resources on its schedule and record the result from

¹ Bioinformatics Research Team, Electronics and Telecommunication Research Institute, 161, Gajeong-dong, Daejeon, 305-350, Korea, E-mail: {aobo, melim,, mnbase, shp}@etri.re.kr

each of the sources. Results finally are transforming into virtual data structure and visualized user friendly.

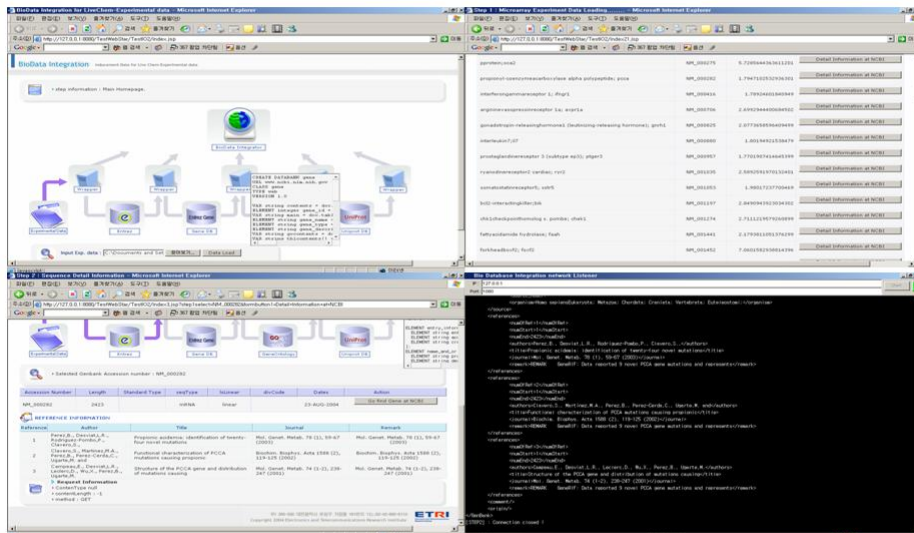


Figure 1: The overview of data integration system for LiveChem®

3 Discussions

Relational Database Management System (RDBMS) based method is effective methods for building and analyzing biological data now. Integration data from multiple resources also raise challenging issue related to data provenance, data ownership, data quality, privacy, and security, which will need to be addressed in the short future. In the near future, the more complicated query and user demand are revealing. Then, the mediator-wrapper based method is going to replace an essential place.

References

[1] Lincoln Stein, Integrating biological databases, Nature review genetics, 4: pp 337-346, 2003
 [2] Zoe Lacroix, Bioinformatics : Managing Scientific Data, MorGan Kaufmann, 2004