ABSTRACT

Between technological breakthroughs and new computational approaches, the quantity of biological data is increasing explosively. As of 2007, there were 1078 biological databases. Providing biologists with central and uniform access to all types of data stored in biological databases is becoming critical. To minimize disruption of current operations, maintain local autonomy and handle heterogeneities, federated databases and Web services have been proposed as viable solutions. This paper explores these issues and reports on our experience with testing multiple approaches for biological database integration. It discusses the trade-offs among performance, support for heterogeneity, robustness and scalability. A significant result of our study is that the most flexible approach, Web Services, performs very competitively. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: Bioinformatics; Distributed Databases; Data Integration; Federated Databases; Web Services,

INTRODUCTION

The quantity of biological data is increasing explosively; most of these data are stored in databases, including the International Nucleotide Sequence Database Collaboration (DDBJ, EMBL and GenBank) and numerous other specialized databases, such as PROSITE, EC-
ENZYME, GDB, Reactome, UniProt, PIR, DIP, Pfam, and PDB. At the end of 2007, there were 1078 biological databases (Galperin, 2008). This tremendous diversity of biological data greatly improves biologists’ ability to study the interactions between the components of a biological system and how these interactions give rise to the function and behavior of that system. Such studies need to access multiple types of data, which are likely to be stored in different, geographically distributed, databases. However, providing biologists with central, uniform access to all types of data is not a trivial task. It would be a poor solution, if not impossible, to make a single database to include all biological data (Stein, 2003). To provide better query facilities and expedite the research process in an automatic way, data integration is essential, and is one of the most important bioinformatics research areas (Stevens et al., 2001). There are three main layers of data integration. The bottom layer is the database and database management system (DBMS), in the form of distributed databases, multi-databases, data warehouses or federated databases. The middle layer is software that supports the distributed applications and includes Service-Oriented Architecture (SOA), Common Object Request Broker Architecture (CORBA), Distributed Component Object Model (DCOM) and Java Remote Method Invocation (RMI). The upper layer puts responsibility for integration into the application and includes Link Integration or Query-Based Integration (Stein, 2003; Karp, 1995). Based on Peter Karp’s four approaches (Hypertext Navigation, Data Warehouse, Unmediated MultiDB Queries and Federated Databases) (Karp, 1995) and Lincoln Stein’s three approaches (Link integration/Web Services, View integration, Data Warehousing) (Stein 2003), one could argue that there are five basic data integration approaches in use: Link Integration, Query-Based Integration, Data Warehouse Integration, Federated Database Integration and Web Service Integration (Stein, 2003; Karp, 1995), as briefly described below. Among them, Federated Database and Web Services are prominent technologies because each minimizes the disruption of current operations, maintains local autonomy, handles heterogeneities, and is scalable.

- **Link Integration** means that end users get comprehensive and relevant information through hyperlinks from the first data source that the end user begins to search. Although very successful, this approach faces problems such as the vulnerability of naming clashes, ambiguities, and the instability of volatile hyperlinks maintained by different hosts (Stein, 2003). The Sequence Retrieval System (SRS) (Zdobnov et al., 2002) is a variation of the Link integration approach. It is designed to retrieve information stored in multiple hosts using the language ODD (Etzold and Argos, 1993). It stores the data indexes in a global repository, while leaving the actual data in their own repositories. The disadvantage is that integration and data location are not transparent to end users (Stein, 2003; Karasavvas et al., 2004; Etzold and Argos, 1993).

- **Query-Based Integration** means that end users can retrieve information from multiple repositories, which can be databases or flat files, through a single query. In this approach, the query is often not SQL. The main problems with this approach are the complexity of queries, that a new query language or language extension must be learned, and that integration and data location are not transparent to end users. For example, in BioKleisli (Davidson et al., 1997) using a query language called “Collection Programming Language (CPL)” (Peter et al., 1994), end users need to manually choose the database and specify how to use it. Other examples of query-based integration include Object-Protocol Model (OPM) (Chen and Markowitz, 1995; Markowitz et al., 1999), which uses OPM*QL as the query language, and P/FDM (Kemp, 2002), which uses Daplex as the query language.
Related Content

Factors Motivating the Acceptance of New Information and Communication Technologies in UK Healthcare: A Test of Three Models
www.irma-international.org/chapter/factors-motivating-acceptance-new-information/22719/

The Implementation Puzzle of CRM Systems in Knowledge Based Organizations
www.irma-international.org/article/implementation-puzzle-crm-systems-knowledge/1243/

Wheelchairs Embedded Control System Design for Secure Navigation with RF Signal Triangulation
www.irma-international.org/article/wheelchairs-embedded-control-system-design-for-secure-navigation-with-rf-signal-triangulation/86273/

Managing Information Security on a Shoestring Budget
www.irma-international.org/article/managing-information-security-shoestring-budget/44539/

Organizational Innovation Strategy
www.irma-international.org/chapter/organizational-innovation-strategy/28130/