Chapter 2.18
BACIIS:
Biological and Chemical Information Integration System

Zina Ben Miled
Indiana University Purdue University, USA

Nianhua Li
Indiana University Purdue University, USA

Omran Bukhres
Indiana University Purdue University, USA

ABSTRACT

Life science Web databases are becoming increasingly essential in conducting everyday biological research. With more than 300 life science Web databases and the growing size of the life science data, searching and managing these complex data requires technology beyond that of traditional database systems. The open research issues include the interoperability of geographically distributed autonomous databases, which are generally only Web-accessible, and the seamless semantic-based integration of these databases with total transparency to the user. In this paper, the implementation of a Biological and Chemical Information Integration System (BACIIS) is presented. BACIIS supports the integration of multiple heterogeneous life science Web databases and allows the execution of global applications that extend beyond the boundaries of individual databases. This paper discusses the architecture of BACIIS. It also discusses the techniques used to extract and integrate data from the various life science Web databases.

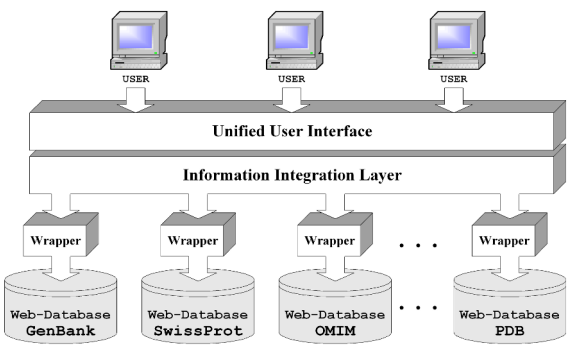
INTRODUCTION

The confederation of geographically distributed life science databases is a critical challenge facing biological and biomedical research. Databases are the intermediaries between experimental observation and the ability to extract biological knowledge. The challenge in managing life science data is due

to the large volume of data, the rate at which the data are increasing in size, and the heterogeneity and complexity of the data format. There are currently more than 300 life science Web databases (Baxevanis, 2002) providing access to scientific data and literature via the Web. These databases use different nomenclatures, file formats, and data access interfaces. Furthermore, they may include redundant and conflicting data.

This paper presents a system (BACIIS) that uses the mediator-wrapper approach to support the integration of life science Web databases. BACIIS constructs a tightly coupled federation of databases and uses ontology as a global schema. Currently, BACIIS (Figure 1) integrates seven life science Web databases: GenBank (Benson, 1998), SWISS-PROT (Bairoch, 1998), PIR (Baker, 1998), PROSITE (Sigrist, 2002), ENZYME (Bairoch, 2002), PDB (Berman, 2000), and OMIM (McKusick, 1998). GenBank is an annotated collection of all publicly available DNA sequences, which includes approximately 20 million DNA sequences. SWISS-PROT is an annotated, nonredundant protein sequence database that includes high-quality annotation of proteins, such as the description of the function of a protein, its domain structure, post-translational modifications, and variants. PIR is a protein sequence and structure database. ENZYME is a database that contains information relative to the nomenclature of enzymes. PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns, and profiles that help to identify to which known protein family a new sequence belongs. PDB is a database of 3-D biological macromolecular structure data. The OMIM database is a catalog of human genes and genetic disorders, which contains textual information, pictures, and reference information.

Figure 1. Information integration for life science Web database



This paper specifically shows how a mediator-wrapper approach can be successfully used to integrate Web databases in the context of the restrictions imposed by the life science domain. These restrictions include:

- The databases are autonomous and only Web accessible. Therefore, data extraction has to be performed following the data presentation protocol that is adopted by the component databases.
- Access to the databases can change often; different databases return their result in a different format, and there is a large number of life science databases. Therefore, any data extraction mechanism needs to take these aspects into consideration in such a way that it easily can adapt to changes in the databases, and databases can be added to the integration system easily.
- Resolving heterogeneity is an important aspect in the integration of life science databases and has to be addressed with minimal user intervention in order for the integration system to be useful to the scientists.

BACIIS is a tightly coupled federation of life science Web databases. Its architecture, shown in Figure 2, is a three-tier software architecture (Kossmann, 2000) that consists of a mediator, wrappers (Ambite, 1998; Levy, 1998; Wiederhold, 1992), and an ontology that is used as a global schema. The mediator transforms data from its format in the component database to the internal format used by the integration system (Ambite,