Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 17–29
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

An agent- and ontology-based system for integrating public gene, protein, and disease databases

R. Alonso-Calvo [a], V. Maojo [a,*], H. Billhardt [b], F. Martin-Sanchez [c], M. García-Remesal [a], D. Pérez-Rey [a]

[a] *Biomedical Informatics Group, Artificial Intelligence Laboratory, School of Computer Science, Universidad Politecnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain*
[b] *Rey Juan Carlos University, Madrid, Spain*
[c] *Medical Bioinformatics Department, Institute of Health Carlos III, Majadahonda, Madrid, Spain*

Received 29 November 2005
Available online 20 March 2006

Abstract

In this paper, we describe OntoFusion, a database integration system. This system has been designed to provide unified access to multiple, heterogeneous biological and medical data sources that are publicly available over Internet. Many of these databases do not offer a direct connection, and inquiries must be made via Web forms, returning results as HTML pages. A special module in the OntoFusion system is needed to integrate these public 'Web-based' databases. Domain ontologies are used to do this and provide database mapping and unification. We have used the system to integrate seven significant and widely used public biomedical databases: OMIM, PubMed, Enzyme, Prosite and Prosite documentation, PDB, SNP, and InterPro. A case study is detailed in depth, showing system performance. We analyze the system's architecture and methods and discuss its use as a tool for biomedical researchers.

Keywords: Bioinformatics; Biomedical informatics; Heterogeneous databases; Data integration; Genomic databases

1. Introduction

At the time of writing this paper, more than 700 biological databases (DBs) were publicly available [1]. These databases are the result of a large number of biological research projects that have produced a huge amount of heterogeneous information about genes, proteins and genetic diseases—e.g., nucleotide polymorphisms, gene mutations, protein sequences and structures, and others. Public DBs are maintained by different institutions and research centers that collect these biological data. Often, different public DBs include related data types—e.g., Prosite, Swiss-Prot, and PDB store information related to proteins. In other cases, different organizations store their own information—e.g., gene polymorphisms and mutations DBs—but this disparate information is not integrated. In this regard, there is now a need and challenge to integrate information from Web-based public DBs and other private, local DBs for efficient use in biomedical research. Whether the publicly available information is integrated or not will have a significant impact on future clinical applications of genomic research.

One of the barriers to the integration of biological and medical databases is that they are designed and maintained differently by different organizations, such as the US National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and others. Furthermore, not all of the data sources are directly available. Many of them cannot be transparently accessed—as it is the case, for example, of databases stored at local database management system (DBMS). Instead, these remote sources are

* Corresponding author.
E-mail address: vmaojo@infomed.dia.fi.upm.es (V. Maojo).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.02.014

usually accessed by querying their DBs through Web-based interfaces—e.g., HTML forms—. In this paper, we refer to these remote sources as "public Web-based DBs."

In this paper, we present the use of the OntoFusion system for integrating public Web-based biological and disease-related information databases. OntoFusion is a system for integrating databases that are either publicly available on the Internet or are directly accessible through DBMS. It uses a multiagent-based architecture and its integration approach is founded on the use of ontologies. OntoFusion has been developed within the INFOGENMED project, with funding from the European Commission [2]. This finished project aimed to create tools to allow transparent and integrated access to biomedical information sources. The rest of the paper is organized as follows. Section 2 gives background on approaches to database integration and extracting information from public Web-based DBs in the biomedical field. In Section 3, we present our approaches to database integration and to the problem of accessing Web-based DBs. In Section 4, we provide a general overview of the OntoFusion system, and describe the part related to the integration of public Web-based DBs in more detail. Section 5 presents a case study, where OntoFusion is used to integrate seven different public Web-based DBs containing biomedical data. Finally, Section 6 provides some discussion and Section 7 concludes the paper.

2. Background

2.1. DB integration approaches

An earlier report [3] suggested that three different approaches to DB integration should be considered: information linkage, data translation, and query translation.

The first approach, information linkage, establishes relationships among different data sources by using cross-references. It facilitates the navigation over different data sources. The main drawback of this approach is that it does not actually integrate the information. This approach is used in many public biological DBs, like MEDLINE, PDB, Prosite, and others.

The data translation approach intends to create a central data repository containing all the data from the databases that are integrated. Data from different sources are translated to a unified conceptual schema and stored at the central repository. Queries are launched to this repository. This approach can be seen as the creation of a central data warehouse for multiple DBs. Its advantage is that it provides efficient and transparent access to the data. However, the effort needed to maintain such a data warehouse is considerable. Furthermore, any changes to the structure of the integrated DBs may require changes to the unified conceptual model.

The query translation approach does not maintain a central data repository. Instead, it divides the user queries into different sub-queries—one for each DB within the system. Then, mediators or wrappers execute the sub-queries in the respective DBs. Results from different DBs are gathered and returned to the user. Depending on the data conceptualization model used, four different categories can be identified: (i) pure mediation, (ii) single conceptual schema, (iii) multiple conceptual schemas, and (iv) hybrid approach. We have analyzed a number of systems and placed them into these categories.

Systems based on pure mediation employ wrappers and mediators to execute user queries (e.g., TSIMMIS [4], DISCO, DIOM, HERMES, Bio Kleisli, and Bio Data Server [5]). These mediators contain all the information needed to retrieve the requested data and to present them to the user. The systems do not explicitly conceptualize the structure of the accessible data, and, thus, the approach is less intuitive for users than other approaches based on data conceptualization.

The single conceptual schema approach uses a global conceptualization model for the data from all integrated databases. The advantage of this approach is that users can specify their queries with regard to a single global conceptual schema. However, as with data warehouses, any changes to the set of integrated DBs may call for modifications to the global conceptualization model. Examples of such systems are SIMS [6], Pegasus [7], Garlic [8], TAMBIS [9,10], ARIADNE [11], BACIIS [12], and Discovery Link [13].

The multiple conceptual schema approach does not rely on a global conceptualization model of the data. Instead, each DB is described by an individual conceptual schema. Additions, modifications, and removals of DBs only affect their conceptual schemas, not the whole system. User queries may be expressed using terms from specific domain ontologies. However, it cannot be generally assumed that the individual schemas employ the terms of such domain ontologies, and, thus, some relevant results may not be found when a query is executed. An example of a system based on multiple conceptual schemas is OBSERVER [14].

Finally, systems using the hybrid query translation approach use individual conceptual schemas to describe each database, but assure that these schemas have been created using a common global conceptualization or domain ontology. As with the previous approach, the incorporation of new DBs, or the modification or removal of DBs, does not require changes to the whole system. Moreover, users can specify their queries with respect to the domain ontology, and it is assured that these queries are transferred to the correct databases. Examples of hybrid query translation systems are PICSEL [15], COIN [16], MECOTA [17], BUSTER [18], and SEMEDA [19].

A different approach to integrating databases that cannot be classified within the above taxonomy is schema matching. Actually, schema matching is performed as an individual step in all the approaches included in the above classification—with the exception of information linkage.

Schema matching identifies conceptually equivalent objects in two or more schemas and creates a unified