Biological Data Integration using SOA

Noura Meshaan Al-Otaibi and Amin Yousef Noaman

***Abstract*—**Nowadays scientific data is inevitably digital and stored in a wide variety of formats in heterogeneous systems. Scientists need to access an integrated view of remote or local heterogeneous data sources with advanced data accessing, analyzing, and visualization tools. This research suggests the use of Service Oriented Architecture (SOA) to integrate biological data from different data sources. This work shows SOA will solve the problems that facing integration process and if the biologist scientists can access the biological data in easier way. There are several methods to implement SOA but web service is the most popular method. The Microsoft .Net Framework used to implement proposed architecture.

***Keywords*—**Bioinformatics, Biological data, Data Integration, SOA and Web Services.

I. Introduction

RECENT years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced, and protein and gene interaction data are accumulating [1]. These biological data is available in a wide variety of formats, annotated, and stored in flat files and relational or object-oriented data bases. The value of any kind of data is greatly enhanced when it exists in a form that allows it to be integrated with other data. An important aspect of bioinformatics consists in building a scientific digital library, integrated view of all data of interest widely distributed and constantly updated in heterogeneous remote public data sources or local private ones. Access to heterogeneous biological data sources is mandatory to scientists. A single query may involve flat files such as GenBank or SwissProt, web resources, or the references data source PubMed [2, 3].

Integration of biological data is just one phase of the entire molecular biology research and genomic hypothesis discovery process [6].

Several works showed that the integration of heterogeneous bio-molecular data sources can significantly improve the performances of data mining and computational methods for the inference of biological knowledge from the available data. Also, Integration of biological data allows uniform access of federation of several data sources [4]. Despite the importance, the following challenges make data integration one of the longest standing problems facing the database research community: how to solve the system heterogeneity; how to build the global model; how to solve the semantic heterogeneity; and how to deal with queries automatically, etc [8].

Contact us at nmalotaibi@kau.edu.sa , anoaman@kau.edu.sa

SOA is a novel architecture aimed to build collaborative computing systems. SOA is essentially a distributed architecture, with systems that span computing platforms, data sources, and technologies [8, 5]. SOA provides a standard method to integrate both data sources and software applications by regarding them as interoperable services. Thus, client applications will combine these services to implement their intended tasks [4]. The implementation of SOA using Web services technologies is the current state of the art in systems integration. This research will elaborate on implementing integration of biological data by using SOA. Also the research will discuss the following questions: Does using SOA solve the above challenges that facing integration process? Does the biologist scientists can access the biological data in easier way than using other integration approach?

II. Background

A. Data Integration

Convergent advances in biochemistry techniques, biotechnologies, and information technology and computer science provided the basis for the development of bioinformatics and made available huge and growing amounts of biological data. Nowadays public database infrastructure spans a very large collection of heterogeneous biological data, opening new opportunities for molecular biology, bio-medical and bioinformatics research, but raising also new problems for their integration and computational processing. Indeed the integration of multiple data types is one of the main topics in bioinformatics and functional genomics [4]. The process of heterogeneous database integration may be defined as "the creation of a single, uniform query interface to data that are collected and stored in multiple, heterogeneous databases." [7]. The example of databases that hold the biological data: Swiss rot and PIR focus on protein sequences, while Protein Data Bank (PDB) stores protein structures, Embank stores DNA sequences, BIND specialize in protein–protein interactions [4]. The main goal of integration is to provide mechanisms that can unify a number of (computer) systems. Three important aspects of system integration are distribution, autonomy and heterogeneity.

- *Distribution:* Often the source of database is distributed. The user need not know the location and other details of each available resource. Such details are usually handled automatically by the integrated system [12].

- *Autonomy:* It is very often the case that integrated resources belong to different organizations or research groups. Each

data sources are working autonomous without any control by another integrated system [12].

• *Heterogeneity:* In an open and diverse environment it is very common that some or all of the data sources are different from each other. The differences are either semantic or technical. Technical heterogeneity difference occurs because of different hardware platforms, operating systems, database management systems (query languages, data models) and programming languages. The semantic heterogeneity is conceptual differences that occur in the data models/schemas of the data sources i.e., the organization of data and the relationships between such data. For examples there are synonyms when attributes of two schemas have different names but refer to the same concept [12].

B. Service Oriented Architecture

The SOA was proposed initially as an emerging paradigm for business process integration inside or across organization boundaries [5]. SOA is an approach to defining integration architectures based on the concept of a service. A basic tenet of SOA is that the use of explicit service interfaces and interoperable, location-transparent communication protocols means that services are loosely coupled with each other. Services are software modules that are accessed by name via an interface, typically in a request-reply mode. Services can be invoked independently by either external or internal service requesters to process simple functions. Service consumers are software that embeds a service interface proxy (the client representation of the interface) [13].

III. Literature Review

As Internet accessible biomedical databases proliferate there is an increased need for tools capable of integrating information available from a variety of sources. Clinicians and researchers could benefit from a more consolidated and unified view of the available biomedical data. Systems biology researchers need to integrate disparate genetic information from multiple public sources to merge with their own experimental data [9]. A wide variety of technologies, techniques and systems have been explored and exploited over the past 15 years [10].

In following subsection we review some approaches (mediator and data warehousing) used for integrated biological data. Then the proposed approach (SOA) will be introduced to solve the problem of integration.

A. Mediator approach

Mediator-based integration concentrates on query translation. A mediator in the information integration context is a system that is responsible for reformulating at runtime a query given by a user on a single mediated schema into a query on the local schema of the underlying data sources. Typically, each individual source will also require the definition of a "wrapper" component, which will be used to export a view of the local data in a useful format for mediation [4, 6]. This approach required mapping to capture the relationship between the source descriptions and the mediator and thus allow queries on the mediator to be translated to queries on the data sources. Specifying this correspondence is a crucial step in creating a mediator, as it will influence both how difficult the query reformulation is and how easily new sources can be added to or removed from the integration system.

The two main approaches for establishing the mapping between each source schema and the global schema are global-as-view (GAV) and local-as-view (LAV). In the GAV approach the mediator relation nothing but a query over the data sources. The GAV approach greatly facilitates query reformulation, however handling the addition or removal of a source in a GAV mediator is much more difficult as it requires a modification of the mediator schema to take into account the changes. In a LAV-based mediator every source relation is defined over the relations and the schema of the mediator. It is therefore up to the individual sources to provide a description of their schema in terms of the global schema, making it very simple to add or remove sources but also complicating the query reformulation and processing role of the mediator [6].

This approaches is satisfied the integration of biological data. The scientist can access the different sources by sending the query and receiving result. Also, each source preserve own data autonomies. Also, this approaches is flexible, it allow adding or removing any sources. But there are problems in this approach the mapping between local and global schemas need to manually specify. Also this approach have complex schema to satisfied mapping between local and global view. Another drawback of this approach, sources must be available during query executions [6].

B. Data warehouse

Data warehouse is bringing all data from multiple sources into a local warehouse and executing all queries on the data contained in the warehouse rather than in the actual sources. The first step in data warehousing is to develop a unified data model that can accommodate all the information that is contained in the various source databases. The next step is to develop a series of software programs that will fetch the data from the source databases, transform them to match the unified data model and then load them into the warehouse. The warehouse can answering any question that can handle by source database also have integrated knowledge not in individual source database. Systems that rely on the data warehouse architecture are usually restricted to consider a few source databases, but can achieve a higher degree of integration of the data sources [1, 6]. The data warehouse achieves the high degree of integration of biological data sources. Also it reduces the response time to answer the queries because requests send to single place. But the major drawbacks of data warehouse are update issue. The all data insert and update in source data base must be re-imported in