Chapter 91
An Approach for Biological Data Integration and Knowledge Retrieval Based on Ontology, Semantic Web Services Composition, and AI Planning

Muhammad Akmal Remli
Universiti Teknologi Malaysia, Malaysia

Safaai Deris
Universiti Teknologi Malaysia, Malaysia

ABSTRACT

This chapter describes an approach involved in two knowledge management processes in biological fields, namely data integration and knowledge retrieval based on ontology, Web services, and Artificial Intelligence (AI) planning. For the data integration, Semantic Web combining with ontology is promising several ways to integrate a heterogeneous biological database. The goal of this work is to construct an integration approach for gram-positive bacteria organism that combines gene, protein, and pathway, thus allowing biological questions to be answered. The authors present a new perspective to retrieve knowledge by using Semantic Web services composition and Artificial Intelligence (AI) planning system, Simple Hierarchical Order Planner 2 (SHOP2). A Semantic Web service annotated with domain ontology is used to describe services for biological pathway knowledge retrieval at Kyoto Encyclopedia of Gene and Genomes (KEGG) database. The authors investigate the effectiveness of this approach by applying a real world scenario in pathway information retrieval for an organism where the biologist needs to discover the pathway description from a given specific gene of interest. Both of these two processes (data integration and knowledge retrieval) used ontology as the key role to achieve the biological goals.

DOI: 10.4018/978-1-4666-3604-0.ch091

INTRODUCTION

Post genomic era of biological experiment and research has produced a large number of biological data available to the scientific community through the Internet. Two important entities that realize this phenomenon are Human Genome Project (HGP) and the popularity of the Web or also called World Wide Web. The process to produce digital data from wet lab experiments has been tremendously meaningful. It allows researchers to do a lot of further and detail researches and contribute new knowledge and analysis. The approach called "High-Throughput Experiment" is one of the successful techniques that make data available over the Internet. There are many organizations and contributors provide data on the Internet ranging from genomics, proteomics, metabolamics, and many other "*omics*" studies. For instance, National Cancer Bioinformatics Institutes (NCBI), specializing in gene data, particularly provides very large information about gene. The gene, protein, and pathway are the common components and each of them is different in term of data format, ambiguity, resources, and relations. The most important phase of research in biological field relies on data gathering and analysis. These processes are performed in the first phase of system biology research to enable the biologists to understand the wide range of biological knowledge, for instances, to understand gene function, which gene is related to pathway and what type of biological process is involved. In system biology (Kitano, 2002), a study in biological process at system level is significant instead of focusing on molecular level in bioinformatics. Understanding of genes and proteins are important research. However, it also required to capture the whole biological systems in order to understand how they work simultaneously. Commonly, the public biological data stored in the Web are categorized by organisms. For instance, the most well known organism that is having continuously study is *Escherichia coli* (*E. coli*), one of the most popular gram-negative bacteria that can cause food poisoning in human (Cooke, 1985). Researchers can access all related data like gene, protein, and pathway regarding the organism over the Internet depending on their research purpose. It is clear that integration of interested organisms is demanded depending on how biological research is conducted. Due to the growth of biological data has increased over the Internet, the integration of these data in order to build new knowledge become more daunting and challenging.

Once integrated knowledge base has been built, retrieval process will take part. There are two general methods to retrieve knowledge in Semantic Web domain, namely, query method and Web service method. The query method involves several queries formulated by user to get information and knowledge in repository databases. In addition, the query formulation will resulting an improved quality and significance output since it can be performed locally and independently. However, this method encounter various problems, including domain expert is needed to formulate the query language and how to connect every single biological component such as gene, protein and pathway is very challenging. Thus, the most common and popular method to retrieve knowledge is by using Web services. Web service is a piece of computer software that acts as an '*information transformer,*' which produces output when input parameter is given by client programs. The growth of Web service in bioinformatics has increased and already matured in biological domain, which is intended to perform various bioinformatics tasks. To date, there are over 2000 services published in BioCatalouge, the life science Web services registry (Bhagat, et al., 2010). This number will increase rapidly because the dynamic nature of such software is able to operate in heterogeneous and distributed computing power. Current Web services are syntactic in terms of service description, which is expressed in Web Service Description Language (WSDL). Thus, Semantic Web is used to annotate service description using