A System for Integration of Heterogeneous Biological XML Data

Noboru Matoba
noboru-m@is.aist-nara.ac.jp
Masatoshi Yoshikawa
yoshikawa@is.aist-nara.ac.jp
Junko Tanoue
junko-ta@is.aist-nara.ac.jp
Shunsuke Uemura
uemura@is.aist-nara.ac.jp
Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan

Keywords: XML, database, data integration, annotation, Gene Ontology, MaXML

1 Introduction

Biological data exists all over the world as various web services, which provide biologists with many useful information. However, when users actually make use of them, they need to access each web service (database) one by one. If they want to compare many different kinds of data, they have to do cumbersome tasks. Furthermore, there are many other problems which make such comparison difficult: firstly, from a viewpoint of biological data: (1) data is omnipresent all over the world, (2) the vocabulary is not unified, and secondly, from web services (database): (1) each database has its own user interface and schema, (2) data is updated frequently, (3) the required knowledge about the data currently supported and the operation method of the service itself differs from system to system.

XML (Extensible Markup Language) [3] is an emerging technology for solving not all but many of these problems. XML is rapidly spreading as a standard data format for exchanging and sharing data on the WWW. In the field of biology, many kinds of XML data has already been provided. However, many of biological XML data is used only for distribution, and in many cases, web service is not supported. Typical web service supported is translation of XML data into HTML data to display the data on browsers. Obviously, these are not the services which fully exploit the advantage of XML. Therefore, it is an immediate subject to design systems which can utilize XML data effectively. In this poster, we propose a system for integrating biological XML data using XML.

2 Method and Materials

Comparing and retrieving data on several species will be becoming important task. However, retrieving data from several heterogeneous databases is not easy, because each of which has its own schema. Our goal is providing users with a system which supports a mechanism to access heterogeneous biological XML data. In our system, users do not need to understand the details of schemas or structures of target databases. Nor they need a special knowledge about operation of XML or a system. They can issue complex demands by same interface as existing web services.

Our prototype system consists of three parts, as illustrated in Figure 1: (A) Client. Clients use an ordinary web browser. A user can operate the system without special knowledge. It sends requests from a web browser to the XML Database System via a web server. (B) XML Database System. The XML Database System has XML data, and stores both an XML Query Engine and the Meta Knowledge. The XML Query Engine processes a demand from the client, and sends applicable data to the XSL Transform in XML format. (C) XSL Transform. The XSL Transform changes XML data from the XML Database System into HTML form, and returns it to the client.
Our system supports XML Query [4]. Accordingly, we consider that our system can modularize a series of processing steps until it returns a result from complicated queries over many kinds of data and can deal with updating and an addition of XML data easily. Moreover, our system offers user interface reflecting a user’s purpose by generating the HTML file as a display format translated from the XML file as a data.

Currently, we are developing a system which stores XML data of GO (Gene Ontology) [1] and MaXML (Mouse annotation XML) [2] which are available on the Web. By unifying these two data, users can easily understand the hierarchical relationship by using functional classification about mouse cDNA.

3 Discussions

Gene Ontology attracts attention from many organizations, and we expect that many data which refers to Gene Ontology is offered in future. If these data is distributed in XML form, a user can easily perform functional comparison of gene between different species, and can acquire the key of discovery of new biological knowledge. We are planning to exhibit this system on the Internet, and to support other XML data.

References