XML Databases for Bioinformatics

Mark Graves, Berlex Biosciences

Tutor

I have been involved in software development for fifteen years and in some aspect of database development for most of that time. The past eight years have been in bioinformatics, including a DOE Distinguished Postdoctoral Fellowship at Baylor College of Medicine (BCM). For two years, I was an Instructor at BCM where I developed software and research database management systems to support the Human Genome Project. For the past four years, I have been working in the biotech and pharmaceutical industry, and I am currently the solution architect for a multi-company, international database project at Berlex Biosciences (a division of Schering AG). I am also the author of the forthcoming book "Designing XML Databases" from Prentice Hall.

My academic bioinformatics teaching experience has been team teaching a bioinformatics course at Baylor College of Medicine. I also taught course sections on scientific databases at BCM, theoretical computer science at University of Michigan, artificial intelligence and Lisp programming at Georgia Tech, and Pascal at University of Delaware. I also have experience teaching Logo programming to children ages 4-6.

Tutorial

XML has been heralded as a panacea which will revolutionize the web, invigorate the economy, and change the world's perception of data. These are amazing expectations for a new file format. In this tutorial, I discuss what XML can and cannot do and what are the ramifications of the technology's strengths and weaknesses to its use in bioinformatics. The goal of the tutorial is to provide an basic understanding of XML-related technologies and databases and how they can be used in bioinformatics.

The tutorial consists of three parts. The first part of the tutorial is an overview of XML-related technologies, including XML Path, XSL, XML Schema, DOM, and SAX parsers. XML is a relatively simple language to understand, but the family of XML-related standards and how they interoperate can be somewhat daunting to approach unassisted. The goal of the first part of the tutorial is to explain the simple aspects of XML, provide an overview of the technologies, and describe which ones may be worthy of further investigation for specific bioinformatics applications. The remainder of the tutorial explores in more depth two specific applications of XML to bioinformatics.

The second part of the tutorial is focused on databases and their use with the web. Bioinformatics databases can be created using flat files or relational database management systems (DBMS), such as Oracle, DB2, or mysql. Existing databases and data formats may be translated into XML, and techniques for translating that data into XML are presented. Particular emphasis is placed on designing good XML document structures which capture the flexible data structures inherent to bioinformatics. Technologies are described which extract data from relational databases to format the data as XML and which translate XML documents into other formats. The goal of the second part of the tutorial is to explain how to use XML to web-enable existing

databases. The emphasis of this part of the tutorial is on using XML with an existing database.

In the third part of the tutorial, new database systems are presented which support the storage and access of XML documents. These systems can form the foundation for bioinformatics databases that supports large quantities of data securely stored in a flexible format. They can use native storage facilities or leverage storage systems already found in flat files, relational DBMSs or object-oriented DBMSs depending upon the requirements for the system. Particular emphasis is placed on efficiently storing XML documents in systems built on relational DBMS in a way which supports querying the documents (such as annotations or literature) and combining those queries with data stored in other relational tables (such as sequence or expression data). The goal of the third part of the tutorial is to explain how to create a new database that uses XML to capture data which does not easily fit within the rigid structure of relational databases. The emphasis in on building a new database that leverages both XML and existing DBMS technology.

Audience

The audience will consist of bioinformatics software developers, scientists interested in learning more about XML, and anyone with large datasets who wish to maintain or share them more effectively. The proposed tutorial will be accessible to anyone with a basic understanding of bioinformatics, databases, and web technologies. It is advanced in content as it assumes the attendee is interested in working with large quantities of bioinformatics data using web technologies. However, I do not assume the attendee is familiar with any particular technology.

The audience participants should be familiar with the basics of relational database management systems and database design. They do not need to know XML or SQL, though that knowledge would provide a firmer foundation for the concepts of the tutorial.

Outline

A half-day tutorial slot is requested.

I. Introduction & Overview

II. XML Technologies

A. XML

B. W3C Standards

C. XML Path

D. XSL

E. XML Schema

F. DOM

G. SAX parsers

H. Applications of Standards
I. Applications of Standards to Bioinformatics
J. Technology Questions

III. Web-enabled databases
A. Overview
B. Web servers
C. Flat file XML Documents
D. XML+XSL->HTML
E. XML+Java->GUI
F. Relational DBMS
G. Relational data -> XML
H. Other formats -> XML
I. Other databases -> XML
J. Complex flat file organizations
K. Hybrid web-enabled databases
L. Conclusions & Applications

IV. XML Database Systems
A. DBMS Architecture
B. Data Models
C. Leveraged Storage Systems
D. XML Data Model
E. XML Storage Systems
F. XML Storage in Relational DBMS
G. Extended Example
H. Hybrid System
I. Extended Hybrid Example
J. Querying
K. Conclusions & Applications

V. Review & Questions