20.453J / 2.771J / HST.958J Biomedical Information Technology
Fall 2008

PARSING AND QUERYING XML DOCUMENTS ASSIGNMENT II

1. BACKGROUND

1.1 *Objectives*

Upon completion of the assignment, the student should be able to:

- Develop Java programs to parse XML documents using SAX
- Understand the schema of SBML and RDF-based markup languages
- Use XQuery to query XML documents through Nux Java API

1.2 *Introduction*

The aim of this assignment is to develop a program that parses XML documents and answer queries on the documents using Java and XQuery. Submissions must be original work.

1.3 *Prerequisites*

Basic knowledge in Java programming is required (See Appendix A).

2. DESCRIPTION OF ASSIGNMENT

BioModels (http://www.ebi.ac.uk/biomodels-main/) is a repository of mathematical models of biological interests. Models are stored under SBML format (http://sbml.org/More_Detailed_Summary_of_SBML), which is encoded in XML.

Download the following models from BioModels:

- ➢ **Kholodenko1999_EGFRsignaling**
  http://www.ebi.ac.uk/biomodels/models-main/publ/BIOMD0000000048.xml

- ➢ **Sasagawa2005_MAPK**
  http://www.ebi.ac.uk/biomodels/models-main/publ/BIOMD0000000049.xml

2.1 *Using Java to parse and query XML documents*

Write a Java program to parse the above documents and answer the below queries:

a. List all species that exists in both models
b. List the 10th reaction in both models according to document order.
c. In **Sasagawa2005_MAPK**, list all reactions that involves the species “MEK” or “SOS”

d. In **Sasagawa2005_MAPK**, list all reversible reactions. For each reaction, show the *Uniprot* ids of the reactants and products
e. For both documents, list all species that are involved in more than 3 reactions

You **must** use the SAX API (see Appendix B) to process the XML documents. Write a Java method to process each query **without** using XPath/XQuery. You may either print the answers on the screen or store them as XML documents.

Submit the following to the course website:

- Java source code
- A brief explanation of your algorithm

2.2 *Using XQuery*

Nux (http://acs.lbl.gov/nux/index.html) is a Java toolkit that allows one to run XQuery queries on XML documents. Write a Java program to answer queries (a) to (e) in Section 2.1 using XQuery. Refer to http://acs.lbl.gov/nux/index.html and Nux documentation to learn how XQuery can be processed using Nux Java API. You may either print the answers on the screen or store them as XML documents.

Submit the following to the course website:

- Java source code
- XQuery file for each query
- A brief explanation of your algorithms

3. APPENDIX A: JAVA PROGRAMMING GUIDES

The fundamentals of Java can be found at http://java.sun.com/docs/books/tutorial/java/index.html. See http://java.sun.com/docs/books/tutorial/getStarted/cupojava/win32.html for a brief tutorial on how to install Java SE Development Kit (JDK), compile a simple Java source file, and run it. Download Eclipse IDE (http://www.eclipse.org), which simplifies the coding/compiling process.

Java developer checklist:

- Java Runtime Environment (http://java.com/en/download/manual.jsp)
- Java SE Development Kit 6.0 (http://java.sun.com/javase/6/download.jsp)
- Eclipse IDE (http://www.eclipse.org/downloads/)

4. APPENDIX B: SAX PARSER GUIDES

Tutorials on SAX Java API can be found at http://developerlife.com/tutorials/?p=29 and http://www.ibm.com/developerworks/edu/x-dw-xusax-i.html.