Web Services for PIR/UniProt Databases

Baris E. Suzek, Hongzhan Huang, Scott Chung, Hsing-Kuo Hua, Peter McGarvey, Zhangzhi Hu, Cathy Wu
Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA 20057-1455

Protein Information Resource (PIR) is an integrated bioinformatics resource that provides protein databases and analysis tools to support genomic and proteomic research. PIR recently joined with the European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB) to establish UniProt—the Universal Protein Resource—to produce a single worldwide resource of protein sequence and function, by unifying the PIR, Swiss-Prot, and TrEMBL database activities (http://www.uniprot.org). The UniProt Knowledgebase (UniProtKB) provides the central database of protein sequences with accurate, consistent, rich sequence and functional annotation. UniProtKB consists of two sections: Swiss-Prot, containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and TrEMBL, containing computationally analyzed records that await full manual annotation. One of the biggest challenges in life sciences research is the discovery, integration and exchange of data coming from multiple research groups. To make the PIR resource widely accessible to the research community and application programs, we are adopting an open-source, common-standard distribution practice and employing industry-standard J2EE technology to develop protein object models and web services. To make the PIR resource interoperable with other bioinformatics databases, we are developing controlled vocabularies and common data elements.

The web services is in the framework of the cancer Biomedical Informatics Grid (caBIG), an infrastructure connecting individuals and institutions to enable the sharing of data and tools for cancer research and developed under the leadership of National Cancer Institute's Center for Bioinformatics (NCICB). PIR, as a participant of caBIG, is developing "Grid-enablement of PIR/UniProt Data Source" project. The goal of this project is to demonstrate how the PIR/UniProt data source can be discovered and consumed in a grid environment by creating an object layer and a web service layer for accessing the data source. The project has an n-tier architecture. *The data layer*, supported by Oracle 9i, stores the UniProtKB data. The *data access layer* utilizing Hibernate provides the mapping between relational database and object model. The *object layer* is developed using a Model Driven Architecture (MDA) approach. The use cases are developed with input from user community. The objects and their relations are designed using Unified Modeling Language (UML) in combination with existing UniProtKB XML schemas. An object-XML mapping tool (Castor) has been used to serialize/deserialize XML data from/to objects. The *web service layer*, supported by Apache Axis, provides language-independent programmatic access to the objects using SOAP protocol. The *web services* will facilitate many query mechanisms to access PIR/UniProt data:

- Identifier searches such UniProtKB ID, RefSeq number
- String-based searches for fields such as protein, gene name or keywords
- Boolean searches

The results are returned in XML and FASTA format for ease data exchange.

To address the issues of data interoperability, PIR is participating in development of common data elements (CDE) as a part of caBIG Vocabulary and Common Data Elements (VCDE) activities. As members of the NIAID Administrative Resource for Proteomic Research Centers, the PIR team and the Virginia Bioinformatics Institute are developing a cyber infrastructure with a central proteomic database for the NIAID Proteomic Research Program. We have established an Interoperability Working Group (IWG) to discuss and address database interoperability issues. Interconnecting with the IWG and caBIG VCDE activities, we also participate in the HUPO PSI, focusing on mass spectrometry (PSI-MS) and general proteomics standards for formats (PSI-ML, XML format for data exchange), minimum reporting requirements (MIAPE), and ontologies (PSI-Ont).