Web Services for Sequence Analysis at DKFZ

Peter Ernst and Sándor Suhai
Husar Bioinformatics Laboratory, Department of Molecular Biophysics
Im Neuenheimer Feld 580, 69120 Heidelberg, Germany
http://genome.dkfz-heidelberg.de

Web Services at DKFZ

Web services start to play more and more an important role in the interconnection of bioinformatics analyses at DKFZ with bioinformatics activities in other institutes. On one hand this means the inclusion of web services on the internet or collaboration partners into analysis pipelines developed at DKFZ. On the other hand, local developments will be made available via the web service technology.

At DKFZ, the web service developments are integrated into the HOBIT platform (Helmholtz Open Bioinformatics Technology; http://hoibit.sourceforge.net). Using this forum, the reinvention of datatypes can be avoided and problems typical for bioinformatics application can be solved in a uniform way. Web services for the database engines EMBoss and EMBoss2 have been developed. These services provide access to a large public database collection via the backends’ query languages.

In order to connect the software on DKFZ servers with the web service technology, interfaces have been developed using the WSDL and SOAP toolkit. Using this toolkit for C/C++ based software has several advantages:

- The service applications are very easily integrated in an existing CGI web setup.
- WSDL descriptions are generated from C header files.
- Good (de)serialization also of complex data types.
- Allows starting the development from an XML schema or a C header file.

Example 1

GOslim: mapping GO terms to GOslims

DESCRIPTION

GO slims are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene products function is required.

IMPLEMENTATION

The GOslim web service at DKFZ performs the calculation of a slim for a given query set of GO numbers. Custom definitions of the slims are possible.

Example 2

SoapDB: sequence databases at DKFZ

SoapDB is a web service module, providing access to public databases installed and updated at DKFZ. It contains interfaces to the EMBoss and the SRS system. Thus, the user of this web service can request database information in a variety of output formats.

Example 3

DomainSweep: automatic functional annotation of proteins

PROBLEM

In order to determine the function of an uncharacterized protein sequence, DomainSweep compares a protein sequence with a range of protein family databases. This leads to the identification of the protein’s domain architecture and thus helps to find correct functional assignments.

The output of DomainSweep comprises a summary of predicted domains and the individual database search results.

IMPLEMENTATION

DomainSweep is implemented under the W3H task framework6. This framework allows the integration of applications and methods to create tailor-made analysis task flows, which can be used for high throughput analysis.

Results are in XML format and as such ready to use in web services. In order to support common bioinformatics XML Schemas like e.g. BioSchemas, the task results can be transformed with XSLT. However, for many areas such schemas still need to be developed.

W3H task system