# WEB SERVICES AT THE EUROPEAN BIOINFORMATICS INSTITUTE

***A. Labarga*, M. Anderson, F. Valentin, R. Lopez***

European Bioinformatics Institute, Hinxton, United Kingdom
e-mail: alabarga@ebi.ac.uk
*Corresponding author

**Keywords:** web services, database integration, bioinformatics workflows

## Summary

*Motivation:* Following the exponentially growing amount of genomic sequence data from the different genome sequencing projects, and lately, of gene expression and protein interactions data, the challenge is now to unravel the gene functions, and to understand the gene regulation processes, and this genome-wide data analysis requires the complex interoperation of multiple databases and analytic tools. On the basis of these observations, the major world institutions for bioinformatics (like EBI, DDBJ, NCBI) have chosen to use the Web Services technology to expose its services in a programmatically accessible manner.

*Results:* We will show some of the tools we are developing a the European Bioinformatics Institute and how they can be used to construct analysis workflows using existing web services to help scientists to perform different tasks related to functional genomics experiments such as protein function prediction, microarray data annotation or literature data mining.
*Availability:* http://www.ebi.ac.uk/Tools/webservices

## Introduction

Today, biological databases are large collections of data that are relatively difficult to maintain outside the centers and institutions that produce them. These data and the corresponding analysis tools are mainly accessed using browser-based World Wide Web interfaces. When large amounts of data need to be retrieved and analyzed, this often proves to be tedious and impractical. Moreover, research is rarely completed just by retrieving or analyzing a particular nucleotide or protein sequence. Database information retrieval and analysis services have to be linked, so that, for example, search results from one database can be used as the base of a search in another, the results of which are then analyzed. When performing these operations using a Web browser, researchers are forced to repeat the troublesome tasks of searching; copying the results for subsequent searches to other databases, and again copying the results for analysis.

Creating a local bioinformatics work environment is possible by downloading and installing the necessary database content and services (such as retrieval and analysis programs). This has the advantage that processes that otherwise require manual operations can be automated. However, the amount of disk space required to store biological sequence databases can be huge, often exceeding several terabytes, requiring several hours, if not days, to complete analysis, even when using a supercomputer. For this reason, creation of a local system is not a suitable option for most individual researchers or institutions.

Web Services technology enables scientists to access biological data and analysis applications as if they were installed on their laboratory computers. Similarly, it enables programmers to build complex applications without the need to install and maintain the databases and analysis tools and without having to take on the financial overheads that accompany these. Moreover, Web Services provide easier integration and interoperability between bioinformatics applications and the data they require.

## Methods and algorithms

To ensure software from various sources work well together, this technology is built on open standards such as Simple Object Access Protocol (SOAP), a messaging protocol for transporting information; Web Services Description Language (WSDL), a standard method of describing Web Services and their capabilities, and Universal Description, Discovery, and Integration (UDDI), a platform-independent, XML-based registry for services. For the transport layer itself, Web Services can use most of the commonly available network protocols, especially Hypertext Transfer Protocol (HTTP).

A client (program) connecting to a web service can read the WSDL to determine what functions are available on the server. Any special data types used are embedded in the WSDL file in the form of an XML Schema. The client can then use SOAP to actually call one of the functions listed in the WSDL.

### *Services available*

Currently, we support SOAP services for both database information retrieval and sequence analysis (Pilai et al. 2005). All available information about EBI web services can be accessed from the web page http://www.ebi.ac.uk/Tools/webservices.

*Sequence and Literature data retrieval*

*WSDbfetch* provides programmatic access to the popular sequence and literature data retrieval tool dbfetch (http://www.ebi.ac.uk/Tools/dbfetch). The databases currently available for data retrieval using this service include EMBL, EMBL-SVA, MEDLINE, UniProt, InterPro, PDB, RefSeq and HGVBase (Pilai et al., 2005)

*CitationExplorer* combines literature search with text mining tools for biology. It provides access to Medline, PubmedCentral, Patent Abstracts and Chinese Biological Abstracts databases. You can get full records from these databases, full text when available, and results are enriched with links to biological databases, synonyms, ontologies, etc.

*OntologyLookup* provides a web service interface to query any ontology available in the Open Biomedical Ontology (OBO) format.

*Sequence analysis*

The European Bioinformatics Institute also provides Web Services for sequence similarity tools (WSFasta, WSWUBlast, WSNCBIBlast, WSMPsrch, WSScanPS, WSScanWise); protein analysis (WSInterProScan); multiple alignment (WSClustalW, WSMuscle and WSTcoffee); and the European Molecular Biology Open Source Software Suite (WSEMBOSS), among others. These Web Services provide the same or even more advanced functionality than the traditional browser-based services described in (Harte et al 2004).

## Results and discussion

One of the main advantages of web services is that researches can construct easily bioinformatics workflows and pipelines combining two or more web services to solve complex biological tasks such as protein function prediction, genome annotation, microarray analysis, etc. These workflows can be created simple scripts, using advanced integration frameworks such as JBI, BPEL, etc or using scientific graphical workflow tools such as Taverna, developed at the EBI, or Triana, developed at the University of Cardiff.

### *Protein function prediction and classification*

EBI web services are used internally to create new composite services such as InterProScan, (Quevillon E., et al. 2004) which combines different databases and protein signature recognition methods to provide automatic classification of proteins and function annotation, or ProFunc (Laskowski et al. 2005) which has been developed to help identify the likely biochemical function of a protein from its three-dimensional structure using a combination of both sequence- and structure-based methods.

### *Gene expression analysis*

In a typical microarray analysis workflow, the user might either upload gene expression data from an external source via the Expression Profiler (Kapushesky et al 2004) web service, or retrieve data from the ArrayExpress public repository at the EBI. The Data Selection method provides a basic statistical overview of the dataset, which can be used to guide the user in selecting genes relevant for further analysis. The Data Transformation methods can impute missing values in the chosen data subset, and perform other data transformations. Following these optional preprocessing steps, data can be subjected to one or more analyses. The user can explore the overall structure of the data via one of the clustering methods available in the Hierarchical and K-groups Clustering components, and the best number and quality of clusters within the data can be evaluated using a novel algorithm, Clustering Comparison. Alternatively, the user can subject data to a supervised method aimed at studying correspondences between groups of samples and genes in the Between Group Analysis component. Users interested in studying a specific gene or a group of genes can annotate them using WSDbfetch and map them to different ontologies using the OntologyLookup service, to get an insight on the molecular functions involved.

### *Genome annotation*

For annotating a novel DNA sequence, users can use the WSGeneMark service to locate the predicted exons. Then they can use the EMBOSS tool sixpak to generate the 6-frame translation and perform a WSWUBlast search against Uniprot, get the homologous sequences with WSDbfetch and align them with WSClustalW. They can also submit the sequences to WSInterProScan for automatic identification of domains and classification.

## Conclusion

Web Services technology brings into the bioinformatics community a new development concept in which users can access all data and applications hosted in the main research centers (EBI, DDBJ, KEGG, NCBI, etc) as if they were installed in their local machines, providing seamless integration between disparate services and allowing the construction of workflows to perform complex tasks.

## References

Harte, N. et al. Public web-based services from the European Bioinformatics Institute Nucleic Acids Res., 32, 3–9 (2004).
Kapushesky M. et al. Expression Profiler: next generation—an online platform for analysis of microarray data. Nucleic Acids Research Vol. 32, Web Server issue (2004)
Labarga, A. et al. Web services at EBI EMBnet.news, 11(4) 18-23 (2005)
Laskowski et al. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89-W93 (2005).
Pillai, S. et al. SOAP-based services provided by the European Bioinformatics Institute Nucleic Acids Res. 33(1):W25-W28 (2005).
Quevillon E., et al. InterProScan: protein domains identifier. Nucleic Acids Research 33: W116-W120 (2005)

## Links

| | |
|---|---|
| EBI Web Services: | http://www.ebi.ac.uk/Tools/webservices |
| EBI services: | http://www.ebi.ac.uk/services |
| Profunc: | http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/ |
| InterProScan: | http://www.ebi.ac.uk/InterProScan |
| Expression Profiler: | http://www.ebi.ac.uk/expressionprofiler |
| Taverna: | http://taverna.sourceforge.net |
| Triana: | http://www.trianacode.org |