Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W25–W28***
doi:10.1093/nar/gki491

SOAP-based services provided by the European Bioinformatics Institute

S. Pillai, V. Silventoinen, K. Kallio, M. Senger, S. Sobhany, J. Tate, S. Velankar, A. Golovin, K. Henrick, P. Rice, P. Stoehr and R. Lopez*

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received February 14, 2005; Revised and Accepted April 25, 2005

ABSTRACT

SOAP (Simple Object Access Protocol) (http://www.w3.org/TR/soap) based Web Services technology (http://www.w3.org/ws) has gained much attention as an open standard enabling interoperability among applications across heterogeneous architectures and different networks. The European Bioinformatics Institute (EBI) is using this technology to provide robust data retrieval and data analysis mechanisms to the scientific community and to enhance utilization of the biological resources it already provides [N. Harte, V. Silventoinen, E. Quevillon, S. Robinson, K. Kallio, X. Fustero, P. Patel, P. Jokinen and R. Lopez (2004) *Nucleic Acids Res.*, 32, 3–9]. These services are available free to all users from http://www.ebi.ac.uk/Tools/webservices.

INTRODUCTION

Today, biological databases are large collections of data that are relatively difficult to maintain outside the centres and institutions that produce them. These data are traditionally accessed using browser-based World Wide Web interfaces. When large amounts of data need to be retrieved and analysed, this often proves to be tedious and impractical. Web Services technology enables scientists to access these 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 (1) 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.

THE TECHNOLOGIES

The European Bioinformatics Institute (EBI) has tried and tested standards such as CORBA (http://www.corba.org) and Web Services. CORBA is standardized and mtature; it uses the Inter-ORB Protocol (IIOP) and can be tunnelled through HTTP but does not natively support HTTP. It is trickier to communicate through firewalls. Web Services uses SOAP (the Simple Object Access Protocol) over HTTP. It interacts with other systems using messages based on eXtensible Markup Language (XML) (http://www.w3.org/XML). A SOAP message can be transferred using almost any application or transport protocol. SOAP uses the Web Services Description Language (WSDL) (http://www.w3.org/TR/wsdl) to describe its interface. A SOAP client can read the WSDL at runtime and dynamically select the proper data-encoding scheme and network transfer protocol. SOAP implementations are available for many programming languages, including Perl and Java, which are popular languages among bioinformaticians.

On the basis of these observations, the EBI has chosen to use the Web Services technology to expose its services in a programmatically accessible manner. All that is required by the bioinformatics programmer is a lightweight program that communicates with existing services running at the EBI. These services have several advantages. As traditional web browsers cannot be used programmatically, these services provide an easy and flexible way to deal with repetitive tasks such as bulk submission with minimal intervention from the user. Web Services clients allow the programmer as well as the service provider to integrate and build more complex analysis workflows using existing EBI services. Also, using these services effectively avoids the need to maintain many programs and databases locally.

SERVICES

In this article, we describe services currently available at the EBI via a SOAP server. These include tools for sequence and literature data retrieval, sequence similarity search services, protein function analysis and structural analysis tools that access the Macromolecular Structure Database (MSD) (2) and a set of Web Services called Soaplab (3) for the European

*To whom correspondence should be addressed at: EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 ISD, UK. Tel: +44 1223 494423; Fax: +44 1223 494468; Email: rls@ebi.ac.uk

Table 1. Web Services available at the EBI

Service	URL	WSDL
WSDbfetch	http://www.ebi.ac.uk/Tools/webservices/WSDbfetch.html	http://www.ebi.ac.uk/ws/urn:Dbfetch?wsdl
WSFasta	http://www.ebi.ac.uk/Tools/webservices/WSFasta.html	http://www.ebi.ac.uk/ws/WSFasta.wsdl
WSWUBlast	http://www.ebi.ac.uk/Tools/webservices/WSWUBlast.html	http://www.ebi.ac.uk/ws/WSWUBlast.wsdl
WSInterProScan	http://www.ebi.ac.uk/Tools/webservices/WSInterProScan.html	http://www.ebi.ac.uk/ws/WSInterProScan.wsdl
MSD Services	http://www.ebi.ac.uk/msd-srv/docs/api	http://www.ebi.ac.uk/msd-srv/docs/api/msd_soap_service.wsdl
Soaplab	http://www.ebi.ac.uk/soaplab	—

Table 2. Methods available in the WSDbfetch service

Methods	Description
getSupporteDBs	Lists the databases available for data retrieval
getSupportedFormats	Lists the format in which data can be obtained for each of the available databases along with the default format
getSupportedStyles	Lists the style in which the result can be obtained
fetchData	Takes the database name followed by a primary or secondary identifier, format and style as parameters and returns the result as a string

Molecular Biology Open Software Suite (EMBOSS) (4). Table 1 lists services available, links to the web pages and WSDL.

Sequence and Literature data retrieval: WSDbfetch

WSDbfetch provides programmatic access to the popular sequence and literature data retrieval tool dbfetch (http://www.ebi.ac.uk/cgi-bin/dbfetch). The databases currently available for data retrieval using this service include EMBL (5), EMBL-SVA (6), MEDLINE, UniProt (7), InterPro (8), PDB (9), RefSeq (10) and HGVBase (11). The data backends currently used are SRS (12), Sequence Version Archive (SVA) and the UniProt consortium server at the EBI (http://www.ebi.uniprot.org), but the service can be easily modified to use other data retrieval systems. It is implemented on Apache Axis (http://ws.apache.org/axis). Users can call the WSDbfetch service from an application written in any programming language that supports SOAP. Data can be retrieved from a database using either a primary or secondary identifier. Each database supports various formats and styles, of which one is set as a default. The results can be obtained as pure ASCII text, HTML with hyperlinks or XML, where available. The various methods available for this service and their descriptions are shown in Table 2. Fully functional Java and Perl client programs are available from http://www.ebi.ac.uk/Tools/webservices/WSDbfetch.html. A sample client in Perl is shown in Figure 1, and Figure 2 illustrates how to run the client.

Sequence searching and protein function analysis: WSFasta, WSWUBlast, WSInterProScan

The EBI provides Web Services for sequence similarity tools such as Fasta (13), WUBlast (14) and the protein function analysis tool InterProScan (15). These are Web Services servers that provide the same functionality as the traditional browser-based services found at http://www.ebi.ac.uk/fasta, http://www.ebi.ac.uk/blast2 and http://www.ebi.ac.uk/InterProScan, respectively. These services are implemented on a Perl-based, SOAP::Lite (http://www.soaplite.com) server. The methods provided for using these services are listed in Table 3. Fully functional Java and Perl client programs are available from http://www.ebi.ac.uk/Tools/webservices/download.html. A sample client in Perl is shown in Figure 3, and Figure 4 illustrates how to run the client. Depending on the input and the databases chosen for the search, jobs may take seconds to complete or up to a few hours. Two modes of job submission exist: synchronous and asynchronous.

Table 3. Methods available in the WSFasta, WSWUBlast and WSInterProScan services

Methods	Description
doFasta/doWUBlast/doIprscan	Input parameters are a set of key-value pairs that correspond to choosing program names, databases, gap values, matrices, job mode etc., and the input sequence in Fasta format. Depending on the job mode chosen, the method returns either a job identifier or the result of the job when completed
polljob	This method is used when a job was submitted in asynchronous mode. This method takes the job identifier and an optional format name as arguments. It returns either the result or the status of the job

Synchronous mode. This mode is equivalent to a user running a command on a console or terminal and waiting for it to complete. This requires the client to be constantly connected to the server. This mode is suitable for database searches that can be executed in up to 5 min (e.g. protein versus protein searches).

Asynchronous mode. In this mode, the user submits a job and receives a job identifier in return. This is the same as running a UNIX command in the background and obtaining a job id. The user can use the 'jobs' command to list processes that are running in the background. Similarly, the user can query or poll the status of an asynchronous mode job and receive the following four states in response: JOB RUNNING (i.e. the job is currently being processed), JOB PENDING (i.e. the job is in a queue waiting processing), JOB NOT FOUND (i.e. the job id is no longer available; job results are deleted after 24 h) and JOB FAILED (i.e. the job failed or no results where found).

Typically, the asynchronous submission mode is recommended when users are submitting batch jobs (e.g. many protein sequences to analyse using InterProScan) or large database searches (e.g. searching the whole of the EMBL nucleotide sequence database). One advantage of this mode is

```
#!/usr/bin/perl
use SOAP::Lite;
my $uri = 'urn:Dbfetch';
my $proxy = 'http://www.ebi.ac.uk/ws/services/Dbfetch';
my $soap = new SOAP::Lite(uri => $uri, proxy => $proxy);
my $result = $soap->fetchData("uniprot:wap_rat", "fasta", "raw");
print $result;
```

Figure 1. A sample Perl client calling the fetchData method.

```
% java DbfetchClient fetchData uniprot:wap_rat fasta raw
>uniprot|P01174|WAP_RAT Whey acidic protein precursor (Whey
phosphoprotein) (WAP).
MRCSISLVLGLLALEVALARNLQEHVFNSVQSMCSDDSFSEDTECINCQTNEECA
QNDMCCPSSCGRSCKTPVNIEVQKAGRCPWNPIQMIAAGPCPKDNPCSIDSDCSG
TMKCCKNGCIMSCMDPEPKSPTVISFQ
```

Figure 2. A sample client invocation showing the method called and result obtained.

```
#!/usr/bin/perl
use SOAP::Lite;
my $WSDL = 'http://www.ebi.ac.uk/ws/WSFasta.wsdl';
my $soap = SOAP::Lite->service($WSDL);
my %params = (program => 'fasta3',
                         database=>'uniprot',
                         searchtype=>1, # indicates asynchronous mode
);
open INPUT, "input.txt"
my $content = <INPUT>;
close INPUT;
my $result = $soap->doFasta(SOAP::Data->name('params')
                  ->type(map=>\%params),
                    SOAP::Data->name(content=>$content)
                  ->type('base64'));
print $result;
```

Figure 3. A sample Perl client for WSFasta calling the doFasta method asynchronously.

```
% perl WSFastaClient.pl --program fasta3 --database uniprot --async \
--file mysequence
fasta-20050412-12273753

% perl WSFastaClient.pl --polljob --jobid fasta-20050412-12273753
fasta-20050412-12273753 : JOB RUNNING
```

Figure 4. A sample client invocation. Note that '\' means a continuous one-line command. The input file (e.g. mysequence) is a Fasta-formatted sequence.

that it is impervious to system or network failure. The results of jobs are stored at the EBI for 24 h after the job has completed.

Structural Analysis

The EBI provides a Web Services interface to tools that access the MSD. This service enables software developers to query the MSD directly from their own application programs and is further described at http://www.ebi.ac.uk/msd-srv/docs/api. The available functions are described in the corresponding WSDL description at http://www.ebi.ac.uk/msd-srv/docs/api/msd_soap_service.wsdl. As well as simple extraction of data from the database, the interface also provides methods for performing complex queries on the MSD relational database remotely.

For protein structure analysis, MSDfold, a protein secondary structure-matching tool, is available as a Web Service. An example client is described at http://www.ebi.ac.uk/msd-srv/docs/api/examples.html.

Soaplab

Soaplab (http://www.ebi.ac.uk/soaplab) is a tool that can automatically generate and deploy Web Services on top of existing command-line analysis programs. It is especially well suited for EMBOSS-type applications. It allows the integration of many applications within a single programming interface. It can also interoperate with other Web Services described earlier (e.g. WSInterProScan) and it can create Web Services on top of existing web resources (e.g. extracting data from a third-party web page and providing its data as a Web Service).

Soaplab in its basic form is a tool for non-programmers who need only to create metadata describing resources (command-line applications, web pages) and let Soaplab generate the rest. The resulting Web Services are uniform and provide a good platform for integration into a workflow such as in Taverna (http://taverna.sourceforge.net). The initial metadata are available from the Soaplab Web Services interface. They make the services self-describing. Soaplab is also a reference implementation of the OMG (Object Management Group, http://www.omg.org/) standard for the Life Sciences Analysis Engine (LSAE).

CONCLUSION

We present here a set of applications that give the user more direct access to data and services from the EBI. From the user's perspective, these are equivalent to installing and maintaining software and databases on local computers. From the programmer's point of view, Web Services provide a robust and flexible environment in which to build applications and provide complex and novel services.

ACKNOWLEDGEMENTS

This work is funded by European Community Contract Nos. QLRI-CT-2001-00015 for 'TEMBLOR' under the specific RTD programme 'Quality of Life and Management of Living Resources' <http://www.cordis.lu/life/> (1998–2002), and IST-2001-32688 for 'ORIEL' under the Shared Cost RTD programme 'Information Society Technologies (2002–2006)'. The Wellcome Trust and core funding from the European Molecular Biology Laboratory (EMBL). Funding to pay the Open Access publication charges for this article was provided by the EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
2. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
3. Senger,M., Rice,P. and Oinn,T. (2003) Soaplab—a unified Sesame door to analysis tools. In Cox,S.J. (ed.), *Proceedings of the UK e-Science All Hands Meeting 2003*, September 2–4, Nottingham, UK, pp. 509–513.
4. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.
5. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.
6. Leinonen,R., Nardone,F., Oyewole,O., Redaschi,N. and Stoehr,P. (2003) The EMBL sequence version archive. *Bioinformatics*, **19**, 1861–1862.
7. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Courne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
10. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.
11. Fredman,D., Munns,G., Rios,D., Sjöholm,F., Siegfried,M., Lenhard,B., Lehväslaiho,H. and Brookes,A.J. (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res.*, **32**, 516–519.
12. Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flatfile data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.
13. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped Blast and Psi-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
15. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.