Published online 13 May 2013 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* ***W597–W600***
doi:10.1093/nar/gkt376

Analysis Tool Web Services from the EMBL-EBI

Hamish McWilliam, Weizhong Li, Mahmut Uludag, Silvano Squizzato, Young Mi Park, Nicola Buso, Andrew Peter Cowley and Rodrigo Lopez*

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

Received January 30, 2013; Revised April 12, 2013; Accepted April 18, 2013

ABSTRACT

Since 2004 the European Bioinformatics Institute (EMBL-EBI) has provided access to a wide range of databases and analysis tools via Web Services interfaces. This comprises services to search across the databases available from the EMBL-EBI and to explore the network of cross-references present in the data (e.g. EB-eye), services to retrieve entry data in various data formats and to access the data in specific fields (e.g. dbfetch), and analysis tool services, for example, sequence similarity search (e.g. FASTA and NCBI BLAST), multiple sequence alignment (e.g. Clustal Omega and MUSCLE), pairwise sequence alignment and protein functional analysis (e.g. InterProScan and Phobius). The REST/SOAP Web Services (http://www.ebi.ac.uk/Tools/webservices/) interfaces to these databases and tools allow their integration into other tools, applications, web sites, pipeline processes and analytical workflows. To get users started using the Web Services, sample clients are provided covering a range of programming languages and popular Web Service tool kits, and a brief guide to Web Services technologies, including a set of tutorials, is available for those wishing to learn more and develop their own clients. Users of the Web Services are informed of improvements and updates via a range of methods.

INTRODUCTION

The European Bioinformatics Institute (EMBL-EBI) provides access to a wide range of biological data resources and bioinformatics applications (1). As well as being available via a web browser, many of these services provide Web Services interfaces based on REST (Representational State Transfer) (2) or SOAP (Simple Object Access Protocol—http://www.w3.org/TR/soap). At present, the BioCatalogue (3), a registry of biological Web Services, lists 724 Web Services provided by EMBL-EBI. The availability of Web Services interfaces allows the integration of data and analysis tools into other tools, applications, web sites, pipeline processes and analytical workflows, while avoiding the need to maintain the databases and programs locally.

Using Web Services, functionality from many databases and analysis tools from a wide range of service providers can be combined to create complex analytical workflows and mashups. This can add value to existing database search or analysis tool results by incorporating data or analysis results from other services. In this article, we describe the current Web Services from the EMBL-EBI for data search, entry retrieval, analysis tools and their use together.

WEB SERVICES

Data search and retrieval

Individual data resources, such as ChEBI (4), ENA (5), Gene Expression Atlas (6) and UniProt (7), provide web interfaces and Web Services (Table 1) tailored to the specifics of their data and the usage patterns required by consumers of their data. These interfaces cater well for searches against the specific resource and address searches and data retrieval operations concerning data from the specific resource; however, this can prove to be an issue when access to multiple data sources is required because each data resource has to be handled separately.

For cases where a more general approach is required, for example, searching across several different databases using one query, the EB-eye (8) Web Service is a possible solution. In addition to searching across the multiple databases available at the EMBL-EBI, the EB-eye Web Service allows navigation through the cross-references network, formed by data entries referencing other entries the same database or in other databases. The EB-eye Web Service can return information held within specified fields of the data entries, and allows searches to be refined by domain or using Boolean logic.

The EB-eye Web Service is limited to retrieving data from selected fields stored in its indexes, so it is complemented by the dbfetch and WSDbfetch (9) services, which

*To whom correspondence should be addressed. Tel: +44 1223 494423; Fax: +44 1223 494468; Email: rodrigo.lopez@ebi.ac.uk

provide single or batch whole entry data retrieval based on entry identifiers. The entry data are often available in a range of data formats, for example, UniProtKB entries are available for retrieval in UniProtKB flat-file format, fasta sequence format, GFF (http://gmod.org/wiki/GFF3), UniProt XML format, UniProt RDF/XML format and SeqXML (http://seqxml.org/0.4/seqxml_doc_v0.4.html).

Analysis tool services

As well as the data search and retrieval services, a range of analysis tool services are also available (Table 2), including sequence similarity search [e.g. FASTA (10) and NCBI BLAST (11)], multiple sequence alignment [e.g. Clustal Omega (12) and MUSCLE (13)], pairwise sequence alignment, protein functional analysis [e.g. InterProScan (14) and Phobius (15)], etc. Most of the analysis tool services are implemented using a job dispatcher framework, JDispatcher (16), which provides a web interface and SOAP & REST Web Services interfaces. The interfaces of these Web Services are largely consistent, easing the learning curve and aiding re-use of supporting services.

Table 1. Data resource-specific Web Services

Topic	Web Services
Genomes	Ensembl BioMart, Ensembl Genomes REST API
Nucleotide sequences	ENA Browser
Protein sequences	PRIDE BioMart, UniProt.org, UniProt BioMart
Small molecules	ChEBI WS, PSICQIC (ChEMBL)
Gene expression	ArrayExpress, Gene Expression Atlas API
Molecular interactions	PSICQIC (IntAct)
Reactions, pathways and diseases	BioModels, PSICQIC (Reactome), Rhea
Protein families	InterPro BioMart
Literature	Europe PMC Web Service
Ontologies	Ontology Lookup Service (OLS), QuickGO, SBO::Web Services, WSMIRIAM

COMBINING SERVICES

Search and data retrieval

The EB-eye and the dbfetch or WSDbfetch Web Services provide a modular approach for performing a database search and retrieving required data in a specified data format. The sample command-line clients provided for these services are capable of being chained together in such a way that these processes can be combined in a single command (Figure 1). For more complex queries or data-retrieval requirements, it may be necessary to use either the sample clients within a script with additional commands to handle any required data transformations, or a client specific to those Web Services, which implements additional logic.

A similar process can be used to combine a sequence similarity search (e.g. FASTA or NCBI BLAST) with data retrieval to automatically pull back the set of hit sequences found for a search. The download functionality provided in the web interface for sequence similarity search results on the EMBL-EBI web site uses this process to obtain the sequences from the dbfetch service and return them to the user.

More advanced workflows and data pipeline processes can be built by combining further analysis tool services. For example, the result from a sequence similarity search can be directly used as input for a multiple sequence alignment, needing only the job identifier to be passed between services in the cases of MView (17) and DbClustal (18).

Web Services in other services

Because Web Services are provided as functionally distinct modules they are well suited for use as components inside web interfaces, where they can provide additional functionality. Some examples are given below.

Sequence similarity search

The sequence similarity search Web Services are used in a number of web sites: the Ensembl Genomes BLAST (19) is provided using the WU-BLAST (20) Web Service;

Table 2. Analysis tool and general data web services

Topic	Web Services
Data retrieval	Dbfetch[a], WSDbfetch[a]
Identifier mapping	PICR[a], UniProt.org ID Mapping[a]
Multi-database search	EB-eye[a]
Multiple sequence alignment	Clustal Omega, ClustalW2, DbClustal, Kalign, MAFFT, MUSCLE, MView, PRANK, T-Coffee
Pairwise sequence alignment	Lalign, EMBOSS tools: matcher, needle, stretcher and water, and the Wise2 tools: GeneWise, PromoterWise and Wise2DBA
Phylogeny	ClustalW2 Phylogeny
Protein functional analysis	InterProScan, Phobius, RADAR
Sequence format conversion	EMBOSS seqret, MView, Readseq
Sequence operations	CENSOR, Seqcksum
Sequence similarity search	FASTA, FASTM, NCBI BLAST, PSI-BLAST, PSI-Search, WU-BLAST
Sequence statistics	SAPS and the EMBOSS tools: pepinfo, pepstats and pepwindow
Sequence translation	EMBOSS tools: backtransambig, backtranseq, sixpack and transeq
Structure analysis	DaliLite, MaxSprout
Text mining	Whatizit[a]

[a]Services not implemented using JDispatcher framework.

```
$ ./EBeyeCliClient.exe --getAllResultsIds uniprot axr3 |
./WSDbfetchCliClient.exe fetchBatch uniprot @- fasta raw
>sp|P93830|IAA17_ARATH Auxin-responsive protein IAA17 OS=Arabidopsis thaliana
GN=IAA17 PE=1 SV=2
MMGSVELNLRETELCLGLPGGDTVAPVTGNKRGFSETVDLKLNLNNEPANKEGSTTHDVV
TFDSKEKSACPKDPAKPPAKAQVVGWPPVRSYRKNVMVSCQKSSGGPEAAAFVKVSMDGA
PYLRKIDLRMYKSYDELSNALSNMFSSFTMGKHGGEEGMIDFMNERKLMDLVNSWDYVPS
YEDKDGDWMLVGDVPWPMFVDTCKRLRLMKGSDAIGLAPRAMEKCKSRA
>tr|A5YXS5|A5YXS5_ARATH Auxin-resistance protein 3 (Fragment) OS=Arabidopsis
thaliana GN=AXR3 PE=4 SV=1
CLGLPGGDTVAPVTGNKRGFSETVDLKLNLNNEPANKEGSTTHDVVTFDSKEKSACPKDP
AKPPAKAQVVGWPPVRSYRKNVMVSCQKSSGGPEAAAFVKVSMDGAPYLRKIDLRMYKSY
DELSNALSNMFSSF
>tr|A5YXT0|A5YXT0_ARATH Auxin-resistance protein 3 (Fragment) OS=Arabidopsis
thaliana GN=AXR3 PE=4 SV=1
CLGLPGGDTVAPVTGNKRGFSETVDLKLNLNNEPANKEGSTTHDVVTFDSKEKSACPKDP
AKPPTKAQVVGWPPVRSYRKNVMVSCQKSSGGPEAAAFVKVSMDGAPYLRKIDLRMYKSY
DELSNALSNMFSSF
>tr|Q56Z92|Q56Z92_ARATH Putative auxin-induced protein, IAA17/AXR3-1
OS=Arabidopsis thaliana GN=At1g04250 PE=4 SV=1
MIDFMNERKLMDLVNSWDYVPSYEDKDGDWMLVGDVPWPMFVDTCKRLRLMKGSDAIGLA
PRAMEKCKSRA
```

Figure 1. Combining the EB-eye and WSDbfetch Web Services to perform a search in UniProtKB for the term 'axr3' and retrieve the corresponding entries in fasta sequence format using the sample .NET clients provided for these services.

the PDBe 'Sequence Search' (http://www.ebi.ac.uk/pdbe/?tab=home&subtab=sequencesearch) uses the FASTA Web Service; and the UniProt.org 'Blast' (http://www.uniprot.org/blast/) uses the NCBI BLAST Web Service.

Text search

The EB-eye Web Service is used to provide the text search functionality for other services at the EMBL-EBI. This is often combined with a query filtering process, which recognizes the input of entry identifiers that the underlying resource can handle, and passes other search terms to the EB-eye Web Service to retrieve a relevant set of identifiers for further processing. The ENA Text Search (http://www.ebi.ac.uk/ena/) uses this process. In Ensembl Genomes, the EB-eye Web Service is used to provide a search facility across the many genome databases in the collection.

Multiple sequence alignment

The 'Align' section (http://www.uniprot.org/align/) of the UniProt.org web site uses the Clustal Omega Web Service to handle the alignment of the input sequences. The webPRANK (21) web interface (http://www.ebi.ac.uk/goldman-srv/webprank/) to the PRANK multiple sequence alignment tool uses the PRANK Web Service provided by JDispatcher to run the analysis.

These Web Services are also being used in desktop applications and as components of workflows implemented within other tools, for example, Blast2GO (http://www.blast2go.com/), BlastStation (http://www.blaststation.com/), Bioclipse (22) and T-COFFEE (23). Additionally published workflow definitions, which consume these services, created using Taverna (24) and other workflow design tools can be found in repositories such as myExperiment (25).

DISCUSSION

The availability of Web Services from the EMBL-EBI allows developers to integrate additional functionality into their programs and web sites without having to worry about maintaining their own copies of the required databases or software involved, or indeed the resources for the storage and execution of the databases and software. This addresses requirements for such functionality, while minimizing the duplication of effort. Reflecting the existing use of bioinformatics tools in general, the modular nature of the Web Services allows users to combine services to create powerful data pipeline and analytical workflows.

At the EMBL-EBI we are seeing the volume and proportion of Web Services traffic continuing to increase. During 2011 the analysis tool services at EMBL-EBI processed ~36 million analysis jobs, of which ~30 million were submitted via the SOAP/REST Web Services interfaces. For 2012 this rose to 50 million jobs of which ~43 million were submitted via the SOAP/REST Web Services interfaces. Regular web browser traffic now comprises <50% of the total web traffic. While much of this change is likely due to the adoption of Web Services in some genome annotation pipelines, this also reflects the usage of Web Services to provide modular functionality as part of applications and in other web sites. Apart from the maintenance of the existing public Web Services, we continue to develop new Web Services and integrate significant data workflows.

We are mindful that Web Services are often used as part of complex pipelines or highly automated processes. These situations present additional requirements regarding the maintenance of interface consistency, etc. Changes to Web Services interfaces and locations can break third-party applications that have grown up around them, and as noted in Schultheis *et al.* (26), there is a range of Web Service persistency and availability from different institutions. At the EMBL-EBI we aim to meet our users' needs for persistency and availability, and we use a range of communication channels (e.g. mailing lists, news feeds and Twitter) to keep users informed of service

changes and availability (see http://www.ebi.ac.uk/Tools/webservices/help/faq).

To aid in the use of our Web Services, we provide significant documentation (e.g. method descriptions, FAQ, tutorials) and sample clients in a range of programming languages (e.g. C#, Java, Perl, PHP, Python, Ruby and VB .NET). During 2013, we plan to further engage with our users to provide further examples of potential data flow combinations in the online documentation. We also provide training courses and helpdesk support for the use of specific Web Services and general help with the development of clients for Web Services available from the EMBL-EBI.

ACKNOWLEDGEMENTS

The authors wish to acknowledge all software developers, database administrators, data curators and users at the EMBL-EBI and elsewhere, who have provided extremely valuable feedback and support throughout.

FUNDING

Funding for open access charge: European Molecular Biology Laboratory (EMBL).

Conflict of interest statement. None declared.

REFERENCES

1. Brooksbank,C., Cameron,G. and Thornton,J. (2010) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, **38**, D17–D25.
2. Fielding,R.T. (2000) Architectural Styles and the Design of Network-based Software Architectures. *Ph.D. Thesis*. University of California, Irvine.
3. Bhagat,J., Tanoh,F., Nzuobontane,E., Laurent,T., Orlowski,J., Roos,M., Wolstencroft,K., Aleksejevs,S., Stevens,R., Pettifer,S. *et al.* (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.*, **38**, W689–W694.
4. de Matos,P., Alcántara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
5. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
6. Kapushesky,M., Adamusiak,T., Burdett,T., Culhane,A., Farne,A., Filippov,A., Holloway,E., Klebanov,A., Kryvych,N., Kurbatova,N. *et al.* (2012) Gene expression atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.
7. The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
8. Valentin,F., Squizzato,S., Goujon,M., McWilliam,H., Paern,J. and Lopez,R. (2010) Fast and efficient searching of biological data resources—using EB-eye. *Brief. Bioinform.*, **11**, 375–384.
9. Pillai,S., Silventoinen,V., Kallio,K., Senger,M., Sobhany,S., Tate,J., Velankar,S., Golovin,A., Henrick,K., Rice,P. *et al.* (2005) SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res.*, **33**, W25–W28.
10. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
11. Altschul,S.F., Madden,T.L., Schäffer,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.
12. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
13. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
14. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
15. Käll,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
16. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–W699.
17. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
18. Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
19. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kähäri,A. *et al.* (2010) Ensembl genomes: extending ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
20. Lopez,R., Silventoinen,V., Robinson,S., Kibria,A. and Gish,W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, **31**, 3795–3798.
21. Löytynoja,A. and Goldman,N. (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**, 579.
22. Spjuth,O., Alvarsson,J., Berg,A., Eklund,M., Kuhn,S., Mäsak,C., Torrance,G., Wagener,J., Willighagen,E.L., Steinbeck,C. *et al.* (2009) Bioclipse 2: a scriptable integration platform for the life sciences. *BMC Bioinformatics*, **10**, 397.
23. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
24. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
25. Goble,C.A., Bhagat,J., Aleksejevs,S., Cruickshank,D., Michaelides,D., Newman,D., Borkum,M., Bechhofer,S., Roos,M., Li,P. *et al.* (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.*, **38**, W677–W682.
26. Schultheis,S.J., Münch,M.-C., Andreeva,G.D. and Rätsch,G. (2011) Persistence and availability of Web services in computational biology. *PLoS One*, **6**, e24914.