W6–W10 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 12 May 2009*
doi:10.1093/nar/gkp302

Web services at the European Bioinformatics Institute-2009

Hamish Mcwilliam*, Franck Valentin, Mickael Goujon, Weizhong Li, Menaka Narayanasamy, Jenny Martin, Teresa Miyar and Rodrigo Lopez

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

Received January 29, 2009; Revised March 30, 2009; Accepted April 16, 2009

ABSTRACT

The European Bioinformatics Institute (EMBL-EBI) has been providing access to mainstream databases and tools in bioinformatics since 1997. In addition to the traditional web form based interfaces, APIs exist for core data resources such as EMBL-Bank, Ensembl, UniProt, InterPro, PDB and ArrayExpress. These APIs are based on Web Services (SOAP/REST) interfaces that allow users to systematically access databases and analytical tools. From the user's point of view, these Web Services provide the same functionality as the browser-based forms. However, using the APIs frees the user from web page constraints and are ideal for the analysis of large batches of data, performing text-mining tasks and the casual or systematic evaluation of mathematical models in regulatory networks. Furthermore, these services are widespread and easy to use; require no prior knowledge of the technology and no more than basic experience in programming. In the following we wish to inform of new and updated services as well as briefly describe planned developments to be made available during the course of 2009–2010.

INTRODUCTION

The European Bioinformatics Institute (EMBL-EBI) provides access to a broad palette of bioinformatics applications and data resources via traditional web-based interfaces, and since 2004 we have also provided programmatic access to these resources using the SOAP (Simple Object Access Protocol—http://www.w3.org/TR/soap/) and REST (Representational State Transfer) (1) Web Services technologies. The use of Web Services for data resources such as Ensembl (2) and UniProt (3), and analytical tools such as BLAST (4) and FASTA (5), allows users to combine applications and data to create analytical workflows which can solve complex problems, as well as allowing systematic analysis of large datasets.

A major advantage of Web Services for the life scientist is that of not having to invest resources in the installation, maintenance and execution of bioinformatics software and the data required by an analysis. Instead these tasks are delegated to the service provider. A disadvantage is the inherent dependency on an Internet connection and in some cases, remoteness, which may slow down the network transactions. Nevertheless, these Web Services are designed to cater efficiently for the user in remote and slow networks and are freely available for use by commercial and academic organisations; see the EMBL-EBI terms of use (http://www.ebi.ac.uk/Information/termsofuse.html).

EMBL-EBI provides SOAP and REST interfaces to a wide range of services for data retrieval and analysis. An overview of these services can be found at http://www.ebi.ac.uk/Tools/webservices/ (Table 1). In the following we describe the current Web Services available from the EMBL-EBI and how these compare with earlier publications (6,7).

SERVICE UPDATES

New data-retrieval services

WSDbfetch is a fundamental service to retrieve database entries given a database name and a set of entry identifiers. Entries can be obtained in a range of data formats and styles suitable for use with common bioinformatics tools. Important developments of this service are: batch retrieval (fetchBatch) and the ability to obtain format and style names for a given database (getDBformats and getFormatStyles). In addition to existing database resources such as EMBL-Bank (8), UniProt, InterPro (9) and PDB (10), new data sources include: Ensembl and InterProMatches for UniProtKB and UniParc.

Many analysis workflows need a method to map identifiers between the sequence databases. This is particularly important to collate annotation from multiple data

*To whom correspondence should be addressed. Tel: +44 1223 492508; Fax: +44 1223 494468; Email: hpm@ebi.ac.uk

Table 1. Web Services available from the European Bioinformatics Institute

Application	Web services
Data retrieval	WSDbfetch, ChEBI WS, Integr8 WS, MSD API, MartService, EB-eye, ArrayExpress, IntAct, SRS, QuickGO, UniProt JAPI.
Analysis tools	InterProScan, EMBOSS, CENSOR, Phobius, Soaplab,
Similarity searches	FASTA, WU-BLAST, NCBI BLAST, PSI-BLAST, PHI-BLAST, PSI-Search, MPSRCH, SCANPS.
Multiple sequence alignments	ClustalW 2, Kalign, MAFFT, MUSCLE, T-Coffee.
Structural analysis	DaliLite, MaxSprout, MSDFold (SSM).
Literature and ontologies	CiteXplore, Whatizit, OLS, WSSBO, MIRIAM, PICR, BioModels.

sources, where the identifiers used for a particular sequence may differ (e.g. NCBI gi's and INSDC (http://www.insdc.org) identifiers). A service which addresses this requirement, for protein sequences, is the Protein Identifier Cross-Reference service (PICR) (11).

The 'EB-eye' offers a simple and efficient way to search more than 360 million entries, from more than 62 databases, including the EMBL-EBI web portal. The web interface caters for all types of users, from novices to experienced researchers, regardless of if they are already familiar with the EMBL-EBI databases. A simple search box is available at the top of every EMBL-EBI web page. The results of a search are first presented as a summary page where the databases are organised into several categories representing knowledge domains currently maintained at EMBL-EBI. The number of entries matching the search terms is displayed for each domain and its member databases. This gives an overview of the results covered by all the data resources. The user then has the choice of selecting a category (or a database) for which to view the corresponding search results. This leads to a new page where results are displayed as a list of entries with short descriptions and links to more information, such as the database's main web site, alternative views of the data and cross-references within EB-eye. Details of how to use the EB-eye can be found in the dedicated help pages at: http://www.ebi.ac.uk/inc/help/search_help.html.

The EB-eye provides a SOAP Web Service (http://www.ebi.ac.uk/Tools/webservices/services/eb-eye/) covering the functionality available in the web interface. A Web Services client can query a database indexed in the EB-eye, retrieve the results and navigate the cross-references network with just a few service calls. Features of the EB-eye, like the text-based search or the cross-reference network navigation, have proven to be useful for the development of database annotation tools, and also in the maintenance of cross-references between databanks.

New analysis tools

Complementing the existing transmembrane and signal prediction methods in the InterProScan (12) service, is the new Phobius (13) service. Phobius provides an integrated method for predicting both transmembrane domains and signal peptide cleavage sites. In addition to providing the location of the predicted features, an optional graphical output is available that shows prediction scoring across the whole protein.

The new CENSOR (14) service identifies both simple and complex sequence repeats (low-complexity regions and complex repeats, such as transposable elements and retroviral inserts) in nucleotide or protein sequences. As well as providing the identification and coordinates for each repeat, the sequence is returned with the repeats masked for use as input with sequence library searching programs such as BLAST or FASTA.

The Soaplab (15) service at EMBL-EBI, which provides access to the EMBOSS suite of analysis tools (16), has been upgraded to use Soaplab2 (http://soaplab.sourceforge.net/soaplab2/). The EMBL-EBI Soaplab service has also been updated to use EMBOSS version 6.0.1 and additional databases have been added to the EMBOSS installation, including, Ensembl, ASTD (17), IntAct (18) sequences, RefSeq (19), PDB and LGICdb (20).

New similarity searches and databanks

In addition to regular maintenance and bug fixes to existing services, such as NCBI BLAST and FASTA, IntAct sequences, Korean Intellectual Property Office (KIPO) protein sequences and EMBL-Bank sliced by methodological classes and taxonomic divisions, are now available. The latter makes it easier to perform focused searches, such as those involved in the characterisation of sequences, including those from novel metagenomics datasets and identification of sequences of the same origin across multiple metagenomic studies.

A significant application update is the introduction of SSEARCH, an accelerated version of Smith and Waterman algorithm (21) from the FASTA package, which provides a more sensitive alternative to BLAST or FASTA and proves to have better performance and scalability than the, equivalent, MPSrch service. In this context, the reader should note that plans are in place to phase out the MPSrch service during the course of 2009–2010,

PSI-Search integrates the Position Specific Iterative strategy for searching distantly related sequences in biologically heterogeneous libraries. This service has been implemented using PSI-BLAST (22) and SSEARCH. In PSI-Search these provide complementary methods for elucidating distant relationships between sequences of distinct origin, but return results that differ in the scoring scheme used in the alignments. PSI-Search will

potentially find extended families of sequences in fewer iterations than the traditional PSI-BLAST approach.

New multiple sequence alignments (MSA) services

The MSA services provide a range of methods that cater for diverse requirements relating to nucleotide and protein sequence alignments. CLUSTAL W (23) is the main workhorse for many users and has been updated to CLUSTAL W version 2 (24). This version of CLUSTAL provides iterative refinement of the alignment and the ability to cluster sequences using UPGMA, which improves performance with large numbers of sequences. T-COFFEE (25), MUSCLE (26), MAFFT (27) and Kalign (28) have all been updated with fixes that provide better memory management and support for additional output formats. MAFFT and Kalign are particularly well suited for large genomic-type alignments where other methods such as CLUSTAL, MUSCLE and T-COFFEE can exhibit memory and performance issues.

New structural analysis services

During 2008, the DALI (29) service was phased out, and has been replaced by PDBeFold (30) (formerly MSDFold), a service for comparing protein structures in 3D. PDBeFold outperforms other similar services by several orders of magnitude in terms of speed and accuracy of results (http://www.ebi.ac.uk/msd-srv/ssm/comparisons/cmp_conclusion.html).

New services in literature and ontologies

Most biological knowledge remains in the scientific literature; integration of electronic literature resources with biological databases and bioinformatics tools is therefore an important part of the EMBL-EBI's role. The addition of the National Agricultural Library Catalogue (AGRICOLA) (http://agricola.nal.usda.gov) database to the CiteXplore (http://www.ebi.ac.uk/citexplore) service supplements the PubMed/MEDLINE (31) data to fill a gap in the coverage of environmental and agricultural science literature. A new feature of the CiteXplore service is the ability to display, for each result, a network of references which cite the article as well as those cited by the article. This is very useful for identifying related literature as well as evaluating the influence of an article.

Ontologies are critical to the efficient exchange of information about biological entities. Many have been devised to describe various biological properties, such as structure, function and taxonomic classification. The Ontology Lookup Service (OLS) (32) presently contains 61 different ontologies, including those related to environmental, chemical, functional and physiological knowledge domains. A complete list can be found at: http://www.ebi.ac.uk/ontology-lookup/ontologyList.do.

Systems biology integrates information from many specialised resources to describe processes in living organisms. The integration of disparate data sources requires standardisation of terms so equivalent semantics can be ascribed. As such, systems biology is a major consumer of ontologies and other structured information. It is also a heavy user as well as provider of Web Services. Examples of EMBL-EBI services in this area include: BioModels (33), the Systems Biology Ontology (http://www.ebi.ac.uk/sbo) and Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) (34), which can be accessed over the web as well as over Web Services APIs.

COMBINING WEB SERVICES

As mentioned earlier, all the services that EMBL-EBI provides are available over traditional browser-based web interfaces, which impose limitations for the analysis of large datasets. However, almost all services provide a programmatic interface that uses Web Services technologies, enabling these services to be used as components in workflows or analytical pipe-lines.

An example that caters well for the analysis of novel types of data, such as in metagenomics, involves running a similarity search against a nucleotide or protein database, identifying close homologues from the results, extracting identifiers in order to obtain complete sequences—potentially with annotation, and using these to create a phylogenetic reconstruction. In practice, this involves running a BLAST search against UniProtKB using the WSWUBlast service and obtaining the ordered list of homologous identifiers from it. These identifiers can then be used to retrieve complete sequences using the WSDbfetch service. These sequences can be used as input to an MSA service, for example: WSClustalW2 or WSMuscle. The resulting alignment can be used as input to protpars or protdist from the PHYLIP (35) package, using the Soaplab services, to complete the phylogenetic reconstruction.

Many other providers of bioinformatic tools and databases have also adopted Web Services technologies. These include: the National Centre of Biotechnology Information (NCBI), DNA DataBank of Japan (DDBJ), Kyoto Encyclopedia of Gene and Genomes (KEGG) and EMBRACE (http://www.embracegrid.org). Given the widespread nature and diversity of these services, users can mix-and-match services as they require. Workbench environments such as Taverna (36) make the rapid development of complex workflows combining diverse services simpler, and provide a useful way to document and share a workflow design.

DISCOVERING WEB SERVICES

Web Services are difficult to find using web search engines, e.g. Google, Yahoo and Live/MSN. Several efforts exist that address this issue, for example: seekda (http://www.seekda.com), BioCatalogue (http://www.biocatalogue.org), the EMBRACE Service Registry (http://www.embraceregistry.org), BioMoby (http://www.biomoby.org) and the DAS Registry (http://www.dasregistry.org). These projects provide extensive searching facilities through their respective portals and encourage the community to submit annotations and additional information about their use.

DOCUMENTATION AND HELP

EMBL-EBI strives to provide extensive documentation and support for users of both web based services and Web Services. For each Web Service, pages exist that describe the service, document the API, provide links to service end-points as well as to sample clients in a variety of programming languages commonly used in bioinformatics to help users getting started. In addition to service specific documentation, tutorials on using Web Services in a range of programming languages, including: C#, Java, Perl, Python, PHP, Ruby and Visual Basic.NET, are provided (http://www.ebi.ac.uk/Tools/webservices/tutorials/intro). User support is available and readers are encouraged to contact EMBL-EBI with problems, comments and suggestions using: http://www.ebi.ac.uk/support/.

FUTURE PLANS

An impediment to using many bioinformatics tools effectively is the lack of detailed information about the parameters, their valid values and the interrelationships between them. To address this issue extensive metadata is to be added to the services. This includes adding operation and data structure documentation to the service description documents, and providing methods for obtaining detailed documentation about the meaning and use of parameter values and their constraints.

In accordance with Web Services best practises, the remaining RPC/encoded services will be migrated to the WS-I (http://www.ws-i.org/) recommended Document/literal style. Also, the REST-style services will be updated to provide service description documents in accordance to the Web Application Description Language (WADL—https://wadl.dev.java.net/) specification to allow automated utilisation of these services.

CONCLUSION

The adoption of Web Services technologies at the EMBL-EBI has had significant implications for the investment in human as well as computational resources. During 2008 the number of computational jobs, such as BLAST, FASTA, InterProScan, etc., exceeded 12 million, more than half of which were using the Web Services APIs. This indicates significant adoption of the technologies by the user community. EMBL-EBI continues to strive to provide access to new services that cater for emergent technologies, such as high throughput sequencing, as well as ensuring that existing services remain well maintained and up-to-date.

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

European Union (contract number 021902 as part of the FELICS Research Infrastructure; contract number LHSG-CT-2004-12092 as part of the EMBRACE project; and contract number IST-2001-32688 as part of the ORIEL Project), the Wellcome Trust; the European Patent Office; the National Institutes of Health (as part of the UniProt project, grant 1 U01 HG02712-01); and core funding from the European Molecular Biology Laboratory (EMBL). Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Fielding,R.T. (2000) Architectural Styles and the Design of Network-based Software Architectures. Ph.D. Thesis, UC Irvine.
2. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*., **37**, D690–697.
3. The UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*., **37**, D169–D174.
4. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
5. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *PNAS*, **85**, 2444–2448.
6. Labarga,A., Valentin,F., Anderson,M. and Lopez,R. (2007) Web services at the European Bioinformatics Institute. *Nucleic Acids Res*., **35**, W6–W11.
7. 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.
8. Cochrane,G., Akhtar,R., Bonfield,J., Bower,L., Demiralp,F., Faruque,N., Gibson,R., Hoad,G., Hubbard,T., Hunter,C. *et al*. (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res*., **37**, D19–D25.
9. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,D., Daugherty,L., Duquenne,L. *et al*. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*., **37**, D211–D215.
10. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
11. Côté,R.G., Jones,P., Martens,L., Kerrien,S., Reisinger,F., Lin,Q., Leinonen,R., Apweiler,R. and Hermjakob,H. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinform*., **8**, 401.
12. 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.
13. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol*., **338**, 1027–1036.
14. Kohany,O., Gentles,A.J., Hankus,L. and Jurka,J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and CENSOR. *BMC Bioinform*., **7**, 474.
15. Senger,M., Rice,P. and Oinn,T. (2003) Soaplab—a unified Sesame door to analysis tools. In Cox,S.J. (ed.), *Proceedings, UK e-Science, All Hands Meeting, 2–4 September*, Nottingham, UK, pp. 509–513.
16. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
17. Koscielny,G., Texier,V.L., Gopalakrishnan,C., Kumanduri,V., Riethoven,J.J., Nardone,F., Stanley,E., Fallsehr,C., Hofmann,O., Kull,M. *et al*. (2008) ASTD: The Alternative Splicing and

Transcript Diversity database. *Genomics*, doi: 10.1016/j.ygeno.2008.11.003.
18. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al*. (2007) IntAct-open source resource for molecular interaction data. *Nucleic Acids Res*., **35**, D561–D565.
19. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.
20. Donizelli,M., Djite,M-A. and Le Novère,N. (2006) LGICdb: a manually curated sequence database after the genomes. *Nucleic Acids Res*., **34**, D267–D269.
21. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
22. 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** (17), 3389–3402.
23. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
24. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al*. (2007) Clustal W and Clustal X version 2. *Bioinformatics*, **23**, 2947–2948.
25. Notredame,C., Higgins,D. and Heringa,J. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol*., **302**, 205–217.
26. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform*., **5**, 113.
27. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics*, **9**, 286–298.
28. Lassmann,T. and Sonnhammer,E.L.L. (2005) Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinform*., **6**, 298.
29. Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res*., **25**, 231–234.
30. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst*., **D60**, 2256–2268.
31. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **37**, D5–D15.
32. Côté,R.G., Jones,P., Martens,L., Apweiler,R. and Hermjakob,H. (2008) The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res*., **36**, W372–W376.
33. Le Novère,N., Bornstein,B., Broicher,A., Courtot,M., Donizelli,M., Dharuri,H., Li,L., Sauro,H., Schilstra,M., Shapiro,B. *et al*. (2006) BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*., **34**, D689–D691.
34. Le Novère,N., Finney,A., Hucka,M., Bhalla,U.S., Campagne,F., Collado-Vides,J., Crampin,E.J., Halstead,M., Klipp,E., Mendes,P. *et al*. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol*., **23**, 1509–1515.
35. Felsenstein,J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
36. 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.