Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W729–W732***
doi:10.1093/nar/gkl320

Taverna: a tool for building and running workflows of services

Duncan Hull*, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R. Pocock[1], Peter Li[2] and Tom Oinn[3]

School of Computer Science, University of Manchester, M13 9PL, UK, [1]School of Computing Science, University of Newcastle, NE1 7RU, UK, [2]School of Chemistry, University of Manchester, M60 1QD, UK and [3]EMBL European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Received February 14, 2006; Revised March 1, 2006; Accepted April 13, 2006

ABSTRACT

Taverna is an application that eases the use and integration of the growing number of molecular biology tools and databases available on the web, especially web services. It allows bioinformaticians to construct workflows or pipelines of services to perform a range of different analyses, such as sequence analysis and genome annotation. These high-level workflows can integrate many different resources into a single analysis. Taverna is available freely under the terms of the GNU Lesser General Public License (LGPL) from http://taverna.sourceforge.net/.

INTRODUCTION

The number of applications and databases providing tools to perform computations on DNA, RNA and proteins are rapidly growing. However, the lack of communication between such tools in molecular biology is commonly a barrier to extracting new knowledge using these resources. Many tools and databases already communicate using the web, as shown by the ever-growing list of servers in this issue of *Nucleic Acids Research*.

Currently, integrating tools and databases available on the web frequently involves either 'screen-scraping' web pages using scripting languages like PERL or manual cut-and-paste of data between applications. Each of these methods has its problems. Screen-scraping is notoriously fragile, because the integrating script is prone to break when the web page or form changes, and for this reason has been likened to 'medieval torture' (1). Cutting and pasting data between applications is another common way to quickly achieve interoperation. However, cut-and-paste procedures are laborious to repeat and verify.

Web services technology provides some solutions for improving this situation. In addition to providing form-based interfaces, tool and database providers can describe their application or database using the standard Web Services Description Language (WSDL). These WSDL descriptions can then be indexed to build a searchable and browsable registry of operations for end-users. Applications can then exchange data, typically using SOAP, a protocol for exchanging XML-based messages over a network, normally using HTTP. For a full description of web service technology, languages and protocols see (2). Using web services has several advantages:

- Tools and databases do not need to be installed locally on the users machine or laboratory server, as they are programmatically accessible over the web.
- Tools created using different programming languages (e.g. Python, PERL, Java, etc.) and platforms (e.g. Unix, Windows, etc.) can be accessed through the same web service interface. This removes the need for the user to know about all the different platforms and programming languages underneath.
- The need for fragile screen-scraping integration scripts is reduced.
- It provides an alternative to time-consuming and laborious 'cut-and-paste' integration between web applications.
- Workflows, or pipelines, of web services can be built to provide high-level descriptions of analyses. These can be created and tested relatively quickly to integrate many different tools and applications in a single analysis

However, there are also several limitations of using web services:

- Since services are provided by autonomous third-parties around the world, they frequently have insufficient or non-existent metadata. Where metadata exists it often provides little indication of the purpose of a service. So for example, inputs can have cryptic names like '`in1`' with

*To whom correspondence should be addressed. Tel: +44 0 161 275 0677; Fax: +44 0 161 275 6204; Email: duncan.hull@cs.man.ac.uk

a datatype of '`string`' which hide complex legacy flat-file formats, and have no immediately obvious function. In the worst case, the only way to work out what task a service performs is to invoke it with some data and examine what comes back from the service. Invoking services relies on knowing exactly what data a service takes as input, information which is not always available. An important consequence of poor service metadata is that many services can be difficult to find in a registry (3).

- Joining services together into pipelines is frequently problematic, as the inputs and outputs are not directly compatible. Consequently, many one-off 'shim' services (4) are required to align closely-related data and enable services to interoperate.
- The web services stack (2) can be difficult to debug. Standard open-source libraries that Taverna uses for creating, documenting and invoking services like WSIF (http://ws.apache.org/wsif/), WSDL4J and Axis (http://ws.apache.org/axis/) can provide poor documentation by default, and cryptic error messages when services fail.
- Services accessed over a network can have unpredictable performance and reliability (http://www.java.net/jag/Fallacies.html). Some services, particularly the more specialist and obscure tools provided by smaller laboratories, can be unreliable, unstable or have licensing issues. Such services are often the 'weakest link' in the chain. When individual services fail, for whatever reason, the whole workflow can not be run. Mirrored replica or redundant services are not always available to address this problem through failover.

Working with both these strengths and limitations, Taverna (5,6), part of the *my*Grid project, is an application that makes building and executing workflows accessible to bioinformaticians who are not necessarily experts in web services and programming. It provides a single point of access to a range of services with programmatic interfaces, primarly web services. As of March 2006, there are around 3000 of these publicly available services in molecular biology, provided by range of third-parties around the world. The potential set of services accessible in Taverna is even larger, as more tool and database providers expose programmatic interfaces to their resources over the web. Currently, building workflows of these services in Taverna, allows users glue these diverse resources together relatively quickly. This can allow rapid exploration of data of hypothesis testing, e.g. on given gene(s) or protein(s).

SERVICES AND WORKFLOWS TAVERNA

There are a wide range of services available in Taverna, first those provided by INSDC (http://www.insdc.org/) member organizations, EMBL-EBI provides standard services (7), the NCBI Entrez Programming Utilities (NCBI services can only be used in Taverna version 1.3.2-RC1 or later) (8) and the DNA Databank of Japan (DDBJ) (9). Additional tools and databases are provided by the Protein Databank of Japan (PDBJ) (10), Kyoto Encyclopedia of Genes and Genomes (KEGG) (11), BioMART (12), PathPort/ToolBus tools (13), BioMOBY (14), BIND (15), SeqVista (16) and Pfam (17) from the Wellcome Trust Sanger Institute. A more comprehensive list and description of the services available can be found at the sourceforge website (http://taverna.sourceforge.net/index.php?doc=services.html). An important feature of Taverna is that it can talk to many different kinds of service, so for example, different services can be added to the services panel.

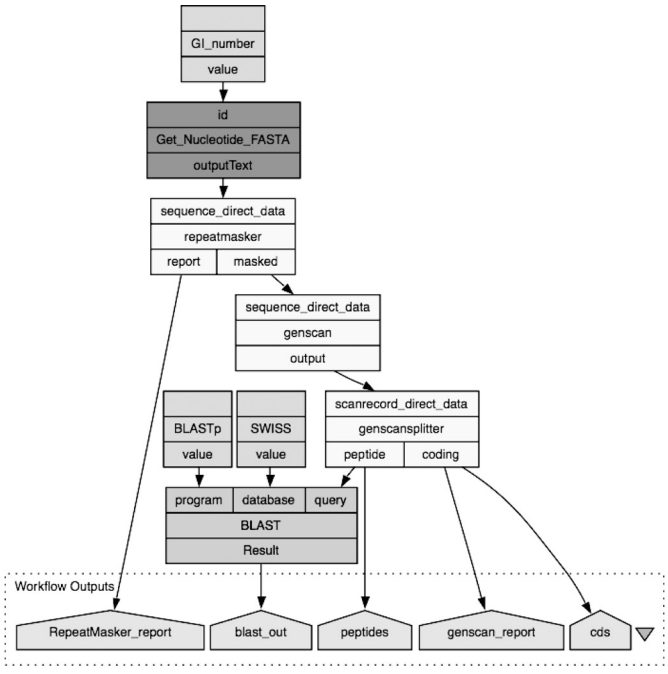

Figure 1. A workflow of services for analysing a draft DNA sequence from GenBank.

A workflow built from some of the services described above, which illustrates the capabilities of Taverna is shown in Figure 1. This workflow starts with an input GenBank identifier (GI number), to retrieve a draft DNA sequence which is then fed into RepeatMasker (http://www.repeatmasker.org/) then GenScan (http://genes.mit.edu/GENSCAN.html.) (18) to predict the location of any genes in the sequence. The report output from GenScan is split, and the part containing the peptide sequence is fed into BLASTp, hosted by the DDBJ. Although not shown in this workflow, the results of the BLAST analysis could be fed into further programs, provided that the user knows how to parse BLAST records and what services could follow. As the services have very little metadata, Taverna cannot currently guide the user during workflow construction. The workflow shown here is a basic gene prediction and characterization pipeline that is part of many workflows created in Taverna, e.g. workflows used inresearch of Williams–Beuren syndrome (19) and Graves disease (20).

RUNNING WORKFLOWS

The workflow shown in Figure 1 can be downloaded from the *my*Grid workflow repository (http://workflows.mygrid.org.uk/repository/narweb.xml.) In order to run this worklfow, which takes >5 min to execute, download Taverna and

consult the user documentation (http://taverna.sourceforge.net/usermanual/docs.word.html) under the heading 'Enacting a predefined workflow'. Other pre-defined workflows can be run by browsing the workflow repository or examples directory of Taverna. Alternatively, arbitrary workflows can be constructed using the services described above, again see the user documentation for details. Each Taverna workflow can have metadata stored inside it using the author and title tags. Additional workflow metadata can be stored seperately from the workflow and identified using a Life Science Identifier (LSID) (21). All workflows have an LSID by default, although the user has to assign metadata to this LSID if they require it.

TAVERNA USERS AND FUTURE WORK

The current version of Taverna, (1.x) has been downloaded around 14 000 (http://taverna.sourcesforge.net/index.php?doc=stats.php) times and has an estimated user base of around 1500 installations. Taverna has been used several different areas of research throught Europe, Asia, Australia and the USA for functional genomics (19,20), metabolic and signalling pathway analysis (5) and chemoinformatics (22).

Based on the experiences of these users, requirements have been gathered for the next release of Taverna, version 2.0. This version is currently being developed and is scheduled for release in 2007. Planned new features include the ability to support higher-throughput and longer-running workflows using Grid technology, a semantically enabled registry with services annotated with terms from a standard ontology, facilities for provenance gathering and a repository of workflows that can be re-used and re-purposed. Taverna 2.0 will also have enhanced results browsing with the ability to incrementally execute workflows and use microarray tools like maxD (23) and the R library (24).

CONCLUSIONS

We present here an application, Taverna, that allow users who are not necessarily expert programmers to design, execute and share workflows of web services. These workflows can be used to perform a range of different analyses in molecular biology and bioinformatics, accessing numerous different databases and tools using standard web protocols.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the rest of the *my*Grid research and development team as well as the early-adopters of the Taverna workbench: Pinar Alper, Andy Brass, Justin Ferris, Paul Fisher, Matthew Gamble, Claire Jennings, Doug Kell, Antoon Goderis, Stuart Owen, Simon Pearce, Martin Senger, Stian Soiland, May Tassabehji, Hannah Tipney, Daniele Turi, Anil Wipat, David Withers, Chris Wroe and Jun Zhao. The authors would also like to thank project partners BioMOBY(Mark Wilkinson), SeqHound and BioMART (Arek Kasprzyk); Industrial partners IBM (Dennis Quan, Sean Martin, Mike Niemi), Sun Microsystems, Cerebra Inc., GlaxoSmithKline, AstraZeneca, Merck KgaA, genetic Xchange and Epistemics Ltd. The development of Taverna has been supported by UK e-Science programme and the Open Middleware Infrastructure Insitute (OMII). Both of these are funded by the Engineering and Physical Sciences Research Council (EPSRC), grant references GR/R67743/01 and EP/D044324/1. Funding to pay the Open Access publication charges for this article was provided by EPSRC grant reference EP/D044324/1.

Conflict of interest statement. None declared.

REFERENCES

1. Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
2. Alonso,G., Casati,F., Kuno,H. and Machiraju,V. (2004) *Web Services: concepts, Architectures and Applications. Data-Centric Systems and Applications*. Springer-Verlag, Berlin and Heidelberg GmBH.
3. Hull,D., Stevens,R. and Lord,P. (2005) Describing Web Services for user-oriented retrieval. *W3C Workshop on Frameworks for Semantics in Web Services, DERI*. Innsbruck, Austria.
4. Hull,D., Stevens,R., Lord,P., Wroe,C. and Goble,C. (2004) Treating shimantic web syndrome with ontologies. In *Proceedings of First Advanced Knowledge Technologies Workshop on Semantic Web Services (AKT-SWS04) KMi*. The Open University, Milton Keynes, UK.
5. Oinn,T., Addis,M., Ferris,J., Marvin,D., Greenwood,M., Carver,T., Pocock,M.R., Wipat,A. and Li,P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.
6. Oinn,T., Greenwood,M., Addis,M., Ferris,J., Glover,K., Goble,C., Goderis,A., Hull,D., Marvin,D., Li,P. *et al*. (2005) Taverna: Lessons in creating a workflow environment for the life sciences. *Concurr. Comput.: Pract. Exp*. In press.
7. Pillai,S., Silventoinen,V., Kallio,K., Senger,M., Sobhany,S., Tate,J., Valenkar,S., Golovin,A., Henrick,K., Rice,P., Stoehr,P. and Lopez,R. (2005) SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res*., **33**, W25–W28.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S., Geer,L.Y. *et al*. (2006) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **34**, 173–180.
9. Miyazaki,S., Sugawara,H., Ikeo,K., Gojobori,T. and Tateno,Y. (2004) DDBJ in the stream of various biological data. *Nucleic Acids Res*., **32**, 31–34.
10. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
11. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Rese*., **34**, 354–357.
12. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
13. Eckart,J.D. and Sobral,B.W. (2003) A life scientist's gateway to distributed data management and computing: the pathport/toolbus framework. *OMICS*, **7**, 79–88.
14. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet Exemplar Case. *Plant Physiol*., **138**, 5–17.
15. Bader,G.D., Betel,D. and Hogue,C.W. (2003) Bind: the biomolecular interaction network database. *Nucleic Acids Res*., **31**, 248–250.
16. Hu,Z., Fu,Y., Halees,A.S., Kielbasa,S.M. and Weng,Z. (2004) Seqvista: a new module of integrated computational tools for studying transcriptional regulation. *Nucleic Acids Res*., **32**, 235–241.
17. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R., Eddy,S.R., Sonnhammer,E.L. and Bateman,A. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.

18. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
19. Stevens,R.D., Tipney,H.J., Wroe,C., Oinn,T., Senger,M., Lord,P.W., Goble,C.A., Brass,A. and Tassabehji,M. (2004) Exploring Williams-Beuren Syndrome Using myGrid. *Bioinformatics*, **20**, i303–i310.
20. Li,P., Hawyward,K., Jennings,C., Owen,K., Oinn,T., Stevens,R., Pearce,S. and Wipat,A. (2004) Association of variations in I kappa B-epsilon with Graves' disease using classical myGrid methodologies. *Proceedings UK e-Science programme All Hands Meeting*, Nottingham, UK, 832–839.
21. Clark,T., Martin,S. and Liefeld,T. (2004) Globally distributed object identification for biological knowledgebases. *Brief Bioinform.*, **5**, 59–70.
22. Wolstencroft,K., Oinn,T., Goble,C., Ferris,J., Wroe,C., Lord,P., Glover,K. and Stevens,R. (2005) Panoply of utilities in taverna. *First International Conference on e-Science and Grid Computing (e-Science'05)*, Melbourne, Australia, 156–162.
23. Hancock,D., Wilson,M., Velarde,G., Morrison,N., Hayes,A., Hulme,H., Wood,A.J., Nashar,K., Kell,D.B. and Brass,A. (2005) maxdload2 and maxdbrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination. *BMC Bioinformatics*, **6**, 264–264.
24. Ihaka,R. and Gentleman,R. (1996) R: A language for data analysis and graphics. *J. Compu. Graph. Statistics*, **5**, 299–314.