BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 21 2012, pages 2858–2859
doi:10.1093/bioinformatics/bts530

Databases and ontologies Advance Access publication September 3, 2012

Updating annotations with the distributed annotation system and the automated sequence annotation pipeline

William Speier[1,2,3,*] and Michael F. Ochs[3]

[1]Medical Imaging Informatics Group, University of California, Los Angeles, CA, [2]Department of Bioengineering, University of California, Los Angeles, CA, and [3]Department of Oncology and Division of Oncology Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Associate Editor: Janet Kelso

ABSTRACT

Summary: The integration between BioDAS ProServer and Automated Sequence Annotation Pipeline (ASAP) provides an interface for querying diverse annotation sources, chaining and linking results, and standardizing the output using the Distributed Annotation System (DAS) protocol. This interface allows pipeline plans in ASAP to be integrated into any system using HTTP and also allows the information returned by ASAP to be included in the DAS registry for use in any DAS-aware system. Three example implementations have been developed: the first accesses TRANSFAC information to automatically create gene sets for the Coordinated Gene Activity in Pattern Sets (CoGAPS) algorithm; the second integrates annotations from multiple array platforms and provides unified annotations in an R environment; and the third wraps the UniProt database for integration with the SPICE DAS client.

Availability: Source code for ASAP 2.7 and the DAS 1.6 interface is available under the GNU public license. Proserver 2.20 is free software available from SourceForge. Scripts for installation and configuration on Linux are provided at our website: http://www.rits.onc.jhmi.edu/dbb/custom/A6/

Contact: Speier@mii.ucla.edu or mfo@jhu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 12, 2012; revised on July 25, 2012; accepted on August 23, 2012

1 INTRODUCTION

High-throughput methods in the biological domain have generated an abundance of data, providing opportunities for mining and learning algorithms. The full potential of these data is not currently realized as heterogeneity in the content, format, and availability makes integration a difficult task. The Automated Sequence Annotation Pipeline (ASAP) was developed to link diverse annotation sources and integrate annotations across platforms (Kossenkov *et al*., 2003). The ASAP system wraps existing web resources in a consistent format and allows users to query these data sources either by creating plans using Perl or by invoking existing plans.

Programmatic interaction with ASAP can be difficult because it is designed as a standalone application. Although the system allows client programs to invoke plans, it requires custom Perl scripts to be written and does not standardize the output format across different annotation resources. Here we present a system allowing client side integration through HTTP using the Distributed Annotation System (DAS) (Dowell *et al*., 2001). The results are formatted according to the DAS 1.6 XML specification to standardize the output. Through this integration, annotations can be automatically accessed by any DAS-compatible system, and annotations can be integrated in a consistent manner across different molecular platforms.

2 METHODS

The DAS integration with ASAP runs as a plugin to ProServer 2.20 (Finn *et al*., 2007). Queries are entered as a formatted URL request through HTTP, which are then parsed and converted into ASAP commands. The plugin then connects to the ASAP installation where the specified annotation jobs are created and executed. This allows users and programs to bypass the graphical interface and run ASAP plans in a simple, consistent manner.

After the annotation plan finishes, the main output is encoded into XML using the DAS protocol. In order to preserve the flexibility of the ASAP system, output files generated by plans are saved on the server and links to the files are included in the DAS XML file. Within ASAP, each ProServer call is treated as a regular ASAP query. Thus, it will keep records of each query, store result files for later use and send email notifications as specified by the ASAP configuration.

3 IMPLEMENTATION

Three example applications are described here to show the system's utility and versatility. A public demonstration server is available at: http://new_bauhaus.onc.jhmi.edu:8080/das/ASAP/

3.1 CoGAPS

The ASAP ProServer integration was implemented in an extension of the Coordinated Gene Activity in Pattern Sets (CoGAPS) algorithm (Fertig *et al*., 2010). CoGAPS finds patterns in microarray data through matrix decomposition and then infers the activity of a pattern by finding the probability that genes are overrepresented.

The original version of CoGAPS requires users to manually specify the gene sets to test. This implementation does this automatically by grouping the input genes based on common transcriptional regulators in the TRANSFAC database (Matys *et al*., 2003). When a user runs the extended CoGAPS implementation

*To whom correspondence should be addressed.

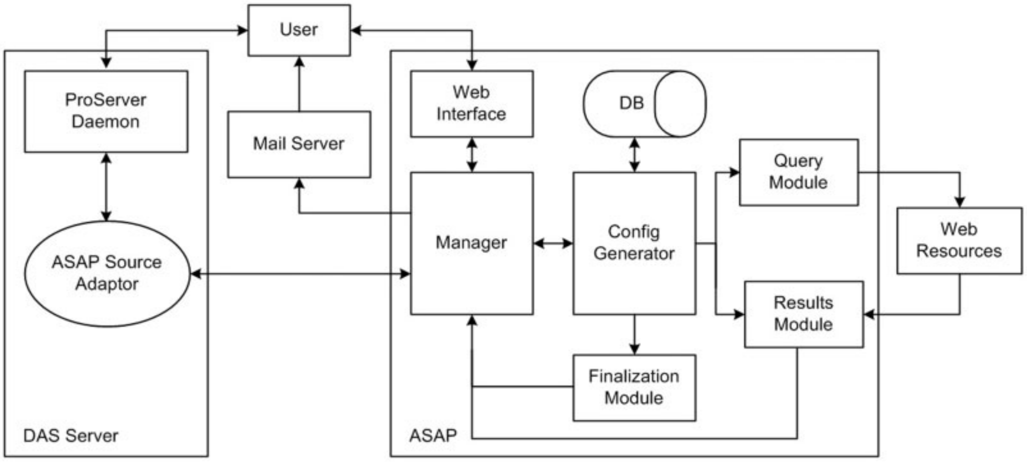

Fig. 1. A schematic representation of ASAP and DAS integration. The user sends an HTTP request to the ProServer daemon, which is then relayed to the ASAP manager where a job is created. When the job is complete, the results are formatted in XML according to the DAS protocol and returned through the ProServer daemon. Additionally, ASAP sends email notifications and saves the results on the server to be accessed later via the web interface

in R, an ASAP query is generated containing the input genes. The resulting XML file containing the information from the regulators and their targets is then parsed to create the sets. Although designed for integration with CoGAPS, this approach is useful for gene sets in general.

3.2 Unified gene annotations

To demonstrate the system's ability to integrate multiple annotations, we created the Unified Gene Annotations plan. This plan annotates four different microarray platforms: Affymetrix, Agilent, Illumina and alternative Affymetrix annotations (Yu *et al*., 2007). This provides a unified view of annotations across different platforms improving consistency as it does not rely on the individual annotations built into Bioconductor packages, which can be unsynchronized in terms of genome build and specific gene annotations. In this example, the Gene Ontology and pathway annotations are linked across platforms and the output is combined into an R environment object.

Since the output of this plan is in an R object, the DAS integration does not attempt to reformat it. Instead, the output file is saved on the server and DAS returns a link imbedded in the usual XML format. Users and interfacing programs can then parse the XML to retrieve the link and download the file. The ASAP server will also store and manage output files so users can return to the output from past jobs.

3.3 SPICE

Standardizing output using the DAS protocol makes ASAP plans immediately compatible with existing DAS clients. Output from any existing plan can be used by these clients and additional plans can be written if specific data are required. As an example, we integrated our system with the SPICE DAS client (Prlić *et al*., 2005).

SPICE is a visualization program that displays protein sequences and structure annotations from DAS servers. We wrote an ASAP plan to provide UniProt (Bairoch *et al*., 2005) accession data required by SPICE. We then added the URL for our ASAP installation to the list of DAS servers, allowing SPICE to contact ASAP using its standard HTTP queries. Although existing servers can already provide UniProt data to SPICE, this integration serves as an example of how plans can be written using this interface to provide ASAP annotations to an existing protocol.

4 CONCLUSION

Data integration is becoming a major goal of many high-throughput projects. The linking of ASAP and DAS provides a method to retrieve diverse annotations and present them through the standard, widely used DAS client interface.

Funding: This work was supported by the National Library of Medicine NLM T15-LM007356.

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. *et al*. (2005) The universal protein resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Dowell,R. *et al*. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Fertig,E. *et al*. (2010) CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomatic data. *Bioinformatics*, **26**, 2792–2793.

Finn,R. *et al*. (2007) Proserver: a simple, extensible Perl DAS server. *Bioinformatics*, **23**, 1568–1570.

Kossenkov,A. *et al*. (2003) ASAP: automated sequence annotation pipline for web-based updating of sequence information with a local dynamic database. *Bioinformatics*, **19**, 675–676.

Matys,V. *et al*. (2003) TRANSFAC: transciptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.

Prlić,A. *et al*. (2005) Adding some SPICE to DAS. *Bioinformatics*, **32**, ii40–ii41.

Yu,H. *et al*. (2007) Transcript-level annotation of Affymetrix probesets improves the interpretation of gene expression data. *BMC Bioinformatics*, **8**, 194.