Bioinformatics, 31(12), 2015, 2056–2058
doi: 10.1093/bioinformatics/btv107
Advance Access Publication Date: 19 February 2015
Applications Note

OXFORD

Databases and ontologies

The GPMDB REST interface

David Fenyö[1] and Ronald C. Beavis[2,*]

[1]Department of Biochemistry and Molecular Pharmacology, New York University Langone Medical Center, New York, NY, USA and [2]Department of Biochemistry and Medical Genetics, The University of Manitoba, Winnipeg, MB, Canada

[*]To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on November 19, 2014; revised on February 2, 2015; accepted on February 11, 2015

Abstract

Summary: The Global Proteome Machine and Database (GPMDB) representational state transfer (REST) service was designed to provide simplified access to the proteomics information in GPMDB using a stable set of methods and parameters. Version 1 of this interface gives access to 25 methods for retrieving experimental information about protein post-translational modifications, amino acid variants, alternate splicing variants and protein cleavage patterns.

Availability and implementation: GPMDB data and database tables are freely available for commercial and non-commercial use. All software is also freely available, under the Artistic License. http://rest.thegpm.org/1 (GPMDB REST Service), http://wiki.thegpm.org/wiki/GPMDB_REST (Service description and help), and http://www.thegpm.org (GPM main project description and documentation). The code for the interface and an example REST client is available at ftp://ftp.thegpm.org/repos/gpmdb_rest

Contact: rbeavis@thegpm.org or david@fenyolab.org

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

1 Introduction

The Global Proteome Machine and Database (GPMDB, http://thegpm.org) is a project to collect and disseminate in2formation derived from raw data generated by proteomics experiments that use tandem mass spectrometry to analyze protein samples (Craig *et al.*, 2004). It currently contains the results of analyzing over 270 000 proteomics datasets and more than 1.8 billion peptide identifications. The system uses a combination of data analysis servers, an XML file storage repository and a set of MySQL database servers to generate and record the results of proteomics experiments. It has become part of a larger, informal system of data repositories and information systems widely used by proteomics investigators (Perez-Riverol *et al.*, 2014).

The original design of the system allowed investigators to use a conventional web browser/HTML system to select information of interest about individual peptides or proteins as well as information about particular datasets using forms and URL links. While this style of interface has proven to be very useful for individuals browsing the information, it has become increasingly cumbersome for automated information retrieval systems, where the HTML pages are being parsed to generate summary information for calculations. The REST interface described here is web service specifically designed to simplify data access and automate some common tasks in bioinformatics and computational biology research.

2 GPMDB REST interface design and implementation

The version 1.0 design was initiated by the results of a request for comment made in June 2012 to the GPMDB user community. Following consultation with the users and a series of test implementations, 25 methods were selected to comprise the initial version of the interface. The base URL for the service, http://rest.thegpm.org/1, has the version number as an integral part of the call: subsequent versions will end in the appropriate version number so the service can be updated without altering the functionality available in a particular version. The methods were selected to be as atomic as possible, so that a complex query could be composed of serial requests, rather than creating a large number of special-purpose methods. Of the major proteomics data resources, PRIDE (Vizcaino *et al.*, 2013)

has recently added a REST interface and PeptideAtlas (Deutsch *et al.*, 2008) does not yet have REST-style web services. Hopefully our design may serve as an example for the creation of a standard set of services that would simplify obtaining information from such systems.

The service was implemented in object-oriented PERL 5 and run by Apache HTTPD server in response to URL-based Common Gateway Interface (CGI) GET requests. All responses to requests are made in Javascript Object Notation (JSON) with the MIME header 'Content-type: application/json'. Three separate modules were created: (i) rest.pl—handles the CGI request and returns the JSON text, (ii) gpm_rest.pl—handles requests for information contained in individual data set analysis files and (iii) gpmdb_rest.pl—handles all requests to the GPMDB MySQL relational database.

The REST interface was designed to return information involving protein biochemistry. Information associated with mass spectra or the details of specific peptide-to-spectrum assignments will continue to be available through the main GPMDB web interface.

3 Rest Service

The 25 methods chosen for the web service fall in four general categories: *interface*, *model*, *peptide* and *protein*. Calls to a particular method have the category type included in the URL, e.g. to call the 'help' method, the string '/interface/help' is appended to the base URL. The specification of all methods and their parameters is available at http://wiki.thegpm.org/wiki/GPMDB_REST.

Information available from GPMDB is held in both a large set of XML files and a relational database system. The XML files (in BIOML format, Fenyö, 1999) contain the original results of analyzing experimental raw data, while the database contains information and indexes derived from those files and external information sources. The *model* category was reserved for accessing information from the XML files directly, while the *peptide* and *protein* categories access information from the relational database.

3.1 Interface methods

The *interface* method category was meant to contain calls associated with generic information regarding the technical specifications of the web service. Two methods were implemented, '*/interface/help*' and '*/interface/version*'. The '*help*' method returns a JSON-formatted text string containing a listing of the available methods and the documentation URL. The '*version*' method returns the service's current implementation date.

3.2 Model methods

The *model* category accesses information about individual data analyses from the original XML file output by the peptide identification search engine (Supplementary Material). It has seven methods, all of which require the XML file accession number as an input parameter. This accession number (a unique file identifier generated at run time) is in the format 'GPMddddddddddd' (d is any digit 0–9). The *model* methods retrieve the metadata about the original sample and identification process, as well as the peptides, proteins, posttranslational modifications and amino acid variants detected by the analysis. These methods query the original XML files directly, without reference to the database. XML files are not either altered or deleted once they are registered with the system: any information updates (e.g. changes to the sample metadata) are performed on the database only and therefore the results of *model* queries should be considered to be permanent records.

3.3 Peptide methods

The *peptide* category retrieves summary information relevant to specific peptide sequences. There are three methods, each of which requires the specification of the peptide sequence of interest in the standard upper-case, single-letter amino acid code, including the rare genome-encoded amino acids pyrolysine (O) and selenocysteine (U). Some legacy peptide sequences may use the B (D or N) and Z (E or Q) ambiguity symbols for residues, if it was specified in the original protein sequence listing FASTA file used for data analysis. This version does not have any provision for wild card or peptide sequence similarity searches: only the specified sequence can be used to obtain information.

The three *peptide* methods allow the retrieval of the total number of observations of a peptide, the number of observations of a peptide as a function of the parent ion charge assigned to the tandem mass spectrum and a list of all of the protein accession numbers that have been assigned to that sequence. The list of accession numbers can be used to formulate more specific queries using one or more of the *protein* methods.

3.4 Protein methods

The *protein* category methods request information relevant to a particular protein sequence, using its accession number and additional query-specific parameters. These accession numbers were all obtained from publicly available protein sequence sources, such as ENSEMBL (Flicek *et al.*, 2014). These accession numbers are assigned at the time of analysis, based on the annotation provided in FASTA files. No attempt has been made to provide translation between different accession number schemes: e.g. proteins assigned to ENSEMBL accessions cannot be retrieved directly using UniProt identifiers. For cases where a peptide sequence is known, the '*/peptide/accessions*' method can be helpful finding a desired protein. A keyword search method (*/protein/keyword*) is also available to generate a list of available accession numbers. Resources that allow the retrieval of accession number mappings to ENSEMBL, such as BIOMART, may also be helpful. There are no plans to include accession number mapping in the REST interface.

Thirteen *protein* methods have been made available. Some of these methods return general information about a protein (*sequence*, *keyword* and *description*). Sequence-specific posttranslational modifications(*modifications)* can be queried as well as sequence variants (*polymorphisms*), with information about the frequency of observation. Due to the complexity of residue modification information, each type of modifications (e.g. phosphorylation, acetylation) must be queried individually. All amino acid variant information for a protein is returned via a single query, in the ':p.' format suggested by den Dunnen *et al.* (2000). The other *protein* methods return summary information about the peptides observed for an accession number and quality control statistics required by for the Human Proteome Project (Lane *et al.*, 2014).

Acknowledgements

The authors would like to thank John Cortens, Brett Phinney, Yasset Perez-Riverol, Paul Rudnick, Attila Csordas and Emanuele Alpi for their helpful suggestions and discussion.

Funding

This work was supported by the George and Fay Yee Centre for Health Care Innovation (to R.C.B.) and the University of Manitoba.

Conflict of Interest: none declared.

References

Craig,R. *et al.* (2004) An open source system for analyzing, validating and storing protein identification data. *J. Proteome Res.*, **3**, 1234–1242.

den Dunnen,J.T. *et al.* (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.*, **15**, 7–12.

Deutsch,E.W. *et al.* (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Reports*, **9**, 429–434.

Fenyö,D. (1999) The biopolymer markup language. *Bioinformatics*, **15**, 339–340.

Flicek,P. *et al.* (2014) ENSEMBL 2014. *Nucleic Acids Res.*, **42**, D749–D755.

Lane,L. *et al.* (2014) Metrics for the Human Proteome Project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.*, **13**, 15–20.

Perez-Riverol,Y. *et al.* (2014) Making proteomics data accessible and reusable: Current state of proteomics databases and repositories. *Proteomics*, **15**, 930–949

Vizcaino,J.A. *et al.* (2013) *The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013*, Nucleic Acids Res., **41**, D1063–D1069.