The GPMDB REST Interface
Ronald C. Beavis (University of Manitoba) & David Fenyö (New York University Medical School)

Introduction

Proteomics databases, such as GPMDB, contain a lot of information but it can be difficult for interested scientists to extract the information they need. The web sites provided for information retrieval can be very helpful to perform searches and extract summaries as graphical displays. However, if you want to ask a simple, specific question, web sites tend to return far more information than required and the resulting web pages are rather complex if you only wish to extract numerical information.

One way to make it easier to access the information is to use an alternate to an HTML web site, called a REST API. Coupled with a simplified method of representing text information called JSON, it is possible to create a useful way to query the database that is simple and produces easy-to-read responses.

The simplicity and brevity of REST/JSON APIs has made them a popular choice for exchanges between a server and a smart-phone app.

REST description

For example, if you want to know how many times the peptide SPY5VEPVADMLMGLFFR has been observed, the normal web interface generates a response with 850x more characters than the REST API.

GPMDB REST query

GET /protein/sequence/acc=ACC
The most current sequence available for ACC.

GPMDB Web query

GET /peptide/count_z/seq=SEQUENCE
Count of the total number of spectra assigned to peptides in the ACC

Design & implementation

The design of this Application Programming Interface was initiated by the results of a Request for Comment (June 2012) on the GPMDB RFC web page. As a result of user consultations and test implementations, 25 methods were finalized as the first version of the API. The service uses the following base URL:

http://rest.thegpm.org/1

that has the version number directly in the URL. The base URL for future versions will end in an appropriate version number. This convention allows the service to be updated without changing the available previous versions. The methods were chosen to be as atomic as possible, enabling the design of complex queries as a set of serial requests. The alternative of creating a larger number of special-purpose methods was rejected. This design may serve as an exemplar for the creation of a standard set of services across similar platforms making information access from these systems more uniform.

The service was written using object-oriented Perl. The production service is being run using an Apache HTTPD server using URL-based Common Gateway Interface (CGI) GET requests. All responses are made in Javascript Object Notation (JSON) including the MIME header “Content-type: application/json”. Three separate software objects were created: 1. rest.pl – interprets the CGI request and returns JSON objects; 2. gpm_rest.pl – services requests for information obtained from individual data set analysis files; and 3. gpmdb_rest.pl – services all requests to the GPMDB MySQL RDB.

The REST interface was designed to return information involving protein biochemistry. Information about mass spectra or the details of peptide-to-spectrum assignments is available through the main GPMDB web interface.

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 value:

http://rest.thegpm.org/1/interface/help

GET /interface/version
The API’s current version number

GET /model/petdata/gpm=GPM
Some text about available methods

GET /model/proteins/gpm=GPM
A list of protein accession numbers (ACCs) found in the data file GPM

GET /model/protein_savs/gpm=GPM&acc=ACC
A list of the single amino acid variants (SAVs) found in GPM for the protein ACC.

GET /model/protein_peptides/gpm=GPM&acc=ACC
A list of the peptides found in GPM for ACC

GET /model/protein_savs/acc=ACC
A list of the single amino acid variants (SAVs) found in GPM for ACC

GET /model/protein_sequence/gpm=GPM&acc=ACC
The protein sequence corresponding to ACC in GPM

GET /model/protein_species/acc=ACC1,ACC2,...
A list of protein ACCs with WORD in their text descriptions.

GET /peptide/sequences/seq=SEQ
A list of the protein ACCs observed with the peptide sequence (SEQ).

GET /peptide/sequence/seq=SEQ
The total number of spectra assigned to SEQ.

GET /peptide/sequence/seq=SEQ&parent=SEQ1(=SEQ2,...)
A list of the spectra count for specific parent ion charges assigned to SEQ.

GET /protein/best/e=acc=ACC
The lowest E-value assigned to ACC.