Published online 30 April 2018 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* ***W563–W570***
doi: 10.1093/nar/gky294

An update on PUG-REST: RESTful interface for programmatic access to PubChem

Sunghwan Kim†, Paul A. Thiessen†, Tiejun Cheng, Bo Yu and Evan E. Bolton*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20894, USA

Received January 29, 2018; Revised March 15, 2018; Editorial Decision April 08, 2018; Accepted April 09, 2018

ABSTRACT

PubChem (https://pubchem.ncbi.nlm.nih.gov) is one of the largest open chemical information resources available. It currently receives millions of unique users per month on average, serving as a key resource for many research fields such as cheminformatics, chemical biology, medicinal chemistry, and drug discovery. PubChem provides multiple programmatic access routes to its data and services. One of them is PUG-REST, a Representational State Transfer (REST)-like web service interface to PubChem. On average, PUG-REST receives more than a million requests per day from tens of thousands of unique users. The present paper provides an update on PUG-REST since our previous paper published in 2015. This includes access to new kinds of data (e.g. concise bioactivity data, table of contents headings, etc.), full implementation of synchronous fast structure search, support for assay data retrieval using accession identifiers in response to the deprecation of NCBI's GI numbers, data exchange between PUG-REST and NCBI's E-Utilities through the List Gateway, implementation of dynamic traffic control through throttling, and enhanced usage policies. In addition, example Perl scripts are provided, which the user can easily modify, run, or translate into another scripting language.

INTRODUCTION

PubChem (https://pubchem.ncbi.nlm.nih.gov) (1–3) is an open chemical database and a public repository for information on chemical substances and their biological activities, developed and maintained by the U.S. National Library of Medicine, part of the U.S. National Institutes of Health. It is one of the largest online resources in the open chemical information domain, with >235 million depositor-provided substance descriptions, 94 million unique chemical structures, and one million biological assays, covering more than ten thousand unique protein target sequences. PubChem is used by millions of unique users per month, serving as a key resource for many research fields such as cheminformatics, chemical biology, medicinal chemistry, and drug discovery.

As described in our previous paper (4), PubChem provides multiple programmatic access routes to its data and services. One of them is PUG-REST, a Representational State Transfer (REST)-like web service interface to PubChem. It is designed to handle synchronous tasks, in which the output of the requests can be returned to the user immediately, as opposed to a queuing system that requires use of a polling scheme to check for completion. PUG-REST does not require prior knowledge of a PubChem-specific XML specification or use of a SOAP envelope. Instead, (almost) all necessary information for a PUG-REST request is encoded into a single-line uniform resource locator (URL). This simplicity makes it easy to use and learn, relative to other programmatic access interfaces provided by PubChem. PUG-REST also provides convenient access to information on PubChem records not possible with the other PubChem programmatic interfaces. As a result, PUG-REST is the most heavily used programmatic interface to PubChem contents with millions of daily requests from tens of thousands of IP addresses (Figure 1).

The concepts and syntax of PUG-REST are well described in our previous paper (4), and additional information can be found in the following PubChem Help documents:

- PUG-REST (https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest)
- PUG-REST Tutorial (https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest-tutorial)

The present paper provides an update on PUG-REST since the previous paper (4). This includes full implementation of fast (synchronous) structure search, identifier list exchange between Entrez-Utilities and PUG-REST through the List Gateway, use of accession identifiers for target proteins,

*To whom correspondence should be addressed. Tel: +1 301 451 1811; Fax: +1 301 480 4559; Email: bolton@ncbi.nlm.nih.gov
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

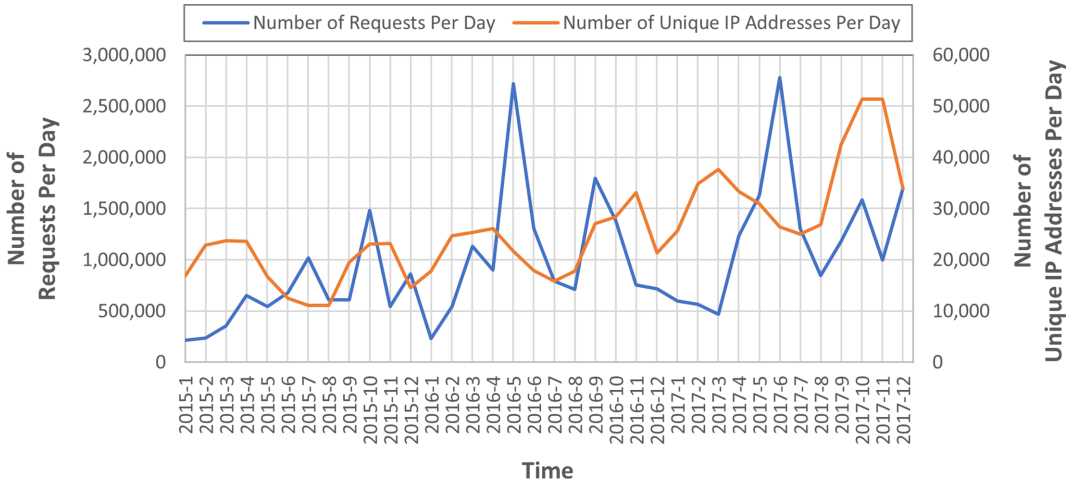


Figure 1. Monthly average of the number of PUG-REST requests per day and unique IP addresses per day.

access to new types of information (e.g. concise bioactivity data, table of contents headings, etc.), dynamic traffic control, and enhanced usage policies. In addition, example Perl scripts (example01.pl through example07.pl) are provided as supplementary materials to demonstrate how to use PUG-REST.

PUG-REST SYNTAX

As the concepts and syntax of PUG-REST requests are explained elsewhere (4), only a brief description is given here. Each request to PubChem requires three pieces of information:

- Input: a list of identifiers for PubChem records, provided explicitly or implicitly (via query)
- Operation: the operation to be performed with the input identifiers
- Output: the format of the output [XML, JSON(P), ASN, SDF, CSV, PNG and TXT]

In PUG-REST, these pieces of information are encoded in a single-line URL (Figure 2). Some types of requests may need or allow additional information to be provided after an '?' character at the end of the request URL via an '&'-separated list of optional name-value pairs. The full specification of the input, operation, and output parts of the PUG-REST request URL is presented in Figure 3. More detailed descriptions of these specifications are given in the PUG-REST Help document (https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest).

It should be noted that special characters [e.g. '/' (forward slash)], or multi-line inputs [e.g. a molecular structure input in the structure-data file (SDF) format] are not compatible with the (HTTP GET) URL-based PUG-REST syntax. This can be easily circumvented by sending data via an HTTP POST mechanism, as demonstrated in example scripts provided in the supplementary material (example03.pl and example04.pl).

It is also worth mentioning that PUG-REST now supports only Hypertext Transfer Protocol Secure (HTTPS) requests, as implied in the URL syntax beginning with 'https' (as opposed to 'http') in Figure 2 (https://go.usa.gov/xnvZZ). This is to comply with the HTTPS-only standard of the U.S. Federal Government, which mandates the use of the HTTPS (over HTTP) at all publicly accessible Federal government websites. HTTPS enhances the privacy and security of the user over unencrypted data transfer via HTTP (https://https.cio.gov/).

FAST (SYNCHRONOUS) STRUCTURE SEARCH

PUG-REST is designed for short, synchronous requests, where the output is returned immediately (ideally within 100 ms, but with a maximum response time of 30 s). However, when PUG-REST was initially developed, certain types of tasks took inherently much longer than 30 s, and therefore it was not possible to handle them as synchronous requests. Examples of such slow tasks included chemical structure search (e.g. similarity search, substructure/superstructure search). Because chemical structure search was commonly requested through other PubChem web services (i.e. PUG and PUG-SOAP), the demand for these tasks was also present for PUG-REST. Therefore, PUG-REST initially supported these tasks in an 'asynchronous' way (Figure 4). That is, the requested task was queued and a numeric key associated with the queued task was returned to the client. This key was then used to poll the server to check the status of the queued task and retrieve the result when it was completed.

Recent improvements in PubChem's structure search system have made chemical structure search and molecular formula search fast enough to be handled synchronously. As shown in Figure 4, these synchronous searches are invoked by adding the 'fast' prefix to the name of the old asynchronous search, except for similarity search. In the past, only two-dimensional (2-D) similarity search (invoked by 'similarity') was supported in PUG-REST as an asynchronous operation. Now it supports synchronous operations for both 2-D and three-dimensional (3-D) similarity searches (invoked by 'fastsimilarity_2d' and 'fastsimilarity_3d', respectively). For example, the following URLs

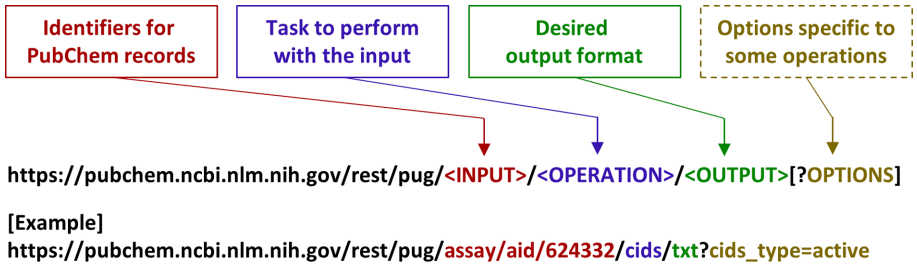


Figure 2. Syntax of a PUG-REST web service request URL, with an example that retrieves (in a text file) compounds tested to be active in an assay.

Input Specification

<input> = <domain>/<namespace>/<identifiers>
<domain> = substance | compound | assay | <other inputs>

- **If <domain> = compound**
  <namespace> = cid | name | smiles | inchi | sdf | inchikey | formula | <structure search> | <xref> | listkey | <fast search>
  <structure search> = {substructure | superstructure | similarity | identity}/{smiles | inchi | sdf | cid}
  <fast search> = {fastidentity | fastsimilarity_2d | fastsimilarity_3d | fastsubstructure | fastsuperstructure}/{smiles | smarts | inchi | sdf | cid} | fastformula
- **If <domain> = substance**
  <namespace> = sid | sourceid/<source id> | sourceall/<source name> | name | <xref> | listkey
  <source name> = any valid PubChem depositor name
- **If <domain> = assay**
  <namespace> = aid | listkey | type/<assay type> | sourceall/<source name> | target/<assay target> | activity/<activity column name>
  <assay type> = all | confirmatory | doseresponse | onhold | panel | rnai | screening | summary | cellbased | biochemical | invivo | invitro | activeconcentrationspecified
  <assay target> = gi | proteinname | geneid | genesymbol | accession

<other inputs> = sources / [substance, assay] |sourcetable | conformers | annotations/[sourcename/<source name>| heading/<heading>]

<identifiers> = comma-separated list of positive integers (e.g. cid, sid, aid) or identifier strings (source, inchikey, formula); in some cases only a single identifier string (name, smiles, xref; inchi, sdf by POST only)
<xref> = xref / {RegistryID | RN | PubMedID | MMDBID | ProteinGI | NucleotideGI | TaxonomyID | MIMID | GeneID | ProbeID | PatentID}

Operation Specification

- **If <domain> = compound**
  <operation> = record | <property> | synonyms | sids | cids | aids | assaysummary | classification | <xrefs> | description | conformers
  <property> = property / [comma-separated list of property tags]
- **If <domain> = substance**
  <operation> = record | synonyms | sids | cids | aids | assaysummary | classification | <xrefs> | description
- **If <domain> = assay**
  <operation> = record | concise | aids | sids | cids | description | targets/<target type> | <doseresponse> | summary | classification
  <target_type> = {ProteinGI, ProteinName, GeneID, GeneSymbol}
  <doseresponse> = doseresponse/sid

<xrefs> = xrefs / [comma-separated list of xrefs tags]

Output Specification

<output> = XML | ASNT | ASNB | JSON | JSONP [?callback=<callback name>] | SDF | CSV | PNG | TXT

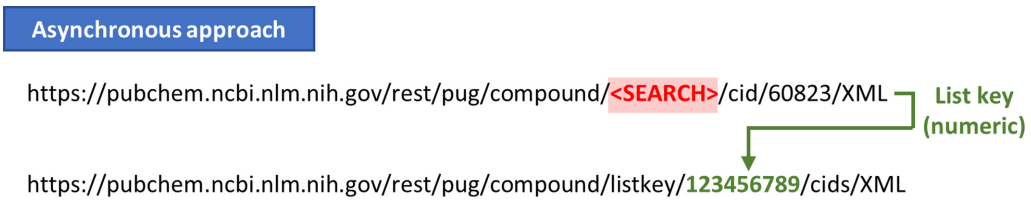
Figure 3. Specification of the input, operation, and output parts of a PUG-REST request.

perform synchronous 2-D and 3-D structure similarity searches, respectively, using CID 446157 (Crestor) as a query and download the hit compounds into a text file:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/fastsimilarity_2d/cid/446157/cids/TXT

https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/fastsimilarity_3d/cid/446157/cids/TXT

In 2-D similarity search, molecular similarity is computed using the Tanimoto equation (5–7) and the PubChem substructure fingerprint (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf). In 3-D similarity



	Keyword for <SEARCH>	
Search type	**Asynchronous approach**	**Synchronous approach**
Identity Search	Identity	fastidentity
2-D Similarity Search	similarity	fastsimilarity_2d
3-D Similarity Search[a]	–	fastsimilarity_3d
Substructure search	substructure	fastsubstructure
Superstructure search	superstructure	fastsuperstructure
Molecular Formula Search	formula	fastformula

[a] 3-D similarity search is not supported in the asynchronous approach.

Figure 4. Construction of PUG-REST requests for chemical structure search and molecular formula search in synchronous and asynchronous approaches.

search, the similarity between conformers are evaluated with the shape-Tanimoto (ST) and color-Tanimoto (CT) scores, which quantify the similarity between their conformers in 3-D shape and functional group orientations, respectively (8,9). The ST and CT scores are calculated using the Gaussian-shape overlay method by Grant and Pickup (10–12), and implemented in the Rapid Overlay of Chemical Structures (ROCS) (13). Up to ten conformers per compound are considered during the 3-D similarity search. More details about PubChem's 3-D similarity search are described elsewhere (8,9,14,15).

Most of the optional parameters used in asynchronous operations can also be used in their synchronous variants. A notable exception is the 'listkey' parameter, which allows one to restrict the structure search to hits from a prior search (specified by the numeric list key). This optional parameter is not (currently) supported in synchronous structure searches. Note also that the 3-D similarity search cannot accept a customized threshold. The threshold used for the 3-D similarity search is (currently) fixed to ST = 0.80 and CT = 0.50, which is identical to the threshold used for PubChem 3-D neighbor computations (16,17). In contrast, the threshold for 2-D similarity search is adjustable using the option 'Threshold'. For instance, the following URL performs a synchronous 2-D similarity search using a Tanimoto threshold of 0.80:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/fastsimilarity_2d/cid/446157/cids/TXT?Threshold=80

Note that the 'Threshold' option takes the integer value of 80 (multiplying 0.80 by 100).

For most queries, the new fast chemical search approach in PUG-REST provides a dramatic benefit of programmatic simplicity, speed, and convenience. In contrast, it is worth noting that some complicated chemical structure searches will not complete rapidly, and may fail as a result. Although one may try the older queued-approach in asynchronous way to circumvent this issue, it is more likely that the query itself is not very specific enough, resulting in too many hits that the search system cannot handle (for example, substructure search using benzene as a query, which would lead to millions of hit compounds). Therefore, in such cases, it is highly recommended to provide a more specific query that may lead to less hits.

LIST GATEWAY

Entrez (18–20) is a data retrieval system that provides integrated access to the three PubChem databases (Compound, Substance, and BioAssay) (1,2) as well as tens of other NCBI databases in a wide range of biomedical data domains, including nucleotide and protein sequences, gene records, 3-D molecular structures, and the biomedical literature (21). Programmatic access to data within the Entrez system is provided through a set of programs, called Entrez Programming Utilities (also known as E-Utilities: https://www.ncbi.nlm.nih.gov/books/NBK25501/). While appropriate for searching or accessing text and numeric data, E-

Figure 5. Schematic diagram for the function of the List Gateway.

Utilities is not suitable for handling other types of data specific to PubChem (such as chemical structure queries, and bioactivity data tables). In addition, Entrez is limited in its extensibility and does not contain all PubChem contents. This is a primary reason why PubChem introduced additional programmatic access protocols, such as PUG, PUG-SOAP and PUG-REST. Therefore, E-Utilities and PubChem-specific programmatic access routes complement each other.

To get desired data from PubChem, one may need to make multiple web service requests in a sequential order such that the output from a previous request is used as the input to a subsequent request. In such cases, it is beneficial to temporarily store the results from intermediate steps on a server for future use. Although this can be done using a list key in PUG-REST and an Entrez history in E-Utilities, they are not compatible with each other, meaning that the list key from a PUG-REST request cannot be used in E-Utilities and the Entrez history from a E-Utilities request cannot be used in PUG-REST.

The PubChem List Gateway (https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest-tutorial$_List_Gateway) is a common gateway interface (CGI) that converts between the list key from a PUG-REST request and the Entrez history from an E-Utilities request (Figure 5). This program allows one to use the result from a PUG-REST request as an input to a subsequent E-Utilities request, or vice versa (i.e. it makes a bridge between the two systems). An Entrez history is specified using three parameters, database (DB), Query Key, and WebEnv. The list gateway takes these three parameters for an Entrez history, and returns a list key, which can be used in a subsequent PUG-REST request. As an example, the following URL shows how to convert from a Entrez history to a PUG-REST list key:

https://pubchem.ncbi.nlm.nih.gov/list_gateway/list_gateway.cgi?action=entrez_to_pug&entrez_db=DB&entrez_query_key=QUERYKEY&entrez_webenv=WEBENV

where QUERYKEY and WEBENV are the query key and WebEnv values for an Entrez history, respectively, and DB is the name of the PubChem database in Entrez ('pccompound' for Compound, 'pcsubstance' for Substance, and 'pcassay' for BioAssay). Conversely, the list key from a PUG-REST request can be converted into the three parameters (DB, Query Key and WebEnv) that specify an Entrez history, via the following URL:

https://pubchem.ncbi.nlm.nih.gov/list_gateway/list_gateway.cgi?action=pug_to_entrez&pug_listkey=LISTKEY

where LISTKEY is a PUG-REST list key.

Two use cases of the list gate way are included as examples in the supplementary material (exmaple05.pl and example06.pl). In example05.pl, an Entrez search is performed to retrieve all compounds whose molecular weights are within a given range, and their molecular properties (including molecular formula, isomeric SMILES, etc.) are downloaded through PUG-REST. In example06.pl, a 2-D similarity search is performed using a CID query through PUG-REST, and then the hit compounds are filtered in Entrez to identify those which satisfy the Lipinski's rule of 5 (22).

ASSAY DATA RETRIEVAL THROUGH ACCESSION

At one point in time, GI numbers, which stands for 'GenInfo Identifiers' (https://www.ncbi.nlm.nih.gov/genbank/sequenceids/), were the only primary identifiers for target protein sequences tested against in bioassays archived in PubChem. However, as described elsewhere (21,23,24), NCBI recently phased out the practice of assigning GI numbers to sequence records. Accordingly, changes have been made to PUG-REST to allow one to access assay data using the 'accession' identifiers as an input. It should be noted that a protein often has multiple versions representing slightly different sequences. Each of these versions are indicated with the corresponding version number following the accession (in the form of 'accession.version'). To exemplify the complication of this issue, the identifiers for the protein target sequences for BioAssay records AID 38693 and AID 1159673 are compared in Figure 6. The two assays were performed against two sequences specified with GI numbers 1018618719 and 113830, respectively, which correspond to different versions of the human androgen receptor (P10275.3 and P10275.2). In PUG-REST, these closely related sequences associated with the same accession are treated equal. For example, when an 'accession' number is provided in a PUG-REST request, all versions of the protein represented by that accession are considered. For example, the following URL retrieves assays targeting the human androgen receptor (accession: P10275):

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/target/accession/P10275/aids/TXT

Target Protein Identifiers & length	AID 38693	AID 1159673
Accession	P10275	P10275
Accession.Version	P10275.3	P10275.2
GI	1018618719	113830
# Amino Acids	920	919

Figure 6. Comparison of the identifiers and lengths of the target protein sequences for AIDs 38693 and 1159673.

When an 'accession.version' identifier is provided, the version number is ignored. Therefore, these two URLs return the same list of AIDs as the above URL does:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/target/accession/P10275.2/aids/TXT
https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/target/accession/P10275.3/aids/TXT

Because of the archival nature of PubChem, it continues supporting the use of a GI-number in a PUG-REST request. However, it is internally converted into an accession. As a result, these two GI-numbers return the same AID list as the three accession-based requests above:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/target/gi/113830/aids/TXT
https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/target/gi/1018618719/aids/TXT

As a result, all five PUG-REST requests provide the same list of AIDs.

It should also be noted that different sequence identifiers often correspond to an identical sequence. In PUG-REST, the protein sequence specified by an input identifier (regardless of GI or accession) is automatically expanded to all its identical sequences (which may be represented by different identifiers).

The accession identifier for the target protein sequence of an assay can be retrieved through a PUG-REST request URL like the following.

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/3364/targets/ProteinAccession/XML

Currently, PUG-REST does not provide a direct way to get the accession.version identifier of the protein targets for an assay. However, this can be done by getting the GI number of an assay target through PUG-REST, followed by the E-Utilities' E-Fetch call for retrieval of the accession from the GI number. For example, the request below returns the GI number of the protein target of AID 3364:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/3364/targets/ProteinGI/XML

The returned GI number 2811086 is then used in the following request to get the accession of the target protein sequence:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=2811086&rettype=acc

CONCISE DATA TABLE

PUG-REST now allows one to get the 'concise' bioactivity data table for an assay or assays (relative to the full bioactivity data table). The concise data table contains key information commonly provided for PubChem assays, such as tested substances/compounds, bioactivity outcomes, protein/gene target identifiers, activity values (e.g. IC_{50}, EC_{50}, K_d and K_i), assay types (e.g. primary, confirmatory or summary), the PubMed PMIDs for a source article (for literature-derived assays), and a flag that indicates RNAi assays. For example, the concise data table for AID 260895 can be accessed through the following URL:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/aid/260895/concise/CSV

Because assays archived in PubChem are very heterogeneous, it is not trivial to extract common data attributes from the full data table of different assays. With that said, the concise data table is useful when the user wants to get aggregated information from multiple assays. For example, the following PUG-REST request returns bioactivity data for all assays whose target is EGFR gene products:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/assay/target/genesymbol/egfr/concise/CSV

DATA SOURCES

Some advanced users are keen on where PubChem data come from. The PubChem Sources page (https://pubchem.ncbi.nlm.nih.gov/sources/) allows users to find who provided what information to PubChem. This information is also available programmatically through PUG-REST. For example, the list of all sources for substance records can be retrieved (in JSON) using the following URL:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/sourcetable/substance/json

Similarly, the list of all assay record sources can be obtained (in CSV) via the URL:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/sourcetable/assay/csv

PubChem also collects a wide range of annotations through integration of authoritative data sources. Using PUG-REST, one can get information on what sources provide which annotations, or what kind of annotation is integrated from a given source into PubChem. For example, the following request URL returns all annotation sources of the solubility data in PubChem:

https://pubchem.ncbi.nlm.nih.gov/rest/pug/annotations/heading/solubility/json

One can also get the kinds of annotations collected from a given source (e.g. HSDB):

https://pubchem.ncbi.nlm.nih.gov/rest/pug/annotations/sourcename/hsdb/json

DYNAMIC TRAFFIC CONTROL THROUGH THROTTLING

All PUG-REST requests are sent to PubChem web servers. Excessive numbers of requests by users can result in unexpected PubChem service disruptions. Established web usage policies exist to help ensure all users can access the PubChem resource. In addition, NCBI-wide usage policies specify per-user limits on the count of web requests per second (https://www.ncbi.nlm.nih.gov/home/about/policies/#scripting). To help maximize uptime and request handling speed, PubChem web servers employ a dynamic, web-request throttling approach that enforces usage policies (see below for more details). In addition, during periods of excessive demand, these policies may be dynamically changed to maintain accessibility to all users. All PUG-REST web-requests are accompanied by specialized HTTP headers providing specifics on the extent one is approaching a usage limit (see the example script example07.pl). Requests exceeding limits are rejected (HTTP 503 error). If the user continuously exceeds the limit, they will be blocked for a period of time.

Information is provided in the HTTP header response, indicating the traffic status for the system state and the extent the user is approaching limits (see the sample script example07.pl). For example, the HTTP header response contains a line similar to the following:

```
X-Throttling-Control: Request Count sta-
 tus: Green (0%), Request Time status:
 Green (0%), Service status: Green (20%)
```

The first two status indicators (Request Count status and Request Time status) give information on your usage of the service in one of four states:

1. Green: less than 50% of the permitted request limit has been used
2. Yellow: between 50% and 75% of the request limit has been used
3. Red: more than 75% of the request limit has been reached
4. Black: the limit has been exceeded and requests are being blocked

The third indicator (Service status) shows the concurrent usage of the service in one of four states:

1. Green (idle): low concurrent usage is being applied to the service at present
2. Yellow (moderate): a moderate number of concurrent requests are being handled
3. Red (busy): a significant number of concurrent requests are being handled
4. Black (overloaded): an excessively high number of concurrent requests are being handled

It is important to note that there are many instances of PubChem services running in parallel. Each instance receives traffic from a load balancer, which distributes the requests across the system. Thus, when a stream of requests is sent to PubChem, the responses will be relative to the PubChem server instance handling the request. One server instance can become overloaded while others may not, depending on the overall nature of requests sent to that server. When providing many requests, one should moderate the speed at which requests are sent according to the worst-case usage feedback received. This will prevent uneven rejection of requests by PubChem services.

USAGE POLICIES

PubChem usage policies, which includes web-based programmatic services, specify limits on web requests as following:

- No more than five requests per second.
- No more than 400 requests per minute.
- No more than 300 second running time on PubChem servers per minute.

Violation of usage policies may result in blocked user requests and being temporarily blocked from accessing PubChem services. It should be noted that these limits can be lowered through the dynamic traffic control at times of excessive load. Throttling information is provided in the HTTP header response, indicating the system-load state and the per-user limits.

DISCUSSION

The use of PUG-REST is steadily increasing in terms of both the number of requests and the number of unique IP addresses accessing it. To make PUG-REST as reliable as possible, and PubChem overall, it was essential to implement a dynamic throttling system through PUG-REST (and other PubChem services). Note that PUG-REST is not intended as a replacement to bulk data download. In addition, if you (the user) believe you need to submit millions of web requests, please consider contacting PubChem first, as there may be a better, more efficient way to do what you want.

DATA AVAILABILITY

All PubChem data, tools, and services, including PUG-REST, are provided to the public free of charge.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

Special thanks go to the entire PubChem team, NCBI staff (especially to the help desk and systems support teams), to the hundreds of data contributors for making their data openly accessible within PubChem, and to you, the PubChem users.

FUNDING

Intramural Research Program of the National Library of Medicine, National Institutes of Health. Funding for open access charge: Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.
2. Wang,Y., Bryant,S.H., Cheng,T., Wang,J., Gindulyte,A., Shoemaker,B.A., Thiessen,P.A., He,S. and Zhang,J. (2017) PubChem BioAssay: 2017 update. *Nucleic Acids Res.*, **45**, D955–D963.
3. Kim,S. (2016) Getting the most out of PubChem for virtual screening. *Expert Opin. Drug Discov.*, **11**, 843–855.
4. Kim,S., Thiessen,P.A., Bolton,E.E. and Bryant,S.H. (2015) PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.*, **43**, W605–W611.
5. Chen,X. and Reynolds,C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **42**, 1407–1414.
6. Holliday,J.D., Salim,N., Whittle,M. and Willett,P. (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **43**, 819–828.
7. Holliday,J.D., Hu,C.Y. and Willett,P. (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screening*, **5**, 155–166.
8. Kim,S., Bolton,E.E. and Bryant,S.H. (2012) Effects of multiple conformers per compound upon 3-D similarity search and bioassay data analysis. *J. Cheminf.*, **4**, 28.
9. Bolton,E.E., Chen,J., Kim,S., Han,L., He,S., Shi,W., Simonyan,V., Sun,Y., Thiessen,P.A., Wang,J. *et al.* (2011) PubChem3D: a new resource for scientists. *J. Cheminf.*, **3**, 32.
10. Grant,J.A. and Pickup,B.T. (1995) A gaussian description of molecular shape. *J. Phys. Chem.*, **99**, 3503–3510.
11. Grant,J.A. and Pickup,B.T. (1996) A gaussian description of molecular shape. *J. Phys. Chem.*, **100**, 2456.
12. Grant,J.A., Gallardo,M.A. and Pickup,B.T. (1996) A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.*, **17**, 1653–1666.
13. Rush,T.S. 3rd, Grant,J.A., Mosyak,L. and Nicholls,A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.*, **48**, 1489–1495.
14. Bolton,E.E., Kim,S. and Bryant,S.H. (2011) PubChem3D: conformer generation. *J. Cheminf.*, **3**, 4.
15. Kim,S., Bolton,E.E. and Bryant,S.H. (2013) PubChem3D: conformer ensemble accuracy. *J. Cheminf.*, **5**, 1.
16. Bolton,E.E., Kim,S. and Bryant,S.H. (2011) PubChem3D: similar conformers. *J. Cheminf.*, **3**, 13.
17. Kim,S., Bolton,E.E. and Bryant,S.H. (2016) Similar compounds versus similar conformers: complementarity between PubChem 2-D and 3-D neighboring sets. *J. Cheminf.*, **8**, 62.
18. Geer,R.C. and Sayers,E.W. (2003) Entrez: making use of its power. *Briefings Bioinf.*, **4**, 179–184.
19. McEntyre,J. (1998) Linking up with Entrez. *Trends Genet.*, **14**, 39–40.
20. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
21. NCBI Resource Coordinators. (2018) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **46**, D8–D13.
22. Lipinski,C.A., Lombardo,F., Dominy,B.W. and Feeney,P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.*, **46**, 3–26.
23. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2017) GenBank. *Nucleic Acids Res.*, **45**, D37–D42.
24. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.