Chemistry Central Journal

ChemistryCentral

Poster presentation

Open Access

A virtual file system for the PubChem chemical structure and bioassay database

Wolf-D Ihlenfeldt

Address: Xemistry GmbH, Auf den Stieden 8, D-35094 Lahntal, Germany
from 3rd German Conference on Chemoinformatics
Goslar, Germany. 11-13 November 2007

Published: 26 March 2008

Chemistry Central Journal 2008, **2**(Suppl 1):P26 doi:10.1186/1752-153X-2-S1-P26

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/P26

The PubChem chemical structure and bioassay database (http://pubchem.ncbi.nlm.nih.gov) has established itself as one of the premier information sources for chemical structures and assay data accessible via the Internet. PubChem provides a convenient interactive Web interface for the execution and result display of standard structure and text-based queries. However, the capabilities for formulating complex queries, and to access and download specific data sets, are limited. Because of the necessity to integrate PubChem into the existing Entrez framework, instead of designing a new streamlined interface for handling its peculiar data content, in many cases queries are awkward to set up and execute if they become more complex.

Recently, PubChem has published a specification for its Power User Gateway (PUG). This interface allows enterprising users to retrieve selected PubChem data via an XML-based Web service. The text-oriented cluster of Entrez databases has been accessible for a long time via similarly structured access modules termed Eutils. Eutils also support access to the text components of the PubChem deposition structure and standardized compound databases. Finally, individual record downloads and similar operations are possible via creatively crafted Web URLs, mimicking those encountered in interactive sessions.

By virtue of these mechanisms, PubChem does in principle provide a larger degree of programmatic accessibility than other chemistry Web databases. Nevertheless, they are extremely difficult to use in their native, raw form.

In order to address this difficulty, we have implemented a virtual file I/O module for the Cactvs Chemoinformatics Toolkit. It provides access to the PubChem compound database as a virtual file. The supported feature set starts with simple record-based I/O and extends to the execution of structure queries of higher complexity than possible via the PUG. Users of the toolkit may now script the same toolkit commands for the PubChem database as they can for a local read-only structure file. Behind the scenes, the I/O module leverages the described three access mechanisms in an optimized fashion, re-routing as many of the operations needed to perform the command to the PubChem servers as possible, but also transparently utilizing downloaded records for local processing in case operations are requested which exceed the capabilities of the systems operating on the PubChem site. A strength of our implementation is that it fully understands the native ASN.1 data specification for the database contents and is therefore not restricted to working on the sometimes coarse approximations of structure configuration which are available as SD-file records.

With this work, we think we are presenting the first usable solution for scripted access to the PubChem database. Given its emerging importance for drug-related research, we hope that this software will be generally useful for a broad audience.