Chemistry Central Journal

ChemistryCentral

Software

Open Access

PubChemSR: A search and retrieval tool for PubChem

Junguk Hur[1] and David J Wild*[2]

Address: [1]Bioinformatics Program, University of Michigan, Ann Arbor, MI 48109, USA and [2]Indiana University School of Informatics, 901 E. 10th Street, Bloomington, IN 47406, USA

Email: Junguk Hur - juhur@umich.edu; David J Wild* - djwild@indiana.edu

* Corresponding author

Published: 15 May 2008

Chemistry Central Journal 2008, **2**:11 doi:10.1186/1752-153X-2-11

Received: 19 March 2008
Accepted: 15 May 2008

This article is available from: http://journal.chemistrycentral.com/content/2/1/11

© 2008 Hur and Wild
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Recent years have seen an explosion in the amount of publicly available chemical and related biological information. A significant step has been the emergence of PubChem, which contains property information for millions of chemical structures, and acts as a repository of compounds and bioassay screening data for the NIH Roadmap. There is a strong need for tools designed for scientists that permit easy download and use of these data. We present one such tool, PubChemSR.

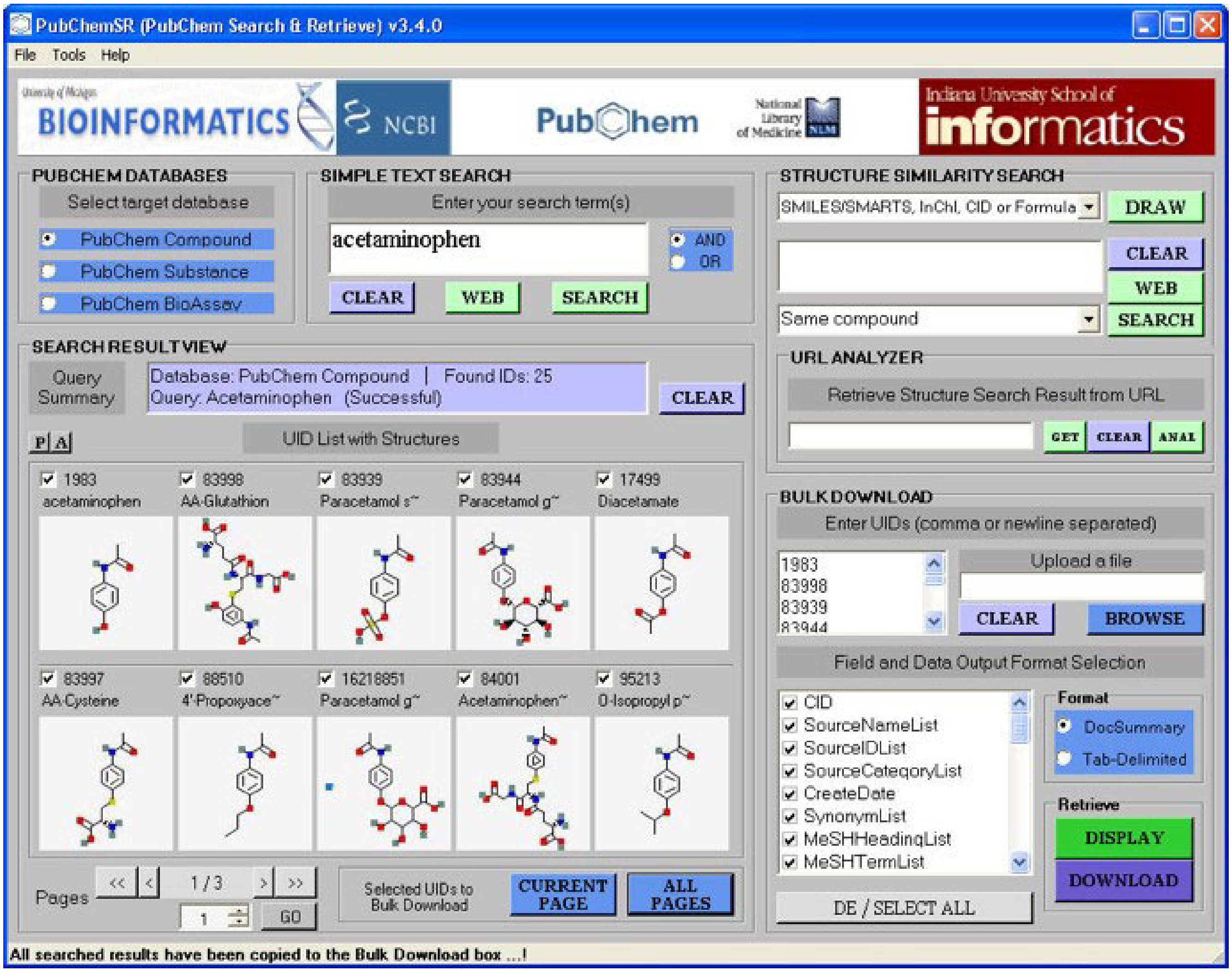
Implementation: PubChemSR (Search and Retrieve) is a freely available desktop application written for Windows using Microsoft *.NET* that is designed to assist scientists in search, retrieval and organization of chemical and biological data from the PubChem database. It employs SOAP web services made available by NCBI for extraction of information from PubChem.

Results and Discussion: The program supports a wide range of searching techniques, including queries based on assay or compound keywords and chemical substructures. Results can be examined individually or downloaded and exported in batch for use in other programs such as Microsoft Excel. We believe that PubChemSR makes it straightforward for researchers to utilize the chemical, biological and screening data available in PubChem. We present several examples of how it can be used.

Background

Recent years have seen an explosion in the amount of chemical and related biological information in freely-accessible databases [1,2] The most widely known of these is PubChem [3], a repository of over 40 million chemical substances (at the time of writing) with associated property, literature reference and biological activity information. In addition to being a resource of information about compounds, this database is the primary repository for High Throughput Screening results generated by the Molecular Libraries Screening Centers Network (MLSCN) [4], part of the NIH Roadmap.

While PubChem has a straightforward web-based user interface for searching, it is quite limited in its facilities for download and processing of search results. For example, one can download data for a particular PubChem entry in XML [5] and a few other formats, but it is not possible to download aggregate search results in a manner that is straightforward for a non-computational scientist. Yet the greatest utility of this information is clearly in aggregate: with structural information for compounds tested in a particular bioassay, one can create a QSAR model; by comparing compounds active in one assay with those active in a second, one can make judgments about selec-



acetaminophen

Database: PubChem Compound | Found IDs: 25
Query: Acetaminophen (Successful)

1983
83998
83939
83944

All searched results have been copied to the Bulk Download box ...!

Figure 1

Main window of PubChemSR. Shown is a simple search result of 'acetaminophen' at the PubChem compound database. PubChemSR has retrieved 25 records by 'acetaminophen' at the PubChem compound database. The bottom left panel shows the structure of the retrieved compounds (10 per page). Clicking 'Current Page' or 'All Pages' will copy the selected (checked) UIDs to the Bulk-download section for further downloading of the full or selected data field.

tivity; by downloading properties for compounds similar to a query one can investigate the behavior of a series of compounds rather than individual compounds. There is thus a need for tools to be developed that allow easy search, access and download of information in PubChem, and in particular which allow one to move information *en bloc* to one's own computer for further processing. The development of PubChemSR was thus driven by the desire to have at hand such features as:

- Easy search and retrieval of detailed compound, substance and bioassay information, including substructure and similarity searching

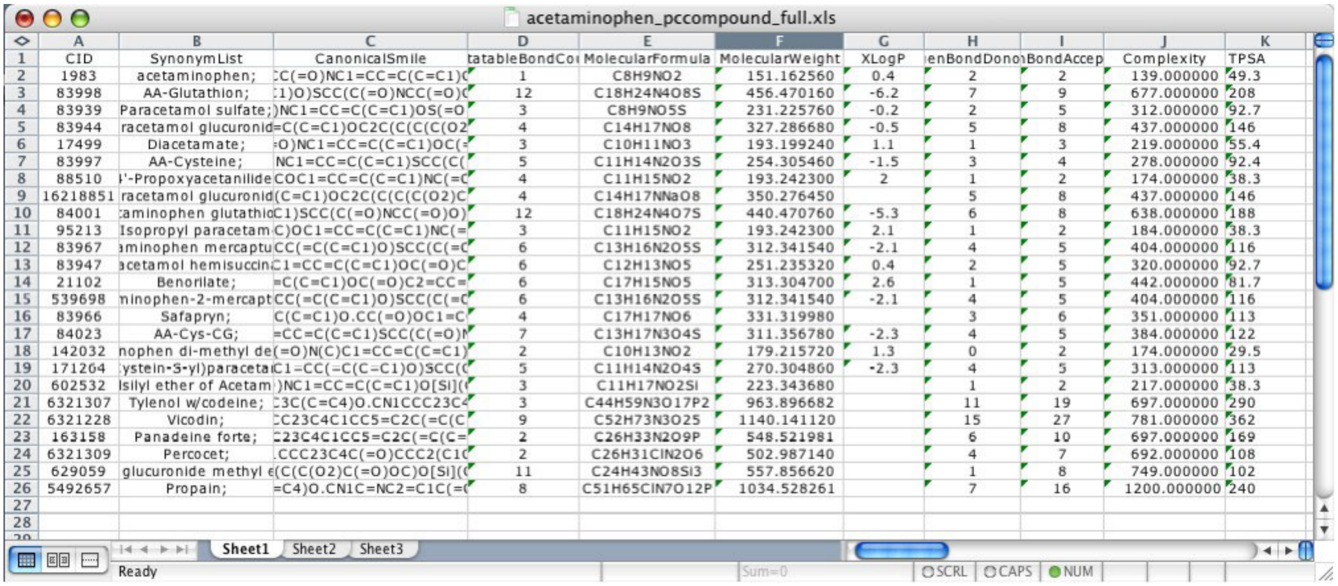
- Interactive refinement of searches

- Facility to export information to simple text or Microsoft Excel files and to specifically include or exclude individual data fields

- The ability to easily retrieve compounds that are active or inactive (or both) in particular bioassays

Implementation

PubChemSR (Search and Retrieve) is written in Microsoft .NET Visual Basic 2005 [6] and retrieves information from the PubChem database using the NCBI Entrez [7] web



acetaminophen_pccompound_full.xls

	A	B	C	D	E	F	G	H	I	J	K
1	CID	SynonymList	CanonicalSmile	tatableBondCo	MolecularFormula	MolecularWeight	XLogP	enBondDono	BondAccep	Complexity	TPSA
2	1983	acetaminophen;	CC(=O)NC1=CC=C(C=C1)C	1	C8H9NO2	151.162560	0.4	2	2	139.000000	49.3
3	83998	AA-Glutathion;	:1)O)SCC(C(=O)NCC(=O)C	12	C18H24N4O8S	456.470160	-6.2	7	9	677.000000	208
4	83939	Paracetamol sulfate;	)NC1=CC=C(C=C1)OS(=O	3	C8H9NO5S	231.225760	-0.2	2	5	312.000000	92.7
5	83944	racetamol glucuronid	=C(C=C1)OC2C(C(C(C(O2	4	C14H17NO8	327.286680	-0.5	5	8	437.000000	146
6	17499	Diacetamate;	=O)NC1=CC=C(C=C1)OC(=	3	C10H11NO3	193.199240	1.1	1	3	219.000000	55.4
7	83997	AA-Cysteine;	NC1=CC=C(C=C1)SCC(C(	5	C11H14N2O3S	254.305460	-1.5	3	4	278.000000	92.4
8	88510	I'-Propoxyacetanilide	COC1=CC=C(C=C1)NC(=C	4	C11H15NO2	193.242300	2	1	2	174.000000	38.3
9	16218851	racetamol glucuronid	(C=C1)OC2C(C(C(C(O2)C	4	C14H17NNaO8	350.276450		5	8	437.000000	146
10	84001	:aminophen glutathio	C1)SCC(C(=O)NCC(=O)O)	12	C18H24N4O7S	440.470760	-5.3	6	8	638.000000	188
11	95213	Isopropyl paracetam	C)OC1=CC=C(C=C1)NC(=	3	C11H15NO2	193.242300	2.1	1	2	184.000000	38.3
12	83967	aminophen mercaptu	CC(=C(C=C1)O)SCC(C(=C	6	C13H16N2O5S	312.341540	-2.1	4	5	404.000000	116
13	83947	acetamol hemisuccin	C1=CC=C(C=C1)OC(=O)C	6	C12H13NO5	251.235320	0.4	2	5	320.000000	92.7
14	21102	Benorilate;	=C(C=C1)OC(=O)C2=CC=	6	C17H15NO5	313.304700	2.6	1	5	442.000000	81.7
15	539698	minophen-2-mercapt	CC(=C(C=C1)O)SCC(C(=C	6	C13H16N2O5S	312.341540	-2.1	4	5	404.000000	116
16	83966	Safapryn;	C(C=C1)O.CC(=O)OC1=C	4	C17H17NO6	331.319980		3	6	351.000000	113
17	84023	AA-Cys-CG;	=CC=C(C=C1)SCC(C(=O)N	7	C13H17N3O4S	311.356780	-2.3	4	5	384.000000	122
18	142032	nophen di-methyl de	(=O)N(C)C1=CC=C(C=C1)	2	C10H13NO2	179.215720	1.3	0	2	174.000000	29.5
19	171264	:ystein-S-yl)paraceta	C1=CC(=C(C=C1)O)SCC(C	5	C11H14N2O4S	270.304860	-2.3	4	5	313.000000	113
20	602532	lsilyl ether of Acetam	)NC1=CC=C(C=C1)O[Si]((	3	C11H17NO2Si	223.343680		1	2	217.000000	38.3
21	6321307	Tylenol w/codeine;	C3C(C=C4)O.CN1CCC23C4	3	C44H59N3O17P2	963.896682		11	19	697.000000	290
22	6321228	Vicodin;	CC23C4C1CC5=C2C(=C(C	9	C52H73N3O25	1140.141120		15	27	781.000000	362
23	163158	Panadeine forte;	C23C4C1CC5=C2C(=C(C=	2	C26H33N2O9P	548.521981		6	10	697.000000	169
24	6321309	Percocet;	.CCC23C4C(=O)CCC2(C1C	2	C26H31ClN2O6	502.987140		4	7	692.000000	108
25	629059	glucuronide methyl e	(C(C(O2)C(=O)OC)O[Si]((	11	C24H43NO8Si3	557.856620		1	8	749.000000	102
26	5492657	Propain;	=C4)O.CN1C=NC2=C1C(=C	8	C51H65ClN7O12P	1034.528261		7	16	1200.000000	240
27											
28											

Figure 2

Property data exported into Microsoft Excel. Selected property-related fields of the 25 'acetaminophen' related compounds were exported into an Excel file. The filtering, sorting and graphing features of Excel can then be used to examine this data.

service version 1.5a via a SOAP interface [8]. It is compatible with Windows XP and the newer Windows Vista. We chose .NET [9] as it enables the maximum flexibility in design of user interface, and makes use of the SOAP protocol straightforward. The major limitation of this approach is that the program can only be used in a Windows environment.

The Microsoft .NET Framework is a software component which provides a plethora of pre-coded solutions to common software development requirements, and manages the execution of applications written for the framework. The deployment size of an application is small since the application can be executed in the runtime environment with .NET framework installed on a user's side.

SOAP (Simple Object Access Protocol or lately also know as Service Oriented Architecture Protocol) is a protocol allowing XML (Extensible Markup Language) based communication over computer networks using the World Wide Web's Hypertext Transfer Protocol (HTTP). One advantage of using SOAP is that it allows easier communication through firewalls and proxies since SOAP runs through HTTP requests that ensure unblocked communication with other programs anywhere. SOAP is one of the languages that enable the deployment of web services for remote access and execution of code. Web services have proven useful in both bioinformatics, and more recently, in cheminformatics [10] for the flexible interaction of distributed data and computation components.

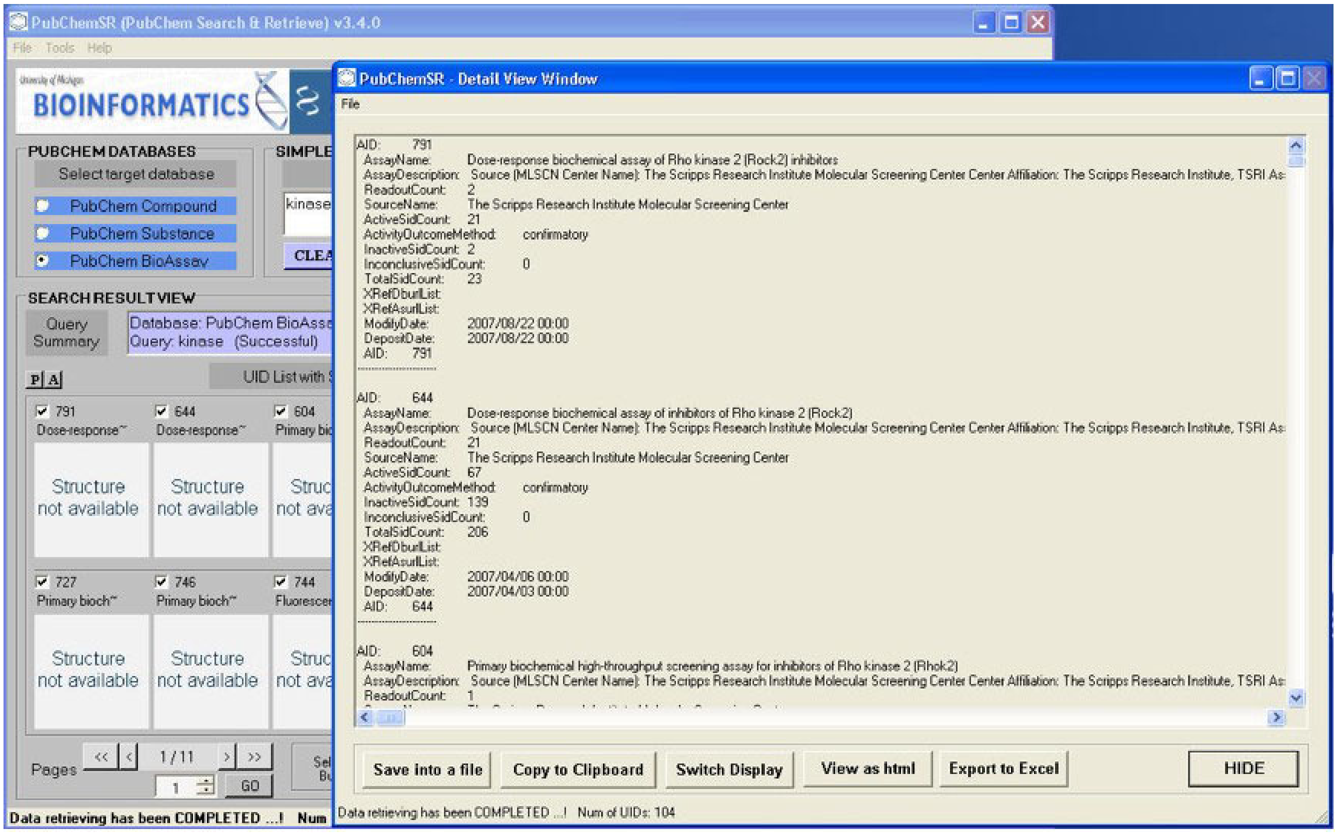
NCBI provides a collection of web services that allow programmable access and query to the Entrez data. These Entrez Programming Utilities, or eUtils, include EInfo, ESearch, EPost, ESummary, EFetch, ELink, EGQuery, ESpell and they are all wrapped into SOAP interface for easier communication. This is the primary mechanism used by PubChemSR for data retrieval. For structure search and BioAssay data retrieval that is not supported through the NCBI SOAP interface, PubChemSR performs such tasks in the background by directly accessing the NCBI's web server.

The JME (Java Molecular Editor) [11] written by Peter Ertl of Novartis is used to draw structure queries and to convert them into SMILES strings. PubChemSR allows users to interact with the JME applet at the PubChemSR web page [12] or the standalone version that comes with the PubChemSR distribution package. The latter requires the JAVA runtime environment to be available on a user's machine [13].

Results and discussion

Search Modes

PubChemSR employs a GUI (Graphical User Interface) with reasonably self-explanatory sections and buttons. It currently supports the three different search modes: simple text search mode (in the main window), structure search mode (in the main window), and batch search mode (through the Tools menu). The simple text search mode and structure search mode provide the same search



Database: PubChem BioAssa
Query: kinase (Successful)

Data retrieving has been COMPLETED ...! Num

```
AID:      791
  AssayName:       Dose-response biochemical assay of Rho kinase 2 (Rock2) inhibitors
  AssayDescription:  Source (MLSCN Center Name): The Scripps Research Institute Molecular Screening Center Center Affiliation: The Scripps Research Institute, TSRI As
  ReadoutCount:    2
  SourceName:      The Scripps Research Institute Molecular Screening Center
  ActiveSidCount:  21
  ActivityOutcomeMethod:     confirmatory
  InactiveSidCount:  2
  InconclusiveSidCount:      0
  TotalSidCount:   23
  XRefDburlList:
  XRefAsurlList:
  ModifyDate:      2007/08/22 00:00
  DepositDate:     2007/08/22 00:00
  AID:    791
-----------------------

AID:      644
  AssayName:       Dose-response biochemical assay of inhibitors of Rho kinase 2 (Rock2)
  AssayDescription:  Source (MLSCN Center Name): The Scripps Research Institute Molecular Screening Center Center Affiliation: The Scripps Research Institute, TSRI As
  ReadoutCount:    21
  SourceName:      The Scripps Research Institute Molecular Screening Center
  ActiveSidCount:  67
  ActivityOutcomeMethod:     confirmatory
  InactiveSidCount:  139
  InconclusiveSidCount:      0
  TotalSidCount:   206
  XRefDburlList:
  XRefAsurlList:
  ModifyDate:      2007/04/06 00:00
  DepositDate:     2007/04/03 00:00
  AID:    644
-----------------------

AID:      604
  AssayName:       Primary biochemical high-throughput screening assay for inhibitors of Rho kinase 2 (Rhok2)
  AssayDescription:  Source (MLSCN Center Name): The Scripps Research Institute Molecular Screening Center Center Affiliation: The Scripps Research Institute, TSRI As
  ReadoutCount:    1
```

Data retrieving has been COMPLETED ...! Num of UIDs: 104

Figure 3
Browsing the results after a search of the bioassay database for kinase-related assays.

functionality as the NCBI's Entrez or PubChem basic structure search, while the batch search mode extends the batch Entrez in ways enabling users to run a list of queries and merge the results into a single file.

URL Analyzer

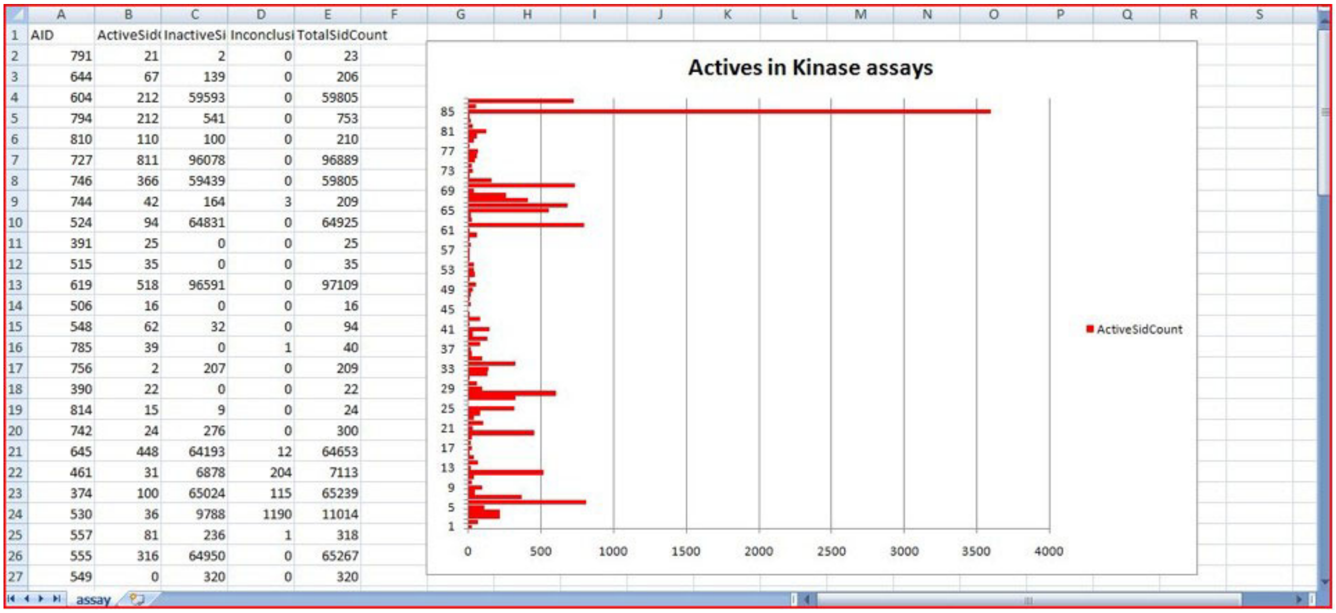
URL Analyzer can retrieve search results and display them in the search result view panel after users perform searches in their web-browser. The full URL of the results web-page can be copied into the clipboard using 'Copy' or 'Ctrl+C'. The user can then paste the URL into the URL analyzer by clicking the *Get* button or by pasting into the box. The *Anal* button will check the URL and retrieve the search results into the preview panel. This feature becomes extremely useful when a search can not be completed within a specified time (default is 120 seconds) or is not supported in PubChemSR. Such examples include structure searches for similar/substructure compounds or advanced structure searches supporting additional filters like chemical property or BioActivity.

Bulk Download

Bulk download enables users to download information on compounds *en masse* and only export the desired data fields for each compound. Needed are a list of UIDs (Unique Identifiers: CID for compounds, SID for substances, and AID for BioAssay), which can be obtained through the simple text search or be uploaded from a file. The buttons in the 'Retrieve' panel will either directly save the data into a text file or display them first in a separate window giving further options to export the data into Microsoft Excel or HTML file.

Other features

Several other available features are offered by the program including *term correction for misspelled queries* – misspelled queries can be automatically corrected via NCBI E-spell web-service; *selectable data field* – for bulk download, the results can be filtered to only include fields of interest to the user; *preview with picture* – the search result view panel provides a summary of the results ten compounds at a time with preview of structure and selected data fields; and *BioAssay retriever* – retrieves the actual bioassay activity data and exports them along with selected compounds/substance data fields to Microsoft Excel or text files.



AID	ActiveSid(	InactiveSi	Inconclusi	TotalSidCount
791	21	2	0	23
644	67	139	0	206
604	212	59593	0	59805
794	212	541	0	753
810	110	100	0	210
727	811	96078	0	96889
746	366	59439	0	59805
744	42	164	3	209
524	94	64831	0	64925
391	25	0	0	25
515	35	0	0	35
619	518	96591	0	97109
506	16	0	0	16
548	62	32	0	94
785	39	0	1	40
756	2	207	0	209
390	22	0	0	22
814	15	9	0	24
742	24	276	0	300
645	448	64193	12	64653
461	31	6878	204	7113
374	100	65024	115	65239
530	36	9788	1190	11014
557	81	236	1	318
555	316	64950	0	65267
549	0	320	0	320

Figure 4
An analysis of kinase-related assays in Excel. Here a graph is used to compare the numbers of active compounds in each of the assays.

Examples of Use

There are many ways that PubChemSR can be used to simplify the process of obtaining information from PubChem. Below are listed a few examples of how it can be employed for common tasks.

Comparing chemical properties of related compounds

It is often useful to compare the properties of compounds in a particular structural class. This is very easy to do using the refinement and Excel export functions. Figures 1 and 2 show respectively a search for 'acetaminophen' using PubChemSR, and an Excel spreadsheet created by exporting selected property-related fields from the program. This kind of comparison may also be done with a substructure or similarity search instead of a simple text search.

Browsing bioassays related to kinases, and downloading active compounds in specific assays

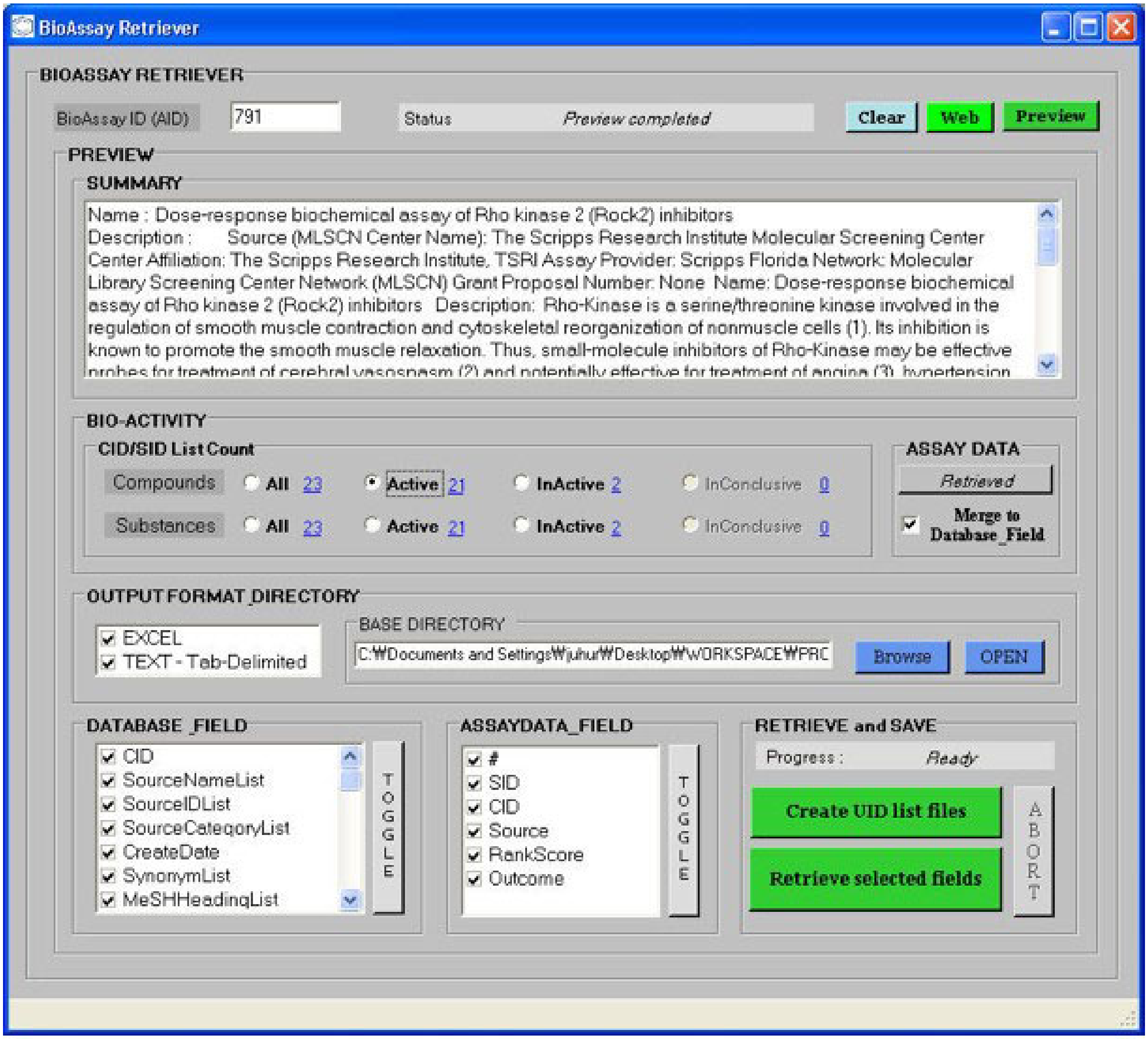
Using a text search on the PubChem BioAssay database, one can find all of the assay descriptions that contain particular keywords such as "*Kinase*". One can then export all of these descriptions to Excel or a text file, or browse them from within the program (as shown in Figure 3) In particular, one can download statistics of assays (counts of active and inactive structures and so on) and use Excel to analyze these (see Figure 4). Upon finding assays of interest, one can retrieve all of the compounds (and related information) that have been flagged as showing activity in that assay by supplying the assay ID to the bioassay retriever as shown in Figure 5. These compounds can then be exported just as with a regular compound search.

Creating a SMILES and activity file for SAR study of an assay

SMILES is a linear text string representation of the 2D chemical structure of a compound. A SMILES file usually contains the SMILES string and name for a compound. When a third column is added that contains biological activity values for a compound, it is a useful format for input into a variety of cheminformatics techniques that can automatically determine structure-activity relationships (SAR) in compounds. Using the BioAssay Retriever, one can download just the SMILES, name, and biological assay results for compounds and then create a simple tab-delimited file that can be loaded into cheminformatics tools.

Conclusion

We believe PubChemSR is an extremely useful and straightforward tool that bridges a gap between the needs of bench scientists and the rich information resource of PubChem. We have shown how it can be used to export and explore compound, property and bioassay information in the database. PubChemSR is not intended to replace the web-based PubChem interface, and there are certain features which are only available in the web-based PubChem interface such as structure clustering or structure-activity analysis in detailed BioAssay summary pages. PubChemSR has been designed to aid users, especially



BIOASSAY RETRIEVER

BioAssay ID (AID): 791

Status: *Preview completed*

PREVIEW

SUMMARY

Name : Dose-response biochemical assay of Rho kinase 2 (Rock2) inhibitors
Description : Source (MLSCN Center Name): The Scripps Research Institute Molecular Screening Center Center Affiliation: The Scripps Research Institute, TSRI Assay Provider: Scripps Florida Network: Molecular Library Screening Center Network (MLSCN) Grant Proposal Number: None Name: Dose-response biochemical assay of Rho kinase 2 (Rock2) inhibitors Description: Rho-Kinase is a serine/threonine kinase involved in the regulation of smooth muscle contraction and cytoskeletal reorganization of nonmuscle cells (1). Its inhibition is known to promote the smooth muscle relaxation. Thus, small-molecule inhibitors of Rho-Kinase may be effective probes for treatment of cerebral vasospasm (2) and potentially effective for treatment of angina (3), hypertension

BIO-ACTIVITY

CID/SID List Count

Compounds	○ All 23	● Active 21	○ InActive 2	○ InConclusive 0
Substances	○ All 23	○ Active 21	○ InActive 2	○ InConclusive 0

ASSAY DATA

Retrieved

☑ Merge to Database_Field

OUTPUT FORMAT DIRECTORY

☑ EXCEL
☑ TEXT - Tab-Delimited

BASE DIRECTORY: C:₩Documents and Settings₩juhur₩Desktop₩WORKSPACE₩PRC

DATABASE_FIELD

☑ CID
☑ SourceNameList
☑ SourceIDList
☑ SourceCategoryList
☑ CreateDate
☑ SynonymList
☑ MeSHHeadingList

ASSAYDATA_FIELD

☑ #
☑ SID
☑ CID
☑ Source
☑ RankScore
☑ Outcome

RETRIEVE and SAVE

Progress : *Ready*

Figure 5
Retrieving active and inactive compounds for a particular assay using the bioassay retriever in PubChemSR.

non-computationally experienced, to search, retrieve, export, and manipulate the PubChem data in more efficient and convenient ways.

Availability and requirements

Project name: PubChemSR

Project home page: http://cheminfo.informatics.indiana.edu/PubChemSR/; http://sourceforge.net/projects/pubchemsr/

Operating system: Windows XP or Vista

Programming language: Microsoft Visual Basic .NET

Other requirements: Microsoft .Net 2.0

License: GNU General Public License version 3 http://www.gnu.org/licenses/gpl.html.

Any restrictions on use by non-academics: The tool may not be used for commercial purposes

Authors' contributions

The program was fully developed by JH initially under the supervision of DW. Both contributed to this paper.

Acknowledgements

We thank Dr. Peter Ertl at Novartis Institutes of Biomedical Research for letting us incorporate his JME Molecular Editor into PubChemSR. We also thank Dr. Gary Wiggins for his insightful discussion and recommendation, and Eunyong Park for helping us integrate the structure-search functionality into PubChemSR.

References

1. Baykoucheva S: **A New Era in Chemical Information: PubChem, DiscoveryGate, and Chemistry Central.** *ONLINE* Sep/Oct edition. 2007, **31:**.
2. Irwin JJ, Shoichet BK: **ZINC – a free database of commercially available compounds for virtual screening.** *Journal of chemical information and modeling* 2005, **45(1):**177-182.
3. **The PubChem Project** [http://pubchem.ncbi.nlm.nih.gov]
4. Austin CP, Brady LS, Insel TR, Collins FS: **NIH Molecular Libraries Initiative.** In *Science Volume 306*. Issue 5699 New York, NY; 2004:1138-1139.
5. **Extensible Markup Language (XML)** [http://www.w3.org/XML/]
6. **Visual Basic Developer Center** [http://msdn.microsoft.com/vbasic/]
7. **Entrez Programming Utilities** [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html]
8. **SOAP Specifications** [http://www.w3.org/TR/soap/]
9. **Microsoft .NET Homepage** [http://www.microsoft.com/net/]
10. Dong X, Gilbert KE, Guha R, Heiland R, Kim J, Pierce ME, Fox GC, Wild DJ: **Web service infrastructure for chemoinformatics.** *Journal of chemical information and modeling* 2007, **47(4):**1303-1307.
11. Ertl P, Jacob O: **WWW-based chemical information system.** *Journal of Molecular Structure: THEOCHEM* 1997, **419:**113-120.
12. **PubChemSR JME Editor** [http://cheminfo.informatics.indiana.edu/PubChemSR/JME/PubChemSRJME.html]
13. **JAVA** [http://java.sun.com/]

Publish with **Chemistry**Central and every scientist can read your work free of charge

"Open access provides opportunities to our colleagues in other parts of the globe, by allowing anyone to view the content free of charge."

W. Jeffery Hurst, The Hershey Company.

- available free of charge to the entire scientific community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.chemistrycentral.com/manuscript/

ChemistryCentral