


SOFTWARE

Open Access



# A practical Java tool for small-molecule compound appraisal

Parisa Amani<sup>1</sup>, Todd Sneyd<sup>1</sup>, Sarah Preston<sup>2</sup>, Neil D Young<sup>2</sup>, Lyndel Mason<sup>1</sup>, Ulla-Maja Bailey<sup>1</sup>, Jonathan Baell<sup>3</sup>, David Camp<sup>4</sup>, Robin B Gasser<sup>2</sup>, Alain-Dominique Gorse<sup>5</sup>, Paul Taylor<sup>6</sup> and Andreas Hofmann<sup>1,2\*</sup> 

## Abstract

**Background:** The increased use of small-molecule compound screening by new users from a variety of different academic backgrounds calls for adequate software to administer, appraise, analyse and exchange information obtained from screening experiments. While software and spreadsheet solutions exist, there is a need for software that can be easily deployed and is convenient to use.

**Results:** The Java application cApp addresses this need and aids in the handling and storage of information on small-molecule compounds. The software is intended for the appraisal of compounds with respect to their physico-chemical properties, analysis in relation to adherence to likeness rules as well as recognition of pan-assay interference components and cross-linking with identical entries in the PubChem Compound Database. Results are displayed in a tabular form in a graphical interface, but can also be written in an HTML or PDF format. The output of data in ASCII format allows for further processing of data using other suitable programs. Other features include similarity searches against user-provided compound libraries and the PubChem Compound Database, as well as compound clustering based on a MaxMin algorithm.

**Conclusions:** cApp is a personal database solution for small-molecule compounds which can handle all major chemical formats. Being a standalone software, it has no other dependency than the Java virtual machine and is thus conveniently deployed. It streamlines the analysis of molecules with respect to physico-chemical properties and drug discovery criteria; cApp is distributed under the GNU Affero General Public License version 3 and available from <http://www.structuralchemistry.org/pcsb/>. To download cApp, users will be asked for their name, institution and email address. A detailed manual can also be downloaded from this site, and online tutorials are available at <http://www.structuralchemistry.org/pcsb/capp.php>.

**Keywords:** Compound appraisal, Molecular properties, Personal database

## Background

Screening of organic small-molecule compounds has been a pivotal activity in the pharmaceutical industry as part of the drug discovery process. In the last decade, compound screening has increasingly been established and employed by academic laboratories due to many disease areas not being tackled by commercially oriented pharmaceutical industry, and also due to the availability of advanced technologies for the probing of biological systems [1].

The use of chemical tools and compound screening has therefore found new user clienteles, not all of whom are expert medicinal chemists and thus familiar with the properties of organic molecules. Recently, Baell and colleagues [2] highlighted a significant problem arising from the massively increased, non-expert compound screening in that molecules with promiscuous activities (pan-assay interference compounds, PAINs) are frequently being reported in the literature as (potential) hits in an indiscriminating fashion.

The concept of chemical spreadsheets is well established, and several different products have been developed in the past [3] that will store chemical data and

\*Correspondence: a.hofmann@griffith.edu.au

<sup>1</sup> Structural Chemistry Program, Eskitis Institute, Griffith University, Nathan, QLD, Australia

Full list of author information is available at the end of the article

present in a tabular form. Most such software is available from commercial providers, but there have also been freeware products, and increasingly web services provided by databases, such as ChemSpider [4] and the CCD Vault [5].

In the recent past, the concept of workflow has been implemented in many bio- and chemo-informatics approaches [6, 7]. Here, activities are classified into generic tasks that can be addressed by modular algorithms and thus combined by the end-user in a flexible fashion. Products in this category include the commercially available Pipeline Pilot (Accelrys, US) or InforSense (InforSense, UK). A freeware alternative is KNIME (Knime.com, Switzerland), based on the open source Eclipse platform, and CDK-Taverna [8] which builds on the Java libraries of the Chemistry Development Kit (CDK) [9].

Our own experience in collaborative work among medicinal chemistry, structural biology and biochemistry laboratories shows that data exchange, collection, archiving and publishing is very much done on a case-by-case basis, whereby simple tasks are often done repetitively and in many cases redundantly. Although the above spreadsheet or workflow software is able to deal with the requirements arising from drug screening projects in the academic setting, the actual deployment of such software by end-users is often hampered by access/availability, difficulty of installation and/or the perceived or real difficulty to learn how to use the software.

We set out to design a platform-independent Java application, based on our in-house developed collection PCSB [10], that should appeal to non-expert laboratories engaged in the handling of medium-sized compound libraries. Particular attention has been paid to making the learning and use of this software as convenient as possible. The portable Java application cApp enables the appraisal of compounds sourced from the commonly used formats of SMILES (simplified molecular-input line entry system; see specifications at [11]), InChI (International Chemical Identifier; see specification at [12]) and SDF (Structure Data Format; see Chemical Table File specification from December 2011 at [13]) files with respect to adherence to likeness rules. Compounds can also be input or manipulated via the embedded JChemPaint [14] chemical editor. Particular innovative features built into cApp are the identification of PAIN components in the appraised compounds, direct queries of the PubChem Compound Database [15] as well as similarity searches initiated with one mouse click.

## Implementation

cApp has been implemented in Java for maximum portability, capitalising on existing chemo- and bio-informatic

Java libraries, namely the CDK [9], JChemPaint [14] and PCSB [10]. The data structure within cApp rests on the custom-programmed *Compound* object that handles all data relating to individual small-molecule compounds for this software. Access to the PubChem Compound Database is through the PubChem Power User Gate (PUG), which is an XML-based communication gateway to interrogate the database.

## Results

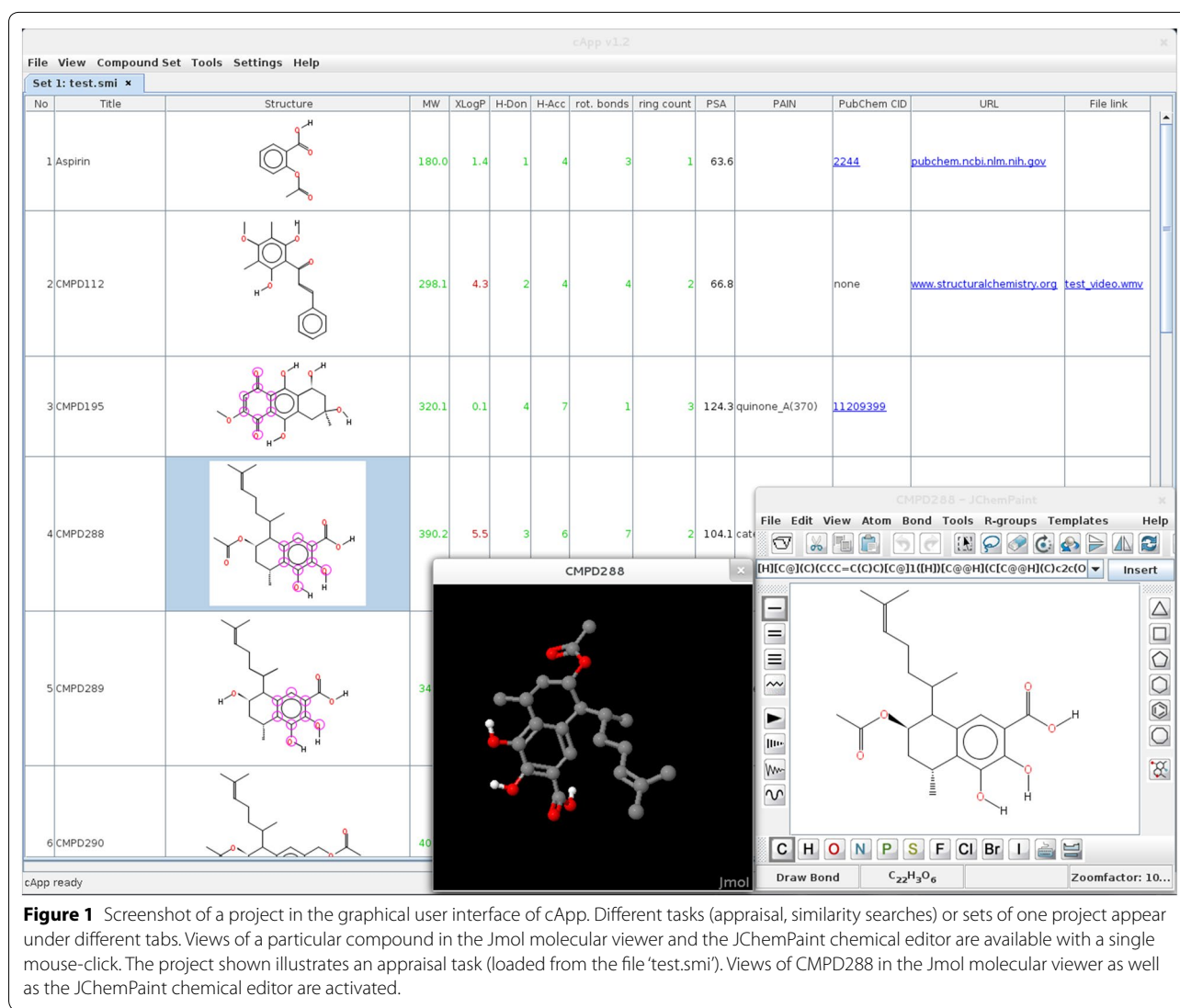
### Software features

cApp is a personal compound database software that allows the user to compare chemical descriptors and similarities of compounds, but also to annotate compound lists with their own data and information. A cApp project comprises all data and compound sets of a software session; a compound set is a particular list of compounds. In the GUI, a compound set is displayed as a table on a particular tab (see Figure 1). Automatically generated HTML, PDF and ASCII presentations of compound sets are identified by their set number. Conceptually, its functionality is divided up into tasks, presentation of results and convenience features. In the present version, the tasks of compound appraisal, similarity search and clustering can be performed. The compound appraisal task calculates physico-chemical properties and structural features, an analysis for compliance with various likeness criteria (drug-, lead- or fragment-like) [16] and the identification of PAINs components [17] using the SMSD maximum common subgraph (MCS) Tanimoto coefficient as criterion [18]. Similarity searches against user-provided libraries can be conducted using an MCS approach which builds on the CDK Fingerprint Tanimoto coefficient [18] or the PubChem Compound Database. For compound clustering, a MaxMin algorithm with subsequent k-Means clustering [19] has been implemented, based on the CDK Fingerprint Tanimoto coefficient as property. The user can annotate compounds with extra information by adding three types of data in additional columns containing either free text, a file link or a URL. Linked files and web content are available with a mouse click from the cApp GUI via the user's preferred web browser.

The individual features of cApp are described in a detailed manual that is available together with the application (see also the Additional file 1). Online tutorials for typical scenarios have been prepared and can be accessed at the project web site.

### Assessment of similarity with pan-assay interference compounds (PAINs)

Baell and Holloway [17] have identified a set of chemical substructures that are frequently observed as effectors



in compound screening and thus deemed to be promising. In the compound appraisal task, cApp conducts SMARTS queries using 480 PAINs substructure filters that have been translated from the original rules in Sybyl Line Notation (sln) by Dr Rajarshi Guha (<http://blog.rguha.net/?p=850>). This conversion of the PAINs substructure filters from sln to SMARTS does not reproduce the original rules perfectly. For the present version of cApp, we have combined the three filters sets obtained from [20] into one set (pains.smt).

We have subjected a library of 50,000 compounds from the ChemBridge catalogue to PAINs filtering using the same SMARTS filters in cApp and PipelinePilot [21]. We also compared the results of PAINs-filtering in cApp with

those obtained by the original sln rules. The results from this benchmarking indicate that there are small variations in the queries conducted by different software (see Table 1).

## Conclusions

With cApp, we have developed a personal, small-molecule database management software that should appeal to the non-expert user due to its ease of installation, intuitive handling and convenient execution of tasks. In future versions, we plan to include additional functionality, such as identification of duplicate entries, and direct query capability of further public compound repositories, such as ChEMBL and others.

**Table 1 Comparison of PAINs identification by different software/methodologies using a library of 50,000 compounds from the ChemBridge catalogue**

Software	cApp v1.2	Sybyl	Matching entries
Rules	pains.smt [20]	sln [17]	
No of PAINs	5,790	6,001	5,788
Hits identified only in one approach	2	213	
Software	cApp v1.2	Pipeline Pilot	Matching entries
Rules	pains.smt [20]	pains.smt [20]	
No of PAINs	5,790	5,994	5,782
Hits identified only in one approach	8	212	

### Availability and requirements

Project name: cApp.

Project home page: <http://www.structuralchemistry.org/pcsb/capp.php>.

Operating system(s): Platform independent.

Programming language: Java.

Other requirements: Java 1.7 or higher.

License: GNU AGPL v3.

Any restrictions to use by non-academics: None.

### Additional file

**Additional file 1:** The software manual accompanies this paper as supplementary information.

### Authors' contributions

AH, RBG, SP and PA designed the project with critical input from all authors; PA, TS, AH and PT wrote and compiled the code; all authors tested the software. PA and AH wrote the paper with contributions from all authors. All authors read and approved the final manuscript.

### Author details

<sup>1</sup> Structural Chemistry Program, Eskitis Institute, Griffith University, Nathan, QLD, Australia. <sup>2</sup> Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, Australia. <sup>3</sup> Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences (MIPS), Monash University, Parkville, VIC, Australia. <sup>4</sup> Griffith School of Environment, Griffith University, Nathan, QLD, Australia. <sup>5</sup> Queensland Facility for Advanced Bioinformatics, Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD, Australia. <sup>6</sup> School of Biological Sciences, The University of Edinburgh, Edinburgh, Scotland, UK.

### Acknowledgements

AH's research is funded by the National Health and Medical Research Council (NHMRC), the Australian Research Council (ARC) and the Rebecca L Cooper Medical Research Foundation. RBG's research is funded mainly through the ARC, NHMRC, Melbourne Water Corporation and Yougene Bioscience, and supported by a Victoria Life Sciences Computation Initiative (VLSI; grant number VR0007) on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government. We gratefully acknowledge advice from Duncan Bucknell (<http://www.duncanbucknell.com/>).

### Compliance with ethical guidelines

### Competing interests

The authors declare that they have no competing interests.

Received: 31 March 2015 Accepted: 27 May 2015

Published online: 16 June 2015

### References

- Hofmann A, Wang CK, Osman A, Camp D (2010) Merging structural biology with chemical biology: structural chemistry at Eskitis. *Struct Chem* 21:1117–1129
- Baell J, Walters MA (2014) Chemical con artists foil drug discovery. *Nature* 513:481–483
- Apodaca R (2008) Your favorite chemical spreadsheet. *Depth First*. <http://www.depth-first.com/articles/2008/09/12/your-favorite-chemical-spreadsheet>
- ChemSpider. <http://www.chemspider.com>
- CCD Vault. <http://www.collaboratedrug.com>
- Tiwari A, Sekhar AKT (2007) Workflow based framework for life science informatics. *Comput Biol Chem* 31:305–319
- Warr WA (2012) Scientific workflow systems: Pipeline Pilot and KNIME. *J Comput Aided Mol Des* 26:801–804
- Kuhn T, Willighagen EL, Zielesny A, Steinbeck C (2010) CDK-Taverna: an open workflow environment for cheminformatics. *BMC Bioinform* 11:159
- Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr Pharm Des* 12:2111–2120
- Hofmann A, Wlodawer A (2002) PCSB—a program collection for structural biology and biophysical chemistry. *Bioinformatics* 18:209–210
- OpenSMILES. <http://www.opensmiles.org>
- The IUPAC International Chemical Identifier (InChI). <http://www.iupac.org/inchi>
- Chemical Table File specification.L. <http://www.download.accelrys.com/freeware/ctfile-formats/ctfile-formats.zip>
- Krause S, Willighagen E, Steinbeck C (2000) JChemPaint—using the collaborative forces of the internet to develop a free editor for 2D chemical structures. *Molecules* 5:93–98
- Bolton E, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. In: *Annual reports in computational chemistry*, vol 4. Elsevier, Oxford, pp 217–240
- Barker J, Hestekamp T, Whittaker M (2008) Integrating HTS and fragment-based drug discovery. *Drug Discov World* 9:69–75
- Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* 53:2719–2740
- Asad Rahman S, Bashton M, Holliday GL, Schrader R, Thornton JM (2009) Small Molecule Subgraph Detector (SMSD) Toolkit. *J. Cheminform* 1:12
- Gorse D, Rees A, Kaczorek M, Lahana R (1999) Molecular diversity and its analysis. *Drug Discov Today* 4:257–264
- Guha R (2010) PAINs SMARTS filters. <http://blog.rguha.net/?p=850>. Accessed 9 Feb 2015
- BIOVIA (2013) Pipeline Pilot V9.1. Dassault Systèmes, San Diego