doi:10.1093/nar/gkr320

ChemMine tools: an online service for analyzing and clustering small molecules

Tyler W. H. Backman[1], Yiqun Cao[2] and Thomas Girke[1,*]

[1]Department of Botany and Plant Sciences and [2]Department of Computer Science and Engineering, University of California Riverside, Riverside, CA 92521, USA

Received February 27, 2011; Revised April 10, 2011; Accepted April 20, 2011

ABSTRACT

ChemMine Tools is an online service for small molecule data analysis. It provides a web interface to a set of cheminformatics and data mining tools that are useful for various analysis routines performed in chemical genomics and drug discovery. The service also offers programmable access options via the R library ChemmineR. The primary functionalities of ChemMine Tools fall into five major application areas: data visualization, structure comparisons, similarity searching, compound clustering and prediction of chemical properties. First, users can upload compound data sets to the online Compound Workbench. Numerous utilities are provided for compound viewing, structure drawing and format interconversion. Second, pairwise structural similarities among compounds can be quantified. Third, interfaces to ultra-fast structure similarity search algorithms are available to efficiently mine the chemical space in the public domain. These include fingerprint and embedding/indexing algorithms. Fourth, the service includes a Clustering Toolbox that integrates cheminformatic algorithms with data mining utilities to enable systematic structure and activity based analyses of custom compound sets. Fifth, physicochemical property descriptors of custom compound sets can be calculated. These descriptors are important for assessing the bioactivity profile of compounds *in silico* and quantitative structure—activity relationship (QSAR) analyses. ChemMine Tools is available at: http://chemmine.ucr.edu.

INTRODUCTION

Cheminformatics tools for analyzing small molecule screening data play an important role in many fields including chemical biology, chemical genomics, drug discovery and agrochemical research (1–3). Informatics resources in these areas are essential for exploring the structure, properties and bioactivity of biologically relevant molecules. To provide these capabilities, software tools are required for analyzing the structural similarities, physicochemical properties and bioactivity profiles of natural and synthetic compounds to gain insight into their modes of action in biological systems. This information is important for the development of effective small molecule probes for studying the functions of protein and cellular networks in chemical genomics and drug discovery research (4). In addition, similar informatics resources are required for identifying the structural and physicochemical relationships among compounds from metabolic or signaling pathways (5–7). The rapidly growing relevance of chemical genomics approaches for modern biology research has significantly increased demand for small molecule mining systems in academia (8).

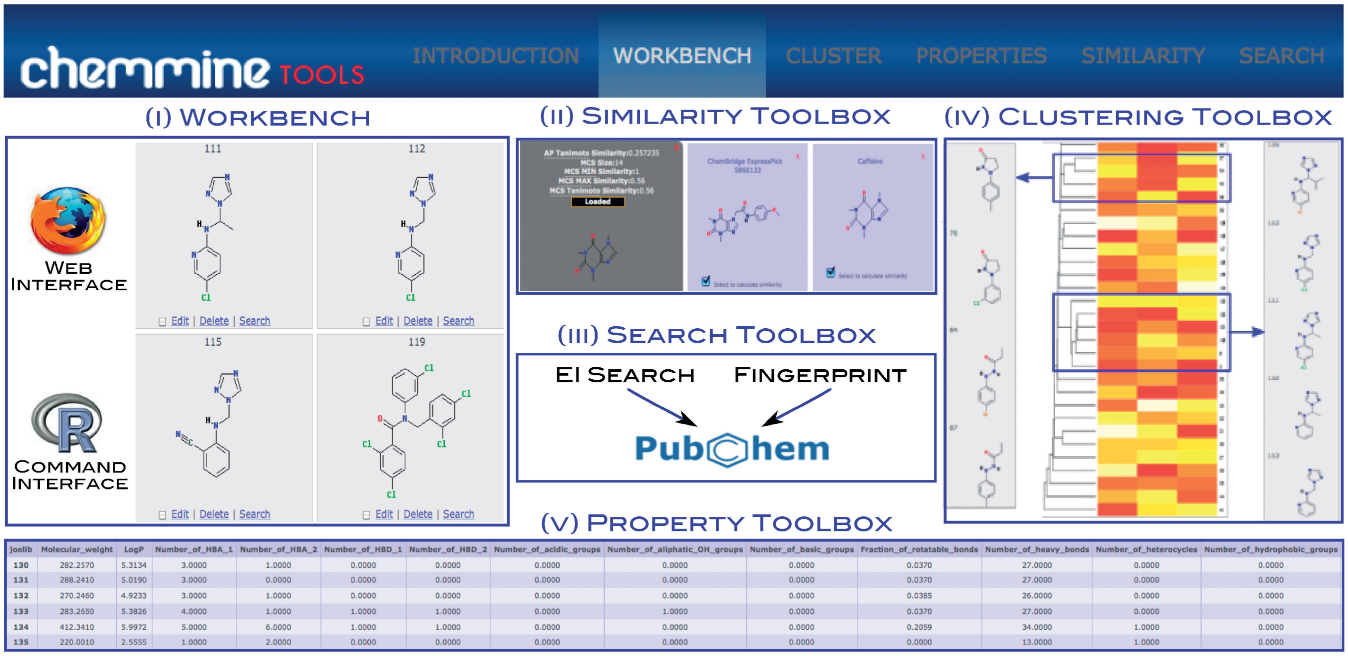
Currently, the structures of over 30 million distinct small molecules are available in open-access databases, including *PubChem*, *ChemBank* and many others (9–15). In addition, preliminary bioactivity data from hundreds of high-throughput screening (HTS) experiments against a wide spectrum of target sites have become available for almost one million compounds in the bioassay sections of various public databases (see below; 9,10,15,16). To efficiently analyze these resources, the development of novel compound data mining and cheminformatic web services is essential.

While there has been extensive development of public domain small molecule databases in recent years (6,9–11, 13–24), the number of open access web services for analyzing public or custom small molecule data is extremely limited at this point (25,26). Thus far, most development has been focused on standalone software applications targeted toward computational rather than experimental scientists. These include *Open Babel* (27,28), the *Chemistry Development Kit* (29,30), the *Chemical Descriptors Library* (31) and *JOELib* (32).

*To whom correspondence should be addressed. Tel: +1 951 905 5232; Fax: +1 951 827 4437; Email: thomas.girke@ucr.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.



joelib	Molecular_weight	LogP	Number_of_HBA_1	Number_of_HBA_2	Number_of_HBD_1	Number_of_HBD_2	Number_of_acidic_groups	Number_of_aliphatic_OH_groups	Number_of_basic_groups	Fraction_of_rotatable_bonds	Number_of_heavy_bonds	Number_of_heterocycles	Number_of_hydrophobic_groups
130	282.2570	5.3134	3.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0370	27.0000	0.0000	0.0000
131	288.2410	5.0190	3.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0370	27.0000	0.0000	0.0000
132	270.2460	4.9233	3.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0385	26.0000	0.0000	0.0000
133	283.2650	5.3826	4.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0370	27.0000	0.0000	0.0000
134	412.3410	5.9972	5.0000	6.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.2059	34.0000	1.0000	0.0000
135	220.0010	2.5555	1.0000	2.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	13.0000	1.0000	0.0000

Figure 1. Illustration of the functionalities provided by *ChemMine Tools*. The utilities of the five application domains (i–v) are listed in more detail in Table 1.

Examples of software designed for non-expert users in this field are Chembench (33) for online quantitative structure—activity relationship (QSAR) modeling and KNIME (34) for designing data analysis pipelines.

Here, we present *ChemMine Tools* as an online portal to a variety of cheminformatics, visualization, search and clustering tools for small molecule data. The utilities provided by this service are useful for various analysis and data mining routines of small molecule screening experiments in chemical genomics and related areas. An easy to use web interface makes these tools accessible to experimental scientists without an extensive computational background.

METHODS

Conceptually, the *ChemMine Tools* online service is divided into five application domains (Figure 1 and Table 1): (i) a Compound workbench for data imports and result management; (ii) a Structure Similarity toolbox to quantify the similarities among compounds; (iii) a Search toolbox for retrieving similar compounds from PubChem; (iv) a Clustering toolbox for accessing clustering and data visualization tools; and (v) a Property toolbox for predicting physicochemical properties of compounds. To construct robust data analysis workflows, the back-end of the server employs a modular design architecture with object-oriented methods and container classes assuring compatible input/output flows and parameter settings among the different data processing units. Currently, the server integrates over 30 cheminformatics and data mining tools that were developed by this or related open source projects. The modular organization of the *ChemMine Tools* service has several advantages. For instance, it maximizes the transparency and maintainability of the system, and simplifies the addition of new features and analysis methods upon user request. The web interface of *ChemMine Tools* is written in Python using the object-oriented and highly scalable *Django* web framework. Modern *JavaScript/Ajax* utilities are embedded to generate interactive and customizable high-content web pages. Moreover, the *ChemMine Tools* project is dedicated to an open access and resource sharing policy. All of its online services and downloadable software components are freely available without restrictions. The following subsections give a detailed description of the underlying algorithms and software tools used by the individual *ChemMine Tools* services.

DISCUSSION OF SERVICES

Compound workbench

A central feature of *ChemMine Tools* is its Compound workbench. It provides a flexible online workspace to upload, manage and visualize small molecule data. Compounds can be imported by reading them from local files, copy and paste, *PubChem* queries (see Search toolbox) or by interacting with the service through the *ChemmineR* library (35) within the statistical programming environment *R*. The latter is an extension of the *ChemMine Tools* project to provide a programmable interface to more advanced users. Alternatively, compounds can be drawn online with the *JME Molecular Editor* (36) and then added to the Compound workbench. Currently, the import utility supports the structure data format (SDF) and simplified molecular

Table 1. List of services provided by *ChemMine Tools*

Functions	Program	Input	Output	Comments
(i) Compound workbench				
Structure import/export	*Open Babel*	Mouse clicks	SMILES/SDF	One or many compounds
Format interconversions	*Open Babel*	SDF/SMILES	SMILES/SDF	One or many compounds
Bioactivity data import	*JavaScript/Ajax*	Tabular data	Table/heat map	SAR table
Structure depictions	*CACTVS*	SMILES/SDF	Image file (GIF)	One or many compounds
Structure drawing	*JME Molecular Editor*	Mouse clicks	SMILES/SDF	Single compound
Database import	*SOAP*	XML/SDF	SMILES/SDF	PubChem
Scriptable access from *R*	*ChemmineR*[a]	SDF, tabular data	Online viewing	SAR table
(ii) Similarity toolbox				
Fragment-based similarity	*Atom Pairs*[a]	SDF/SMILES	Similarity coefficients	Pairwise comparisons
Maximum common substructure	*MCS*[a]	SDF/SMILES	MCS (SDF), similarity coefficient	Pairwise comparisons
(iii) Search toolbox				
Embedding and indexing	*EI Search*[a]	Mouse clicks, SDF/SMILES	Ranked compound list	Database search
Fingerprint search	*PubChem PUG*	Mouse clicks, SDF/SMILES	Ranked compound list	Database search
(iv) Clustering toolbox				
Binning clustering	*cmp.cluster*[a]	SDF/SMILES, custom table	Cluster table	
Hierarchical clustering	*hclust*	SDF/SMILES, custom table	Tree, distance matrix	Optional heat map
Multidimensional scaling	*cmdscale*	SDF/SMILES, custom table	Scatter plot	Interactive
(v) Property toolbox				
Physicochemical descriptors	*JOELib*	SDF/SMILES	Property table	38 descriptors

The names of software tools, libraries and environments are italicized.
[a]Programs developed by the *ChemMine Tools* project. Acronyms defined in text.

input line entry system (SMILES). After the import, one can organize and annotate the compounds or view their structure images in single or batch modes. These images are generated in real time from the underlying structure definition data using the structure depiction tool of the *CACTVS* software suite (11) which runs on the server side. To revisit instances of compound sets, users can save their workbench for later use by downloading the compounds to local files. The compound download function also serves as a format conversion tool to interconvert structure representations between SDF and SMILES formats using utilities from the *Open Babel* project (27,28). Once the user has populated the Compound workbench with structures, it serves as a central submission system to all downstream analysis services.

Similarity toolbox

In many small molecule screening data analysis routines it is important to compute objective similarity measures among compounds as a means to compare and prioritize structurally related lead compounds. To provide this functionality, *ChemMine Tools* has implemented two algorithms for computing similarity coefficients among compound structures. The first employs atom pairs as structural descriptors (37) and the widely used Tanimoto coefficient as a similarity measure (see below for more details). Alternatively, users can choose other similarity coefficients, such as Tversky or Dice (38). The second algorithm identifies the maximum common substructure (MCS) shared among compound pairs (39). Subsequently, the size of both compounds and the size of their shared MCS is used to calculate the available similarity coefficients. The underlying MCS algorithm often provides the most accurate and sensitive similarity measure, especially for compounds with large size differences (40,41).

Search toolbox

To efficiently mine much of the chemical structure and bioactivity space available in the public domain, the *ChemMine Tools* service provides text and structure similarity search methods that interface with the *PubChem* database (15) via its *SOAP*-based *Power User Gateway* (*PUG*) data exchange feature. During an analysis session, instantaneous search functionality is often important for retrieval of detailed property and annotation information for compounds of interest, or to identify related structures. In *ChemMine Tools*, structural similarity searches can be performed with *PubChem's* fingerprint search engine or via the *EI Search* method. The latter was developed in house as part of this project to provide ultra-fast structure similarity search functionality using an embedding/indexing (EI) algorithm (42). When the fingerprint method is chosen, the query is sent to *PubChem*, where the structure search is performed and the results are returned to the *compound workbench*. In contrast to this, *EI Search* is specific to the *ChemMine Tools* project and thus, runs locally on its servers. These two tools possess complementary strengths and weaknesses in identifying weak similarities among compounds (42).

Clustering toolbox

Clustering of compounds by structural or property similarity can be a powerful approach to correlating compound features with biological activity. Clustering tools are also widely utilized for diversity analyses to identify structural redundancies and other biases in compound libraries. *ChemMine Tools'* clustering workbench provides an online interface to three clustering algorithms which include hierarchical clustering, multidimensional scaling (MDS) and binning clustering (35). The following provides a short overview of these tools, while a more detailed outline of the underlying theory and clustering schemes is available in the online tutorial. When clustering by structural similarity, the required similarity measures are computed by first generating the atom pair descriptors (features) for each compound which are then used to calculate a similarity matrix based on the common and unique features observed among all compound pairs using the Tanimoto coefficient. The Tanimoto coefficient has a range from 0 to 1 with higher values indicating greater similarity than lower ones. For the subsequent clustering steps, the similarity matrix is converted into a distance matrix by subtracting the similarity values from 1. The hierarchical and MDS clustering methods provided by *ChemMine Tools* are based on the *R* programs *hclust* and *cmdscale*, respectively; the third method utilizes an internally developed C++ implementation. These three programs complement one another with respect to their data outputs and visualization options. Hierarchical clustering organizes compounds by similarity in a tree with branch lengths proportional to the item-to-item (compound-to-compound) similarities, while the MDS output encodes this information in a scatter plot. These two methods do not directly provide assignments of compounds to discrete similarity groups; assignments are generated downstream of the actual clustering process using various post-processing methods, such as tree cutting approaches. The binning clustering output provides these groupings directly for a user-definable similarity cutoff. For instance, if a Tanimoto coefficient of 0.6 is chosen then compounds will be joined into groups that share a similarity of this value or greater using a 'single linkage' rule for cluster joining. Final results are presented as interactive visualization pages to simplify the interpretation of the (often complex) clustering results. The hierarchical clustering result page uses the *Google Maps API* to generate zoom- and click-able trees aligned with molecular structure images. Moreover, heat maps of user uploaded data containing compound property, activity or other information can be viewed alongside the tree. A similar system is used to present the MDS results as click-able scatter plots with cursor-over viewing of compound structures. The binning clustering results are presented in a table view containing (among other information) the cluster identifiers and the corresponding compound depictions.

Property toolbox

Predictions of small molecule physicochemical properties are important for assessing their 'druglikeness' and 'leadlikeness' *in silico* (43,44). They are also useful for enriching compound collections with desirable properties. For instance, the famous 'Lipinski Rule of Five' (45) is often applied to enrich compound collections with druglike candidates. This rule filters for compounds with $\leq$5 hydrogen bond donors, $\leq$10 hydrogen acceptors, a molecular weight $\leq$500 daltons and an octanol-water partition coefficient log $P \leq 5$. Physicochemical property data are essential for predicting bioactive and other properties of small molecules using modern machine learning approaches. These data are fundamental to the development of QSAR models (25). *ChemMine Tools* provides an online interface to the property prediction module of the *JOELib* package (32). This service can calculate 38 physicochemical property values, including Lipinski descriptors for custom compound sets. The resulting property tables can be downloaded or further processed on *ChemMine Tools* by sending them to the Clustering toolbox. There, they can be used to cluster compounds by similar property profiles, as described above, or the data can be visualized as a heat map next to the hierarchical clustering trees.

CONCLUSION AND FUTURE DEVELOPMENT

ChemMine Tools is an online service for compound analysis in the chemical genomics field. The service is unique in that it integrates a large number of cheminformatic programs with clustering and visualization functionalities. Additional outstanding features of *ChemMine Tools* include: (i) its commitment to publicly developed open source software throughout its infrastructure; (ii) its strong dedication to the development of new cheminformatic tools and their free distribution in the community; and (iii) the integration of its many components into a unified online and downloadable software infrastructure which maximizes their utility for diverse tasks with different levels of complexity and customization needs. An intuitive web interface makes these tools accessible to scientists with limited computational background, while simultaneously providing a programmable interface for advanced users. To the best of our knowledge, there are currently no related online services available that provide a comparable suite of functionalities. Overlaps exist, however they are limited to isolated functionalities. For instance, *ChemDB* and VCCLab (13,43) can be used for property predictions and structure format interconversions of single compound queries; and *PubChem* supports structure-based clustering for compounds retrieved from its own database.

In the future, many additional utilities will be added to the *ChemMine Tools* service including the addition of MCS-based search functionality within the Similarity toolbox to support more complex graph-based search strategies against custom compound sets imported into the Compound workbench. Existing functionalities for analyzing bioactivity data will also be expanded by adding a Bioactivity toolbox that will contain regression, machine learning and QSAR modeling tools.

ACKNOWLEDGEMENTS

We thank the community projects—*Open Babel*, *JOELib*, *CACTVS* and *R*—for providing excellent software and data resources that are used by *ChemMine Tools*. We also thank Peter Ertl for providing the *JME Molecular Editor*. TG acknowledges support from the core facilities at the Institute for Integrative Genome Biology (IIGB) at UC Riverside. Additionally, we thank our systems administrator Aleksandr Levchuk for assistance in debugging these tools, and expertly maintaining the necessary computational resources.

FUNDING

National Science Foundation (grant numbers ABI-0957099, 2010-0520325 and IGERT-0504249). Funding for open access charge: National Science Foundation (grant number: ABI-0957099).

Conflict of interest statement. None declared.

REFERENCES

1. Strausberg,R.L. and Schreiber,S.L. (2003) From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science*, **300**, 294–295.
2. Haggarty,S.J. (2005) The principle of complementarity: chemical versus biological space. *Curr. Opin. Chem. Biol.*, **9**, 296–303.
3. Oprea,T.I., Tropsha,A., Faulon,J.L. and Rintoul,M.D. (2007) Systems chemical biology. *Nat. Chem. Biol.*, **3**, 447–450.
4. Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
5. Hattori,M., Okuno,Y.Y., Goto,S. and Kanehisa,M. (2003) Heuristics for chemical compound matching. *Genome Inform.*, **14**, 144–153.
6. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
7. Rahman,S.A., Bashton,M., Holliday,G.L., Schrader,R. and Thornton,J.M. (2009) Small molecule subgraph detector (SMSD) toolkitl. *J. Cheminform.*, **1**, 12.
8. Olah,M.M., Bologa,C.G. and Oprea,T.I. (2004) Strategies for compound selection. *Curr. Drug. Discov. Technol.*, **1**, 211–220.
9. Austin,C.P., Brady,L.S., Insel,T.R. and Collins,F.S. (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.
10. Seiler,K.P., George,G.A., Happ,M.P., Bodycombe,N.E., Carrinski,H.A., Norton,S., Brudz,S., Sullivan,J.P., Muhlich,J., Serrano,M. *et al*. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.
11. Ihlenfeldt,W.D., Voigt,J.H., Bienfait,B., Oellien,F. and Nicklaus,M.C. (2002) Enhanced CACTVS browser of the open NCI database. *J. Chem. Inf. Comput. Sci.*, **42**, 46–57.
12. Girke,T., Cheng,L.C. and Raikhel,N. (2005) ChemMine. A compound mining database for chemical genomics. *Plant Physiol.*, **138**, 573–577.
13. Chen,J.H., Linstead,E., Swamidass,S.J., Wang,D. and Baldi,P. (2007) ChemDB update–full-text search and virtual chemical space. *Bioinformatics*, **23**, 2348–2351.
14. Irwin,J.J. and Shoichet,B.K. (2005) ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model*, **45**, 177–182.
15. Li,Q., Cheng,T., Wang,Y. and Bryant,S.H. (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today*, **15**, 1052–1057.
16. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
17. Voigt,J.H., Bienfait,B., Wang,S. and Nicklaus,M.C. (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.*, **41**, 702–712.
18. Couzin,J. (2003) Molecular medicine. NIH dives into drug discovery. *Science*, **302**, 218–221.
19. Wang,R., Fang,X., Lu,Y. and Wang,S. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.
20. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcántara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
21. Block,P., Sotriffer,C.A., Dramburg,I. and Klebe,G. (2006) AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.*, **34**, D522–D526.
22. Kuhn,M., von Mering,C., Campillos,M., Jensen,L.J. and Bork,P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.
23. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
24. Goede,A., Dunkel,M., Mester,N., Frommel,C. and Preissner,R. (2005) SuperDrug: a conformational drug database. *Bioinformatics*, **21**, 1751–1753.
25. Spjuth,O., Willighagen,E.L., Guha,R., Eklund,M. and Wikberg,J.E. (2010) Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J. Cheminform.*, **2**, 5.
26. Zhu,Q., Lajiness,M.S., Ding,Y. and Wild,D.J. (2010) WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminform.*, **2**, 6.
27. Guha,R., Howard,M.T., Hutchison,G.R., Murray-Rust,P., Rzepa,H., Steinbeck,C., Wegner,J. and Willighagen,E.L. (2006) The blue obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model*, **46**, 991–998.
28. O'Boyle,N.M., Morley,C. and Hutchison,G.R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 5.
29. Steinbeck,C., Hoppe,C., Kuhn,S., Floris,M., Guha,R. and Willighagen,E.L. (2006) Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.*, **12**, 2111–2120.
30. Guha,R. (2007) Chemical Informatics functionality in R. *J. Stat. Softw.*, **18**, 1–16.
31. Sykora,V.J. and Leahy,D.E. (2008) Chemical descriptors library (CDL): a generic, open source software library for chemical informatics. *J. Chem. Inf. Model*, **48**, 1931–1942.
32. Wegner,J.K., Fröhlich,H. and Zell,A. (2004) Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA). *J. Chem. Inf. Comput. Sci.*, **44**, 931–939.
33. Walker,T., Grulke,C.M., Pozefsky,D. and Tropsha,A. (2010) Chembench: a cheminformatics workbench. *Bioinformatics*, **26**, 3000–3001.
34. Berthold,M.R., Cebron,N., Dill,F., Gabriel,T.R., Kotter,T., Meinl,T., Ohl,P., Sieb,C., Thiel,K. and Wiswedel,B. (2007) *KNIME: The Konstanz Information Miner*. Springer, New York.
35. Cao,Y., Charisi,A., Cheng,L.C., Jiang,T. and Girke,T. (2008) ChemmineR: a compound mining framework for R. *Bioinformatics*, **24**, 1733–1734.
36. Ertl,P. (2010) Molecular structure input on the web. *J. Cheminform.*, **2**, 1.
37. 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.
38. 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.
39. Cao,Y., Jiang,T. and Girke,T. (2008) A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, **24**, 366–374.

40. Raymond,J.W. and Willett,P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, **16**, 521–533.
41. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
42. Cao,Y., Jiang,T. and Girke,T. (2010) Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing. *Bioinformatics*, **26**, 953–959.
43. Tetko,I.V., Gasteiger,J., Todeschini,R., Mauri,A., Livingstone,D., Ertl,P., Palyulin,V.A., Radchenko,E.V., Zefirov,N.S., Makarenko,A.S. *et al.* (2005) Virtual computational chemistry laboratory–design and description. *J. Comput. Aided Mol. Des.*, **19**, 453–463.
44. Monge,A., Arrault,A., Marot,C. and Morin-Allory,L. (2006) Managing, profiling and analyzing a library of 2.6 million compounds gathered from 32 chemical providers. *Mol. Divers*, **10**, 389–403.
45. Lipinski,C.A., Lombardo,F., Dominy,B.W. and J,F.P. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliver. Rev.*, **23**, 3–25.