Published in final edited form as:
Drug Discov Today. 2014 November ; 19(11): 1751–1756. doi:10.1016/j.drudis.2014.08.008.

PubChem applications in drug discovery: a bibliometric analysis

Tiejun Cheng, **Yongmei Pan**, **Ming Hao**, **Yanli Wang***, and **Stephen H. Bryant***
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

Abstract

A bibliometric analysis of PubChem applications is presented by reviewing 1132 research articles. The massive volume of chemical structure and bioactivity data in PubChem and its online services has been used globally in various fields including chemical biology, medicinal chemistry and informatics research. PubChem supports drug discovery in many aspects such as lead identification and optimization, compound–target profiling, polypharmacology studies and unknown chemical identity elucidation. PubChem has also become a valuable resource for developing secondary databases, informatics tools and web services. The growing PubChem resource with its public availability offers support and great opportunities for the interrogation of pharmacological mechanisms and the genetic basis of diseases, which are vital for drug innovation and repurposing.

Keywords

PubChem applications; drug discovery; public database; bibliometric analysis

Introduction

PubChem (http://pubchem.ncbi.nlm.nih.gov), hosted by the National Center for Biotechnology Information (NCBI), National Institutes of Health, is a public repository for chemical structures and their bioactivities [1–5]. It has three interconnected databases: Substance, BioAssay (containing depositions of chemical samples, biological results for small molecules and RNAi reagents) and Compound (containing derived unique chemical structures). PubChem grows rapidly and is now arguably the largest chemical biology database available to the public (Figure 1), which offers open access to over 50 000 users daily via the NCBI Entrez system, as well as web-based and programmatic tools. Moreover, PubChem is closely integrated with other literature and biomedical databases such as PubMed, Protein, Gene, Structure and Taxonomy.

*Corresponding authors: Wang, Y. (ywang@ncbi.nlm.nih.gov); Bryant, S.H. (bryant@ncbi.nlm.nih.gov).

Teaser: Bibliometric analysis of PubChem applications revealed how the community took advantage of the PubChem resources and suggested potential opportunities and challenges for drug discovery and repurposing.

In this work, we aim to provide a comprehensive review on community utilization of the PubChem resource to advance drug discovery and other research. Our analysis, based on the 1132 recent publications, shows that PubChem was widely employed for referencing and obtaining small molecule structures and annotations, and supporting lead identification via HTS and virtual screening of PubChem compounds. The large collection of bioactivity data and molecular target information in PubChem BioAssay has facilitated research areas such as SAR studies, compound–target profiling, drug activity evaluation and polypharmacology studies. The free access of chemical and biological data in PubChem also stimulated the development of secondary databases and informatics tools that interact with PubChem.

Bibliometric analysis of PubChem applications

Our analysis was based on the biomedical literature retrieved from the NCBI PubMed Central (PMC) and PubMed databases. As of 31 December 2013, a keyword search of 'PubChem' in the above two databases returned 1703 unique hits, including 1456 PMC full articles and additionally 247 PubMed abstracts, for which full articles were obtained subsequently. We manually inspected each article to identify the applied PubChem resource and its utilization in each study. We excluded 184 articles published by the research groups who had deposited data into PubChem to avoid the analysis being biased toward such frequent and 'power' users of PubChem. We further limited our investigation to research articles only, which excluded 171 reviews, commentaries, perspectives, meeting abstracts and so on. Another 216 articles that simply cited PubChem as one representative public chemical biology database or had little relevance to PubChem utilization were excluded as well. As a result, a total of 571 articles were excluded (Table S1 in the supplementary material online), and the remaining 1132 research articles were used for subsequent analysis (Table S2 in the supplementary material online). It is evident that PubChem received increasing citations as a function of time (Figure 2). A closer view shows that such citations were from over 270 peer-reviewed journals and by worldwide researchers (Figure 3), indicating the impact of PubChem upon the global community.

To facilitate further investigation and illustration, we classified the applications of PubChem into the following four categories: (i) data retrieval, exchange and service utilization; (ii) secondary resources and tools involving PubChem; (iii) applications in informatics research; and (iv) applications in wet-lab experiments. Note that an article could be assigned to multiple categories based on its content (154 articles in total; Table S2 in the supplementary material online). We will illustrate each category in detail in the following sections with a specific highlight on PubChem applications in drug discovery. Owing to page limitations, only a small number of articles are cited in this manuscript, and citations for equivalent or similar work can be found in Table S2 (supplementary material online).

Data retrieval, exchange and service utilization

The vast amount of molecular and bioactivity data in PubChem was extensively retrieved and analyzed by chemists, pharmacologists and biologists using the PubChem search engine, various web services and online and programmatic tools.

Data retrieval

Users approached PubChem for a wide variety of data, and the most frequently accessed were as follows: (i) chemical structures downloaded in SMILES, InChI or SDF format and structural images linked via web interface; (ii) basic small molecule information, such as annotations, synonyms, MeSH, formula, physicochemical properties (e.g. logP, molecular weight and numbers of hydrogen-bond donors and acceptors) and pharmacological information; (iii) PubChem fingerprints (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt) that were often used for similarity search, chemical space analysis, structure clustering and SAR studies; (iv) bioassay datasets including HTS biochemical and cell-based assays for SAR studies and data mining. Examples of the applications for such retrieved data are presented throughout the manuscript.

Data exchange

The PubChem identifiers (i.e. CID for Compound record, SID for Substance record and AID for BioAssay record) were commonly adopted as a means for data and information exchange among various studies and were indexed or recognized by various databases or tools. Many articles annotated compounds of interest with SIDs or CIDs, which efficiently facilitated the transferring and comparison of structural information. Moreover, selected Elsevier journals recommend authors to supply an explicit summary of studied compounds with corresponding CID numbers. As a result, online articles were enriched with relevant information such as molecular weight and formula, as well as the chemical structure and associated bioactivity data extracted from PubChem. Likewise, AIDs offered an easy solution for sharing biological experiment details and test results, and worked as a convenient reference to bioassay datasets downloaded from PubChem.

Service utilization

PubChem provides many online tools and web services allowing data retrieval and analysis. Although system log files can offer a thorough view of PubChem usage, examples described in research articles can tell exactly how the services were utilized. Our review shows that: (i) the structure search tool was heavily used for identity, similarity or substructure search for compounds of interest (e.g. looking for a known drug molecule or seeking analogs or derivatives for a given lead compound); (ii) the structure clustering tool was applied for grouping, organizing and analyzing a set of compounds based on 2D or 3D structural features; (iii) the BioActivity SAR service and other assay-related tools were utilized to perform bioactivity data retrieval and analysis, or to explore compound-target-assay associations; and (iv) the power user gateway with flexibility and batch capability was applied as a programmatic tool to access PubChem data.

Secondary resources and tools involving PubChem

The rich contents in PubChem have fostered the development of secondary databases and tools by extracting data from PubChem, accepting PubChem identifiers or cross-reference linking to PubChem. A list of over 180 resources and tools involving PubChem is provided (Table S3 in the supplementary material online). Notably, the majority of them are in the

public domain, serving as additional and valuable resources for chemical biology research and drug discovery.

Databases derived from PubChem

There were stirred interests in establishing secondary databases or datasets by compiling data from PubChem [6–8]. For example, COMMODE is a large-scale collection of molecular descriptors for the entire PubChem compound database [6]. Maximum unbiased validation (MUV) is a benchmark dataset generated from PubChem BioAssay that is designed for virtual screening [7]. SuperCYP is a database of Cytochrome P450 (CYP) enzymes with associated CYP–drug interactions collected from PubChem and other sources [8]. The majority of these databases and datasets have links back to PubChem allowing their users to retrieve additional annotations from PubChem, meanwhile providing value-added curation to PubChem data.

Tools and web services compatible with PubChem

Users showed strong and diverse needs in developing tools or web applications for customized tasks such as PubChem-based data search and retrieval, for example PubChemSR [9] and PubChemDB [10], although similar functionalities are already offered at PubChem. Many tools have built-in features for working with PubChem data directly and interactively. For example, Avogadro [11] is an advanced chemical editor, visualization and analysis platform that allows retrieving structures from PubChem by compound synonyms. The ChemMine tools [12] support the batch import and similarity search of structures with PubChem CIDs. Such tools eliminated the need of manual work and lowered the barrier for data exchange across resources and studies. In addition, PubChem's open access to the large-scale HTS data with great diversity enabled the development of semantic tools, such as BioAssay ontology (BAO) for the description and characterization of bioassays and HTS results [13].

Applications in informatics research

Informatics research has greatly benefited from the publicly available data in PubChem. Studies using PubChem for data mining or analysis as well as for the application or validation of informatics tools emerged rapidly following the launch of PubChem (Figure 4).

Chemical space analysis

PubChem represents an extremely large chemical space of over 50 million unique structures to the public. Based on the molecular and structural properties, these compounds were often analyzed for similarity, diversity, novelty, scaffold topology, as well as for clustering, ontology, classification, tautomerism and synthetic accessibility [14–17]. For example, van Deursen *et al.* analyzed and visualized the chemical space of the drug-like and lead-like compounds from PubChem by using 42 structural descriptors [14]. Singh *et al.* performed a chemical space analysis across PubChem and other databases in terms of physicochemical properties, structural properties and scaffolds to evaluate the consistency, complementarity, uniqueness and overlaps among different databases [15]. Bioactivity information was sometimes combined with molecular descriptors for chemical space calculation. Krein and

Sukumar explored the chemical space of compounds from a PubChem bioassay dataset by considering structural properties and the derived SAR [16]. Lounkine *et al.* performed an activity-aware fingerprint-based clustering by using 462 dose–response assays from the PubChem BioAssay [17].

Compound–target network and drug polypharmacology

PubChem BioAssay was employed to establish the connections among compounds, targets, related gene expression profiles, signaling pathways and other biological systems to gain insights into drug polypharmacology study that could be helpful to future drug design [18–20]. As an example, Covell performed data mining of the NCI-60 anticancer screen datasets in PubChem for exploring the correlations among the gene expression, chemoactivity profile and biological pathways [18]. On a larger scale, Chen *et al.* investigated the drug polypharmacology behavior by performing a cross-assay analysis of PubChem BioAssay followed by mapping the obtained bioassay network to other biological networks (e.g. drug–target network) [19]. In addition, Hu *et al.* systematically analyzed 1085 confirmatory bioassays from PubChem and generated a drug bioactivity profile across a wide range of biological targets, which might serve as a reference for drug selectivity and promiscuity [20].

SAR model generation and validation

The chemical structures, bioassay datasets and molecular properties including structural fingerprints calculated with PubChem were applied for the generation and validation of *in silico* SAR models, which can be used for structural optimization and virtual-screening-based lead identification by predicting drug activities, toxicities, adverse effects and other properties. The generated models included binary machine learning based models [e.g. support vector machines (SVM), Bayesian and recursive partitioning models], as well as quantitative SAR models such as 2D-descriptor-based linear regression and 3D-based comparative molecular field analysis (CoMFA) models [21–23]. For example, Periwal *et al.* constructed machine learning models for screening antitubercular activity by using a HTS confirmatory bioassay dataset from PubChem [21]. Pouliot *et al.* generated logistic regression models for prediction of adverse drug reactions (ADRs) by correlating ADRs with PubChem bioassay screening results, which were retrospectively verified by established drugs [22]. Nagamine *et al.* validated their SVM-based classification strategy for enhancing protein–ligand interaction with over 19 million PubChem compounds followed by *in vitro* validation [23]. SAR models were also employed for lead optimization to identify more active compounds or chemical probes [24,25]. For example, Wendt *et al.* designed and verified selective inhibitors of hypoxia-inducible factor 1 (HIF-1) guided by a topomer CoMFA model generated from a PubChem HTS dataset [24]. Chou *et al.* developed two chemical probes for p97 ATPase inhibition based on the SAR study of the self-developed quinazoline analogs previously deposited in PubChem BioAssay [25].

Informatics method development and validation

The abundant structure and bioassay data in PubChem have motivated the development and validation of various informatics methods and algorithms [26–28]. For instance, Feldman *et*

al. applied PubChem structures to a chemical ontology algorithm that was developed based on chemical functional groups [26]. Matlock *et al.* demonstrated their scaffold discovery framework by employing the PubChem HTS datasets to maximize the active scaffolds that increased the number of active molecules confirmed by experiments [27]. Butkiewicz and co-workers assembled a group of PubChem HTS datasets for benchmarking ligand-based virtual screening involving the major families of drug target proteins [28].

Applications in wet-lab experiments

PubChem was widely utilized for lead identification and optimization, an essential step for the discovery and design of drug candidates. With the bioactivity information for thousands of protein and gene targets, especially the large volume of HTS datasets that are currently lacking in the public sector, PubChem greatly facilitated the compound activity profiling and polypharmacology studies. Furthermore, PubChem played an important part in elucidating the identity of unknown biomarkers, metabolites and other compounds.

Lead identification and optimization

PubChem chemical structures were employed to construct screening libraries for identifying lead compounds as potential drug candidates. Libraries were typically compiled by extracting diverse compounds from PubChem according to the descriptor-based chemical space analysis [29], by performing similarity or substructure search against PubChem compounds [30], by selecting compounds tested in specific bioassay datasets or by downloading structures manually chosen according to user criteria [31]. Such libraries were virtually screened via docking, SAR model or similarity-based search for lead compounds, with predictions further validated by experimental assays [31–36]. For example, Ren *et al.* performed a hierarchical multistage virtual screening of the entire PubChem database based on SVM model, pharmacophore and molecular docking and discovered novel Pim-1 kinase inhibitors that were confirmed by *in vitro* assays [34]. Srinivasan *et al.* predicted novel human apurinic/apyrimidinic endonuclease-1 inhibitors by docking-based virtual screening of a structurally diverse sub-library of PubChem and validated them with *in vitro* assays [35]. Lin and colleagues conducted a similarity search against PubChem based on previously identified hits and subsequently obtained a promising drug candidate that is currently in clinical trials for the treatment of various cancers [36]. In addition to virtual screening, PubChem compounds with certain functional annotations were also selected for experimental screening to identify active outcomes targeting particular biological systems [37,38]. For instance, Ho *et al.* selected active compounds with the functions of cell growth inhibition, antiproliferation and apoptosis induction from the PubChem BioAssay, and tested their β-catenin signaling activity with *in vitro* assays [38].

PubChem structures were also downloaded for characterizing ligand–protein interaction through docking, molecular dynamics (MD) simulation, binding energy calculation or other molecular modeling techniques [39–41]. Such studies were often performed after the identification of lead compounds by experimental chemists and pharmacologists aiming to investigate specific compound–target interactions that could benefit further drug optimization or *de novo* drug design. The public availability of the large compound collection in PubChem has greatly enabled such investigations.

PubChem as a reference resource

The search tools of PubChem allow efficient large-scale retrieval of bioactivity data and assay descriptions that provide an ideal complement to the conventional literature resources in several ways:

- Bioactivity profile evaluation of studied compounds for their selectivity, diversity, novelty and cytotoxicity using the PubChem BioAssay data across assays, targets, cell lines, cellular functions, signaling pathways and so on [42–46]. For instance, Vang *et al.* searched the PubChem bioassay datasets among various drug targets and eliminated nonselective inhibitors of lymphoid tyrosine phosphatase [46].
- Compound inspection for active or inactive outcomes against a particular target, cell line or other biological systems to evaluate the structure diversity or novelty of compounds as compared with prior art in PubChem [38,47]. For example, Rickard *et al*. confirmed the novelty of the identified nucleotide-binding oligomerization domain 2 (NOD2) inhibitors by searching the screening results in the PubChem BioAssay database and other literature [47].
- Utilization or adaption of experimental protocols deposited in PubChem BioAssay by biologists for devising their own experiments [48].

Unknown identity annotation

PubChem with molecular structure and property data helped analytic chemists tremendously with the identification of unknown biomarkers, metabolites and other molecules by searching PubChem using exact mass, fragment, molecular formula and other information derived from analytical techniques [e.g. mass spectrometry (MS) combined with separation equipment such as liquid or gas chromatography (LC or GC)] [49–53]. Cheng and Guengerich identified orphan CYP substrates with HPLC-MS-derived formulae that matched known structures in PubChem and other databases, with the identities confirmed by comparing their HPLC elution time and MS fragmentation pattern with standard compounds [49]. Derewacz *et al.* elucidated metabolite identities by querying PubChem and other databases with MS-obtained accurate mass combined with the analysis of fragmental spectra, LC and retention time [50].

Concluding remarks

PubChem grows rapidly with chemical structure, bioactivity and molecular target data deposited on a daily basis. This open and valuable resource increasingly attracts worldwide interest from academic and industrial sectors. As a result, citations on the PubChem resource also grow quickly, covering multidisciplinary research fields such as informatics studies, biomedical and pharmaceutical research, and database and web-service development. PubChem data and services were extensively utilized, ranging from straightforward information search and retrieval to in-depth data mining for drug discovery studies. It is observed that computational applications dominated over experimental research in the first several years, but the latter caught up gradually (Figure 4). Our review on these applications could help to provide insight into the impact of PubChem resource on the medicinal,

pharmacological and chemical biological research community, meanwhile highlighting areas that have been largely overlooked.

Many challenges remain for PubChem users, as well as for PubChem to improve its data and services. This review indicates sections among the PubChem resources that have not been fully explored, and highlights fields that are worthwhile for further research investigation or future improvement of PubChem: (i) the chemical probes available in PubChem, which were generated by the Molecular Libraries Initiative as small molecule tools, are to be exploited for unraveling complex biological and disease related systems (http://www.ncbi.nlm.nih.gov/books/NBK47352/); (ii) the RNAi screening data in PubChem remained largely unnoticed, which together with small molecule bioassays can provide useful insights to the biological systems under investigation, as well as to understand the genetic basis of diseases [54,55]; (iii) integration of PubChem assay targets including proteins, genes and pathways with genomic data and disease information represents other interesting but less explored research areas such as polypharmacology, drug repurposing and personalized medicine [56]; (iv) text mining on bioassay data with rich descriptions on disease and targets, and recently added patent information toward data integration for exploring drug-target-disease relationships is currently scarce; (v) the HTS data in nature are often highly imbalanced and noisy, making it challenging for data mining and modeling. Despite a number of previous attempts [57–59], it still demands efforts from researchers and PubChem for developing methods to handle these issues.

Serving as a public repository, PubChem has been continuously making progress in the past ten years with multiple milestones achieved for collecting data, developing BioAssay data model, building information platforms and integrating with other biomedical resources. Further enhancements in information integration, efficient search tools, data quality control, annotations and classifications on chemical structure and bioactivity data will certainly be appreciated by the research community. By contrast, the entire community including researchers, funding agencies and open access journals can all come together to take important roles in the continuous development of PubChem (e.g. by sharing research data and depositing them into PubChem).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported (in part) by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Wang Y, et al. PubChem BioAssay: 2014 update. Nucleic Acids Res. 2014; 42:D1075–1082. [PubMed: 24198245]
2. Wang Y, et al. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009; 37:W623–633. [PubMed: 19498078]

3. Wang Y, et al. An overview of the PubChem BioAssay resource. Nucleic Acids Res. 2009; 38:D255–266. [PubMed: 19933261]
4. Bolton, EE., et al. PubChem: integrated platform of small molecules and biological activities. In: Ralph, AW.; David, CS., editors. Annual Reports in Computational Chemistry. Vol. 4. Elsevier; 2008. p. 217-241.
5. Li Q, et al. PubChem as a public resource for drug discovery. Drug Discov Today. 2010; 15:1052–1057. [PubMed: 20970519]
6. Dander A, et al. [COMMODE] a large-scale database of molecular descriptors using compounds from PubChem. Source Code Biol Med. 2013; 8:22. [PubMed: 24225386]
7. Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. J Chem Inf Model. 2009; 49:169–184. [PubMed: 19434821]
8. Preissner S, et al. SuperCYP: a comprehensive database on cytochrome P450 enzymes including a tool for analysis of cyp-drug interactions. Nucleic Acids Res. 2010; 38:D237–243. [PubMed: 19934256]
9. Hur J, Wild DJ. PubChemSR: a search and retrieval tool for PubChem. Chem Cent J. 2008; 2:11. [PubMed: 18482452]
10. Southern MR, Griffin PR. A Java API for working with PubChem datasets. Bioinformatics. 2011; 27:741–742. [PubMed: 21216779]
11. Hanwell MD, et al. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform. 2012; 4:17. [PubMed: 22889332]
12. Backman TW, et al. ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res. 2011; 39:W486–491. [PubMed: 21576229]
13. Visser U, et al. BioAssay ontology (BAO): a semantic description of bioassays and high-throughput screening results. BMC Bioinformatics. 2011; 12:257. [PubMed: 21702939]
14. van Deursen R, et al. Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem. J Comput Aided Mol Des. 2011; 25:649–662. [PubMed: 21618008]
15. Singh N, et al. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. J Chem Inf Model. 2009; 49:1010–1024. [PubMed: 19301827]
16. Krein MP, Sukumar N. Exploration of the topology of chemical spaces with network measures. J Phys Chem A. 2011; 115:12905–12918. [PubMed: 21882847]
17. Lounkine E, et al. Activity-aware clustering of high throughput screening data and elucidation of orthogonal structure–activity relationships. J Chem Inf Model. 2011; 51:3158–3168. [PubMed: 22098146]
18. Covell DG. Integrating constitutive gene expression and chemoactivity: mining the NCI60 anticancer screen. PLoS One. 2012; 7:e44631. [PubMed: 23056181]
19. Chen B, et al. PubChem as a source of polypharmacology. J Chem Inf Model. 2009; 49:2044–2055. [PubMed: 19708682]
20. Hu Y, Bajorath J. What is the likelihood of an active compound to be promiscuous? Systematic assessment of compound promiscuity on the basis of PubChem confirmatory bioassay data. AAPS J. 2013; 15:808–815. [PubMed: 23605807]
21. Periwal V, et al. Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. BMC Pharmacol. 2012; 12:1. [PubMed: 22463123]
22. Pouliot Y, et al. Predicting adverse drug reactions using publicly available PubChem BioAssay data. Clin Pharmacol Ther. 2011; 90:90–99. [PubMed: 21613989]
23. Nagamine N, et al. Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. PLoS Comput Biol. 2009; 5:e1000397. [PubMed: 19503826]
24. Wendt B, et al. Toluidinesulfonamide hypoxia-induced factor 1 inhibitors: alleviating drug-drug interactions through use of PubChem data and comparative molecular field analysis guided synthesis. J Med Chem. 2011; 54:3982–3986. [PubMed: 21574568]

25. Chou TF, et al. Structure-activity relationship study reveals ML240 and ML241 as potent and selective inhibitors of p97 ATPase. ChemMedChem. 2013; 8:297–312. [PubMed: 23316025]
26. Feldman HJ, et al. CO: a chemical ontology for identification of functional groups and semantic comparison of small molecules. FEBS Lett. 2005; 579:4685–4691. [PubMed: 16098521]
27. Matlock MK, et al. Scaffold network generator: a tool for mining molecular structures. Bioinformatics. 2013; 29:2655–2656. [PubMed: 23918250]
28. Butkiewicz M, et al. Benchmarking ligand-based virtual high-throughput screening with the PubChem database. Molecules. 2013; 18:735–756. [PubMed: 23299552]
29. Xie XQ, Chen JZ. Data mining a small molecule drug screening representative subset from nih PubChem. J Chem Inf Model. 2008; 48:465–475. [PubMed: 18302356]
30. Chen H, et al. Prediction of molecular targets of cancer preventing flavonoid compounds using computational methods. PLoS One. 2012; 7:e38261. [PubMed: 22693608]
31. Fernández JR, et al. Identification of small molecule compounds with higher binding affinity to guanine deaminase (cypin) than guanine. Bioorg Med Chem. 2010; 18:6748–6755. [PubMed: 20716488]
32. Mast N, et al. *In silico* and intuitive predictions of CYP46A1 inhibition by marketed drugs with subsequent enzyme crystallization in complex with fluvoxamine. Mol Pharmacol. 2012; 82:824–834. [PubMed: 22859721]
33. Huang JW, et al. Fragment-based design of small molecule X-linked inhibitor of apoptosis protein inhibitors. J Med Chem. 2008; 51:7111–7118. [PubMed: 18956862]
34. Ren JX, et al. Discovery of novel Pim-1 kinase inhibitors by a hierarchical multistage virtual screening approach based on SVM model, pharmacophore, and molecular docking. J Chem Inf Model. 2011; 51:1364–1375. [PubMed: 21618971]
35. Srinivasan A, et al. Identification and characterization of human apurinic/apyrimidinic endonuclease-1 inhibitors. Biochemistry. 2012; 51:6246–6259. [PubMed: 22788932]
36. Lin FY, et al. Dual dehydrosqualene/squalene synthase inhibitors: Leads for innate immune system-based therapeutics. ChemMedChem. 2012; 7:561–564. [PubMed: 22290830]
37. O'Leary DA, et al. Identification of small molecule and genetic modulators of AON-induced dystrophin exon skipping by high-throughput screening. PLoS One. 2009; 4:e8348. [PubMed: 20020055]
38. Ho JY, et al. Ovatodiolide targets beta-catenin signaling in suppressing tumorigenesis and overcoming drug resistance in renal cell carcinoma. Evid Based Complement Alternat Med. 2013; 2013:161628. [PubMed: 23781255]
39. Tiwari G, Mohanty D. An *in silico* analysis of the binding modes and binding affinities of small molecule modulators of PDZ-peptide interactions. PLoS One. 2013; 8:e71340. [PubMed: 23951139]
40. Alam A, et al. Novel anti-inflammatory activity of epoxyazadiradione against macrophage migration inhibitory factor: Inhibition of tautomerase and proinflammatory activities of macrophage migration inhibitory factor. J Biol Chem. 2012; 287:24844–24861. [PubMed: 22645149]
41. Saraswati S, et al. Tylophorine, a phenanthraindolizidine alkaloid isolated from tylophora indica exerts antiangiogenic and antitumor activity by targeting vascular endothelial growth factor receptor 2-mediated angiogenesis. Mol Cancer. 2013; 12:82. [PubMed: 23895055]
42. Tholander F, Sjoberg BM. Discovery of antimicrobial ribonucleotide reductase inhibitors by screening in microwell format. Proc Natl Acad Sci U S A. 2012; 109:9798–9803. [PubMed: 22665797]
43. Mattmann ME, et al. Identification of (R)-N-(4-(4-methoxyphenyl)thiazol-2-yl)-1-tosylpiperidine-2-carboxamide, ML277, as a novel, potent and selective k(v)7.1 (KCNQ1) potassium channel activator. Bioorg Med Chem Lett. 2012; 22:5936–5941. [PubMed: 22910039]
44. Firestine SM, et al. Identification of inhibitors of N5-carboxyaminoimidazole ribonucleotide synthetase by high-throughput screening. Bioorg Med Chem. 2009; 17:3317–3323. [PubMed: 19362848]
45. Crowe A, et al. Identification of aminothienopyridazine inhibitors of tau assembly by quantitative high-throughput screening. Biochemistry. 2009; 48:7732–7745. [PubMed: 19580328]

46. Vang T, et al. LYP inhibits T-cell activation when dissociated from CSK. Nat Chem Biol. 2012; 8:437–446. [PubMed: 22426112]
47. Rickard DJ, et al. Identification of benzimidazole diamides as selective inhibitors of the nucleotide-binding oligomerization domain 2 (NOD2) signaling pathway. PLoS One. 2013; 8:e69619. [PubMed: 23936340]
48. Kumar A, et al. Chemical correction of pre-mrna splicing defects associated with sequestration of muscleblind-like 1 protein by expanded r(CAG)-containing transcripts. ACS Chem Biol. 2012; 7:496–505. [PubMed: 22252896]
49. Cheng Q, Guengerich FP. Identification of endogenous substrates of orphan cytochrome P450 enzymes through the use of untargeted metabolomics approaches. Methods Mol Biol. 2013; 987:71–77. [PubMed: 23475668]
50. Derewacz DK, et al. Antimicrobial drug resistance affects broad changes in metabolomic phenotype in addition to secondary metabolism. Proc Natl Acad Sci U S A. 2013; 110:2336–2341. [PubMed: 23341601]
51. Hall LM, et al. Development of $Ecom_{50}$ and retention index models for nontargeted metabolomics: Identification of 1,3-dicyclohexylurea in human serum by HPLC/mass spectrometry. J Chem Inf Model. 2012; 52:1222–1237. [PubMed: 22489687]
52. Wang M, et al. Calcium-deficiency assessment and biomarker identification by an integrated urinary metabonomics analysis. BMC Med. 2013; 11:86. [PubMed: 23537001]
53. Hill DW, et al. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. Anal Chem. 2008; 80:5574–5582. [PubMed: 18547062]
54. Hasson SA, et al. High-content genome-wide RNAi screens identify regulators of parkin upstream of mitophagy. Nature. 2013; 504:291–295. [PubMed: 24270810]
55. Yin Z, et al. A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. Nat Cell Biol. 2013; 15:860–871. [PubMed: 23748611]
56. Dakshanamurthy S, et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem. 2012; 55:6832–6848. [PubMed: 22780961]
57. Hao M, et al. An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. Anal Chim Acta. 2014; 806:117–127. [PubMed: 24331047]
58. Calhoun BT, et al. Automatically detecting workflows in PubChem. J Biomol Screen. 2012; 17:1071–1079. [PubMed: 22693105]
59. Schorpp K, et al. Identification of small-molecule frequent hitters from AlphaScreen high-throughput screens. J Biomol Screen. 2013; 19:715–726. [PubMed: 24371213]

Highlights

- We investigated PubChem applications by a bibliometric analysis of research articles
- PubChem resource was accessed globally for various research areas
- PubChem was widely utilized in informatics and wet-lab studies in drug discovery
- PubChem is a valuable resource for secondary database and tool development
- Our analysis highlighted fields in PubChem that remain to be further explored

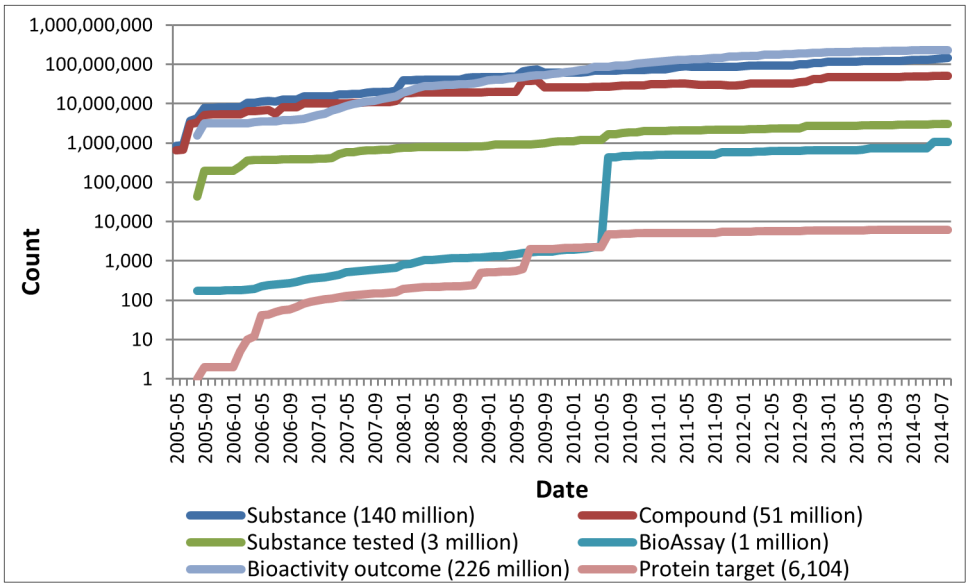

Figure 1.

Growth of PubChem data content. The numbers in parenthesis are the statistics by 7th July, 2014.

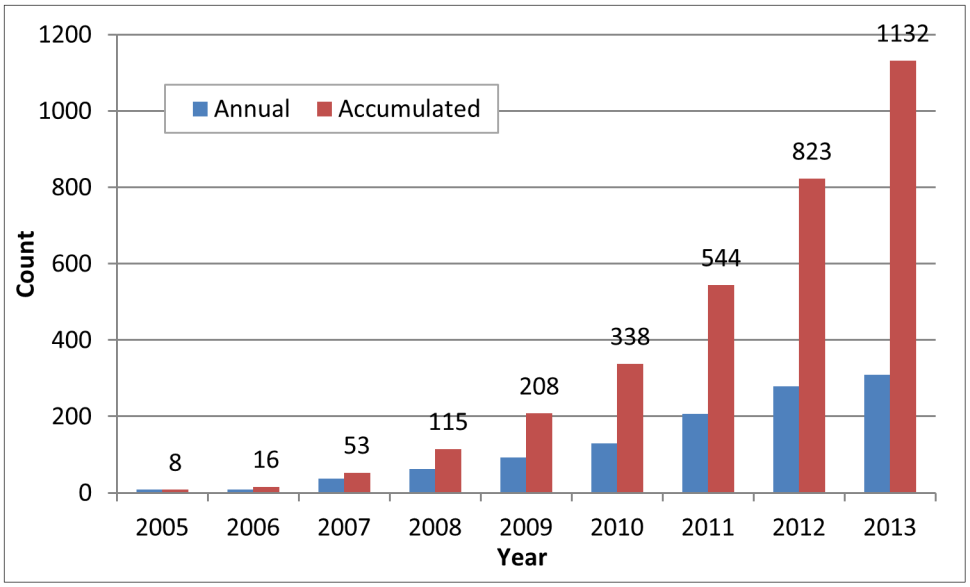

Figure 2.
Distribution of research articles citing PubChem by year.

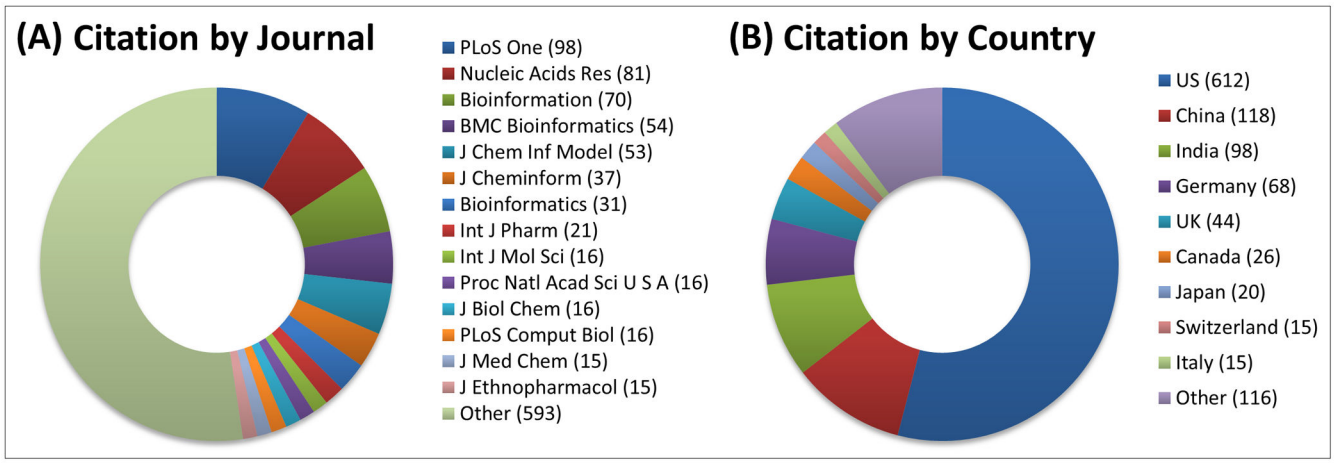

Figure 3.

Distribution of research articles citing PubChem in terms of journal **(a)** and country **(b)**. Respective article counts are shown in parentheses.

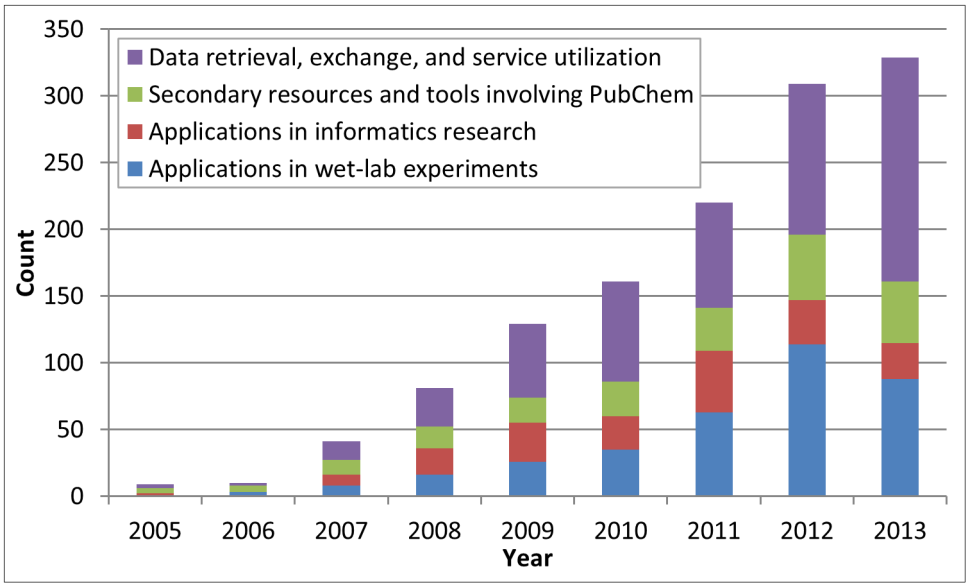

Figure 4.

Distribution of research articles citing PubChem by year and application category. A single article can be assigned to multiple categories for this plot.