PubChem as a public resource for drug discovery

Qingliang Li, Tiejun Cheng, Yanli Wang and Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

PubChem is a public repository of small molecules and their biological properties. Currently, it contains more than 25 million unique chemical structures and 90 million bioactivity outcomes associated with several thousand macromolecular targets. To address the potential utility of this public resource for drug discovery, we systematically summarized the protein targets in PubChem by function, 3D structure and biological pathway. Moreover, we analyzed the potency, selectivity and promiscuity of the bioactive compounds identified for these biological targets, including the chemical probes generated by the NIH Molecular Libraries Program. As a public resource, PubChem lowers the barrier for researchers to advance the development of chemical tools for modulating biological processes and drug candidates for disease treatments.

PubChem [1,2] (http://pubchem.ncbi.nlm.nih.gov) is a public repository for chemical structures and their biological properties. The bioactivity results in PubChem are contributed by more than a hundred organizations, with the majority of data coming from the screening center network under the NIH Molecular Libraries Program (MLP) [3]. This program aims to expand the use of small molecules as chemical probes, which offer dynamic, reversible and tunable perturbations for biological systems [4], to study the functions of genes and proteins in physiology and pathology. Unlike the pharmaceutical industry and biotechnology companies, which primarily focus on the 'druggable genome' [5,6] to screen the 'drug-like' small molecules against limited types of targets (such as kinases, G-protein-coupled receptors, enzymes, ion channels and nuclear hormone receptors), an extensive collection of biological targets and chemical compounds are being investigated by the MLP to address a wide scope of biological issues, from identifying inhibitors of a specific enzyme to looking for small molecules that affect protein–protein interactions or modulate splicing events [3]. With the rapid growth in data capacity, PubChem is becoming a valuable resource for drug development and has attracted considerable interest from researchers in both academia and industry.

PubChem consists of three interconnected databases: Substance, BioAssay and Compound. The Substance database contains the descriptions of molecules (primarily small molecules) provided by depositors; the BioAssay database contains the screening results of substances by assay providers; and the Compound database contains unique chemical structures derived by structural standardization of the records in the Substance database. Currently, more than 25 million unique chemical structures, which were derived from a collection of 70 million substances, are in the Compound database. As of April 2010, the BioAssay database comprised more than 2700 bioassays associated with more than one million compounds tested against several thousand molecular targets. In addition, several bioassays from RNAi screening experiments have been deposited in the BioAssay database.

A review of this public resource will allow the community to better understand the information content and utilize the data in PubChem, which might ultimately help to advance the development of new chemical tools and drug candidates by enabling researchers to study structure–activity relationships, investigate the interaction mechanisms between small molecules and their targets [7], and gain insights into the chemical and biological space in their research area. Here, we provide a comprehensive summary

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

1359-6446/06/$ - see front matter. Published by Elsevier B.V. doi:10.1016/j.drudis.2010.10.003

of the protein targets in PubChem with respect to their functional classification, availability of 3D structure and biological pathway. The potency, selectivity and promiscuity of the bioactive compounds (including the chemical probes developed by the MLP), which are associated with those protein targets, are also investigated.

Bioassay targets in PubChem

Target identification is one of the key steps of drug development [8,9]. Tremendous efforts have been made in recent decades by pharmaceutical industries and biotechnology companies that focus on the druggable genome [5,6] to identify novel drug targets for drug discovery; however, only a few drug targets are successfully used in current therapies [10]. The human genome project has identified approximately 20,000–25,000 genes and an even larger number of transcripts and proteins, which provide great opportunities for drug target investigation [11]. Currently, PubChem records two major types of molecular targets for research (i.e. protein targets for small molecules and gene targets from RNAi reagents), which represent a great diversity of types of assays, including, for example, enzyme inhibitor identification, protein–protein interactions, tumor cell growth inhibition and even organismal phenotypes. Because the protein targets are of particular interest to researchers in drug discovery and the majority of bioassays in PubChem focus on enzymes or other proteins, we focus on the analysis of protein targets in this study. A collection of 2206 protein targets was compiled from PubChem at the time of this work.

Functional families

To look into the potential functions of these bioassay targets, we performed sequence similarity search against the annotated functional domains in the NCBI Conserved Domain Database (CDD; http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) [12] by using the reverse-position-specific BLAST tool [13]. We found that the 2206 protein targets fell into 671 unique protein superfamilies (Fig. 1a). Approximately 15% of them belonged to the protein kinase superfamily. Other superfamilies such as nuclear receptor, trypsin-like serine protease, src homology protein and zinc-dependent metalloprotease comprised approximately 2–3% of the bioassay targets. The rest of the superfamilies (67%, 450 out of 671) contained only one or two bioassay targets for each member. In particular, the high-throughput screening assays under the MLP contributed 450 protein targets, scattering into 312 protein superfamilies (Fig. 1c). Although the protein kinase superfamily still dominated this subset, it accounted for 5% of the MLP target set.

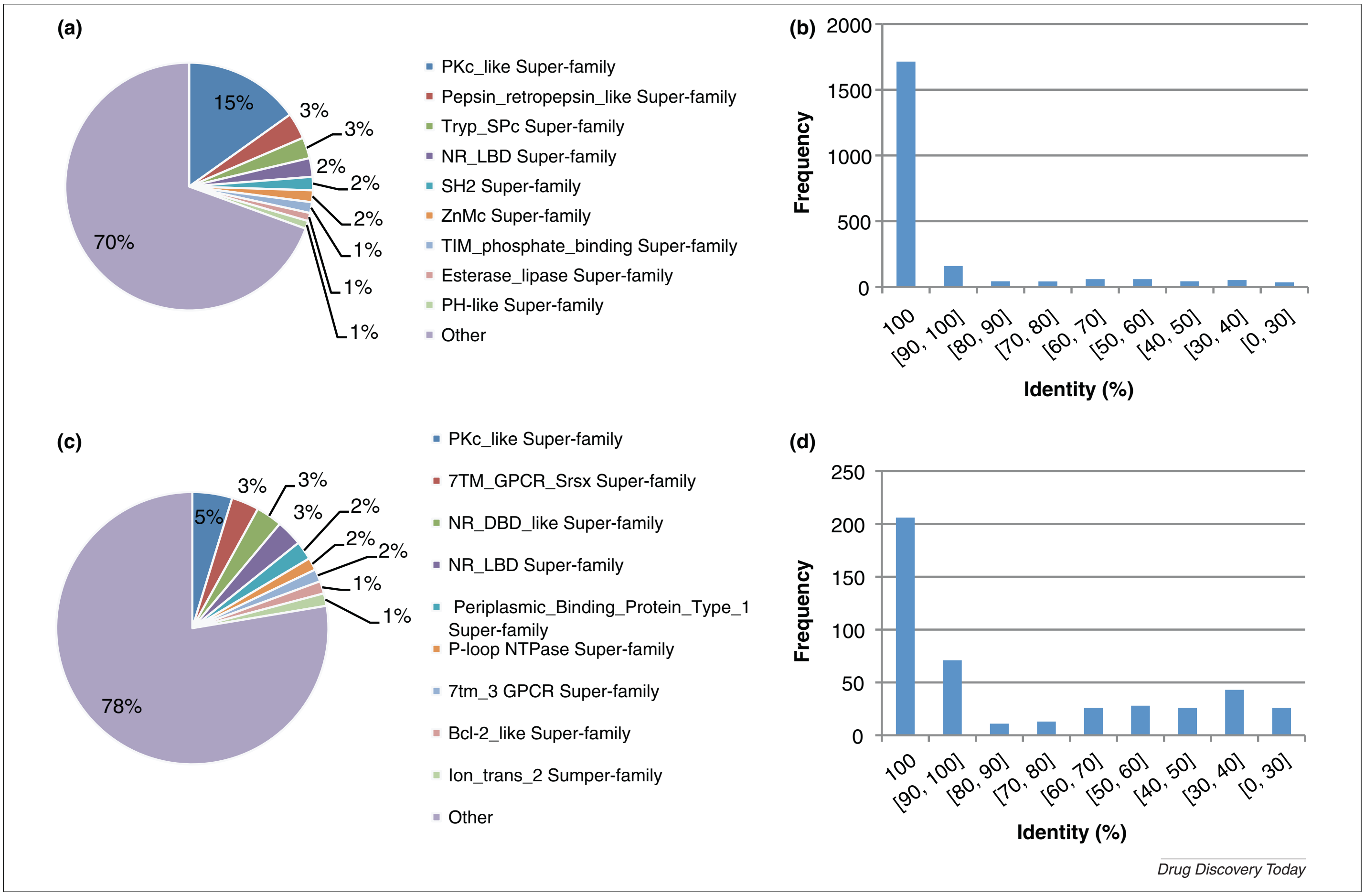

FIGURE 1

Linking protein superfamilies and 3D structures to PubChem bioassay targets. **(a)** and **(c)** represent superfamily annotations; **(b)** and **(d)** indicate the availability of related 3D structures derived from homologous analysis at each sequence similarity level. (a) and (b) denote the entire set of protein targets in PubChem; (c) and (d) represent the protein targets that are involved in high-throughput screening assays from the MLP.

The other superfamilies, such as seven-transmembrane G-protein-coupled receptors and DNA-binding domain of nuclear receptors accounted for 2–3% on average. These results suggest that the bioassay targets in PubChem represent a broader functional diversity than the known druggable targets; thus, PubChem enables researchers to study the mechanisms of protein–ligand interactions on a wider scope and to identify novel molecular targets for potential treatments.

Three-dimensional structures

The 3D structures of macromolecular targets are important to the study of the mechanisms of protein–protein and protein–ligand interactions. To link the protein targets to relevant 3D structures, we used the BLAST tool [14,15] to search against the protein sequences derived from the Protein Data Bank (PDB; http://www.pdb.org) [16]. We found that 78% of these targets have corresponding 3D structures with 100% sequence identity in the PDB database (Fig. 1b). When looking into the possibility of inferring related structures from the similarity search, another 8% of these targets found related structures in the PDB database with sequence identity over 90%. Given the fact that protein structures tend to be highly conserved at this level of sequence identity, this analysis suggests that more than 86% of the molecular targets in PubChem have related structural information in PDB. Conversely, less than 2% of these targets could not be linked to any relevant 3D structures or were only able to be linked to the related protein structures with sequence identity below 30%. As for the 450 protein targets from the MLP, more than 60% either have corresponding 3D structures with sequence identity of 100% or can be linked to related structures with sequence identity of 90% or above (Fig. 1d).

Related pathways

Most diseases occur because of the misregulation of multiple genes that are involved in mutual interactions – including genes, transcripts and proteins – in a dynamic network. During the past decade, high-throughput technologies have been widely used in biological research and generated a tremendous amount of experimental data, which make it possible to study the functions of genes or proteins at a biological system level. Drug development is inherently a complicated process because drugs and their targets are engaged in a complex system, which is far from being thoroughly understood. Moreover, approximately 35% of known drugs or drug candidates are active against more than one target [17], which makes the interactions more sophisticated. Therefore, it is essential to investigate the connections of the drug, drug target and disease in the context of a biological system.

In this study, we mapped 507 (23%) of the 2206 protein targets from PubChem to 287 pathways in the KEGG database (http://www.genome.jp/kegg/) [17–20]. We observed that some pathways, such as the mitogen-activated protein kinase signaling pathway, were related to multiple protein targets in PubChem. In addition, some bioassay targets were involved in multiple KEGG pathways. A list of top 20 pathways that contain multiple bioassay targets and top 20 targets that are involved in multiple pathways are provided in Tables S1 and S2, respectively, in the supplementary material online. Targets involved in the same pathway are likely to have similar roles in regulating a specific biological process. Thus, selectively inhibiting or activating a target in the same pathway might effectively modulate a specific biological process or restore the function from a disease state back to a normal one. Thus, the wealth of bioactivity data in PubChem might facilitate research into chemical biology and drug development at the system level.

Bioactive compounds in PubChem

The characteristics of small molecules make them useful, not only as drugs that modulate physiological functions but also as chemical tools that interrogate the functions of novel genes, pathways and cells [3]. The purpose of the MLP is to develop chemical probes for modulating biological processes and facilitate the development of new drugs by offering the capacity of high-throughput screening to the public sector [3]. Currently, more than one million compounds have been tested against several thousand targets and deposited in PubChem. Approximately two hundred thousand of them were reported active, among which there were 116 chemical probes generated by the MLP projects at the time of this article.

Potency

A large fraction of the bioactive compounds (91,022) in PubChem were assayed with a confirmed potency measurement, which were associated with 1771 out of the 2206 protein targets in total. The distribution of bioactivity potency was analyzed, with the results showing that nearly 10% of the compounds have a potency of ≤ 1 μM (Fig. 2a). These compounds were associated with more than 60% of the 1771 targets (i.e. each of these targets had at least one bioactive compound with a potency of ≤ 1 μM). We found, however, that approximately 40% of the targets had no active compound with a potency better than 10 μM (Fig. 2a), which indicates that there are great chances to develop highly potent compounds for these targets through further study by medicinal chemistry approaches. When focusing on the 116 MPL chemical probes, we found that most of them demonstrated much higher potency in the range of 0.001–1 μM (Fig. 2b). The MLP probes are discussed in detail in the section 'Chemical probes'.

Selectivity and promiscuity

It is essential to understand the selectivity and promiscuity of small molecules when fully exploiting the therapeutic potential and minimizing the toxic effects of drugs or drug candidates [17,21,22]. To evaluate these properties of a compound, a straightforward approach is to investigate the bioactivity profile by screening this compound across a broad panel of targets; however, this could be expensive when applied to a large compound library. As more data are available in PubChem, however, it will be possible to derive such bioactivity profiles for a particular chemical compound, as well as to investigate the selectivity and promiscuity against a specific target by combining the assay results contributed by many organizations. In particular, the projects under the MLP, which share a common library of more than 340,000 compounds, make it feasible to systematically derive target profiling information for many bioactive compounds.

We performed an across-target activity analysis for all of the 189,807 active compounds in PubChem to identify the selective

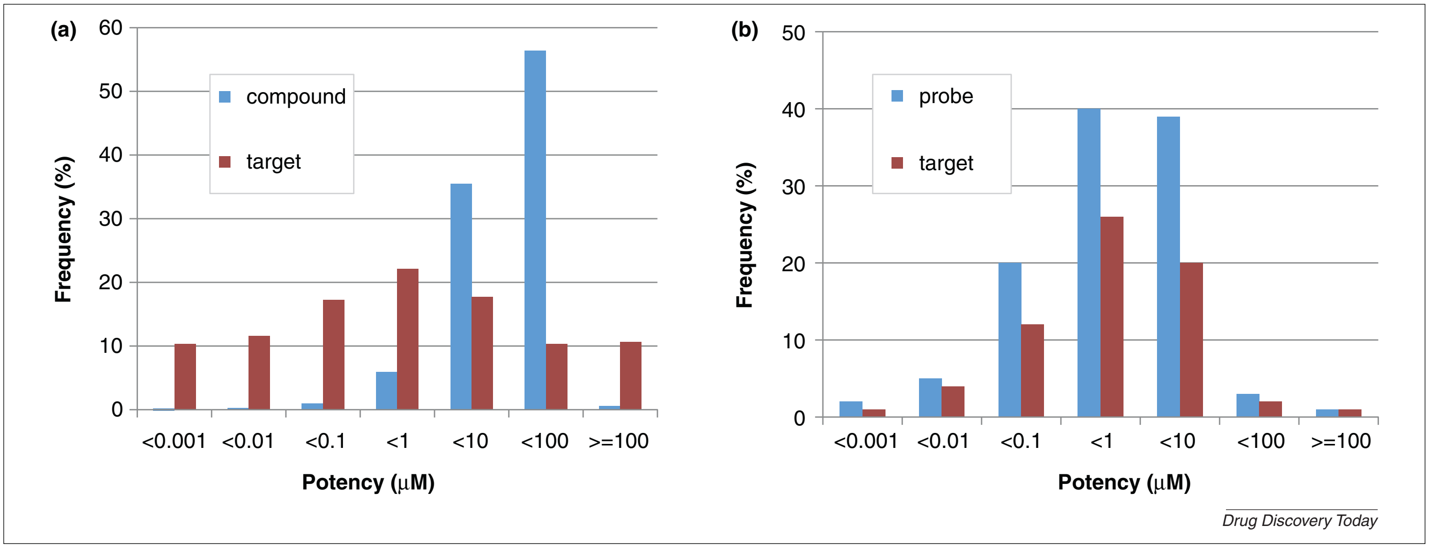

FIGURE 2

The distribution of the potency of the bioactive compounds and protein targets in PubChem. **(a)** Blue bars show the distribution of the potency for entire bioactive compounds in PubChem in seven potency groups; red bars denote the frequency of the protein targets with most potent compounds falling in respective potency group. **(b)** Blue bars represent the distribution of potency for chemical probes identified by the MLP; red bars show the frequency of the protein targets with most potent chemical probes falling in respective potency group.

and promiscuous compounds, following the procedure described previously [23]. As a result, 38% (71,627) of those compounds were observed as potentially selective with bioactivity outcome reported active against a single target, and the rest of them (62%) demonstrated active against multiple targets, with a portion of them hitting multiple but otherwise related targets (Fig. 3a). Many bioassay targets in PubChem are biologically related, as revealed by sequence homology analysis [1]. In particular, the MLP projects usually take a secondary screening against related targets in the search for compounds with higher specificity. Thus, it is not surprising to often observe common hits for related targets. However, there are many other causes of the promiscuity of a compound [24]. To address this issue, the MLP has developed several profiling bioassays for evaluating aggregation effects, filtering chemical reactivity and identifying interference molecules, including screenings for luciferase inhibitors by multiple laboratories. In summary, all of the information has made PubChem a valuable resource for studying the promiscuity of chemical com-

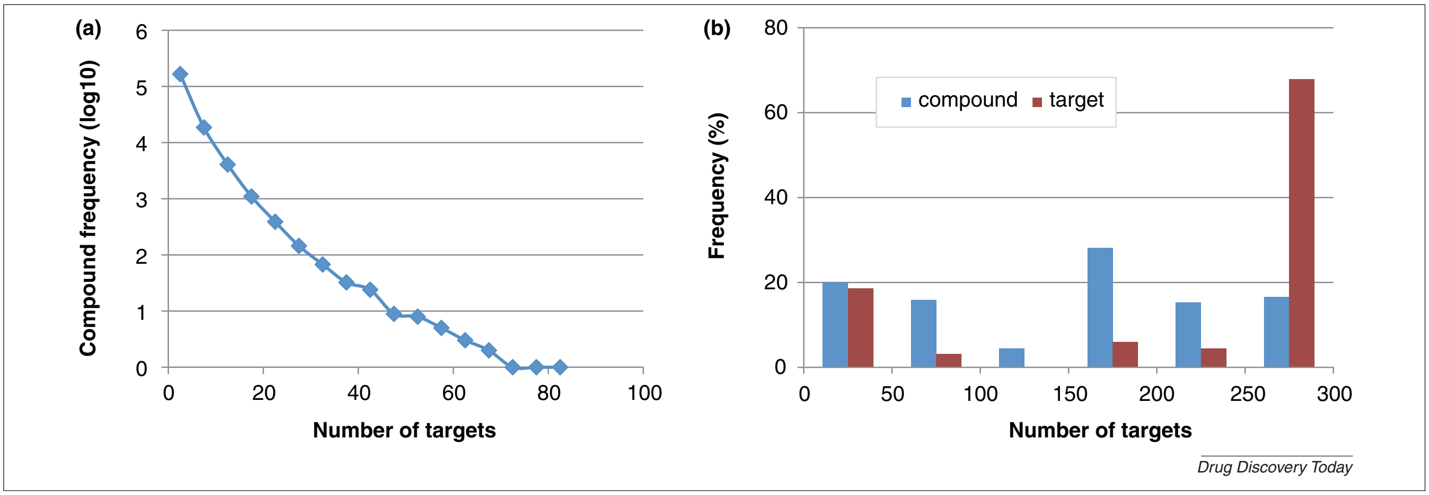

FIGURE 3

An overview of the selectivity property of bioactive compounds in PubChem. **(a)** The distribution of compounds' across-target activity. The *x* axis represents the number of distinct active protein targets associated with a compound, and the *y* axis represents the frequency of compounds at each across-activity level. This shows that majority compounds are associated with one or a few protein targets, and a small portion of them interact with a large amount of targets. **(b)** The blue bars represent the distribution of the number of tested targets for selective compounds (only active to one protein target in PubChem). Compounds are divided into six selectivity groups. This suggests that the majority of the selective compounds have been tested across more than 150 targets. The red bars denote the frequency of the protein targets associated with the compounds in the respective selectivity group.

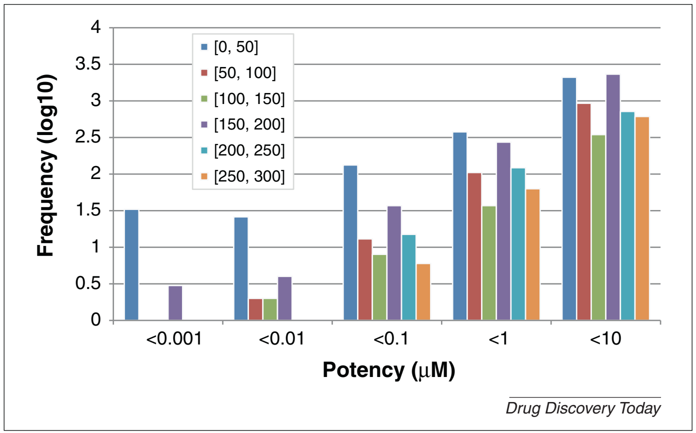

FIGURE 4

The distribution of the selective compounds in PubChem. The *x* axis represents the potency in micromole; the *y* axis represents the frequency in the scale of log10. The color of the bars denotes the range of targets against which the selective compounds were tested. For example, the range of [100, 150] means that this group of compounds was tested against 100–150 targets and selectively against one of them.

pounds and investigating the polypharmacology properties of chemical compounds in system-based drug discovery [25,26].

As it would be necessary to assess the selectivity and promiscuity properties in the context of tested targets, we looked into those potentially selective compounds (71,627) and observed that approximately 80% of them were tested against at least 50 distinct protein targets, and a significant portion (60%) was highly selective, as tested against more than 150 targets (Fig. 3b). We also observed that 14% (316) of the 2206 targets were associated with at least one of these selective compounds. Among this subset of targets, more than 60% of them were associated with highly selective compounds that were tested broadly across more than 250 distinct protein targets (Fig. 3b). These results indicate that compounds with potentially high selectivity are available for a great portion of protein targets in PubChem. In addition, we evaluated the potency of these selective compounds by dividing them into several selectivity groups based on the number of targets tested (Fig. 4). This analysis provides further insights into both the selectivity and the potency of the bioactive compounds in this subset. It enables one to apply a certain selectivity threshold to identify the compounds with a desired potency and to track down the molecular target associated with the compound as well, which might serve as a starting point for a medicinal chemist to further optimize the bioactive compound towards a chemical probe or a drug candidate.

Chemical probes

At the time of this work, the MLP project has generated 116 chemical probes. The detailed descriptions of the characterizations of the probes are publicly available for the community to review (http://mli.nih.gov/mli/mlp-probes/). These MLP chemical probes were associated with 67 individual protein targets, which fell into 89 CDD superfamilies (some targets belonged to more than one superfamily) according to the CDD functional domain annotations. Among them, 36 protein targets had corresponding 3D structures with sequence identity of 100% in the PDB database and 41 were mapped to 155 relevant conserved pathways in the KEGG database. The distribution of the bioactivity potency of these MLP chemical probes with their corresponding targets is shown in Fig. 2b. The chemical probes with potency in the range of 0.001–1 μM have been found for more than 60% of the protein targets (43 out of 67), which indicates varying quality of the probes with respect to potency. Compared to other bioactive compounds in PubChem, the MLP probes demonstrate higher potency and considerably better selectivity for the respective targets in general. As several literature-based bioactivity databases become publicly available [27–29], it is also possible to gain insights into the novelty of the MLP probes by comparing them with the prior art. Detailed information of the MLP chemical probes, including bioactivity potency, biological pathways and related 3D structures of their targets, is provided in Table S3 in the supplementary material online.

Recently, there have been intensive discussions on the criteria and principles of defining a chemical probe, and some contradictory opinions have been raised [21,30]. Although only a portion of the MLP chemical probes seem to have medium or high quality based on a crowdsourcing evaluation [31] and most of them have low citation rates by the bibliometric method [30,32], it would probably take more time to find out their merits in future studies. Researchers in both academia and industry can help, however, and are highly encouraged to assess and improve the MLP chemical probes through their own research. To this end, the efforts undertaken by the MLP to further characterize the probes and make the data publicly accessible through PubChem would help make this happen.

Concluding remarks

PubChem is growing rapidly with new data being deposited on a daily basis, which makes it both feasible and imperative to evaluate the properties of a particular bioactive compound, a drug candidate or even a known drug on a large scale to identify potentially new functions or off-target effects. It is starting to emerge as a valuable resource to explore the functions of genes and proteins in physiology and pathology. A summary of public services and tools are listed in Table S4 in the supplementary material online to facilitate use of the data in PubChem.

As a public molecular information resource at NIH, the free availability of PubChem will undoubtedly lower the barrier for researchers from chemical biology, medicinal chemistry and drug discovery to advance the development of new chemical tools for interrogating biological functions and potential drug candidates for disease treatments. It also provides great opportunities for researchers in bioinformatics and cheminformatics to tackle the problems in those research fields with computational approaches.

Acknowledgements

We thank the National Institutes of Health Fellows Editorial Board for providing editorial assistance. This work is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.drudis.2010.10.003.

References

1 Wang, Y. *et al.* (2009) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* 38, D255–D266

2 Wang, Y. *et al.* (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633

3 Austin, C.P. *et al.* (2004) NIH Molecular Libraries Initiative. *Science* 306, 1138–1139

4 Editorial, (2009) Perfecting probes. *Nat. Chem. Biol.* 5, 435

5 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730

6 Russ, A.P. and Lampel, S. (2005) The druggable genome: an update. *Drug Discov. Today* 10, 1607–1610

7 Drews, J. (2006) What's in a number? *Nat. Rev. Drug Discov.* 5, 975

8 Lindsay, M.A. (2003) Target discovery. *Nat. Rev. Drug Discov.* 2, 831–838

9 Harland, L. and Gaulton, A. (2009) Drug target central. *Expert Opin. Drug Discov.* 4, 857–872

10 Overington, J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.* 5, 993–996

11 Mayr, L.M. and Bojanic, D. (2009) Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* 9, 580–588

12 Marchler-Bauer, A. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.* 33, D192–D196

13 Marchler-Bauer, A. *et al.* (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 30, 281–283

14 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

15 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

16 Bernstein, F.C. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535–542

17 Xie, L. *et al.* (2009) Drug discovery using chemical systems biology: identification of the protein–ligand binding network to explain the side effects of CETP inhibitors. *PLoS Comput. Biol.* 5, e1000387

18 Kanehisa, M. *et al.* (2009) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360

19 Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30

20 Kanehisa, M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–D357

21 Frye, S.V. (2010) The art of the chemical probe. *Nat. Chem. Biol.* 6, 159–161

22 Rix, U. and Superti-Furga, G. (2009) Target profiling of small molecules by chemical proteomics. *Nat. Chem. Biol.* 5, 616–624

23 Han, L. *et al.* (2009) A survey of across-target bioactivity results of small molecules in PubChem. *Bioinformatics* 25, 2251–2255

24 Feng, B.Y. *et al.* (2005) High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* 1, 146–148

25 Yildirim, M.A. *et al.* (2007) Drug–target network. *Nat. Biotechnol.* 25, 1119–1126

26 Chen, B. *et al.* (2009) PubChem as a source of polypharmacology. *J. Chem. Inf. Model.* 49, 2044–2055

27 Liu, T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201

28 Overington, J. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr.. *J. Comput. Aided Mol. Des.* 23, 195–198

29 Harmar, A.J. *et al.* (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.* 37, D680–D685

30 Workman, P. and Collins, I. (2010) Probing the probes: fitness factors for small molecule tools. *Chem. Biol.* 17, 561–577

31 Oprea, T.I. *et al.* (2009) A crowdsourcing evaluation of the NIH chemical probes. *Nat. Chem. Biol.* 5, 441–447

32 Bologa, C. (2010) Promiscuity and PubChem: a retrospective analysis. In *Proceedings of the Society for Biomolecular Screening 16th Annual Conference and Exhibition* pp. 119