Published online 4 June 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W623–W633***
doi:10.1093/nar/gkp456

PubChem: a public information system for analyzing bioactivities of small molecules

Yanli Wang, Jewen Xiao, Tugba O. Suzek, Jian Zhang, Jiyao Wang and Stephen H. Bryant*

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

Received March 24, 2009; Revised May 12, 2009; Accepted May 14, 2009

ABSTRACT

PubChem (http://pubchem.ncbi.nlm.nih.gov) is a public repository for biological properties of small molecules hosted by the US National Institutes of Health (NIH). PubChem BioAssay database currently contains biological test results for more than 700 000 compounds. The goal of PubChem is to make this information easily accessible to biomedical researchers. In this work, we present a set of web servers to facilitate and optimize the utility of biological activity information within PubChem. These web-based services provide tools for rapid data retrieval, integration and comparison of biological screening results, exploratory structure–activity analysis, and target selectivity examination. This article reviews these bioactivity analysis tools and discusses their uses. Most of the tools described in this work can be directly accessed at http://pubchem.ncbi.nlm.nih.gov/assay/. URLs for accessing other tools described in this work are specified individually.

INTRODUCTION

PubChem (1) (http://pubchem.ncbi.nlm.nih.gov) is an open repository for chemical structures and their biological test results. It was launched in September 2004 as part of a research program under the NIH Molecular Libraries Roadmap Initiative, which aimed to discover chemical probes through high-throughput screening of small molecules that modulate the activity of gene products (2,3). PubChem contains three related databases: Substance, Compound and BioAssay. The Substance database (primary accession-SID) contains contributed sample descriptions provided by depositors, whereas the Compound database (primary accession-CID) contains unique chemical structures derived from the substance depositions. The PubChem BioAssay database (primary accession-AID) contains bioactivity screens of chemical substances described in PubChem, and serves as the public repository for the biological screening results contributed by the NIH Molecular Library Program (4,5), other research organizations and industrial companies (6–10). A BioAssay data entry contains contributed bioactivity descriptions and test results, such as percentage of activity inhibition, generated by one assay protocol. About 30 academic institutions, government agencies, research laboratories, as well as industrial assay vendors have deposited biological test results, which were either generated by HTS screenings or extracted from literature, to the PubChem BioAssay repository. The PubChem BioAssay database currently contains more than 1400 bioassay depositions and 45 millions of biological activity outcomes for over 700 000 compounds of unique chemical structures. With the accelerated growth of biological screening results in both volume and complexity, the need for computational tools to retrieve and analyze such rich data becomes more imperative.

The goal of the PubChem BioAssay system is to provide services to make the screening data of this large scale collection easily accessible to the public, and provide data analysis tools to facilitate the utilization of this valuable information. To this end, information content of PubChem can be accessed through the NCBI Entrez system. One can search, in Entrez's PubChem Compound database, a compound with a chemical synonym, and subsequently link to a list of screening experiments involving the given compound via the 'BioAssays' link (1). One can also find, in Entrez's PubChem BioAssay database, all bioassay tests for a specific target by querying the name of the protein target. The Entrez 'Limits' facility allows one to construct a specific query based on one's research need. Furthermore, PubChem provides a set of web-based tools to integrate the chemical and biological activity information, and support the navigation and in-depth data analysis that facilitates identification of chemical probes and biological interesting targets contained within PubChem databases. For these tools,

*To whom correspondence should be addressed. Tel: +1 301 435 7792; Fax: +1 301 480 9241; Email: bryant@ncbi.nlm.nih.gov

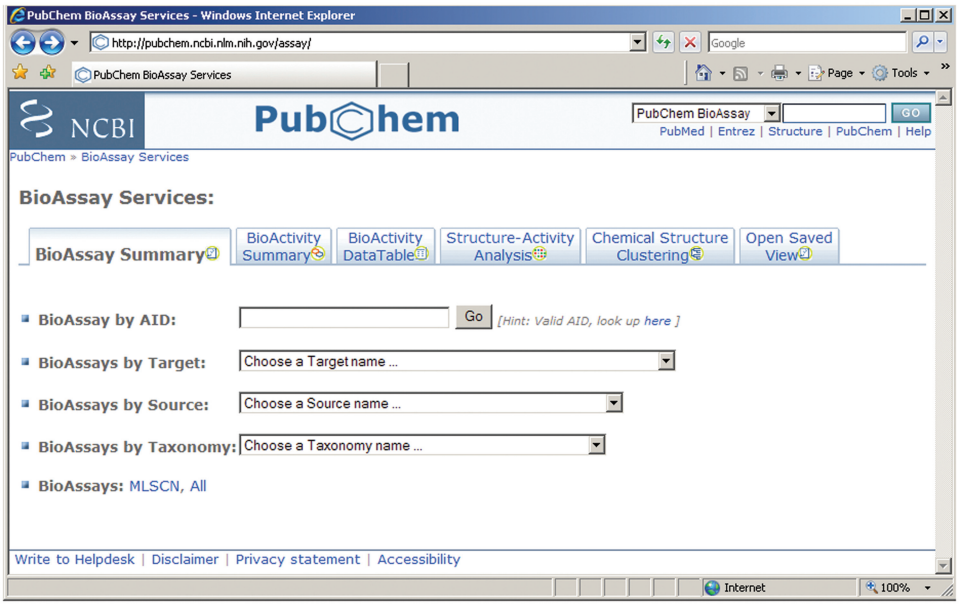

Figure 1. Common gateway of PubChem BioActivity Analysis Service. It provides a central entry point for accessing bioassay records, and tools including BioAssay Summary, BioActivity Summary, Data Table and Structure–Activity Analysis. Files saved for recording analysis status can be imported using the 'Open Saved View' tab.

PubChem implements a queuing mechanism to balance web requests.

The BioAssay information system provides access to review the deposited bioassay record in detail. It also provides exploratory data analysis tools by using the 'summary' biological test results. Screening descriptions and test results in PubChem BioAssay database are diverse and assay-specific. PubChem allows depositions of as many readouts as it needs to be, and provides tools to search, select and retrieve such deposited information. Despite this flexibility, PubChem requires a summary result for each tested chemical sample for defining the bioactivity outcome and bioactivity score. PubChem bioactivity outcome summary includes five categories, e.g. chemical probe, active, inactive, inconclusive and unspecified. For the dose–response screening test, the primary endpoint, such as IC50, is denoted as 'active concentration' summary, and needs to be reported in micromolar units. Such summary results allow one to classify and rank hits of a screening test. More importantly, having a bioactivity outcome summary for each tested sample allows PubChem to provide tools for rapid comparative analysis across multiple screening reports for given chemical samples. PubChem has integrated the bioactivity analysis tools (Figure 1) using these summary results to provide a comprehensive review of biological tests, compare biological activity data from multiple screenings, and explore structure–activity relationship.

There are a number of publicly available databases providing bioactivity information and data mining tools. Among the well known resources, BindingDB (9), IUPHAR (10) and PDBBind (11), focus on collecting and curating bioactivity data from literature. ChemBank (12) is a database for biological screening data generated by HTS experiments. It provides cross-experiment analysis through a heatmap visualization service; however, classification functionality has not been reported. GLIDA (13) primarily focuses on the integration of information between GPCRs and their ligands. It provides a useful correlation map to facilitate the study of correlation patterns between GPCR targets and their respective ligands using a few distinct types of bioactivity information. PubChem provides a similar structure–activity analysis tool, yet it allows one to derive and compare activity profile of the compound based on the quantitative bioactivity data across a broad range of targets. Overall, the PubChem bioactivity analysis services are clearly advantageous in several respects. These services are seamlessly integrated and allow users to set and refine research focus as needed to be in the process of data analysis. They leverage on the powerful facilities provided by the NCBI Entrez system and allow one to utilize the information relationship among the data content between PubChem and other NCBI databases. Some other unique features of the PubChem bioactivity analysis services include the facilities supporting test result drill-down, registration-free bulk download for screening results, structures of tested chemicals, as well as similarity matrices used in various data analysis. Multiple entry points are provided for these services to support chemistry, bioassay or molecular target centric analysis.

BioAssay SUMMARY

PubChem BioAssay Summary service (Figure 2) is the primary service for presenting depositor-provided information, represented by a PubChem BioAssay accession, AID. This includes a summary of data attribution, assay

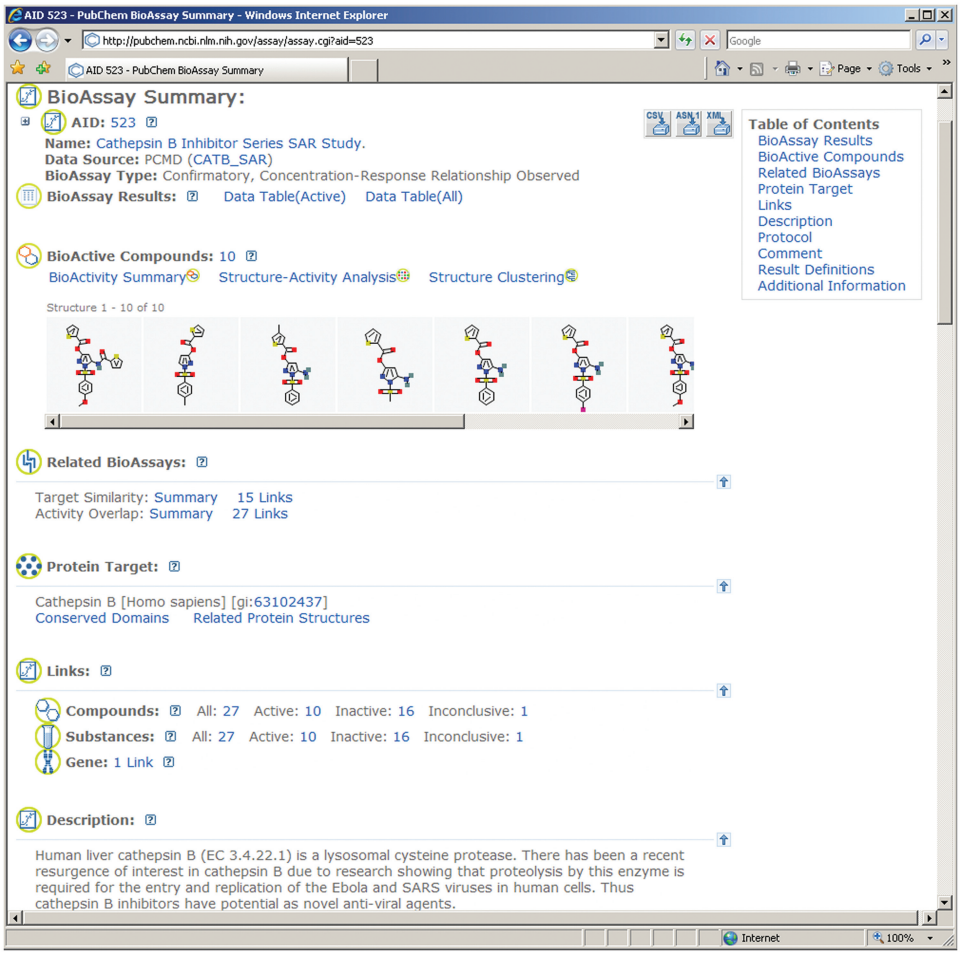

Figure 2. The BioAssay Summary view for PubChem assay AID 523, a screening experiment aimed to identify inhibitors of human liver cathepsin B, a lysosomal cysteine protease which is associated with several human diseases. PubChem provides similar summary for each of the bioassay records, which can be accessed at the same url shown by this screen shot with the respective AID.

description, experimental protocol, depositor comments, screening outcome methodology and definitions of reported readouts. It also includes depositor supplied cross links to tested substance samples, hit compounds, protein and gene target, PubMed publications, and other information resources. Overall, it provides a comprehensive description of a bioassay report, helps user to understand the scientific goal of the experiment, the biological background of the testing system, assay technology exploited, discoveries made by the screening test, as well as threshold used for deriving bioactivity outcome and possible factors of artifacts. PubChem BioAssay database tracks and archives each update of an assay submission. This service enables one to review the update history of a bioassay deposition by providing a list of the versions, deposition date and modification date. One may check such information by clicking on the '+' sign to the left of AID and retrieve the information content for each update by specifying the version ID using the 'BioAssay Version' selection tool. Biological test results for the active compounds can be accessed through the 'Data Table (Active)' link, and those for all tested compounds can be accessed through the 'Data Table (All)' link.

Another goal of the BioAssay Summary service is to present annotations provided by PubChem. The assays in the PubChem BioAssay database are related in many ways, for example, through testing on biologically related targets, reporting overlapped hits, or reporting counter screening tests for the same assay project. Such assay relationships are identified and summarized through the links under the 'Related BioAssays' section on the page. Annotations for protein targets are provided by linking to the protein classification and known 3D protein structure resources at NCBI. Following the result from an individual-screening experiment, this BioAssay Summary service serves as a central entry point to evaluate the screening results, discover further biological properties of small molecules by using additional screening results within PubChem, and begin analysis of active compounds by allowing users to access various bioactivity analysis tools using the links under the 'BioActive Compounds' section, including the BioActivity Summary tool and Structure–Activity Analysis tool as described below. The BioAssay Summary server can be accessed at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=, for each PubChem BioAssay deposition by providing a valid

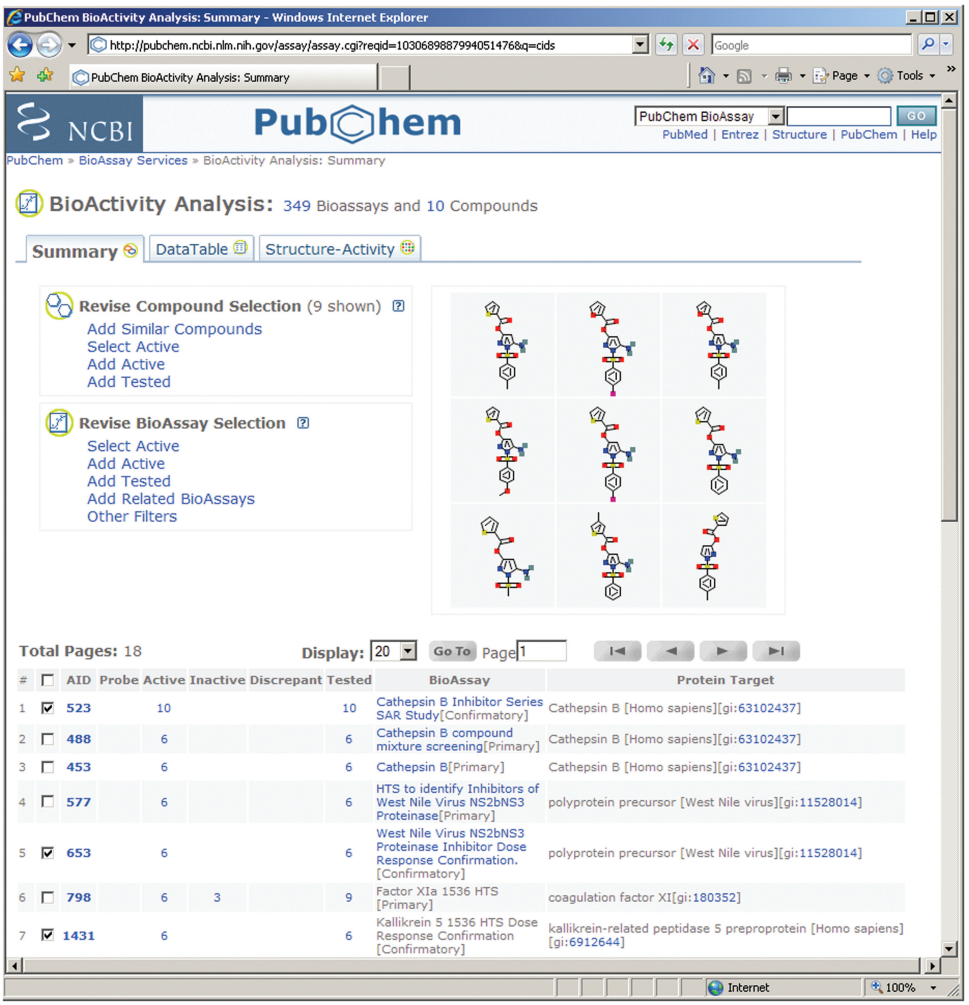

Figure 3. A view of BioActivity Summary analysis for the 10 compounds active in the secondary assay 'Cathepsin B Inhibitor Series SAR Study' (AID 523). This analysis shows that these compounds are tested in several hundreds of screening experiments, and some compounds are considered active in other biological tests as well. A further analysis by revising the assay focus using the 'Select Active' tool will tell the number of assays having one or multiple hits overlapping with those in AID 523. To review compounds and the test results from the 'HTS to identify Inhibitors of West Nile Virus' assay, AID 577, for example, one may click on '6' shown under the 'Active' column to retrieve assay results given in AID 577 for the six compounds which are considered active in that assay. Alternatively, one may select assays using the check box on left of AID, and invoke the 'Data Table' tab to retrieve test results from the selected assays.

accession number, e.g. AID. A complete list of PubChem BioAssay records can be obtained at http://www.ncbi.nlm.nih.gov/sites/entrez?db=pcassay&term=all[filt].

BioActivity SUMMARY TOOL

The PubChem BioActivity Summary tool (Figure 3) allows one to aggregate all available screening results, and readily examine and compare biological outcomes across multiple assays for one or more tested compounds or substances. It reports and summarizes the available screening bioactivity outcomes for a single or a set of chemical samples. This service updates the screening results, on a daily basis, with information from new bioassay depositions or updated results, and therefore provides a comprehensive overview of the biological profile for tested small molecules using up-to-date screening results within PubChem. Moreover, this service provides functionalities to tailor the compounds and assays to a focused data set that meets one's research goal. It thus offers a platform for constructing a panel of assays and compounds for further exploratory structure–activity analysis and target selectivity evaluation by linking to additional bioactivity analysis tools.

Depending on users' goals, the summary of bioactivity can be switched between compound-centric and substance-centric views. If centered on substance descriptions, this particular tool provides a summary view of all available biological tests and the respective bioactivity outcome contributed by a single organization. If the substances are deposited by the MLSMR (NIH Molecular Libraries Small Molecule Repository), they can be tested by multiple screening centers within the NIH Molecular Library Program. If centered on compound descriptions, the service provides a comprehensive summary view of biological activity by aggregating all screening data across multiple

contributors for the unique chemical structures. Given the difference in scope between compounds and substances, we will consider only the case of compounds in more detail below.

The BioActivity Summary tool can be invoked from any BioAssay Summary page. In this case, the active compounds of the bioassay will become the default focus. For example, in the 'Cathepsin B inhibitor Series SAR Study' assay, AID 523(http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=523), 10 compounds are identified as potential cathepsin B inhibitors (Figure 2). One may perform the BioActivity Summary analysis for the active compounds using the 'BioActivity Summary' link under the 'BioActive Compounds' section (Figure 2). The result of such analysis is shown in Figure 3. These 10 active compounds are tested in over 300 assays, deposited in PubChem at the time of this writing, including AID 523. In the summary table, assays are sorted by the count of active compounds. Each assay is summarized, one per line, with the count of active, inactive and tested compounds. The assay name, bioactivity outcome method type, e.g. primary versus confirmatory, and protein target specification give a summary of what screenings are performed for the compounds under analysis. More importantly, one can see that some of the compounds that are active in AID 523 also show activity in a number of other assays. The confirmatory assays ranked near the top of the table include assay targets from several other protease-related proteins, such as West Nile Virus NS2bNS3 proteinase, Factor XIa, kallikrein-related peptidase 5, Cathepsin G and Factor XIIa. This suggests that some of these 10 compounds are nonspecific inhibitors for the Cathepsin B target and demonstrate cross reactivity against a group of biologically related proteins.

The BioActivity Summary service provides several powerful selection-revise features that enable one to rapidly revise the focus of the analysis by modifying the set of selected compounds and assays. As depicted in Figure 3, one may choose to expand the current set of compounds to include all tested compounds or active compounds within a given set of bioassays, or expand the compounds to include those having similar 2D structures to the ones in the current set. Alternatively, one may choose to limit the compounds to only those found active in the assay subset. Similarly, the selected assay set can be expanded to include those that are active or tested for a given set of compounds. Additionally, bioassays with similar bioactivity profiles and bioassays with similar protein target sequence can be added to selection. For example, to focus on a subset of the input compounds that are active in one or more of the selected assays, user would click the 'Select Active' link in the 'Revise Compound Selection' section to exclude the less interesting compounds, or to explore additional assay screens where the given compounds are considered active by using the 'Add Active' link in the 'Revise BioAssay Selection' section. One may also choose to focus on confirmatory assays or assays with specific molecular targets using the filtering features provided in the 'Other Filters' pop-up menu.

One of the common entry points for accessing the BioActivity Summary tool is from a single PubChem compound summary record. Invoking the BioActivity Summary tool from a compound summary record will readily generate an overview of all biological screenings performed for that compound. From the BioActivity Summary page users can expand the analysis by including compound similar in 2D chemical structures via 'Add Similar Compounds' link in the 'Revise Compound Selection' section. This operation adds compounds with significant 2D structural similarity and allows users to collect and examine the bioactivity data among tested analogs. A further request of all bioassays tested for the analog series using the 'Add Tested' link in the 'Revise BioAssay Selection' section may reveal additional important screenings where the structural analogs are tested. Subsequent analysis using the Structure–Activity Analysis tool (to be described) enables further evaluation on the SAR and bioactivity profile of such analog series.

Other entry points include NCBI Entrez 'DocSum' reports for PubChem substance, compound and bioassay records, where the BioActivity Summary tool can be invoked for each individual record as well as for the entire data set resulted from an Entrez search. This can be done by using the explicit 'BioActivity Analysis' link, or clicking the double six-member ring icon from the 'Tools' area. For example, one may start, in Entrez's PubChem Compound database, with a compound submitted to PubChem by a journal article reporting specific enzyme inhibitors. To verify the discussed inhibition activity of the enzyme inhibitor, one can compare the reported bioactivity information to the biological tests deposited in PubChem. Alternatively, one may start a structure search with a given substructure using the service provided at http://pubchem.ncbi.nlm.nih.gov/search/search.cgi, and launch the BioActivity Summary tool for the resulting compound set to the link described above. In another case, one may search PubChem BioAssay database for all available screening tests for a particular target, then use the BioActivity Summary tool to examine the bioactivity outcomes from each screening experiment, compare the hit list and compile a library of bioactive compounds for the target. Users can also choose to access this analysis tool through the common gateway of PubChem BioActivity Analysis Service provided at: http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=bioactivity. From this entry point, the dataset of assay, substance/compound, which are subject to the analysis, can be specified by entering an ID list, providing a text file contains the IDs (comma separated, or one ID per line), or referring to an Entrez search history. This entry point provides the flexibility to focus the analysis on a well defined compound and assay dataset, for example, to evaluate the toxicity properties for the compounds of one's interest using only toxicity profiling assays available at PubChem. In the case that only compound input is specified, a summary for all available biological test results will be provided.

In addition to providing a bioactivity overview, this service serves as the starting point in the bioactivity analysis

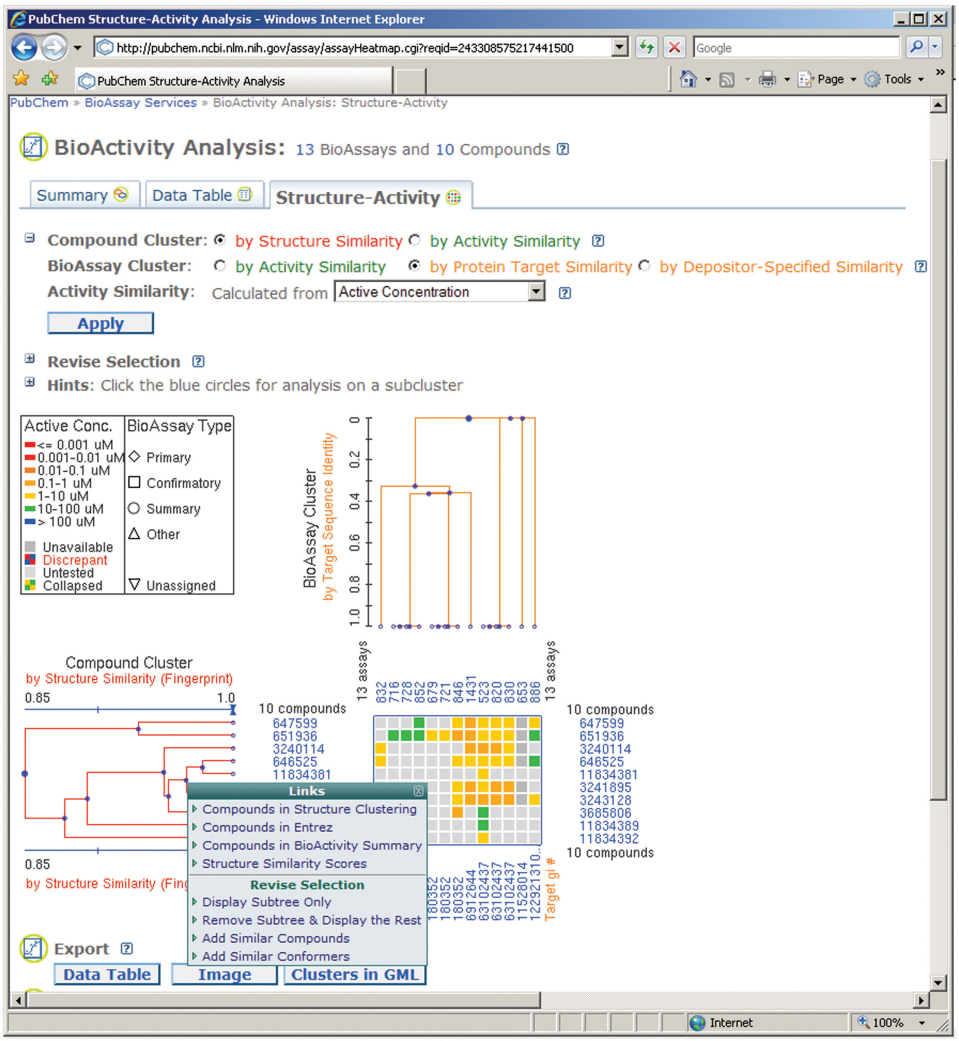

Figure 4. Screen shot of PubChem structure-activity analysis for the 10 active compounds in AID 523 and several confirmatory assays against a few protein targets. Assay clusters, based on assay target sequence similarity, are shown in the horizontal dimension, while compound clusters, based on 2D structure similarity, are shown in the vertical dimension. Each cell in the heatmap is colored based on the reported active concentration value (e.g. IC50) according to the legend contained within the figure. The PubChem Compound accession, e.g. CID, is shown to the right of each leaf of the compound cluster dendrogram. A chemical structure display can be invoked upon mouse-over the respective CID. PubChem BioAssay accession, e.g. AID, is shown beneath each leaf of the assay cluster dendrogram, while GI numbers for the respective assay targets are provided below the heatmap and hyperlinked to the corresponding Entrez protein records.

process, which leads users to further analysis using the Structure–Activity Analysis tool and Data Table tool which will be described in the following sections. PubChem Data Table tool supports the retrieval of assay data from multiple depositions. Prior to such analysis, multiple assays can be specified using the checkbox on the page provided with the BioActivity Summary tool. The results of the BioActivity Summary analysis are saved on a temporary server, and will be available only for a limited period of time, usually 48 hours. The status of this analysis, however, can be saved through the 'Save View' feature to facilitate scientific communication. Analysis can be resumed by importing the status file through the web server at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=qfile under the common gateway of PubChem BioActivity Analysis Service.

Overall, the BioActivity Summary service aims to provide insights into the activity profile of the compounds using multiple screening test results, and offer an efficient platform to define and collect an interesting set of compounds and panel of assays to perform further analysis.

STRUCTURE–ACTIVITY ANALYSIS TOOL

The Structure–Activity Analysis tool (Figure 4) allows one to perform exploratory analysis by simultaneously clustering compound and assay information using a single linkage clustering methodology (14). The Structure–Activity Analysis Tool is designed to help rapidly identify interesting subsets of compounds and bioassays using various similarity concepts. It enables users to compare and contrast screening results by bioactivity profile or assay target

similarity, or analyze the activity of compound analog series in a panel of assays to identify SAR if any, and to suggest structure features critical for improving biological activity potency.

The results are presented through the use of an interactive heatmap display. With this web-based service, a group of compounds and assays may be clustered using various means. Chemical structures may be clustered based on 2D structure similarity (as measured by Tanimoto score using the PubChem dictionary-based fingerprint [ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt]) or biological response, as measured by reported bioactivity outcome, bioactivity score, or active concentration (e.g. IC50, EC50, AC50, etc.). Assays can be clustered based on similarity in biological response of the compound set, similarity in sequence of assay targets, or similarity provided by depositors. Some of the similarity data (e.g. protein target similarity) are pre-computed to obtain optimal analysis performance.

'Revise Selection' features, similarly to those in the BioActivity Summary tool, are provided to define and modify the focus of the analysis. These features can be accessed by clicking on the '+' sign shown on the left of the Revise Selection section (Figure 4). Furthermore, facilities for further analyzing the clustering results and navigating between various PubChem tools, including the Entrez search system, are provided and may be accessed throughout the heatmap display. For example, as shown in Figure 4, in the dendrogram display, one may click on a blue circle attached to a node of a compound cluster to invoke a feature menu, and to display the sub-tree, prune the sub-tree, or add compounds similar to those contained in a sub-tree. Users can also retrieve chemical structure similarity score matrix used in the clustering and send the selected compounds to one of PubChem tools or to Entrez system. Using similar functionalities associated with assay cluster, one may retrieve assay target similarity score matrix, or revise assay selection to include the various types of related bioassays. Users can also perform a number of operations on a combined group of assays and compounds. To define such a subset, one can zoom in the heatmap display by clicking on two cells in the heatmap. The operations available include sending the compounds in the sub-cluster to the 'Structure Clustering' service to visualize the chemical structure classes, sending assays, or compounds to the Entrez system to, for example, check the availability of protein 3D structure complex or look for information on biological mechanism using linked PubMed articles, etc. Using this feature, one can further compare multiple test results in details by retrieving all readouts using the Data Table tool (to be described).

A common entry point for the Structure–Activity Analysis service is from the previously described BioActivity Summary service, which can be used to narrow down the compound and assay set, thus to prepare an appropriate input for the SAR study. This tool can also be accessed from the BioAssay Summary service for a given bioassay record to analyze the set of active compounds in a single assay through identification of chemical structure clusters that exhibit similar biological response. In this particular application, one may want to take advantage of the integrated tools to further expand the assay scope, for example, to combine the screening results for related targets using the 'Add Related BioAssays' functionality provided in the Revise Selection section, and attempt to search compounds demonstrating high selectivity towards a particular target. This service can be also accessed through the common gateway of the PubChem BioActivity Analysis Service at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=heat, where assays, substances or compounds can be flexibly specified based on one's research need.

Results from such structure–activity analysis, including image, data table and similarity matrix can be exported in respective formats including the Graph Modelling Language format for the dendrograms. Similarly to the BioActivity Summary service, the results of this analysis are saved temporarily, and can be accessed for only a limited period of time, usually 48 hours. Users can use the 'Save View' button to save the analysis in a status file and use this file to communicate the results with collaborators, who can open the status file at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=qfile to see the exact same analysis page at a later time.

An example to illustrate the usage for this Structure–Activity Analysis Tool is shown in Figure 4, which demonstrates how this tool allows chemists and biologists to easily access and combine related screening data for identification of potential chemical probes with desired SAR property and target selectivity. To start the analysis, compounds active in AID 523 and a few confirmatory assays are selected from the BioActivity Summary table as shown in Figure 3. Each of the selected assays has a specified protein target. In this particular analysis, compounds are clustered based on 2D structure similarity and assays are clustered based on target sequence similarity. Each cell is colored according to the reported potency of the compound in the corresponding assay. The heatmap presentation allows one to identify instantly a cluster of compounds that demonstrate interesting SAR. Meanwhile, this analysis allows one to examine the selectivity and target specificity of the compounds by comparing the biological responses against a series of related targets.

DATA TABLE TOOL

The Data Table tool supports rapid search and retrieval of test results for a single or multiple assay records. It is integrated with the BioActivity Summary and Structure–Activity Analysis services. It also links to the Test Result Select tool, Histogram tool and Scatter Plot tool (to be described) when analyzing results for a single assay record. Assay data retrieval, particularly for multiple bioassay depositions simultaneously, can be time consuming. This service implements a queuing and caching mechanism to optimize the web server performance.

Test results are shown in tabular format (Figure 5). Each row in the data table displays a chemical structure, SID, CID, bioactivity outcome, score and associated assay

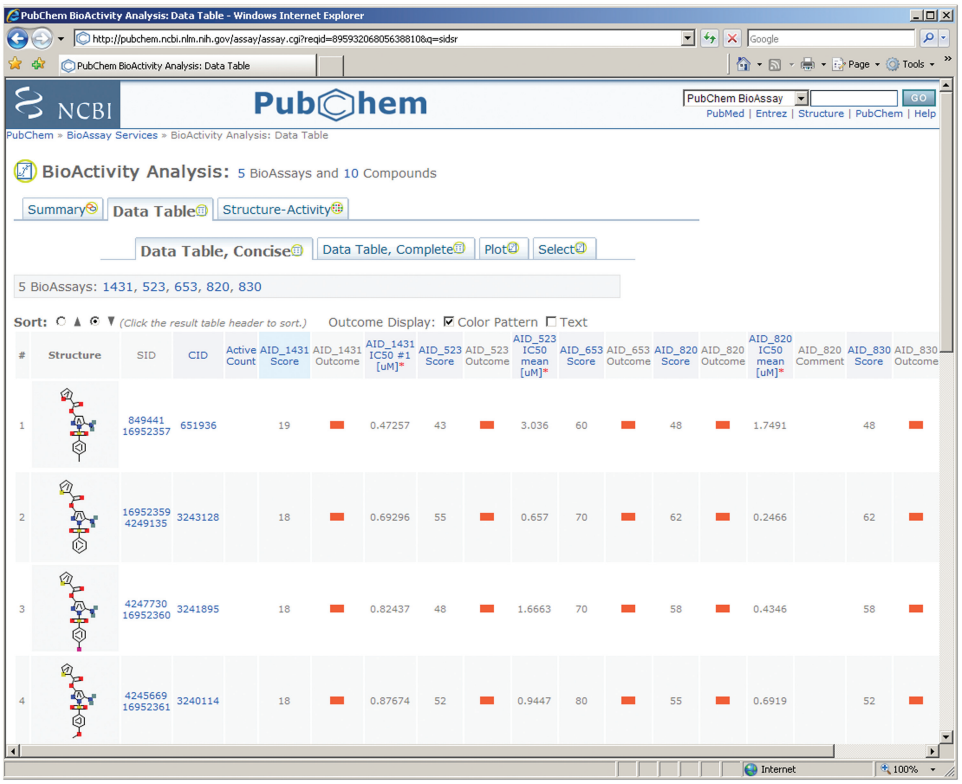

Figure 5. Concise view of data table for multiple screening test results. The compound centric view merges test results from multiple bioassay depositions, and as a result, one CID may point to multiple SID. Bioactivity outcomes are presented using graphical icons, with 'active' results represented with icon colored in red. Complete assay results can be obtained through the 'Data Table, Complete' link. One may access additional data analysis tools using the provided links. One may also save the results or analysis status using the features provided in this page (not shown).

readouts. Bioactivity outcome is highlighted using icon (rectangle shape), with 'active' outcome colored in red and 'inactive' outcome colored in blue. Each CID is linked to the respective PubChem compound summary server, while SID(s), which can be multiple with the merged data presentation from the 'compound view' (see description below), is linked to the substance 'DocSum' report in Entrez. Cross references for a specific test result, such as the GenBank accession number for the protein target of a compound, or PubMed ID of an article where the activity information is extracted, can be reported along with assay readouts in the data table. In such case, they are hyperlinked to the respective Entrez records.

'Concise Data' view, including bioactivity outcome, score and active concentration if provided, is shown in the data table as the default for rapid review, while complete results including all provided readouts can be obtained upon request using the 'Data Table, Complete' link. Data table can be sorted flexibly based on SID/CID, bioactivity score, or a user-selected readout. Data table can be paginated vertically for navigating the test results for different chemical samples, and can be paginated horizontally in the case of multiple assay data retrieval, for navigating the test results from different bioassay depositions. These features are necessary to support the analysis of screening results within PubChem, which sometimes contain hundreds of thousands of chemical samples, and a few hundreds of readout fields.

Assay results can be downloaded in multiple formats including CSV, XML and ASN.1. Chemical structure data can be downloaded as SD files, as well as in many other data formats, such as SMILES, InCHI, XML and ASN.1. Given the large data volume, when downloading test results simultaneously for multiple bioassay records, only the data fields explicitly shown on the current data table can be downloaded at each time. The PubChem BioAssay system currently supports the retrieval and display of readout fields for up to 20 assay records at one time. If users want to request the next 20 bioassay records, they can use the pagination functionality to retrieve and download additional test results. No such limitation is set for data rows, e.g. this service allows downloading all shown test result fields for the entire set of chemical samples.

Data table can be presented in various ways with multiple assay results depending on the specified view focus. Each test result in PubChem BioAssay is associated with a substance sample, represented as a SID unique to the assay depositor. For chemicals with identical structures, test results provided by different depositors are associated with distinct SIDs (with results from NIH MLP program as an exception). This happens sometimes even within the data from the same depositor when samples of such

chemicals are provided in different batches and as such submitted to PubChem as different substance records. As a result, the 'substance view', which can be selected using the 'Group Results by: substance' feature under the 'Result Display Option' section provided on the page, would present these data points on separate rows in the data table display, with each one associated with the respective SID. The 'compound view', however, would merge and associate the test results with the respective unique chemical structure, represented as a CID, and as a result, data points from multiple tests would be presented on the same row. Furthermore, compound libraries supplied by different data source may not encode structures identically. Similar functionalities are developed to merge or collapse test results based on chemical structure associations, such as based on the 'same parent' association among molecules with different salt forms, or based on the 'same connectivity' association for molecules having the same connectivity but with mixed stereo information. Conceivably, such functionality may help in general to obtain comprehensive activity profile for closely related chemical structures.

TEST RESULT SELECT TOOL

The Test Result Select tool can be accessed at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=t&aid=, for each PubChem Bioassay deposition by providing a valid accession number, e.g. AID. It is also linked from the data table page. This service allows one to search and select assay readouts, filter chemical substances/compounds by bioactivity outcome, bioactivity score or active concentration. It lets one specify readout fields to be displayed in data table. It also allows one to request the most recent or all versions of test results within an assay record. This service utilizes NCBI Entrez Eutils facility, communicates with Entrez history system, and enables one to retrieve test results for the compounds obtained from a specific Entrez search. These features provide means for drilling down and partitioning assay test results, which facilitates to rapidly identify chemicals with desired biological property, reduce the data set under investigation to manageable sizes, and make it suitable for subsequent in-depth analysis. Such functionalities also allow one to define 'active' compounds on the fly by setting the desired threshold on screening results from one or multiple assays and derive a focused data set prior to analysis with other bioactivity analysis tools.

HISTOGRAM AND SCATTER PLOT TOOL

Histogram Plot and Scatter Plot tools can be accessed at http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=p&aid=, for each PubChem Bioassay deposition by providing a valid accession number, e.g. AID. They are also linked from the data table page. These tools enable preliminary data analysis using numerical readouts from one assay. Plots are displayed with results of different types of bioactivity outcome labeled with distinct colors and symbols. Interactive functionalities are built in these tools which allow one to identify interesting subset of compounds, and perform subsequent analysis by zooming in the subset or linking to data table. These tools may help PubChem users to evaluate activity threshold, examine distribution of the selected activity readout, and provide an alternative, yet more powerful means for identifying biological interesting chemicals and drilling down test results. Further development for these tools will enable them to support analysis across multiple screening results.

BLAST SEARCH TOOL

A specialized BLAST service is developed at NCBI by integrating the molecular target information in PubChem BioAssay database with BLAST (15) search tool. This service enables one to search assay targets and link to the biological test results contained in PubChem. To this end, protein targets in PubChem BioAssay database are composed as a BLAST search database. This service can be accessed using the link, 'Search protein or nucleotide targets in PubChem BioAssay', on the BLAST home page (http://blast.ncbi.nlm.nih.gov/Blast.cgi) under the 'Specialized BLAST' section. The BLAST service for protein bioassay target search takes a protein sequence submission, by a GenBank accession, GI number, an amino acid sequence in FASTA format, as well as a UniProt accession, and performs a sequence similarity comparison against the protein sequence identity group of each of the available protein targets within the PubChem BioAssay database. PubChem annotates screening results with molecular targets and provides cross links to GenBank protein records whenever possible. This new BLAST service utilizes such annotations, facilitates the search of biological targets and their small molecule ligands within PubChem, and provides an additional path for biologists to discover the screening results and chemical probes contained in PubChem. BLAST server now further annotates the BLAST search results, it highlights the proteins for which biological screening data are available at PubChem and directs users to the respective bioassay records for each particular target (Figure 6). PubChem allows the deposition of bioassay with nucleotide target as well. PubChem currently contains only a few data entries with RNAi screening test results, but anticipates more screening data targeting on genes by collaborating with organizations involved with RNAi screening research including the RNAi Global Initiative Consortium.

CONCLUSION

PubChem provides a set of web servers for accessing, retrieving and analyzing biological test results archived in PubChem databases. BioAssay database grows rapidly with the contribution of the on-going NIH Molecular Library Program and many organizations supporting open access. PubChem BioAssay system has made continuously progress by expanding and optimizing data retrieval and analysis utilities. These PubChem BioAssay

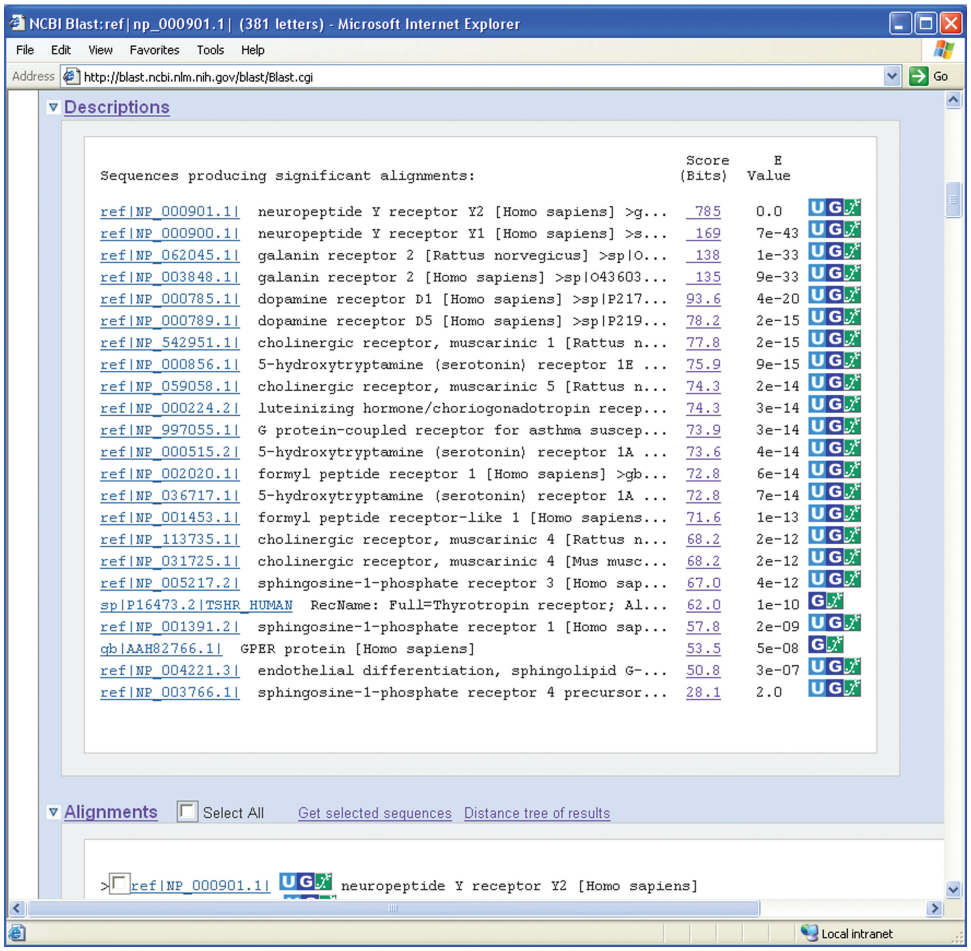

Figure 6. A screen shot of BLAST search results against PubChem BioAssay target database. This report highlights a BLAST hit if biological screening data is available in PubChem for that protein. Each of such BLAST hits is linked to the respective PubChem BioAssay records through a 'dose–response curve' icon.

utilities will be further enhanced towards two directions: develop additional methods to support data analysis and scale up the system for efficient management of the growth of information content and diversity.

ACKNOWLEDGEMENTS

The authors thank Evan Bolton and other colleagues at NCBI for carefully reading the manuscript.

FUNDING

Intramural Research program (National Institutes of Health). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, 217–241. Chapter 12.
2. Zerhouni,E. (2003) Medicine. The NIH roadmap. *Science*, **302**, 63–72.
3. Zerhouni,E.A. (2006) Clinical research at a crossroads: the NIH roadmap. *J. Investig. Med.*, **54**, 171–173.
4. Austin,C.P., Brady,L.S., Insel,T.R. and Collins,F.S. (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.
5. Lazo,J.S., Brady,L.S. and Dingledine,R. (2007) Building a pharmacological lexicon: small molecule discovery in academia. *Mol. Pharmacol.*, **72**, 1–7.
6. Driscoll,J.S. (1984) The preclinical new drug research program of the National Cancer Institute. *Cancer Treat. Rep.*, **68**, 63–76.
7. Zaharevitz,D.W., Holbeck,S.L., Bowerman,C. and Svetlik,P.A. (2002) COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition. *J. Mol. Graph. Model.*, **20**, 297–303.
8. Richard,A.M., Gold,L.S. and Nicklaus,M.C. (2006) Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Curr. Opin. Drug Discov. Devel.*, **9**, 314–325.
9. 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.
10. Harmar,A.J., Hills,R.A., Rosser,E.M., Jones,M., Buneman,O.P., Dunbar,D.R., Greenhill,S.D., Hale,V.A., Sharman,J.L., Bonner,T.I. *et al.* (2009) IUPHAR-DB: the IUPHAR database of G protein-coupled receptors and ion channels. *Nucleic Acids Res.*, **37**, D680–D685.
11. 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.

12. 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.
13. Okuno,Y., Tamon,A., Yabuuchi,H., Niijima,S., Minowa,Y., Tonomura,K., Kunimoto,R. and Feng,C. (2008) GLIDA: GPCR—ligand database for chemical genomics drug discovery – database and tools update. *Nucleic Acids Res*., **36**, D907–D912.
14. Numata,K., Bannai,H., Tamada,Y., de Hoon,M., Imoto,S. and Miyano,S. (2005) Memory-efficient clustering algorithms for microarray gene expression data. *Genome Informatics 2005 Poster and Software Demonstrations*, P049.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.