Published online 19 October 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D1045–D1053***
doi: 10.1093/nar/gkv1072

BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology

Michael K. Gilson*, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang and Jenny Chong

Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0736, USA

Received August 20, 2015; Revised October 02, 2015; Accepted October 05, 2015

ABSTRACT

BindingDB, www.bindingdb.org, is a publicly accessible database of experimental protein-small molecule interaction data. Its collection of over a million data entries derives primarily from scientific articles and, increasingly, US patents. BindingDB provides many ways to browse and search for data of interest, including an advanced search tool, which can cross searches of multiple query types, including text, chemical structure, protein sequence and numerical affinities. The PDB and PubMed provide links to data in BindingDB, and vice versa; and BindingDB provides links to pathway information, the ZINC catalog of available compounds, and other resources. The BindingDB website offers specialized tools that take advantage of its large data collection, including ones to generate hypotheses for the protein targets bound by a bioactive compound, and for the compounds bound by a new protein of known sequence; and virtual compound screening by maximal chemical similarity, binary kernel discrimination, and support vector machine methods. Specialized data sets are also available, such as binding data for hundreds of congeneric series of ligands, drawn from BindingDB and organized for use in validating drug design methods. BindingDB offers several forms of programmatic access, and comes with extensive background material and documentation. Here, we provide the first update of BindingDB since 2007, focusing on new and unique features and highlighting directions of importance to the field as a whole.

INTRODUCTION

Launched on the web in 2000, BindingDB (www.bindingdb.org) is the first publicly accessible database of measured protein–ligand affinity data (1,2). It is designed to support access to focused data sets, such as the affinity data associated with a particular drug target, as well as expansive analyses taking advantage of the comprehensiveness of a large and growing data set. For example, a medicinal chemist might be satisfied with a compact set of affinity data for a specific protein, in order to help with the design of new or improved drugs targeting this protein; a computational chemist might be interested in machine-readable compound and affinity data across a set of related targets, to test and parameterize algorithms for computer-aided drug design; and a pharmacologist might be interested in affinity data across as many human proteins as possible, to help screen a drug candidate for side effects or develop hypotheses for the mechanism of action of a new bioactive compound. When last described in the NAR Database Issue in 2007 (2), BindingDB provided about 20 000 measured binding affinities. Today, BindingDB supports these and other applications with over a million binding data, for nearly 500 000 small molecules and thousands of proteins, coupled with a range of new tools for finding, viewing, downloading and integrating the information in this vast data set.

The BindingDB website is heavily and consistently used, at about 9000 user sessions per month, with an average of six pages viewed per session. This represents a strong increase from 2007, when there were about 2500 sessions per month, with an average of four pages per session. Usage is worldwide, as about one quarter of hits come from the United States, one-eighth each come from China and India, and the remaining half come from European Union member states, Brazil, Japan, South Korea, Canada, Taiwan, Russia and other countries. In addition to these web access statistics, the fact that BindingDB data are downloaded thousands of times a year points to substantial off line usage, not measurable by web hits.

Here, we provide an update of BindingDB in 2015, focusing on its unique data sets and features, and its dramatically expanded data holdings and functionalities. The pa-

*To whom correspondence should be addressed. Tel: +1 858 822 0622; Fax: +1 858 822 7726; Email: mgilson@ucsd.edu

per is organized as follows: Section 3 describes the data, how they are collected, and how they are linked with related data in other bioinformatics resources; Section 4 discusses connections between BindingDB and other databases; Section 5 discusses the range of methods currently available to browse, query and access the data; Section 6 presents tools and data sets specialized for specific user communities; Section 7 summarizes programmatic methods of accessing BindingDB, such as its RESTful API; Section 8 describes didactic materials, introductory tutorials and documentation to help users understand the data and use the features of BindingDB; and Section 9 touches on key directions for future work to secure and enhance access to protein–ligand affinity data.

DATA COLLECTION

Continuous data collection efforts since BindingDB was launched in 2000 have led to steady growth of the database. BindingDB's data holdings currently stand at about 1.1 million measured protein-small molecule affinities, involving about 490 000 small molecules and several thousand proteins; further details regarding the composition and origins of this data set may be found elsewhere (3). With a few exceptions, each BindingDB data entry is derived from one publication (e.g. a scientific article or a patent) and contains at least one protein target, at least one compound, at least one associated compound–protein affinity, and information on the source publication. Entries collected directly from the literature by BindingDB staff also include extended information on experimental conditions like temperature, pH and buffer composition. BindingDB also contains a modest collection of measured protein–protein, protein–peptide and host–guest affinities. The remainder of this section describes the sources of the data collection.

BindingDB staff collect many protein–ligand interaction data not available through other public databases. One aspect of this ongoing curation effort is the capture of roughly 35 000 measurements per year from recent US patents (www.bindingdb.org/bind/ByPatent.jsp). Most of these protein–ligand affinity data are absent from the scientific literature, and the BindingDB patent data set, unlike SureChEMBL (www.surechembl.org), extracts quantitative activity data for the patented compounds. In addition, BindingDB curators collect data on an ongoing basis from about a dozen scientific journals not covered by other database efforts. Primarily focused on the areas of chemical biology and biochemistry, these journals often publish early studies of proteins that later emerge as promising drug targets. They include ACS Chemical Biology, ChemBioChem, Chemical Biology and Drug Design, Chemistry and Biology, and Nature Chemical Biology, ACS BioChemistry, Bioorganic Chemistry and the Journal of Biological Chemistry. It is worth noting that BindingDB collects not only the quantitative affinity data from the source materials, but also key experimental conditions, notably the temperature, buffer composition and pH.

BindingDB also harvests selected data from the PubChem (4,5), ChEMBL (6), PDSP Ki (7) and CSAR (8) (www.csardock.org) databases, so that users can access an integrated collection of protein–ligand affinity data through BindingDB's unified interface (see below). It is worth noting, however, the data available from other databases typically are somewhat less detailed; in particular, they rarely include experimental details, like pH and temperature. BindingDB staff use a combination of automated and manual procedures to import only measurements with a well-defined protein target and a quantitative measure of affinity or relative affinity (usually an IC50, Ki or Kd value), and to flesh out entries as needed to conform to BindingDB's relatively stringent specifications. For example, phenotypic assay data in ChEMBL are not imported to BindingDB, since they do not correspond unequivocally to binding affinities for specific proteins. Similarly, only confirmatory assay results are collected from PubChem, as these are actual affinity data derived from proper compound titrations. In contrast, compound screening data derived from large-scale experiments with compounds at only a single concentration are not imported from PubChem, because they do not provide quantitative affinities and are more liable to experimental error than the confirmatory assay data. Importation of the data from PDSP Ki involved substantial automated and manual curation, such as supplementation of each entry with a protein sequence and experimental details from associated publications.

BindingDB invites direct deposition of binding data by experimentalists. The most straightforward approach uses a simple, preformatted Excel file, which may be downloaded from the BindingDB website (www.bindingdb.org/bind/contributedata.jsp). The user fills out the spreadsheet and uploads it via the BindingDB website, and it is automatically parsed for review and entry by BindingDB curators. Alternatively, data may be deposited via the same web forms used by BindingDB curators (www.bindingdb.org/deposition/index.jsp). Either way, direct deposition is attractive because it eliminates transcription steps where errors may creep in, and, if widely adopted, it would crowdsource the time consuming process of data collection. Thus, although only a negligible fraction of existing BindingDB data have come by this route, routine deposition by experimentalists remains a goal.

Several measures enhance the reliability of the data provided by BindingDB. First, data curated from original sources by BindingDB staff are checked, and corrected if needed, by a second staff member. Data imported from other databases, such as PubChem and ChEMBL, are automatically checked for completeness and certain easily detected errors, and any data flagged by these procedures are reviewed manually and corrected if needed. For example, data imported from other databases occasionally list implausibly small pKi values, such as 1×10^{-8}, usually because a curator mistakenly entered a Ki value as a pKi. Finally, automated scripts are used to identify the corresponding author of each source article, where possible, and then email the author a link to his or her data on the BindingDB website, with an invitation to check the data and report any errors. In addition to transcription errors, authors sometimes also report errors in the actual published article. In order to avoid confusion in cases where BindingDB presents corrected data, the BindingDB display marks the data as having been corrected.

Although BindingDB focuses primarily on experimental data, it includes some—clearly marked—results from computational modeling. In particular, SDfiles may be downloaded not only with 2D compound structures, but also with 3D structures computed with the program Vconf (www.verachem.com/products/vconf), which are suitable starting points for further conformational analysis or docking calculations. Additionally, computed protein–ligand poses, generated by automated docking with the program Surflex (9), are provided for protein–ligand complexes for which the crystal structure of a congeneric ligand (i.e. one with a large shared substructure) is available to guide the docking process. The full set of dockings are provided on a single page (www.bindingdb.org/bind/surflex_entry.jsp), and are also presented in search results when available.

INTEGRATION WITH OTHER DATABASES

BindingDB is connected with a number of other databases via either unidirectional (from BindingDB) or bidirectional (to and from BindingDB) hyperlinks. The bidirectional links between BindingDB and the RCSB (10) Protein Data Bank (PDB) (11,12) are probably the most heavily used: the home pages for many crystal structures in the PDB are hyperlinked to the corresponding protein–small molecule affinity data in BindingDB, and, conversely, many protein–ligand affinity data in BindingDB are linked to co-crystal structures in the PDB. It is worth noting, however, that only a fraction of BindingDB's affinity data are associated with crystal structures. This is because, while medicinal chemistry publications frequently list on the order of 20 protein–ligand affinities, only a few corresponding cocrystal structures appear in the PDB. This point clarifies a key difference between BindingDB and the BindingMOAD (13) and PDBBIND (14,15) databases: the latter provide affinities only for protein–ligand interactions associated with cocrystal structures, and hence contain well below 50 000 protein–ligand affinities. BindingDB provides many more affinities because it is not limited to cases where a crystal structure is available.

BindingDB also provides direct links from data sets to published articles and US Patents online, as well as to article entries in PubMed (www.ncbi.nlm.nih.gov/pubmed). Conversely, PubMed users may navigate directly from an article listing to the corresponding data in BindingDB, via the LinkOut/More Resources option below the PubMed abstract. Many other precomputed links allow users to navigate from BindingDB data to related information in other databases. These include:

- Matching proteins in UniProt (16,17), BindingMOAD and DrugBank (18)
- Matching compounds in ChEMBL, PubChem and UniChem (19), as well as in the ZINC (20) database of commercially available compounds
- Antibodies against protein targets, in AntibodyPedia (www.antibodypedia.com)
- Links from the MarinLit (pubs.rsc.org/marinlit) database of marine natural products to data for matching compounds in BindingDB
- Links from BindingDB protein targets to biomolecular pathway pages in Reactome (21–23)

BindingDB also participates in the Protein Standard Interface Common Query Interface (PSICQUIC; www.ebi.ac.uk/Tools/webservices/psicquic/view/main.xhtml) (24), a web service which provides standardized access to multiple molecular interaction databases.

DISCOVERING AND VIEWING DATA IN BindingDB

BindingDB aims not only to collect data, but also to make it available in a user-friendly and flexible manner. The growth of BindingDB's data collection, the potential for new uses of the data, and the emergence of new software technologies, have allowed the creation and deployment of innovative tools for query, browsing, download, analysis, visualization and integration. The present section highlights a set of these tools that are particularly useful or distinctive. First, however, we describe the primary report table, a central data reporting format which is used by multiple browsing and query tools.

Figure 1 displays the table header and a sample row of such a table, for the androgen receptor binding to the steroidal compound DHT. Moving from left to right, the first column identifies the protein target and the institution (e.g. a university or company) where the measurement was made. The second column identifies the ligand and optionally shows the corresponding SMILES (25) or InChI (26) strings. The next three columns respectively, provide links to additional information about the target, the ligand, and the target–ligand combination. For example, there are links to more information about the androgen receptor in the PDB, in several pathway databases, such as Reactome, in DrugBank, and in AntibodyPedia; the ligand occurs in multiple compound databases, including in ZINC with an availability code of '1', meaning that it is purchasable; and more information about this ligand–target pair may be found in the PDB and the source article. The Ligand Links column also provides quick access to BindingDB data for chemically similar compounds via the 'Similars' link.

The gray tiles to the right then provide the quantitative binding data, along with the pH and temperature at which the measurement was made, if available. The Citation and Details link goes to a page with additional information on the measurement, while the color icon links to an interactive 3D view of an available crystal or modeled structure (see below). In some cases, multiple measurements are available for a given protein and ligand; the 'More data for this Ligand-Target Pair' is a link to such data in BindingDB, if it exists. The 'Make Data Set' button above each table of results can be used to collect and download the displayed data in machine-readable formats, either SDfile or a tab-separated value (TSV) file with SMILES strings.

BindingDB provides a wide range of tools to query and browse data. Most of these are available through the menu of options at the lefthand side of the page, where they are categorized as target-oriented, compound-oriented, citation oriented, etc. Due to space limitations, the following subsections highlight only a few of the commonly used and more specialized tools, and users are encouraged to visit the website to explore the additional capabilities provided there.

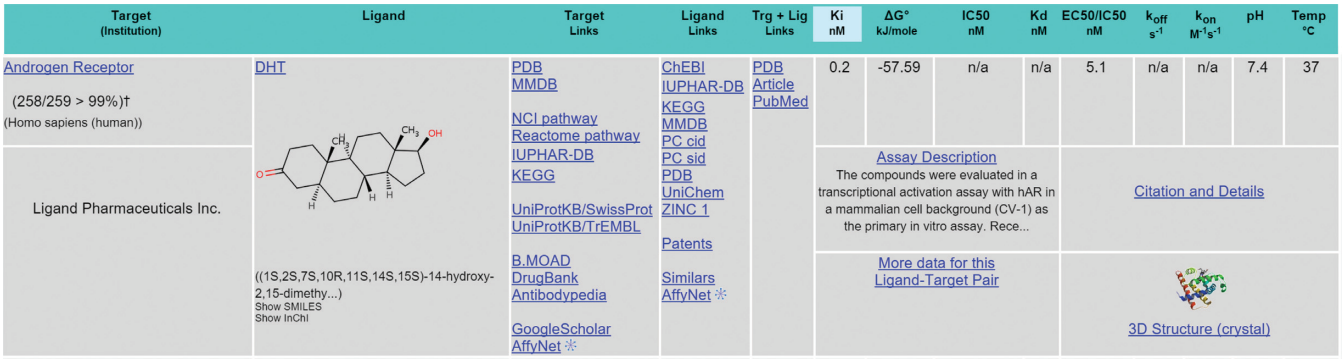

Figure 1. Sample data row in a standard BindingDB report. See text for details.

Query and browsing tools

Protein targets. The Simple Search window near the top of the home page provides an easy way to carry out a broad, text-based search. The report page groups the resulting hits according to whether the text matches occurred in the name of a protein target, a compound, the title of an article or the text of an assay description, and each hit can then be followed to the associated data in BindingDB.

Another easy way in to the data set, for users who are interested in a particular protein target, is to browse by target name (www.bindingdb.org/bind/ByTargetNames.jsp), and then follow the links provided to all information on a given target, data subsetted by measurement type, such as Ki or IC50, listings of the articles from which the data were drawn, or direct download of all the data for a target of interest in the form of a machine-readable SDfile or a TSV file with SMILES strings.

The target-oriented tools also allow, among other things, searching for data on a given target or targets with a user-specified range of Ki or IC50 values; searching for rate constant and calorimetric data; searching for data measured within a given pH range; and searching for enzymatic data with a desired substrate, a desired compound or a compound within a desired molecular weight range. A specialized Pathways path merges information from Reactome pathways with BindingDB data to help users find data for proteins within pathways of interest.

Compound search. BindingDB provides two tools for users interested in quick access to information on compounds of relatively high general interest. One is a page of FDA-approved products (www.bindingdb.org/bind/ByFDAdrugs.jsp), which identifies the active ingredients, the proteins against which they have been tested, links to the corresponding binding data, and prepared downloads with these data in machine-readable SDfile and TSV formats. The second is a tabbed page of 'Important Compounds' in simplified format (www.bindingdb.org/ByLigandName), with separate tabs providing information, drawn from BindingDB, DrugBank, ZINC, PubChem and other sources, for existing drugs, experimental drugs, other bioactive compounds of interest and peptides.

A tool to search for compounds according to chemical structure, entered with a drawing or a SMILES or InChI string (www.bindingdb.org/bind/chemsearch/marvin/index.jsp), allows users to select among exact match, chemical similarity and chemical substructure, while applying additional criteria for compound potency against any target, and whether the compound exists as a ligand in the PDB. Also, one may use check boxes to define a limited set of target proteins of interest. Finally, the 'Find my Compound's Targets' page, detailed below, can be used to query BindingDB with multiple compounds uploaded in the form of an SDfile or a text file with one SMILES string in each line; again, options are provided to search by exact match, similarity and substructure, and with user-specified potency criteria. BindingDB uses current ChemAxon default settings to compute chemical fingerprints, and chemical similarities are then computed with the Tanimoto metric.

Obtaining data by article, patent, author, institution, etc. BindingDB provides the capability, apparently unique among public protein–ligand databases, of direct access to all of the binding data from an article or patent of interest. The article page (www.bindingdb.org/bind/ByJournal.jsp) allows users to select a curated article by journal name, volume and page number, and display it as a BindingDB table or download it as a machine-readable file in SDfile or TSV format; see, e.g. www.bindingdb.org/bind/ByJournal.jsp?journal = J%20Med%20Chem. Similar browsers are available to retrieve data from selected US Patents (www.bindingdb.org/bind/ByPatent.jsp), PubChem BioAssays (www.bindingdb.org/bind/ByPCBioAssay.jsp) and articles listed by PubMed IDs (www.bindingdb.org/bind/ByPubMed.jsp). These tools provide a straightforward way for users to collect the data from sources of interest, in machine-readable formats suitable for importation to a local database or for setting up calculations like QSAR or simulations. For other useful perspectives of the BindingDB data collection, one may browse or search for data from a specific author (http://www.bindingdb.org/bind/ByAuthor.jsp) or from a specific company, university or institute (www.bindingdb.org/bind/ByInstitution.jsp). It is worth noting that a given institution may appear with multiple names, chiefly because of variations in author usage.

BindingDB embeds journal reference data in its web pages, in the form of ContextObjects in Spans (COinS; ocoins.info). This allows users with suitable browser add-

ons, such as the free Zotero (www.zotero.org) reference manager, which integrates with Firefox, to collect all of the journal references associated with the data displayed on a BindingDB page with one or two mouse-clicks.

Biomolecular pathways and protein fold families. Reactome is a manually curated and peer-reviewed pathway database, which serves as a good starting point for researchers seeking to learn more about signaling cascades, metabolic pathways, etc. In order to enrich the BindingDB user experience with this information, the Reactome Pathway Browser (www.reactome.org/PathwayBrowser/) has been repurposed as a point of entry for BindingDB users seeking to obtain binding data for proteins associated with a particular biological pathway of interest (www.bindingdb.org/Pathways/pathways.jsp). A similar browsing tool provides users with a view into BindingDB based on protein fold families (www.bindingdb.org/SCOP/scop.jsp), based on the widely used SCOP classification (27).

Advanced search with multiple queries. In response to user requests, a powerful advanced search capability has been developed and deployed (www.bindingdb.org/as/search.html). This allows users to specify any number of heterogeneous search criteria, and to limit the search to specific sources within the BindingDB data set; i.e. data curated by BindingDB; data from US Patents; and/or data harvested from ChEMBL, PubChem, PDSP Ki or CSAR. Available query criteria, which may be combined with a Boolean 'and' or 'or', include target name, sequence, molecular weight and source organism; compound name, SMILES string, molecular weight, similarity or substructure; binding potency, such as Ki and IC50; and article or patent information, such as publication year and author. For example, the search in Figure 2 combines text in the target name, chemical substructure and potency: tinyurl.com/puvu2ed. The results are provided in the standard table format (above), along with additional reports categorizing the data found according to target, compound, article, patent, author, etc.

Downloading and reusing data

All BindingDB data are available for download and reuse, chiefly in the form of SDfiles and TSV files with SMILES strings. Documentation of both formats is provided at the Info (https://www.bindingdb.org/bind/info.jsp) and Download (https://www.bindingdb.org/bind/chemsearch/marvin/SDFdownload.jsp?all_download=yes) pages, and the SDfiles are available with 2D chemical structures or computed 3D structures. All data curated by BindingDB are provided under a Creative Commons Attribution License which allows reuse, redistribution and republication, so long as BindingDB is cited. All data curated by ChEMBL are provided, as required by ChEMBL, under a more restrictive Creative Commons Attribution-Share Alike License, which, in addition to citation of ChEMBL, also requires that any collection or product based on ChEMBL data be redistributed under the same Share Alike license. BindingDB users are asked to complete a free registration, via SSL encrypted pages, in order to download files. The registrations are important when we seek funding to continue BindingDB, and they enable a brief survey of BindingDB users once every few years to learn what is working, what needs work, and what directions would be most useful.

As noted above, many of BindingDB's query and browsing tools allow the user to download the resulting data for local use and redistribution. In addition, the Download page provides downloadable files of all of the data in BindingDB, as an Oracle data dump or an SDfile or TSV file, as well as of major subsets of BindingDB data. For example, users who already have ChEMBL data may wish to obtain only the data curated by BindingDB, only the patent data curated by BindingDB, etc. Additional available downloads include lists of target protein sequences, mappings of BindingDB entries to other databases, etc.

Customized data collection

Users with a long term interest in one or a few protein targets may use the myBDB feature of BindingDB to add them to a personalized list of proteins, which may then be accessed and viewed in a compact format. In addition, users may receive email notifications when new data for their targets appear. The myBDB feature is linked to one's BindingDB user registration.

SPECIALIZED TOOLS AND DATA SETS

It was recognized early on that an expansive database of protein–ligand interaction data would have emergent capabilities, such as the ability to formulate hypotheses for the protein targets of bioactive compounds, to raise flags regarding possible off-target effects of drugs and drug candidates (28), and to generate suggestions for old compounds to test against new protein targets. In addition, easy availability of a large set of quantitative binding data should provide a valuable basis for the evaluation and parameterization of software for computer-aided ligand design. BindingDB supports such applications with specialized tools and data sets, as described in the following subsections.

Find my compound's target

The experimental observation that a compound is bioactive, such as through phenotypic screening, immediately poses the question of its mechanism of action. BindingDB's large data set can be used to generate hypotheses regarding the target or targets of bioactive compounds, based on a simple transitivity principle: a compound A, which is chemically similar to another compound B, has a good chance of binding the same proteins that B binds. Thus, the targets of compound B may explain the mechanism of compound A. Similarly, if A is a drug, it may generate off-target side effects by binding some of the same targets as compound B.

BindingDB's Find My Compound's Target (FMCT) tool (www.bindingdb.org/bind/chemsearch/marvin/BatchStructures.jsp) is designed to support such analyses. Here, one specifies a single compound of interest as a chemical drawing, SMILES string or InChI string; or multiple compounds of interest by uploading an SDfile or a text file with one SMILES string per line. These are the A compounds. One then adjusts the chemical similarity

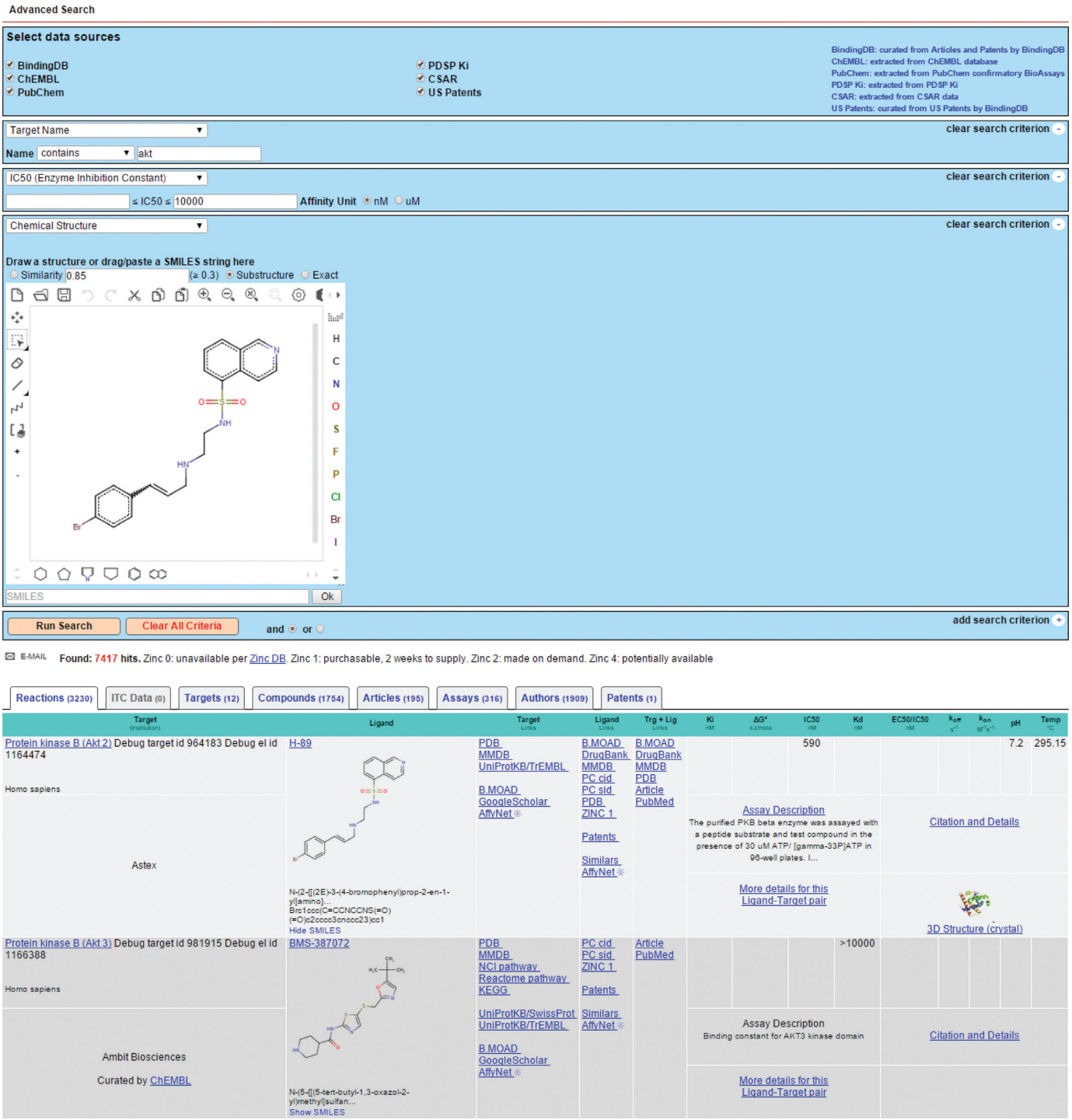

Figure 2. Example of advanced search, based on name of protein, numerical range of IC50 and chemical structure similarity. Results are shown at the bottom of the figure.

criteria to be applied in finding the B compounds, and sets an optional filter for the affinity with which the B compounds must bind a protein target in order for the protein to be reported as a possible target of the A compounds. The results of the query are presented in a table (Figure 3), where each row lists one suggested protein target, the maximum similarity of query compound A to any compound B binding this target, links to the compound(s) B and their associated affinity data, and downloadable files with the compound(s) B and their affinities for the target. Figure 3 shows part of the table of results for a query with the cough suppressant cloperastine, which has been analyzed with the Similarity Ensemble Approach (SEA (29)). Clearly, this compound has high potential to bind a variety of

My Compound's Targets

	Target Name	Uploaded compounds generating hits	Max Similarity	Hits (All Compounds)	Ki Data	IC50 Data	Kd Data	EC50 Data	Add to myBDB	Download
1	Acetylcholine-binding protein (Ls-AchBP)	1	0.75	1	1	0	0	0		2D 3D TSV
2	CHRM5	1	0.70	1	1	0	0	0		2D 3D TSV
3	Dopamine D4 receptor	1	0.70	1	1	0	0	0		2D 3D TSV
4	Dopamine receptors; D3 & D4	1	0.70	1	1	0	0	0		2D 3D TSV
5	Dopamine Transporter (DAT)	1	0.70	28	44	38	0	0		2D 3D TSV
6	Histamine H1 receptor	1	0.70	4	9	1	3	0		2D 3D TSV
7	Muscarinic acetylcholine receptor	1	0.70	1	2	0	0	0		2D 3D TSV
8	Muscarinic acetylcholine receptors; M1 & M2	1	0.70	1	2	0	0	0		2D 3D TSV
9	Muscarinic receptor M1 and M2	1	0.70	3	8	0	0	0		2D 3D TSV
10	Muscarinic receptor M2 and M3	1	0.70	1	2	0	0	0		2D 3D TSV
11	Norepinephrine transporter	1	0.70	1	1	0	0	0		2D 3D TSV
12	Norepinephrine Transporter (NET)	1	0.72	7	7	7	0	0		2D 3D TSV
13	Serotonin Transporter (SERT)	1	0.70	12	15	12	0	0		2D 3D TSV
14	Sigma opioid receptor	1	0.70	2	0	3	0	0		2D 3D TSV

Figure 3. Sample results table from the Find My Compound's Target tool, with cloperastine as the query compound.

membrane-bound receptors and transporters in the central nervous system, and the sigma receptor prediction (last row) is consistent with recent experimental confirmation (30). Further examples and analysis of BindingDB's FMCT tools are available in a separate publication (31).

Find compounds for my targets

Users looking for candidate compounds to bind a new protein target of interest may take advantage of the 'Find Compounds for My Targets' (FCMT: www.bindingdb.org/as/targetsearch.html) tool, which is based on a mirror image of the reasoning used to find targets for compounds, above. Thus, if a protein A is similar in amino acid sequence to a protein B, it is likely to bind some of the same compounds as protein B. Here, one pastes one or more protein sequences into text windows, specifies a sequence identity cutoff, and then searches for compounds which bind any protein in BindingDB whose sequence identity with the query protein(s) satisfies the cutoff. The results may be sorted by affinity, the number of similar targets found or other parameters, and may be downloaded in the form of an Excel (xls) file containing the SMILES strings and binding information for the compound hits.

Virtual compound screening

BindingDB provides web-based tools that allow users to upload a set of active compounds, or to pick a set of BindingDB compounds that bind a protein target of interest, and then use these actives as a basis for computationally screening a compound catalog for other compounds likely to have the same activity. The available methods for this ligand-based screening service are Maximum Similarity, Binary Kernel Discrimination (32) and a Support Vector Machine (33) approach.

Validation data sets for computer-aided drug design

The BindingDB data collection can be used to train and evaluate software for computer-aided ligand design. This application is specifically supported by a page with 700 focused, downloadable data sets drawn from the larger BindingDB data collection (http://www.bindingdb.org/validation_sets/index.jsp). Each data set comprises 10–50 congeneric compounds spanning a range of affinities for a single protein target. For each data set, the PDB holds a protein–ligand cocrystal structure of at least one of the compounds in the series. This index structure provides a basis for modeling the rest of the compounds in the series, and, as noted above, docked structures are provided for many of them, in order to provide users with an initial look at a plausible range of ligand poses. The publications which yielded the data in each data set are also provided at the validation set web page, and it is worth emphasizing that most data sets are derived from only 1–3 articles published by a single institution, so the affinities in each data set are usually traceable to a single experimental assay procedure. This uniformity is beneficial, as it avoids concerns about comparing affinities measured by different assays at different institutions. Each data set also has a place for users to leave public comments, such as if a particular data set is found to be particularly useful or problematic.

PROGRAMMATIC ACCESS

The BindingDB server provides two different types of programmatic access. One is a RESTful API (https://www.bindingdb.org/bind/BindingDBRESTfulAPI.jsp),

which provides two types of access to binding data, returned in XML format. In one, the user provides the UniProt ID of a protein of interest, and BindingDB returns a list of BindingDB compounds (BindingDB Monomer IDs and SMILES strings) and their affinities for the query protein. In the other, the user provides a SMILES string and a chemical similarity cutoff, and BindingDB returns a list of similar compounds, along with their protein targets and the corresponding binding affinities.

The BindingDB website also supports various queries with structured URLs (https://www.bindingdb.org/bind/SearchTemplates.jsp), which return a BindingDB web page with the search results. For example, the following URL returns the binding data associated with PDB entry 1OKZ: www.bindingdb.org/jsp/dbsearch/PrimarySearch_pdbids.jsp?pdbids_submit = Search&pdbids = 1OKZ; note that the full list of PDB entries represented in BindingDB is also available: https://www.bindingdb.org/bind/BindingDB_PDB.txt. Similarly, one may recover all of the data associated with a given article (i.e. PubMed ID), and there is a corresponding list of PubMed IDs for which BindingDB holds data: www.bindingdb.org/bind/BindingDB_PubMed.txt. Structured URLs area also available to query BindingDB by protein target name, protein sequence (via BLAST (34)), BindingDB compound monomerID and SMILES string.

Finally, BindingDB has been successfully linked to KNIME (35), a graphical workbench for data analysis, analytics, visualization and reporting (www.knime.org/). KNIME allows data analyses to be set up by interactively linking functional 'nodes' into custom pipelines which can then be used and shared. Over 1000 nodes are currently available, including many for chemical computing and analysis, so connecting BindingDB with this analysis environment holds great potential to facilitate the flexible use and reuse of binding data in a variety of applications. As detailed elsewhere (31), KNIME workflows have been constructed that query BindingDB in two modes. One mode is to access the RESTful API; the second is to access with a local, downloaded table containing essentially all BindingDB data. The latter approach is particularly useful for large scale queries or work where confidentiality is of high concern. Ten KNIME workflows and their associated documentation, including a full KNIME implementation of the Find My Compound's Target tool (Section 6.1), are available for free download and reuse (https://www.bindingdb.org/bind/chemsearch/marvin/SDFdownload.jsp?all_download = yes).

DOCUMENTATION AND INSTRUCTIONAL RESOURCES

The BindingDB website provides a range of documentation, chiefly through the Info page, https://www.bindingdb.org/bind/info.jsp, including:

- Didactic background regarding the science of protein–ligand binding and how affinities are measured, suitable for suitable for students and interested scientists from other fields: https://www.bindingdb.org/bind/BindingDB-Intro2a.pdf
- A 'how to' page with explanations of common searches and answers to frequently asked questions: www.bindingdb.org/bind/FAQ2.jsp
- A glossary of BindingDB terminology: www.bindingdb.org/bind/glossary.jsp
- Documentation of BindingDB's structured URLs (Section 7) and download file formats
- Link to the BindingDB on Wikipedia: en.wikipedia.org/wiki/BindingDB
- Four video tutorials regarding commonly used BindingDB tools, accessible via links on the BindingDB main page.

DIRECTIONS

Continuing the flow of new data into BindingDB and other protein–ligand interaction databases is both fundamental and non-trivial. Great strides have been made through the manual extraction of data from scientific articles and patents, notably by ChEMBL and BindingDB. However, ensuring the stability of these efforts may require a broader recognition, by the scientific community and funding agencies, that the collection of protein–ligand interaction data is a priority on par with collection of other types of data, such as protein structures and gene sequences, due to its critical value in the translation of basic science to therapeutics. It is also worth noting that the need for manual curation would be greatly reduced if journals were to establish a regular flow of the data being published in their articles into a common data collection portal, from which the information could be transferred into the relevant databases. Such data flows may be facilitated by the growing use of electronic lab notebooks and other research data systems. Ultimately, authors should be able to move the data they are publishing into both journals and databases with a few keystrokes.

Today, the availability of large, publicly accessible databases of protein–ligand interaction data, such as BindingDB, is actively transforming the landscape of molecular informatics (3), enabling new linkages among diverse fields like medicinal chemistry, drug discovery, systems biology and genomics. Indeed, there are still exciting opportunities for new tools to take advantage of these data, through seamless integration of data resources and computational tools both on- and off-line. For example, the growing field of systems pharmacology (36–38) would benefit from advanced tools to markup pathways with compounds known to modulate the proteins involved, and to smoothly access target–compound data in support of ligand-design and the prediction of side effects. In addition, there is potential for tighter integration with journals, by linking articles to associated binding and structural data in machine-readable format, and by using these data to bring electronic articles to life with structural views, structure-activity spreadsheets and charts, and facile integration with data for similar compounds and related proteins. The continued expansion of the data collections, and of capabilities in computing, networking, visualization and informatics, point only to growth in the power and utility of BindingDB and its related resources.

ACKNOWLEDGEMENT

We thank the National Institutes of Health (NIH) for supporting this project via grant R01GM070064. These findings are solely of the authors and do not necessarily represent the views of the NIH.

FUNDING

National Institutes of Health (NIH) [R01GM070064]. Funding for open access charge: National Institutes of Health (NIH) [R01GM070064].

Conflict of interest statement. M.K.G. has an equity interest in, and is a cofounder and scientific advisor of VeraChem LLC.

REFERENCES

1. Chen,X., Liu,M. and Gilson,M.K. (2002) BindingDB: A Web-accessible molecular recognition database. *Comb. Chem. High Throughput Screen*, **4**, 719–725.
2. 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 Acid Res.*, **35**, D198–D201.
3. Nicola,G., Liu,T. and Gilson,M.K. (2012) Public domain databases for medicinal chemistry. *J. Med. Chem.*, **55**, 6987–7002.
4. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
5. Wang,Y., Suzek,T., Zhang,J., Wang,J., He,S., Cheng,T., Shoemaker,B.A., Gindulyte,A. and Bryant,S.H. (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**, D1075–D1082.
6. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
7. Roth,B.L., Kroeze,W.K., Patel,S. and Lopez,E. (2000) The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrasment of riches? *Neuroscientist*, **6**, 252–262.
8. Carlson,H.A. and Dunbar,J.B. (2011) A Call to Arms: What You Can Do for Computational Drug Discovery. *J. Chem. Inf. Model.*, **51**, 2025–2026.
9. Jain,A.N. (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, **46**, 499–511.
10. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucl Acids Res.*, **28**, 235–242.
11. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
12. Rose,P.W., Prlić,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
13. Ahmed,A., Smith,R.D., Clark,J.J., Dunbar,J.B. and Carlson,H.A. (2015) Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Res.*, **43**, D465–D469.
14. Wang,R., Fang,X., Lu,Y., Yang,C.-Y. and Wang,W. (2005) The PDBBind database: Methodologies and updates. *J. Med. Chem.*, **48**, 4111–4119.
15. Liu,Z., Li,Y., Han,L., Li,J., Liu,J., Zhao,Z., Nie,W., Liu,Y. and Wang,R. (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.
16. Pundir,S., Magrane,M., Martin,M.J., O'Donovan,C. and The UniProt Consortium. (2015) Searching and Navigating UniProt Databases: Searching and Navigating UniProt Databases. In: Bateman,A, Pearson,WR, Stein,LD, Stormo,GD and Yates,JR (eds). *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc, Hoboken, NJ, pp. 1.27.1–1.27.10.
17. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
18. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
19. Chambers,J., Davies,M., Gaulton,A., Hersey,A., Velankar,S., Petryszak,R., Hastings,J., Bellis,L., McGlinchey,S. and Overington,J.P. (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminformatics*, **5**, 3.
20. Irwin,J.J., Sterling,T., Mysinger,M.M., Bolstad,E.S. and Coleman,R.G. (2012) ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.*, **52**, 1757–1768.
21. Haw,R.A., Croft,D., Yung,C.K., Ndegwa,N., D'Eustachio,P., Hermjakob,H. and Stein,L.D. (2011) The Reactome BioMart. *Databas*, **2011**, bar031.
22. Milacic,M., Haw,R., Rothfels,K., Wu,G., Croft,D., Hermjakob,H., D'Eustachio,P. and Stein,L. (2012) Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome. *Cancers*, **4**, 1180–1211.
23. Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G., Caudy,M., Garapati,P., Gillespie,M., Kamdar,M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
24. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S.L., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.
25. Weininger,D. (1988) SMILES, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.*, **28**, 31–36.
26. Heller,S.R., McNaught,A., Pletnev,I., Stein,S. and Tchekhovskoi,D. (2015) InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics*, **7**, 23.
27. Andreeva,A., Howorth,D., Chandonia,J.-M., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2007) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
28. Gilson,M.K. (2001) Editorial: Molecular recognition databases. *Biopolymers*, **61**, 97–98.
29. Keiser,M.J., Roth,B.L., Armbruster,B.N., Ernsberger,P., Irwin,J.J. and Shoichet,B.K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
30. Gregori-Puigjané,E., Setola,V., Hert,J., Crews,B.A., Irwin,J.J., Lounkine,E., Marnett,L., Roth,B.L. and Shoichet,B.K. (2012) Identifying mechanism-of-action targets for drugs and probes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 11178–11183.
31. Nicola,G., Berthold,M.R., Hedrick,M.P. and Gilson,M.K. (2015) Connecting proteins with drug-like compounds: Open source drug discovery workflows with BindingDB and KNIME. *Database*, **2015**, bav087.
32. Harper,G., Bradshaw,J., Gittins,J.C., Green,D.V. and Leach,A.R.J. (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.*, **41**, 1295–1300.
33. Jorissen,R. and Gilson,M. (2005) Virtual screening of molecular databases using a Support Vector Machine. *J. Chem. Inf. Mod.*, **45**, 549–561.
34. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **214**, 1–8.
35. Berthold,M.R., Cebron,N., Dill,F., Gabriel,T.R., Kötter,T., Meinl,T., Ohl,P., Sieb,C., Thiel,K. and Wiswedel,B. (2007) KNIME: The Konstanz Information Miner. In: Preisach,C, Burkhardt,H, Schmidt-Thieme,L and Decker,R (eds). *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, Berlin; Heidelberg, 319–326.
36. Arrell,D.K. and Terzic,A. (2010) Network Systems Biology for Drug Discovery. *Clin. Pharmacol. Ther.*, **88**, 120–125.
37. van der Greef,J. and McBurney,R.N. (2005) Rescuing drug discovery: in vivo systems pathology and systems pharmacology. *Nat. Rev. Drug Discov.*, **4**, 961–967.
38. Berger,S.I. and Iyengar,R. (2009) Network analyses in systems pharmacology. *Bioinformatics*, **25**, 2466–2472.