DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bat044, doi:10.1093/database/bat044

Original article

The CARLSBAD Database: A Confederated Database of Chemical Bioactivities

Stephen L. Mathias, Jarrett Hines-Kay, Jeremy J. Yang, Gergely Zahoransky-Kohalmi, Cristian G. Bologa, Oleg Ursu and Tudor I. Oprea*

Department of Internal Medicine, Translational Informatics Division, University of New Mexico School of Medicine, 1 University of New Mexico, MSC09 5025, Albuquerque, NM 87131, USA

***Corresponding author:** Tel: +(505) 925 4756; Fax: +(505) 925 7625; Email: toprea@salud.unm.edu

Submitted 8 January 2013; Revised 12 May 2013; Accepted 21 May 2013

Citation details: Mathias,S.L., Hines-Kay,J., Yang,J.J. *et al*. The CARLSBAD Database: A Confederated Database of Chemical Bioactivities. *Database* (2013) Vol. 2013: article ID bat044; doi:10.1093/database/bat044

Many bioactivity databases offer information regarding the biological activity of small molecules on protein targets. Information in these databases is often hard to resolve with certainty because of subsetting different data in a variety of formats; use of different bioactivity metrics; use of different identifiers for chemicals and proteins; and having to access different query interfaces, respectively. Given the multitude of data sources, interfaces and standards, it is challenging to gather relevant facts and make appropriate connections and decisions regarding chemical–protein associations. The CARLSBAD database has been developed as an integrated resource, focused on high-quality subsets from several bioactivity databases, which are aggregated and presented in a uniform manner, suitable for the study of the relationships between small molecules and targets. In contrast to data collection resources, CARLSBAD provides a single normalized activity value of a given type for each unique chemical–protein target pair. Two types of scaffold perception methods have been implemented and are available for datamining: HierS (hierarchical scaffolds) and MCES (maximum common edge subgraph). The 2012 release of CARLSBAD contains 439 985 unique chemical structures, mapped onto 1,420 889 unique bioactivities, and annotated with 277 140 HierS scaffolds and 54 135 MCES chemical patterns, respectively. Of the 890 323 unique structure–target pairs curated in CARLSBAD, 13.95% are aggregated from multiple structure–target values: 94 975 are aggregated from two bioactivities, 14 544 from three, 7 930 from four and 2214 have five bioactivities, respectively. CARLSBAD captures bioactivities and tags for 1435 unique chemical structures of active pharmaceutical ingredients (i.e. 'drugs'). CARLSBAD processing resulted in a net 17.3% data reduction for chemicals, 34.3% reduction for bioactivities, 23% reduction for HierS and 25% reduction for MCES, respectively. The CARLSBAD database supports a knowledge mining system that provides non-specialists with novel integrative ways of exploring chemical biology space to facilitate knowledge mining in drug discovery and repurposing.

Database URL: http://carlsbad.health.unm.edu/carlsbad/.

Introduction

As the number of chemicals and screening efforts multiply, the number of bioactivity databases offering information on biological activity of small molecules is increasing. They represent a rich source of information in our quest to map the chemical space of bioactive molecules to phenotypic and target space. We estimate that the space of publicly available bioactivity data indexes over at least 1.15 million unique chemicals, annotated onto >15 000 targets (1), with potentially an equal number of phenotypic screens. The exact magnitude of this space could be derived only if one could uniformly process these data into a single database and harmonize chemicals, targets, bioassays and bioactivities. Each of the many sources and databases available has its own interface and data query style, with both

Database, Vol. 2013, Article ID bat044, doi:10.1093/database/bat044

strengths and weaknesses. Such multitude of sources, interfaces and styles is likely to make it difficult for scientists who are not expert in data mining to gather all facts, make connections and appropriate decisions that would lead their own research to the best possible outcome.

This difficulty is best illustrated by considering the chemical biology of estrogen: estrogen-related macromolecular targets include at least five nuclear receptors (estrogen receptors ERα and ERβ; estrogen-related receptors: ERRα, ERRβ and ERRγ), one G-protein coupled receptor (G-protein estrogen receptor, or GPR30), aromatase, several sulfotransferases and sulfatases, as well as the sex hormone steroid-binding globulins. All these targets are associated with and recognize a common chemical pattern (CCP), namely, a *para*-substituted phenol at the 'A' ring. Non-steroidal scaffolds are known to bind one, or several, of the above targets. The steroidal scaffold would be identified by CCP perception tools; however, other chemical signatures as well as non-steroidal CCPs would require more complex methods. Most chemists would not immediately associated estrogen biology with all the above targets, whereas biologists would be less likely to associate estrogen-related targets with non-steroidal chemical signatures.

To address some of these harmonization challenges, and to achieve consistency and coherence among disparate chemical—target—bioactivity pairs, we proposed to develop the unified database, CARLSBAD (Confederated Annotated Research Libraries of Small molecule Biological Activity Data). A chemical relational database, CARLSBAD integrates subsets of bioactivity data (that is, chemicals tested for bioactivity on selected targets) from the following databases: ChEMBL (2), IUPHAR (3), PDSP (4), PubChem (5) and WOMBAT (6).

For the scientist interested in evaluating hundreds of thousands of bioactive compounds, the ability to identify global trends at the CCP or at the target level may be more relevant than, for example, the exact Ki of Propranolol to the three β adrenergic receptor subtypes under a particular set of experimental conditions. Conceptually, for any given 'compound A' that shows activity on 'target W' in the 10–100 nM range according to three independent groups, and only milimolar activity according to a fourth group, most users interested in global trends would reasonably conclude that compound A displays good bioactivity on target W. The opposite trend may be encountered as well: if 'compound B' shows double digit micromolar activity on 'target Y' according to two independent groups, and shows nanomolar activity according to a third group, it could be reasonably assumed that compound B is not potent on target Y. Although detailed resolution of experimental data may be lost during data processing into CARLSBAD, this database aims to provide a 'bird's eye view' of the entire bioactivity landscape, one that is useful for multi-disciplinary research.

The focus on high-quality subsets of data from the five aforementioned databases was a major determinant for CARLSBAD, which aggregates chemical bioactivity information for drug discovery and repurposing activities from five different sources, shown earlier in the text. Only bioactivities that can be normalized to negative log molar values were processed for inclusion in the aggregated database. No single-point bioactivity values or phenotypic/cellular assay data were captured. In the current release, CARLSBAD includes only activity values associated with protein targets from human, mouse and rat. All activity data from the source databases that satisfy the aforementioned criteria are stored in the CARLSBAD database.

For the purpose of data mining, patent analytics and decision making, a single (highest confidence) activity value for any given bioactivity type, e.g. inhibition constant, Ki, or effective concentration at which 50% of the response is obtained (EC50) is calculated, and returned for each unique chemical–protein target pair ('CARLSBAD activity'). CARLSBAD activities correspond to unique four-tuples (chemical–protein–species–activity type). For example, the cholesterol-lowering drug lovastatin has only one activity of type Ki on the human HMG CoA reductase protein—the rate-limiting enzyme in the metabolic pathway that produces cholesterol—stored in the CARLSBAD database. To generate these unique four-tuples, we introduced 'confidence levels' to establish a hierarchy for data sources during aggregation. When multiple activity values of the same type (e.g. Ki) with equal confidence levels were found, the mean value was indexed.

One of the key distinguishing features of CARLSBAD is that CCPs are pre-calculated and stored for all chemical structures in the database. CCPs were derived using the maximum common edge subgraph (MCES) and hierarchical scaffold (HierS) algorithms, as discussed further. The choice of MCES and HierS for CCPs is due to the complementarity of these methods, as each method perceives chemical scaffolds and structural features for small molecules in a different manner. The cross-indexing of bioactivity, target and CCP data enables scientists to perform multiple tasks related to data mining, hypothesis generation and chemical biology space exploration. Classes of structural features that might be responsible for invoking certain biological responses can thus be examined within the CARLSBAD platform. Alternatively, biological targets could be categorized based on their preference toward particular CCPs.

Methods

Database implementation and schema

The CARLSBAD database is implemented as a PostgreSQL relational database with entities such as substance, compound, activity, target and so forth, and the various

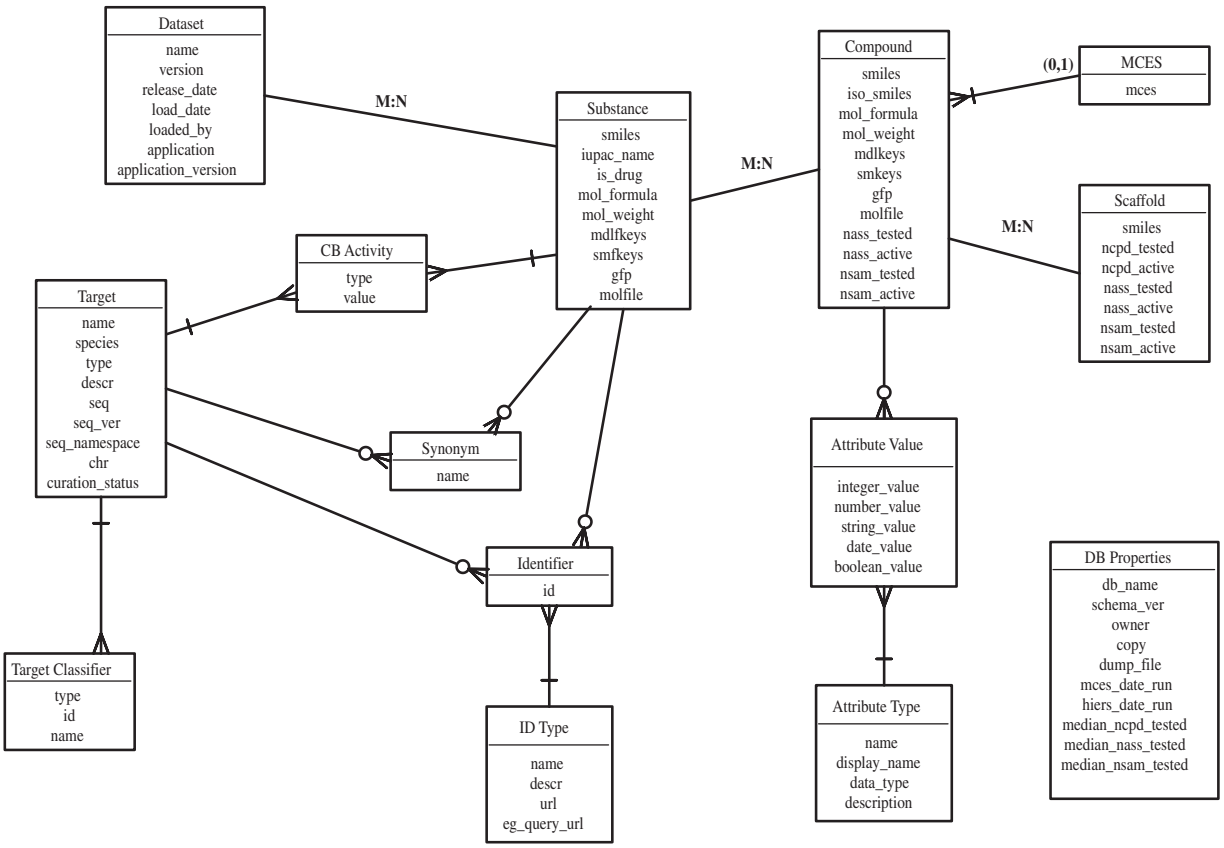

Figure 1. CARLSBAD database entity relationship diagram.

relationships between them (Figure 1). The CHORD chemical cartridge from gNova Scientific Software (http://www.gnova.com/) is used to provide fast chemical functionalities such as SMILES canonicalization (7), chemical fingerprints and structure searching. CHORD is based on the OEChem toolkit, available from OpenEye Scientific Software (http://www.eyesopen.com/).

Data sources, extraction and curation

Separate extract, transform and load (ETL) pipelines were built for each of the data sources. The sections later in the text detail the specific source of data used and the extraction criteria applied for each.

ChEMBL. A MySQL dump of ChEMBL v13, 2012–02–21 was downloaded from the website and used to create a local MySQL staging database that served as the source from which data were extracted and used to populate the CARLSBAD database (2). ChEMBL data passing the following filters were loaded into CARLSBAD. Only activities from publications were loaded; activities associated with pharmacokinetic, cellular and *in vivo* assays, and any other activities not associated with a protein target were not imported; activities not associated with human, rat and mouse targets were skipped; and activities without values or units that could be converted to −log(molar) were also skipped. Activities of the following type were loaded: EC50, IC50, pEC50, pIC50, Log EC50, Log IC50, Ki, Kb, Kd, pKi, pKb, pKd, Log Ki, Log Kb, LogKd, ED50, IC80, IC90, A2, D2, pA2, pD2 and Km. Also, activities with units expressed in molarity, as well as activities with an associated structure were loaded. Additionally, activity values were converted to molar wherever necessary and converted to negative log where appropriate.

IUPHAR. Data were programmatically extracted from the IUPHAR website (http://www.iuphar-db.org/) and used to populate a local MySQL staging database (3). This staging database was constructed during February 2011 and served as the source from which data were extracted and used to populate the CARLSBAD database. Only activities with the following classes were loaded: agonists, antagonists, pore blockers, activators, allosteric regulators, gating inhibitors and channel blockers. In addition, midpoints or medians were used for affinities expressed as ranges. Activities not associated with human, rat and mouse targets as well as activities with unknown affinities or units were excluded.

Database, Vol. 2013, Article ID bat044, doi:10.1093/database/bat044

PDSP. The text file (kidb110121.txt) was downloaded from the website (http://pdsp.med.unc.edu/indexR.html) (4). UniProt IDs were added to this file by the group of Stephan Schurer, University of Miami. This file was used as the source from which data were extracted and used to populate the CARLSBAD database. Only PDSP data passing the following filters were loaded into CARLSBAD. Activities associated with structures not parseable by gNOVA, and activities with qualified values (i.e. >x) were skipped.

PubChem. Only the subset of PubChem derived from the Molecular Libraries Probe Network (PubChem MLP) was used (5). The PubChem Assays and Substances to be loaded into CARLSBAD were selected using the Entrez EUtils API to search pcassay with the following queries/filters: 'Molecular Libraries Probe Production Centers Network[SourceCategory]', confirmatory[Filter] and pcassay_protein_target[Filter]. Substance structures were retrieved as SMILES using the PubChem Power User Gateway (PUG). Assay data were loaded from xml and csv files downloaded from the PubChem ftp site. PubChem MLP data passing all of the following filters were loaded into CARLSBAD. Only activities associated with human, rat or mouse targets were loaded. Only activities with the following result types were loaded: various versions of EC50, AC50, IC50, Ki and Potency. Activities without values or units were skipped. Only activities with units expressed in molarity were loaded. Only activities with an associated structure were loaded. Additionally, activity values were converted to molarity if necessary, and activity values were converted to negative Log10 if necessary.

WOMBAT. Version 2011.2 (SDF and activities.tab files) was used as the source from which data were extracted and used to populate the CARLSBAD database (6). Only activities of the following types were loaded: EC50, ED50, IC50, IC80, IC90, Ki, Kb, Kd, Km, A2 and D2. In addition, the following data were skipped: activities not associated with a known target; activities not associated with human, rat and mouse targets; activities associated with targets without an associated UniProt identifier; activities from primary screening; activities labeled 'inactive'; and activities with descriptive values (e.g. 'active').

When pairing structures with targets and bioactivities in a similar effort (6), Tikkainen and Franke observed that only 3.6% (i.e. 410 of 11 278) of the scientific articles with activity indexed in more than one database matched each other. Indeed, data discrepancies are ubiquitous as far as data curation is concerned (8). The processing log for CARLSBAD is summarized in Table 1: of 975 117 unique structure–target pairs in the database, 84 794 were found unique to WOMBAT and, therefore, have not been processed into CARLSBAD. For the remaining 890 323 structure–target pairs, 124 231 (13.95%) were aggregated from multiple structure–target values: 94 975 from two bioactivities, 14 544 from three, 7930 from four and 2214 from five bioactivities, respectively. The highest number of consolidated bioactivities is 109, with the second highest number being 106. As data aggregation is the intended purpose for CARLSBAD, we focused on eliminating extremes in the bioactivity spectrum, and aggregating values towards a mean value. Hierarchical processing (i.e. confidence levels) was used in ~25% of the cases (192 736 + 38 670 substance–target pairs) when generating the CARLSBAD activity.

Chemical curation

In the CARLSBAD database, chemical substances are distinguished from compounds in a manner analogous to the PubChem terminology. In this paradigm, compounds represent the abstract structure of any of the components of the substance. Chemical structures are stored as canonical SMILES (7) using CHORD (gNova/OpenEye). The corresponding SDF format is also stored if present in the input database. In addition, 26 chemical descriptors are calculated and stored for each unique compound. These descriptors (e.g. molecular weight, number of rings and so forth) are provided for convenience to users interested in specific subsets of chemical space. A key feature of the CARLSBAD database is the common chemical patterns (CCPs), which are calculated and associated with the corresponding chemical structures. Later in the text, we briefly describe the methods used to calculate the two types of CCPs and how they are stored.

HierS

HierS, the hierarchical scaffold grouping algorithm (9), is based on the molecular framework concept described by Bemis and Murcko (10). The 'scaffold' concept is central in medicinal chemistry and provides a chemically intuitive manner to visualize chemical classes, as ring-based linkages are central structural features in most (>90%) drug molecules. The algorithm relates any two compounds by their common shared scaffolds. HierS has two advantages: (i) speed and (ii) HierS scaffolds are considered by some to be more meaningful than the typical maximum common substructure (MCS). To our knowledge, there is no currently available implementation of HierS in any commercial or open-source package. These tools are implemented in an open-source Java library (http://code.google.com/p/unm-biocomp-hscaf/) built on the JChem toolkit from ChemAxon.

MCES. The maximum common edge subgraph (MCES) concept (11) can be used to compute similarity between two molecular graphs and has been widely used in many applications (12–17). However, MCES computation is NP complete, and several heuristics have been proposed to reduce computational time, although computational time required for large chemical datasets is prohibitive. Thus, in

Table 1. CARLSBAD database consolidation process summary

Processed substance–target pairs	Number of bioactivities
975 117	Initial aggregated data
975 110	Valid processed pairs
84 794	WOMBAT only
658 917	Only one activity on record
192 736	Only one activity type (each entered)
38 670	Multiple activity types processed
932 881	total activities loaded

the CARLSBAD database, additional heuristics based on common ring systems/scaffolds were applied to further reduce computational time and make feasible MCES computation for large libraries. The CARLSBAD database contains 435 578 compounds with >99% compounds containing at least one ring. As ring system determination using HierS (9) is efficient and fast, scaffold information determined using HierS was used to group compounds based on the number of common scaffolds shared between them. Once this preliminary heuristic was applied, pairwise MCES between compounds sharing the same set of scaffolds was computed. Thus, the MCES algorithm was run on the CARLSBAD database and used to identify clusters of compounds with shared maximum common substructures.

Target curation

Representation of targets varies greatly across source databases, and this creates several challenges. In particular, targets are named and identified in different ways, which makes it difficult to know whether a target from one data source is the same as a target from another source, i.e. target matching and 'unification'. As the goal was to have one target record in CARLSBAD for each unique protein represented in assays, a target curation step was performed after each data source was loaded, where newly loaded targets were annotated with data from UniProt (18), to expedite the target unification process. Targets identified in the source data by SwissProt or UniProt IDs were annotated with name, description, sequence and other identifiers (NCBI gi, RefSeq, Gene, UniGene and PDB) from UniProt. This allowed a comparison for target redundancy by sequence and identifiers after each data source was loaded to be made. Data from UniProt were also used to annotate targets in the CARLSBAD database with the following classifiers: InterPro, Pfam and PROSITE domains; GO terms; and UniProt family.

Web interface

A browsing and query interface to the CARLSBAD database is available (http://carlsbad.health.unm.edu/carlsbad/) (Figure 2). This web interface is delivered via the open source Apache web server. The application is written in the Perl programming language and uses Marvin Java applets from ChemAxon for drawing and displaying chemical structures. Users can query from structures by name, structure and/or properties; and for targets by name, species, type and/or identifier.

Discussion

The availability of massive amount of molecular bioactivity data creates rich new opportunities, yet for typical scientists involved in biomedical discovery research, the difficulty of processing and analyzing that data can often be a barrier. With the occasional, less experienced end-user in mind, we have developed a small molecule bioactivity database that facilitates navigation in the small molecule/bioactivity space. The unique features and underlying data structure of the CARLSBAD database are designed to support polypharmacology-driven drug discovery scenarios, such as drug repurposing, side effect/off-target prediction and lead identification workflows.

The net result of chemical, bioactivity and target aggregation, curation and harmonization is summarized in Table 2: the number of substances, i.e. chemicals tested for bioactivity, is smaller than the one obtained by summing the five databases by 17.27%. A similar trend is observed when examining bioactivities (34.35% reduction) and CCPs (23.1% reduction using HierS and 25% using MCES). The aforementioned values are the result of machine-based harmonization and consolidation of multiple data objects in chemical, bioactivity and CCP space. An independent study by Tiikkainen and Franke (19), comparing ChEMBL (release 14) and WOMBAT 2012.01, showed >394 000 unique bioactivities in WOMBAT, compared with nearly 3.3 million bioactivities in ChEMBL; and 2755 unique targets in ChEMBL, compared with 1486 unique targets in WOMBAT. The harmonization trends suggest that a consolidated database is preferable to a federated collection, at least in this case, when seeking to evaluate global bioactivity trends. This solution was, for example, implemented in the 'Merz Virtual Bioactivity Database', which integrates ChEMBL and WOMBAT, among other data sources (8, 19).

Comparing the databases, it is apparent that ChEMBL is the most populated in terms of substances, bioactivities and CCPs, followed by WOMBAT and PubChem/MLP. This is to be expected, given their chemogenomic purpose. Two of the databases dedicated to pharmacology, IUPHAR and PDSP, are significantly smaller. An in-depth comparison

Database, Vol. 2013, Article ID bat044, doi:10.1093/database/bat044

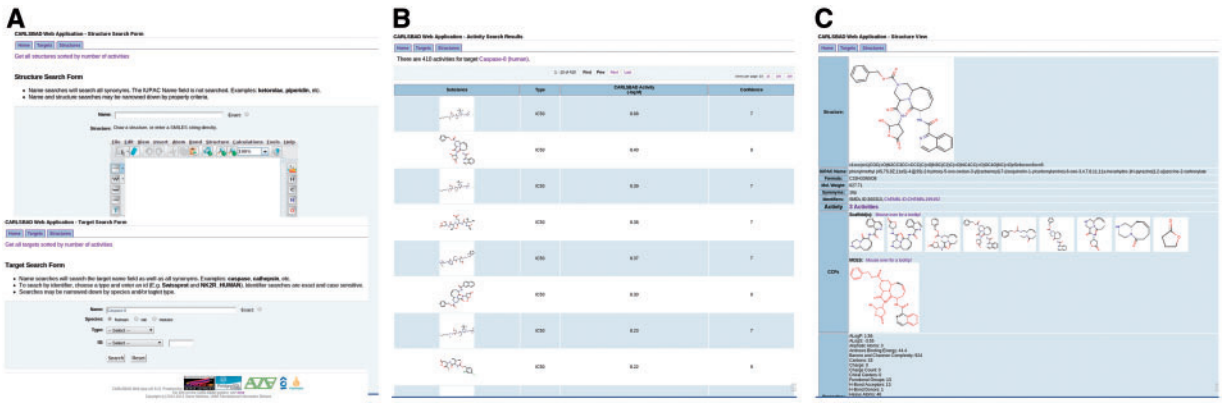

Figure 2. CARLSBAD Web Application. (**A**) Query forms. (**B**) Activity search results. (**C**) Substance view.

Table 2. Overview of the numbers of substances, activities and CCP data in the original databases, as well as the consolidated CARLSBAD database

Source	Version	Release date	Structures	Activities	CCPs
ChEMBL	13	2012–02–21	267 744	798 755	182 496 scaff 32 794 mces
IUPHAR		2011	2297	6049	2704 scaff 652 mces
PDSP	kidb110121		3499	22 202	3422 scaff 823 mces
PubChem MLP		2011–11–04	133 435	320 311	83 570 scaff 20 867 mces
WOMBAT	2011.2		124 873	273 572	88 135 scaff 17 086 mces
Total			*531 848*	*1 420 889*	*360 327 scaff 72 222 mces*
CARLSBAD	2012.1		439 985	932 881	277 140 scaff 54 135 mces

with respect to targets, bioactivities and chemistry coverage for some of these databases has been performed (8). Each of these databases provided relevant contributions in terms of CARLSBAD aggregation.

Chemical errors were addressed with focus on the high-value, high-confidence IUPHAR and PDSP subsets. We found only one PDSP structure that was not parsable by gNova/OEChem; it was manually corrected. For IUPHAR, we extensively curated >2700 small molecules and peptides from IUPHAR's 'Ligand List' (http://www.iuphar-db.org/DATABASE/LigandListForward: retrieved fourth February 2011). This curation involved reading the original ligand references to resolve ligand names, 2D structures and biological activities, including >700 peptides for which structural information was not then available in IUPHAR-DB (20). In the future, the teams supporting the CARLSBAD and IUPHAR-DB projects will work together to ensure the consistency of data between the two resources.

When aggregating data in CARLSBAD, we did not explicitly address biology or bioactivity errors. By cross-referencing PubMed IDs for literature-based data (i.e. PDSP, IUPHAR-DB, ChEMBL and WOMBAT), we found that identical articles are covered by these resources, yet data are not always identical. Indeed, up to 3% errors in target protein identity, up to 2.7% errors in bioactivity values, and up to 7% errors in chemical structure depiction were found in comparing three data sources (19). In CARLSBAD, these tuples were harmonized by providing median values wherever possible, and by representing 'higher curation' values where possible, when multiple conflicting values were found. For example, bioactivity results from IUPHAR-DB were given the highest priority, as they summarize the significant curation effort made by members of the IUPHAR Nomenclature Committee. Overall, this situation occurred in <10% of the database. With respect to data generated by the NIH Molecular Libraries Initiative (21), only data from PubChem was uploaded into CARLSBAD, as stated earlier in the text. Thus, any bioactivity value 'feedback loop', i.e. propagation of errors from one database to another, was avoided by importing non-overlapping sets of data.

Chemical space overlap between structures in the CARLSBAD database and drugs approved for human use was determined using structure identity comparison with an in-house curated database of drug structures approved worldwide (DRUGSDB), which includes discontinued drugs

as well (22, 23). A total of 1435 unique chemical structures for active pharmaceutical ingredients (i.e. 'drugs') were identified in CARLSBAD of ~4000 small organic molecules from DRUGSDB. These chemical structures were flagged accordingly for user convenience and can be used to explore biological activity space of known drugs.

CARLSBAD represents only a first step in our effort to assist non-expert scientists to navigate chemical biology data. For example, all protein targets related to estrogen biology can be identified via a single CARLSBAD query. However, their inter-connectedness via chemicals and CCPs is intended to be explored in the networked environment provided by Cytoscape (24). The CARLSBAD network extraction tool (SNAKE), the Cytoscape plugin and the process of visualizing networks of connected protein targets, chemical structures and bioactivities, are described elsewhere (Hines-Kay *et al*., submitted for publication).

Summary

CARLSBAD is a database focused on high-quality subsets aggregated from several bioactivity databases, which are integrated in a uniform interface and manner, suitable for chemical biology and drug discovery studies, as well as large scale, 'big data' informatics and knowledge mining. In contrast to the original data collections, CARLSBAD provides a single normalized activity value of a given type for each unique chemical–protein target pair. Aggregation accounted for ~25% of the >975 000 structure–target pairs processed, up to and including 109 bioactivities for a single chemical. CARLSBAD data processing resulted in a net 17.3% reduction in terms of unique chemicals, 34.3% reduction in terms of unique bioactivities and >23% reduction in terms of CCPs, respectively, suggesting that data consolidation is preferable to a federated database system, at least where bioactivity is concerned. We implemented two types of scaffold perception for common chemical pattern detection HierS and MCES, respectively. The 2012 release of CARLSBAD contains 439 985 unique chemical structures, mapped onto 1 420 889 unique bioactivities and annotated with 277 140 HierS scaffolds and 54 135 MCES patterns, respectively. It also contains bioactivities and tags for 1435 unique active pharmaceutical ingredients. The CARLSBAD database can be accessed using SNAKE; our dedicated subnet extraction tool, and Cytoscape, via the CARLSBAD plugin (Hines-Kay *et al*., submitted for publication).

Access and linking to CARLSBAD

The CARLSBAD database can be accessed by accessing the dedicated website, http://carlsbad.health.unm.edu/carlsbad/. For more information on the CARLSBAD platform, visit the project homepage, http://carlsbad.health.unm.edu/. If you are interested in establishing links to CARLSBAD, please notify us via email at info-carlsbad@poblano.health.unm.edu.

Funding

National Institutes of Health (R21GM095952-01 to T.I.O.). Funding for open access charge: R21GM095952-02.

Conflict of interest. T.I.O. is Founder and CEO of Sunset Molecular Discovery LLC, which markets the WOMBAT database.

References

1. Kim Kjærulff,S., Wich,L., Kringelum,J. *et al*. (2013) ChemProt-2.0: visual navigation in a disease chemical biology database. *Nucleic Acids Res.*, **41**, D464–D469.
2. Gaulton,A., Bellis,L.J., Bento,A.P. *et al*. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
3. Sharman,J.L., Mpamhanga,C.P., Spedding,M. *et al*. (2011) IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic. Acids Res.*, **39**, D534–D538.
4. Roth,B.L., Lopez,E., Patel,S. *et al*. (2000) The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *Neuroscientist*, **6**, 252–262.
5. Bolton,E.E., Wang,Y., Thiessen,P.A. *et al*. (2008) PubChem: integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.*, **4**, 217–241.
6. Olah,M., Rad,R., Ostopovici,L. *et al*. (2008) WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery. In: Schreiber,S.L., Kapoor,T.M. and Wess,G. (eds), *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*. Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 760–786.
7. Weininger,D., Weininger,A. and Weininger,J.L. (1989) SMILES 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97–101.
8. Tiikkainen,P. and Franke,L. (2012) Analysis of commercial and public bioactivity databases. *J Chem. Inf. Model.*, **52**, 319–326.
9. Wilkens,S.J., Janes,J. and Su,A.I. (2005) HierS: hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.*, **48**, 3182–3193.
10. Bemis,G.W. and Murcko,M. A. (1996) The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.*, **39**, 2887–2893.
11. Raymond,J.W., Gardiner,E.J. and Willett,P. (2002) RASCAL: calculation of graph similarity using maximum common edge subgraphs. *Comput. J.*, **45**, 631–644.
12. Sheridan,R.P., Hunt,P. and Culberson,J.C. (2005) Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.*, **46**, 180–192.
13. Stahl,M. and Mauser,H. (2005) Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model.*, **45**, 542–548.
14. Gardiner,E.J., Gillet,V.J., Willett,P. *et al*. (2007) Representing clusters using a maximum common edge substructure algorithm applied to

reduced graphs and molecular graphs. *J. Chem. Inf. Model.*, **47**, 354–366.

15. Boücker,A. (2008) Toward an improved clustering of large data sets using maximum common substructures and topological fingerprints. *J. Chem. Inf. Model.*, **48**, 2097–2107.
16. Hariharan,R., Janakiraman,A., Nilakantan,R. *et al*. (2011) MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. *J. Chem. Inf. Model.*, **51**, 788–806.
17. Boström,J. (2012) Symmetric Kv1.5 blockers discovered by focused screening. *ACS Med. Chem. Lett*, **3**, 769–773.
18. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
19. Tiikkainen,P. and Franke,L. (2013) Estimating error rates in bioactivity databases. *J Chem. Inf. Model*.
20. Sharman,J.L., Benson,H.E., Pawson,A.J. *et al*. (2013) IUPHAR-DB: updated database content and new features. *Nucleic Acids Res.*, **41**, D1083–D1088.
21. Austin,C.P., Brady,L.S., Insel,T.R. *et al*. (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.
22. Oprea,T.I., Nielsen,S.K., Ursu,O. *et al*. (2011) Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Mol. Inf.*, **30**, 100–111.
23. Manallack,D.T., Prankerd,R.J., Nassta,G.C. *et al*. (2013) A chemogenomic analysis of ionization constants - implications for drug discovery. *Chem. Med. Chem.*, **8**, 242–255.
24. Shannon,P., Markiel,A., Ozier,O. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.