Drug Discovery Today: Technologies

Vol. 14, 2015

From chemistry to biology database curation

Chemical databases: curation or integration by user-defined equivalence?

Anne Hersey *, Jon Chambers, Louisa Bellis, A. Patrícia Bento, Anna Gaulton, John P. Overington

European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

There is a wealth of valuable chemical information in publicly available databases for use by scientists undertaking drug discovery. However finite curation resource, limitations of chemical structure software and differences in individual database applications mean that exact chemical structure equivalence between databases is unlikely to ever be a reality. The ability to identify compound equivalence has been made significantly easier by the use of the International Chemical Identifier (InChI), a non-proprietary line-notation for describing a chemical structure. More importantly, advances in methods to identify compounds that are the same at various levels of similarity, such as those containing the same parent component or having the same connectivity, are now enabling related compounds to be linked between databases where the structure matches are not exact.

Section editor:
Antony Williams – Royal Society of Chemistry, Wake Forest, NC, USA.

Introduction

Because of the pressures in the pharmaceutical industry of increasing drug development costs, greater requirements for safer medicines and desire for prescribers to show value for money, over the past 5–10 years the industry has changed such that increasingly early drug discovery is being undertaken by SMEs (small to medium sized enterprises) and in academic groups. These groups do not have the large chemical and biological databases of the large Pharma companies and so are more reliant on the data available in public domain databases. The availability of open access databases on the Internet has greatly increased over the past 5 years. This itself brings advantages of chemical structure diversity but also disadvantages of lack of standardisation, particularly in chemical structures, as organisations have evolved their own business rules for standardising chemical structures and have limited resources for curation activities. This paper will outline some of the valuable resources available to drug discovery researchers, highlight some of the issues around curation and standardisation and discuss some of the methods and tools available to overcome some of these issues.

Open and public domain databases

The available public domain databases that are specifically aimed at drug discovery scientists all have their own specialist content and, in general, this is complementary. For example, vendor information, patented compounds, data on marketed drugs, as well as bioactivity data for both efficacy and liability

*Corresponding author: A. Hersey (ahersey@ebi.ac.uk)

 http://dx.doi.org/10.1016/j.ddtec.2015.01.005

Table 1. Examples of some publicly available databases containing chemical information including a description of their content and the number of compounds they contain

Database	Content	Size (no. of compounds)	URL	Reference
Bioactivity data				
ChEMBL	Bioactivity data from the medicinal chemistry literature	1 360 000	https://www.ebi.ac.uk/chembldb	[25]
PubChem	Biological screening results on small molecules	49 000 000	https://pubchem.ncbi.nlm.nih.gov/	[14]
Patents				
IBM	Chemicals from full text patents	2 500 000	http://www-935.ibm.com/services/us/gbs/bao/siip/	
SureChEMBL	Chemicals from full text patents	12 400 000	https://www.surechembl.org	
Drugs				
DRUGBANK	Drug data and drug target information	7700	http://www.drugbank.ca	[26]
FDA/USP SRS	Substances present in FDA regulated products	34 000	http://fdasis.nlm.nih.gov/srs/srs.jsp	
Availability				
ZINC	Commercially available compounds	22 700 000	http://zinc.docking.org	[27]
emolecules	Commercially available compounds	5 900 000	http://www.emolecules.com	
Other				
ChEBI	Database and ontology of Chemical Entities of Biological Interest	27 000	https://www.ebi.ac.uk/chebi/	[20]
PDB	Data on biological macromolecular structures	16 000	https://www.ebi.ac.uk/pdbe/	[28]

Note: All numbers from Apr 2014.

targets and crystal structures of small molecules bound to protein targets, can all be found in public databases. The number of compounds ranges from the comparatively small manually curated sets, such as ChEBI, to large patent databases, such as SureChEMBL, where the data is extracted from patents using 'name to structure' and 'image to structure' software and for which manual curation would be an prohibitively expensive task. Table 1 summarises some of these databases for which the chemical structures, identifiers and in many cases additional data can be freely downloaded.

As well as their own primary content, some databases also take depositions from other databases, or directly from depositors. For example: PubChem includes data from ChEMBL and ChEBI, alongside an extensive set of user depositions; ChEMBL includes some data from PubChem, and ZINC contains data from ChEMBL. Similarly, the Open PHACTS drug discovery platform [1] includes data from ChEBI, ChEMBL and DrugBank and tracking data provenance under these circumstances is challenging. ChemSpider [2] is a chemical structure database that currently integrates data on about 30 million chemicals from more than 470 other databases of varying content. The difference between ChemSpider and the aforementioned databases, is that while users can search for compounds online and get links to data in the originating databases it isn't possible to download the ChemSpider compounds as a complete dataset. Given the trend for aggregating database content, it is particularly important that database providers supply attributions to the data so that the provenance can be determined.

Much has been written about the quality of data, both chemical and biological, in public databases and the impact of incorrect structures on modelling [3–5]. Data quality will undoubtedly be varied for different databases and it would be unreasonable to expect to see the same degree of curation on a database of 10 million compounds and one with only a few thousand. What is certain is that all data providers will curate compounds to the best of their abilities and as far as their budget allows. However, even with resource and time to do it, there are several factors that will lead to the same compound appearing to have different structures in different databases.

Sources of structure differences

One difficulty encountered when trying to curate chemical structures is that often there is no definitive source for a structure. Until recently, marketed drugs were a key example of this and although attempts are now being made to create a definitive database of structures [6] it seems that more curation is needed [3].

In scientific publications, compound structures are often drawn in a form that has relevance to the context of the paper. For example, in a docking paper an acidic or basic molecule might be drawn as a negatively or positively charged molecule as this is the relevant form for binding to the protein. Other papers might report bioactivity data and display the parent form of the molecule even though the dosed substance was its salt. Using trivial names such as USAN (United States Adopted Name) and INN (International Nonproprietary Name) for drugs to try and identify their structures is also fraught with difficulties. An INN is, in most cases, filed for a parent structure whilst since 2004, different USANs need to be filed for both the parent and the salt structures. Prior to 2004, only the marketed form (often a salt or ester)

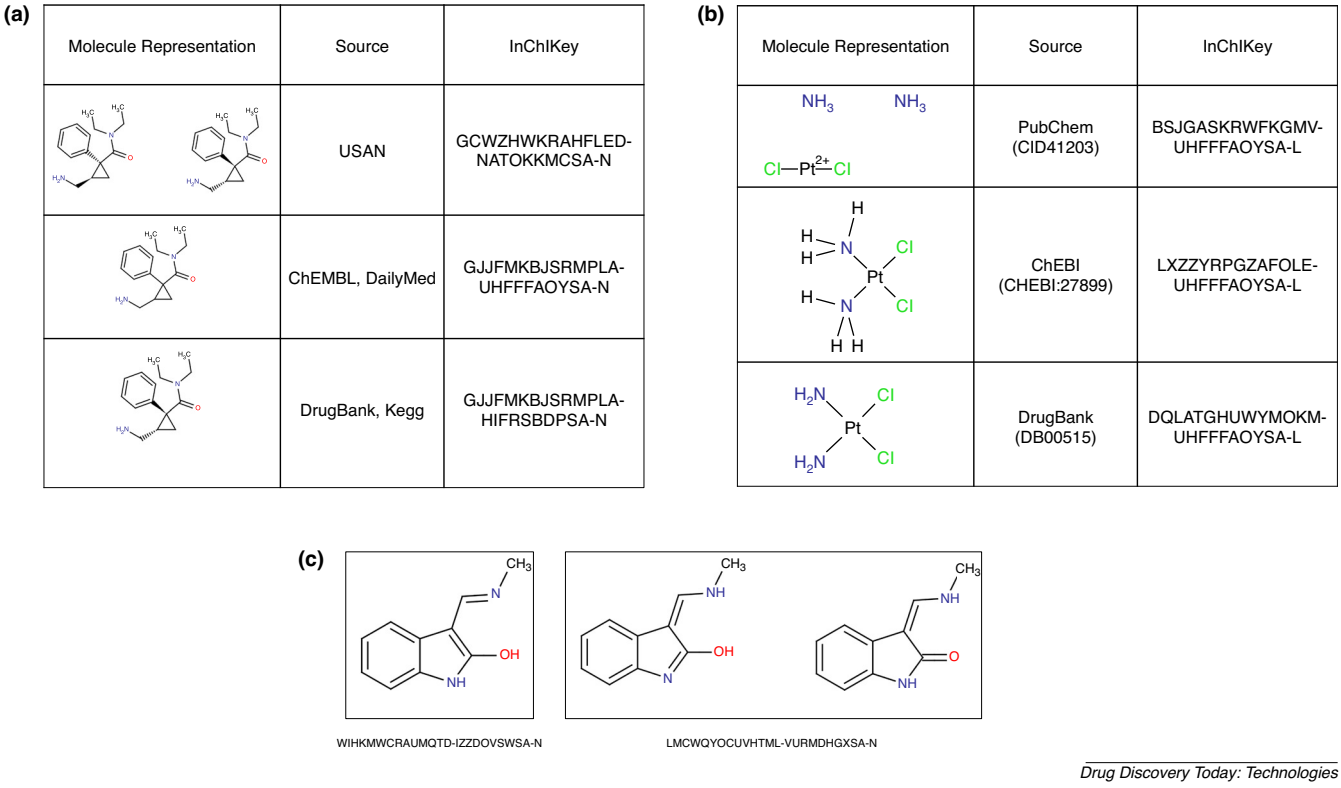


Figure 1. Examples of different structure representations and the effect of these representations on the InChIKey. **(a)** Examples of database structure representations for Milnacipran. **(b)** Examples of database structure representations for Cisplatin. **(c)** Examples of tautomers with matching and non-matching Standard InChIs.

would require an USAN. Taking Sildenafil Citrate as an example, the respective USAN is 'Sildenafil Citrate' whereas there is no INN filed for this salt. However, an INN has been recommended for the parent of this salt and that is 'Sildenafil'. While a clear mapping between a structure and a synonym is possible when examining the original source of the synonym, confusion may arise when examining a source of compiled synonyms, such as the USP Dictionary [7]. In this dictionary, for USANs adopted prior to 2004, the INN is often recorded against the USAN structure, which in most cases is a salt. Hence for this example, not only the USAN, but also the INN Sildenafil, will be recorded against the structure of Sildenafil Citrate. This illustrates how the use of these trivial names can lead to confusion and mismatches with different structures being given the same name.

Currently, most chemical structures in the scientific literature appear as images and not in a structure readable format. This means that the process of extracting the chemical structures and loading them into a database entails redrawing the structure from the image in the paper, or using image to structure software, or a combination of both. Inevitably, this will introduce structure errors, which won't be prevented unless journal editors insist on structure files being submitted for chemical structures [8].

Software limitations

Almost all public databases are using the v2000 molfiles [9] as the preferred way of storing chemical structures and these have some limitations in their ability to represent certain types of compounds. Firstly, they cannot represent compounds that have two stereogenic centres and are a mixture of two enantiomers but do not contain any of the diastereoisomers. The drug Milnacipran is a typical example of this as it is a mixture of the 1R, 2S and 1S, 2R enantiomers. But how best to represent this? It can be drawn on paper as a mixture (Fig. 1a) and this is how it appears in the American Medical Association's USAN document [10]. However, if it is stored as a molfile with two components in a database this creates problems, especially when calculating properties. Representing it as a racemate with no stereochemistry shown or as a single enantiomer is also arguably incorrect. So, what is seemingly the same compound is represented in Drugbank and PubChem as a single enantiomer but in ChEMBL and DailyMed as a racemate. A second structural class that is not well described in v2000 molfiles is that of coordination compounds, such as the drug Cisplatin. Here there is no method for adequately representing the dative bonds. Again, the molecule can be drawn on paper but in databases it is represented in various ways to try and overcome the

inadequacy of the molfile representation. Some of these representations are shown in (Fig. 1b).

Although v2000 molfiles are still the format of choice for storing structures in databases, thereby enabling substructure and similarity searching, when comparing the presence or absence of compounds across public databases it is most commonly the Standard InChI (or Standard InChIKey) that is used [11]. Most of the time this works extremely well and it has the advantage that the Standard InChI is tautomer independent, but again there are currently known limitations. For example, some types of 1,5 tautomers, such as the structures shown in Fig. 1c, would not be identified as the same compound from the Standard InChI. Also relative stereochemistry can be captured in a non-Standard InChI [12] but as it is the Standard InChI that is used to determine structure uniqueness this doesn't help with database mapping for compounds, such as the Milnacipran example shown above. Interestingly the Standard InChI is also unable to distinguish between Cisplatin and Transplatin as it does not recognise the cis- and trans-geometric isomerism of the platinum.

Business rules for standardisation

Most database providers have their own set of business rules that they use for standardising chemical structures. These tend not to be so strict as those deployed by pharmaceutical companies in their registration systems, where a key driver is the ability to prove novelty for intellectual property purposes. For publicly available databases, basing their business rules on guidelines such as those produced by the FDA for their substance registration system [13] is adequate for most purposes. Database providers will, however, have preferences and rules for whether compounds are 'merged' at a parent level, how nitro groups and sulphoxides are standardised and whether tautomers are canonicalised and if so, how this is done. Taking a simple case such as the representation of a nitro group, whether it is standardised as the pentavalent or charge separated form does not matter, but what is important is that it is standardised consistently throughout a database.

Some database providers make their rules and standardisation software available for people to use online. The PubChem Standardisation Service is an example of this [a]. This is the same set of validation and standardisations that they apply to deposited structures. The process consists of a validation step where, for example, the structure is checked for valid atom types, valence checks are performed and functional groups such as nitro groups are converted to a consistent representation. This is followed by a standardisation step in which converted to a canonical tautomeric form, aromatic structures are kekulised, placement of stereo bonds are standardised and all implicit hydrogens are converted to explicit hydrogens.

The Royal Society of Chemistry (RSC) has also recently made their Chemical Validation and Standardisation Platform (CVSP) available for people to use on their own compounds [15]. This also applies a series of validation and standardisations and is based on GGA's Indigo and OpenEye's toolkits as well as some in-house standardisation libraries. In contrast to the PubChem Standardisation Service, the RSC have recognised that database providers can have different requirements for standardisations and CVSP has been written such that a user can select their own preferences for the set of standardisation rules to use and apply just those to their compounds.

Other database providers, such as ChEMBL, have implemented their validation and standardisation process using pipelining tools such as Pipeline Pilot or Knime. These tools also allow flexibility such that new components can be added or adapted as business needs change. The ChEMBL database providers routinely include a salt stripping process in their standardisation based on a dictionary of pharmaceutically relevant salts. This enables bioactivity data, whilst recorded against the experimental salt, to be grouped at the parent level.

Software vendors are also aware of the needs of chemoinformaticians to be able to standardise large sets of compounds, whether it is for database creation or for preprocessing prior to analysis. Off the shelf solutions such as ChemAxon's standardizer [16] or Biovia's Cheshire [17] are also now available. These also have the flexibility for users to select the appropriate standardisations that meet their business needs and then to process molecules and standardise them in a consistent manner.

Tautomerisation is a complex topic for database providers and a few years ago was the subject of a whole issue of the Journal of Computer Aided Molecular Design [18]. In this, Wendy Warr [19] outlined the various approaches taken by 27 software vendors and database providers to treat tautomers. There is still no consensus on whether tautomers should be canonicalised and if so how it is done. Currently, some database providers canonicalise tautomers while others do not. In general, although databases do show a single tautomeric representation, ChEBI being a notable exception [20]. The 'preferred' tautomer used in a database matters in the sense that molecules drawn as different tautomers will not be recognised as the same structure in substructure or similarity searches. Also, the physicochemical properties calculated on a compound will generally be different for different tautomeric forms.

The ideal situation would be that all database providers standardise their compounds in the same way as this would greatly enhance data exchange and integration but this is probably some way off. The good news is that the Standard InChI and Standard InChIKey are independent of many of these structure representations, including tautomers. Thus, this enables users to use the Standard InChIKey as an identifier for the occurrence of a compound in different databases,

in most cases, irrespective of its stored structural representation or tautomeric form.

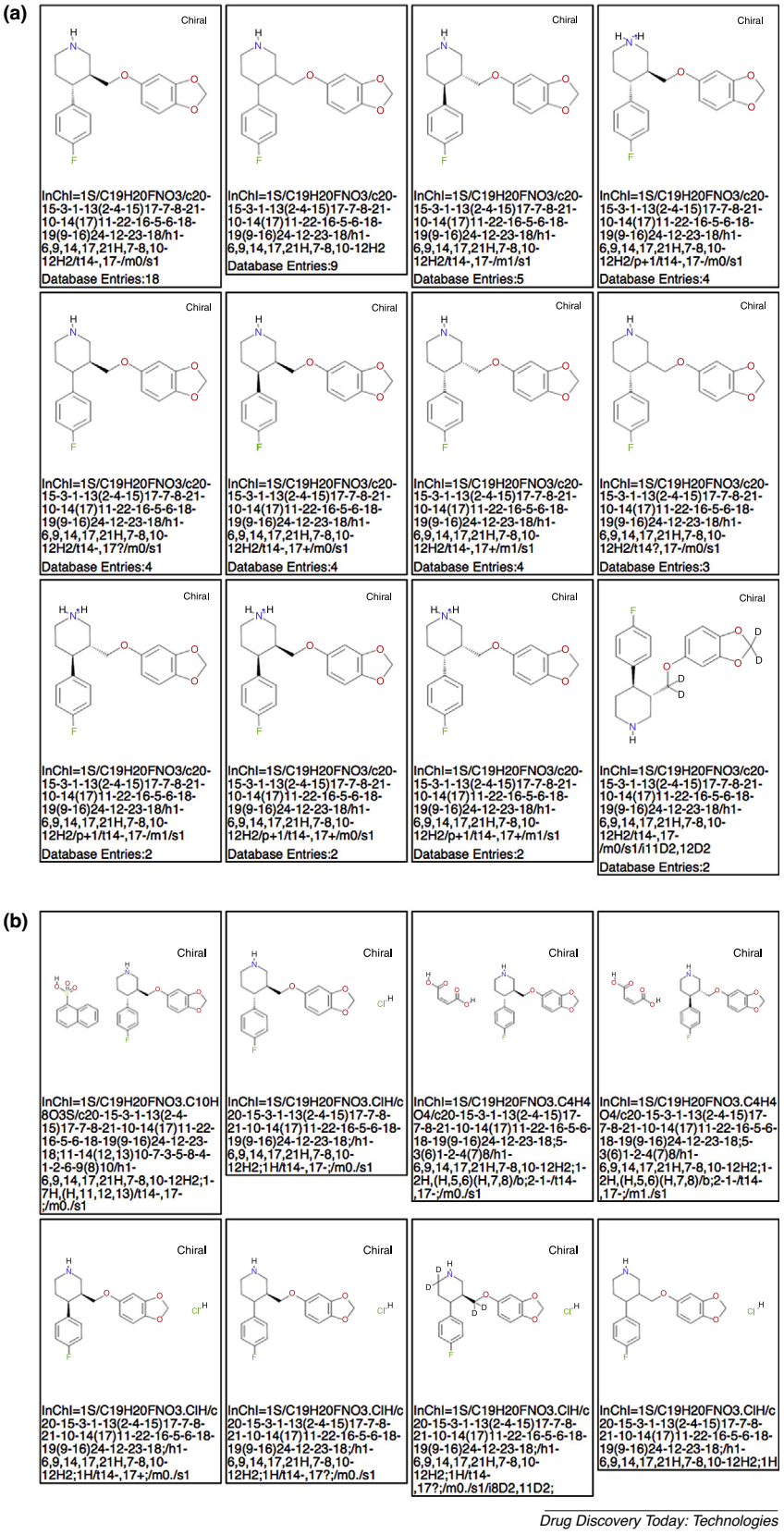
Matching across databases

For all the reasons discussed above, there will be intended and unintended differences between the chemical structure representation of compounds in different databases and this is something that users of these databases need to be aware of and accept when using them. Lipinski et al., for example, highlighted the difficulty in identifying structures and hence bioactivity data in different databases for the NIH Molecular Library Probes [21] which are a relatively well known and well characterised compound set. However, there are now approaches being taken by some data providers, such as PubChem, Open PHACTS and ChEMBL, to help link compounds where the structures have been incompletely or incorrectly represented. One goal of the Open PHACTS project is to enable the easy searching and retrieval of different data types across varied data sources. In the chemistry space they are developing methods that allow the user to decide whether they consider different tautomers or different stereoisomers for example as 'the same physical entity' [1]. The National Institutes of Health (NIH) in the USA has developed a Chemical Structure Lookup Service (CSLS) that identifies which databases a particular structure occurs in. This search can be done on the basis of different levels of specificity such as tautomerism, counter ions, isotopes, charges and stereochemistry, and currently searches across 74 million structures in over 100 databases (http://cactus.nci.nih.gov/cgi-bin/lookup/search). Using their methodology they were able to identify the number of unique parent molecules in a large number of databases and then the number of compounds that were tautomers of each other [22]. On average they found that 0.3% of the structures in the individual databases were tautomers, although this did vary from database to database. They also showed in an analysis of 103 million original structure records from about 150 databases that once the parent structures were generated, and the differences just due to tautomers removed, only 70.6 million structures remained. This is a reduction in unique compounds of about 30%.

UniChem [23] is a mapping service developed at EMBL-EBI based on Standard InChIKeys that can be used to map compounds between databases. Currently, it contains 97 million structure records, of which 63 million are unique compounds, and it provides compound mappings across 22 different databases. As with the CSLS system, it shows tautomer independent matches, as defined by the Standard InChIKey, and has recently been enhanced to enable connectivity mapping of compounds between databases (https://www.ebi.ac.uk/unichem/widesearch/widesearch) [24]. This is achieved simply by identifying the differences or similarities in the InChI layers. For example, it will show where there is the racemate of a compound in one database and a specific enantiomer in another, a salt in one database and the parent in another or an isotopic difference between compounds in different databases.

As a way of exemplifying the use of this and showing the variability of structural forms in databases, the connectivity mapping obtained using UniChem for Paroxetine and some of its related forms is shown in Fig. 2. This is particularly useful as it means that the user can easily identify compounds across databases that vary only by the individual differences or combination of differences that they are interested in. These differences can be stereochemistry, salt forms, isotopic substitution or charge. The reasons for these discrepancies can be due to genuine differences, errors in structure drawing or disparity in business rules, such as storing parent versus salt forms or use of charges on basic nitrogen for docking studies. It is worth pointing out that this will not identify tautomeric differences in structures, as the Standard InChI that is used for the mapping is tautomer independent.

To highlight the full benefits of a connectivity mapping approach we have compared the differences in the number of compounds mapped between all the databases in UniChem, using connectivity versus exact mapping (Table 2). Unsurprisingly, in virtually all cases, the connectivity mapping results in higher numbers of compound matches between databases. The difference in some cases is as high as 20%. For example, it can be seen that matching compounds on the basis of connectivity results in 67.15% of the DrugBank compounds matching a compound in ChEMBL, whereas comparing exact matches it is only 53.58%. In terms of numbers this is an additional 780 compounds that have the same connectivity but have different stereochemistry, charge among others. In the example of SureChEMBL to ChEMBL, the connectivity mapping results in an extra 3% of matches, which equates to identification of an extra 1600 compounds in the patent literature for which there is data on a related compound in ChEMBL. There are several use cases where these mappings can potentially be very useful to both users and database providers. Firstly, while different salt forms will have different InChIs and so not be identified as 'the same', from a bioactivity perspective it doesn't generally matter what the salt form is and so being able to link these compounds is useful. An example of the potential of this is for the marketed drug Sildenafil. The drug is sold as the citrate salt but in ChEMBL, for example, most of the bioactivity data has been determined on the parent (22 data points for the citrate salt versus >1000 on the parent). It is also a tool for potentially identifying which compounds should be considered for curation. For example, if database A has no exact matches to a compound in any other database (or a small number of matches) but many connectivity matches, this might suggest there is a stereochemistry error in the structure in database A. It might also enable users to identify common



Drug Discovery Today: Technologies

Figure 2. **(a)** Examples of Paroxetine-like structures in different databases. Paroxetine is the first structure. The Standard InChI for each structure is shown as is the number of database entries for that particular structure. **(b)** Database examples of salts and mixtures of compounds where one component is a connectivity match of Paroxetine.

Table 2. Percentage of exact compound to compound matches between databases determined using Standard InChIKey matches from UniChem. The percentage of compound connectivity matches (also from UniChem) is shown in brackets. The percentages are calculated as the percent of source 'X' (header row) which overlaps with source 'Y' (first column). Full descriptions of the sources are available at https://www.ebi.ac.uk/unichem/ucquery/listSources

	chembl	drugbank	pdb	iuphar	pubchem_dotf	kegg_ligand	chebi	nih_ncc	zinc	emolecules	ibm
chembl	100 (100)	53.58 (67.15)	41.26 (50.18)	65.60 (81.90)	62.64 (77.15)	42.27 (50.86)	29.14 (41.98)	91.09 (98.74)	1.3 (3.11)	6.95 (7.48)	2.78 (3.33)
drugbank	0.25 (0.34)	100 (100)	23.41 (31.38)	14.01 (20.57)	9.21 (17.57)	9.13 (13.79)	7.22 (13.07)	35.60 (72.98)	0.01 (0.02)	0.04 (0.06)	0.08 (0.12)
pdb	0.50 (0.65)	60.92 (79.16)	100 (100)	9.12 (14.66)	7.93 (13.30)	11.69 (15.71)	10.28 (16.21)	15.43 (35.09)	0.01 (0.03)	0.07 (0.09)	0.14 (0.18)
iuphar	0.08 (0.11)	3.61 (5.20)	0.9 (1.47)	100 (100)	3.29 (6.72)	2.10 (3.21)	1.19 (2.39)	7.23 (21.44)	0.00 (0.01)	0.01 (0.02)	0.03 (0.03)
pubchem_dotf	0.26 (0.36)	8.19 (15.89)	2.71 (4.77)	11.38 (24.06)	100 (100)	3.79 (6.96)	3.25 (6.26)	39.49 (57.10)	0.00 (0.02)	0.03 (0.05)	0.06 (0.11)
kegg_ligand	0.44 (0.54)	20.16 (28.48)	9.93 (12.85)	18.05 (26.22)	9.42 (15.89)	100 (100)	22.05 (34.66)	32.82 (63.92)	0.02 (0.04)	0.10 (0.13)	0.19 (0.24)
chebi	0.59 (0.62)	30.99 (37.81)	16.96 (18.59)	19.88 (27.42)	15.69 (20.02)	42.85 (48.58)	100 (100)	49.37 (71.30)	0.02 (0.04)	0.14 (0.15)	0.27 (0.29)
nih_ncc	0.05 (0.06)	4.03 (8.40)	0.67 (1.60)	3.18 (9.78)	5.03 (7.27)	1.68 (3.57)	1.30 (2.84)	100 (100)	0.00 (0.00)	0.01 (0.01)	0.01 (0.02)
zinc	20.85 (37.04)	21.59 (45.75)	16.13 (27.42)	16.34 (54.28)	16.02 (41.43)	24.37 (48.33)	18.47 (35.64)	31.43 (91.36)	100 (100)	54.79 (90.29)	3.31 (5.43)
emolecules	26.56 (30.71)	34.49 (49.76)	21.26 (29.60)	39.35 (63.17)	31.66 (42.23)	37.21 (49.48)	26.68 (42.04)	84.28 (96.93)	13.04 (31.09)	100 (100)	4.22 (4.80)
ibm	5.08 (6.58)	32.93 (44.84)	20.90 (27.23)	38.67 (52.12)	25.89 (49.14)	33.53 (44.55)	24.35 (39.30)	43.53 (83.84)	0.38 (0.90)	2.02 (2.31)	100 (100)
atlas	0.04 (0.05)	3.84 (5.32)	1.37 (1.94)	4.10 (6.22)	2.14 (3.60)	2.42 (3.22)	1.71 (2.91)	10.70 (20.75)	0.00 (0.00)	0.01 (0.01)	0.01 (0.02)
patents	2.35 (3.08)	29.39 (38.94)	14.10 (19.2)	29.68 (41.46)	21.41 (41.89)	27.39 (37.83)	18.87 (31.1)	45.06 (84.81)	0.12 (0.30)	0.66 (0.78)	13.12 (13.56)
fdasrs	1.05 (1.07)	26.84 (33.96)	13.47 (15.08)	24.11 (36.50)	40.63 (48.20)	32.85 (38.26)	20.27 (27.28)	58.27 (79.24)	0.04 (0.09)	0.28 (0.28)	0.49 (0.60)
surechembl	17.59 (20.83)	50.26 (70.61)	36.62 (50.80)	60.58 (75.61)	66.50 (81.71)	45.14 (61.77)	28.31 (48.1)	79.13 (94.84)	0.89 (2.02)	4.41 (4.78)	62.67 (68.35)
pharmgkb	0.05 (0.06)	9.65 (11.62)	1.26 (2.00)	7.53 (10.92)	3.56 (7.09)	3.36 (4.43)	2.18 (3.70)	23.22 (45.26)	0.00 (0.00)	0.01 (0.02)	0.02 (0.03)
hmdb	0.32 (0.50)	27.44 (32.99)	7.74 (10.96)	15.66 (23.61)	9.28 (17.63)	21.49 (31.56)	13.90 (26.89)	36.99 (74.09)	0.01 (0.04)	0.10 (0.13)	0.20 (0.27)
selleck	0.12 (0.14)	7.26 (13.63)	2.38 (3.97)	5.08 (12.76)	12.06 (15.66)	3.17 (5.70)	2.67 (5.15)	42.28 (58.63)	0.00 (0.01)	0.03 (0.03)	0.02 (0.04)
pubchem_tpharma	26.68 (32.04)	55.71 (72.99)	45.17 (56.8)	67.93 (88.76)	76.26 (90.56)	55.26 (69.18)	43.98 (64.46)	81.64 (96.51)	0.52 (1.16)	2.46 (2.75)	6.85 (8.48)
pubchem	94.66 (96.73)	94.66 (97.4)	83.09 (89.66)	93.63 (99.17)	96.33 (98.49)	96.27 (98.43)	93.85 (97.31)	95.13 (97.63)	41.11 (57.83)	93.63 (95.03)	79.82 (82.32)
mcule	24.16 (27.89)	16.68 (22.66)	11.11 (14.29)	12.97 (17.26)	8.04 (13.53)	14.94 (19.74)	10.55 (16.76)	42.00 (62.11)	14.73 (31.55)	76.06 (81.94)	2.22 (2.31)
nmrshiftdb2	0.08 (0.11)	1.81 (3.35)	1.48 (2.23)	1.35 (3.11)	0.21 (1.40)	2.98 (3.95)	2.25 (3.93)	0.97 (4.6)	0.00 (0.01)	0.03 (0.03)	0.07 (0.07)

	atlas	patents	fdasrs	surechembl	pharmgkb	hmdb	selleck	pubchem_tpharma	pubchem	mcule	nmrshiftdb2
chembl	82.84 (91.11)	7.82 (9.45)	42.05 (50.13)	1.91 (2.39)	86.94 (95.63)	10.68 (15.5)	82.59 (93.46)	9.35 (11.25)	2.6 (3.07)	5.45 (6.42)	27.37 (35.23)
drugbank	36.09 (50.84)	0.46 (0.61)	5.05 (8.11)	0.03 (0.04)	75.49 (90.39)	4.32 (5.21)	24.39 (45.92)	0.09 (0.13)	0.01 (0.02)	0.02 (0.03)	2.81 (5.32)
pdb	33.43 (46.70)	0.57 (0.76)	6.60 (9.08)	0.05 (0.08)	25.73 (39.27)	3.17 (4.37)	20.84 (33.71)	0.19 (0.26)	0.03 (0.04)	0.03 (0.04)	5.99 (8.93)
iuphar	9.91 (15.00)	0.12 (0.16)	1.17 (2.20)	0.01 (0.01)	15.14 (21.44)	0.63 (0.94)	4.39 (10.85)	0.03 (0.04)	0.00 (0.00)	0.00 (0.01)	0.54 (1.25)
pubchem_dotf	17.89 (31.08)	0.30 (0.59)	6.79 (10.4)	0.03 (0.04)	24.75 (49.87)	1.30 (2.52)	36.03 (47.67)	0.11 (0.15)	0.01 (0.01)	0.01 (0.01)	0.29 (2.01)
kegg_ligand	50.14 (63.39)	0.95 (1.22)	13.64 (18.84)	0.05 (0.07)	58.00 (71.07)	7.47 (10.28)	23.54 (39.57)	0.20 (0.26)	0.03 (0.03)	0.03 (0.05)	10.22 (12.92)
chebi	68.93 (80.39)	1.27 (1.41)	16.37 (18.83)	0.06 (0.08)	73.15 (83.16)	9.39 (12.27)	38.51 (50.18)	0.31 (0.33)	0.05 (0.05)	0.05 (0.06)	14.99 (18.03)
nih_ncc	11.39 (22.81)	0.08 (0.15)	1.24 (2.18)	0.00 (0.01)	20.56 (40.52)	0.66 (1.35)	16.08 (22.73)	0.02 (0.02)	0.00 (0.00)	0.01 (0.01)	0.17 (0.84)
zinc	34.91 (72.74)	6.32 (10.94)	26.85 (48.34)	1.55 (2.78)	34.48 (83.29)	6.76 (15.34)	37.24 (80.61)	2.93 (4.86)	18.12 (21.92)	53.31 (86.67)	24.1 (35.25)
emolecules	75.14 (89.73)	8.45 (9.78)	42.85 (52.78)	1.83 (2.26)	76.35 (94.63)	12.66 (17.06)	81.16 (88.98)	3.30 (3.98)	9.83 (12.4)	65.52 (77.51)	35.45 (40.06)
ibm	51.92 (67.68)	79.78 (82.26)	35.72 (55.33)	12.46 (15.54)	70.93 (81.42)	12.10 (16.43)	28.51 (53.50)	4.39 (5.9)	4.01 (5.17)	0.91 (1.05)	39.71 (44.92)
atlas	100 (100)	0.07 (0.10)	1.18 (1.79)	0.00 (0.01)	18.47 (22.31)	0.67 (0.89)	7.67 (12.63)	0.01 (0.02)	0.00 (0.00)	0.00 (0.01)	1.03 (1.78)
patents	43.04 (58.19)	100 (100)	27.53 (43.31)	2.26 (2.80)	71.30 (79.3)	9.26 (12.48)	29.41 (53.72)	1.66 (2.22)	0.80 (1.00)	0.31 (0.36)	18.98 (22.69)
fdasrs	58.72 (71.51)	2.29 (2.85)	100 (100)	0.14 (0.17)	65.51 (81.29)	9.47 (12.12)	63.54 (75.20)	0.57 (0.57)	0.06 (0.06)	0.09 (0.11)	20.06 (24.39)
surechembl	55.17 (84.22)	68.97 (74.61)	51.11 (70.68)	100 (100)	72.90 (92.39)	13.33 (19.32)	80.37 (94.70)	31.42 (36.66)	19.23 (21.72)	2.29 (2.7)	18.74 (41.74)
pharmgkb	22.18 (27.41)	0.14 (0.16)	1.57 (2.50)	0.00 (0.01)	100 (100)	1.52 (1.83)	10.89 (21.16)	0.02 (0.02)	0.00 (0.00)	0.01 (0.01)	0.73 (1.53)
hmdb	40.23 (54.05)	0.92 (1.24)	11.31 (18.33)	0.04 (0.07)	75.36 (90.14)	100 (100)	24.44 (46.22)	0.19 (0.31)	0.04 (0.06)	0.03 (0.05)	9.49 (13.22)
selleck	21.44 (35.83)	0.14 (0.25)	3.56 (5.33)	0.01 (0.02)	25.36 (48.87)	1.14 (2.17)	100 (100)	0.04 (0.05)	0.00 (0.00)	0.01 (0.01)	0.54 (2.09)
pubchem_tpharma	80.32 (90.96)	15.71 (19.34)	65.63 (75.92)	9.75 (11.99)	86.33 (97.88)	17.68 (26.89)	88.09 (97.08)	100 (100)	7.46 (8.86)	1.08 (1.41)	56.39 (70.21)
pubchem	91.42 (97.70)	96.6 (96.91)	92.61 (95.28)	76.13 (78.58)	96.92 (99.25)	43.3 (53.74)	95.02 (98.64)	95.16 (97.9)	100 (100)	94.19 (97.89)	88.98 (98.75)
mcule	37.57 (48.54)	4.50 (4.74)	15.26 (22.13)	1.10 (1.35)	45.56 (58.10)	5.12 (6.33)	26.29 (40.41)	1.68 (2.15)	11.48 (13.51)	100 (100)	18.44 (20.83)
nmrshiftdb2	6.21 (10.71)	0.19 (0.22)	2.43 (3.67)	0.01 (0.02)	3.69 (7.48)	0.96 (1.32)	1.16 (4.43)	0.06 (0.08)	0.01 (0.01)	0.01 (0.02)	100 (100)

compounds where the business rules for compound standardisation are different between databases. Obviously it has to be left up to the user to make their own assessment as to whether these differences are important in the context of their own research interests.

Conclusions

The number of chemical databases in the public domain and the number of chemical structures within them is now large and likely to continue increasing in future years, particularly as automated extraction methods such as image to structure software becomes more reliable. Is it now time to accept that however diligent database providers are, there will always be differences in structure representations and indeed some errors in the structures that cannot be fixed with a realistic level of resource? Should we therefore turn our attention to encouraging the use and development of tools that enable the mapping together of related compounds rather than concentrate our efforts on ever more curation?

Conflict of interest

The authors declare no conflict of interest

Acknowledgements

This work was supported by a Wellcome Trust Strategic Award for Chemogenomics (WT086151/Z/08/Z); European Molecular Biology Laboratory; European Commission/IMI IMI eTox Award (115002); European Commission FP7 BioMedBridges Award (284209); European Commission ESFRI EU-OpenScreen (261861).

References

[1] Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, et al. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 2012;17(21–22):1188–98.
[2] Pence HE, Williams A.. ChemSpider. An online chemical information resource. J Chem Educ 2010;87(11):1123–4.
[3] Williams AJ, Ekins S, Tkachenko V. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. Drug Discov Today 2012;17(13–14):685–701.
[4] Tiikkainen P, Bellis L, Light Y, Franke L. Estimating error rates in bioactivity databases. J Chem Inf Model 2013;53(10):2499–505.
[5] Fourches D, Muratov E, Tropsha A. Trust but verify on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 2010;50:1189–204.
[6] Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, et al. The NCGC Pharmaceutical Collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. Sci Transl Med 2011;3(80):ps16.
[7] USP dictionary of USAN and international drug names 2010. United States Pharmacopeial; 2010.
[8] Hersey A, Senger S, Overington JP. Open data for drug discovery: learning from the biological community. Future Med Chem 2012;4(15):1865–7.
[9] Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Comput Sci 1992;32(3):244–55.
[10] United States Adopted Names. http://www.ama-assn.org/ama/pub/physician-resources/medical-science/united-states-adopted-names-council.page.
[11] Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI – the worldwide chemical structure identifier standard. J Cheminform 2013;5:7.
[12] InChI FAQs from the InChI Trust. http://www.inchi-trust.org/fileadmin/user_upload/html/inchifaq/inchi-faq.html.
[13] Food and Drug Administration Substance Registration System Standard Operating Procedure Version 5c. http://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf.
[14] Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. Annual reports in computational chemistry, vol. 4. Elsevier; 2008. p. 217–40 [chapter 12].
[15] Karapetyan K, Batchelor C, Sharpe D, Tkachenko, Williams AJ. The Chemical Validation and Standardization Platform (CVSP). Large-scale automated validation of chemical structure datasets in preparation (private communication from AJ Williams, submitted to Journal of Cheminformatics).
[16] ChemAxon Standardiser. https://docs.chemaxon.com/display/standardizer/Home.
[17] Biovia (formerly Accelrys) Cheshire. http://accelrys.com/products/pdf/cheshire.pdf.
[18] Perspectives in drug discovery and design: tautomers and tautomerism.Martin YC, editor. J Comput Aided Mol Des 2010;24(6–7).
[19] Warr WA. Tautomerism in chemical information management systems. J Comput Aided Mol Des 2010;24(6–7):497–520.
[20] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res 2013;41(Database Issue):D456–63.
[21] Lipinski CA, Litterman NK, Southan C, Williams AJ, Clark AM, Ekins S. Parallel worlds of public and commercial bioactive chemistry data. J Med Chem 2014 [Dec 4 Epub ahead of print].
[22] Sitzmann M, Ihlenfeldt W, Nicklaus MC. Tautomerism in large databases. J Comput Aided Mol Des 2010;24:521–51.
[23] Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform 2013;5:3.
[24] Chambers J, Davies M, Gaulton A, Papadatos G, Hersey A, Overington JP. UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. J Cheminform 2014;6:43.
[25] Gaulton A, Bellis LJ, Patrícia Bento A, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(Database Issue):D1100–07.
[26] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 2011;39(Database Issue):D1035–41.
[27] Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 2012;52(7):1757–68.
[28] Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, et al. PDBe: protein data bank in Europe. Nucleic Acids Res 2014;42(Database Issue):D285–91.