DOI 10.1186/s13321-015-0083-5

Journal of Cheminformatics

DATABASE

Open Access

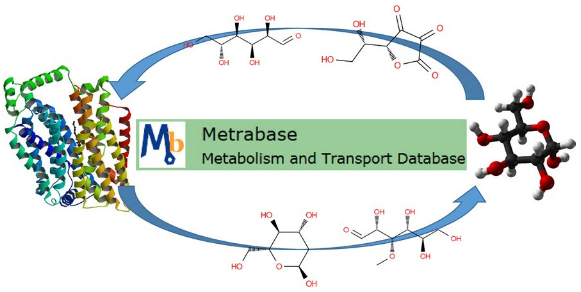
Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling

Lora Mak[1,2], David Marcus[1,2], Andrew Howlett[1], Galina Yarova[3], Guus Duchateau[4], Werner Klaffke[4,5], Andreas Bender[1] and Robert C Glen[1,6*]

Abstract

Both metabolism and transport are key elements defining the bioavailability and biological activity of molecules, i.e. their adverse and therapeutic effects. Structured and high quality experimental data stored in a suitable container, such as a relational database, facilitates easy computational processing and thus allows for high quality information/knowledge to be efficiently inferred by computational analyses. Our aim was to create a freely accessible database that would provide easy access to data describing interactions between proteins involved in transport and xenobiotic metabolism and their small molecule substrates and modulators. We present Metrabase, an integrated cheminformatics and bioinformatics resource containing curated data related to human transport and metabolism of chemical compounds. Its primary content includes over 11,500 interaction records involving nearly 3,500 small molecule substrates and modulators of transport proteins and, currently to a much smaller extent, cytochrome P450 enzymes. Data was manually extracted from the published literature and supplemented with data integrated from other available resources. Metrabase version 1.0 is freely available under a CC BY-SA 4.0 license at http://www-metrabase.ch.cam.ac.uk.

Graphical Abstract:



Keywords: Database, Substrate, Inhibitor, Inducer, Transport protein, Membrane protein, Metabolism, CYP, Drug transporter, Metabolite transporter

*Correspondence: rcg28@cam.ac.uk
[1] The Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
Full list of author information is available at the end of the article

Background

Absorption, distribution, metabolism and excretion (ADME) properties of small molecules are vitally important for their function and potential toxicity, hence understanding their interaction with transport proteins as well as metabolizing enzymes is fundamental to the discovery and development of e.g. safe, efficacious medicines and skin products. However, this data is largely dispersed across the literature, with fractions of the data available in current open and proprietary databases. Numerous valuable publicly available transporter-related data sources exist that are either specific to membrane transporters (e.g. TP-search [1], UCSF-FDA TransPortal [2], TSdb [3], Human Transporter Database [4], HMTD [5], TCDB [6], SLC Tables [7] and Membrane Proteins of Known 3D Structure [8]) or contain transporter data as part of a broader collection of biological and pharmacological data (e.g. ChEMBL [9], HMDB [10], DrugBank [11], Transformer [12], KEGG [13], Recon X [14], PharmGKB [15], UniProt [16], CTD [17] and TTD [18]), but most of these databases do not include the additional related data (e.g. tissue expression) and metadata required for in-depth cheminformatics analyses. Ligand-based studies, however, often remain the main or only option for modeling transport proteins as targets, since there are not many 3D structures of small molecule transporters available, especially for human proteins. Moreover, since these are very flexible proteins [19–21], analyses involving the dynamics of the multiple transporter systems is often complex and often beyond the scope of presently available simulation capabilities. Nonetheless, quick and convenient access to chemical structures is often missing in these resources and sometimes only chemical names are provided, e.g. UCSF-FDA TransPortal (focused on FDA-approved drugs) and Transformer (focused on biotransformations of xenobiotics). In QSAR modeling and other cheminformatics applications however, molecular properties are typically calculated using 1D, 2D or 3D representations of molecular structures and thus access to accurate molecular structures is required. Tools that convert chemical names into structures (e.g. NCI/CADD Chemical Identifier Resolver [22] or OPSIN [23]) can be used, but due to the complexity of chemical naming they may fail or derive incorrect/inconsistent structures, hence thorough manual checking of structures is highly beneficial. Those few relevant freely and easily accessible resources that do provide chemical structures for download, such as ChEMBL, HMDB and Drugbank, all have different aims and contents (collections of bioactivity data, metabolite data and drug data, respectively). Furthermore, their transporter substrate-related content and transporter specific metadata is in general limited. In a recent study that is most related to the current work, Sedykh et al. [24] collected and published a significant amount of data pertaining to substrates and modulators of transport proteins with the aim of building predictive models for several transporters. Our efforts overlap, extend and complement theirs. In a single resource, which is not limited to drugs and only positive results, we include cheminformatics-specific data and functionality and also protein tissue expression levels and transporter locations in tissues that are often missing in other comparable databases.

Here we describe Metrabase, in which we offer easily accessible data on small molecule transport and metabolism, publicly available and useful for computational analyses and modeling.

Both metabolism and transport are key elements defining the bioavailability and biological activity of molecules, i.e. their adverse and therapeutic effects. Membrane transport proteins are a large and diverse group of proteins that are responsible for transporting a very diverse collection of molecules, including ions and small molecules across biological membranes. This can be observed in a relatively low-throughput fashion (transporters or carriers, including ATP-powered pumps) and also in a high-throughput fashion, e.g. up to 10^8 water molecules or ions per second (channels) [25]. Metrabase includes a selection of efflux and influx transporters from the ABC [26] and SLC [7] families of transporters. They have attracted significant interest in recent years [27] and are also included in the revised guidance of the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) [28] for the approval of drugs. In the area of metabolism, Metrabase contains a limited set of the cytochrome P450 isozymes (CYPs) that have been selected as a major group of xenobiotic metabolizing enzymes [29]. Since drug absorption and excretion are influenced by both transport and metabolism, transporters and drug metabolizing enzymes can serve as a coordinated export system for toxic compounds and drugs, i.e. as a protective mechanism. Metrabase aims to provide comprehensive structural, physicochemical and biological data that can be used to infer the relationships between these transporters/enzymes and their ligands and its first release is presented here.

Construction and content

Metrabase v1.0 includes interaction data on 20 transporters and 13 CYPs. However, the major focus of this first version of the database is on transport proteins: specifically, on their interactions with small molecules that were experimentally found to be (or not to be) substrates. Data held in Metrabase was manually extracted from the published literature and supplemented with data integrated from other available resources (Figure 1). The

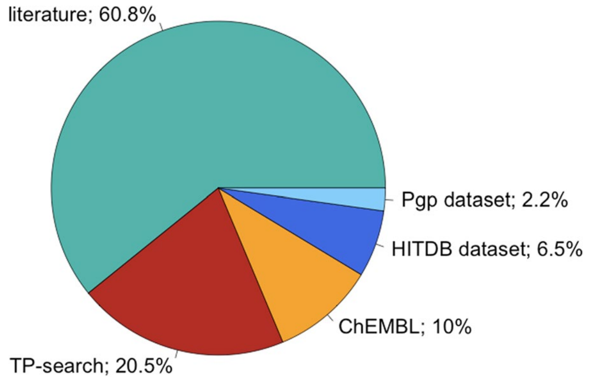


Figure 1 Fractions of Metrabase interaction records per data source. The total number of records is 11,649. 'Pgp dataset' and 'HITDB dataset' refer to the data taken from Poongavanam et al. [31] and Sedykh et al. [24] respectively. All the interaction data contained in Metrabase was extracted from the published literature: where datasource is 'literature', the data was extracted by members of the Metrabase Development Team.

literature search, carried out using resources and tools such as PubMed, SciFinder, Web of Knowledge, Google, Google Scholar and iHOP [30], primarily involved scanning for reports of transporter (non-)substrates. Protein synonyms given in the 'protein_synonyms' database table were used to ensure a comprehensive coverage of the search space where needed. If the search returned publications describing action types other than substrate or non-substrate, the data was or was not extracted at the curator's discretion. Manual data extraction was carried out by thirteen supervised undergraduate second and third year Natural Sciences students of the University of Cambridge (with specialization in chemistry and biochemistry). Each curator underwent induction including reading background reviews on transporters and metabolism and an explanation of the literature review process. The curation process involved reading scientific articles (abstracts and full texts) to ensure that the correct action type was assigned to the interacting small molecule and then transferring the data of interest into an electronic form. Two curators were involved in data extraction and verification from a single publication, each person performing one or the other task, but not both tasks for the same record. Special attention was given to the accuracy of chemical structures: three curators verified a single structure. Protein, compound, action type, publication and organism (to which the protein belongs) were extracted as a minimum, while the additional data (as described below) were extracted at a curator's discretion if the data was available in the publication.

A summary of the transporter-related Metrabase content is shown in Table 1. The principal content is associated with 13 of 20 transporters (MDR1, BCRP1, PEPT1, MRP1-4, OCT1, OATP1B1, OATP2B1, OATP1B3, OATP1A2 and ASBT), while the remaining 7 transporters (OATP3A1, OATP4A1, MCT1, OATP2A1, OSTα/OSTβ, GLUT1 and LAT1) have many fewer associated records (less than 20 substrates/non-substrates). This initial selection of transporters was mainly directed by the guidelines of the International Transporter Consortium [27]—membrane transporters considered to have a role in drug absorption and disposition (distribution and elimination), therapeutic efficacy and safety (i.e. toxicity) were included. The total number of transporter-related activity records is substantial, totaling 11,143 records. The CYP-related content of Metrabase v1.0 is currently, however, limited to only 506 records for all the 13 CYP isoenzymes (3A4, 2S1, 2W1, 4B1, 2D6, 2C19, 2C9, 2E1, 1A2, 1A1, 1B1, 2C8 and 3A5). This selection of CYPs includes the major drug metabolizing isoforms (about 95% of the drugs are affected by them) and the isoforms expressed in the skin. The latter category includes mainly CYPs for which there is very little relevant experimental data published. The total numbers of compounds interacting with the 20 transporters and 13 CYPs are 3,307 and 212 respectively; these were compiled from 1,211 publications. On the biological side, Metrabase v1.0 also includes 1,087 records of transporter tissue expression levels manually extracted from 66 publications. We include the expression data, because the expression levels influence the capacity of transporters under different physiological and pathophysiological conditions. If, for example, we want to increase drug absorption using a transporter, then not only its substrate specificity, but also expression profile needs to be considered. In drug development, transporters expressed in the intestine, liver, kidney and blood–brain barrier are of particular interest. The overall database structure is shown in Figure 2.

Activities

The key information held in the 'activities' table of the database covers the interactions between proteins and chemical compounds, indicating the compound action type as either substrate, non-substrate, inducer, non-inducer, repressor, inhibitor, non-inhibitor, stimulator or binder. In this work we define them as follows. Substrates are substances that are transported (by transporters) or catalyzed (by enzymes), while non-substrates are substances that were experimentally tested and found not to be substrates of a particular transporter or enzyme. Modulators (inducers, stimulators, inhibitors and repressors) are substances that alter the ability of a transporter to transport a substrate or alter the ability of an enzyme to metabolize a substrate. Inducers, non-inducers and repressors (associated with

Table 1 A summary of the transporter-related Metrabase content

Gene	Protein	Sub	nSub	−Mod	nInh	+Mod	nInd
ABCB1	MDR1	566	469	311	113	43	18
ABCG2	BCRP1	310	197	616	24	15	2
SLC15A1	PEPT1	247	95	275	26	7	–
ABCC2	MRP2	160	139	170	150	102	25
SLC22A1	OCT1	166	95	293	68	3	–
ABCC1	MRP1	98	91	4	4	–	–
SLCO1B1	OATP1B1	97	41	341	50	5	1
SLCO2B1	OATP2B1	48	75	136	203	23	–
SLCO1B3	OATP1B3	59	36	260	55	10	–
ABCC3	MRP3	66	24	47	4	17	5
SLCO1A2	OATP1A2	56	24	50	1	1	–
SLC10A2	ASBT	54	19	11	–	–	–
ABCC4	MRP4	47	19	–	–	–	–
SLCO3A1	OATP3A1	5	7	6	–	–	–
SLC16A1	MCT1	8	3	30	1	–	–
SLCO4A1	OATP4A1	7	1	4	–	–	–
SLC2A1	GLUT1	5	1	21	41	–	–
SLCO2A1	OATP2A1	5	1	9	3	–	–
SLC51A/B	OSTα/β	4	2	–	–	–	–
SLC7A5	LAT1	2	–	6	–	–	–

The figures quoted show the number of included compounds for their corresponding actions. Sub, nSub, −Mod, nInh, +Mod and nInd refer to substrate, non-substrate, negative modulators (inhibitor and repressor), non-inhibitor, positive modulators (stimulator and inducer) and non-inducer respectively.

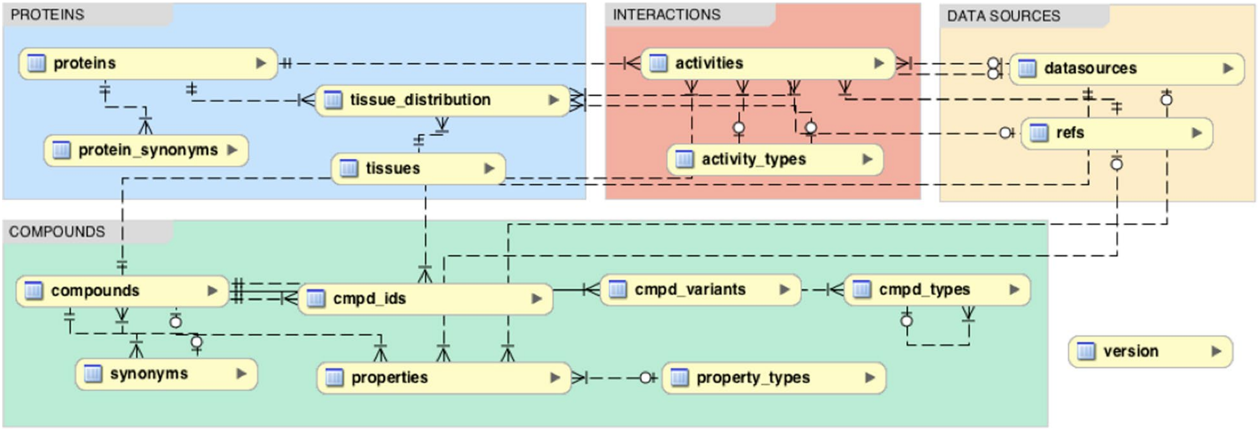


Figure 2 Metrabase schema. The database tables are grouped according to their context. There are four main tables: 'activities', 'compounds', 'proteins' and 'refs'. Chemical structures and proteins are stored in the 'compounds' and 'proteins' tables respectively. Interactions between chemical compounds and proteins are stored in the 'activities' table, while the 'refs' table contains the publications' citation information (bibliographic fields). An entity relationship diagram showing all the database fields is provided for download along with the database.

increased, unchanged and decreased levels of expression, respectively) belong to the compound action types related to protein expression, while inhibitors, non-inhibitors and stimulators (decreased, unchanged and increased protein activity, respectively) belong to the group of action types related to protein activity (Table 2). Care must be taken with respect to the current status of the inhibition records, since depending on the threshold used (e.g. percentage inhibition) some of the compounds annotated as 'inhibitors' can be regarded as 'non-inhibitors' and vice versa. This is an active area of research and is planned to be further resolved in subsequent releases of Metrabase.

Table 2 A list of the action types in Metrabase

	Action types
Protein activity (transport or catalysis)	Substrate Non-substrate
Compound activity (affecting protein activity/expression)	Inhibitor/repressor (negative modulators) Stimulator/inducer (positive modulators) Non-inhibitor/non-inducer (inactive compounds)

The 'activities' table was initialized with data taken from the TP-Search [1] and ChEMBL [9] databases, comprising 2,388 TP-Search records and 1,167 records holding action type annotations in ChEMBL versions 12 and 14. We also integrated the relevant data (1,009 records) from the recent publications of Sedykh et al. [24] and Poongavanam et al. [31]. Every interaction record is linked to the publication that provided the relevant data point. The current release holds only human-related records, but the importance of data related to orthologs is indisputable and therefore scheduled to be included in subsequent releases. There are five key activity fields: cmpd_id, protein_id, action_type, ref_id and species. Other 'activities' fields holding additional extracted data, such as assay descriptions, relevant experimental measurements, cell lines, compound concentrations and the substrates used in inhibition assays, may have only been partially completed in this release.

Compounds

The total number of compounds with recorded interaction data for both transporters and enzymes is 3,438. Their structures are available in MDL molfile format and as absolute SMILES strings (in Kekulé form). The standard InChI and InChI Key strings were computed using v1.04 of the InChI software [32]. The great majority of the compounds are small organic molecules (containing just the following atoms: C, H, O, P, S, N, F, Cl, Br and I) and other types (coordination complexes, inorganic compounds, metalloid-containing compounds, selenium-containing compounds and polymers) are listed in the 'cmpd_types' table. Annotations of compounds as other types and subtypes, such as 'natural product' and 'sesquiterpene', can also easily be added. Stereoisomers, different forms (e.g. a cyclic isomer of glucose) and multi-component structures (e.g. irinotecan hydrochloride; also including mixtures, such as ivermectin) are listed in the 'cmpd_variants' table. Where it was not clear which component(s) of a multi-component structure are responsible for the action, the multi-component structure was annotated as a mixture (here only the mixture has been tested and so the biological activity may reside in either one or all of its individual components). Chemical structures were verified using ChemSpider [33] and SciFinder [34] as "gold standards". The 'properties' table contains selected molecular properties, which were calculated/predicted for all structures (molecular mass) or just the small organic single-component structures (a selection of the constitutional descriptors, log P and log D) using ChemAxon's Calculator (*cxcalc*) [35]. Additional properties, either calculated/predicted or experimental, can easily be added (and defined in the 'property_types' table before the records can be inserted). The database is therefore also expandable by other users who may wish to insert their own data or add metadata to the records.

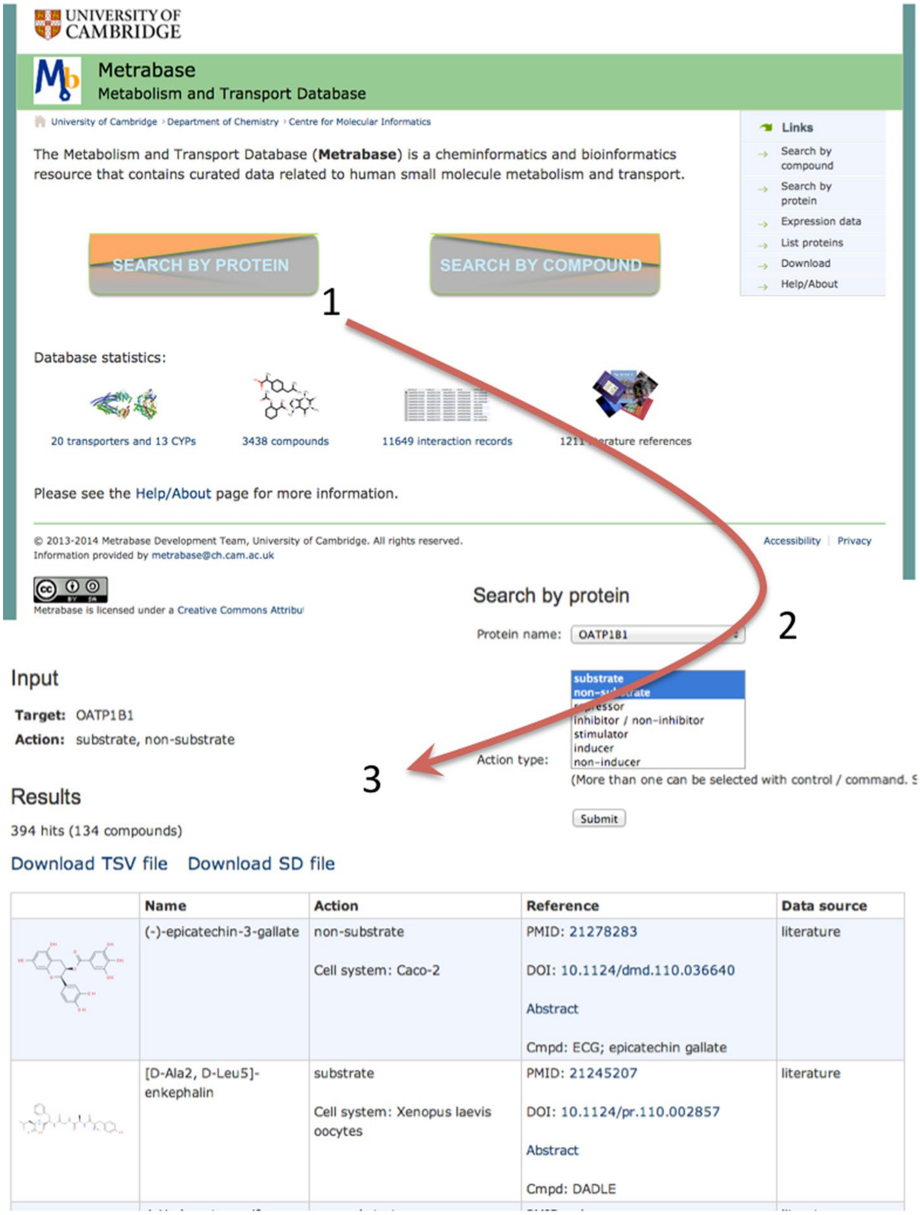
Proteins

The proteins contained in Metrabase are categorized as either transporters or enzymes and are provided with symbols and names approved by the HUGO Gene Nomenclature Committee (HGNC) [36], as well as UniProt [16] accession identifiers. Protein sequences for the indicated isoforms were included from UniProt. Metrabase also contains qualitative information about protein expression levels across healthy human tissues. Part of this data is based on immunohistochemistry using tissue microarrays and comes from the *normal_tissue.csv* file of the Human Protein Atlas (HPA) v9.0 [37]. All other expression records contain data extracted from the literature.

Utility and discussion

Database access

Metrabase is accessible via an online user interface at http://www-metrabase.ch.cam.ac.uk, and allows users to search the database by protein or compound. Search by protein is the main feature (Figure 3). It allows retrieval of all compounds interacting with the selected protein and shows their action types and links to the literature references from which the data originated. Search by compound allows name, structure, substructure and similarity searches. These are supported by the ChemDoodle 2D sketcher [38] for structure drawing and the Mychem cartridge (for MySQL) [39], which performs cheminformatics-specific functions. The similarity search employs the FP2 fingerprint of Open Babel [40] and the exact search compares InChI strings. Furthermore, users can retrieve tissue expression data and access information held in the UniProt and HGNC databases following the links provided in the list of the proteins. A database dump is also available for download enabling users to install their own local copy of the database. MySQL Workbench can be used as an interface: even though knowledge of SQL may not be required to display records of all tables and save them in other formats, such as TSV



The Metabolism and Transport Database (**Metrabase**) is a cheminformatics and bioinformatics resource that contains curated data related to human small molecule metabolism and transport.

Database statistics:

20 transporters and 13 CYPs

3438 compounds

11649 interaction records

1211 literature references

Please see the Help/About page for more information.

© 2013-2014 Metrabase Development Team, University of Cambridge. All rights reserved.
Information provided by metrabase@ch.cam.ac.uk

Metrabase is licensed under a Creative Commons Attribu

Search by protein

Protein name: OATP1B1

Action type: substrate, non-substrate, repressor, inhibitor / non-inhibitor, stimulator, inducer, non-inducer

(More than one can be selected with control / command. S

Input

Target: OATP1B1
Action: substrate, non-substrate

Results

394 hits (134 compounds)

	Name	Action	Reference	Data source
	(-)-epicatechin-3-gallate	non-substrate Cell system: Caco-2	PMID: 21278283 DOI: 10.1124/dmd.110.036640 Abstract Cmpd: ECG; epicatechin gallate	literature
	[D-Ala2, D-Leu5]-enkephalin	substrate Cell system: Xenopus laevis oocytes	PMID: 21245207 DOI: 10.1124/pr.110.002857 Abstract Cmpd: DADLE	literature

Figure 3 Search by protein example: retrieval of compounds tested for transport by OATP1B1. The Metrabase homepage is shown at the *top*. Results of searches are displayed in tables and can be downloaded in TSV (tab-separated values) and SD files.

or CSV, it is needed to fully exploit the database in this manner.

The transporter substrate dataset

We aim to provide a version of the transporter substrate dataset (MBTPsubDS) as a supplement to each Metrabase release (Figure 4). The first version of MBTPsubDS, based on Metrabase v1.0 and including datasets MBTPsubDS1_0 (Additional file 1) and MBTPsubDS1_0a (Additional file 2), was created using an in-house Python script and contains unique substrate and non-substrate records, which were processed to facilitate human transporter data analysis and predictive modeling. Mixtures, coordination complexes and other metal-containing compounds, inorganic compounds, polymers and metalloid-containing (including selenium-containing) compounds (as annotated in the 'cmpd_types' table) were removed. Only small molecules defined as those with a molecular mass of ≤1 kDa were retained. Stereoisomers were not considered (of course, this is an area for future

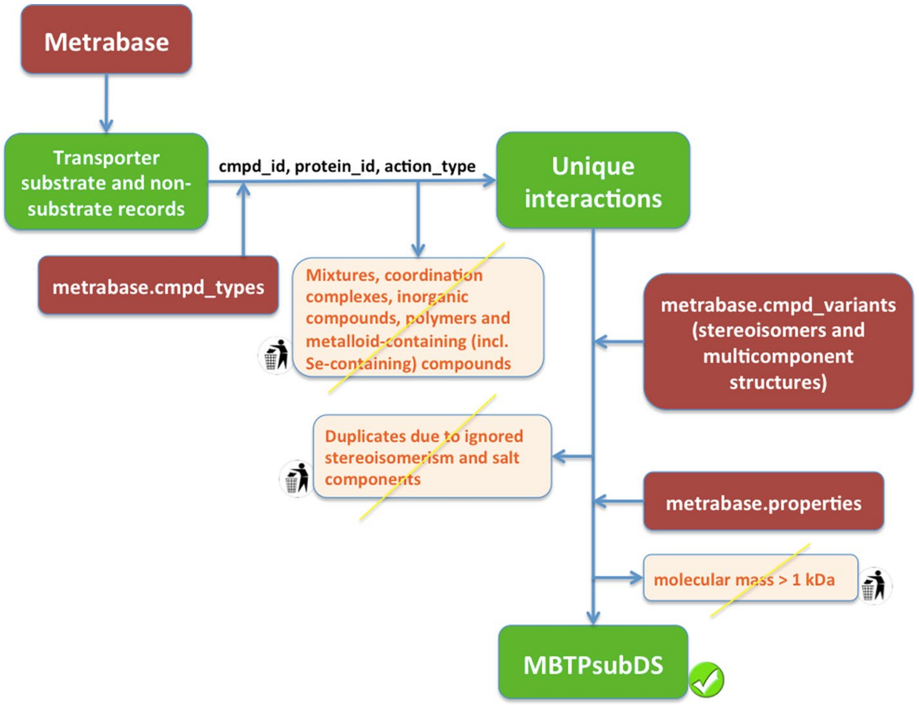

Metrabase
Transporter substrate and non-substrate records
cmpd_id, protein_id, action_type
Unique interactions
metrabase.cmpd_types
Mixtures, coordination complexes, inorganic compounds, polymers and metalloid-containing (incl. Se-containing) compounds
metrabase.cmpd_variants (stereoisomers and multicomponent structures)
Duplicates due to ignored stereoisomerism and salt components
metrabase.properties
molecular mass > 1 kDa
MBTPsubDS

Figure 4 A diagram depicting the steps in the creation of the transporter substrate dataset. The first version of MBTPsubDS was generated using an in-house Python script to process the data retrieved from Metrabase. Mixtures, metal-, metalloid- and selenium- containing compounds and polymers were identified using the metrabase.cmpd_types table and compounds with molecular mass >1 kDa were identified using the metrabase.properties table.

development as stereoisomers can have substantially different bioactivities) and all the remaining multi-component structures were replaced by their single-component counterparts as given in the 'cmpd_variants' table. Effectively, this corresponds to the usually applied procedure of keeping the largest component of each structure, except that this procedure is based on either the number of atoms or molecular mass and can thus lead to retention of unwanted components. Finally, all the duplicates formed due to ignored stereoisomerism and salt components were removed.

Datasets MBTPsubDS1_0 and MBTPsubDS1_0a contain 2,901 and 2,913 small molecule-transporter interaction records respectively. The latter dataset is a slightly extended version of the former dataset. All the interactions involving conflicting action types (where a compound was found to be both a substrate and a non-substrate of a single transporter) were removed from the former dataset, whereas the latter dataset contains a few of the compound-protein pairs corresponding to the resolved conflicting action types (where upon inspection we have defaulted to considering the compound as either a substrate or a non-substrate depending on the number of publications supporting each type). Compounds of the MBTPsubDS dataset make a diverse collection of molecular structures covering wide ranges of property values, however they are mostly lipophilic compounds, with 85% of the compounds possessing calculated log $P > 0$. Distributions of log P and several other selected molecular properties are shown in Figure 5.

Usage

Structured and high quality experimental data stored in a container, such as a relational database, can facilitate easy computational processing, thus allowing for high quality information/knowledge to be inferred by computational analyses. In cheminformatics and other disciplines dealing with small molecules, in particular the correctness of chemical structures is vital [41, 42]. As stated by Young et al. [41], small structural errors can lead to significant errors in predictions and so the predictivity of QSAR models could be substantially increased by thorough curation.

Data mining using resources like Metrabase can be employed to understand, for example, transporter function or the promiscuous nature of compounds binding to transporters competitively or non-competitively. Transporters can be highly promiscuous proteins transporting structurally quite diverse compounds across cell membranes at rates of hundreds to tens of thousands

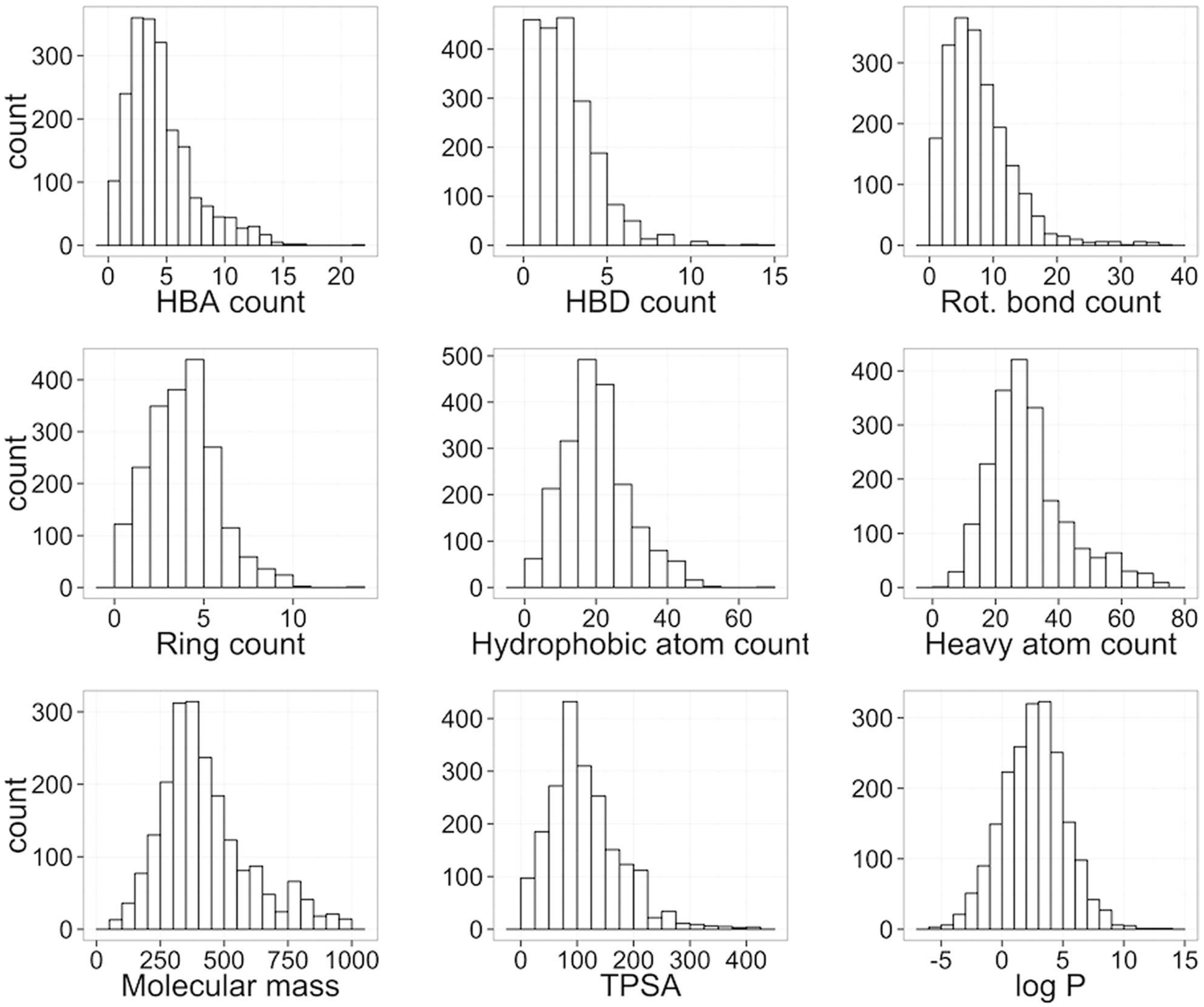
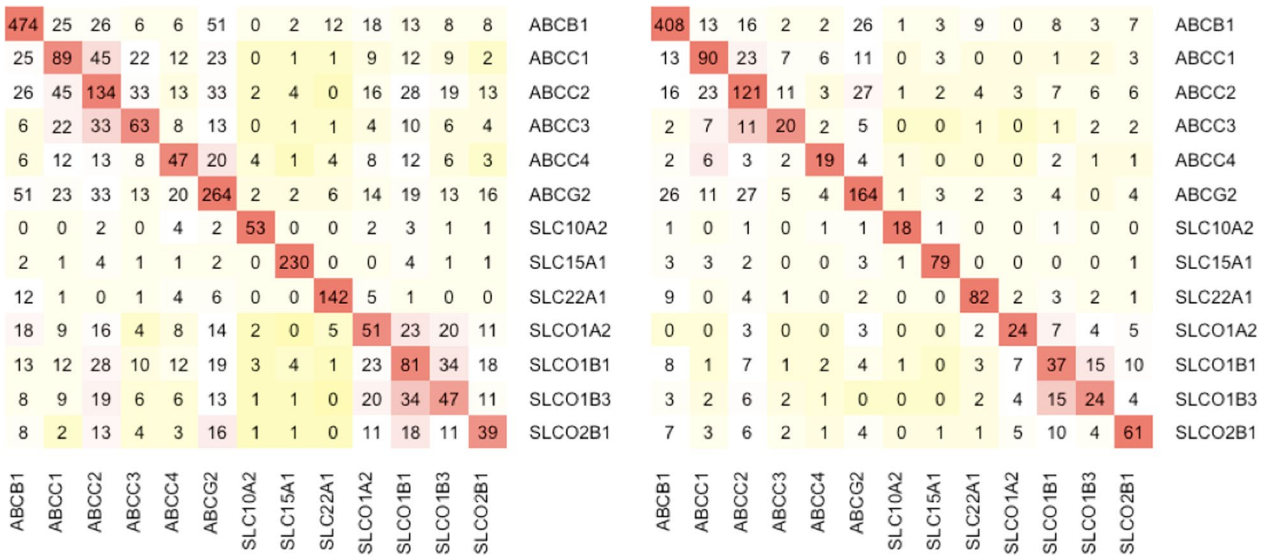


Figure 5 Histograms showing distributions of selected properties for all compounds contained in the MBTPsubDS1_0a dataset. HBA, HBD and TPSA stand for hydrogen bond acceptor, hydrogen bond donor and topological polar surface area respectively.

per second [25]. We can infer the relationships between transporters and their ligands, to see what kind of substrates they prefer and do not prefer, and why. The intended purpose of the database is also to facilitate in silico predictive modeling, e.g. tissue distribution of a new compound (and thus its toxicity potential assessed within the scope of the predicted tissue accumulation).

MBTPsubDS1_0 and MBTPsubDS1_0a datasets were created using the data contained in Metrabase v1.0 to allow a structural analysis of the transporter substrates and generation of transporter models that can predict whether or not a compound is a substrate of a range of efflux and uptake transporters (this work will be published elsewhere). All the required information (e.g. stereoisomers, metal complexes or multi-component structures linked to their single-component counterparts) is held in the database, and thus only a processing script written in a suitable programming language is sufficient to produce a version of MBTPsubDS.

The number of substrates (or non-substrates) shared between transporters is shown in Figure 6. It can be seen that in the current dataset there is not enough data to explore the extent to which similarities in transporter sequences determine similarities in compounds transported, since having a more complete data matrix would be required—a selection of compounds tested against a selection of transporters in an all-to-all fashion would be ideal. Furthermore, this dataset contains quite a diverse set of transporters with only OATP1B1 and OATP1B3 sharing a high sequence similarity/identity



	ABCB1	ABCC1	ABCC2	ABCC3	ABCC4	ABCG2	SLC10A2	SLC15A1	SLC22A1	SLCO1A2	SLCO1B1	SLCO1B3	SLCO2B1
ABCB1	474	25	26	6	6	51	0	2	12	18	13	8	8
ABCC1	25	89	45	22	12	23	0	1	1	9	12	9	2
ABCC2	26	45	134	33	13	33	2	4	0	16	28	19	13
ABCC3	6	22	33	63	8	13	0	1	1	4	10	6	4
ABCC4	6	12	13	8	47	20	4	1	4	8	12	6	3
ABCG2	51	23	33	13	20	264	2	2	6	14	19	13	16
SLC10A2	0	0	2	0	4	2	53	0	0	2	3	1	1
SLC15A1	2	1	4	1	1	2	0	230	0	0	4	1	1
SLC22A1	12	1	0	1	4	6	0	0	142	5	1	0	0
SLCO1A2	18	9	16	4	8	14	2	0	5	51	23	20	11
SLCO1B1	13	12	28	10	12	19	3	4	1	23	81	34	18
SLCO1B3	8	9	19	6	6	13	1	1	0	20	34	47	11
SLCO2B1	8	2	13	4	3	16	1	1	0	11	18	11	39

	ABCB1	ABCC1	ABCC2	ABCC3	ABCC4	ABCG2	SLC10A2	SLC15A1	SLC22A1	SLCO1A2	SLCO1B1	SLCO1B3	SLCO2B1
ABCB1	408	13	16	2	2	26	1	3	9	0	8	3	7
ABCC1	13	90	23	7	6	11	0	3	0	0	1	2	3
ABCC2	16	23	121	11	3	27	1	2	4	3	7	6	6
ABCC3	2	7	11	20	2	5	0	0	1	0	1	2	2
ABCC4	2	6	3	2	19	4	1	0	0	0	2	1	1
ABCG2	26	11	27	5	4	164	1	3	2	3	4	0	4
SLC10A2	1	0	1	0	1	1	18	1	0	0	1	0	0
SLC15A1	3	3	2	0	0	3	1	79	0	0	0	0	1
SLC22A1	9	0	4	1	0	2	0	0	82	2	3	2	1
SLCO1A2	0	0	3	0	0	3	0	0	2	24	7	4	5
SLCO1B1	8	1	7	1	2	4	1	0	3	7	37	15	10
SLCO1B3	3	2	6	2	1	0	0	0	2	4	15	24	4
SLCO2B1	7	3	6	2	1	4	0	1	1	5	10	4	61

Figure 6 The number of shared substrates and non-substrates between the 13 selected transporters. Transporters can be highly promiscuous with overlapping compound recognition patterns. These two matrices show the number of shared substrates (*left*) and non-substrates (*right*) within the MBTPsubDS1_0a dataset.

of 87.1%/78.5% (global pairwise sequence alignment was obtained using EMBOSS Stretcher v6.6.0 [43] with default parameters) followed by the ABCC subfamily (ABCC1 and ABCC3 have sequence similarity/identity of 73.4%/56.7%). We could perhaps perceive from Figure 6 initial support for the assumption that it is more likely that a substrate of SLCO1B1 will also be a substrate of SLCO1B3 than other transporters. However, with many data points missing, this could also be a consequence of experimental assays being carried out more often for transporters belonging to the same family (if more than one transporter was explored in the experiment).

The following two lists summarize the main intended uses of Metrabase (an example is shown in Figure 3):

- read across using similarity to compounds in the database,
- given a protein, show interacting compounds (including those where interaction was tested but not observed),
- given a compound, show interacting proteins (including those where interaction was tested but not observed),
- given a protein, list its tissue expression levels (qualitative only),
- given a tissue, list protein expression levels for this tissue (qualitative only);

and its benefits:

- action type is specified for all Metrabase compounds,
- negative action types (incl. inactive compounds) are included,
- chemical structures are included in the database, along with compound names,
- cell lines are specified for substrate records linked to primary literature references and we aim to expand this for all the 'activities' records,
- records are linked to the publications, which the data was extracted from (allowing thus for easier verification and for getting further information) and
- Metrabase is a publicly available and fully accessible resource.

Use case 1: identification of drug–drug interactions for atorvastatin

Atorvastatin (Lipitor) is a highly prescribed statin for the treatment of high blood cholesterol and for prevention of incidents that are associated with cardiovascular diseases. The mechanism of action of atorvastatin is inhibition of HMG-CoA reductase, and it is also known to be associated with other proteins, among them transporters [44]. In this use case we show how to identify the transporters with which atorvastatin is associated, and cross-reference them with potential inhibitors of these transporters, which could result in high accumulation of the drug and potential toxic drug–drug interactions by increased bioavailability of the drug [45].

Method

1. In the “Search by compound” section, run an exact search for “atorvastatin”.
2. The results show that atorvastatin was found to be a substrate of six transporters (MDR1, MCT1, OATP1B1, OATP1B3, OATP1A2, OATP2B1) and also an inhibitor of five transporters (MDR1, BCRP1, OATP1B1, OATP1B3, OATP2B1) with seven references indicating that it is an OATP1B1 substrate.
3. In “Search by protein” select “OATP1B1” and both “inhibitor” and “inducer”.
4. The results reveal a number of potential drug–drug interactions, among them several references to potentially strong inhibitors, such as cyclosporine.

The results of an exact search for cyclosporine (“Search by compound”) reveal a large number of references in support of cyclosporine being an inhibitor of several transporters, including 12 references for OATP1B1 inhibition, suggesting that the dose of atorvastatin should be lowered when co-administered with cyclosporine.

Use case 2: Exploring potential drug-food interactions with green tea

Green tea is known to contain high amounts of flavan-3-ols, including catechins of which (−)-epigallocatechin-gallate (EGCG) is the most abundant catechin [46]. In this use case we are employing the molecular structure of catechin to search for similar compounds with annotated transporter interactions. If found this might indicate potential adverse nutrient-drug interactions instigated by consumption of green tea coupled together with drug intake.

Method

1. In the “Search by compound” section run a similarity search using the SMILES representation of the catechin molecular structure, (c1cc(c(cc1[C@@H]2[C@H](Cc3c(cc(cc3O2)O)O)O)O)O), and Tanimoto coefficient (TC) cut-off of 0.3.
2. The results show 317 hits found in Metrabase that exhibit structural similarity to the catechin chemical structure. The hits include known catechins such as EGCG (TC = 0.823) and epicatechin (TC = 1.0), drugs such as nadolol (TC = 0.511) and troglitazone (TC = 0.457), and other natural products such as phlorizin (TC = 0.457).
3. Metrabase shows that EGCG has been found to be extensively associated with transporters: as a substrate of MRP1/2, OATP1B3 and OATP1A2, as a non-substrate of MDR1, OATP1B1, OATP2B1 and as an inhibitor of several OATPs such as OATP1A2, OATP1B1 and OATP2B1.
4. Nadolol is annotated as a substrate of MDR1 and OATP1A2 and as an inhibitor of BCRP1.
5. This potential interaction suggests that uptake of nadolol may be affected by co-administration with green tea and thus result in low plasma concentration due to the inhibition of its uptake transporter.

This nutrient-drug interaction was highlighted in a recent publication (not included in this version of Metrabase), which reported low intestinal uptake of nadolol by inhibition of OATP1A2-mediated uptake caused by green tea ingestion [47].

Future development plans

Further data curation, comprehensive data coverage and improvements are planned for subsequent Metrabase releases, as well as inclusion of additional biochemical annotations (especially with respect to sequence variants due to their importance in phenotypic variation, e.g. in drug metabolism). Moreover, we aim to include species other than human, as well as protein expression levels for diseased tissues, as this is highly relevant in the context of drug discovery. Even though the first release of Metrabase includes calculated molecular properties only, the ‘properties.type’ field values of ‘c’ and ‘e’ are proposed to designate calculated and experimental properties respectively, for possible future addition of experimental properties. The additional ‘activities’ fields, although incomplete as indicated in the Activities section above, were included in the first release and are planned to be completed in the subsequent releases. The selection of transporters, which was mainly steered by the guidelines of the International Transporter Consortium [27], is planned to be extended. We first aim to include additional data points for the 13 Metrabase CYPs before extending the protein list to include other Phase I, II and III enzymes involved in xenobiotic metabolism. Further improvements to the web interface can also be expected.

Conclusions

Metrabase offers structured and freely accessible manually extracted data on interactions between transport and metabolism related proteins and chemical compounds (the first version’s emphasis being on the transport proteins). It provides not only actions and measured activities, but also chemical structural information, tissue expression data and negative action types, which are essential in modeling activity. In particular, by creating accessible data that can be used in e.g. biochemistry, pharmacology and toxicology, we hope diverse research communities will find Metrabase useful and valuable. The availability of the raw data in an easily accessible form will allow other research workers in the field to easily include this data in collections of transporter/metabolism datasets as well as allow easy integration with other programs and services.

Availability and requirements

Metrabase is accessible via an online user interface at http://www-metrabase.ch.cam.ac.uk. Users can also install their own local copy of the database, for which a MySQL server installation is required. The Download section provides access to a MySQL dump of Metrabase, its schema, a user manual providing comprehensive information and versions of the transporter substrate dataset, i.e. MBTPsubDS. Metrabase version 1.0 is freely available under a Creative Commons Attribution-Share-Alike 4.0 International license. However, the integrated data retains the licensing of the original data sources. The TP-Search and ChEMBL records may have been modified and augmented based on literature revision or reinterpretation, while the HPA records were included unmodified. The 'datasource id' and 'datasource version' fields indicate the source of each relevant Metrabase record.

Additional files

Additional file 1: The transporter substrate dataset. The Additional file 1 contains data for compounds that were found to be (or nor to be) substrates of transporters (all records involving the conflicting action types were excluded; MBTPsubDS1_0.csv).

Additional file 2: The transporter substrate dataset. The Additional file 2 contains data for compounds that were found to be (or nor to be) substrates of transporters (MBTPsubDS1_0a.csv).

Abbreviations

ABC: ATP-binding cassette; ADME: absorption, distribution, metabolism and excretion; ATP: adenosine triphosphate; CSV: comma-separated values; CYP: cytochrome P450; EGCG: (−)-epigallocatechin-gallate; EMA: European Medicines Agency; FDA: US Food and Drug Administration; HGNC: HUGO Gene Nomenclature Committee; HMG-CoA: 3-hydroxy-3-methylglutaryl-coenzyme A; HPA: Human Protein Atlas; InChI: IUPAC International Chemical Identifier; MBTPsubDS: Metrabase transporter substrate dataset; (Q)SAR: (quantitative) structure–activity relationship; SLC: solute carrier; SQL: structured query language; TSV: tab-separated values.

Authors' contribution

RCG and WK conceived the project. LM developed the database and carried out data integration and processing. LM and DM participated in data collection and curation. DM defined the use cases. AH developed the web interface. LM drafted the manuscript and DM, AB and RCG revised it critically. GY, GD, WK, AB and RCG acted as actively involved advisors throughout the project and carefully proofread the manuscript. All authors have read and approved the final manuscript.

Author details

[1] The Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. [2] European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [3] Unilever Research & Development, 40 Merritt Blvd, Trumbull, CT 06611, USA. [4] Unilever Research & Development, Olivier van Noortlaan, 3133 AT Vlaardingen, The Netherlands. [5] Haus der Technik e.V., Hollestrasse 1, 45127 Essen, Germany. [6] Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, SW7 2AZ London, UK.

Acknowledgements

We thank Unilever for funding and Eisai Ltd. for their contribution toward summer student funding. We thank Claire Dickson, Joseph Dixon, Ivan Lam, Richard Lewis, Callum Picken, Claudia Pop, Heyao Shi, Emma Stirk, Yasmin Surani, Paddy Szeto, Nathaniel Wand, Julian Willis and Jing Xiangyi (at the time 2nd and 3rd year undergraduate students of the University of Cambridge) for their contributions toward the Metrabase development during the summers of 2011–2013. We also thank the authors of the TP-Search, Human Protein Atlas and ChEMBL databases for their permission to distribute their data as part of Metrabase. We thank Louise Birch and Luis Castro (Eisai Ltd.) for constructive discussions. We thank ChemAxon for an academic license.

Compliance with ethical guidelines

Competing interests

The authors declare that they have no competing interests.

Received: 8 March 2015 Accepted: 10 June 2015
Published online: 23 June 2015

References

1. Ozawa N, Shimizu T, Morita R, Yokono Y, Ochiai T, Munesada K et al (2004) Transporter database, TP-search: A web-accessible comprehensive database for research in pharmacokinetics of drugs. Pharm Res 21(11):2133–2134
2. Morrissey KM, Wen CC, Johns SJ, Zhang L, Huang SM, Giacomini KM (2012) The UCSF-FDA TransPortal: a public drug transporter database. Clin Pharmacol Ther 92(5):545–546
3. Zhao M, Chen Y, Qu D, Qu H (2011) TSdb: a database of transporter substrates linking metabolic pathways and transporter systems on a genome scale via their shared substrates. Sci China Life Sci 54(1):60–64
4. Ye AY, Liu QR, Li CY, Zhao M, Qu H (2014) Human transporter database: comprehensive knowledge and discovery tools in the human transporter genes. PLoS One 9(2):e88883
5. Yan Q, Sadée W (2000) Human membrane transporter database: a web-accessible relational database for drug transport studies and pharmacogenomics. AAPS PharmSci 2(3):E20
6. Saier MH, Reddy VS, Tamang DG, Vastermark A (2014) The transporter classification database. Nucl Acids Res 42(D1):D251–D258
7. Hediger MA, Clemencon B, Burrier RE, Bruford EA (2013) The ABCs of membrane transporters in health and disease (SLC series): introduction. Mol Aspects Med 34(2–3):95–107
8. White SH (2004) The progress of membrane protein structure determination. Protein Sci 13(7):1948–1949
9. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A et al (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucl Acids Res 40:D1100–D1107
10. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y et al (2013) HMDB 3.0—the human metabolome database in 2013. Nucl Acids Res 41(D1):D801–D807
11. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y et al (2014) DrugBank 4.0: shedding new light on drug metabolism. Nucl Acids Res 42(D1):D1091–D1097
12. Hoffmann MF, Preissner SC, Nickel J, Dunkel M, Preissner R, Preissner S (2013) The transformer database: biotransformation of xenobiotics. Nucl Acids Res 42(Database issue):D1113–D1117
13. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucl Acids Res 42(D1):D199–D205
14. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK et al (2013) A community-driven global reconstruction of human metabolism. Nat Biotech 31(5):419–425
15. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF et al (2012) Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 92(4):414–417

16. The UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucl Acids Res 41:D43–D47
17. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Saraceni-Richards C et al (2013) The comparative toxicogenomics database: update 2013. Nucl Acids Res 41(D1):D1104–D1114
18. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X et al (2009) Update of TTD: therapeutic target database. Nucl Acids Res 38(suppl 1):D787–D791
19. Linton K, Higgins C (2007) Structure and function of ABC transporters: the ATP switch provides flexible control. Pflug Archiv Eur J Phys 453(5):555–567
20. Dahl SG, Sylte I, Ravna AW (2004) Structures and models of transporter proteins. J Pharmacol Exp Ther 309(3):853–860
21. Ravna A, Sylte I (2012) Homology modeling of transporter proteins (Carriers and Ion Channels). In: Orry AJW, Abagyan R (eds) Homology Modeling, vol 857. Humana Press, pp 281–299
22. NCI/CADD chemical identifier resolver. http://cactus.nci.nih.gov/chemical/structure
23. Lowe DM, Corbett PT, Murray-Rust P, Glen RC (2011) Chemical name to structure: OPSIN, an open source solution. J Chem Inf Model 51(3):739–753
24. Sedykh A, Fourches D, Duan J, Hucke O, Garneau M, Zhu H et al (2013) Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. Pharm Res 30(4):996–1007
25. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J (2000) Molecular cell biology. W. H. Freeman, New York
26. Rees DC, Johnson E, Lewinson O (2009) ABC transporters: the power to change. Nat Rev Mol Cell Biol 10(3):218–227
27. Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KLR, Chu X et al (2010) Membrane transporters in drug development. Nat Rev Drug Discov 9(3):215–236
28. Maeda K, Sugiyama Y (2013) Transporter biology in drug approval: regulatory aspects. Mol Aspects Med 34(2–3):711–718
29. Ortiz de Montellano PR (ed) (2005) Cytochrome P450: structure, mechanism, and biochemistry, 3rd edn. Kluwer Academic, Plenum, NewYork
30. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36:664
31. Poongavanam V, Haider N, Ecker GF (2012) Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors. Bioorg Med Chem 20(18):5388–5395
32. The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi, http://www.inchi-trust.org
33. ChemSpider, Royal Society of Chemistry. http://www.chemspider.com
34. SciFinder, Chemical Abstracts Service: Columbus, OH. https://scifinder.cas.org
35. Marvin 6.1.3, ChemAxon. http://www.chemaxon.com
36. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA (2013) Genenames.org: the HGNC resources in 2013. Nucl Acids Res 41(D1):D545–D552
37. Uhlen M, Bjoerling E, Agaton C, Szigyarto CAK, Amini B, Andersen E et al (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. Mol Cell Proteomics 4(12):1920–1932
38. ChemDoodle Web Components, iChemLabs LLC. http://web.chemdoodle.com
39. Mychem 0.9.1. http://mychem.sourceforge.net
40. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminf 3:33
41. Young D, Martin T, Venkatapathy R, Harten P (2008) Are the chemical structures in your QSAR correct? QSAR Comb Sci 27(11–12):1337–1345
42. Fourches D, Muratov E, Tropsha A (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50(7):1189–1204
43. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16(6):276–277
44. Neuvonen PJ, Niemi M, Backman JT (2006) Drug interactions with lipid-lowering drugs: mechanisms and clinical relevance. Clin Pharmacol Ther 80(6):565–581
45. Mueller F, Fromm MF (2011) Transporter-mediated drug–drug interactions. Pharmacogenomics 12(7):1017–1037
46. Naldi M, Fiori J, Gotti R, Periat A, Veuthey J-L, Guillarme D et al (2014) UHPLC determination of catechins for the quality control of green tea. J Pharm Biomed Anal 88:307–314
47. Misaka S, Yatabe J, Mueller F, Takano K, Kawabe K, Glaeser H et al (2014) Green tea ingestion greatly reduces plasma concentrations of nadolol in healthy subjects. Clin Pharmacol Ther 95(4):432–438