The Reactive Metabolite Target Protein Database Archive

Robert P. Hanzlik, Yakov M. Koen and Matthew J. Garrett

Department of Medicinal Chemistry, University of Kansas
2034 Becker Drive, Lawrence, KS 66047

Abstract

The biotransformation of drugs and xenobiotic chemicals in living cells sometimes generates chemically reactive metabolites (CRMs) that diffuse through the cell and react covalently with cellular constituents, especially proteins. In some cases this covalent binding may impair the function of the protein(s) leading to cellular injury and even cell death. The elucidation of the chemistry involved in the bioactivation, the details of the ensuing protein modification, and the connection between protein binding and toxicity, became a major pursuit in a number of laboratories starting in the early 1970s. An attempt to collect all the relevant data about the proteins targeted by CRMs led to the construction of the Reactive Metabolite Target Protein Database (TPDB) at the University of Kansas. This curated, searchable database became freely available online in January 2006, and was maintained online until July 2020, when its maintenance ceased. This manuscript describes the nature and extent of the data that was compiled and the searches that could be implemented within the TPDB. The results of analyses of that data have been published elsewhere. All of the data originally compiled, including a list of all pertinent literature references, is now available in this manuscript and the associated Appendices at http://hdl.handle.net/1808/30592.

List of Appendices (Download all from http://hdl.handle.net/1808/30592)

1. Home page.pdf
2. Help page.pdf
3. Distribution of studies.xlsx
4. Chemical structures and abbreviations.pdf
5. List of hits and proteins.xlsx
6. List of protein "orthologs" (≥90% similar).xlsx
7. Commonality matrix, orthologs excluded.xlsx
8. Commonality matrix, orthologs included.xlsx
9. Targets proteins ranked by hit frequency.xlsx
10. Chemicals ranked by number of targets.xlsx
11. References cited in TPDB.xlsx

Introduction

Nature provides a great number of molecules that have potent effects on living cells. In many cases the effects are mediated by the chemical interacting with a specific enzyme, or a specific pharmacological receptor, to initiate significant changes in the biochemical composition and thus the physiological status of the cell. In parallel with chemicals expressing biological activity via these "specific" mechanisms, other chemicals, occasionally of natural origin but more commonly man made, can evoke equally profound biological change despite a lack of comparable specific cellular receptors or enzymes with which to interact. In a majority of these cases the mechanism seems to involve general chemical reactivity, either intrinsic to the molecule or arising as a result of its biotransformation (metabolism) within the cell or organism that it affects.

Many simple organic molecules that can injure or even kill living cells by virtue of undergoing conversion to a chemically-reactive metabolite (CRM). Such metabolites, often electrophilic in nature, may then express their reactivity toward various nucleophilic sites on important biological molecules, especially proteins. Such protein covalent binding (CVB), almost always detected using radioactivity in the parent molecule, may interfere with the normal functioning of the target protein, thereby harming the cell, possibly after an amplification event involving a signaling cascade. Because of its involvement in metabolizing drugs and xenobiotic chemicals, the liver is often a target for CRM-associated toxicity. Bromobenzene is an example of a simple organic chemical that is relatively innocuous chemically but nevertheless quite hepatotoxic in vivo. Direct evidence for the covalent modification of hepatocellular proteins by bromobenzene (BB) metabolites was first put forth in 1970, and a putative epoxide metabolite of bromobenzene was postulated to be a precursor for both its chemically stable metabolites and the accompanying CVB (i.e., arylation of liver proteins). Over the next several years the hepatotoxicity of the common analgesic acetaminophen (APAP) was also demonstrated to involve protein covalent binding concomitant with metabolic activation, in this case via a quinonimine intermediate.

As this putative mechanism for cytotoxicity was being generalized to other simple chemicals and drugs, questions arose as to the identities of the protein(s) that became covalently

modified, and the role of this adduction in the ensuing toxicity. Early studies using biochemical methods for protein identification gave way to faster methods based on mass spectroscopy, and the number of identified target proteins reported in the literature started to grow. To organize this growing stream of information the Hanzlik laboratory at the University of Kansas started a comprehensive listing of reactive metabolite target proteins using a spreadsheet, but it soon became clear that a database approach was needed. The result became known as the Reactive Metabolite Target Protein Database or simply the TPDB.

The TPDB was made available as a searchable online database in 2006. It eventually grew to contain over 500 individual proteins targeted by CRMs derived from one or more of 60 different drugs and chemicals. The rate of increase in the number of reported target proteins was greatest between 2007 and 2014, after which the rate slowed considerably as leading researchers shifted their attentions to other topics (or retired). During this time several types of global analyses of the data in the TPDB were carried out in attempts to find "important" target proteins related to cell killing. Ultimately, while some commonly targeted proteins were identified, no specific proteins appearing to be *causally* related to cell injury or death were clearly identified. By 2020 the value of maintaining the TPDB online had decreased to the point that it seemed that a static but published archive of all the information it contained would be sufficient for those interested. This article gives a brief description of the contents, structure and operation of the original online version of the TPDB, and a collection of appendices that constitute the archive of the information it contained.

The online TPDB

The Reactive Metabolite Target Protein Database was introduced in 2006 as a curated, searchable, online database listing mammalian proteins identified as being targets for covalent modification by chemically reactive metabolites (CRMs) of drugs and chemicals. Its most recent version was accessible publicly at https://targetprotein.ku.edu/index.php. The home page banner appeared as shown in Figure 1. In 2020 the entire website, other than the software that ran the internal searches, was archived and made available at http://hdl.handle.net/1808/30592. This article describes first how the original TPDB website was structured, what kinds of data were

available and how searches were run. It also describes the data contained in the archive and how it is presented in a series of Appendices.

TPDB home page

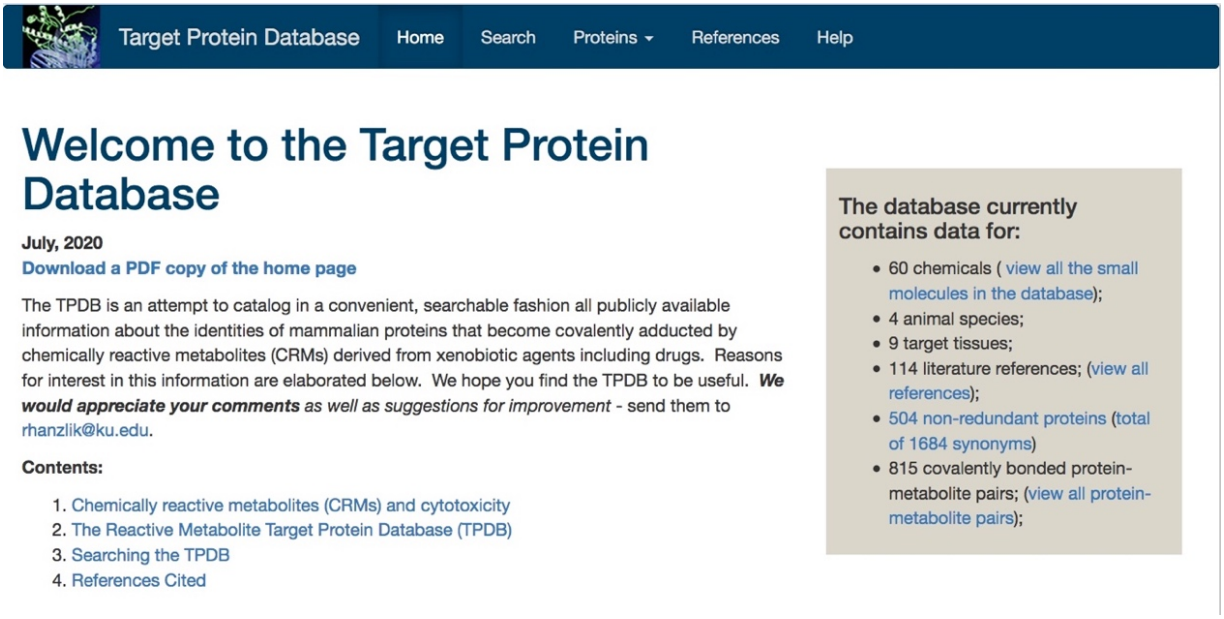

Figure 1. The TPDB home page banner, containing useful links to summary information, search functions and literature references. See Appendix for the complete home page including a discussion of topics 1-4 listed as contents.

The home page of the online TPDB began with the banner shown in **Figure 1**, but the information on the home page extended far below the banner. This information gave a narrative explanation of the history of the field of chemically reactive metabolites, the origins of the TPDB, guidance on how to search the TPDB, and a short list of references relevant to the narrative. The banner has links to the search pages, the reference list, and the Help page. As shown in the side bar of Figure 1, at the time this *archive* was created, the TPDB reported on 110 published studies of 60 different drugs or chemicals implemented across nine different tissues and two cellular systems derived from four species. It listed 815 individual protein-CRM adducts derived from 504 non-redundant proteins, most of which were known by multiple synonyms in the literature. In addition, 173 of the 504 targets had at least one ortholog with ≥90% sequence similarity found in a different tissue or species. This information was collected

from 114 literature references. Finally, a link to a Help page was provided. All of these sections are preserved in the 11 Appendices associated with this archive.

TPDB search page

The search page of the online TPDB is shown in **Figure 2**. This page provided a number of flexible but simple search options as well as several more sophisticated options. The Help page describes them in detail, but all are straightforward and intuitive. Search results were displayed as tables that could be printed or saved in Excel, CSV or PDF formats. For example, a search for all bromobenzene target proteins generated a list of 46 proteins (45 rat proteins plus one mouse protein), while a similar search for acetaminophen targets revealed 49 proteins (10 rat, 38 mouse and one human). The default search, in which each pull-down menu is set to any(all), produces a listing of all the protein hits listed in the TPDB.

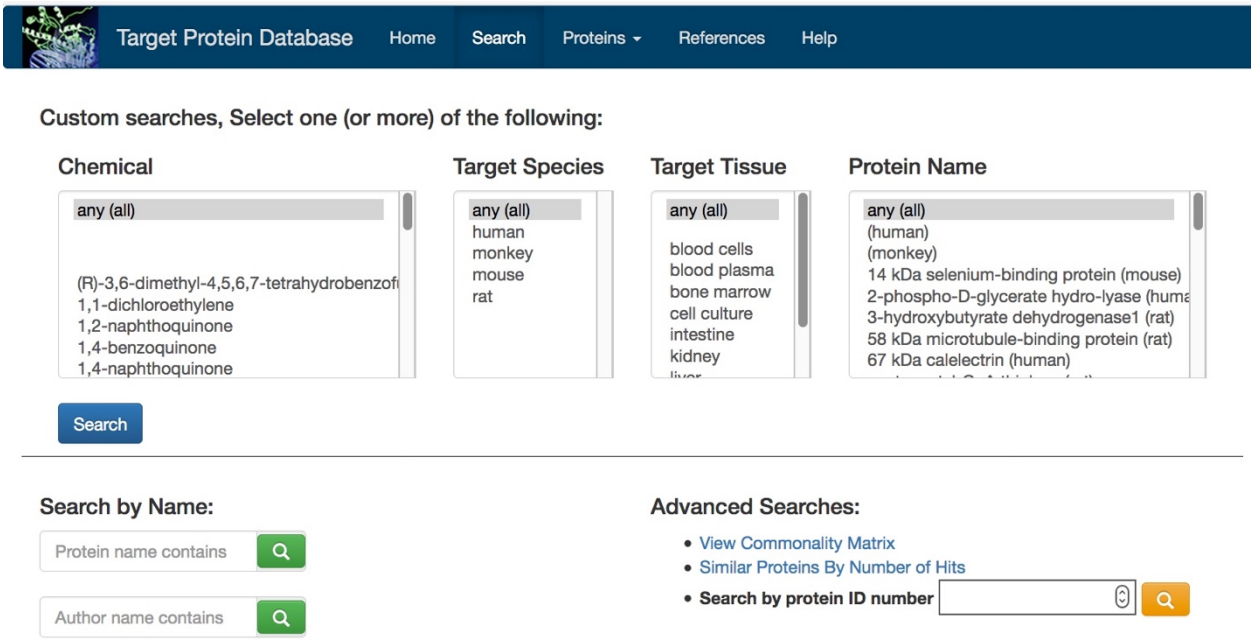

Figure 2. View of the TPDB search page.

Commonality among target proteins

As the number of target proteins reported began to increase over time it was natural to ask if certain proteins were frequent targets for multiple CRMs. Implicit in this question was the hope that one might find a common pathway or mechanism of action for cell killing by CRMs.

Our first attempt to ask this question systematically involved something called the Commonality Matrix, a portion of which is shown in **Figure 3**. In this matrix the rows and columns are labeled with the chemicals in the database, the diagonal elements in the matrix give the number of target proteins reported for a particular chemical, while the off-diagonal elements report the number of target proteins in common between any two chemicals. Each of the numbers in the table is a link that would generate a downloadable listing of the relevant proteins along with their MW, pI and links to their entries in SwissProt, NCBI and PDB.

A	1,2-NQ	1,4-NQ	4-BP	ABA	ACHDO	ACRM	AG	AMAP	AMOX	APAP	ATRZ	BB	BDIOL	BENZ
1,2-NQ	2													
1,4-NQ		8												
4-BP			14								2	9		
ABA				3	1				1	1				
ACHDO				1	4									
ACRM						5				1				
AG							1							
AMAP								7		6		1		1
AMOX				1					12	1				
APAP				1		1		6	1	49	1	3	1	3
ATRZ			2							1	9	5	1	
BB			9					1		3	5	46	1	
BDIOL										1	1	1	2	
BENZ								1		3				14

B	1,2-NQ	1,4-NQ	4-BP	ABA	ACHDO	ACRM	AG	AMAP	AMOX	APAP	ATRZ	BB	BDIOL	BENZ
1,2-NQ	2													
1,4-NQ		8	4					1		2	1	2		1
4-BP		4	14					2		4	2	9		1
ABA				3	1				1	1				
ACHDO				1	4									
ACRM						5				2				
AG							1							
AMAP		1	2					6		5		2		1
AMOX				1					12	1		2		1
APAP		2	4	1		2		5	1	48	3	10	1	4
ATRZ		1	2							3	9	5	1	1
BB		2	9					2	2	10	5	44	1	3
BDIOL										1	1	1	2	2
BENZ		1	1					1	1	4	1	3	2	14

Figure 3. A portion of the Commonality Matrix generated without ortholog consideration (A) and with ortholog consideration (B). See Appendix 4 for a listing of chemical structures and their abbreviations.

The partial commonality matrix shown in Figure 3A indicates that acetaminophen (APAP) and bromobenzene (BB) have 49 and 46 target proteins, respectively, and that these two chemicals have only three target proteins in common. The low number of common targets is an artifact of not considering information about protein orthologs between different species (a complete listing of which is given in Appendix 6). Rat proteins often have counterparts in the mouse which are not identical but may be closely similar (i.e., orthologous) based on sequence and function. Taking protein similarity into account, the partial commonality matrix shown in Figure 3B indicates 48 targets for APAP and 44 targets for BB, which is different from the number of targets revealed in Figure 3A or by the direct search using the pull-down menus (see above). These small discrepancies are due to the fact that among the 49 APAP targets listed there

is one pair of orthologs, while among the 46 BB targets listed there are two pairs of orthologs (one rat-rat and one rat-mouse).

The two forms of the commonality matrix also show differences in their off-diagonal elements. Not considering protein orthologs across different species (Fig. 3A), it appears that APAP and BB have only three protein targets in common, two in rat and one in mouse. However, taking rat-mouse orthologs into consideration (Fig. 3B) expands the list of common targets of APAP and BB from three to 11 (again including one pair of orthologs). The issue of orthologs among target proteins will become important again later (see below).

In either form, with or without taking orthologs into consideration, the commonality matrix analysis shows one other striking feature of the entire listing of CRM target proteins, and that is that a great many off-diagonal elements are either very small numbers or zero. There are several reasons for this. One is that for some chemicals the total number of known targets is itself quite small relative to other chemicals. This is especially true for earlier studies done before the adoption of 2D gels coupled with mass spectrometry. Another factor is the overall sensitivity of the identification method used; some CRMs result in lower levels of covalent binding, making it more difficult to detect their protein binding. Another factor is the scale and enzymatic activity of the biological system used to generate adducts. Exposures conducted in rats or mice generally yield adequate amounts of adducted tissue proteins to analyze, whereas studies done in tissue explants or cell culture generally are more limiting. Finally, different research groups have tended to have "favorite" chemicals to study, usually in relation to some specific toxicological question. Once they elucidated the target proteome of their favorite chemical, there was no systematic going back to look specifically for commonality of targeting, to "fill in the blanks" so to speak. Thus, the overall result is that while more than 800 acts of protein adduction involving more than 500 different target proteins are recorded in the TPDB, *we still know very little about overall protein adduction by CRMs,* and listing all the adducts still does not establish that any of them, per se, actually *cause* cytotoxicity (although it is widely believed that they do).

Global ranking of target proteins

With the recognition that a given target protein might exist within a set of orthologs across different species or tissues, a search more global than just the pair-wise commonality matrix was implemented in software. Combing through individual target proteins that were hit by multiple chemicals revealed some interesting findings not previously reported as such. For example, as expected, perhaps, based on its general properties as a "binding protein" plus its abundance and ease of isolation, human serum albumin and its rat and mouse counterparts (which are called serum albumin [precursor] but are not strictly orthologs in the sense of being ≥90% similar to HSA) collectively accrued 22 chemical hits. The next largest groupings of rat/mouse/human target orthologs included cellular thyroid hormone binding protein (also known as protein disulfide-isomerase [precursor]), and microsomal protein disulfide-isomerase A3, each with 13 hits. These highly susceptible ortholog groups are followed by one group hit by nine chemicals, seven groups hit by eight chemicals, three groups hit by seven chemicals, six groups hit by six chemicals, eight groups hit by five chemicals and ten groups hit by four chemicals. A ranking that includes all target proteins and sub-groupings thereof is given in Appendix 9.

Ranking of chemicals by number of targets reported

The total number of target proteins hit by any one CRM can be seen at a glance along the diagonal of the commonality matrix made *without* considering orthologs. The rank position of any one chemical in terms of its number of targets depends on its efficiency of conversion to a CRM, the efficiency with which the CRM is intercepted by small non-protein scavengers like glutathione, and other factors as discussed for the commonality matrix. After correcting for the occasional identification of the same target by different research groups it is found that just 17 chemicals hit a total of 635 targets (range 10 to 88, average 37.9), while the next 12 highest-ranked chemicals have just 88 known targets (range 5 to 9, average 7.7). The other 31 chemicals have just 4 or fewer known targets. A complete listing, corrected for multiple identifications of the same target by different research groups, but not corrected for orthologs, is provided in Appendix 10. The overlying caveat, of course, is that we don't know what we don't know about adduction events that are below current limits of detection.

Database Structure

Data for the TPDB was stored in a Mysql 5.7 database. The following is a description of each table and its data.

- Biological_system – Relational table with each of the biological systems available for interaction mapping.
- Cache – Simple key/value table used by the web application for quick caching of results. No effect on the core data.
- DisplayText – Dynamic text used by the web application that can be changed by the admin. No effect on the core data.
- Evidence – Relational table used to describe the evidence supporting each interaction in the database. Consists of a unique id, type, and method.
- Evidence_reference – Many-To-Many relational table that ties one or more evidence entries to one or more references.
- Id_mapping – Source to lookup a given molecule and find where it is listed in vendor databases, eg, chembridge, Prestwick, etc.
- Interaction – Core relational table that ties protein, molecule, evidence, biological system together, as well as indicating the tissue and whether or not its published.
- Molecule_table – Relational table that has all molecules used in the system.
- Protein – relational table that has all the proteins that are in the system, as well as the species to which they are tied.
- Reference – Relational table that lists all references for evidences and interactions.
- Similar – Relational table that lists each protein id and any other protein ids that are ≥90% similar to it.
- Synonyms – Relational table that lists synonym names for the proteins.