

PROCEEDINGS

Open Access

DetoxiProt: an integrated database for detoxification proteins

Zhen Yang^{1†}, Ying Yu^{2†}, Lei Yao¹, Guangui Li², Lin Wang², Yiyao Hu¹, Haibin Wei³, Li Wang⁴, Riadh Hammami⁵, Roxanne Razavi⁶, Yang Zhong^{7*}, Xufang Liang^{2*}

From Asia Pacific Bioinformatics Network (APBioNet) Tenth International Conference on Bioinformatics – First ISCB Asia Joint Conference 2011 (InCoB/ISCB-Asia 2011)
Kuala Lumpur, Malaysia. 30 November - 2 December 2011

Abstract

Background: Detoxification proteins are a class of proteins for degradation and/or elimination of endogenous and exogenous toxins or medicines, as well as reactive oxygen species (ROS) produced by these materials. Most of these proteins are generated as a response to the stimulation of toxins or medicines. They are essential for the clearance of harmful substances and for maintenance of physiological balance in organisms. Thus, it is important to collect and integrate information on detoxification proteins.

Results: To store, retrieve and analyze the information related to their features and functions, we developed the DetoxiProt, a comprehensive database for annotation of these proteins. This database provides detailed introductions about different classes of the detoxification proteins. Extensive annotations of these proteins, including sequences, structures, features, inducers, inhibitors, substrates, chromosomal location, functional domains as well as physiological-biochemical properties were generated. Furthermore, pre-computed BLAST results, multiple sequence alignments and evolutionary trees for detoxification proteins are also provided for evolutionary study of conserved function and pathways. The current version of DetoxiProt contains 5956 protein entries distributed in 628 organisms. An easy to use web interface was designed, so that annotations about each detoxification protein can be retrieved by browsing with a specific method or by searching with different criteria.

Conclusions: DetoxiProt provides an effective and efficient way of accessing the detoxification protein sequences and other high-quality information. This database would be a valuable source for toxicologists, pharmacologists and medicinal chemists. DetoxiProt database is freely available at <http://lifecenter.sgst.cn/detoxiprot/>.

Background

Living organisms are exposed to the external environment and are continuously confronted with a variety of chemical compounds and xenobiotics. Many of these naturally occurring substances or man-made chemicals can be deleterious when inhaled or absorbed, or even hazardous to organisms, (e.g. plant toxins to prevent herbivory). Artificial chemicals include anti-cancer

drugs, antibiotics produced by fungi, insecticide, herbicide and other environmental pollutants [1]. Some kinds of toxins can lead the cell to produce large amounts of reactive oxygen species (ROS), though in some cases, ROS can be natural by-products of metabolic processes. ROS can react with many cellular molecules and cause damage or contribute to the development of various pathologies [2]. Thus, self-protective mechanisms including catalytic biotransformation have evolved as an adaptation against various toxic chemical species. Cells possess various kinds of detoxification proteins that are capable of metabolizing a wide variety of toxins, mainly classified into Phase I and Phase II xenobiotic-metabolizing enzymes (XEMs) as well as antioxidant enzymes

* Correspondence: yangzhong@fudan.edu.cn; tliangxf@jnu.edu.cn

† Contributed equally

²College of Life Science and Technology, Jinan University, Guangzhou, 510632, China

⁷Institute of Biodiversity Science and Geobiology, Tibet University, Lhasa, 850000, China

Full list of author information is available at the end of the article

according to their functional mechanisms [3-5]. These proteins react directly with toxins to reduce their toxicity and enhance their solubility so that they can be easily eliminated from the cell.

Phase I reactions include oxidation, reduction and hydrolysis. A major function of the Phase I reaction is the transformation of substrates so that they can be easily eliminated or modified by reactive groups for subsequent Phase II reaction. The most important Phase I system is the cytochrome P450 enzyme (CYP). It is an important class of heme-containing monooxygenase that mainly catalyzes the oxidative reactions of the adjunction of an atom from molecular oxygen into a substrate. This superfamily can be functionally classified into three groups, with more than 50 protein members have been identified in humans [6]. The genetic association between CYP superfamily and various diseases has been identified, such as cancer, fatty liver disease and Parkinson's disease [7-9]. Phase II reactions include glutathione (GSH) conjugation, glucuronidation, sulfation, acetylation, and methylation. Small polar molecules, such as glutathione, UDP-glucuronic acid (UDPGA) and acetyl coenzyme A (AcCoA), can be conjugated with the toxic compounds [10-12]. Glutathione conjugation is the primary Phase II reaction. The glutathione S-transferase (GST) catalyzes the conjugation of the thiol group of GSH, the tripeptide gamma-Glu-Cys-Gly, with a wide variety of electrophiles. It has been widely demonstrated that the polymorphism of GSTs is related to cancer susceptibility and patient survival [13]. ROS produced by exogenous toxins or generated during the endogenous metabolic process can react with the oxidation enzymes and was eventually reduced by non-enzymatic antioxidant defenses, such as ascorbic acid (vitamin C), alpha-tocopherol (vitamin E) and GSH, or enzymatic antioxidant defenses, such as catalase (CAT), peroxidase (POD) and superoxide dismutase (SOD) [14,15]. Relatively high levels of oxygen in the brain make it one of the vulnerable organs to ROS damage. Many neurodegenerative diseases are associated with oxidative damage such as Parkinson's disease [16]. Thus, these proteins may constitute the major protective mechanism of the brain from damage due to oxidative stress [17].

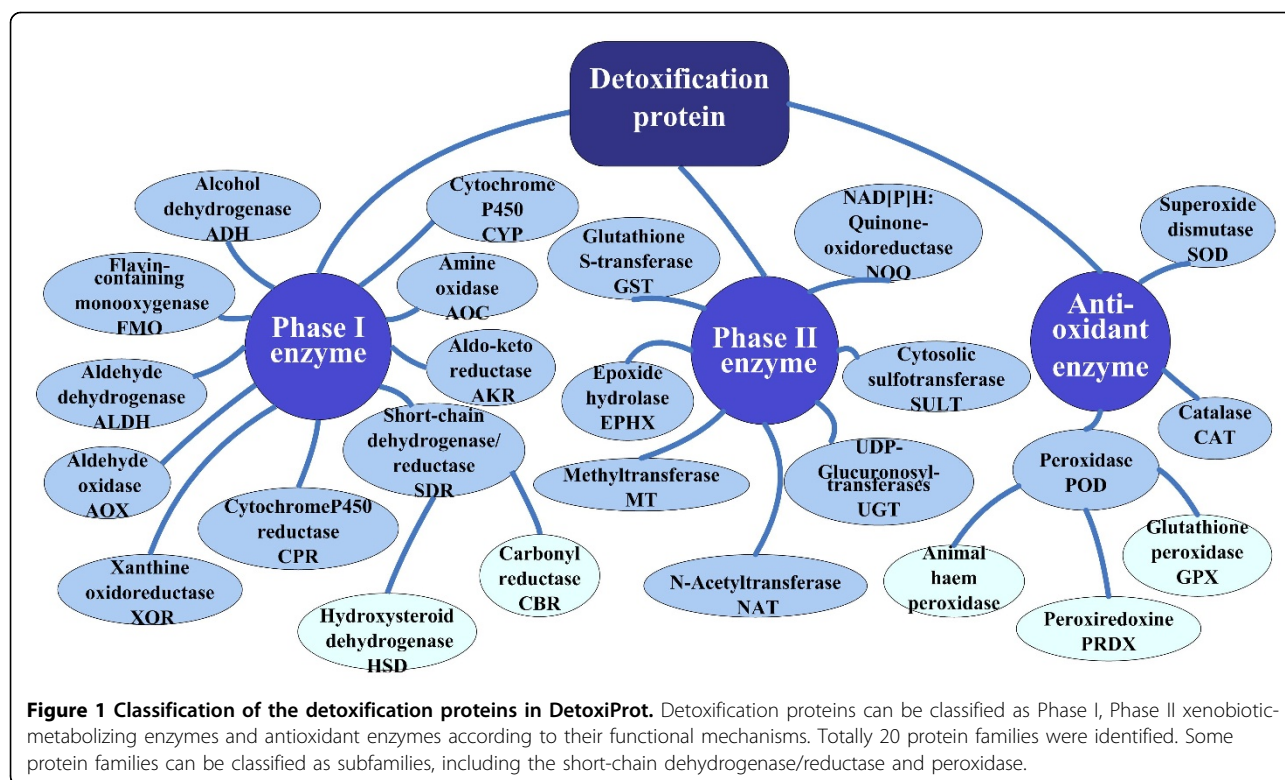
Detoxification proteins have been found from invertebrates to vertebrates. More than one hundred genes encoding these proteins have been identified in humans. Large efforts have been made to understand the structural, functional and evolutionary basis of detoxification proteins, as well as their roles in disease development [18-20]. Many detoxification protein-related databases have been established to provide this kind of information, such as PeroxisomeDB [21] and PeroxiBase [22]. These databases mainly focus on peroxidase related

proteins, which constitute only part of the steps in cellular detoxification. There is currently no system for the classification and annotation of all these proteins. Thus, the DetoxiProt database was created to facilitate the efficient annotation and acquisition of information about the detoxification proteins. The aim of the DetoxiProt is to provide a useful platform for detoxification protein researchers in the fields of physiology, pharmacology, toxicology, food security, farm chemical development and environmental pollution related research. It provides relevant information for the function and evolutionary analysis of these proteins. In addition to the chemical features of the genes, the DetoxiProt features 3D structures, sub-cellular location, tissue expression and especially, the inducer, inhibitor and substrates of these enzymes. Presently, the database contains 5956 entries encompassing 20 different detoxification protein families.

Construction and content

We extensively searched various public databases and related literature to collect information on detoxification proteins. Data collection was performed by the following approaches: we first performed a systemic search for relevant literature and determined the precise classification of these proteins. Accordingly, 20 protein families were identified (Figure 1). For each protein family, a summary description was compiled displaying the gene symbol name, InterPro database ID and the protein introduction [23]. Then, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [24], the UniProt Knowledgebase (UniProtKB) [25] and NCBI gene database were inquired to identify all the members of each family. Other public databases including Ethanol-Related Gene Resource (ERGR) [26], Aldo-Keto Reductase (AKR) Superfamily homepage [27] and PeroxiBase [22] were subsequently searched for further confirmation and supplementation. Then we queried through a list of the protein names as the keywords against SwissProt and NCBI protein database for collection of protein sequences [28]. For the completeness of the database, we used the BLAST embedded in NCBI to search homolog sequences. The final set of detoxification proteins were manually compiled, with a total of 5956 proteins from 628 organisms collected.

Extra information from other databases and literature, as well as the predicted protein features were incorporated into each detoxification record. Protein cofactors, tissue distribution and cellular location were retrieved from the references and publicly available databases such as NCBI and UniProtKB, while experimentally determined protein structures were retrieved from protein databank (PDB) [29]. DetoxiProt also provides the



predicted protein physiological-biochemical properties including the molecular weight, theoretical pI, aliphatic index and Grand average of hydropathicity (GRAVY), which computed by the ProtParam tool from the Expasy server (<http://www.expasy.org/tools/protpar-ref.html>). To provide fast protein classification, protein domains were predicted by various web servers including the Pfam [30], Prosite [31] and SMART [32]. External database links such as NCBI Protein, UNIPROT, KEGG etc, were also included.

Due to the importance of detoxification proteins for medicinal chemistry, we collected information from the literature the about the inducer, inhibitor and substrate for each protein from nine model organisms (*Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Pan troglodytes*, *Rattus norvegicus* and *Xenopus laevis*). The nomenclatures of the toxins were revised and unified according to MeSH and PubChem [28]. Thus, each entry was manually annotated to ensure data quality. A detailed strategy for protein information collection and data integration can be seen in Figure 2.

Database architecture

DetoxiProt was implemented as a relational database using MySQL. Five major tables were used to store data. Web interface was implemented using PHP language. The database is running on a Red Hat Enterprise Linux

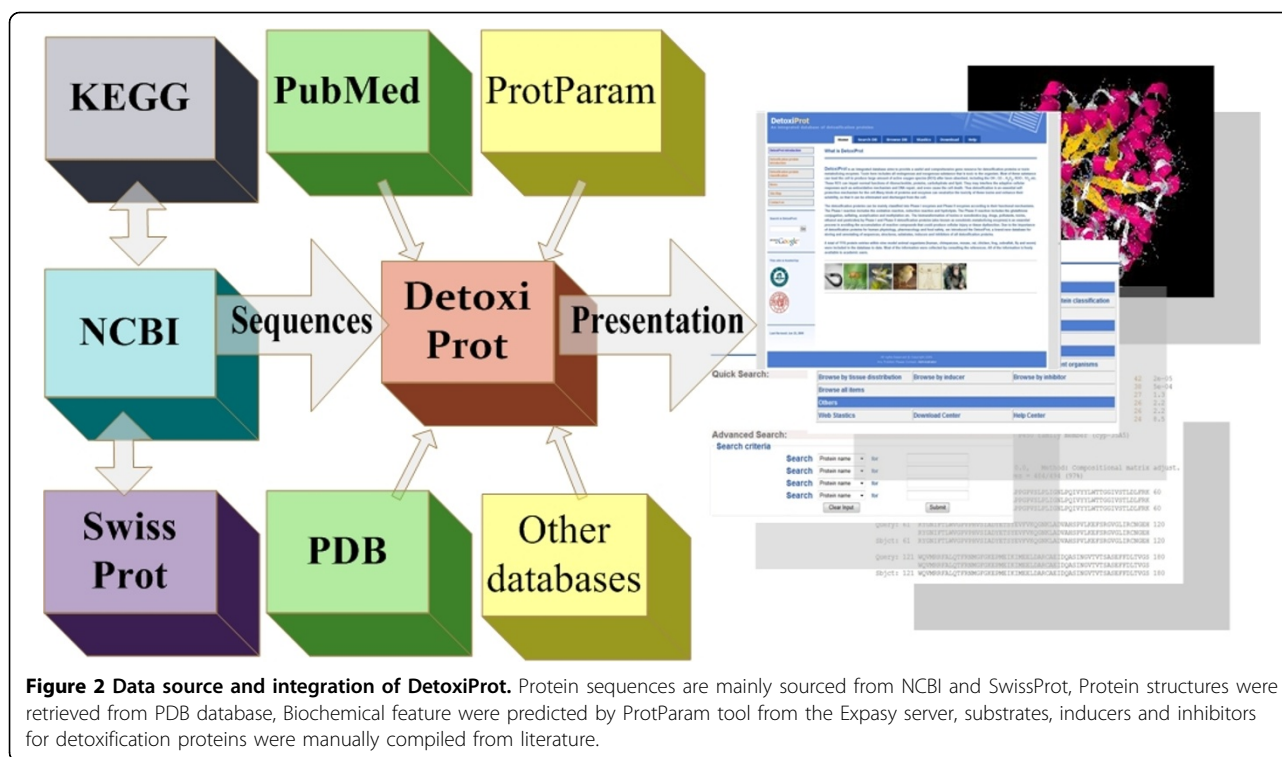
3 with Apache as HTTP server (please see additional file 1).

Utility

DetoxiProt provides a user-friendly interface allowing easy access to data. The “Home” page contains general information about the database and detoxification proteins. The “detoxification protein classification” page summarizes detailed information about the biological activity of each kind of protein. Other useful tools are also provided:

Basic and advanced searches

The “Search DB” page permits users to perform a quick search and advanced search in the database. Gene symbol name, NCBI gene ID, UNIPROT ID and PDB ID are supported for the quick search. Advanced search tool allow users to combine multiple criteria including protein name, organism, tissue distribution, sub-cellular location, etc. This method helps users to locate specific proteins of interest rapidly and accurately. For example, a medicinal chemist or a pharmacologist may be particularly interested in the specific human cell membrane proteins that are inhibited by some kind of toxin, so the options of protein name, organism, cellular location and inhibitor in each of the pull-down lists can be selected separately and simultaneously, and records limited by specific keywords will be displayed.



Peptide mapping and protein domain search

A tool of peptide mapping is designed for mapping the functional peptides to the entire sequence of detoxification protein. User defined peptide sequences can be exactly searched from protein sequences and the position information provided. This is helpful for understanding the distribution of the active peptides across entire proteins. Another useful tool to infer the functional diversity of detoxification protein is the protein domain search. The concept that conserved sequences among sequences may be a symbol of functionally similar protein domains underlies many classification methods of protein families. Thus we predict the domains within each protein by Pfam, Prosite and SMART web servers. A simple keyword search tool was provided to search database entries by domain or family name of interest to the user.

Structural similarity search

DetoxiProt provides a structural similarity search tool, the 3D-BLAST [33,34], to search detoxification proteins that have similar 3D-structures. This method allows for a quick and accurate search based on protein structure. It identifies 23 states of structural alphabet, which are used to represent pattern profiles of the protein backbone fragment. Protein structures are represented as structural alphabet sequence databases (SADB). Users may upload their own structure files in PDB format, after which the 3D-Blast program uses BLAST to search

for the longest common substructures, known as structural alphabet high-scoring segment pairs (SAHSP), between the structural alphabet sequence translated from query structure and that of other structures in the database. A statistical significance score (E-value) is calculated to indicate the reliability of the prediction of the alignment. Depiction of the results for different search methods is provided (please see additional file 2).

Browse database

Different browsing methods are implemented in the database to allow users to navigate by specific criteria. DetoxiProt provides seven different browsing methods: browse all data or specific dataset including the classification, species, tissue distribution, substrates, inducers and inhibitors. Cascading browse styles from specific dataset to protein list, then to detailed information are provided. The “detailed information” page includes general information, protein information, predicted protein features, protein structure, external database links and references. The 3D illustration of a protein structure may be viewed in interactive mode via the Jmol applet (<http://www.jmol.org/>) when the structure file is available. This tool would be particularly useful for protein structure investigation and drug development. While protein sequences may be individually download in FASTA format, the complete set of protein sequences may be retrieved from the “Download” page. Figure 3 shows an example of the detailed information on the “browse” page.

DetoxiProt
 An integrated database of detoxification proteins

Home
Search DB
Browse DB
Tools
Download
Help

Home

- DetoxiProt introduction
- Detoxification protein introduction
- Detoxification protein classification
- Home
- Site Map
- Contact us

Search DB

- Key Word Search
- Blast
- Peptide Mapping
- Protein Domain Search
- Structural Search

Browse DB

- Review all
- Review by classification
- Review by organisms
- Review by tissue distribution
- Review by substrate
- Review by inducer
- Review by inhibitor

Glutathione S-transferase

Introduction:
 GSTs are the principal phase II enzymes involved in metabolic detoxification processes. Its main function is to catalyze the conjugation of the third group of glutathione (GSH), the tripeptide gamma-Glu-Cys-Gly, with a wide variety of electrophiles, generally resulting in greater solubility and easier removal of the hydrophilic substrate from the cells.

Gsts can be classified as three groups including the cytosolic GST, mitochondrial GST and the microsomal GSTs based on their cellular locations. Totally seven classes of cytosolic GSTs have been classified in mammalian organisms, designated alpha, mu, pi, sigma, theta, omega, and zeta. Other classes of cytosolic GSTs, named beta, delta, epsilon, lambda, phi, rho, and tau also have been identified in nonmammalian organisms including plants, insects, bacteria, fungi, haloblasts, fish, etc. Moreover, the mitochondrial GST isoenzyme which was designated kappa has been investigated in mammalian species. The kappa GSTs are quite distinct from cytosolic GSTs. The microsomal GST was now termed MAPEG (membrane-associated protein involved in eicosanoid and glutathione metabolism). The MAPEG include 5-lipoxygenase activating protein (FLAP) and leukotriene (LT) C4 synthase, MGS1, MGS2, MGS3 and prostaglandin (PG) E synthase 1 (PGES1).

Reference:
 Blanchelet B, Feig X, Slagh BB. Marine glutathione S-transferases. *Mar Biotechnol* (NY). 2007 Sep Oct; 9(5):513-42. PMID:1782821
 Froux C. Glutathione transferases in the genomics era: new insights and perspectives. *Froux C. Biomed Eng.* 2006 Sep; 23(4):143-60. PMID:1623810
 Hines RN. The ontogeny of drug metabolism enzymes and implications for adverse drug events. *Pharmacol Ther.* 2008 May; 118(2):250-47. PMID:18406487
 Jakobsson PA, Bergstrom R, Mancini J, Ford-Hutchinson A, Persson B. Membrane-associated proteins in eicosanoid and glutathione metabolism (MAPEG): a widespread protein superfamily. *Am J Physiol Cell Physiol.* 2009 Feb; 296(2):C20-4. PMID:19575221
 Li J, Xia Z, Ding J. Thioesterase-like domain of human kappa class glutathione transferase reveals sequence homology and structure similarity to the theta class enzyme. *Protein Sci.* 2005 Sep; 14(9):2561-9. PMID:16081649

Quick Search: Search by New Entry

Advanced Search: Search criteria

Blast in Database

Enter sequence below in FASTA format

Or load it from local disk

Algorithm Parameters: Expect 10 Max target sequences 100

3D-Blast in Database

Please upload the protein structure file in PDB format from local disk

Algorithm Parameters: Expect 10 Max target sequences 100

Peptide Matching

Type peptide sequence below in FASTA format (Up to 10 peptide sequence were allowed)

Protein Domain Search

Please enter the domain name below

General Information

DetoxiProt: DP00451
 Symbol Name: GSTM2
 Protein Name: glutathione S-transferase mu 2 isoform 1
 Protein Family: Glutathione S-transferase
 Protein Class: Phase II
 EC Number: 2.5.1.18
 Organism: Homo sapiens
 Taxonomic group: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominoidea; Homo

Protein Information

Chromosome location: chromosome: 1; Location: 1p13.3
 Tissue Distribution: muscle
 Subcellular Location: cytoplasm
 Cofactor: N/D
 Substrate: dinitrochlorobenzene <1-chloro-2,4-dinitrobenzene>
 1,2-dichloro-4-nitrobenzene
 Inducer: wild ginseng
 butanoate
 Inhibitor: 6-(7-allyl-2,1,3-benzoxadiazol-4-ylthio)hexanol

Protein Features

Molecular Weight: 25744.6
 Theoretical pI: 5.99
 Aliphatic Index: 81.01
 GRAVY: -0.463

Database	Entry	Identifier	Start	End	Score	E value
PFAM:	PF02798.13	GST_N	3	82	83.8	6.8e-24
PFAM:	PF00043.18	GST_C	104	192	62.8	1.9e-17
Prosite:	PSS0404	GST_NITER	1	88	31.845	0
Prosite:	PSS0405	GST_CTER	90	208	25.358	0

Protein length: 218

Protein Sequence:

```

1  MPWILGVNI  RGLAISIRL  LEYSSSYEE  KEYINGDAPD  YDRSDILNEK  FKLGLDFPRL
61  PYLIDGTHKI  TDSNALLRYI  ARKHNLCDEE  EKEDIEDIL  ENDFHDSRHO  LAKLKYDPOF
121  EKLEPELYDA  LPEMLLYYS  FLQGPFLG  DKITFYDFA  YDVLNNGYF  EPCLDAFVN
181  LKDFLSRFEG  LEKISAYMS  SRFLRPYFT  KMAVRQIK
    
```

Top BLAST Hits:

Entry ID	Score	E value	Identity	Species	Description
DP04080	441	e-125	99%	Homo sapiens	PREDICTED: glutathione S-transferase M2 (muscle) isoform 3
DP03957	441	e-125	99%	Homo sapiens	glutathione transferase M2
DP05062	433	e-123	94%	Homo sapiens	glutathione S-transferase mu 2
DP00977	400	e-113	84%	Homo sapiens	glutathione S-transferase mu 7
DP03255	394	e-111	83%	Homo sapiens	PREDICTED: similar to glutathione S-transferase M1

Protein Structure

1HNA.pdb 1HNB.pdb 1HNC.pdb
 1XW5.pdb 1YKC.pdb 2AB8.pdb
 2CAJ.pdb 2STU.pdb 3GTU.pdb

ruban Color
 color ribbons none
 color ribbons structure
 color ribbons amino
 color ribbons skyblue
 color ribbons yellow
 wireframe on/off
 spm on/off
 spacefill # off 20% 25% 100%

External Database links

NCBI Protein Entry: NP_000839
 Gene ID: 2946
 UNIPROT: P28161
 KEGG: hsa2946

Gene Ontology:

Biological process: metabolic process
 Cellular component: cytoplasm
 Molecular function: glutathione transferase activity
 Traceable author statement

[1] Applah Opong R, Commandeur JH, Istyastono E, Bogaards JJ, Vermeulen HP Inhibition of human glutathione S-transferases by curcumin and analogues. *Xenobiotica* 2009 Apr; 29(4):302-11 PMID:19350453
 [2] Zhu R, Ju Z, Zhou K, Biera HP, Santo A, Gabrielson NL, Li Y Cytotoxicity: dithiolethione-mediated coordinated induction of total cellular and mitochondrial antioxidants and phase 2 enzymes in human primary cardiomyocytes: cytoprotection against oxidative/electrophilic stress and doxorubicin toxicity. *Exp Biol Med* (Maywood) 2009 Apr; 234(4):415-29 PMID:19176875
 [3] Romero L, Ng L, Kirby GM Chemical inducers of rodent glutathione s-transferases down-regulate human GSTA1 transcription through a mechanism involving variant hepatic nuclear factor 1.C. *Mol Pharmacol* 2006 Jul; 70(1):277-86 PMID:16608922
 [4] Cacciatore L, Cacciarri AM, Cocco A, De Maria F, Di Stefano A, Luisi G, Pinnen F, Ricci G, Sozio P, Turella P Potent isozyme-

Bibliography browse page

DetoxiProt ID	Name	Species	Classification	Phase	Detail
DP00451	GSTA1	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00452	GSTA2	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00453	GSTA3	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00454	GSTA4	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00455	GSTA5	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00456	GSTK1	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00457	GSTM1	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00458	GSTM2	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00459	GSTM3	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00460	GSTM4	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00461	GSTM5	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00462	GSTO1	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00463	GSTO2	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00464	GSTP1	Homo sapiens (human)	Glutathione S-transferase	Phase II	view
DP00465	GSTT1	Homo sapiens (human)	Glutathione S-transferase	Phase II	view

Figure 3 DetoxiProt Interface. (A) Description page for each protein class, (B) Protein query page of the database, (C) Bibliography browse page, (D) an example for search result of human GSTA1, protein entry DP00451, detailed information includes general information, protein features, protein sequences, protein structures and external database links.

Comparative genomics tools

A customized BLAST tool is embedded for sequence similarity search [35]. To identify potential paralogs or orthologs of each detoxification protein, pre-computed top-matched hits by BLAST search against other proteins in DetoxiProt were listed in the “detailed information” page. This would facilitate phylogenetic relationships studies of the genes across different organisms throughout evolution. Putative ortholog genes may be identified from the best reciprocal hits. For each of the nine model organisms, we generated multiple sequence alignments for each protein family by MUSCLE [36]. Corresponding phylogenetic trees were built by Maximum Likelihood (ML) method embedded in RAxML [37]. The nonparametric bootstrap test was performed for 100 replicates. A Java application of Archaeopteryx tree viewer is embedded for browsing and manipulation of the phylogenetic trees [38]. An example illustrating the result of cytosolic sulfotransferases of *Mus musculus* is available in the additional file 3. Both the multiple sequence alignment and phylogenetic tree pages can be found in the “tools” page.

Discussion

Although information concerning specific types of detoxification proteins, for example, peroxidase or peroxisome related proteins, proteins involved in ethanol or drug metabolism, can be found in some databases including the PeroxisomeDB, PeroxiBase, ERGR and SuperTarget [39], a complete collection of all these categories of proteins is still deficient. Here, the key categories of proteins involved in detoxification are identified and classified. In addition, the targets identified for the detoxification proteins may provide clues to determine the possible effective enzymes in specific drug metabolism pathway. In total, 20 different protein families were identified to be involved in the functionality of detoxification after a literature review. For complete analysis of the database, we inspected the

phylogenetic distribution of protein families across model organisms (please see additional file 4). All of the three major detoxification protein classes were found in the nine model organisms. For the protein families, relatively complete datasets were found in mammals. However, the aldehyde oxidases and xanthine oxidoreductases are missing in chimpanzees, the catalases are missing in chickens and seven protein families are missing in amphibia. In contrast, a relatively small number of proteins were found in invertebrates including the *Drosophila melanogaster* and *Caenorhabditis elegans*. The amount of detoxification genes took up about 1.5% of genomic genes in most model organisms. At the gene-family level, the cytochrome P450 family and the glutathione S-transferase family are the most abundant groups, with 68 and 48 members identified in humans, respectively. A relatively small number of proteins were found in non-model organisms, probably due to the incomplete annotation of the genes in these species.

In order to give an overview of the function of detoxification proteins, we collected all the 388 human genes and then explored their annotation of molecular function in Gene Ontology by means of the GOEAST [40]. Among 368 genes identified to have catalytic activities, more than 70% have the oxidoreductase activity, and about 30% have transferase activity and the function of ion binding. Other genes function for lipid binding, carboxylic acid binding and selenium binding (please see additional file 5). To provide an overview of protein and toxins relationships, we constructed a network to describe links of the relationship between human detoxification proteins and chemical compounds or environmental factors (here unified as toxins), which include the substrates, inducers and inhibitors compiled from the literature (please see additional file 6). Within this network, 18 detoxification proteins (13.6%) are associated with only one kind of toxin, while 78 detoxification proteins (59.1%) are associated with more than five kinds of toxins (Figure 4A). CYP1A1, CYP1A2, CYP3A4, CYP2B6 and UGT1A1 demonstrated the highest

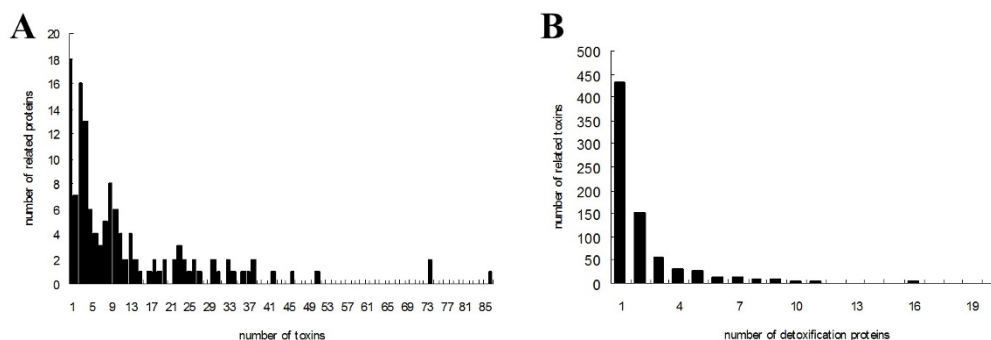


Figure 4 Distribution of human detoxification proteins and related toxins documented in DetoxiProt. (A) Histogram of the number of toxins associated with individual detoxification protein, (B) Histogram of the number of detoxification proteins associated with individual toxin.

connectivity, which suggests the importance of the cytochrome P450 superfamily for elimination of toxins and maintenance of balance in the internal environment. CYPs represent the most important Phase I toxin/drug metabolizing enzymes. They are responsible for the metabolism of more than 80% of the clinical drugs. Inter-individual variation of CYP genes constitutes an important factor related to cancer susceptibility [41]. Similar to CYPs, The Phase II enzyme UGT is another important kind of drug metabolic protein. Association between UGT genes and diseases including cancer and fatty liver disease are already identified [42,43]. For the toxins, the metal ions, estradiol, tobacco smoke, polycyclic aromatic hydrocarbon and rifampin take the top five positions in our developed network. Figure 4B shows the number of proteins associated with toxins. 430 kinds of toxins (57.3%) were shown to be associated with only one protein. Although some bias may exist due to the literature investigated, nevertheless, this network provides a useful tool for inferring the relationships between detoxification protein and toxins.

Conclusions

In summary, DetoxiProt is a powerful and reliable database that provides comprehensive information about detoxification proteins. Various browsing manners and powerful search engines were implemented providing users with useful ways to explore information in the database. We expect that DetoxiProt will serve as a useful tool for researchers in the detoxification protein field. It could also lead to better understanding of the function and evolution of detoxification proteins. The database will be continually updated, and as the increasing numbers of novel sequences become available, newly identified detoxification proteins will be added to the database. We also hope to adopt text-mining tools to help us to pre-screen protein-toxin relationships. These methods will continue to enhance the completeness of the database.

Availability and requirements

The DetoxiProt may be accessed through <http://lifecenter.sgst.cn/detoxiprot/>. All the information is freely available to users. To browse the protein 3D structures and phylogenetic trees, the Java Runtime Environment (JRE) plug-in is required.

Additional material

Additional file 1: Database structure for DetoxiProt. Five major table were used to store data, including the Main Info, predicted Domain, Mutual Blast result, Gene Ontology and Reference.

Additional file 2: Display of the results for different search method. (A) Result for query result list, (B) Result for BLAST, (C) Result for peptide mapping, (D) Key word search result for protein domains, (E) Result for 3D-BLAST.

Additional file 3: Illustration of the phylogenetic tree of Cytosolic sulfotransferases of *Mus musculus* by the Archaeopteryx tree viewer. Multiple Sequence Alignment was performed by MUSCLE, phylogenetic tree was generated by Maximum Likelihood (ML) method embedded in RAxML. The nonparametric bootstrap test was performed for 100 replicates.

Additional file 4: Phylogenetic distribution of detoxification proteins in model organisms. A relative small number of protein families were found in invertebrates than that in vertebrates.

Additional file 5: Gene ontology enrichment analysis of human detoxification genes. This analysis was performed using GOEAST: <http://omicslab.genetics.ac.cn/GOEAST/>

Additional file 6: A bipartite network demonstrating the relationship between detoxification proteins and related toxins. This network comprises 1633 edges and 882 nodes. Red nodes represent detoxification proteins and green nodes are toxins. The nodes size represents degree of connectivity of the nodes. Key proteins (CYP1A1, CYP1A2, CYP3A4, CYP2B6 and UGT1A1) have highest degree of connectivity were labelled, with 86, 74, 74, 51 and 46 edges respectively.

Acknowledgements

This work was supported by the National '863' Program of China [2007AA09Z437]; the National Natural Science Foundation of China [30670367 and 31000583]; the National Basic Research Program of China [Project No. 2009CB118702]; and the China Postdoctoral Science Foundation (20100470746). We would like to thank Larissa Wenren (Harvard University) and Joan Kirtland (University of Virginia) for their comments and suggestions.

This article has been published as part of *BMC Genomics* Volume 12 Supplement 3, 2011: Tenth International Conference on Bioinformatics – First ISCB Asia Joint Conference 2011 (InCoB/ISCB-Asia 2011): Computational Biology. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2164/12?issue=S3>.

Author details

¹School of Life Science, Fudan University, HanDan Road 220#, Shanghai, 200433, China. ²College of Life Science and Technology, Jinan University, Guangzhou, 510632, China. ³Key Laboratory of Nutrition and Metabolism, Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences; Shanghai, 200031, China. ⁴Shanghai Center for Bioinformation Technology, Shanghai, 200235, China. ⁵Department of Food Sciences and Nutrition, Pavillon Paul-Comtois, Université Laval, 2425 rue de l'Agriculture, Québec, QC, Canada G1V 0A6. ⁶Department of Biology, Queen's University, Kingston, Ontario, Canada K7L 3N6. ⁷Institute of Biodiversity Science and Geobiology, Tibet University, Lhasa, 850000, China.

Authors' contributions

XL and YZ conceived the concept of DetoxiProt. GL and LW collected and compiled data from literature and public databases. YY and ZY constructed the database and compiled the draft of the manuscript. LY incorporated protein structure manipulative tools for the database. YH, HW and LW participated in the design of the study and integrated protein data. RH and RR provided valuable suggestions and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 30 November 2011

References

1. Gonzalez FJ: Cytochrome P450 humanised mice. *Hum Genomics* 2004, **1**(4):300-306.
2. Kovacic P, Somanathan R: Mechanism of teratogenesis: electron transfer, reactive oxygen species, and antioxidants. *Birth Defects Res C Embryo Today* 2006, **78**(4):308-325.

3. Hines RN, McCarver DG: **The ontogeny of human drug-metabolizing enzymes: phase I oxidative enzymes.** *J Pharmacol Exp Ther* 2002, **300**(2):355-360.
4. Limón-Pacheco J, Gonshebbat ME: **The role of antioxidants and antioxidant-related enzymes in protective responses to environmentally induced oxidative stress.** *Mutat Res* 2009, **674**(1-2):137-147.
5. McCarver DG, Hines RN: **The ontogeny of human drug-metabolizing enzymes: phase II conjugation enzymes and regulatory mechanisms.** *J Pharmacol Exp Ther* 2002, **300**(2):361-366.
6. Ingelman-Sundberg M: **Human drug metabolising cytochrome P450 enzymes: properties and polymorphisms.** *Naunyn Schmiedebergs Arch Pharmacol* 2004, **369**(1):89-104.
7. Elbaz A, Dufouil C, Alperovitch A: **Interaction between genes and environment in neurodegenerative diseases.** *C R Biol* 2007, **330**(4):318-328.
8. Kohjima M, Enjoji M, Higuchi N, Kato M, Kotoh K, Yoshimoto T, Fujino T, Yada M, Yada R, Harada N, et al: **Re-evaluation of fatty acid metabolism-related gene expression in nonalcoholic fatty liver disease.** *Int J Mol Med* 2007, **20**(3):351-358.
9. Kumarakulasingham M, Rooney PH, Dundas SR, Telfer C, Melvin WT, Curran S, Murray GI: **Cytochrome p450 profile of colorectal cancer: identification of markers of prognosis.** *Clin Cancer Res* 2005, **11**(10):3758-3765.
10. Abel EL, Opp SM, Verlinde CL, Bammler TK, Eaton DL: **Characterization of atrazine biotransformation by human and murine glutathione S-transferases.** *Toxicol Sci* 2004, **80**(2):230-238.
11. Lin SY, Yang JH, Hsia TC, Lee JH, Chiu TH, Wei YH, Chung JG: **Effect of inhibition of aloe-emodin on N-acetyltransferase activity and gene expression in human malignant melanoma cells (A375.S2).** *Melanoma Res* 2005, **15**(6):489-494.
12. Mackenzie PI, Gregory PA, Gardner-Stephen DA, Lewinsky RH, Jorgensen BR, Nishiyama T, Xie W, Radomska-Pandya A: **Regulation of UDP glucuronosyltransferase genes.** *Curr Drug Metab* 2003, **4**(3):249-257.
13. Di Pietro G, Magno LA, Rios-Santos F: **Glutathione S-transferases: an overview in cancer research.** *Expert Opin Drug Metab Toxicol* 2010, **6**(2):153-170.
14. Wu C, Wang L, Liu C, Gao F, Su M, Wu X, Hong F: **Mechanism of Cd²⁺ on DNA cleavage and Ca²⁺ on DNA repair in liver of silver crucian carp.** *Fish Physiol Biochem* 2008, **34**(1):43-51.
15. Zhao H, Xu X, Na J, Hao L, Huang L, Li G, Xu Q: **Protective effects of salicylic acid and vitamin C on sulfur dioxide-induced lipid peroxidation in mice.** *Inhal Toxicol* 2008, **20**(9):865-871.
16. Seet RC, Lee CY, Lim EC, Tan JJ, Quek AM, Chong WL, Looi WF, Huang SH, Wang H, Chan YH, et al: **Oxidative damage in Parkinson disease: Measurement using accurate biomarkers.** *Free Radic Biol Med* 2010, **48**(4):560-566.
17. Power JH, Blumbergs PC: **Cellular glutathione peroxidase in human brain: cellular distribution, and its potential role in the degradation of Lewy bodies in Parkinson's disease and dementia with Lewy bodies.** *Acta Neuropathol* 2009, **117**(1):63-73.
18. Danielson PB: **The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans.** *Curr Drug Metab* 2002, **3**(6):561-597.
19. Sheehan D, Meade G, Foley VM, Dowd CA: **Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily.** *Biochem J* 2001, **360**(Pt 1):1-16.
20. Zamocky M, Furtmuller PG, Obinger C: **Evolution of catalases from bacteria to humans.** *Antioxid Redox Signal* 2008, **10**(9):1527-1548.
21. Schluter A, Fourcade S, Domenech-Estevéz E, Gabaldon T, Huerta-Cepas J, Berthommier G, Ripp R, Wanders RJ, Poch O, Pujol A: **PeroxisomeDB: a database for the peroxisomal proteome, functional genomics and disease.** *Nucleic Acids Res* 2007, **35**(Database issue):D815-822.
22. Koua D, Cerutti L, Falquet L, Sigrist CJ, Theiler G, Hulo N, Dunand C: **PeroxiBase: a database with new tools for peroxidase family classification.** *Nucleic Acids Res* 2009, **37**(Database issue):D261-266.
23. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.
24. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36**(Database issue):D480-484.
25. **The Universal Protein Resource (UniProt) 2009.** *Nucleic Acids Res* 2009, **37**(Database issue):D169-174.
26. Guo AY, Webb BT, Miles MF, Zimmerman MP, Kendler KS, Zhao Z: **ERGR: An ethanol-related gene resource.** *Nucleic Acids Res* 2009, **37**(Database issue):D840-845.
27. Hyndman D, Bauman DR, Heredia WV, Penning TM: **The aldo-keto reductase superfamily homepage.** *Chem Biol Interact* 2003, **143-144**:621-631.
28. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetverin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37**(Database issue):D5-15.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**(1):235-242.
30. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(Database issue):D211-222.
31. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuče BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ: **The 20 years of PROSITE.** *Nucleic Acids Res* 2008, **36**(Database issue):D245-249.
32. Letunic I, Doerks T, Bork P: **SMART 6: recent updates and new developments.** *Nucleic Acids Res* 2009, **37**(Database issue):D229-232.
33. Tung CH, Huang JW, Yang JM: **Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database.** *Genome Biol* 2007, **8**(3):R31.
34. Yang JM, Tung CH: **Protein structure database search and evolutionary classification.** *Nucleic Acids Res* 2006, **34**(13):3646-3659.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389-3402.
36. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
37. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688-2690.
38. Zmasek CM, Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17**(4):383-384.
39. Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiss A, Jensen LJ, et al: **SuperTarget and Matador: resources for exploring drug-target relationships.** *Nucleic Acids Res* 2008, **36**(Database issue):D919-922.
40. Zheng Q, Wang XJ: **GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W358-363.
41. Zhou SF, Liu JP, Chowbay B: **Polymorphism of human cytochrome P450 enzymes and its clinical impact.** *Drug Metab Rev* 2009, **41**(2):89-295.
42. Yao L, Qiu LX, Yu L, Yang Z, Yu XJ, Zhong Y, Yu L: **The association between TA-repeat polymorphism in the promoter region of UGT1A1 and breast cancer risk: a meta-analysis.** *Breast Cancer Res Treat* 2010, **122**(3):879-882.
43. Lin YC, Chang PF, Hu FC, Chang MH, Ni YH: **Variants in the UGT1A1 gene and the risk of pediatric nonalcoholic fatty liver disease.** *Pediatrics* 2009, **124**(6):e1221-1227.

doi:10.1186/1471-2164-12-S3-S2

Cite this article as: Yang et al: **DetoxiProt: an integrated database for detoxification proteins.** *BMC Genomics* 2011 **12**(Suppl 3):S2.