mUbiSiDa: A Comprehensive Database for Protein Ubiquitination Sites in Mammals

Tong Chen[1], Tao Zhou[2], Bing He[1], Haiyan Yu[1], Xuejiang Guo[2]*, Xiaofeng Song[1]*, Jiahao Sha[2]

1 Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 2 State Key Laboratory of Reproductive Medicine, Department of Histology and Embryology, Nanjing Medical University, Nanjing, China

Abstract

Motivation: Protein ubiquitination is one of the important post-translational modifications by attaching ubiquitin to specific lysine (K) residues in target proteins, and plays important regulatory roles in many cell processes. Recent studies indicated that abnormal protein ubiquitination have been implicated in many diseases by degradation of many key regulatory proteins including tumor suppressor, oncoprotein, and cell cycle regulator. The detailed information of protein ubiquitination sites is useful for scientists to investigate the mechanism of many cell activities and related diseases.

Results: In this study we established mUbiSida for mammalian Ubiquitination Site Database, which provides a scientific community with a comprehensive, freely and high-quality accessible resource of mammalian protein ubiquitination sites. In mUbiSida, we deposited about 35,494 experimentally validated ubiquitinated proteins with 110,976 ubiquitination sites from five species. The mUbiSiDa can also provide blast function to predict novel protein ubiquitination sites in other species by blast the query sequence in the deposit sequences in mUbiSiDa. The mUbiSiDa was designed to be a widely used tool for biologists and biomedical researchers with a user-friendly interface, and facilitate the further research of protein ubiquitination, biological networks and functional proteomics. The mUbiSiDa database is freely available at http://reprod.njmu.edu.cn/mUbiSiDa.

Introduction

Protein ubiquitination, known as the important protein post-translational modification of targeting proteins by ubiquitins for their subsequent degradation in the ATP-dependent ubiquitin proteasome system (UPS), plays an important role in cell activity [1–5]. Currently, many regulatory functions of protein ubiquitination have been discovered, including the regulation of DNA repair and transcription, control of signal transduction, and implication of endocytosis and sorting. There are also sufficient evidences showing that the deregulation of protein ubiquitination have been implicated in many diseases by degradation of many key regulatory proteins, including tumor suppressor, oncoprotein, and cell cycle regulator [6–8]. Therefore protein ubiquitination has captured the attention of researchers in life science fields.

Protein ubiquitination is implemented by ubiquitin binding to the lysine site of a target protein. The location, numbers, and distribution of ubiquitination site are important information for scientists to investigate the mechanism of UPS and relevant diseases [9,10]. Therefore recently scientists began to focus on collecting and managing the available protein ubiquitination information. The current available ubiquitination database like UbiProt provides information almost exclusively on yeast protein ubiquitination due to the previous experimental limits [11]. For human, hUbiquitome mainly focused on the relationship between ubiquination enzymes and substrates, and only 279 substrates were included in the database [12]. With the recent new technology advances, such as proteomic technology, mass spectrometry technology, the growing number of the experimentally confirmed ubiquitination sites of mammalian protein laid the solid basis for investigating the regulatory mechanism of protein stability. Therefore it is imperative for scientific community to construct the comprehensive database for depositing and retrieving mammalian protein ubiquitination sites.

Consequently, in this study we constructed a user-friendly database, mUbiSida, which meets the above requirements. The dataset in mUbiSiDa are mainly collected from published papers. In total, we searched and obtained 104 references containing experimentally validated 35,494 mammalian ubiquitinated proteins from 5 species. Over 95% of these sites are from human and mouse. This comprehensive database enables not only the information retrieval of protein ubiquitination sites, but also the study of cross-regulation between post-translational modification [13,14]. In addition, the mUbiSiDa can facilitate biologists in further exploration on the molecular mechanisms of protein stability related cell process.

In conclusion, the aim of the mUbiSiDa database is to provide a user-friendly web interface to browse, search, retrieve and update information on mammalian protein ubiquitinated sites, and to promote the further research of protein ubiquitination, biological networks and functional proteomics.

Materials and Methods

Data sources and database construction methods

Our data are collected from two main resources: Firstly, published literatures from PubMed. Through keyword search results in PubMed, we have obtained a mass of related literatures, after which experimentally identified mammalian ubiquitylation sites were downloaded from these literatures Several main references were listed in the section of References [15–26]. A well-known international database UniProt was also searched for protein ubiquitination [11]. Next, we manually reviewed entries of UniProt, and added those ubiquitinated proteins with reference support to the database, and eliminated those predicted lysine modifications without references support in Uniprot. Based on above data collection, we constructed a database containing 35,494 proteins from 5 species with 110, 976 ubiquitination sites. And currently there are 30,322, 5,168, 2, 1 and 1 counts for Human, Mouse, Bovine, Rat and Pig, respectively. The data collecting pipeline and the structure of mUbiSiDa were shown in Figure 1. The distribution of all experimentally validated mammalian protein ubiquitination sites collected in mUbiSiDa is shown in Figure 2. Proteins with less than or equal to 5 sites occupy 85.6% of the entries in mUbiSiDa. Proteins with greater than 5 and not greater than 10 sites occupy 10.0% of the entries. And proteins with greater than 10 sites occupy only 4.4%.

mUbiSiDa was constructed and configured upon a typical LAMP (Linux + Apache + MySQL + PHP) platform. Apache5.0.51b was firstly used to build up a webserver. All dataset were stored in MySQL 5.0, and web interface was achieved by PHP scripts (PHP version 5.2) on Linux, powered by an Apache server. WebPages were designed with html and JavaSript techniques. Website and database were connected and all kinds of function were achieved by PHP techniques.

Results

Functions and utility

All functions of mUbiSiDa were shown in Home page (Figure 3). As an interactive, comprehensive and user-friendly database, mUbiSiDa mainly provides the following functions for users: Search, Advanced Retrieval, Browse, Resource, Interaction and Blast Search.

Search function

In the text field of the Search page, the users can input the query strings, such as protein ID, protein name, or others, and then the obtained result pages will be the list of protein entries that matched the query strings. Those keywords matching the query strings in the result pages will be yellow highlighted. The users can view the detailed information about their interested protein by clicking the protein ID on the left column. For further retrieval, combined search is also available, but the most two query strings combined by '+' are only supported in the current version homepage of database. In addition, on each result page, mUbiSiDa also provides customized function to reorganize the feature list of protein entries according to user's interests.
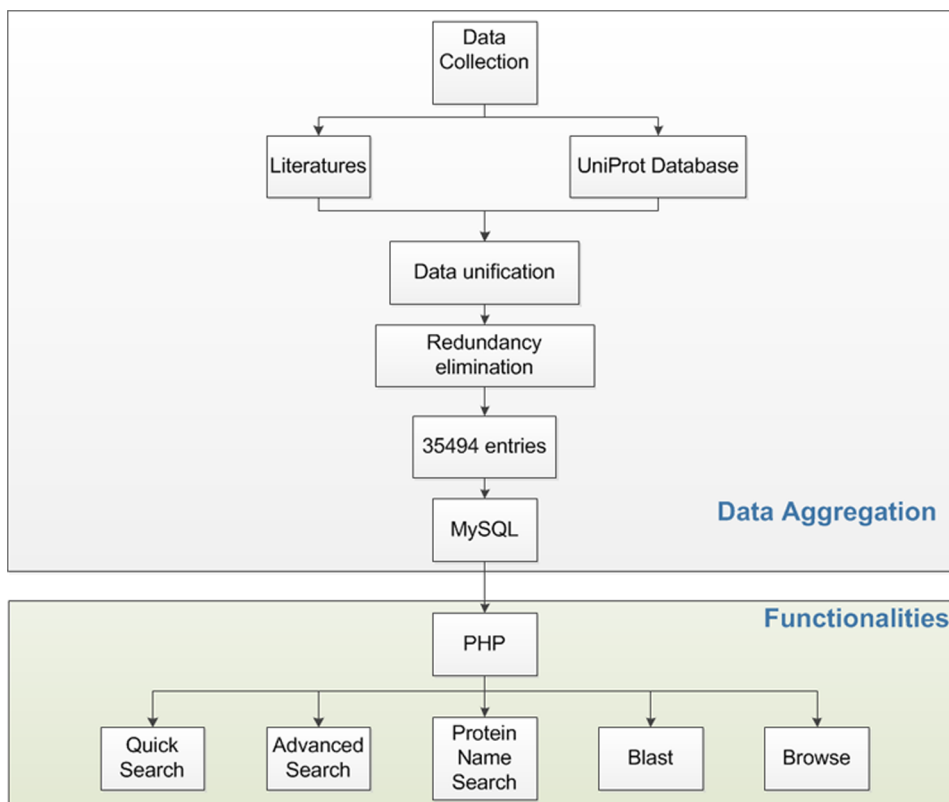

Figure 1. Data collection and the structure of mUbiSiDa.
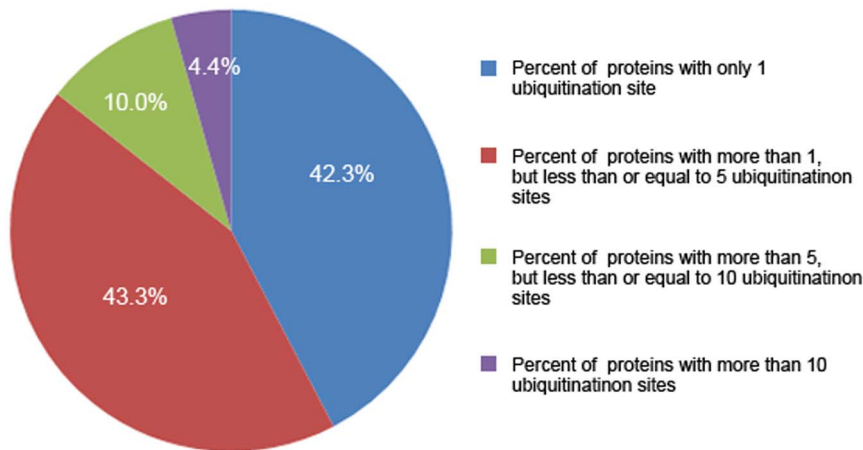doi:10.1371/journal.pone.0085744.g001

Figure 2. Statistics of protein ubiquitination sites.
doi:10.1371/journal.pone.0085744.g002

Advanced retrieval

There are three options in Advanced Retrieval: (1) Advanced Search (2) Protein Name Search, (3) Sequence Blast, which is efficient access for the users to obtain their interested information. The users can obtain the specific protein ubiquitination information by entering more restricted words in six text fields combined with 'AND', 'OR' and 'BUT' in "Advanced Search" query page. We specifically designed "Protein Name Search" query page for the users conveniently to obtain their interested protein when knowing their names. There are four text fields for users to input query

Figure 3. Home page of mUbiSiDa.
doi:10.1371/journal.pone.0085744.g003

General information

Uniprot Accession Number	Q86WH2-1
Organism	Homo sapiens (Human)

Gene Ontology (GO) annotation

Gene Ontology (GO)

Biological process	Accession Number	Description
	0007165	signal transduction

	Accession number	Description
	0005737	cytoplasm

Ubiquitination annotation(Features)

Ubiquitination site	K207; K217;				
		position/full length	Ubiquitination sequence	Graphical view	Evi
Ubiquitination sequence	No.1	207/238	FLRILDKEEDEQL	——————|——	Exp
	No.2	217/238	EQLQNLKRRYTAY	————————|—	Exp

Protein squence

Length	238
Sequence	MSSGYSSLEE DAEDFFFTAR TSFFRRAPQG KPRSGQQDVE KEKETHSYLS KEEIKEKVHK YNLAVTDKLK MTLNSNGIYT GFIKVQMELC KPPQTSPNSG KLSPSSNGCM NTLHISSTNT VGEVIEALLK KFLVTESPAK FALYKRCHRE DQVYACKLSD REHPLYLRLV AGPRTDTLSF VLREHEIGEW EAFSLPELQN FLRILDKEED EQLQNLKRRY TAYRQKLEEA LREVWKPD

Half-life prediction

| SProtP Human | >Q86WH2-1
MSSGYSSLEEDAEDFFFTARTSFFRRAPQGKPRSGQQDVEKEKETHSYLSKEEIKEKVHKYNLAVTDKLKMT
LNSNGIYTGFIKVQMELCKPPQTSPNSGKLSPSSNGCMNTLHISSTNTVGEVIEALLKKFLVTESPAKFALY
KRCHREDQVYACKLSDREHPLYLRLVAGPRTDTLSFVLREHEIGEWEAFSLPELQNFLRILDKEEDEQLQNL
KRRYTAYRQKLEEALREVWKPD

Submit |

Multi-species alignment analysis

Multi-species Alignment Analysis for Homo sapiens	Analysis

Reference

Content	1. PubMed: 21906983 Woong Kim,Eric J. Bennett, Edward L. Huttlin, Ailan Guo, Jing Li, Antho Sowa,Ramin Rad, John Rush, Michael J. Comb, J. Wade Harper, and Ste Quantitative Assessment of the Ubiquitin-Modified Proteome. Molecular

Figure 4. Detailed information page.
doi:10.1371/journal.pone.0085744.g004

strings, part or all of the text fields can be used for combined search. When the ID text field is empty, the other three text fields were designed to be fuzzy query. When the ID text field is not empty, it was set to be precise search. The ID is supposed to be UniProt ID.

"Blast search" was also specifically designed for the users to predict the potential ubiquitination site information of new protein in other mammalian species, which is orthologs or paralogs deposited in the database. As a tool for comparing primary biological sequence information, Sequence Blast enables a user to compare a query sequence with all sequences of mUbiSiDa, and identify sequences of mUbiSiDa that resemble the query sequence above a certain threshold. By means of this method, one or few results which sequence is most similar to the input sequence will be obtained after blast process. The Sequence Blast is a high efficient tool for retrieving possible ubiquitinated lysine sites of proteins input by users. Not only will the protein ID, but also related references are posted after using blast. Users can view detailed information of the result by clicking the proteins ID. This tool extends the function of the mUbiSiDa to be used to predict protein ubiquitination sites for most mammalian species.

Data browse and download

The dataset in mUbiSiDa can be browsed in four options: (1) Browse data by organism, (2) Browse data by biological Process, (3) Browse data by cellular component, (4) Browse data by molecular function.

In "Browse Data by Organism" page, the total dataset in mUbiSiDa are grouped by organism, which is listed in a table with 5 kinds of organisms collected by mUbiSiDa. The corresponding entry number of each organism is also listed. The users can view all associated entries for one organism by clicking on right numbers. This function is quite useful for users to view all the relevant information in particular species.

The total dataset in mUbiSiDa can also be viewed by Gene Ontology (GO). In the result pages, the gene ontology IDs classified as biological process, cellular component, and molecular function for each protein deposited in mUbiSiDa will be listed. The gene ontology terms, and the corresponding protein entries numbers are also listed. The user can select and click the gene ontology ID to view all proteins of the selected gene ontology ID. The matched results can help the users with a convenient way to view ubiquitinated proteins grouped by GO terms. The user can benefit from this function to do cell research on regulation of protein's function or biological processes.

All the dataset in mUbiSiDa can be downloaded at "Resources" web page.

Submission of new protein ubiquitination site

In order to maintain an up-to-date and comprehensive resource, we designed the submission page for the users to submit their own data to mUbiSiDa. The submission page requires that the users should offer the following submission items: protein name, ID, organism, ubiquitination site(s), PubMed ID, GO ID, sequence, and other information.

Detailed information in each entry page

Users are allowed to have an access to the detailed information of each entry by clicking the ID on the left column in the result page. All information was carefully checked manually to ensure the accuracy of each entry. The each entry page is divided into six sections. The first section (Figure 4) provide basic protein information including "Uniprot Accession Number", "Organism", "Protein names" and "Gene names". The Ontology section mainly contains the Gene Ontology (GO) of each entry, which was

chiefly downloaded manually from UniProt. The most important part of mUbiSiDa is shown in Ubiquitination annotation (Features) section (Figure 4) including the location in the sequence about ubiquitination sites. In this section, all ubiquitination sites of lysine in the sequence are listed in the Graphical view. On the right of each ubiquitination sites, colored PubMed ID with hyperlink to Reference section is also provided as the strong evidence of protein ubiquitination site. All lysine ubiquitination sites were labeled and red colored in "Protein sequence" section. This makes it easy for users to recognize all ubiquitination sites. Besides, "Half-life Prediction" section and "Multi-species Alignment Analysis" section were added here for study of short-lived proteins and homologous sites in non-human species.

Discussion and Conclusions

mUbiSiDa, the comprehensive mammalian ubiquitination site database, aiming to systematically aggregate all experimentally verified ubiquitinated protein, and provide biologists for analyzing protein stability and mechanism of protein degradation. Most data in this database are collected from human and mouse. It needs emphasizing that potential ubiquitinated lysine site from other organisms can be predicted according to the results of BLAST SEARCH or Multi-species Alignment Analysis. As a tool designed to be used by biologists and researchers, mUbiSiDa will be continually improved and updated to ensure the convenience and utility of the service, accuracy of the information and innovative in style and function.

To obtain an overview of the biological features of human and mouse ubiquitinated proteins, we firstly classified these proteins according to Gene Ontology (See File S1). For human, most proteins are located on intracellular part (61.9%). And the largest parts for molecular function and biological process were binding (33.9%) and metabolic process (27.2%) respectively. The distribution of mouse ubiquitinated proteins was similar to human, suggesting that the biological functions of ubiquitinated proteins may be conserved among mammals. We further performed functional enrichment analysis based on human and mouse ubiquitinated proteins using WebGestalt [27]. Among the top ten enriched pathways for human proteins, the most significant pathway is "Genes involved in Gene Expression", which is also enriched among mouse proteins. Thus, we showed that mUbiSiDa could provide resources for further analysis of the functions of ubiquitination.

In the future work, efficiency in search process and new style of search need to be improved to satisfy different needs. The function of mammalian ubiquitination sites prediction will be subsequently added to mUbiSiDa in the future version. With continuous improvement, mUbiSiDa is expected to make a contribution to the researches on regulation of protein's function or biological processes.

Supporting Information

File S1 Gene Ontology classification and pathway enrichment analysis results for human and mouse ubiquitinated proteins.
(XLSX)

Author Contributions

Conceived and designed the experiments: XFS XJG JHS. Performed the experiments: TC TZ. Analyzed the data: BH HYY. Wrote the paper: TC XFS XJG.

References

1. Hershko A, Ciechanover A (1998) The ubiquitin system. Annu Rev Biochem, 67:425–479.
2. Hershko A (2005) The ubiquitin system for protein degradation and some of its roles in the control of the cell division cycle. Cell Death Differ,12:1191–1197.
3. Pickart CM, Eddins MJ (2004) Ubiquitin: structures, functions, mechanisms. Biochim. Biophys Acta, 1695:55–72.
4. Sorokin AV, Kim ER, Ovchinnikov LP (2009) Proteasome system of protein degradation and processing. Biochemistry, 74(13):1411–1442.
5. Walsh CT, Garneau TS, Gatto GJ Jr (2005) Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl, 44: 7342–7372.
6. Ciechanover A, Orian A, Schwartz AL (2000) The ubiquitin-mediated proteolytic pathway: mode of action and clinical implications. J Cell Biochem Suppl, 34: 40–51.
7. Schwartz AL, Ciechanover A (2009) Targeting proteins for destruction by the ubiquitin system: implications for human pathobiology. Annu Rev Pharmacol Toxicol, 49:73–96.
8. Stewart HL, Alfred LG, William EM (2006) Protein Degradation by the Ubiquitin–Proteasome Pathway in Normal and Disease States. JASN, 17(7):1807–1819.
9. Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, et al. (2010). Identification, analysis, and prediction of protein ubiquitination sites. Proteins: Structure, Function, and Bioinformatics 78 (2): 365–380.
10. Tzachi H, Agnes TP, Ariel A, Yaakov L (2012) The origins and evolution of ubiquitination sites. Mol BioSyst, 8, 1865–1877.
11. Chernorudskiy AL, Garcia A, Eremin EV, Shorina AS, Kondratieva EV, et al. (2007) UbiProt: a database of ubiquitylated proteins. BMC Bioinformatics, 8: 126.
12. Du Y, Xu N, Lu M, Li T (2011) hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. Database (Oxford). 2011:bar055. doi:10.1093/database/bar055
13. Wang Z, Udeshi ND, Slawson C, Compton PD, Sakabe K, et al. (2010) Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. Sci. Signal., 3, ra2.
14. Yang XJ, Seto E (2008) Lysine acetylation: codified crosstalk with other post-translational modifications. Mol Cell, 31:449–461.
15. Denis NJ, Vasilescu J, Lambert JP, Smith JC, Figeys D (2007) Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry. Proteomics, 7(6):868–874.
16. Jeon HB, Choi ES, Yoon JH, Hwang JH, Chang JW, et al. (2007) A proteomics approach to identify the ubiquitinated proteins in mouse heart. Biochem Biophys Res Commun, 357(3):731–736.
17. Kirkpatrick DS, Weldon SF, Tsaprailis G, Liebler DC, Gandolfi AJ (2005) Proteomic identification of ubiquitinated proteins from human cells expressing His-tagged ubiquitin. Proteomics, 5(8):2104–2111.
18. Matsumoto M, Hatakeyama S, Oy amada K, Oda Y, Nishimura T, et al. (2005) Large-scale analysis of the human ubiquitin-related proteome. Proteomics, 5(16): 4145–4151.
19. Meierhofer D, Wang X, Huang L, Kaiser P (2008) Quantitative analysis of global ubiquitination in HeLa cells by mass spectrometry. J Proteome Res 7, 4566–4576.
20. Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, et al. (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol, 21(8): 921–926.
21. Tan F, Lu L, Cai Y, Wang J, Xie Y, et al. (2008) Proteomic analysis of ubiquitinated proteins in normal hepatocyte cell line Chang liver cells. Proteomics 8: 2885–2896.
22. Xu G, Paige JS, Jaffrey SR (2010) Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. Nat Biotechnol 28, 868–873.
23. Vasilescu J, Smith JC, Ethier M, Figeys D (2005) Proteomic analysis of ubiquitinated proteins from human MCF-7 breast cancer cells by immunoaffinity purification and mass spectrometry. J Proteome Res 4, 2192–2200.
24. Oshikawa K, Matsumoto M, Oyamada K, Nakayama KI (2012) Proteome-wide identification of ubiquitylation sites by conjugation of engineered lysine-less ubiquitin. J Proteome Res 11, 796–807.
25. Emanuele MJ, Elia AE, Xu Q, Thoma CR, Izhar L, et al. (2011) Global identification of modular cullin-RING ligase substrates. Cell 147, 459–474.
26. Na CH, Jones DR, Yang Y, Wang X, Xu Y, et al. (2012) Synaptic protein ubiquitination in rat brain revealed by antibody-based ubiquitome analysis. J Proteome Res. 11, 4722–4732.
27. Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res, 41 (Web Server issue), W77–83.