Databases and ontologies

PRISM: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact

Fuyi Li[1,2], Cunshuo Fan[3], Tatiana T. Marquez-Lago[4], André Leier[4], Jerico Revote[1], Cangzhi Jia[5,6], Yan Zhu[7], A. Ian Smith[1], Kuo-Chen Chou[8], Geoffrey I. Webb[2], Quanzhong Liu[3,*], Leyi Wei[9,*], Jian Li[7] and Jiangning Song[1,2,*]

[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, VIC 3800, Australia, [2]Monash Centre for Data Science, Faculty of Information Technology, Monash University, VIC 3800, Australia, [3]College of Information Engineering, Northwest A&F University, Yangling 712100, China, [4]Department of Genetics and Department of Cell, Developmental and Integrative Biology, School of Medicine, University of Alabama at Birmingham, AL, USA, [5]College of Science, Dalian Maritime University, Dalian 116026, China, [6]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore, [7]Biomedicine Discovery Institute and Department of Microbiology, Monash University, VIC 3800, Australia, [8]Gordon Life Science Institute, Boston, MA 02478, USA and [9]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300000, China.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Post-translational modifications (PTMs) play very important roles in various cell signalling pathways and biological process. Due to PTMs' extremely important roles, many major PTMs have been thoroughly studied, while the functional and mechanical characterization of major PTMs is well-documented in several databases. However, currently available databases only focus on protein sequences, while the real 3D structures of PTMs have been largely ignored. Therefore, studies of PTMs 3D structural signatures have been severely limited by the deficiency of the data.

Results: Here, we developed PRISM, a novel publicly available and free 3D structure-based database for PTMs. PRISM provides comprehensive and interactive online knowledge focused on 3D structural information of PTM proteins. The first version of PRISM encompasses 17,145 non-redundant modification sites on 3,919 related protein 3D structure entries pertaining to 37 different types of PTMs. Our entry web page organizes in a comprehensive manner detailed PTM annotation on the 3D structure and biological information in terms of mutations affecting PTMs information, secondary structural features and residue accessibility features of PTM sites, domain context information, predicted disordered regions and sequence alignments. In addition, high-definition JavaScript packages are employed to enhance information visualization in PRISM. PRISM equips a variety of interactive and customizable search options and data browsing functions; these capabilities allow users to access data via keyword, ID, and advanced options combination search in an efficient and user-friendly way. A download page is also provided to enable users to download the SQL file, computational structural features, and PTM sites' data. We anticipate PRISM will swiftly become an invaluable online resource, assisting both biologists and bioinformaticians to conduct experiments and develop applications supporting discovery efforts in the sequence-structural-functional relationship of PTMs and providing important insight into mutations and PTM sites interaction mechanisms.

Availability: The PRISM database is freely accessible at http://prism.erc.monash.edu/. The database and web interface are implemented in MySQL, JSP, JavaScript, and HTML with all major browsers supported.

Contact: Jiangning.Song@monash.edu; weileyi@tju.edu.cn; liuqzhong@nwafu.edu.cn.

1 Introduction

The relatively modest number of human genes (~25,000) does not reflect the downstream proteomic complexity that is necessary for development and sustaining life. This essential complexity is achieved through protein post-translational modifications (PTMs) which exhibit tremendous chemical diversity and dynamics, playing fundamental roles in various cell signalling pathways, fundamental cellular and biological processes. For example, phosphorylation, one of the most studied PTMs (Li, et al., 2018), is involved in regulation of protein degradation (Swaney, et al.,

2013), signal transduction (McCubrey, et al., 2000), protein-protein interaction (Nishi, et al., 2011), and signalling pathways (Duan and Walther, 2015). Glycosylation, another important PTM, has been proved to play a crucial role in protein folding, trafficking (Marino, et al., 2010; Moharir, et al., 2013), subcellular localization, and degradation (Dwek, 1998; von der Lieth, et al., 2004). Other PTMs, such as lysine PTMs, are key for protein stability (Polevoda and Sherman, 2002), transcription (Glozak, et al., 2005), cellular metabolism (Zhao, et al., 2010), DNA repair and replication (Pickart, 2001) etc. Likewise, due to the fundamental importance of PTMs in cell biology, a number of aberrant PTM sites, which are often introduced by mutations, are highly relevant to various human diseases, including somatic cancers (Fleuren, et al., 2016; Karaca, et al., 2015; Pinho and Reis, 2015). Altogether, it is critical to identify and characterize PTMs and their functional roles, which can in turn help elucidate the pathogenesis of PTMs' related diseases.

A key step to understand mechanisms and functional roles of PTMs is to identify the sequences/structures of their natural substrates and the cor responding PTM sites. Given the fact that PTMs are strongly associated with mutations and diseases, significant efforts have been placed in using advanced high-throughput experimental techniques, such as mass spectr ometry (MS) or MS-based techniques, to detect PTMs of proteins in prot eomics data (Medzihradszky, 2005). Therefore, large-scale modified prot eomics data have been archived in recent years. To date, over 660 PTM t ypes have been discovered (http://www.uniprot.org/docs/ptmlist.txt).

Several popular online databases have been developed, to collectively deposit and provide access to different types of experimentally validated PTMs data for public use. These databases include PhosphoSitePlus (Hornbeck, et al., 2012; Hornbeck, et al., 2015), dbPTM (Huang, et al., 2016; Lee, et al., 2006), SysPTM (Li, et al., 2009; Li, et al., 2014), PLMD (Xu, et al., 2017), Phospho.ELM (Dinkel, et al., 2011), and UniProt (UniProt Consortium, 2018). PhosphoSitePlus was built by manually collecting and organizing data from various PTMs and model organisms, though primarily phosphorylation, ubiquitination and acetylation, from mainly human and mouse proteins. dbPTM has been maintained over 10 years, and integrates multiple PTMs information, including functional and structural knowledge of PTMs; dbPTM also stores drug binding data that shows associations to PTM sites. SysPTM contains both PTM data from public databases and MS experimental data from the literature. We note Pfam domains, KEGG pathways, and Gene Ontology information are also integrated into SysPTM. PLMD is a lysine-specific PTM database, and it encompasses 20 types of lysine PTMs in both eukaryotes and prokaryotes. Phospho.ELM is a phosphorylation-specific database, designed to provide general and kinase-specific phosphorylation sites collected from scientific literature and phosphoproteomic experiments. Lastly, UniProt is one of the largest and most regularly updated databases available, providing comprehensive annotations on protein sequences, functions and other relevant biological information. Currently, the UniProt database contains the information of over 660 types of PTMs.

These online resources collect protein sequences with experimentally validated post-translational modification sites, where the positions of PTM sites on sequences have also been documented. These publicly available online resources not only help biologists to conduct valuable experimental analyses, but also provide bioinformaticians with a wealth of resources to build novel prediction tools for studying sequences and functions of PTMs. All these, together, speed up PTMs' research process.

However, there are few databases have been developed for systematically archiving and conveniently accessing 3D structural data of PTMs. The latest version of dbPTM provides the related PDB (Protein Data Bank) ID of PTM proteins, but the corresponding positions of PTM sites position on PDB structures are not available. Furthermore, dbPTM does not provide a download function for PTM related PDB information. Recently, Gao *et al*. (Gao, et al., 2017) developed BioJavaMolFinder, a software designed to identify protein modifications stored in 3D structures within the PDB database. BioJavaMolFinder scanned PDB database and identified the proteins with modification, and provided the results on the PDB website. However, BioJavaMolFinder solely focuses on modification residues in PDBs, while the relationships among protein sequence-structure-function are not provided. Therefore, there is a generalized lack of user friendly tools for both biologists and bioinformaticians to investigate the sequence-structural-function information of PTM proteins from existing protein sequence-based databases, limiting research efforts. For instance, although there are many bioinformatics prediction tools based on protein sequence data to investigate PTMs (Chen, et al., 2015; Jia, et al., 2018; Li, et al., 2018; Li, et al., 2015; Taherzadeh, et al., 2018; Xue, et al., 2008), there are few prediction studies using PDB structures to build prediction models (Chuang, et al., 2012; Durek, et al., 2009; Li, et al., 2016). This is mainly due to the fact that the process of mapping from protein sequences to protein structures is quite time-consuming and operationally complex.

To bridge the existing knowledge gap, we developed a new web-based database, PRISM (PRoteIn Structure Modification database) to provide a comprehensive annotation of the position of PTM sites in real 3D structures. We extracted experimentally validated PTMs information from several currently available protein sequence-based databases, and subsequently mapped the PTM sites to the corresponding 3D structures. We then filtrated the blast results and stored these results in PRISM. Aside, it has been reported there are 21% disease-related amino acid substitutions associated with missense single nucleotide variants (SNVs) are located in PTM sites (Krassowski, et al., 2018). Accordingly, we additionally extracted and collected PTM-associated disease and population mutations from ActiveDriverDB (Krassowski, et al., 2018) database, so as to facilitate the investigation of structural properties of PTMs and to encourage construction of bioinformatics tools with structural features. To this effect, we also calculated secondary structure based features and residue accessibility features of PTM sites using DSSP (Joosten, et al., 2011) and NACCESS (Hubbard, 1992) software, respectively.

With all the above in consideration, PRISM provides users with the protein domain context, sequence-structural alignment and protein disorder information in each entry page. Moreover, to facilitate users' access to data of interest, keyword/ID search, advanced combinatorial schemas search and "browse" functions are provided. PRISM also allows users to download the SQL file, the results files calculated from DSSP and NACCESS, and PTM annotation files in the download page.

2 Methods

2.1 PTM data collection

Figure 1 (**A**) shows the flowchart of PRISM steps, including data collection and processing. There are three major steps involved in entry collection and processing, described in what follows.

First, we extracted, manually checked and integrated 37 types of PTM annotation data from six major online, public free available PTM databases, including PhosphoSitePlus (Hornbeck, et al., 2015), dbPTM (Huang, et al., 2016), SysPTM (Li, et al., 2014), PLMD (Xu, et al., 2017), Phospho.ELM (Dinkel, et al., 2011), and UniProt (UniProt Consortium, 2018). We only considered and integrated experimentally validated PTM sites data for PRISM; accordingly, all PTM sites labelled

as computational PTM annotations were discarded. For UniProt data, we only considered PTMs with ECO code ECO:0000269 (https://www.uniprot.org/help/evidences), which means the data is manually curated and involving published experimental data.

Secondly, all the remaining protein sequences with PTM sites were mapped to the PDB database (Berman, et al., 2000) using PSI-BLAST (Altschul, et al., 1997) against PDB sequence database (pdbaa) with the e-value of 10^{-3} and three iterations.

Finally, the mapped PDB entries were selected according to the following criteria:

(1) With 100% sequence identity; only PDB sequences aligned with the entry protein sequences with 100% sequence identity were retained.

(2) X-ray resolution structures only; thus, nuclear magnetic resonance (NMR) and electron microscopy (EM) structures were excluded.

(3) PDB chains with missing atoms were removed.

(4) For each query blast sequence, the PDB chain with more PTM sites that can be aligned was selected;

(5) For protein sequences with more than one mapped PDB structure, the structure with the highest resolution was selected.

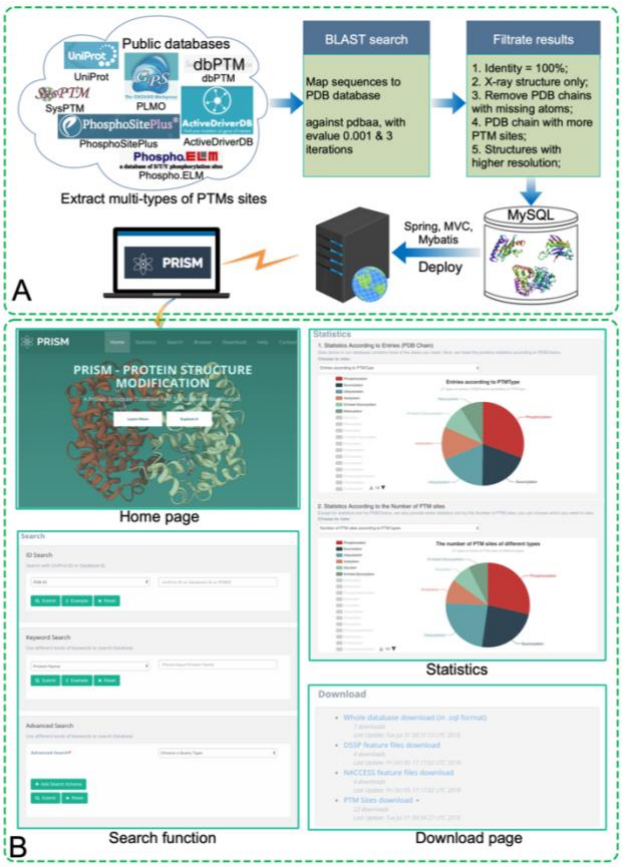


Fig. 1. (**A**) PRISM data collection and processing flowchart; (**B**) Sample web interfaces. There are four major web pages in PRISM: "Home" page, "Search" page, "Statistics" page, and "Download" page.

After these four steps, a total number of 1,7145 PTM sites on 3,919 PDB chains for 37 types of PTMs were extracted and stored into PRISM. The statistics of each type of PTM data are shown in **Table S1**. A discussion of the statistical analysis is included in Section 3.1.

2.2 PTM-associated Mutation collection

Single amino acid variants (SAVs) are the most abundant form of known genetic variations associated with human disease (Yip, et al., 2004). It is reported that about 60% of Mendelian diseases are caused by amino acid substitutions (Botstein and Risch, 2003). A potential effect of SAVs on protein function is the disruption of PTMs (Kim, et al., 2015). For example, the potential for phosphorylation sites to be affected by amino acid variants is high and there have been numerous examples of diseases associated naturally occurring variants that impact the phosphorylation status of proteins (Wagih, et al., 2015). Thus, the mutations around PTMs sites are crucial in the study of health and disease. We have combined mutations associated with various diseases (including but not limited to cancer) and population variants affecting PTMs from ActiveDriverDB, which is a comprehensive collection of annotations of disease and population mutations contained in TCGA. The collected mutations were then mapped to the entries in PRISM. Please refer to **Figure 2** (**C**) "Mutations affecting PTM sites" panel.

2.3 3D structure-based features calculation

Previous studies have shown that 3D structure-based features play an important role in improving the predictive performance of PTM sites (Durek, et al., 2009; Li, et al., 2016). Therefore, we calculated two commonly used protein 3D structure features, secondary structure features and solvent accessible area features, through the use of DSSP (Joosten, et al., 2011) and NACCESS (Hubbard, 1992), respectively. The result files can be downloaded from the download web page of PRISM (please refer to **Figure 1** (**B**) (http://prism.erc.monash.edu/pdb/download)).

2.3.1 Secondary structure features

Different types of features were extracted from the DSSP calculation result files for each PTM site, including secondary structure, bond angles (virtual bond angle, *KAPPA*), torsion angles (virtual torsion angle, *ALPHA*; IUPAC peptide backbone torsion angles, *PHI PSI*), atom coordinates (C_α atom coordinates, *X-CA Y-CA Z-CA*), and the number of water molecules (*ACC*) features. There are eight features for each PTM site shown on the corresponding entry page.

2.3.2 Solvent accessible area (ASA) features

Since solvent accessible area (ASA) has been widely used in many bioinformatics fields, including hot spots identification (Pan, et al., 2018), catalytic sites prediction (Song, et al., 2018), and glycosylation sites prediction (Li, et al., 2016), etc. Here, we employed the NACCESS program to calculate ASA features. The resulting file of NACCESS contains two types of ASA features, absolute ASA features and relative ASA features. Each type includes five different features: all-atoms, total-side, main-chain, non-polar, and all-polar ASA features. Thus, there are totally 10 ASA features for each PTM site. Please refer to the "Secondary structure and accessibilities features" panel of **Figure 2** (**C**).

2.4 Database access and download functionality

To obtain a database with abundant data, it is essential to apply a good search algorithm to provide users with more efficient and convenient access. As shown in the "Search function" panel of **Figure 1** (**B**), PRISM provides searching the entries in various ways, based on users' favour. The entries can be searched through three different methods, including ID search, Keyword search, and Advanced search. The Advanced search method in PRISM allows multiple search schemas specific to different types of data combined, e.g. PTM types and PDB ID, or PDB Chain, UniProt Acc and X-Ray resolution range. The search function is described in detail in Section 3.1, and an example is given to illustrate the usage of the advanced search function.

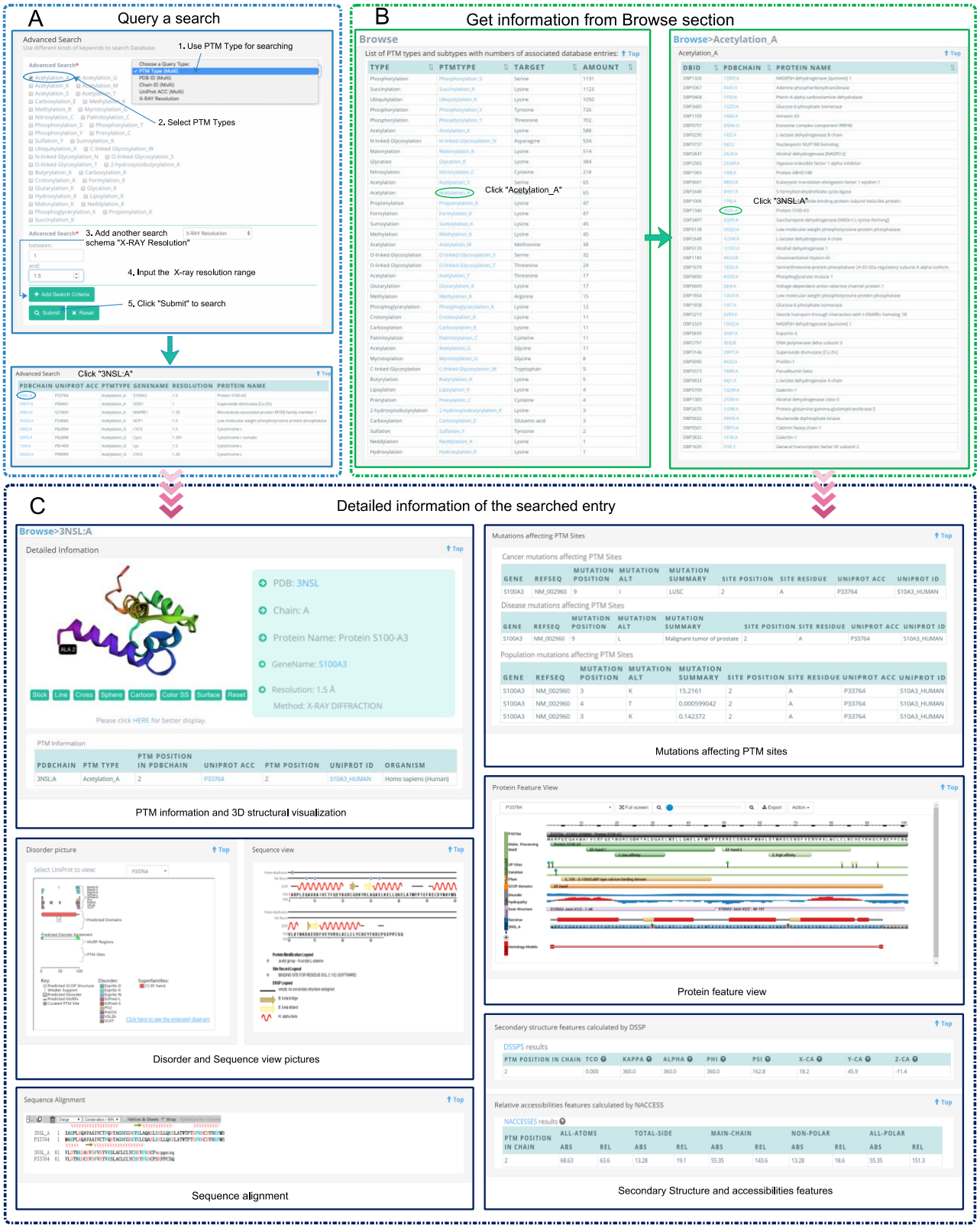


Fig. 2. Two different methods for data accessing in PRISM. **(A)** An example showing how to use the advanced function of PRISM. Here, two different search schemas were combined, PTM types (Acetylation A is selected) and the range of X-ray resolution (between 1 and 1.5); (**B**) An example illustrates how to explore the PTM data through the Browse webpage (Acetylation A and PDB chain 3NSL:A is selected); (**C**) Typical search result webpage using the PDB chain 3NSL:A as an example. The entry webpage consists of six major sections showing detailed information: PTM information and 3D structural visualization, mutations affecting PTM sites information, protein feature view, secondary structure and accessibilities features, disorder and sequence view pictures and sequence alignment.

PRISM also allows users to explore data embedded in the database through the "Browse" page. **Figure 2** (**B**) illustrates an example of how to explore the data this way. When users click the "Browse" icon on the top page, the web page redirects them to "Browse" page, which displays a table with four columns including type (e.g. Acetylation), PTM type (e.g. Acetylation_A), target Amino acid (e.g. Alanine), and the number of entries (e.g. 65). The table lists all the PTM types stored in PRISM, as shown in the left sub-figure of **Figure 2** (**B**). Then, when users click the PTM type (e.g. Acetylation_A), the web page redirects them to the browse page of this PTM. There users find a three-column table that includes database ID (e.g. DBP1340), PDB chain (e.g. 3NSL:A), and protein name (e.g. Protein S100-A3). Please refer to the right sub-figure of **Figure 2** (**B**) for more details. Both the "Search" results page and "Browse" page allow results to be sorted, thereby enabling users to easily find the information they need, immediately.

According to Chou's 5-step rule (Chen, et al., 2018; Chen, et al., 2018; Chou, 2011; Li, et al., 2018; Li, et al., 2018; Song, et al., 2018; Song, et al., 2018; Song, et al., 2018), in order to build a powerful bioinformatics prediction tool, the first step comes down to constructing a valid benchmark dataset and an independent test dataset. Accordingly, PRISM also provides the download function to facilitate users download of PTMs data in PRISM. As shown in the "Download page" panel of **Figure 1** (**B**), PRISM allows users to download the SQL file of the database, calculated DSSP feature files, calculated NACCESS feature files, and PTM sites data. These files can greatly facilitate bioinformaticians in their efforts to build machine-learning based bioinformatics methods for predicting PTMs sites straight from the protein structure level.

2.5 Implementation

The information stored in PRISM resides in a MySQL relational database consisting of 49 tables. All the routines in PRISM were written in the Java programming language. A highly user-friendly web front-end to the data was implemented using the jQuery framework. The basic operations of web interfaces are supported by JavaScript and jQuery, and the web page frame was built utilizing the bootstrap framework (https://getbootstrap.com/). Apache Tomcat 7 resides on a Linux server handles serving request of data from users, utilizing SSM (Spring, SpringMVC and MyBatis) framework for data searching, operation, web page hopping and viewing. The Linux server equipped with an 8-core CPU, 30 GB hard disk and 6 GB memory.

For the statistics page, the ECharts (https://ecomfe.github.io/echarts-doc/public/en/index.html) package (Li, et al., 2018) was used to achieve a user interactive pie charts display of statistical results. PRISM provides 11 different statistical results; users can view different statistics' pie charts by selecting the different statistical methods from the drop-down menus. Additionally, when users move the mouse over a sector of a pie chart, the corresponding statistical results are displayed right where the cursor points. By default, the pie chart shows the statistical results of the top-five types. However, users can select the types to be displayed by clicking the legend on the left.

For the entry page, 3Dmol.js (Rego and Koes, 2015) was employed for protein 3D structure visualization. 3Dmol.js is a high-performance, pure object-oriented JavaScript library, enabling interactive visualization of protein 3D structures. We opted to use 3Dmod.js because it does not need to install browser plug-ins or Java. In PRISM, each entry is a specific PDB chain. Therefore, only the corresponding chain of the PDB structure is displayed, and the PTM sites on the chain are labelled at the corresponding positions. Please refer to the "PTM information and 3D structural visualization" panel in **Figure 2** (**C**), where the PTM site is labelled at position 2 of the chain A as "ALA 2". A Scalable Vector Graphics (SVG) library, the Protein Feature view (http://andreasprlic.github.io/proteinfeatureview/), was applied to show variation data, Pfam domains, functional sites and some basic structural information. This visualization library enables users to quickly identify PTMs that lie within functional domains and to visualize the topology of the PTMs. The "Protein feature view" panel in **Figure 2** (**C**) illustrates this function. The RCSB-sequence viewer plug-in (https://github.com/biojava/rcsb-sequenceviewer) was also employed to generate the sequence view picture of the PDB chain. The sequence view shows the protein modification generated by BioJavaMolFinder, PDB chain sequence and secondary structures from DSSP. In addition, we also extracted the protein disorder figures from the disordered protein predictions database D2P2 (Oates, et al., 2013). An example of the sequence view and protein disorder pictures are shown in the "Disorder and Sequence view pictures" panel in **Figure 2** (**C**). Lastly, Alignment-To-HTML (Gille, et al., 2014) was used for interactive visualization for the sequence alignments between the PDB chain sequence and the corresponding UniProt sequences.

3 Results

3.1 Search functionality

As we discussed in section 2.4, PRISM provides several search algorithms, enabling users to get easy, comprehensive access to data. Here we introduce the search function of PRISM in detail, and present a specific example to show how to use the advanced search function.

PRISM data can be searched by three different search methods. The first method is the ID search, where the users can search PRISM entries by using a PDB ID (e.g. '3NSL'), UniProt ACC (e.g. 'P33764') or Database ID (e.g. 'DBP1340'). The second method is the keyword search, where users can search entries by entering a protein name or gene name. For these two search methods, a corresponding 'Example' button is provided, to guide users to conduct a search. With the click of the 'Example' button, users can automatically get an example for a search with ID/keyword. The third method is the Advanced Search. There are five different query schemas in the Advanced search field, including (i) PTM types, (ii) PDB IDs, (iii) Chain IDs, (iv) UniProt ACCs; and (v) range of X-Ray resolution. Users can search with individual schema or combinations of multiple schemas by selecting the search schema in the dropdown list of the 'Advanced search' menu and using the "Add Search Schema" button to search with multiple search schemas. **Figure 2** (**A**) illustrates an example using the combination of PTM types ("Acetylation_A" was selected) and X-ray resolution range (1 to 1.5 were filled) to conduct the search.

Searched results will be shown in a table on a separate webpage. The table will contain six columns, showing basic information of searched queries, including PDB ID, chain ID, PTM positions in PDB chain, the UniProt accession number (Acc), PTM positions in UniProt sequence, organism, PTM type, gene name, PDB resolution, and protein name (as shown in **Figure 2** (**A**)). Moreover, detailed information of each entry can be browsed in a new webpage by clicking the PDB Chain in the result table. For each entry, there are eight major types of information provided, including protein 3D structure visualization, PTM information, mutations affecting PTM sites information, secondary structure based features calculated by DSSP, relative accessibilities features calculated by NACCESS, protein feature view, sequence alignment, disorder picture, and protein sequence view. To provide a demonstration of the annotation information for each entry in PRISM, 'PDBChain=3NSL:A'

(PRISM ID='DBP1340') was used as an example query. The resulting page with annotation information is shown in **Figure 2** (**C**).

3.2 Database contents

With the data contained in PRISM, we conducted a statistical analysis of the PTM information distribution across the top eight species with the largest number of entries/PTM sites. Two bar charts (**Figure 3** (**A**) and (**B**)) show statistical results of the eight species with the largest number of entries (PDB chain) and PTM sites, respectively. The most abundant source with 11,560 PTM sites on 2,171 PDB chains was Human, followed by *E. coli* with 2,246 PTM sites on 549 entries, Yeast with 922 PTM sites on 275 entries, and Mouse with 877 PTM sites on 219 entries.

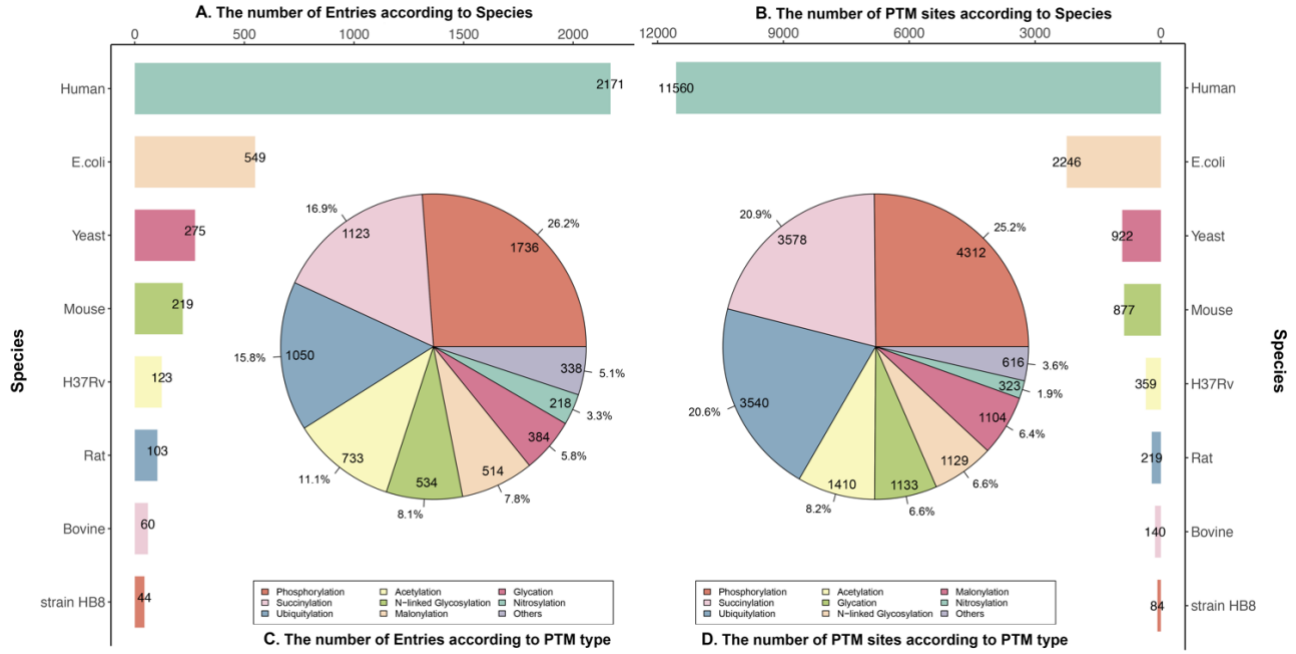


Fig. 3. Statistical analysis results in terms of species and PTM types. (**A**) Statistical results of the entries (PDB chains) from the eight largest species, including *Homo sapiens*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Mus musculus*, *Mycobacterium tuberculosis*, *Rattus norvegicus*, *Bos taurus*, and *Thermus thermophiles*; (**B**) Statistical analysis of the PTM sites from the eight largest species; (**C**) Distributions of eight largest PTM types in terms of the number of PTM entries; and (**D**) Distributions of eight largest PTM types in terms of the number of PTM sites.

Two pie charts in **Figure 3** (**C** and **D**) indicate the distribution of entries and PTM sites of the eight largest number of PTM types, respectively. In general, the top eight largest types of PTM are the same, but the order is not exactly the same, in terms of the number of PTM proteins (entries) and the number of PTM sites. For example, the top two PTM types in terms of the number of proteins and PTM sites are the same, i.e. Phosphorylation and Acetylation respectively. There are 1736 phosphorylated proteins with 4312 phosphorylation sites and 1123 acetylated proteins with 3578 acetylation sites. The third largest type in terms of the PTM proteins number is Glycation (1050), while the third largest type in terms of the PTM sites number is Malonylation (3540).

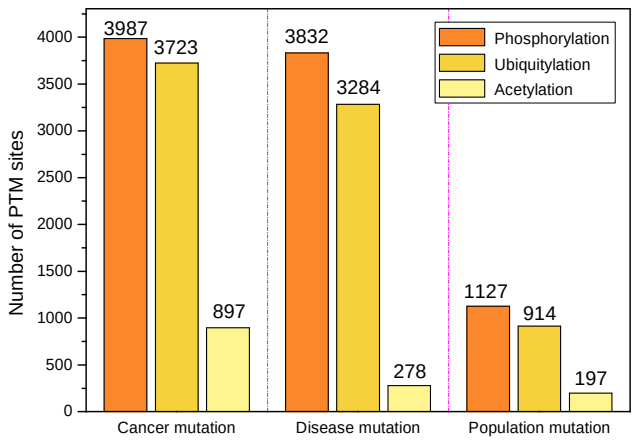


Fig. 4. The top three PTMs in terms of the number of PTM sites associated with mutations. There are three types of mutations, including somatic cancer mutations, inherited disease mutations, and population mutations.

The bar chart in **Figure 4** illustrates the number of PTM sites associated with mutations. There are three types of mutations, including somatic cancer mutations, inherited disease mutations, and population mutations. **Supplementary Tables 2 and 3** list the cancer types and disease types of somatic cancer mutations and disease mutations, respectively.

The top three types of PTMs associated with three types of mutations are the same. They are: (i) Phosphorylation with 3987 cancer mutations, 3832 disease mutations, and 1127 population mutations; (ii) Ubiquitylation with 3723 cancer mutations, 3284 disease mutations, and 914 acetylation mutations; and (iii) Acetylation with 897 cancer mutations, 278 disease mutations and 197 population mutations.

4 CONCLUSIONS

PRISM is a freely available web-based resource providing 3D structure information of protein post-translational modification sites. PRISM retrieves comprehensive PTM information from six major protein sequence-based PTM databases. The collected protein sequences with PTM sites were then mapped to protein 3D structures. PRISM not only includes PTM sites data but also integrates and annotates the disease/cancer/population mutation affecting the PTM sites. Other information, such as protein secondary structure properties, solvent accessibility area features, protein disorder region, and protein domain context, etc., were also collected and annotated in PRISM.

PRISM combines the strength of Java, Spring, SpringMVC, MyBatis, and JavaScript to achieve user-friendly interface design and efficient data query/display. Multiple data querying methods and data browsing pages enabled in PRISM can be explored, with a number of options. PRISM is also equipped with several bioinformatics plug-ins such as 3Dmol.js, Protein Feature view, and RCSB-sequence viewer to better visualize the comprehensive annotation data contained in PRISM. As new datasets accumulate rapidly, as well as computational resources, it is essential to maintain the database current. Thus, we plan to develop a website background management system to facilitate timely and automatized updates of the database.

PRISM is unique among community-wide online resources due to the comprehensiveness of PTM information on real 3D structures contained in it. We anticipate that PRISM will become a useful resource for both experimental and computational analyses of protein post-translational modifications.

Acknowledgements

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) (1144652 and 1127948), the Australian Research Council (ARC) (LP110200333 and DP120104460), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965), and a Monash Major Inter-Disciplinary Research (IDR) Grant. TML and AL's work was supported in part by the Informatics Institute of the School of Medicine at UAB. J.L. is an NHMRC Principal Research Fellow.

Conflict of Interest: none declared.

References

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389-3402.

Berman, H.M., *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):235-242.

Botstein, D. and Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;33 Suppl:228-237.

Chen, Z., *et al.* Large-scale comparative assessment of computational predictors for lysine post-translational modification sites. *Brief Bioinform* 2018:bby089-bby089.

Chen, Z., *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;34(13):2499-2502.

Chen, Z., *et al.* Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features. *Brief Bioinform* 2015;16(4):640-657.

Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J Theor Biol* 2011;273(1):236-247.

Chuang, G.Y., *et al.* Computational prediction of N-linked glycosylation incorporating structural properties and patterns. *Bioinformatics* 2012;28(17):2249-2255.

Dinkel, H., *et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* 2011;39(Database issue):D261-267.

Duan, G. and Walther, D. The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput Biol* 2015;11(2):e1004049.

Durek, P., *et al.* Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* 2009;10:117.

Durek, P., *et al.* Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* 2009;10.

Dwek, R.A. Biological importance of glycosylation. *Dev Biol Stand* 1998;96:43-47.

Fleuren, E.D.G., *et al.* The kinome 'at large' in cancer. *Nat Rev Cancer* 2016;16(2):83-98.

Gao, J., *et al.* BioJava-ModFinder: identification of protein modifications in 3D structures from the Protein Data Bank. *Bioinformatics* 2017;33(13):2047-2049.

Gille, C., Birgit, W. and Gille, A. Sequence alignment visualization in HTML5 without Java. *Bioinformatics* 2014;30(1):121-122.

Glozak, M.A., *et al.* Acetylation and deacetylation of non-histone proteins. *Gene* 2005;363:15-23.

Hornbeck, P.V., *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 2012;40(Database issue):D261-270.

Hornbeck, P.V., *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 2015;43(Database issue):D512-520.

Huang, K.Y., *et al.* dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic Acids Res* 2016;44(D1):D435-446.

Hubbard, S. NACCESS: program for calculating accessibilities. *Department of Biochemistry and Molecular Biology, University College of London* 1992.

Jia, C.Z., Zuo, Y. and Zou, Q. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* 2018;34(12):2029-2036.

Joosten, R.P., *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res* 2011;39(Database issue):D411-419.

Karaca, M., *et al.* Mutation of androgen receptor N-terminal phosphorylation site Tyr-267 leads to inhibition of nuclear translocation and DNA binding. *PLoS One* 2015;10(5):e0126270.

Kim, Y., *et al.* Detection and analysis of disease-associated single nucleotide polymorphism influencing post-translational modification. *BMC Med Genomics* 2015;8 Suppl 2:S7.

Krassowski, M., *et al.* ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. *Nucleic Acids Res* 2018;46(D1):D901-D910.

Lee, T.Y., *et al.* dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res* 2006;34(Database issue):D622-627.

Li, D., *et al.* ECharts: A declarative framework for rapid construction of web-based visualization. *Visual Informatics* 2018.

Li, F., *et al.* Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics* 2018:bty522-bty522.

Li, F., *et al.* GlycoMine(struct): a new bioinformatics tool for highly accurate mapping of the human N-linked and O-linked glycoproteomes by incorporating structural features. *Sci Rep* 2016;6:34595.

Li, F., *et al.* GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* 2015;31(9):1411-1419.

Li, F., *et al.* Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: a comprehensive revisit and benchmarking of existing methods. *Brief Bioinform* 2018:bby077-bby077.

Li, H., *et al.* SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol Cell Proteomics* 2009;8(8):1839-1849.

Li, J., *et al.* SysPTM 2.0: an updated systematic resource for post-translational modification. *Database (Oxford)* 2014;2014:bau025.

Marino, K., *et al.* A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* 2010;6(10):713-723.

McCubrey, J.A., *et al.* Serine/threonine phosphorylation in cytokine signal transduction. *Leukemia* 2000;14(1):9-21.

Medzihradszky, K.F. Peptide sequence analysis. *Methods Enzymol* 2005;402:209-244.

Moharir, A., *et al.* The role of N-glycosylation in folding, trafficking, and functionality of lysosomal protein CLN5. *Plos One* 2013;8(9):e74299.

Nishi, H., Hashimoto, K. and Panchenko, A.R. Phosphorylation in protein-protein binding: effect on stability and function. *Structure* 2011;19(12):1807-1815.

Oates, M.E., *et al.* D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res* 2013;41(Database issue):D508-516.

Pan, Y., *et al.* Computational identification of binding energy hot spots in protein-RNA complexes using an ensemble approach. *Bioinformatics* 2018;34(9):1473-1480.

Pickart, C.M. Ubiquitin enters the new millennium. *Mol Cell* 2001;8(3):499-504.

Pinho, S.S. and Reis, C.A. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer* 2015;15(9):540-555.

Polevoda, B. and Sherman, F. The diversity of acetylated proteins. *Genome Biol* 2002;3(5):reviews0006.

Rego, N. and Koes, D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* 2015;31(8):1322-1324.

Song, J., *et al.* PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 2018;34(4):684-687.

Song, J., *et al.* PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. *J Theor Biol* 2018;443:125-137.

Song, J., *et al.* iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief Bioinform* 2018:bby028-bby028.

Swaney, D.L., *et al.* Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nat Methods* 2013;10(7):676-682.

Taherzadeh, G., *et al*. Predicting lysine-malonylation sites of proteins using sequence and predicted structural features. *J Comput Chem* 2018;39(22):1757-1763.

UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;46(5):2699.

von der Lieth, C.W., *et al*. Bioinformatics for glycomics: status, methods, requirements and perspectives. *Brief Bioinform* 2004;5(2):164-178.

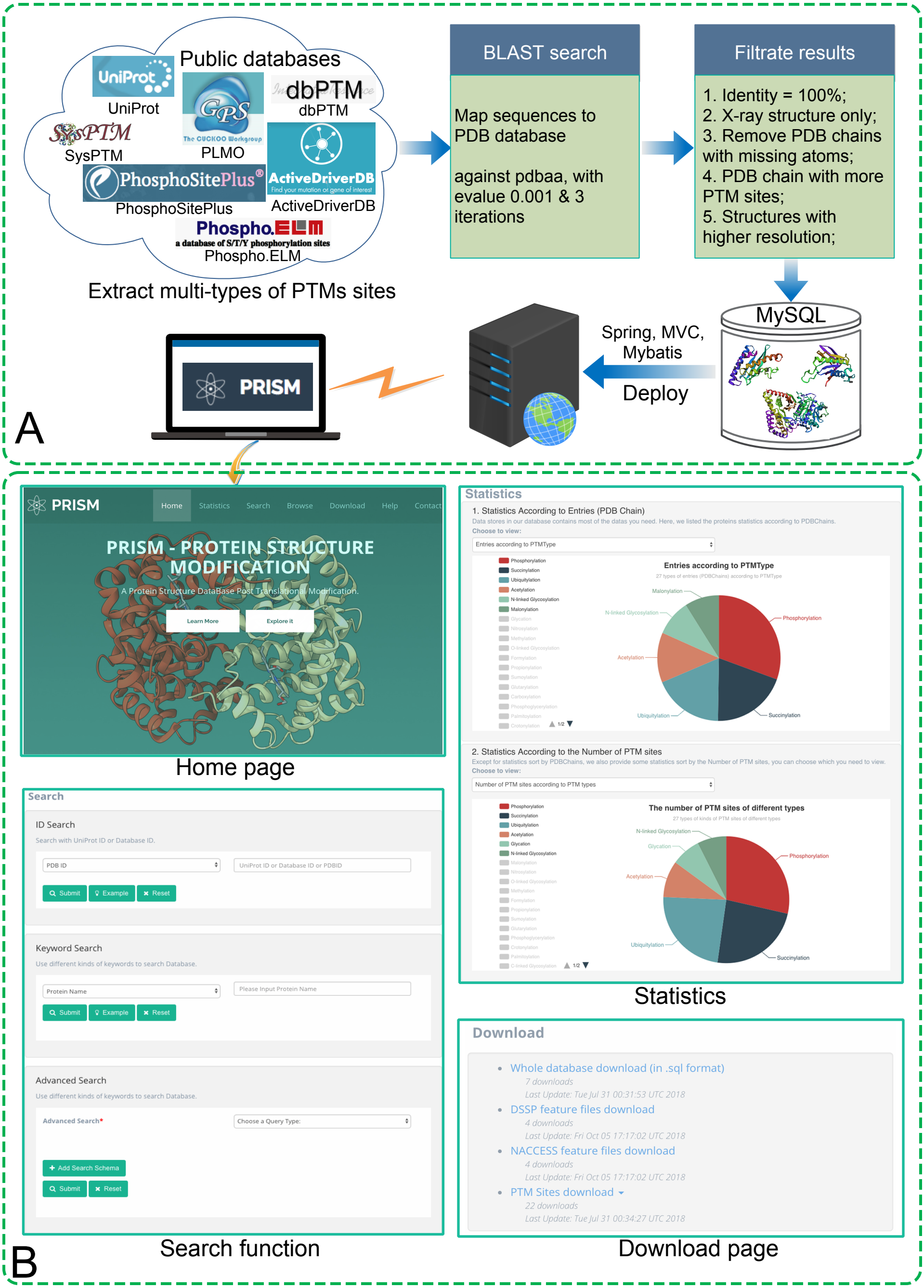
Wagih, O., Reimand, J. and Bader, G.D. MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat Methods* 2015;12(6):531-533.

Xu, H., *et al*. PLMD: An updated data resource of protein lysine modifications. *J Genet Genomics* 2017;44(5):243-250.

Xue, Y., *et al*. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics* 2008;7(9):1598-1608.

Yip, Y.L., *et al*. The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum Mutat* 2004;23(5):464-470.

Zhao, S., *et al*. Regulation of cellular metabolism by protein lysine acetylation. *Science* 2010;327(5968):1000-1004.


Public databases
UniProt
dbPTM
SysPTM
PLMO
PhosphoSitePlus
ActiveDriverDB
Phospho.ELM
Extract multi-types of PTMs sites
BLAST search
Map sequences to PDB database
against pdbaa, with evalue 0.001 & 3 iterations
Filtrate results
1. Identity = 100%;
2. X-ray structure only;
3. Remove PDB chains with missing atoms;
4. PDB chain with more PTM sites;
5. Structures with higher resolution;
MySQL
Spring, MVC, Mybatis
Deploy
PRISM
A
PRISM - PROTEIN STRUCTURE MODIFICATION
A Protein Structure DataBase Post Translational Modification.
Home page
Statistics
1. Statistics According to Entries (PDB Chain)
2. Statistics According to the Number of PTM sites
Statistics
Search
ID Search
Keyword Search
Advanced Search
Search function
Download
Whole database download (in .sql format)
DSSP feature files download
NACCESS feature files download
PTM Sites download
Download page
B

A Query a search

Advanced Search
Use different kinds of keywords to search Database.

1. Use PTM Type for searching

Choose a Query Type:
- ✓ PTM Type (Multi)
- PDB ID (Multi)
- Chain ID (Multi)
- UniProt ACC (Multi)
- X-RAY Resolution

Advanced Search*

☑ Acetylation_A ☐ Acetylation_G
☐ Acetylation_K ☐ Acetylation_M
☐ Acetylation_S ☐ Acetylation_T
☐ Carboxylation_E ☐ Methylation_K
☐ Methylation_R ☐ Myristoylation_G
☐ Nitrosylation_C ☐ Palmitoylation_C
☐ Phosphorylation_S ☐ Phosphorylation_T
☐ Phosphorylation_Y ☐ Prenylation_C
☐ Sulfation_Y ☐ Sumoylation_K
☐ Ubiquitylation_K ☐ C-linked Glycosylation_W
☐ N-linked Glycosylation_N ☐ O-linked Glycosylation_S
☐ O-linked Glycosylation_T ☐ 2-hydroxyisobutyrylation_K
☐ Butyrylation_K ☐ Carboxylation_K
☐ Crotonylation_K ☐ Formylation_K
☐ Glutarylation_K ☐ Glycation_K
☐ Hydroxylation_K ☐ Lipoylation_K
☐ Malonylation_K ☐ Neddylation_K
☐ Phosphoglycerylation_K ☐ Propionylation_K
☐ Succinylation_K

2. Select PTM Types

Advanced Search* X-RAY Resolution

3. Add another search schema "X-RAY Resolution"

between: 1
and: 1.5

4. Input the X-ray resolution range

+ Add Search Criteria

Submit Reset

5. Click "Submit" to search

Advanced Search — Click "3NSL:A" — Top

PDBCHAIN	UNIPROT ACC	PTMTYPE	GENENAME	RESOLUTION	PROTEIN NAME
3NSL:A	P33764	Acetylation_A	S100A3	1.5	Protein S100-A3
2WYT:A	P00441	Acetylation_A	SOD1	1	Superoxide dismutase [Cu-Zn]
2R8U:A	Q15691	Acetylation_A	MAPRE1	1.35	Microtubule-associated protein RP/EB family member 1
5KQG:A	P24666	Acetylation_A	ACP1	1.5	Low molecular weight phosphotyrosine protein phosphatase
2B4Z:A	P62894	Acetylation_G	CYCS	1.5	Cytochrome c
5DF5:A	P62898	Acetylation_G	Cycs	1.301	Cytochrome c somatic
1I54:A	P81459	Acetylation_G	cyc	1.5	Cytochrome c
3ZOO:A	P99999	Acetylation_G	CYCS	1.35	Cytochrome c

B Get information from Browse section

Browse

List of PTM types and subtypes with numbers of associated database entries: Top

TYPE	PTMTYPE	TARGET	AMOUNT
Phosphorylation	Phosphorylation_S	Serine	1131
Succinylation	Succinylation_K	Lysine	1123
Ubiquitylation	Ubiquitylation_K	Lysine	1050
Phosphorylation	Phosphorylation_Y	Tyrosine	726
Phosphorylation	Phosphorylation_T	Threonine	702
Acetylation	Acetylation_K	Lysine	588
N-linked Glycosylation	N-linked Glycosylation_N	Asparagine	534
Malonylation	Malonylation_K	Lysine	514
Glycation	Glycation_K	Lysine	384
Nitrosylation	Nitrosylation_C	Cysteine	218
Acetylation	Acetylation_S	Serine	65
Acetylation	Acetylation_A	[illegible]	65
Propionylation	Propionylation_K	Lysine	47
Formylation	Formylation_K	Lysine	47
Sumoylation	Sumoylation_K	Lysine	45
Methylation	Methylation_K	Lysine	45
Acetylation	Acetylation_M	Methionine	38
O-linked Glycosylation	O-linked Glycosylation_S	Serine	32
O-linked Glycosylation	O-linked Glycosylation_T	Threonine	24
Acetylation	Acetylation_T	Threonine	17
Glutarylation	Glutarylation_K	Lysine	17
Methylation	Methylation_R	Arginine	15
Phosphoglycerylation	Phosphoglycerylation_K	Lysine	12
Crotonylation	Crotonylation_K	Lysine	11
Carboxylation	Carboxylation_K	Lysine	11
Palmitoylation	Palmitoylation_C	Cysteine	11
Acetylation	Acetylation_G	Glycine	11
Myristoylation	Myristoylation_G	Glycine	8
C-linked Glycosylation	C-linked Glycosylation_W	Tryptophan	5
Butyrylation	Butyrylation_K	Lysine	5
Lipoylation	Lipoylation_K	Lysine	4
Prenylation	Prenylation_C	Cysteine	4
2-hydroxyisobutyrylation	2-hydroxyisobutyrylation_K	Lysine	3
Carboxylation	Carboxylation_E	Glutamic acid	3
Sulfation	Sulfation_Y	Tyrosine	2
Neddylation	Neddylation_K	Lysine	1
Hydroxylation	Hydroxylation_K	Lysine	1

Click "Acetylation_A"

Browse>Acetylation_A

Acetylation_A — Top

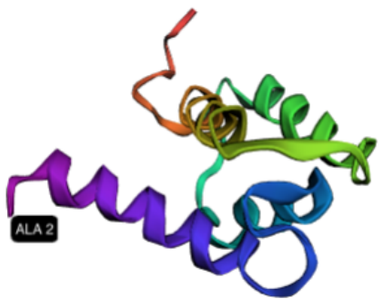
DBID	PDBCHAIN	PROTEIN NAME
DBP1326	1QRD:A	NAD(P)H dehydrogenase [quinone] 1
DBP3367	4X45:A	Adenine phosphoribosyltransferase
DBP0468	1F93:A	Pterin-4-alpha-carbinolamine dehydratase
DBP3485	1GZD:A	Glucose-6-phosphate isomerase
DBP1109	1A8A:A	Annexin A5
DBP0797	2NN6:G	Exosome complex component RRP40
DBP0290	1I0Z:A	L-lactate dehydrogenase B chain
DBP3737	5IJO:J	Nucleoporin NUP188 homolog
DBP2847	2ALR:A	Alcohol dehydrogenase [NADP(+)]
DBP2583	2XUM:A	Hypoxia-inducible factor 1-alpha inhibitor
DBP1965	1IMJ:A	Protein ABHD14B
DBP0681	4BVX:B	Eukaryotic translation elongation factor 1 epsilon-1
DBP3448	3HXT:A	5-formyltetrahydrofolate cyclo-ligase
DBP1006	1TRJ:A	[illegible] nucleotide-binding protein subunit beta-like protein
DBP1340	3NSL:A	Protein S100-A3
DBP3497	2Q99:A	Saccharopine dehydrogenase [NAD(+) L-lysine-forming]
DBP0138	5KQG:A	Low molecular weight phosphotyrosine protein phosphatase
DBP2648	1LDM:A	L-lactate dehydrogenase A chain
DBP0139	1CDO:A	Alcohol dehydrogenase 1
DBP1185	4ZG4:B	Unconventional myosin-Vc
DBP1679	1B3U:A	Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform
DBP0890	4GPZ:A	Phosphoglycerate mutase 1
DBP0609	2JK4:A	Voltage-dependent anion-selective channel protein 1
DBP1854	1DG9:A	Low molecular weight phosphotyrosine protein phosphatase
DBP1858	1IAT:A	Glucose-6-phosphate isomerase
DBP2215	2V8S:V	Vesicle transport through interaction with t-SNAREs homolog 1B
DBP3329	1DXQ:A	NAD(P)H dehydrogenase [quinone] 1
DBP0699	3A6P:A	Exportin-5
DBP2791	3E0J:B	DNA polymerase delta subunit 3
DBP3146	2WYT:A	Superoxide dismutase [Cu-Zn]
DBP0096	4X25:A	Profilin-1
DBP0315	1B8R:A	Parvalbumin beta
DBP0833	4AJ1:A	L-lactate dehydrogenase A chain
DBP0709	1GZW:A	Galectin-1
DBP1305	2FZW:A	Alcohol dehydrogenase class-3
DBP2670	1L9M:A	Protein-glutamine gamma-glutamyltransferase E
DBP0632	3WX8:A	Nucleoside diphosphate kinase
DBP0501	1BPO:A	Clathrin heavy chain 1
DBP3832	1A78:A	Galectin-1
DBP1635	5IY6:T	General transcription factor IIF subunit 2

Click "3NSL:A"

C Detailed information of the searched entry

Browse>3NSL:A

Detailed Infomation — Top



Stick Line Cross Sphere Cartoon Color SS Surface Reset

Please click HERE for better display.

- PDB: 3NSL
- Chain: A
- Protein Name: Protein S100-A3
- GeneName: S100A3
- Resolution: 1.5 Å
  Method: X-RAY DIFFRACTION

PTM Information

PDBCHAIN	PTM TYPE	PTM POSITION IN PDBCHAIN	UNIPROT ACC	PTM POSITION	UNIPROT ID	ORGANISM
3NSL:A	Acetylation_A	2	P33764	2	S10A3_HUMAN	Homo sapiens (Human)

PTM information and 3D structural visualization

Mutations affecting PTM Sites — Top

Cancer mutations affecting PTM Sites

GENE	REFSEQ	MUTATION POSITION	MUTATION ALT	MUTATION SUMMARY	SITE POSITION	SITE RESIDUE	UNIPROT ACC	UNIPROT ID
S100A3	NM_002960	9	I	LUSC	2	A	P33764	S10A3_HUMAN

Disease mutations affecting PTM Sites

GENE	REFSEQ	MUTATION POSITION	MUTATION ALT	MUTATION SUMMARY	SITE POSITION	SITE RESIDUE	UNIPROT ACC	UNIPROT ID
S100A3	NM_002960	9	L	Malignant tumor of prostate	2	A	P33764	S10A3_HUMAN

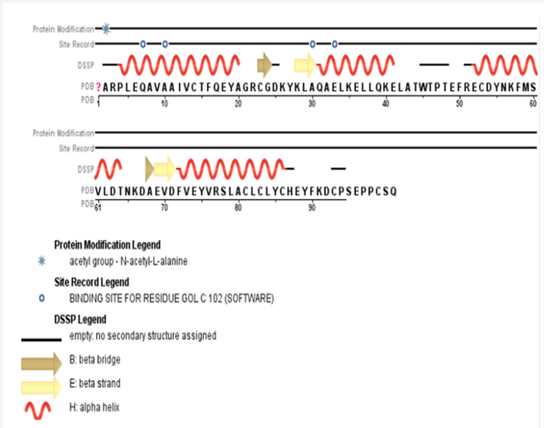
Population mutations affecting PTM Sites

GENE	REFSEQ	MUTATION POSITION	MUTATION ALT	MUTATION SUMMARY	SITE POSITION	SITE RESIDUE	UNIPROT ACC	UNIPROT ID
S100A3	NM_002960	3	K	15.2161	2	A	P33764	S10A3_HUMAN
S100A3	NM_002960	4	T	0.000599042	2	A	P33764	S10A3_HUMAN
S100A3	NM_002960	3	K	0.142372	2	A	P33764	S10A3_HUMAN

Mutations affecting PTM sites

Disorder picture — Top

Sequence view — Top

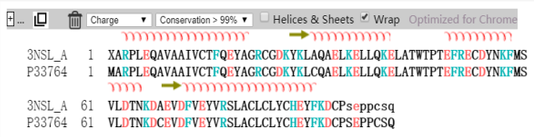


Disorder and Sequence view pictures

Protein Feature View — Top

Protein feature view

Sequence Alignment — Top



Sequence alignment

Secondary structure features calculated by DSSP — Top

DSSPS results

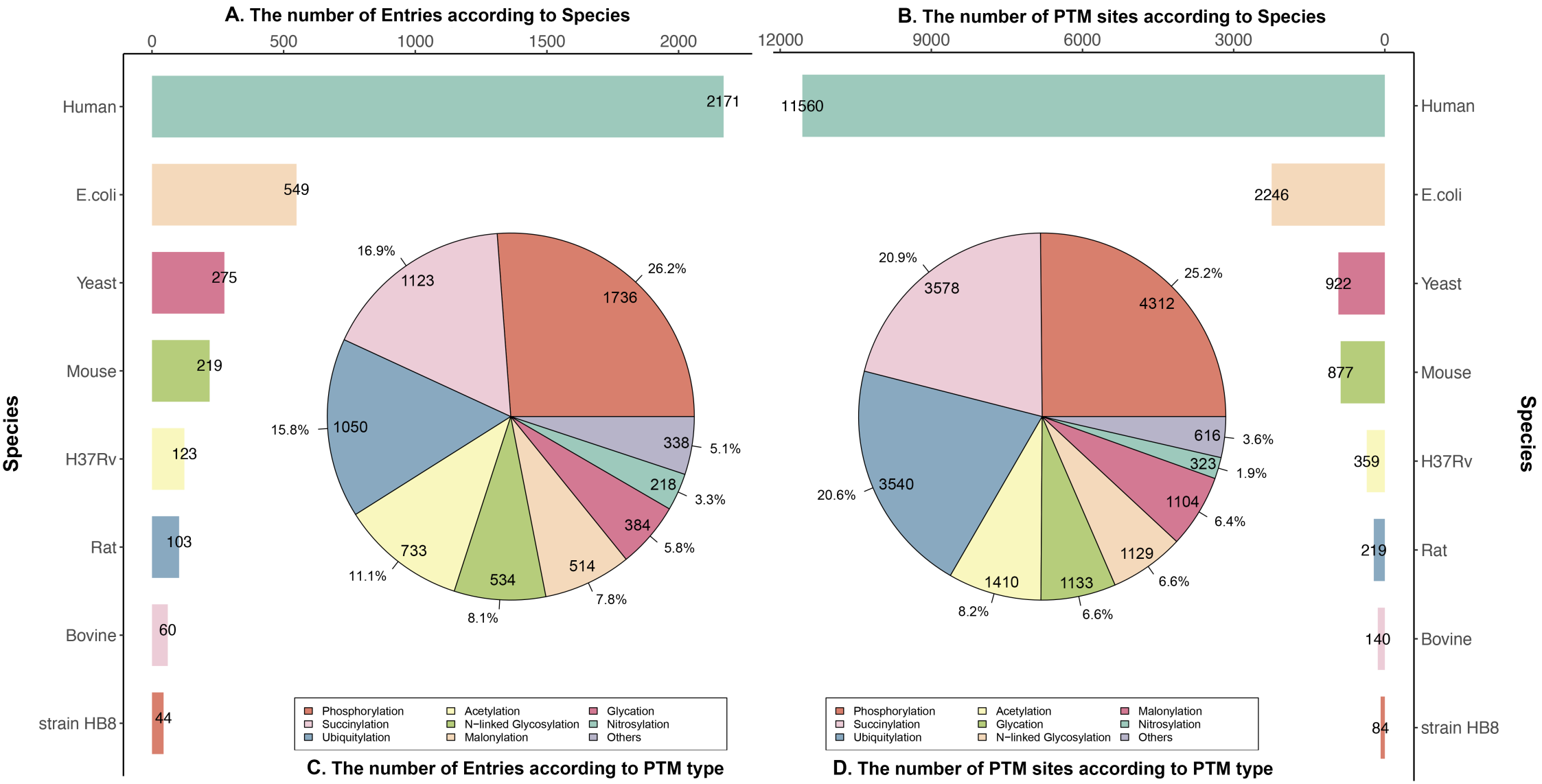
PTM POSITION IN CHAIN	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA
2	0.000	360.0	360.0	360.0	162.8	18.2	45.9	-11.4

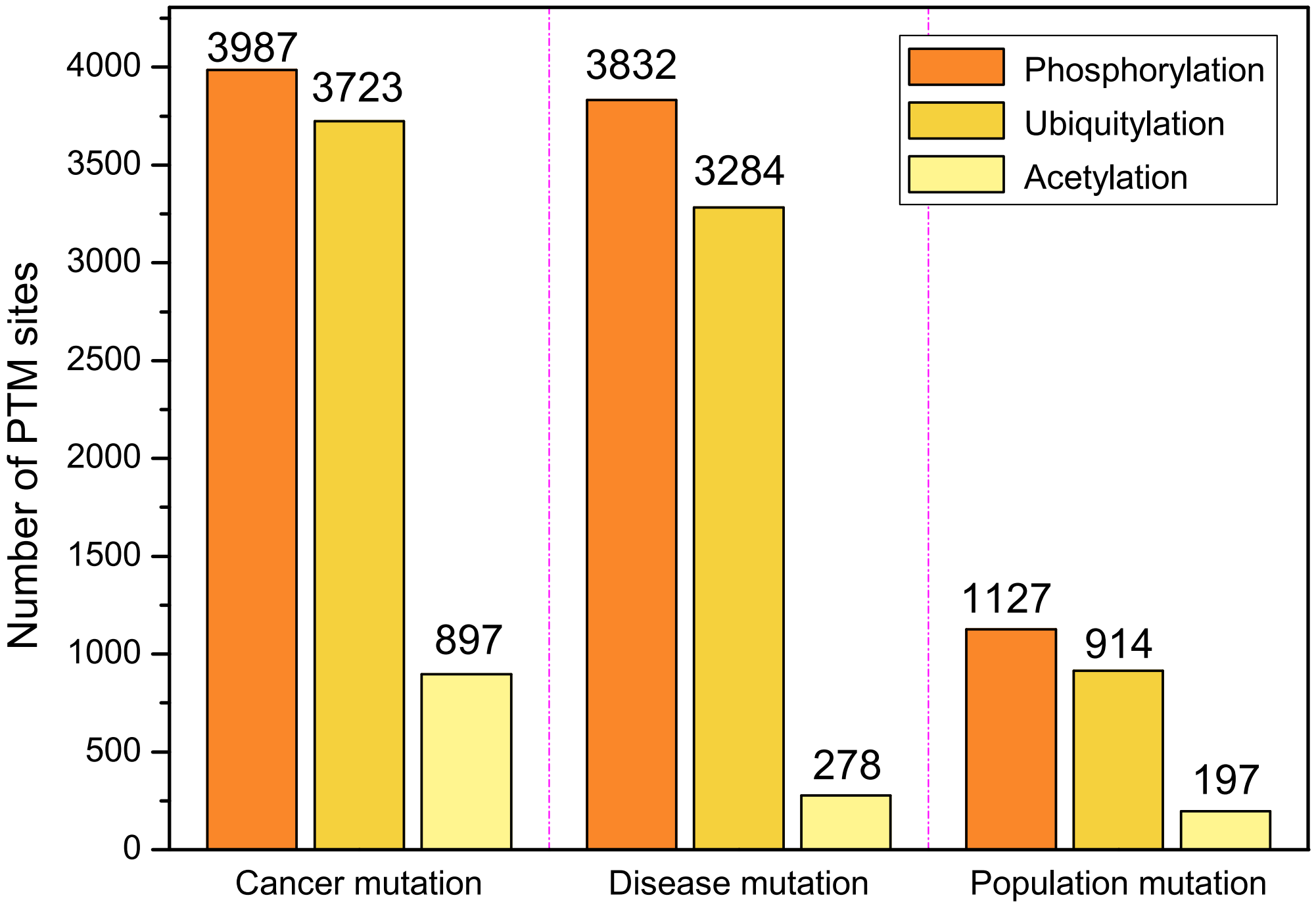
Relative accessibilities features calculated by NACCESS — Top

NACCESSES results

PTM POSITION IN CHAIN	ALL-ATOMS		TOTAL-SIDE		MAIN-CHAIN		NON-POLAR		ALL-POLAR	
	ABS	REL	ABS	REL	ABS	REL	ABS	REL	ABS	REL
2	68.63	63.6	13.28	19.1	55.35	143.6	13.28	18.6	55.35	151.3

Secondary Structure and accessibilities features


A. The number of Entries according to Species
0
500
1000
1500
2000
Species
Human
E.coli
Yeast
Mouse
H37Rv
Rat
Bovine
strain HB8
2171
549
275
219
123
103
60
44
B. The number of PTM sites according to Species
12000
9000
6000
3000
0
11560
2246
922
877
359
219
140
84
C. The number of Entries according to PTM type
1736
26.2%
1123
16.9%
1050
15.8%
733
11.1%
534
8.1%
514
7.8%
384
5.8%
218
3.3%
338
5.1%
Phosphorylation
Succinylation
Ubiquitylation
Acetylation
N−linked Glycosylation
Malonylation
Glycation
Nitrosylation
Others
D. The number of PTM sites according to PTM type
4312
25.2%
3578
20.9%
3540
20.6%
1410
8.2%
1133
6.6%
1129
6.6%
1104
6.4%
323
1.9%
616
3.6%
Phosphorylation
Succinylation
Ubiquitylation
Acetylation
Glycation
N−linked Glycosylation
Malonylation
Nitrosylation
Others


Number of PTM sites
0
500
1000
1500
2000
2500
3000
3500
4000
3987
3723
897
3832
3284
278
1127
914
197
Phosphorylation
Ubiquitylation
Acetylation
Cancer mutation
Disease mutation
Population mutation