Genomics Proteomics Bioinformatics 13 (2015) 36–39

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

RESOURCE REVIEW

Web Resources for Mass Spectrometry-based Proteomics

Tao Chen [1,a], Jie Zhao [2,b], Jie Ma [1,c], Yunping Zhu [1,*,d]

[1] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China*
[2] *Biological Information College, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*

Received 2 January 2015; revised 22 January 2015; accepted 28 January 2015
Available online 23 February 2015

Handled by Siqi Liu

KEYWORDS

Mass spectrometry;
Proteomics;
Web resources

Abstract With the development of high-resolution and high-throughput mass spectrometry (MS) technology, a large quantum of proteomic data is continually being generated. Collecting and sharing these data are a challenge that requires immense and sustained human effort. In this report, we provide a classification of important web resources for MS-based proteomics and present rating of these web resources, based on whether raw data are stored, whether data submission is supported, and whether data analysis pipelines are provided. These web resources are important for biologists involved in proteomics research.

Introduction

The advancement of tandem mass spectrometry (MS) has made it possible to identify hundreds of thousands of proteins in MS-based experiments [1]. With the development of a wide range of methods for spectrometry and data analysis, MS-based proteomics has gained popularity in biomedical research. The vastly-expanding research using tandem MS technology is continually generating large amounts of proteomics data. Collecting these datasets is undoubtedly becoming crucial to the research community. Proteomics data repository contains a proteome with high coverage and sufficient data content for statistical analysis, and provides extensive observational data for genome annotation projects as well. However, maintaining such data repository is challenging due to the diversity and quantum of data as well as varying needs of different users. In this report, we describe web data repositories for MS-based proteomics and rate them based on their score against parameters such as storage of raw data, data submission support, and provision of data analysis pipelines. The main features of these resources are shown in **Table 1**. Based on their focus areas within proteomic research, we classified these resources into 3 categories: general proteomics data repositories, quantitative proteomics data repositories, and proteomics data repositories focusing on protein post-translational modifications (PTMs).

* Corresponding author.
E-mail: zhuyunping@gmail.com (Zhu Y).
[a] ORCID: 0000-0001-8191-4135.
[b] ORCID: 0000-0002-6275-7506.
[c] ORCID: 0000-0002-8934-922X.
[d] ORCID: 0000-0002-7320-7411.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.01.004

Table 1 List of major MS-based proteomics resources

Category	Name	Link	Main features	Rating	Refs.
General	PRIDE	http://www.ebi.ac.uk/pride/archive	Supports raw data storage and data submission	★★★★★	[3]
	PeptideAtlas	http://www.peptideatlas.org	Supports raw data storage, data submission, and data analysis	★★★★★	[1]
	Human Proteinpedia	http://www.humanproteinpedia.org	Supports raw data storage and data submission	★★★★☆	[4]
	iProX	http://iprox.hupo.org.cn	Supports raw data storage, data submission, and data analysis	★★★★★	
	Tranche	https://proteomecommons.org/tranche	Supports raw data storage and data submission	★★★☆☆	[5]
	GPMDB	http://www.thegpm.org	Supports data analysis	★★★☆☆	[6]
	MOPED	http://moped.proteinspire.org	Stores protein expression information from MS-based proteomics experiments	★★★☆☆	[7]
	YPED	http://yped.med.yale.edu	An integrated bioinformatics suite and database for proteomics research	★★★★☆	[8,9]
Quantitative PTMs-focused	PaxDb	http://pax-db.org	Supports quantitative proteomics data storage	★★★☆☆	[10]
	Phospho.ELM	http://phospho.elm.eu.org	Supports phosphoproteomic MS data storage	★★★☆☆	[11]
	PhosphoSitePlus	http://www.phosphosite.org	Stores raw data and MS-reported PTM sites	★★★★☆	[12]
	dbPTM	http://dbptm.mbc.nctu.edu.tw	Stores raw data and MS/MS peptides associated with PTMs	★★★★☆	[13,14]
	PHOSIDA	http://www.phosida.com	Supports raw data storage and phosphoproteomic MS data storage	★★★★☆	[15,16]

Note: These web resources are rated based on their score against parameters such as storage of raw data, data submission support, and provision of data analysis pipelines.MS, mass spectrometry; PTM, post-translational modification.

General proteomics data repositories

Proteomics IDEntifications database

The Proteomics IDEntifications (PRIDE) database created by the European Bioinformatics Institute (EBI) is a web resource that collects MS-based proteomics data. By the end of 2014, PRIDE accumulated data for 41,835 proteins, 269,806 unique peptides, and about 101 million spectra [2]. PRIDE is one of the most popular proteomic data repositories that have played an important role in the nascent Human Proteome Project (HPP) [3].

PeptideAtlas

PeptideAtlas is a database that stores various formats of output files and metadata from MS-based experiments [1], it also allows users to submit raw data. These raw data are periodically analyzed for identification and statistical analysis purposes. The results are made available back to the researchers by web-based presentation systems. PeptideAtlas can help plan targeted proteomics experiments, improve genome annotation, and support data mining projects [1].

Human Proteinpedia

Human Proteinpedia is a resource to integrate, store, and share proteomic data [4]. It is a platform for collecting human proteomic data using a distributed annotation system, which allows the research community to contribute protein annotations. By the end of 2014, Human Proteinpedia has covered 15,231 proteins, 1,960,352 peptides, and about 5 million spectra [2]. It also provides a panorama of the human proteome.

iProX

iProX is an integrated proteome resources center based in China, which is built to support the worldwide sharing of proteomics data. Currently, iProX comprises an experiment data submission system and a proteome database. The iProX submission system is a public platform that was set up following the data-sharing policy of the ProteomeXchange consortium. Raw data and standardized meta-data from proteomics experiments can be collected and shared by using controlled vocabularies to describe the Minimum Information About a Proteomics Experiment (MIAPE). Registered users can choose to submit their proteomics datasets to iProX via public or private modes. Datasets submitted via the public mode are openly accessible, whereas private datasets can only be accessed by the authorized users. On the other hand, the iProX proteomics database was developed as a structured storage platform for data deposited in the system. iProX facilitates data analysis and sharing. Up till now, it has covered 46 projects, 190 subprojects, and 6441 data files.

Tranche

Tranche is a data repository targeting storage and sharing of information for proteomics researchers. It supports re-use and dissemination of both data and software. To reduce data

redundancy and achieve load balancing, it adopts peer-to-peer networking. It also uses a client–server model to ensure authentication and reliability. A client tool is required to upload and download datasets. It has several important features including pre-publication encryption, data pedigree, data integrity, immutability, and versioning. Tranche provides interfaces for PRIDE, Human Proteinpedia, and PeptideAtlas to store and disseminate large MS-based data files [5].

Global Proteome Machine Database

The Global Proteome Machine Database (GPMDB) is a resource for collecting diverse tandem mass spectra. It also includes peptide and protein identifications that are important for further MS computational research [6]. GPMDB provides a pipeline for reprocessing raw data submitted by users or imported from other repositories, thus generating XML files that store information about peptide and protein identification. Specifically, identified proteins are organized into separate spreadsheets for each chromosome and mitochondrial DNA. By the end of 2014, GPMDB data spans 136,373 proteins, 1,786,698 peptides, and 1020 million spectra [2]. GPMDB has played an important role in the Chromosome-Centric Human Proteome (C-HPP) Project.

Model Organism Protein Expression Database

The Model Organism Protein Expression Database (MOPED) is a proteomics repository that integrates protein expression information from MS-based proteomics experiments on human specimens and that from model organisms [7]. It also provides new estimates of protein abundance and concentration, and statistical summaries from experiments. Several search and visualization tools are available. By the end of 2014, MOPED has developed into a repository containing 17,141 proteins, 250,000 unique peptides, and approximately 15 million spectra [2], providing researchers with information on complex biological processes and thus supporting biomedical discovery.

Yale Protein Expression Database

The Yale Protein Expression Database (YPED) [8] is an integrated bioinformatics suite and database for proteomics research, which was significantly improved from the first version released in 2007 [9]. YPED supports many kinds of data including those from multiple MS instruments, different search engines, and labeled or label-free quantification. YPED is a web-accessible and user-friendly resource, designed to meet data management, archival, and analysis needs of high-throughput MS-based proteomics research.

Quantitative proteomics data repositories

PaxDb

PaxDb is a meta-database integrating whole-organism data and tissue-resolved data at absolute protein abundance levels for various model organisms. It imports quantitative proteomics data sets exclusively from published experiments and from primary proteomics data resources such as PRIDE and PeptideAtlas, and then analyzes the actual spectral count [10]. By the end of 2014, it included 10,482 proteins; 143,456 peptides, and about 24 million spectra [2]. The launch of PaxDb brings together disparate aspects of biology for high-throughput analysis and supports global comparative analysis across different organism groups.

Proteomics data repositories focusing on protein PTMs

Phospho.ELM

Recent advances in MS techniques have enabled more efficient detection of phosphorylated proteins [9]. The Phospho.ELM is a web-based resource aimed at storing phosphorylation data imported from research papers and phosphoproteomic MS analyses. MS experiments are run on human/mouse cell lines/tissues. Phospho.ELM is used by laboratory scientists and computational biologists to develop public repositories [11]. To date, this web resource covers 42,914 instances, 299 kinases, 3657 references, 11,224 sequences, and 8698 substrates.

PhosphoSitePlus

PhosphoSitePlus (PSP) is a comprehensive and manually-curated resource designed to collect the structure and function of PTMs, primarily of human and mouse origin. PSP supports two kinds of data, including the modified amino acid and surrounding sequences as well as upstream and downstream interactions with regard to functional regions of the protein [12]. The majority of PTM sites in PSP were detected using MS. PSP is useful to life scientists and biomedical researchers. Currently, PSP spans 50,636 proteins, 1,933,888 MS peptides, 438,576 high-throughput MS sites, 20,262 low-throughput sites, and 18,374 curated papers.

dbPTM

dbPTM is a resource which collects data on experimentally-validated protein PTMs. This resource imports PTM sites from public resources such as SwissProt, Phospho.ELM, and O-GLYCBASE [13]. It also extracts identified peptides with PTMs from research papers. dbPTM is an important resource for researchers working on substrate specificity of PTM sites [14]. To date, dbPTM has covered 153,113 phosphorylation experimental sites, 23,673 ubiquitylation experimental sites, 10,385 acetylation experimental sites, 15,678 *N*-linked glycosylation experimental sites, and 3711 *O*-linked glycosylation experimental sites.

Phosphorylation site database

The phosphorylation site database (PHOSIDA) is a database with a collection of a large number of high-confidence phosphorylation sites. MS-based proteomics is used to identify these sites in various species [15]. To date, the database covers 80,062 *N*-glycosylated, phosphorylated, or acetylated sites. Stringent quality criteria based on a very low false positive rate are used to obtain these sites from high-resolution MS data

[16]. PHOSIDA contains PTM sites from human as well as other species, including bacteria.

Concluding remarks

In this report, we have covered some important proteomics data repositories that are useful for the research community. These resources not only provide raw data and identification results, but also support prospective, high-throughput proteomics research. In addition, they also act as data providers for large-scale genome annotation efforts. In the years to come, sharing data and metadata between repositories will become more important. Thus, proteomics repositories need to focus on developing an integrated approach to data accessibility between repositories. On the other hand, with the advent of new instruments, new sample preparation techniques, and new data analysis methods, new forms of data will be continuously generated. The amount of data in the repositories to be shared at present is just a small fraction of the actually-generated proteomics data that will eventually become available. In order to attract more researchers to submit data, the resources will have to standardize the process and simplify the interface for data submission.

Competing interests

The authors declared that there are no competing interests.

Acknowledgements

This research was supported by the Ministry of Science and Technology of China (Grant Nos. 2013CB910801, 2012AA020201, 2012AA020409, and 2014DFB30010), the National Natural Science Foundation of China (Grant Nos. 21105121, 21475150, and 61303073) and Beijing Municipal Natural Science Foundation of China (Grant No. 5122013).

References

[1] Deutsch EW. The PeptideAtlas project. Methods Mol Biol 2010;604:285–96.

[2] Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics 2014. http://dx.doi.org/10.1002/pmic.201400302.

[3] Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res 2013;41:D1063–9.

[4] Kandasamy K, Keerthikumar S, Goel R, Mathivanan S, Patankar N, Shafreen B, et al. Human proteinpedia: a unified discovery resource for proteomics research. Nucleic Acids Res 2009;37:D773–81.

[5] Smith BE, Hill JA, Gjukich MA, Andrews PC. Tranche distributed repository and ProteomeCommons.org. Methods Mol Biol 2011;696:123–45.

[6] Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 2004;3:1234–42.

[7] Kolker E, Higdon R, Haynes W, Welch D, Broomall W, Lancet D, et al. MOPED: Model Organism Protein Expression Database. Nucleic Acids Res 2012;40:D1093–9.

[8] Colangelo CM, Shifman M, Cheung KH, Stone KL, Carriero NJ, Gulcicek EE, et al. YPED: an integrated bioinformatics suite and database for mass spectrometry based proteomics research. Genomics Proteomics Bioinformatics 2015;13:25–35.

[9] Shifman MA, Li Y, Colangelo CM, Stone KL, Wu TL, Cheung KH, et al. YPED: a web-accessible database system for protein expression analysis. J Proteome Res 2007;6:4019–24.

[10] Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, et al. PaxDb, a database of protein abundance averages across all three domains of life. Mol Cell Proteomics 2012;11:492–500.

[11] Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, et al. Phospho.ELM: a database of phosphorylation sites – update 2011. Nucleic Acids Res 2011;39:D261–7.

[12] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, 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:D261–70.

[13] Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH. DbPTM: an information repository of protein post-translational modification. Nucleic Acids Res 2006;34:D622–7.

[14] Lu CT, Huang KY, Su MG, Lee TY, Bretana NA, Chang WC, et al. dbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res 2013;41:D295–305.

[15] Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 2007;8:R250.

[16] Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the post-translational modification database. Nucleic Acids Res 2011;39:D253–60.