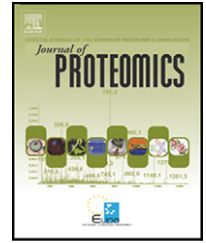




Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot



Review

Proteomic databases and tools to decipher post-translational modifications

Karthik S. Kamath, Meghana S. Vasavada, Sanjeeva Srivastava*

Wadhvani Research Center for Biosciences and Bioengineering, Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Mumbai 400076, India

ARTICLE INFO

Available online 29 September 2011

Keywords:

Posttranslational modifications
Databases
Predictors
Phosphorylation
Glycosylation

ABSTRACT

Post-translational modifications (PTMs) are vital cellular control mechanism, which affect protein properties, including folding, conformation, activity and consequently, their functions. As a result they play a key role in various disease conditions, including cancer and diabetes. Proteomics as a rapidly growing field has witnessed tremendous advancement during the last decade, which has led to the generation of prodigious quantity of data for various organisms' proteome. PTMs being biologically and chemically dynamic process, pose greater challenges for its study. Amidst these complexities connecting the modifications with physiological and cellular cascade of events are still very challenging. Advancement in proteomic technologies such as mass spectrometry and microarray provides HT platform to study PTMs and help to decipher role of some of the very essential biological phenomenon. To enhance our understanding of various PTMs in different organisms, and to simplify the analysis of complex PTM data, many databases, software and tools have been developed. These PTM databases and tools contain crucial information and provide a valuable resource to the research community. This article intends to provide a comprehensive overview of various PTM databases, software tools, and analyze critical information available from these resources to study PTMs in various biological organisms.

© 2011 Elsevier B.V. All rights reserved.

Contents

1. Introduction	128
2. Post-translational modifications	129
3. Techniques to study PTM	130
4. Classification of databases	131
4.1. Phosphorylation databases	132
4.2. Glycosylation databases	133
4.3. Databases of other types of PTMs	136
5. PTM tools	137
5.1. Machine learning processes for prediction	137

* Corresponding author. Tel.: +91 22 2576 7779; fax: +91 22 2572 3480.
E-mail address: sanjeeva@iitb.ac.in (S. Srivastava).

5.2. Strong prediction tools by evolving models	139
5.3. Phosphorylation tools	139
5.4. Glycosylation tools	139
5.5. Other PTM related tools	140
6. Organism specific database and tools	140
7. Conclusions	141
Acknowledgments	141
References	141

1. Introduction

The completion of genome projects has accelerated the analysis of proteome; however, due to the complexity of proteins its study is more challenging than any other biomolecules. This complexity arises due to the biological phenomenon such as gene splicing to form different isoforms and various post-translational modifications (PTMs), which gives rise to enormous number of proteins, about three orders of magnitude higher than the total number of genes encoded in genome [1,2]. As the name indicates for PTMs, the process of protein modifications takes place after translation of mRNA into a protein. All proteins undergo appreciable amount of PTMs to make biologically active form, and this dynamic process occurs in various cell compartments to decide the function of modified protein. About 300 different types of PTMs have been reported till date and many more are still being

reported [2]. PTMs, also designated as ‘cellular switches’, provide diverse role to proteins as per cellular requirements. For instance, ubiquitination is a predominant phenomenon, in which sequential, covalent attachment of ubiquitin on a protein leads to the degradation and decides the fate of protein. Several signaling pathways are majorly regulated through phosphorylation cascades. Hence it is impossible to judge on protein nature and function without having a precise idea about what PTM it undergoes in a given time span.

Initially PTM studies were carried out on selected candidates with mutational screens, western blotting, and tracking with radio labeling; however, recent advancements in mass spectrometry and microarray have enabled HT screening and quantification of PTMs, with sensitivity at subattomolar level [2]. Each run of such HT screening experiment generates large amount of data, which requires intense analysis and interpretation to provide clues for its biological significance.

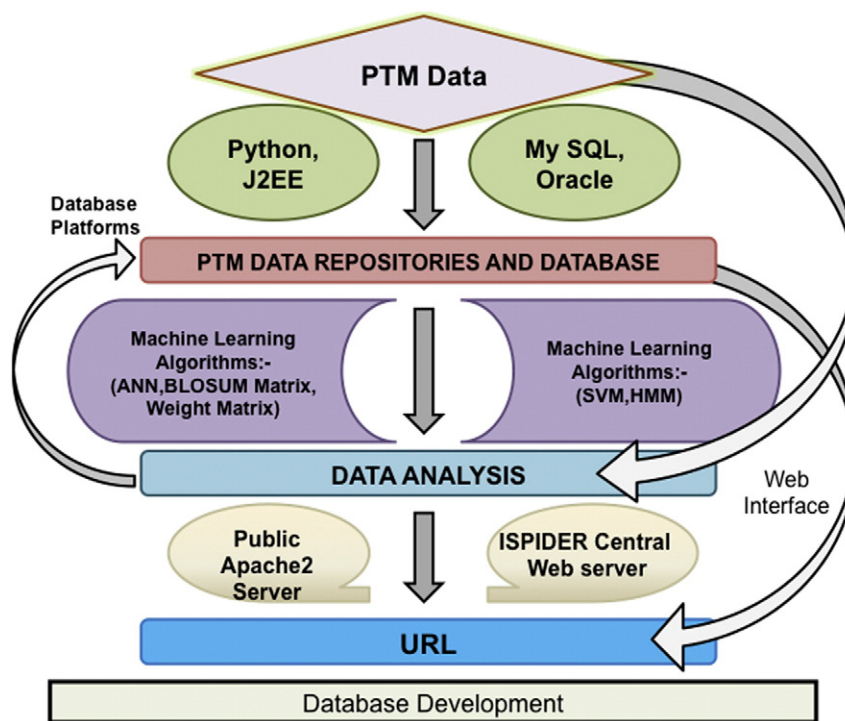


Fig. 1 – Creation of PTM databases and tools: The general scheme of creation of database using database building platforms. PTM data from various sources are continuously annotated in the databases. As a result the curated data are utilized to teach machine learning techniques thereby building a classifier, which mimics biological condition and predicts PTM in the input sequence. These tools and databases are made publically available over World Wide Web.

Therefore, various tools have been developed to enable study of PTMs. The PTM data is shared over World Wide Web in databases, which can be successfully used by the scientific community [3]. Furthermore, the experimentally generated PTM data can be used to teach the computerized machine learning algorithms, which can enable *in silico* prediction of PTMs site and its functions. The computational analysis can save significant time and resources involved in HT screening of PTMs, and therefore, generation of data, curation and developing predictors goes simultaneously (Fig. 1). These PTM databases and tools not only serve as a resource to study the PTMs but also provide an insight into PTM biology and mechanistic insights of complex cellular machinery involved in signaling pathways, metabolism, phylogenetic trees and evolution. This review intends to provide a comprehensive description of various databases and software tools used to study PTM biology. These PTM databases and tools are rapidly evolving and we have made an effort to provide the latest compendium of PTM computational resources and its applications in biological context of different groups of organisms.

2. Post-translational modifications

The Human Genome Project revealed that there are about 30,000 genes in human, which raised an obvious question about number of proteins that outnumber the genes. How millions of proteins are regulated and perform its function are complex questions, which are investigated through proteomics. Splicing and PTMs are suggested to give rise to tremendous complexity, and PTMs diversify the function of proteins by introducing chemical modifications. PTMs act as molecular switches and control biological activity in much orchestrated manner; and its perturbation leads to the deregulation of cell machinery. It is PTM that mostly dictates the fate of a protein with respect to folding, cellular localization and life span. Almost all the proteins undergo one or other type of PTM during or after their synthesis in a well-defined cellular location. Perhaps few PTMs occur after naive protein emerges out of ribosome protein synthesis machinery, leading to the modification of side chain or main skeleton of protein, which in most cases make protein biologically functional. Therefore, linear or one-dimensional genetic message from mRNA is translated into the three dimensional structure of proteins [2].



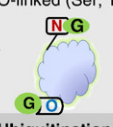
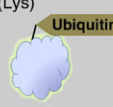
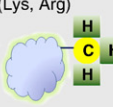

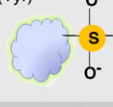


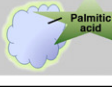
A PTM may be reversible or irreversible. For instance, phosphorylation results in addition of phosphate moiety on the protein backbone. It is a reversible process, whereas proteolytic cleavage of signal peptides is an irreversible one. PTMs are mostly aided by specific enzymes, except a few that undergo auto-modification such as modifications of green florescent protein and auto-phosphorylation of kinases. PTMs can be grouped majorly in two types, covalent attachment of chemical group and covalent cleavage of side chains. Almost all the amino acids undergo the process of PTM, except Leu, Ile, Val, Ala, and Phe [4].

It has been reported that PTMs are highly conserved with respect to evolution. Troponin T, which is part of troponin protein complex, is conserved with respect to phosphorylation in mouse and human [5]. Evolutionary conservation patterns with respect to phosphorylation were studied in *E. coli* and *B. subtilis* by Boris Macek et al. and it was reported that

phosphoproteins and phosphosites are highly conserved in phylogenetic tree as compared to the non-phosphorylated forms [6]. Table 1 provides an overview of commonly reported PTMs and their biological significance.

There is a growing interest of scientific community to decipher the role of PTMs in various biological contexts, which is evident from a simple key word search “posttranslational modification” in Pubmed that retrieved over 36,500 articles. The PTMs are reported to be associated with major human diseases such as cancer, diabetes, cardiovascular disorders etc. In Alzheimer’s disease, the tau proteins aggregate due to abnormal phosphorylation process [7]. The phosphorylation as well as other PTMs such as nitration, glycation, glycosylation,

Table 1 – PTMs processes and biological significance.

Naked protein	PTM Type	Mechanism	Major biological function
	Phosphorylation (Ser, Thr, Tyr)	Enzymatic transfer of phosphate moiety from ATP to a protein, aided by kinases.	Signal transduction, enzyme activity regulation, metabolism.
	Glycosylation N-Linked (Asn), O-linked (Ser, Thr)	Transfer of oligosaccharide to the protein backbone.	Protein stability and solubility, antigenicity, protein targeting, cell-cell interactions.
	Ubiquitination (Lys)	Enzymatic sequential mechanism for covalent attachment of Ubiquitin to the protein.	Protein degradation.
	Methylation (Lys, Arg)	Catalyzed by methyltransferases. Transfer of methyl group to proteins from S-adenosyl L methionine.	DNA methylation, methylation of histone, regulation of gene expression.
	Acetylation (Lys)	Enzymatic transfer of acetyl groups to protein.	Controlling cell signaling processes, modification of histones.
	Sulfation (Tyr)	Sulphate group added to Tyr. Catalyzed by sulfotransferase.	Protein-protein interactions.
	Sumoylation (Lys)	Covalent attachment of SUMO peptide to lysine residues of targeted substrate	Protein stability, protein-protein interactions, transcriptional control.
	Myristoylation (Gly)	Attachment of myristoyl group to N terminal glycine catalyzed by Myristoyltransferase.	Regulation of enzyme activity, membrane binding.
	Palmitoylation (Cys)	Covalent attachment of palmitate residues to cysteine residues of a protein.	Regulating protein-protein interaction, membrane affinity, apoptosis.
			

polyamination, ubiquitination and oxidation is also reported to be associated with various neurodegenerative disorders. Androgen receptors, which are dynamically regulated through various PTMs, are strongly correlated with manifestation of prostate cancer [8]. Hence studying PTMs may enhance our understanding for various human diseases.

3. Techniques to study PTM

Detection of subtle changes, which occur during the PTMs, poses challenge to even advance proteomic techniques. Determination of changes to very minute level and correlation with biological phenomenon remain challenging for modern technologies. PTMs bring in either addition of chemical moieties or removal of few amino acids; therefore, difference in mass must be apparent when measured. For instance, in palmitoylation, addition of palmitic acid on the cystein residue yields the addition of 238 Da whereas; methylation of lysine

residue causes the addition of 14 Da to the total mass. The bulkier modifications such as ubiquitination may result in mass difference of about 1 kDa. There are many approaches ranging from gel-based techniques, mass spectrometry, microarrays, peptide library screening etc. that are currently used to study PTMs (Fig. 2). As per the research questions, either whole proteome or only an enriched part containing proteins with PTMs of specific interest can be screened. Affinity based enrichments, immunopurification and metal affinity chromatography are commonly used strategies for the purification of proteins containing specific PTM. Immobilized metal affinity chromatography (IMAC) purification is a common chemical affinity strategy for the enrichment of phosphoproteins, whereby immobilized Fe^{3+} ions selectively bind to the phosphorylated peptides. Other metal oxide affinity resins such as TiO_2 , Fe_3O_4 are also commonly used [9].

Conventional proteomic approaches such as gel-based techniques have been used to profile global PTMs in a given biological condition. Two dimensional gel electrophoresis (2-

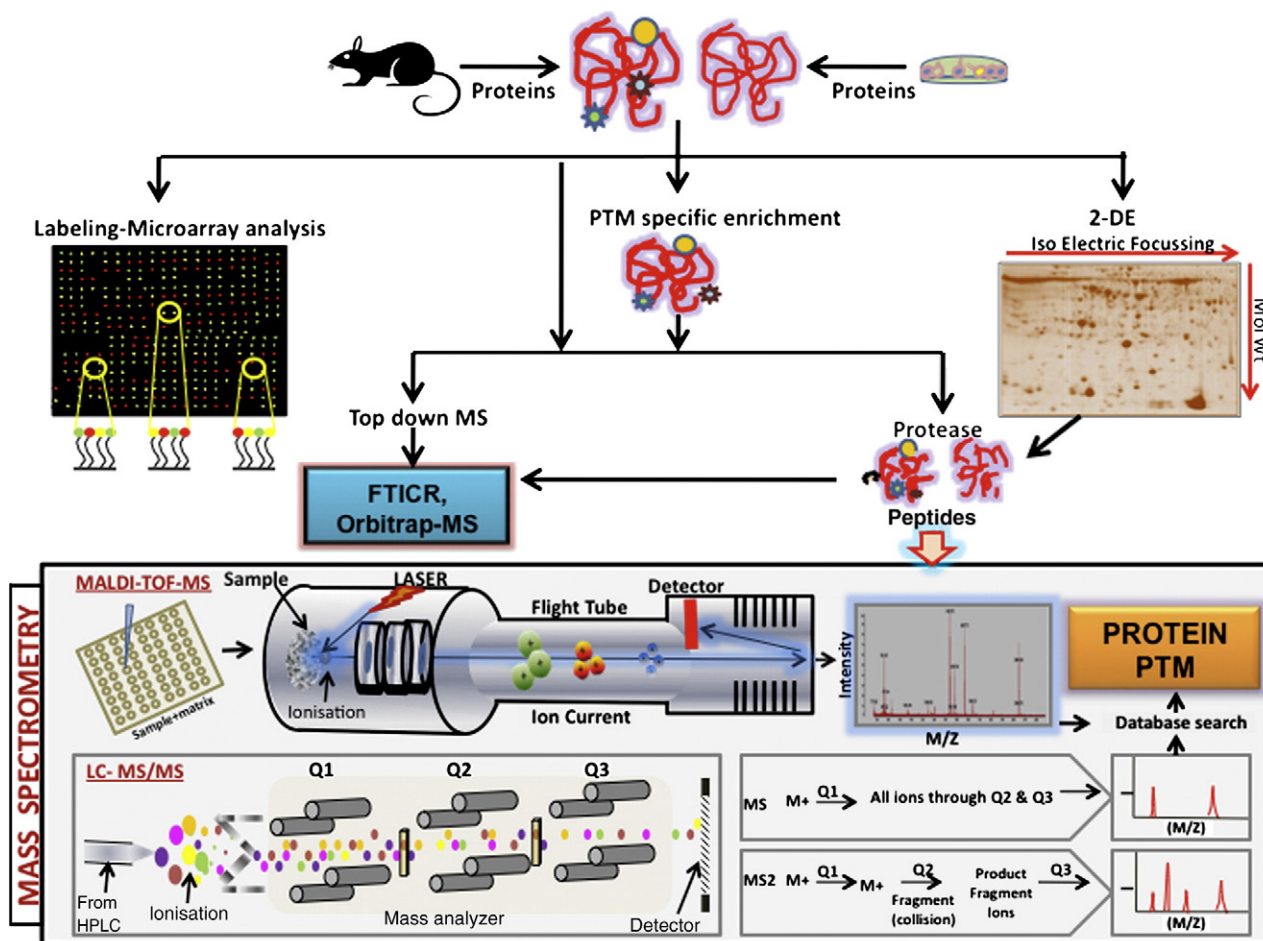


Fig. 2 – Proteomic techniques for profiling PTMs. Present day cutting edge technologies enable high throughput screening of PTMs. Gel-based or gel-free approaches are widely used. Methods such as mass spectrometry, microarray and 2-DE are extensively used for proteome wide screening of PTMs. Similarly, a subclass of proteins with specific PTM can be segregated using various purification strategies. Essentially shotgun method involves analysis of proteolytically digested peptide over mass spectrometer involving soft ionization techniques such as ESI and MALDI. Top-down MS technique typically involves using high-resolution MS platforms such as FTICR for elucidating PTM patterns of intact proteins. Gel-based approach such as 2-DE, helps in separation and visualization of proteins. After protein separation gels can be stained with PTM specific stain.

DE) separates proteins on the basis of isoelectric point followed by molecular weight. After protein separation the gel can be stained with PTM specific stains such as Pro-Q Diamond for phosphoproteins. This approach can be used to compare the differential expression of PTMs in control and treatment by comparing the staining intensity. Zong et al. used 2-DE to understand PTM patterns of murine cardiac 20S proteasomes and their associated proteins. This study revealed phosphorylation, glycosylation, nitrosylation, and oxidation patterns of 20S proteasomes [10]. Although the gel-based approaches are convenient to use, it has a few drawbacks with regard to robustness, sensitivity and gel-to-gel reproducibility. Characterization of these resolved proteins subsequently requires other technique such as mass spectrometry to identify the proteins (Fig. 2).

Advancement in analytical techniques and evolution of various high-resolution mass spectrometers during the last decade has accelerated the large scale screening of PTMs from various biological sources. Advent of shotgun based MS methods has accelerated the rapid and direct identification of proteins and PTMs from complex mixtures. In this approach typically a protein mixture is digested with proteolytic enzyme such as trypsin, and resultant peptides are then separated over liquid chromatography (typically reverse phase chromatography) followed by MS/MS scanning. Identification of proteins and associated PTMs can be done by software and searching against databases. The software matches the submitted tandem MS data with *in silico* MS data in databases [11]. Trypsin being a versatile enzyme is regarded as a general choice as a protease but in specific cases other enzymes such Arg-C, Lys-C are also employed. For instance histones, which are regulated through methylation and acetylation, are digested with Arg-C, Lys-C which provides better coverage than trypsin. The most commonly used PMF analysis tool Mascot is linked to PTM database Unimod. The inbuilt algorithm of Mascot searches the mass difference by matching the input spectra with that of the reference spectra in databases and predicts the type of PTM [12].

Top down mass spectrometry involves analysis of intact proteins using high-resolution MS techniques. High-resolution MS platforms such as FTICR-MS, Orbitrap-MS with PTM friendly dissociation techniques such as Electron capture dissociation (ECD) and Electron Transfer Dissociation (ETD) are mainly used [13]. Data resulted from MS can then be submitted to search tool such as ProSight PTM to characterize the PTM patterns and establish the identity of protein [14]. MS based quantification of PTMs, absolute or relative has gained increasing interest. Labeling of proteins is used as tool for relative quantification. Metabolic labeling methods such as SILAC are now well established for label-based quantification of PTMs. In SILAC cells are grown in different media containing light or heavy amino acids and quantification is performed on the basis of inherent mass difference between light and heavy forms [15]. SILAC was used to study histone modifications specifically lysine acetylation and methylation patterns by Cuomo A et al. [16]. Largest phosphorylation site measurement by Olsen JV et al. used SILAC as tool for quantification and identified 20,443 unique phosphorylation sites [17]. However, the metabolic labeling can only be incorporated into the living cells. Other chemical labeling techniques such

as iTRAQ are also used for PTM quantification, which has flexibility with respect to labeling which involves derivatization of primary amine group with isobaric tags. iTRAQ was used to identify quantitative differential expression patterns of O-GlcNAc sites with respect to Alzheimer's disease [18] and T-Cell signaling cascade related phosphoproteomic changes in diabetic mouse [19]. The label-free quantification of PTMs, which involves comparison of different LC runs on the basis of parameters such as peak area and spectral count is evolving rapidly. A label-free quantification study by Hoffert et al. identified and quantified 714 phosphorylation sites on 223 unique phosphoproteins through LC-MS/MS-neutral loss scanning strategy [20].

Several large-scale PTM screens have resulted to provide enormous and valuable data. A study by Bo Zhai et al. determined large-scale phosphorylation sites (~13,720) in 2702 proteins in *Drosophila* embryos [21]. Using MS approach, Choudhary and colleagues studied lysine acetylation in 1750 proteins and identified 3600 potent acetylation sites [22]. This dataset was further used to train Support Vector Machine (SVM) based acetylation predictor. This predictor is available on line in PHOSIDA resource for public use. One of the largest MS based screening of glycoproteins used high-resolution MS based screening in mouse tissue and revealed 6367 high confident N-glycosylation sites on 2352 proteins [23]. One of the largest human ubiquitylation site screening by Wagner et al. revealed 11,054 ubiquitylation sites on 4273 proteins [24]. We have described only few studies but there are also other large scale screening studies, which have resulted into generation of large PTM data sets. Some of these data are also deposited into databases and further used for developing trained algorithm of PTM predictor.

Another technique, protein microarray, is one of the versatile platforms for HT screening of PTMs. Protein microarrays are miniaturized arrays containing small amounts of immobilized proteins. Kung et al. group developed a lectin-binding assay for screening glycoproteins on yeast proteome arrays that revealed 534 glycoproteins [25]. A comprehensive protein array study by Jason Ptacek et al. elucidated about 4000 phosphorylation sites in 1325 proteins [26]. Mitogen-activated protein kinase–substrate interactions were studied using protein microarrays and study revealed 570 MPK substrates in *Arabidopsis thaliana* [27]. Yeast two hybrid system has also been used for large scale screening of PTMs [2]. Despite having advancements in several HT techniques, studying short-lived and often chemically labile PTMs and its characterization remains challenging due to the dynamic range and detection limit etc. Many advanced technologies have attempted to bridge this gap; however, no single technique can be solely relied for screening all the PTMs in a given biological question.

4. Classification of databases

PTM databases are continuously growing in size due to the advent of high-throughput screening technologies. These PTM databases feature vast variety of data ranging from viruses to humans. Some of these databases are specific to a PTM and others are composed of wide variety of PTMs in a single platform. For instance, databases such as PhosphoBase, O-

glycibase are specifically focused for one type of PTM, whereas Swiss-Prot, HPRD, dbPTM, RESID, PHOSIDA etc. provide detailed information for different types of PTMs (Table 2). These databases and resources have accelerated the analysis, visualization and prediction of PTMs in biological contexts. The data in most of the PTM databases are derived empirically or curated manually through the literature. For example, HPRD is a manually curated database, which has more than 93,710 entries and it is linked to other entities such as PhosphoMotif Finder for further reference [28]. Swiss-Prot (UniProt), one of the largest collections of various PTM types, is non-redundant and enables users to get amass information on a single platform. These databases are collection of variety of information in one place, thereby trying to provide complete biological information to the entries present [29]. Fig. 3 represents general and PTM specific databases.

4.1. Phosphorylation databases

The PTMs act as a biological switch to activate or deactivate molecules by signal transduction pathways. Protein phosphorylation is one of the most-studied PTMs, which accounts for over 30% of all PTMs. Phosphorylation process involves transfer of phosphate moiety from ATP to a protein (serine, threonine or tyrosine residues) by enzyme kinases resulting in formation of ADP. Phosphorylation is biologically significant because this ubiquitous regulatory mechanism controls processes such as cellular growth, differentiation, apoptosis and DNA repair etc. [2,30]. Many of the kinases are being used as potential drug targets to treat some of the major diseases such as cancer. It is predicted that there are approximately 500,000 phosphorylation sites in human proteome alone [30], which emphasizes that there is a great need to decipher the role of protein phosphorylation.

Mass spectrometry is widely used to study phosphoproteome. As discussed in the previous section, several large-scale MS based studies have identified thousands of phosphoproteins and phosphosites. Several databases help in mining phosphorylation data. PhosphoBase was the first

report of a phosphorylation database. This database curated experimentally determined data from literature, major protein sequence databases such as SwissProt, and protein information resource [31]. This database initially had 156 phosphoproteins with 398 phosphorylation sites but since then large number of databases have been created and useful data has been accumulated in these databases. For instance, one such widely used database Phospho.ELM, which is a manually curated database of experimentally verified non-redundant phosphosites of eukaryotic origin, contains 42,575 serine, threonine and tyrosine phosphorylation sites and information for 310 different kinases. For a given input Phospho.ELM provides vast variety of information such as sequence of phosphosite, related kinase, and evolutionary significance about conservation of phosphosites, structure through phospho 3D, binding motif and molecular interaction network. This database is also linked to predictor application called ELM [32]. PHOSIDA (PHosphorylation Site DATabase) database data is derived from MS screening of phosphosites. Mann and colleagues screened about 2244 proteins and obtained 6600 phosphorylation sites in HeLa cells, which were deposited into PHOSIDA. PHOSIDA also has information of phosphoproteins for 8 other organisms. Although this database was initially established as a phosphorylation database, later addition of PTM predictor and motif search made this resource more comprehensive. PHOSIDA is regularly updated through Swissprot and TIGR databases [33].

PhosphoSitePlus is one of the biggest collections of PTM information, which contains information on several types of PTMs of enormous proteins from *in vivo* and *in vitro*, originating from various vertebrates and invertebrates. Previous version of Phosphosite only targeted phosphorylation but PhosphoSitePlus provides wide range of information on phosphorylation as well as other types of PTMs such as acetylation, methylation, and ubiquitination. It contains data from published datasets and previous, valid unpublished data generated from Cell Signaling Technologies. A few of the applications of PhosphoSitePlus such as search of PTM sites based on tissue, disease, cell lines make it comprehensive and

Table 2 – General PTM databases with web links and salient features.

SR no	Database/tool (predictor)	Features
1	Human Protein Reference Database [HPRD] [28] http://www.hprd.org/	HPRD provides details for protein interactions and PTMs (30,047 human proteins and 93,710 PTMs). Extra features such as PhosphoMotif Finder and Human Proteinpedia.
2	dbPTM [53] http://dbPTM.mbc.nctu.edu.tw/	Information on PTM sites from Swiss-Prot, PhosphoELM, UbiProt and O-GLYCBASE.
3	SysPTM [77] http://lifecenter.sgst.cn/SysPTM	Contains 36,466 non-redundant empirical PTM sites and a PTM predictor.
4	RESID [78] http://www.ebi.ac.uk/RESID/	PTM research tool with four tools; PTMBlast, PTMPathway, PTMPhylog maps and PTMCluster. Nearly 50 PTM types, 117,350 experimentally determined PTM sites on 38,674 proteins.
5	Swiss Prot [29] http://www.ebi.ac.uk/uniprot/	Annotations and structures for protein pre-, co- and post-translational modifications including N-terminal, C-terminal and peptide chain cross-links modifications. 559 entries.
6	FindMod [79] http://expasy.org/tools/findmod/	One of the largest comprehensive resources. Experimental PTMs, Putative PTMs, Protein variants.
		Tool that examines PMF data. Finds mass differences between empirical and theoretical peptides. Over 22 types of PTMs are considered.

^a These are general databases (and tools) which provide PTM as well as other details.

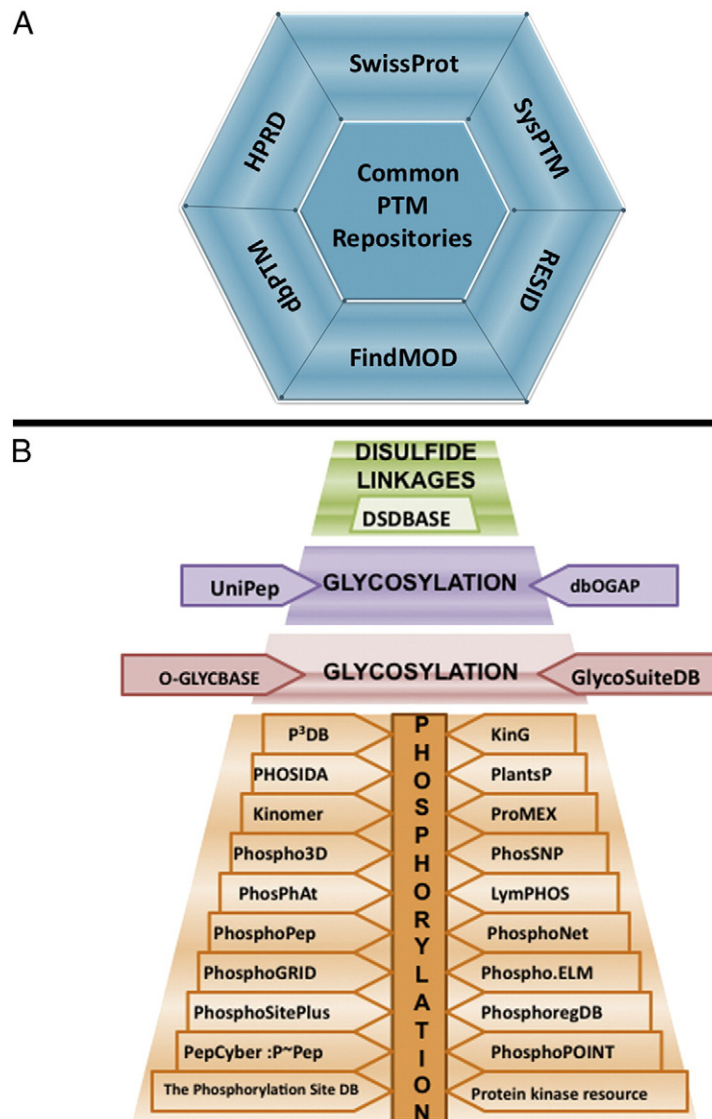


Fig. 3 – General and PTM specific databases: (A) Common PTM repositories share large amount of data pertaining to different types of PTMs on a single platform. SwissProt is one of the largest continuously updated, non-redundant repositories for PTM data. This encompasses enormous data on various types of PTMs (B) Databases for specific PTM store data with respect to one type of PTM. One of the typical databases such as Phospho.ELM has largest collected information on phosphorylation.

informative [34]. Structural repositories such as Phospho3D contain experimentally verified 3D structures of phosphorylation sites. Most of the databases enable various search options such as name of kinase; PDB identification code etc., which make the search process simple [35]. PlantsP, a plant specific database, combines information derived from plant genomic sequences with experimentally derived functional genomics data on plant kinases and phosphatases [36]. PhosphoPep provides an idea about the signal transduction pathways by linking kinases with their upstream and downstream molecules in various organisms such as *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiencie*. This database contains LC-MS/MS generated data, which is searched from standard protein database and validated further. The relevant pathways of proteins of interest can also be visualized [37]. LymPHOS database is focused on phosphoproteome of human lymphocytes

and it elucidates key role of various signal transduction pathways and altered immune response. This database provides options of searching proteins and peptide sequences, with the mass spectral information [38]. A comprehensive view of most widely used databases and tools for phosphorylation research is depicted in Table 3, Figs. 3 and 4.

4.2. Glycosylation databases

Glycosylation involves linking saccharides to proteins in presence of glycosyltransferase enzymes, giving rise to a glycoprotein. Glycosylation process mainly occurs in endoplasmic reticulum and golgi complex of the cellular compartment and major glycoproteins are seen localized to the cell surface. Like-wise glycation, a non-enzymatic process involves attachment of sugar moieties to proteins specifically at lysine

Table 3 – Databases and tools to study phosphorylation.

SR no	Database/tool (predictor)	Features
1	The Phosphorylation Site Database [80] http://www.phosphorylation.biochem.vt.edu/	PTMs of prokaryotes of the domains archaea and bacteria. Directly linked with RESID, O-GLYCBASE etc.
2	Phospho.ELM [32] http://phospho.elm.eu.org/	One of the largest DB of eukaryotic p-site. Contains 42,575 Serine, Threonine and Tyrosine non-redundant p-sites.
3	PhosPhAt [75] http://phosphat.mpimp-golm.mpg.de	DB: p-sites in <i>A thaliana</i> identified by MS. Predictor, trained on the experimental dataset. Predicts p-sites in protein sequence.
4	PHOSIDA [33] http://www.phosida.com/	PTM DB of wide range of organisms. Phosphorylation, acetylation, N-Glycosylation information. 70,095 p-sites on 23,669 proteins
5	PhosphoPep [37] http://www.phosphopep.org/	Predictor, predicts high confidence phosphosites, with the Support vector machines (SVMs) algorithm. It also predicts other types of PTMs such as acetylation. Phosphoproteome resource for <i>S cerevisiae</i> , <i>C. elegans</i> , <i>D. melanogaster</i> (Kc167 cells) and <i>Homo sapiens</i> . Based on MS data. 9000 identified p-sites in yeast, 10,000 phosphorylation sites of nearly 4600 phosphoproteins <i>D. melanogaster</i> and 3980 p-sites in humans.
6	PhosphoPOINT [81] http://kinase.bioinformatics.tw/	Human kinase interactome and p-protein DB. 518 known human serine/threonine/tyrosine kinases. 4195 p-proteins.
7	PhosphoNET–Human Phosphosite KnowledgeBase [82] http://www.phosphonet.ca/	Human p-sites. Over 657,391 p-sites in 23,469 represented proteins.
8	Phospho3D [35] http://cbm.bio.uniroma2.it/phospho3d	DB of 3D structures of p-sites. Retrieves data from the phospho.ELM Database enriches further at structural level.
9	ProMEX [83] http://promex.pph.univie.ac.at/promex/	Mass spectral reference DB for tryptic digested proteins and protein p-sites of 14 plant species including <i>Arabidopsis thaliana</i> . Contains manually validated mass spectra.
10	PhosphoSitePlus [34] [http://www.phosphosite.org]	One of the largest DB of protein phosphorylation majorly in humans and mouse, also has information on other organisms such as rabbit, hamster etc. Includes information on other types of PTMs such as ubiquitination etc.
11	PhosphoregDB [84] http://phosphoreg.imb.uq.edu.au/	Tissue and sub-cellular distribution of mouse protein kinases and phosphatases.
12	PhosphoGRID [85] http://phosphogrid.org/	DB of experimentally verified <i>in vivo</i> p-sites in <i>S cerevisiae</i> . Entry over 5000 p-sites.
13	Plant Protein Phosphorylation Database (P ³ DB) [86] http://digbio.missouri.edu/p3db/	Plant protein phosphorylation DB. Involves data of 4 different plant species. Over 31,000 p-sites from 10,400 proteins.
14	LymPHOS [38] http://www.lymphos.org	Database of phosphoproteome of human lymphocytes. Annotates data generated by MS based experiments.
15	NetworKIN [87,88] http://networkin.info/search.php	Provides latest collection of phosphoproteins through connection with PhosphoELM. Predictor, predicts <i>in vivo</i> kinase–substrate relationships. Adopts inbuilt NetworKIN method for search.
16	Predikin [89] http://predikin.biosci.uq.edu.au/	Predikin predicts protein kinase peptide and likely p-sites for a specific protein kinase, links substrates to kinase sequences.
17	Protein kinase resource [90] http://pkr.genomics.purdue.edu/pkr/Welcome.do	Integrated view of the protein kinase superfamily, structural representation along with complete assistance for kinases study. Provides information on signaling pathway and associated diseases.
18	Kinomer [91] http://www.compbio.dundee.ac.uk/kinomer/	Systematically classified protein kinases. Annotations of kinases for about 43 organisms.
19	KinG [92] http://king.mbu.iisc.ernet.in/	Collection of protein kinases in genomes. Protein Kinases information on <i>D. melanogaster</i> , <i>A. thaliana</i> , <i>H. sapiens</i> , <i>C. elegans</i> and <i>S. cerevisiae</i> . In addition Protein Kinases information on 8 archaeal genomes and 27 bacterial genomes.
20	PlantsP [36] http://plantsp.genomics.purdue.edu/	DB which connects functional genomics data to sequence in context of plant (Kingdom viridiplantae) kinases and phosphatases.
21	PepCyber: P~Pep [93] http://www.pepcyber.org/PPEP/	DB of human Protein–Protein Interactions mediated by Phosphorylation binding domains (PPBDs). 11,269 records of interactions between 387 PPBDs proteins and 1471 substrate proteins.
22	PhosSNP [94] http://phosnp.biocuckoo.org/	Database of phosphorylation related non-synonymous SNPs in Humans.
23	ProMost [95] http://proteomics.mcw.edu/promost	Tool calculating the pI and molecular weight of phosphorylated and modified proteins on 2D gels.
24	PhosphoScore [96] http://dir.nhlbi.nih.gov/papers/lkem/phosphoscore/	Java based software for phosphorylation site assignment tool for MS data.
25	NetPhos [97] http://www.cbs.dtu.dk/services/NetPhos/	Predicts non-kinase specific phosphorylation status based on sets of experimentally validated Ser, Thr and Tyr p-sites. ANN used.
26	NetPhosK [39] http://www.cbs.dtu.dk/services/NetPhosK/	Predicts kinase-specific p-sites based on sets of Ser, Thr and Tyr p-sites. ANN used.

Table 3 (continued)

SR no	Database/tool (predictor)	Features
27	NetPhosBac [73] http://www.cbs.dtu.dk/services/NetPhosBac	Predicts Ser/Thr p-sites in bacterial proteins. ANN used.
28	NetPhosYeast [74] http://www.cbs.dtu.dk/services/NetPhosYeast/	Predicts p-sites in yeast proteins. ANN used.
29	KinasePhos [98] http://KinasePhos2.mbc.nctu.edu.tw/	SVM based predictor for protein kinase-specific p-sites prediction. SVM utilizes features such as solvent accessibility and sequences based amino acid coupling patterns.
30	PostMod [99] http://pbil.kaist.ac.kr/PostMod	Prediction of kinase-specific p-sites. Training data was retrieved from PhosphoELM.
31	DISPHOS [100] http://www.ist.temple.edu/disphos/	Disorder-Enhanced p-site predictor. Uses logistic regression based linear predictor model.
32	NetPhorest [101] http://netphorest.info/	Catalogue of linear motifs involved in phosphorylation based signaling. Information on 179 kinases.
33	pkaPS [102] http://mendel.imp.ac.at/pkaPS/	Prediction of protein kinaseA p-sites. High confidence prediction with 96% sensitivity and 94% specificity.
34	Group-based Prediction system (GPS) [103] http://gps.biocuckoo.org/	Prediction of p-sites for 408 human kinases. Group based scoring system, uses BLOSUM62 for scoring.
35	Scansite [104] http://scansite.mit.edu	Predicts cell signaling interactions using short sequence motifs within proteins that are likely to be phosphorylated by specific protein kinases.
36	PhosphoMotif Finder [105] http://www.hprd.org/PhosphoMotif_finder	Literature based information on kinase and phosphatase substrates and binding motifs.
37	MetaPredPS [106] http://MetaPred.biolead.org/MetaPredPS	Predicts p-sites of major S/T kinase families: CDK, CK2, PKA, and PKC. Makes use of element predictors such as GPS, KinasePhos, NetPhosK, PPSP, PredPhospho and Scansite.
38	CRPhos [107] http://www.ptools.ua.ac.be/CRPhos	Prediction of kinase-specific phosphorylation sites. Uses CRF conditional random fields.
39	PhoScan [108] http://bioinfo.au.tsinghua.edu.cn/phoscan/	Prediction of kinase-specific phosphorylation sites with sequence features. Uses log-odds ratio approach.
40	Prediction of PK-Specific Phosphorylation Site (PPSP) [109] http://ppsp.biocuckoo.org/	Prediction of kinase-specific phosphorylation sites. Uses BDT.
41	PhosphoBlast [110] http://phospho.elm.eu.org/pELMBlastSearch.html	Predictor in PhosphoELM. Tool which identifies specific phosphosite mutations. Matches p-peptides sharing the p-sites within and across species.
42	Motif-X [55] http://motif-x.med.harvard.edu/	Predictor of phosphorylation short linear motif. First ever substrate driven approach to predict motifs.

P-site—phosphorylation site, p-proteins—phosphoproteins, DB—database, MS—mass spectrometry, p-peptides—Phosphopeptides, BDT = Bayesian Decision Theory, CRF = Conditional Random Fields, SVM—Support Vector Machines, ANN = Artificial neural network.

residue. Depending on the linkage between the amino acid and the sugar moiety, there are 4 types of glycosylations, namely; N-linked glycosylation, O-linked glycosylation, C-mannosylation and Glycophosphatidylinositol anchored (GPI) attachments [39]. Glycosylation is involved in various cellular events, which has implications in various biological functions such as antigenicity of immunological molecules, protein's half-life, protein folding, protein targeting, cell–cell interactions and protein stability [39]. Aberrant forms of glycosylation play a major role in various human congenital disorders. Despite technological advances, as compared to phosphorylation, much is yet to be explored to understand the interaction between several glycotransferases and their corresponding substrates. Glycosylation databases provide valuable information curated out of published reports that helps to study glycobiology and its relevance to diseases. General glycosylation databases such as GlycoSuiteDB annotate and collect glycan structures derived from glycoproteins of various biological sources. It contains information of glycan types, linkages and anomeric configurations, mass, composition and the analytical methods used to determine

the glycans structure. Current version; GlycoSuiteDB 8.0 is composed of 9436 entries of which 3238 are unique and 1851 are completely characterized. This database is extensively linked with the ExpASY, GlycoMod, SWISS-PROT and PubMed and provides details on the disease relevant modifications [40].

The glycosylation screening technologies are continuously evolving. MS techniques and protein microarray are regularly used to study glycosylation. One of the recent MS based database screening studies by Dorota et al. revealed 6367 N-glycosylation sites on 2352 proteins derived from four mouse tissues and blood plasma [23]. Massive data for glycosylated proteins thus produced are annotated in various databases. Glycodatabases, depending upon the type of chemistry involved in the attachment of saccharide moieties, are available on World Wide Web (Table 4). O-glycosylation is a process occurring in golgi apparatus, which is an enzymatic attachment of N-Acetylgalactosamine on hydroxyl group of Ser or Thr residues in presence of enzyme N-Acetylgalactosaminyltransferase. Transfer of first sugar moiety leads to the sequential addition of other sugar molecules in

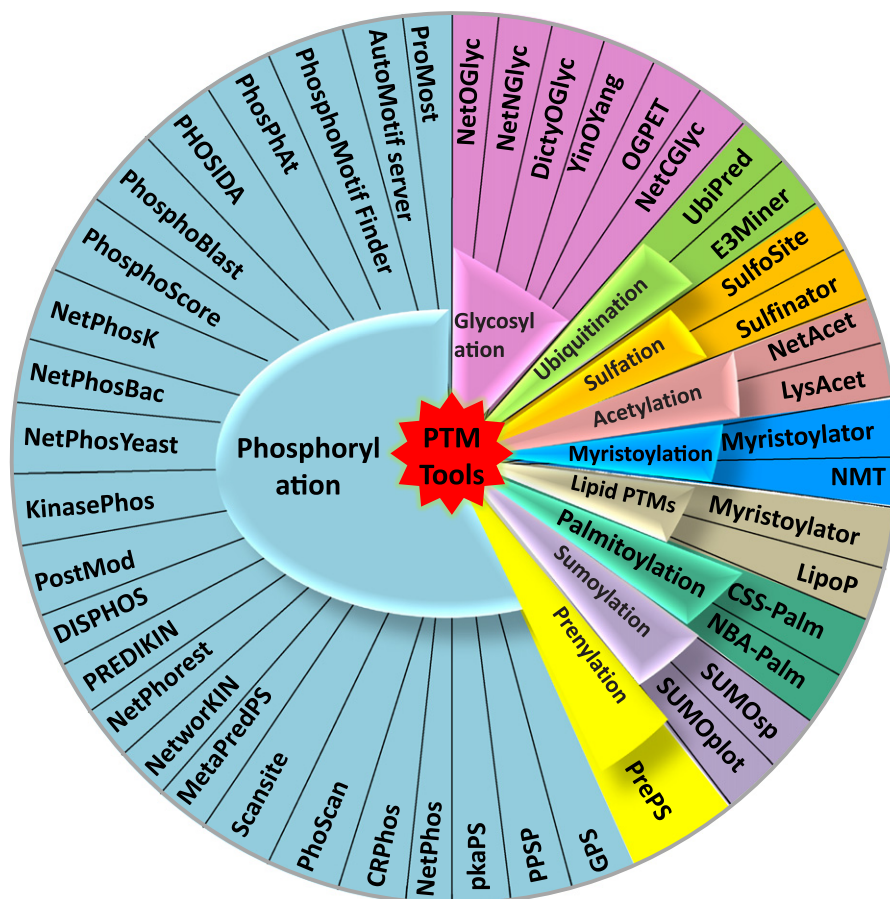


Fig. 4 – Different tools for PTM prediction: A number of PTM predictors are being reported every day. To date compared to tools of other types of PTMs, phosphorylation prediction tools are most studied. The figure segregates various tools as per the type of PTM. Each type of PTM is presented with most popular predictors, which are available on WWW.

synthesis of O-linked glycoprotein [39]. O-GLYCBASE and BOLD databases provide information on O-linked glycosylation [41,42]. One of the initially reported databases, O-GLYBASE contains information related to the glycosylation sites and glycoproteins in context of O-linked glycosylation. There are 242 glycoproteins and 2413 verified, non-redundant O-glycosylation sites in this database that are extracted from SWISS-PROT, PIR and cross-referred to sequence and structural databases. Another O-Linked glycosylation database, the biological O-linked database (BOLD), provides information about glycans at four levels: glycan structure, biological sources, glycan related references, and methods for the identification and characterization of glycans. Very recently Jinlian Wang et al. reported first publicly available largest collection of O-GlcNAcylated proteins and sites called dbOGAP. Currently the database has entry of about 800 glycoproteins with experimentally established O-GlcNAcylation information. This database not only provides information about O-GlcNAcylation but also about associated diseases, biological pathways, cellular components etc. [43].

N-linked glycosylation is one of the commonly observed types of glycosylation in eukaryotic proteins than prokaryotic ones. This phenomenon is localized to endoplasmic reticulum membrane and involves enzymatic transfer of N-Acetylglucosamine to asparagine residue of the proteins. Further addition of different

types of the sugar moieties takes place depending on the subtype of N-glycoproteins. N-linked glycosylation can be mainly of three types, namely high mannose type, hybrid and complex types. The sequence involved in this process is Asn-Xaa-Ser/Thr, where Xaa is not proline. Although this type of glycosylation was initially reported in eukaryotes, it was also observed to be present in one of the gram-negative bacteria, *Campylobacter jejuni* [44]. Similar to the O-linked glycosylation, there are N-linked glycosylation databases such as UniPep, which provide information for about 1522 unique N-linked glycosites. Since N-linked glycoproteins are major secretory, surface linked and plasma proteins, this database intends to enable biomarker discovery. The data originates from various biological sources such as plasma, liver, prostate etc. Each query submitted to the database provides information on protein, trans-membrane sequence, sub-cellular location, predicted N-glycosylated site and biological pathway. Information of sub-cellular location of the protein is provided on the basis of signal peptide, which is present in the corresponding glycoprotein [45].

4.3. Databases of other types of PTMs

Phosphorylation and glycosylation are major PTMs, however, there are other PTMs, which also play an important role in various cellular events; to list a few, acetylation, ubiquitination, sulfonation,

Table 4 – Databases and tools to study glycosylation.

SR no	Database/tool (predictor)	Features
1	O-GLYCBASE [56] http://www.cbs.dtu.dk/databases/OGLYCBASE/	DB contains information on O and C-glycosylated proteins and their g-sites. 242 glycoprotein entries.
2	GlycoSuiteDB [40] http://glycosuitedb.expasy.org/glycosuite/glycodb	DB of curated and annotated glycans. Entries over 9436. Provides vast variety of options such as disease relation, taxonomy, mass etc.
3	UniPep [45] http://www.unipep.org/	DB for human N-linked glycosites. Over 1522 unique N-linked glycosites. A resource aids biomarker discovery.
4	NetOGlyc [56] http://www.cbs.dtu.dk/services/NetOGlyc/	Neural network based Predictor of mucin type GalNAc O-g-sites in mammalian proteins.
5	NetNGlyc [59] http://www.cbs.dtu.dk/services/NetNGlyc/	Predicts N-g-sites in human proteins by examining the sequence environment of Asn-Xaa-Ser/Thr sequences. Neural network used.
6	DictyOGlyc [72] http://www.cbs.dtu.dk/services/DictyOGlyc/	Predicts GlcNAc O-g-sites in <i>Dictyostelium discoideum</i> proteins. Neural network used.
7	YinOYang [39] http://www.cbs.dtu.dk/services/YinOYang/	Predicts potential sites which undergo O- β -GlcNAc attachment and phosphorylation in eukaryotic protein sequences. Neural network used.
8	OGPET [111] http://ogpet.utep.edu/OGPET/	Predicts mucin-type O-glycosylated residues in eukaryotic proteins. Uses in house developed variation profiling scoring systems methods for prediction.
9	NetCGlyc [61] http://www.cbs.dtu.dk/services/NetCGlyc/	Predicts C-mannosylation sites in mammalian proteins. Neural Network used.
10	Oglyc [57] http://www.biosino.org/Oglyc	SVM based predictor of mucin-type O-glycosylation site on mammalian proteins.
11	dbOGAP [36] http://cbsb.lombardi.georgetown.edu/hulab/OGAP.html	DB of O-GlcNAcylated proteins and O-GlcNAcylation sites. Over 800 entries.
12	BPI [60] http://mendel.imp.ac.at/gpi/gpi_server.html	Predicts GPI anchoring sites in the input sequence. Provides taxon option of specific prediction.
13	GPP [58] http://comp.chem.nottingham.ac.uk/glyco/	Prediction of N-linked and O-linked glycosylation using random forest algorithm.
14	NetGlycate [62] http://www.cbs.dtu.dk/services/NetGlycate/	Predicts lysine glycation in mammalian proteins.

DB—Database, g-sites = glycosylation sites.

myristoylation, prenylation, glycosyl-phosphatidylinositol anchoring (GPI) etc. Although these PTMs are being studied in detail there are very few databases and resources to study these PTMs. Detailed list of databases is described in Table 5 and Fig. 3 and only few representative ones are discussed in this section.

Lipid modifications of protein are biologically very important, which are evident through their role in membrane function. Common lipid modifications include prenylation, myristoylation and GPI anchoring. A unique PTM observed in bacteria is attachment of N-acyl S-diacylglyceryl group to N-terminal cystine, which helps to anchor bacterial proteins on hydrophobic surfaces thereby helping in pathogenesis. DOLOP database of bacterial N-acyl S-diacylglyceryl group attachment modifications has entry of about 278 distinct lipoproteins from bacterial genomes [46]. Ubiquitination encompasses attachment C-terminal glycine of 76 amino acid polypeptide ubiquitin to lysine residue of target proteins and is aided by a set of enzymes, which leads the protein to degradation process [47]. One of the repositories of manually curated ubiquitination site is UbiProt. It provides information for the protein substrates of ubiquitylation from the experimental data. Each ubiquitylated protein entry provides details of the source, mode of ubiquitylation with emphasis on conjugation cascade and covers ubiquitylation patterns of organisms such as *S. cerevisiae*, *H. sapiens*, and *M. musculus* [47]. Disulfide bonds in proteins involve oxidation of thiol groups of cysteine residues and play a crucial role in maintaining thermodynamic stability of proteins. This PTM is widely studied for its nature due to its applicability in

biopharmaceutical products. The smallest peptide such as insulin to few large enzymes like tissue plasminogen activator have inter and intramolecular disulfide bonds. In this regard, DSDBASE, provides comprehensive information on native disulfide bond present in proteins with over 2,385,617 entries [48].

5. PTM tools

Experimental results generated from the HT techniques have demonstrated thousands of potential PTM sites but experimental validation of these targets is time consuming. Therefore, PTM predictors play a key role in studying PTM biology. Computational prediction is made plausible due to the inherent nature of PTMs, which are directed by specific enzymes and involve specific sequence motifs recognition. Many predictors use these properties to predict the potential PTMs in a given amino acid sequence. Predictors explore unknown PTM sites and provide valuable information to make meaningful interpretations. Unlike tools, databases require continuous annotation. Otherwise, user may not get information, if the input query is not stored or updated in the database.

5.1. Machine learning processes for prediction

Commonly used local linear alignment tools such as BLAST are unable to precisely predict PTM sites in a given protein sequence, therefore, better prediction algorithms such as weight

Table 5 – Resources to study other types of PTMs.

SR no	Database/tool (predictor)	Features
1	Ubiprot [47] http://ubiprot.org.ru	DB of protein substrates of ubiquitylation. Manually curated database on verified literature.
2	<i>Saccharomyces Cerevisiae</i> Ubiquitination Database(SCUD) [112] http://scud.kaist.ac.kr	DB with information about ubiquitinated proteins and related enzymes in <i>S. cerevisiae</i> . Has entry of over 940 substrates.
3	PlantsUPS [113] http://bioinformatics.cau.edu.cn/plantsUPS	DB of ubiquitin/26S proteasome system of 7 species of higher plants. Vast information on basic gene characterization, protein features and microarray information as well as BLAST hits against various DB data.
4	Database Of Bacterial Lipoproteins (DOLOP) [46] http://www.mrc-lmb.cam.ac.uk/genomes/dolop/	DB of Bacterial Lipoproteins. Probable lipoproteins from 234 bacterial genomes.
5	Disulphide Database DSDBASE [48] http://caps.ncbs.res.in/dsdbase/dsdbase.html	DB of disulphide bonds in proteins, which provides information on native disulfides. Records 2385617 protein substructures that have stereochemical compatibility.
6	E3Miner [114] http://e3miner.biopathway.org	Text mining tool for ubiquitin-protein ligases. Extracting and managing data from MEDLINE abstracts and relevant protein databases.
7	UbiPred [68] http://flipper.diff.org/app/tools/info/2503	Predict ubiquitylation sites in query sequence. SVM built on informative physicochemical property mining algorithm (IPMA).
8	SulfoSite [115] http://sulfosite.mbc.nctu.edu.tw/	Tool to predict protein sulfotyrosine sites. Uses SVM.
9	Sulfinator [71] http://au.expasy.org/tools/sulfinator/	A tool predicts tyrosine sulfation sites in protein sequences. HMM used.
10	NetAcet [116] http://www.cbs.dtu.dk/services/NetAcet/	Predicts N-acetyltransferase A (Nata) substrates (in yeast and mammalian proteins). ANN used.
11	Myristoylator [64] http://web.expasy.org/myristoylator/	Predicts N-terminal myristoylation by neural networks.
12	NMT-The MYR Predictor [117] http://mendel.imp.ac.at/myristate/SUPLpredictor.htm	Predicts N-Myristoylation for higher Eukaryote, viral and fungal query sequences. Self-consistency and Jack-knife test used.
13	LipoP [118] http://www.cbs.dtu.dk/services/LipoP/	Predict lipoprotein signal peptides in Gram-negative Eubacteria. HMM used.
14	NBA-Palm [65] http://www.bioinfo.tsinghua.edu.cn/NBA-Palm	Predicts palmitoylation. SVMs, Naïve Bayes algorithm used.
15	Clustering and scoring strategy CSS-Palm [119] http://csspalm.biocuckoo.org/online.php	Software for predicting Palmitoylation site. Clustering and scoring strategy (CSS) algorithm used.
16	Prenylation Prediction Suite (PrePS) [67] http://mendel.imp.ac.at/sat/PrePS/index.html	Predicts prenylation motifs.
17	SUMOplot [69] http://www.abgent.com/tools/sumoplot_login	Predicts SUMO protein attachment sites and scores sumoylation sites in proteins. BLOSUM62 for matrix, Matthews' correlated coefficient used.
18	SUMOSp [120] http://sumosp.biocuckoo.org/	Predicts sumoylation sites. BLOSUM62 for matrix, Matthews' correlated coefficient used.
19	LysAcet [121] http://www.biosino.org/LysAcet/	Performs lysine acetylation prediction. SVM used.
20	Methylation Modification Prediction Server MeMo [122] http://www.bioinfo.tsinghua.edu.cn/~tigerchen/memo.html	Tool to predict protein methylation modifications mostly Lysine and Arginine. Uses SVMs
21	BPB-PPMS [123] http://www.bioinfo.bio.cuhk.edu.hk/bpbppms/intro.jsp	Predicts the methylation for lysine and arginine residues by using Bi-profile Bayes approach and SVM.

DB—Database, SVM—Support Vector Machines, NMT—N-terminal N-Myristoylation, HMM—Hidden Markov Models.

matrices and machine learning techniques are evolved. Machine learning process involves teaching the system and building the algorithm through which it is made to learn and mimic the biological phenomenon of PTMs with experimentally proved training dataset. Then the same algorithm is trained to predict PTMs for a test set while it is checked for true and false prediction. Hence, unlike simple local alignment tool, which just aligns the input sequence with that of the stored sequence, the machine learning enables input of specific properties and makes prediction similar to an enzyme recognizing its motif to bring in PTM.

Supervised machine learning techniques majorly involves two aspects; training and testing. Training the system is

done through the potential PTM datasets, which are experimentally derived or mined from known repositories containing curated data from literature. The training dataset should have optimized level of positive and negative PTM sites in it. Once a high quality dataset is ready, the input feature for prediction is provided and learning functions are selected. This prediction process is performed on training dataset to generate a classifier, which can be further used for prediction of test dataset. The established classifier is then tested for its performance on an independent test dataset. This is checkpoint for accurate prediction of PTM sites. Inaccurate prediction due to less sensitivity and specificity during the validation brings a need to check the built algorithm. It is

very essential that the data used for training the predictor algorithm must be error free, because an error will subsequently produce erroneous PTM prediction. Hence, trained algorithms provide positive and negative prediction outputs for PTM and non-PTM sites, respectively. Presently predictors tend to possess more sensitivity than specificity but one need to balance between specificity and sensitivity for comprehensive prediction [49,50].

Supervised machine learning approaches typically use some of the learning functions such as Artificial Neural Network (ANN), Support Vector Machine (SVM) and Hidden Markov Models (HMMs) etc., which are unique to each predictor. For instance, the SVM method is designed to maximize the margin to separate two classes so that the trained model can be generalized to predict the data [51]. PHOSIDA enables Support Vector Machine (SVM) based prediction of phosphorylation and acetylation with 78% precision [33]. ANN is inspired by the biological neural networks, in which each data point is represented as neuron and interconnected as biological neuron. Multiple inputs are provided to generate single output from each node and each point is trained to predict PTM site accurately. NetPhosK is one such ANN based kinase specific phosphorylation site predictor, which is based on data specific to six kinases [39].

Several prediction tools for various types of PTMs are available for public use in WWW (Fig. 4). Single prediction server may predict one or many PTMs. FindMod present in ExPASy server predicts about 22 different types of PTMs from input PMF data, Swiss-Prot ID and single-letter amino acid code. Neural network based tool such as NetNGlyc predicts specifically N-glycosylation sites in human proteins. Fengfeng Zhou and colleagues created a user interface, which collects 32 different types of PTMs and reduces effort of searching different resources for multiple PTM prediction [52].

5.2. Strong prediction tools by evolving models

Similar to the advancement in mass spectrometry and other technologies for PTM identification, the PTM prediction tools are also evolving. The present day tools are not restricted to specific sequence motifs but are much more advanced to provide evolutionary aspects, kinase specific information, consideration of flanking amino acids etc. For example, NetPhosK was introduced to predict phosphorylation sites based on simple sequence motifs but later on kinase specific information and a concept of 'evolutionary stable sites' were introduced to make the predictions more precise [39]. The dbPTM considers secondary and tertiary structures, solvent accessibility of the substrate and protein domains for prediction [53], whereas PHOSIDA considers structure of the motif and evolutionary conservation of phosphosites. The BlastP was used to perform homology search of all phosphoproteins over seventy species ranging from bacteria to mouse [33].

Organism specific predictors of PTMs can provide strong prediction model by involving and teaching the system to predict the PTMs with an orientation of biological phenomenon pertaining to a particular organism. More organism specific databases and predictors are described in section 6. With more and more global proteomes being analyzed, larger training datasets are generated and it is anticipated that better predictions would evolve in days to come. In the following

section, a few common PTMs and their corresponding tools for prediction are discussed.

5.3. Phosphorylation tools

Phosphorylation prediction tools outnumber predictors of any other PTMs. Although there are over 500 kinases in human genome, referring to several phosphorylation sites, only a few kinases are extensively studied. Therefore, predictors of phosphorylation sites are very informative. Initially, phosphorylation site prediction was only kinase driven approach, which followed screening by incubation of a specific kinase with large set of peptides and ATP, and resulting phosphorylation patterns were studied to understand kinase substrate interaction [30,54]. Current MS based screening and *in silico* predictions are very efficient and far more rapid. The pattern recognition algorithms such as ANN, SVM etc are trained for this purpose to recognize phosphorylation sites with high sensitivity and specificity.

A typical kinase specific phosphorylation site predictor, such as NetPhosK, predicts the phosphorylation site using artificial network and gives an option to choose from 18 different kinases. Another predictor YinOYang, in conjunction with NetPhos predicts the potential sites on protein that may undergo both glycosylation and phosphorylation [39]. A non-kinase driven motif prediction approach was taken in case of Motif-X, which is a substrate driven predictor for phosphorylation motifs. In this regard, authors developed a statistical algorithm using two database contemporary datasets [55].

5.4. Glycosylation tools

The glycosylation process with respect to enzymes involved and their specificity is less understood. This is because attachment of saccharide chain on protein is complex as compared to other PTMs such as methylation, acetylation or phosphorylation, where transfer of chemical and covalent attachment is more or less a single step process. N and O-linked glycosylations are two mostly studied glycosylation types and hence much of predictors are centered to these types.

NetOglyc is a predictor for mammalian mucin type GalNAc O-glycosylation sites. The authors showed high confidence prediction of glycosylation sites by training ANN with data extracted from O-GLYCBASE. This prediction was based on action of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases complex, its specificity for Serine or Threonine and charged residues in -1 and +3 position. ANN trained in all these features resulted in good prediction (76% prediction of glycosylated and 93% of non-glycosylated sites) [56]. One of the SVM based mucin-type O-glycosylation site predictors of mammalian protein is Oglyc, which considers the sequence of O-glycosylation, physical properties of amino acids and binary way of representing the sequence. Positive and negative datasets for training SVM were obtained from Swiss-Prot/UniProt [57]. Glycosylation prediction program (GPP) uses random forest along with information about pair wise pattern to predict glycosylation sites. Training dataset was extracted from OGLYCBASE. Authors have made an effort to make the prediction better by introducing further information such as hydrophobicity, predicted secondary structure and predicted surface accessibility

[58]. NetNglyc predicts N-Glycosylation sites in human proteins using ANN and it was trained with 469 positive and 309 negative glycosylation sites. During cross validation, the method yielded identification of 86 and 61% of glycosylated and non-glycosylated sites, respectively [59]. Other types of glycosylation such as GPI anchoring, C-mannosylation, and glycations have also attracted researchers attention to develop supervised machine learning based predictors. To name a few, BigPI predicts Glycosylphosphatidylinositol (GPI) attachment sites [60], NetC-Glyc predicts mammalian C-mannosylation sites [61] and Net-Glycate predicts glycation sites for mammalian proteins [62]. Since studying glycosylation sites is complex, availability of glycosylation tools would definitely help researchers to provide predictions of this crucial PTM.

5.5. Other PTM related tools

Majority of PTM tools are centered on two most common PTMs, phosphorylation and glycosylation; however, there are around 300 other types of PTMs, which are also biologically important. Few prominent tools are discussed in this section and detailed predictors for various types of PTMs are described in Table 5. Myristoylation is addition of myristoyl group to the N-terminal of glycine residue of a protein with an enzyme N-myristoyltransferase. Apart from eukaryotes, myristoylation is also found to be associated with viral proteins [63]. Myristoylator, a predictor for myristoylation sites in a given amino acid sequence, is one of the best myristoylation predictors available with false discovery positive error rate of 2.1% [64]. This predictor uses neural network, trained with 390 positive and 327 negative sequences. NBA-palm, a predictor of palmitoylated sites, is designed on basis of Naïve Bayes algorithm. The algorithm was trained with

help of 245 palmitoylated sites from 105 non-redundant proteins [65].

Prenylation is another PTM in which an isoprenoid tail is attached to the end of a substrate protein's cysteine residues in C-terminal. Prenylation helps to localize the protein to cellular membranes and aids in mediating protein–protein interactions [66]. PrePS is a predictor of prenylation site, which is designed on the basis of experimentally established sequence recognition patterns of enzymes such as farnesyltransferase, geranylgeranyltransferase 1 and 2. The web interface predicts prenylation for all three enzymes [67]. UbiPred is ubiquitination predictor tool, which adopts a random forest classifier and predicts potential ubiquitination sites in given query proteins. This classifier was trained on a set of 266 non-redundant empirically verified ubiquitination sites [68]. SUMOplot predicts SUMOylation sites in given protein sequence [69]. Sulfonation is a unique type of PTM, which transfers sulfate group to the protein. It is found to be evolutionarily conserved in proteins of wide range of organisms e.g. *Plasmodium falciparum*, invertebrate such as *Lymnaea stagnalis* and humans, there by linking it in broader evolutionary context [70]. Sulfinator is a Hidden Markov Model based tool available on ExPASy server which predicts tyrosin sulfonation sites [71]. Availability of various PTM tools is very crucial for PTM research and it is important to continuously update and annotate these databases and tools.

6. Organism specific database and tools

Organism specific predictors help to predict the PTM patterns of an organism by using prediction systems, which are trained specifically in context of an organism. This enables incorporation of





ORGANISM	DATABASE/TOOL(PREDICTOR)
PLANTS 	Databases: PhosPhAt, ProMEX, P ³ DB, PlantsP, plantsUPS, Tool: PhosPhAt.
HUMAN 	Databases: Human Protein Reference database(HPRD), Phospho.ELM, PhosphoPOINT, PhosphoNET –Human Phosphosite KnowledgeBase, PhosphoPep, NetworkKIN, LymPHOS, Kinomer, PepCyber :P~Pep, PhosSNP, UniPep, UbiProt. Tools: NetworkKIN, NetPhosK, PostMod, Group-based Prediction system (GPS), PhosphoMotif Finder, PhosphoBlast, NetOGlyc, NetNGlyc, YinOYang, OGPET, NetCGlyc, NetAcet.
BACTERIA, FUNGI 	Databases: The Phosphorylation Site Database, Database Of Bacterial Lipoproteins (DOLOP), PhosphoGRID, Saccharomyces Cerevisiae Ubiquitination Database(SCUD), PhosphoPep. Tools: NetPhosBac, NMT-The MYR Predictor, big-PIPredictor, LipoP, NetPhosYeast.
DROSOPHILA, MOUSE 	Databases: PhosphoPep, PhosphoregDB. Tools: NetPhosK, PhosphoBlast, YinOYang, OGPET.

Fig. 5 – Organism specific PTM databases and prediction tools: Organism specific tools and databases are more precise in providing information about PTMs in context of organism. This approach is proved better for PTM prediction when compared to those of general ones due to higher sensitivity and specificity of prediction. Some of the common organism specific databases and tools are presented.

unique features into training dataset pertaining to a specific organism and makes the prediction more targeted (Fig. 5). Ramneek Gupta and colleagues developed, DictyOGlyc, which is ANN based O-linked GlcNAc glycosylation predictor for secreted and membrane proteins of *Dictyostelium discoideum*. This predictor is useful since *Dictyostelium* serves as a model organism to study glycosylation of eukaryotes [72]. NetPhosBac and NetPhosYeast are two ANN-based kinase predictors for bacterial and yeast proteins, respectively [73,74]. These predictors can be used to screen potential substrates of protein kinases or to distinguish phosphorylated and non-phosphorylated protein forms. PhosphAT, a database of phosphorylation sites in *Arabidopsis*, contains data produced by MS experiments and annotations in database are linked with mass spectrum to provide information in context of their biological significance. This database also contains a predictor, which predicts the potential phosphorylation sites [75]. virPTM is first virus PTM database which contains manually curated information for 329 phosphorylation sites of 53 different human viruses [76]. A detailed list of organism-specific database and tools are listed in Figure 5.

7. Conclusions

Proteomics has enhanced our understanding of diverse and complex PTMs by applying various advanced techniques such as mass spectrometry. Enormous data generated by various HT methods are curated and shared worldwide with help of dedicated databases. The growing number of databases provides researchers various resources; however, due to the complexity of PTMs, no database can provide a holistic solution. To provide accurate information, the databases require continuous update of new dataset. In this regard, Swiss Prot, HPRD, Phospho.ELM etc. provide comprehensive and latest collection of data. Apart from the database, the PTM prediction tools also play an important role due to their rapidness and accuracy of prediction. These databases and tools are powerful resources to reduce the time and effort of researchers for PTM prediction since identifying PTMs *in vitro* is always challenging. However, by using newer and advanced machine learning methods these prediction tools can be made more effective. The comprehensive tables for variety of PTMs described in this article will provide researchers a resource for selecting PTM databases and tools best suited for their PTM research; however, based on the type of application, one must be very careful in choosing the right database or predictor. The user may look into details such as the algorithm used for prediction, the accuracy of prediction etc. Similarly, the databases can be looked into details such as how often and effectively they are updated. Undoubtedly, in years to come, the computational resources will provide great value, and extend its analysis to decipher the role of PTMs in signaling pathways and related ailments.

Acknowledgments

The financial support from Ministry of Human Resource and Development India (MHRD-10MHRD005), Board of Research in Nuclear Sciences India (2009/20/37/4/BRNS) and Department of

Biotechnology India (BT/PR13562/MED/12/451/2010) for the proteomics work performed in SS laboratory is gratefully acknowledged. The help rendered by Renisa D'souza in drawing figures is gratefully acknowledged.

REFERENCES

- [1] Cox J, Mann M. Is proteomics the new genomics? *Cell* 2007;130:395–8.
- [2] Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. *Nat Methods* 2007;4:798–806.
- [3] Wojcik J, Schachter V. Proteomic databases and software on the web. *Brief Bioinform* 2000;1:250–9.
- [4] Walsh CT, Garneau-Tsodikova S, Gatto Jr GJ. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl* 2005;44(45):7342–72.
- [5] Zhang J, Zhang H, Ayaz-Guner S, Chen YC, Dong X, Xu Q, et al. Phosphorylation, but not alternative splicing or proteolytic degradation, is conserved in human and mouse cardiac troponin T. *Biochemistry* 2011;50(27):6081–92.
- [6] Macek B, Gnad F, Soufi B, Kumar C, Olsen JV, Mijakovic I, et al. Mol Cell Proteomics. Phosphoproteome analysis of *E. coli* reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics* 2008;7(2):299–307.
- [7] Martin L, Latypova X, Terro F. Post-translational modifications of tau protein: implications for Alzheimer's disease. *Neurochem Int* 2011;58(4):458–71.
- [8] Gioeli D, Paschal BM. Post-translational modification of the androgen receptor. *Mol Cell Endocrinol* in press. Jul 24. [PMID:21820033].
- [9] Macek B, Mann M, Olsen JV. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu Rev Pharmacol Toxicol* 2009;49:199–221.
- [10] Zong C, Young GW, Wang Y, Lu H, Deng N, Drews O, et al. Two-dimensional electrophoresis based characterization of post-translational modifications of mammalian 20S proteasome complexes. *Proteomics* 2008;8:5025–37.
- [11] Wu CC, MacCoss MJ. Shotgun proteomics: tools for the analysis of complex biological systems. *Curr Opin Mol Ther* 2002;4(3):242–50.
- [12] Jensen ON. Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol* 2006;7:391–403.
- [13] Siuti N, Kelleher NL. Decoding protein modifications using top-down mass spectrometry. *Nat Methods* 2007;4(10):817–21.
- [14] Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, et al. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. *Nucleic Acids Res* 2007;35:W701–6.
- [15] Mann M. Functional and quantitative proteomics using SILAC. *Nat Rev Mol Cell Biol* 2006;7(12):952–8.
- [16] Cuomo A, Moretti S, Minucci S, Bonaldi T. SILAC-based proteomic analysis to dissect the “histone modification signature” of human breast cancer cells. *Amino Acids* 2011;41(2):387–99.
- [17] Olsen JV, Vermeulen M, Santamaria A, Kumar C, Miller ML, Jensen LJ, et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* 2010;3(104):ra3.
- [18] Skorobogatko YV, Deuso J, Adolf-Bryfogle J, Nowak MG, Gong Y, Lippa CF, et al. Human Alzheimer's disease synaptic O-GlcNAc site mapping and iTRAQ expression proteomics with ion trap mass spectrometry. *Amino Acids* 2011;40(3):765–79.
- [19] Iwai LK, Benoist C, Mathis D, White FM. Quantitative phosphoproteomic analysis of T cell receptor signaling in

- diabetes prone and resistant mice. *J Proteome Res* 2010;9(6): 3135–45.
- [20] Hoffert JD, Pisitkun T, Wang G, Shen RF, Knepper MA. Quantitative phosphoproteomics of vasopressin-sensitive renal cells: regulation of aquaporin-2 phosphorylation at two sites. *Proc Natl Acad Sci U S A* 2006;103(18):7159–64.
- [21] Zhai B, Villen J, Beausoleil SA, Mintseris J, Gygi SP. Phosphoproteome analysis of *Drosophila melanogaster* embryos. *J Proteome Res* 2008;7:1675–82.
- [22] Choudhary C, Kumar C, Gnäd F, Nielsen ML, Rehman M, Walther TC, et al. Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* 2009;325:834–40.
- [23] Zielinska DF, Gnäd F, Wiśniewski JR, Mann M. Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* 2010;141(5): 897–907.
- [24] Wagner SA, Beli P, Weinert BT, Nielsen ML, Cox J, Mann M, et al. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol Cell Proteomics* in press. Sep 1. PMID: 21890473.
- [25] Kung LA, Tao SC, Qian J, Smith MG, Snyder M, Zhu H. Global analysis of the glycoproteome in *Saccharomyces cerevisiae* reveals new roles for protein glycosylation in eukaryotes. *Mol Syst Biol* 2009;5:308.
- [26] Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, et al. Global analysis of protein phosphorylation in yeast. *Nature* 2005;438:679–84.
- [27] Popescu SC, Popescu GV, Bachan S, Zhang Z, Gerstein M, Snyder M, et al. MAPK target networks in *Arabidopsis thaliana* revealed using functional protein microarrays. *Genes Dev* 2009;23(1):80–92.
- [28] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database—2009 update. *Nucleic Acids Res* 2009;37:D767–72.
- [29] Farriol MN, Garavelli JS, Boeckmann B, Duvaud S, Gasteiger E, Gateau A, et al. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* 2004;4:1537–50.
- [30] Lemeer S, Heck AJ. The phosphoproteomics data explosion. *Curr Opin Chem Biol* 2009;13:414–20.
- [31] Kreegipuu A, Blom N, Brunak S. PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res* 1999;27: 237–9.
- [32] Diella F, Cameron S, Gemund C, Linding R, Via A, Kuster B, et al. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 2004;5:79.
- [33] Gnäd F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* 2011;39:D253–60.
- [34] <http://www.phosphosite.org/homeAction.do>.
- [35] Zanzoni A, Ausiello G, Via A, Gherardini PF, Helmer-Citterich M. Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res* 2007;35: D229.c–31.c.
- [36] Gribskov M, Fana F, Harper J, Hope DA, Harmon AC, Smith DW, et al. PlantsP: a functional genomics database for plant phosphorylation. *Nucleic Acids Res* 2001;29: 111–3.
- [37] Bodenmiller B, Campbell D, Gerrits B, Lam H, Jovanovic M, Picotti P, et al. PhosphoPep — a database of protein phosphorylation sites in model organisms. *Nat Biotechnol* 2008;26:1339–40.
- [38] Ovelheiro D, Carrascal M, Casas V, Abian J. LymPHOS. Design of a phosphosite database of primary human T cells. *Proteomics* 2009;9:3741–51.
- [39] Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 2004;4:1633–49.
- [40] Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res* 2003;31:511–3.
- [41] Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. O-GLYCBASE Version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 1999;27:370–2.
- [42] Cooper CA, Wilkins MR, Williams KL, Packer NH. BOLD—a biological O-linked glycan database. *Electrophoresis* 1999;20: 3589–98.
- [43] Wang J, Torii M, Liu H, Hart GW, Hu ZZ. dbOGAP — an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics* 2011;12:91.
- [44] Weerapana E, Imperiali B. Asparagine-linked protein glycosylation: from eukaryotic to prokaryotic systems. *Glycobiology* 2006;16(6):91R–101R.
- [45] Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, et al. UniPep — a database for human N-linked glycosites: a resource for biomarker discovery. *Genome Biol* 2006;7: R73.
- [46] Babu MM, Priya ML, Selvan AT, Madera M, Gough J, Aravind L, et al. A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins. *J Bacteriol* 2006;188(8):2761–73.
- [47] Chernorudskiy AL, Garcia A, Eremin EV, Shorina AS, Kondratieva EV, Gainullin MR. UbiProt: a database of ubiquitylated proteins. *BMC Bioinformatics* 2007;8:126.
- [48] Vinayagam A, Pugalenth G, Rajesh R, Sowdhamini R. DSDBASE: a consortium of native and modeled disulphide bonds in proteins. *Nucleic Acids Res* 2004;32:D200–2.
- [49] Eisenhaber B, Eisenhaber F. Prediction of posttranslational modification of proteins from their amino acid sequence. In: Oliviero C, Eisenhaber F, editors. *Methods in molecular biology. Data mining techniques for the life sciences*. Humana Press; 2010. p. 365–84.
- [50] Liu C, Li H. In silico prediction of post-translational modifications. In: Bing Y, Marcus H, editors. *Methods in molecular biology. Silico tools for gene discovery*. Humana Press; 2011. p. 325–40.
- [51] Yang ZR. Biological applications of support vector machines. *Brief Bioinform* 2004;5:328–38.
- [52] Zhou F, Xue Y, Yao X, Xu Y. A general user interface for prediction servers of proteins' post-translational modification sites. *Nat Protoc* 2006;1(3):1318–21.
- [53] 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.
- [54] Mann M, Ong SE, Grønberg M, Steen H, Jensen ON, Pandey A. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol* 2002;20(6):261–8.
- [55] Schwartz D, Gygi SP. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* 2005;23(11):1391–8.
- [56] Julenius K, Mølgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* 2005;15(2):153–64.
- [57] Li S, Liu B, Zeng R, Cai Y, Li Y. Predicting O-glycosylation sites in mammalian proteins by using SVMs. *Comput Biol Chem* 2006;30(3):203–8.
- [58] Hamby SE, Hirst JD. Prediction of glycosylation sites using random forests. *BMC Bioinformatics* 2008;9:500.
- [59] Julenius K, Johansen MB, Zhang Y, Brunak S, Gupta R. Prediction of glycosylation sites in proteins. In: Lieth CW, Luetke T, Frank M, editors. *Bioinformatics for glycobiology*

- and glycomics: an introduction. Wiley-Blackwell; 2009. p. 163–92.
- [60] Eisenhaber B, Bork P, Eisenhaber F. Prediction of potential GPI-modification sites in proprotein sequences. *J Mol Biol* 1999;292(3):741–58.
- [61] Julenius K. NetCGlyc 1.0: prediction of mammalian C-mannosylation sites. *Glycobiology* 2007;17:868–76.
- [62] Johansen MB, Kierner L, Brunak S. Analysis and prediction of mammalian protein glycation. *Glycobiology* 2006;16(9):844–53.
- [63] Boutin JA. Myristoylation. *Cell Signal* 1997;9:15–35.
- [64] Bologna G, Yvon C, Duvaud S, Veuthey AL. N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics* 2004;4:1626–32.
- [65] Xue Y, Chen H, Jin C, Sun Z, Yao X. NBA-Palm: prediction of palmitoylation site implemented in Naive Bayes algorithm. *BMC Bioinformatics* 2006;7:458.
- [66] Houglund JL, Fierke CA. Getting a handle on protein prenylation. *Nat Chem Biol* 2009;5:197–8.
- [67] Maurer SS, Eisenhaber F. Refinement and prediction of protein prenylation motifs. *Genome Biol* 2005;6:R55.
- [68] Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, et al. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 2010;78:365–80.
- [69] Zhou F, Xue Y, Lu H, Chen G, Yao X. A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus. *FEBS Lett* 2005;579:3369–75.
- [70] Medzihradsky KF, Darula Z, Perlson E, Fainzilber M, Chalkley RJ, Ball H, et al. O-sulfonation of serine and threonine Mass spectrometric detection and characterization of a new posttranslational modification in diverse proteins throughout the eukaryotes. *Mol Cell Proteomics* 2004;3:429–43.
- [71] Monigatti F, Gasteiger E, Bairoch A, Jung E. The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics* 2002;18:769–70.
- [72] Gupta R, Jung E, Gooley AA, Williams KL, Brunak S, Hansen J. Scanning the available *Dictyostelium discoideum* proteome for O-linked GlcNAc glycosylation sites using neural networks. *Glycobiology* 1999;9:1009–22.
- [73] Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. NetPhosBac — a predictor for Ser/Thr phosphorylation sites in bacterial proteins. *Proteomics* 2009;9:116–25.
- [74] Ingrell CR, Miller ML, Jensen ON, Blom N. NetPhosYeast. Prediction of protein phosphorylation sites in yeast. *Bioinformatics* 2007;23:895–7.
- [75] Joshua LH, Pawel D, Jan H, Joachim S, Wolfram W, Dirk W, et al. PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res* 2008;36:D1015–21.
- [76] Schwartz D, Church GM. Collection and motif-based prediction of phosphorylation sites in human viruses. *Sci Signal* 2010;3(137):rs2.
- [77] Li H, Xing X, Ding G, Li Q, Wang C, Xie L, et al. SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol Cell Proteomics* 2009;8:1839–49.
- [78] Garavelli JS. The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res* 2003;31:499–501.
- [79] Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, et al. High-throughput mass spectrometric discovery of protein post-translational modifications. *J Mol Biol* 1999;289:645–57.
- [80] Wurgler MSM, King DM, Kennelly PJ. The Phosphorylation Site Database. A guide to the serine-, threonine-, and/or tyrosine-phosphorylated proteins in prokaryotic organisms. *Proteomics* 2004;4:1562–70.
- [81] Yang CY, Chang CH, Yu YL, Lin TC, Lee SA, Yen CC, et al. PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics* 2008;24:i14–20.
- [82] <http://www.phosphonet.ca/Default.aspx?AspxAutoDetectCookieSupport=1>.
- [83] Hummel J, Niemann M, Wienkoop S, Schulze W, Steinhauser D, Selbig J, et al. PromEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics* 2007;8:216.
- [84] Forrest AR, Taylor DF, Fink JL, Gongora MM, Flegg C, Teasdale RD, et al. PhosphoregDB: the tissue and sub-cellular distribution of mammalian protein kinases and phosphatases. *BMC Bioinformatics* 2006;7:82.
- [85] Stark C, Su TC, Breitkreutz A, Lourenco P, Dahabieh M, Breitkreutz BJ, et al. PhosphoGRID: a database of experimentally verified *in vivo* protein phosphorylation sites from the budding yeast *Saccharomyces cerevisiae*. *Database* 2010;2010:bap026 2010 Jan 28 (PMID:20428315).
- [86] Gao J, Agrawal GK, Thelen JJ, Xu D. P3DB: a plant protein phosphorylation database. *Nucleic Acids Res* 2009;37:D960–2.
- [87] Linding R, Jensen LJ, Pasulescu A, Olhovskiy M, Colwill K, Bork P, et al. NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* 2008;36:D695–9.
- [88] Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM, et al. Systematic discovery of *in vivo* phosphorylation networks. *Cell* 2007;129:1415–26.
- [89] Saunders NF, Brinkworth RI, Huber T, Kemp BE, Kobe B. Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics* 2008;9:245.
- [90] Niedner RH, Buzko OV, Haste NM, Taylor A, Gribskov M, Taylor SS. Protein kinase resource: an integrated environment for phosphorylation research. *Proteins* 2006;63:78–86.
- [91] Martin DM, Miranda-Saavedra D, Barton GJ. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. *Nucleic Acids Res* 2009;37:D244–50.
- [92] Krupa A, Abhinandan KR, Srinivasan N. KinG: a database of protein kinases in genomes. *Nucleic Acids Res* 2004;32:D153–5.
- [93] Gong W, Zhou D, Ren Y, Wang Y, Zuo Z, Shen Y, et al. PepCyber:P PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. *Nucleic Acids Res* 2008;36:D679–83.
- [94] Ren J, Jiang C, Gao X, Liu Z, Yuan Z, Jin C, et al. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol Cell Proteomics* 2010;9:623–34.
- [95] Halligan BD. ProMoST: A tool for calculating the pI and molecular mass of phosphorylated and modified proteins on 2 dimensional gels. *Methods Mol Biol* 2009;527:283.
- [96] Rutenberg BE, Pisitkun T, Knepper MA, Hoffert JD. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. *J Proteome Res* 2008;7:3054–9.
- [97] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 1999;294:1351–62.
- [98] Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, et al. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res* 2007;35:W588–94.
- [99] Jung I, Matsuyama A, Yoshida M, Kim D. PostMod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC Bioinformatics* 2010;11:S10.
- [100] Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 2004;32:1037–49.
- [101] Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, et al. Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal* 2008;1:ra2.

- [102] Neuberger G, Schneider G, Eisenhaber F. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol Direct* 2007;2:1.
- [103] Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X. GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res* 2005;33:W184–7.
- [104] Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0. Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* 2003;31:3635–41.
- [105] Amanchy R, Periaswamy B, Mathivanan S, Reddy R, Tattikota SG, Pandey A. A curated compendium of phosphorylation motifs. *Nat Biotechnol* 2007;25:285–6.
- [106] Wan J, Kang S, Tang C, Yan J, Ren Y, Liu J, et al. Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res* 2008;36:e22.
- [107] Dang TH, Van Leemput K, Verschoren A, Laukens K. Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics* 2008;24:2857–64.
- [108] Li T, Li F, Zhang X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins* 2008;70:404–14.
- [109] Xue Y, Li A, Wang L, Feng H, Yao X. PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* 2006;7:163.
- [110] Wang Y, Klemke RL. PhosphoBlast, a computational tool for comparing phosphoprotein signatures among large datasets. *Mol Cell Proteomics* 2008;7:145–62.
- [111] <http://ogpet.utep.edu/OGPET/>.
- [112] Lee WC, Lee M, Jung JW, Kim KP, Kim D. SCUD: Saccharomyces Cerevisiae Ubiquitination Database. *BMC Genomics* 2008;9:440.
- [113] Du Z, Zhou X, Li L, Su Z. plantsUPS: a database of plants' Ubiquitin Proteasome System. *BMC Genomics* 2009;10:227.
- [114] Lee H, Yi GS, Park JC. E3Miner: a text mining tool for ubiquitin-protein. *Nucleic Acids Res* 2008;36:W416–22.
- [115] Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, Hsu PC, et al. Incorporating support vector machine for identifying protein tyrosine sulfation sites. *Sulfosite*. *J Comput Chem* 2009;30:2526–37.
- [116] Kiemer L, Bendtsen JD, Blom N. NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics* 2005;21:1269–70.
- [117] Maurer SS, Eisenhaber B, Eisenhaber F. N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J Mol Biol* 2002;317:541–57.
- [118] Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* 2003;12:1652–62.
- [119] Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X. CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel* 2008;21:639–44.
- [120] Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res* 2006;34:W254–7.
- [121] Xu Y, Wang XB, Ding J, Wu LY, Deng NY. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *J Theor Biol* 2010;264:130–5.
- [122] Chen H, Xue Y, Huang N, Yao X, Sun Z. MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res* 2006;34:W249–53.
- [123] Shao J, Xu D, Tsai SN, Wang Y, Ngai SM. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One* 2009;4(3):e4920.