Scop3P: a comprehensive resource of human phosphosites within their full context

*Pathmanaban Ramasamy[1,2,3,4,5], Demet Turan[1,2], Natalia Tichshenko[1,2], Niels Hulstaert[1,2], Elien Vandermarliere [1,2], Wim Vranken [3,4,5 *], Lennart Martens [1,2 **]*

[1] VIB-UGent Center for Medical Biotechnology, VIB, Ghent 9000, Belgium.
[2]Department of Biomolecular Medicine, Faculty of Health Sciences and Medicine, Ghent University, Ghent 9000, Belgium.
[3] Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, 1050 Brussels, Belgium.
[4] Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[5] Centre for Structural Biology, VIB, 1050 Brussels, Belgium

* Corresponding author. TEL: +32 2 629 19 96; E-mail: wim.vranken@vub.be
** Corresponding author. TEL: +32 9 264 93 58; E-mail: lennart.martens@UGent.be

Abstract

Protein phosphorylation is a key post-translational modification (PTM) in many biological processes, and is associated to human diseases such as cancer and metabolic disorders. The accurate identification, annotation and functional analysis of phosphosites is therefore crucial to understand their various roles. Phosphosites (P-sites) are mainly analysed through phosphoproteomics, which has led to increasing amounts of publicly available phosphoproteomics data. Several resources have been built around the resulting phosphosite information, but these are usually restricted to protein sequence and basic site metadata. What is often missing from these resources, however, is context, including protein structure mapping, experimental provenance information, and biophysical predictions. We therefore developed Scop3P: a comprehensive database of human phosphosites within their full context. Scop3P integrates sequences (UniProtKB/Swiss-Prot), structures (PDB), and uniformly reprocessed phosphoproteomics data (PRIDE) to annotate all known human phosphosites. Furthermore, these sites are put into biophysical context by annotating each phosphoprotein with per-residue structural propensity, solvent accessibility, disordered probability, and early folding information. Scop3P presents a unique resource for visualization and analysis of phosphosites, and for understanding of phosphosite structure-function relationships.

Keywords: Phosphorylation/PTM/Protein structure/Proteomics

Introduction

Post-translational modifications (PTMs) are typically the result of the addition of a small molecule to one or more residues of a protein (Deribe et al, 2010; Santos & Lindner, 2017). PTMs can be reversible as well as irreversible, with more than 200 PTMs currently identified. Protein phosphorylation, a reversible PTM, is one of the best studied PTMs and is involved in many regulatory processes (Lim, 2005; Humphrey et al, 2015). Protein phosphorylation is regulated by three core machineries, namely kinases, phosphatases, and proteins which recognize the phosphorylation signals/P-sites (Cohen, 2002). Kinases, one of the largest protein families, can be considered as the writers of protein phosphorylation: they add a highly negatively charged phosphate group to the side chain of serines, threonines, and tyrosines and less frequently to cysteines and histidines (Peck, 2006). These phosphorylation signals can be recognized by phospho binding proteins – the readers – which interact with phosphorylation signals and proteins (Ubersax & Ferrell, 2007). Phosphatases, the erasers of the phosphate group, function opposite to kinases: they remove the phosphate group from phosphorylated residues (Ubersax & Ferrell, 2007; Shi, 2009).

Several studies have attempted to differentiate between functional and non-functional P-sites based on their evolutionary conservation (Boekhorst et al, 2008; Chen *et al*, 2010), kinase specificity (Landry *et al*, 2009; Xiao *et al*, 2016), PTM cross-talk or based on their interactions (Beltrao *et al*, 2012). However, P-site conservation may not be particularly useful to determine functional importance of a P-site as only a small fraction (~35%) of functional P-sites were reported to be conserved (Holt *et al*, 2009; Amoutzias *et al*, 2012), while some functional P-sites have been identified in poorly conserved regions (Landry *et al*, 2009; Nguyen Ba & Moses, 2010). Most of the conserved but non-functional P-sites have been accumulated due to the off-target effect of kinases (Landry *et al*, 2009) in disordered regions of proteins as these are more accessible to kinases than ordered regions are (Jin & Pawson, 2012).

The dominant means of discovering novel P-sites is mass spectrometry (MS) based phosphoproteomics, which has been the primary driver for the expansion of the known P-sites. With

the increasing public availability of data from such proteomics experiments (Martens & Vizcaíno, 2017), several databases related to one or more PTMs in one or more model organisms have been established. For instance, O-GLYCBASE (Gupta et al, 1999) is an information repository of the glycosylation status of proteins, while the Human Protein Reference Database (HPRD) (Keshava Prasad et al, 2009) contains information on experimentally annotated PTMs of human proteins, and dbPTM (Lu et al, 2013) contains experimentally verified as well as computationally annotated PTMs. Besides these generic PTM databases, PHOSIDA (Gnad et al, 2007), Phospho.ELM (Dinkel et al, 2011), Phospho3D (Zanzoni et al, 2011), PhophositePlus (Hornbeck et al, 2012), and database of Phospho-sites in Animals and Fungi (dbPAF) (Ullah et al, 2016) are all databases that focus on P-sites specifically and contain information about the sequence and/or structural features of experimentally determined P-sites. Many of these sites are also reported in UniProtKB/Swiss-Prot (Bateman et al, 2017), which contains both functional and structural annotations of proteins, but lacks direct access to important structural properties related to P-sites. Most of the abovementioned repositories collect and integrate a large number of P-sites from different sources or MS experiments, but provide very little or no information on the functional role(s) of these P-sites. Moreover, the phospho-peptides measured in different phosphoproteome experiments are typically identified with different search engines and different false discovery rate (FDR) thresholds, which may artificially increase the heterogeneity of P-sites when these are integrated into a database which has no information on the significance of the reported P-sites (Zhang *et al*, 2015). This issue can be mitigated by analyzing the entire data set as one, thus allowing the control of the global FDR threshold (Verheggen *et al*, 2017).

Because of the difficult and time-consuming process of experimental identification of P-sites, several computational P-site prediction algorithms have been developed (Blom *et al*, 1999; Li *et al*, 2008; Dou *et al*, 2014; Gao *et al*, 2016). These predictors are typically trained on data obtained from the above described public resources or on the observation of kinase specificity and sequence features to predict if a particular site can be phosphorylated. However, these predictors provide very little to no information on whether the site is functional or not. They also often neglect the importance of conformational specificity of kinases (Vandermarliere & Martens, 2013) and of structural dynamics upon phosphorylation/de-phosphorylation (Xin & Radivojac, 2012), which can lead to incorrect or non-confident P-site predictions.

Because the function of a protein and the corresponding signaling cascades are highly correlated with protein structure, and because phosphorylation status can result in protein structural re-arrangements (Johnson, 1992; Li *et al*, 2003), it is important to know where a phosphorylation is located on the protein structure to understand its possible regulatory role. Visualizing P-sites mapped onto available protein structures can provide such insight by presenting researchers with an overview of the spread of P-sites over the three-dimensional structure of a protein.

Beyond the structural context itself, there is also the biophysical context of a residue. Indeed, adding or removing a negatively charged phosphate group alters a residue's electrostatic potential (Johnson, 2009; Nishi et al, 2011). This alteration may serve as a recognition site for phospho binding proteins but can also trigger conformational transitions in the phosphoprotein. Phosphorylation can moreover modulate the binding specificity of phosphoprotein binding proteins by offering a wide range of recognition patterns based on conserved amino acid residues close to the P-sites (Liang & Van Doren, 2008; Johnson et al, 2010). Moreover, studies have shown an association between phosphorylation/de-phosphorylation events and order-disorder transitions that are in turn coupled with binding regulation (Scheinin et al, 1990; Johnson, 2009; Iakoucheva et al, 2004; Bozoky et al, 2013).

It is thus clear that a thorough analysis of the biological relevance and possible role of a given P-site needs to take place against the full context of that P-site, which consists of the P-site localization in the protein, its structural characteristics, its experimental provenance, and its biophysical properties.

Here we therefore present Scop3P, a database of human P-sites in their full context. To do so, Scop3P provides annotation for all known human P-sites based on protein sequence, 3D structure, and biophysical predictions. Moreover, Scop3P is unique in that it also provides a reliability measurement for each P-site based on the frequency with which that phosphorylation has been seen across different phosphoproteomics experiments. Importantly, these phosphoproteomics results have been obtained by a uniform, large scale re-analysis of phosphoproteomics data from the PRIDE database (Martens et al, 2005), and have been filtered by a global FDR to high reliability. In addition, Scop3P contains secondary structural propensity (helix, sheet, coils), solvent accessibility, and biophysical properties such as the probability of being a disordered region, backbone dynamics, and functional information related to phosphoproteins. Importantly, every phosphoprotein is also annotated with early folding predictions (Raimondi et al, 2017), which give an idea of residues or regions that are crucial in the folding dynamics of the protein. By providing information on early folding regions/residues, Scop3P adds further unique information on whether or not a phospho acceptor residue is crucial in forming local structural elements that influence the final fold of a protein.

Results

After populating Scop3P with all known phosphosites as obtained from UniProtKB/Swiss-Prot and the reprocessing of complete human phosphoproteomics experiments in PRIDE, the Scop3P database contains 9775 phosphoproteins, covering 48% of the 20,408 human proteins in UniProtKB/Swiss-Prot. Together, these proteins contain a combined 71640 P-sites (Table 1) of which 58758 are unique P-sites (27152: unique Swiss-Prot, 18724: unique PRIDE, 12882: shared). Of these, 40034 (experimental: 30016, by similarity 10018) are obtained from UniProtKB/Swiss-Prot annotations, and 18724 from experimental data in PRIDE. The distribution of all P-sites in Scop3P shows that 69.1% of P-sites are phosphoserine, 18.7% are phosphothreonine, 6.2% are phosphotyrosine, 3.9% are phosphohistidine, and 1.8% are phosphocysteine.

The structural data in Scop3P contains 5198 unique P-sites corresponding to 1963 phosphoproteins represented by 14665 different structures (Table 1). The structures in the database are determined from different methods, including X-ray diffraction (96%), NMR (2.5%), EM (1.1%), neutron diffraction (<1%), and other combinatorial methods (<1%). Scop3P contains 78027 human amino acid variants mapped onto 6731 phosphoproteins covering 47740 variants. 380 of these variants fall on P-sites, and 121 of these are deleterious variants associated with one or more diseases (Table 1).

Web Interface and usage of Scop3P

The information in Scop3P can be accessed by search or browse options (Fig 1A). The user can search for a protein by UniProt accession number, entry name, protein name, and keywords, or for the results of an entire experimental data set by its ProteomeXchange identifier (ID) (Vizcaíno et al, 2014). The results page displays two levels of information: the sequence level, and the structural level (Fig 1A). A quick preview will be displayed in the top panel where all P-sites are mapped in ball and stick notation on the amino acid sequence. The coloring of the ball reflects the frequency of that P-site across the different phosphoproteomics projects from PRIDE. A tooltip (triggered when hovering the mouse over the ball) will give additional information such as modification name, number of unique peptides that contain that modification, number of different PRIDE projects in which this P-site is seen, and the mapped position on the PDB structure. A preview on the right-hand side of the panel will highlight the P-site on the structure upon hovering (Fig 1A).

Data is also rendered as tables and interactive graphs. The table at the bottom of the sequence annotation gives an overview of the P-sites (Fig 1A), their source (obtained from UniProtKB/Swiss-Prot, from PRIDE, or from both, and, if found in UniprotKB/Swiss-Prot, the evidence information for the P-site (*Experimental*, *By Similarity*, or *Combined* for both). The interactive circular graph contains the residue level predicted secondary structural propensity, backbone dynamics, disorder, and early folding values of the protein in context (Fig 1B). Hovering over a residue in the graph will display the values associated with that residue. As a point of reference, the first amino acid of the protein will be colored dark in all rings of the circular plot.

The phospho-peptide table (Fig 1C) provides information on all PRIDE-derived peptides that contain one or more P-sites for a particular protein. Information such as the ProjectID, peptide sequence, start and end positions of that peptide in the protein sequence, and modified residue position in the protein and peptide sequence are displayed. Hovering over the projectID will display project metadata such as the title, species, submission type, tissues, and publication date. By clicking on the drop-down icon, project information such as project id and frequency of that peptide in the corresponding project can be accessed. In addition, the frequency of P-sites (i.e. the number of times the P-site is seen as phosphorylated or unphosphorylated in particular PRIDE projects) will be shown. This serves as an indication of reliability for that P-site.

In the mutation table, known amino acid variants and their associated diseases (if any) are given (Fig 1D). These variant details can also be viewed when hovering over the P-site's ball and stick representation in the top panel.

Finally, in the structure table (Fig 1E), all available structures for the selected protein that contain at least one P-site will be displayed. The idea of visualizing a particular P-site mapped onto multiple protein structures can give insights into the structural context in which the P-site is located in different structural conformations of that protein. In the overview, information such as PDB ID, main chain, interacting chain/molecule, secondary structural propensity, conservation scale, resolution and stoichiometry of the structure, and the position of the P-sites in the PDB structure will be displayed. By clicking on the dropdown icon, secondary structural information such as accessible surface area (ASA), buried surface area (BSA) is given. Upon clicking the PDB ID, a dedicated page for structure will be displayed where the user can color the P-sites based on solvent accessibility, frequency of the site across the different PRIDE projects, and neutral/deleterious polymorphisms (Fig 1E). A complete map of all P-sites mapped on to that particular structure can be viewed upon clicking the "show all P-sites" checkbox.

Discussion

Scop3P is a user-friendly data resource that allows the analysis of experimentally determined human P-sites in the context of protein structure. It integrates information from different knowledge bases, and shows how re-analysis of large scale public proteomics data sets can add an additional level of significance and confidence to the P-sites based on P-site frequency. Moreover, Scop3P also displays all structural and biophysical information that is available for a particular P-site. This provides additional knowledge about a P-site's spatial location, structural propensity, and accessibility in the different available conformations of a protein structure. The value of this latter analysis option is highlighted by the fact that 3464 of the 5198 P-sites with an available structure have more than one such structure available in Scop3P. Moreover, early folding, disordered propensity, and backbone dynamics data will all provide valuable added information for researchers seeking to understand whether any phospho acceptor residues or related residues in close proximity are crucial for protein folding and stability. Interestingly, by re-processing phosphoproteomics data with the aid of the ionbot search engine (https://ionbot.cloud) we could identify and annotate 1268

P-sites that are currently annotated as 'by similarity' in UniProtKB/Swiss-Prot, thus providing solid experimental evidence for these instances. Moreover, by re-processing we identified and annotated 1819 proteins that contain at least one P-site which is not annotated as P-site in UniProtKB. Over time, new phosphoprojects in PRIDE will be reprocessed and added to Scop3P, which might add further confirmed or even wholly novel P-sites to the system. The intention is that Scop3P will over time be extended to include other PTMs than can affect residues that can be phosphorylated (e.g., nitrosylation, sulfation, and glycosylation), which might help understand their potential competition for the same residue. Updates for Scop3P will take place every six months, and will include new data from UniProtKB/Swiss-Prot and PDB, newly reprocessed PRIDE data, and updated data from the various biophysical property predictors.
Because of its broad and unique data contents, Scop3P provides a unique and powerful resource to understand the impact of P-sites on human protein structure and function, and can serve as a springboard for researchers seeking to analyze and interpret a given phosphosite or phosphoprotein.

Materials and Methods

To create Scop3P we collected and integrated all available P-sites from different data sources (Fig 2A-C, Table 1). First, we retrieved all available human P-sites from UniProtKB/Swiss-Prot (Bateman et al, 2017) (release-2018_02) by parsing UniProtKB/Swiss-Prot MOD_RES records. For every P-Site the evidence annotation (experimentally determined, or by similarity) and the associated reference information were also obtained.

Re-processing of human phosphoproteomics data from PRIDE

We retrieved a list of all Human projects which are annotated to contain phosphorylations from PRIDE (Martens et al, 2005). Only those projects which are unlabeled and submitted as complete projects containing high resolution spectra files were considered (see appendix Table S1). These projects were typically originally processed with different search engines and different search settings. In order to obtain uniform data, we collected all '.RAW' files from PRIDE and converted these to Mascot Generic Format (MGF) peak files using ThermoRawFileParser (Hulstaert et al, 2019). The resulting peak files were then searched against the human complement of UniProtKB/Swiss-Prot (release-2018_02, containing 20259 protein sequences) with the target/decoy approach using ionbot (https://ionbot.cloud/; ionbot is based on MS^2PIP (Degroeve *et al*, 2013; Degroeve *et al*, 2015; Gabriels *et al*, 2019) and ReScore (Peters *et al*, 2016; C. Silva *et al*, 2019)). Results were filtered at 1% FDR. The ionbot engine searches for all modifications listed in Unimod (Creasy & Cottrell, 2004) on top of a set of user-defined fixed and variable modifications. The search settings were as follows: carbamidomethylation of cysteine was specified as fixed modification, and oxidation of methionine, phosphorylation of serine (S), phosphorylation of threonine (T), phosphorylation of tyrosine (Y), phosphorylation of cysteine (C) and phosphorylation of histidine (H) were set as variable modifications. Up to two missed cleavages per peptide were allowed. Only identified peptides with q-values <0.01 were considered for further analysis. The total search time was five days on a single Linux server with 24 cores and 30 GB of RAM memory.

Structural properties of the phosphoproteins

For every human phosphoprotein for which at least one structure was available, the modeled segment of the UniProtKB/Swiss-Prot sequence in the protein structure was scanned to check if any P-sites were within range of that segment. If the modeled segment contained at least one such P-site, the corresponding PDB (Berman et al, 2006; Rose et al, 2013) structure was used to map and visualize all matching P-sites. P-sites that fell in missing segments of structures were not considered

for structural mapping. The oligomeric state of the protein structure, solvent accessibility of the P-sites, and exposure level (buried, exposed, or in an interface region) were obtained from the Protein Interfaces, Surfaces and Assemblies (PISA) server (also known as PDBePISA) (Krissinel & Henrick, 2007). The interface details were obtained by taken into consideration the most probable quaternary structure as assigned by PISA. PISA predicts quaternary structures based on the interactions occurring in macromolecular crystals (pair of chain or ligand-chain interactions). The secondary structural assignments for P-sites with a matched structure were retrieved from DSSP (Kabsch & Sander, 1983). The eight-class classification of DSSP was grouped into three states (**helix (H):** 'alpha helix, 3/10 helix, pi helix', **strand (E)**: 'extended strand, residue in isolated beta-bridge', and **loop (C)**: 'turn, bend and the rest'). Every structure is also annotated with its determination method, resolution, and stoichiometry details.

For all phosphoproteins, regardless of existing structure match, the three states of the secondary structural propensity (helix (H), coil (C), sheet (E)) were predicted using Fast Estimator of Secondary structures (FESS), which is a component of the FELLS method (Piovesan et al, 2017). Protein biophysical characteristics such as backbone dynamics, disordered propensity, and early folding properties were predicted using DynaMine (Cilia et al, 2013), DisoMine (http://bio2byte.com/disomine), and EfoldMine (Raimondi et al, 2017), respectively. DynaMine predicts the residue-level backbone flexibility in the form of S^2 values between 0 (highly dynamic) and 1 (stable conformation), which represents how restricted the movement of the atomic bond vector is with respect to the molecular reference frame. For DisoMine, the probability cutoff of 0.5 distinguishes the (predicted) ordered and disordered state of the protein. EfoldMine predicts the early folding (EF) propensity of amino acids based on local interactions, and as such provides insight into which amino acids are likely involved in early stages of protein folding and thus shape the folding landscape of that protein. The EF propensities and binary classification based on a 0.163 probability cutoff were used to distinguish between early folding and non-early folding residues.

Conservation and variation of amino acids in phospho acceptor residues

Known amino acid variations for P-sites (phospho variants) or for sites in their close proximity may result in functional variants, e.g., through a change in kinase specificity, loss and gain of P-sites, and diseases (Ryu et al, 2009). In order to map such variants on both sequence and structure, we retrieved all curated human missense variant details from the Humsavar dataset (release 12-Sep-2018) from UniprotKB/Swiss-Prot (Fig 2A, Table 1) that are classified as disease/polymorphisms/unclassified based on their role in disease. Humsavar data contains all manually curated single amino acid polymorphisms as retrieved from literature that are associated with diseases and phenotypes. In total it contains 72,960 variants of which 40% are associated with diseases. We also obtained the evolutionary conservation from Consurf-DB (Goldenberg et al, 2009) for all P-sites mapped to a structure (Fig 2A, Table 1). These conservation values are given as scales from 1 (variable) to 9 (conserved).

Database construction, integration and content

The web interface was developed using the Spring Boot framework. JQuery, Bootstrap and Tymeleaf were used as front-end technologies. Protein structures are visualized with the aid of NGL Viewer (A. S. Rose & Hildebrand, 2015), and other protein visualizations such as the circular plot and ball and stick representation of P-sites on primary (one dimensional) amino acid sequences are developed through the D3 javascript library. Scop3P data is stored in a relational database running on MySQL 5.7.

Scop3P contains both sequence (Fig 2A,C) and structure (Fig 2B) information, for both phosphoproteins and for individual P-sites (Fig 1A-E). All obtained parameters were mapped to the

amino acid sequences of the human proteins retrieved from UniProtKB/Swiss-Prot. Sequence to structural position mapping was done with the aid of SIFTS (Velankar *et al*, 2013). P-sites which fell in the missing segments of available structures were not considered for structure mapping. Every instance with structure was annotated with secondary structural propensity and evolutionary conservation details as described earlier. Additional residue level biophysical properties such as DynaMine, DisoMine and EfoldMine predictions were annotated to UniProtKB/Swiss-Prot protein sequences. Moreover, to show the reliability of the P-sites, every P-site is annotated with the frequency of phosphorylation as found in the different phosphoproteomics experiments. As a second level of annotation, every P-site is annotated with the number of distinct peptides identified for that particular protein from different PRIDE projects, and every such peptide is then annotated with its ProteomeXchange ID, and its frequency across the different PRIDE projects.

In Scop3P, we aim to map all P-sites of a particular protein to all available three dimensional structures. Thus, if a protein has more than one structure, all structures that contain at least one P-site are retained for mapping and visualization. For each P-site with 3D structure, the assembly and interface details such as the macromolecule chain where the P-site is present (main chain), accessible surface area (ASA), buried surface area (BSA), and information about crystal contacts like complex significance score (CSS) and interacting chains/ligands are also given.

CSS score

PISA assigns a value from 0 to 1 for every complex (CSS) in the biological assembly. This value is calculated as a fractional contribution of the particular interface to the crystal assembly. Hence, in Scop3P, if a particular P-site is present in a multimeric protein (with chain XYZ) at chain X, only the interfaces with higher CSS value for the chain X are considered. For example: if the CSS value for interface XY is higher than XZ then this means that this complex – composed of chain X and chain Y – is the most probable biological assembly as predicted by PISA. Sometimes the interacting molecule will be a ligand which means that the ligand is fixed with the polymer during PISA prediction and the CSS for the main chain of P-site and this ligand is higher.

Data availability

Scop3P is available as a web-interface and can be accessed at **https://iomics.ugent.be/Scop3P**.

Acknowledgements

PR, LM, WV acknowledge funding from the Research Foundation Flanders under grant agreement number G.0328.16N. LM acknowledges funding from the European Union's Horizon 2020 Programme under Grant Agreement 823839 (H2020-INFRAIA-2018-1). NH and LM acknowledge funding from a Concerted Research Action grant from Ghent University under grant agreement number BOF12/GOA/014. EV is a postdoctoral research fellow of the Research Foundation Flanders under grant agreement number 12F0816N. The authors are grateful to all submitters to the PRIDE database for making their proteomics data publicly available.

Author contributions

LM conceived and designed the project. PR performed the data extraction and integration. DT, NT and NH developed the web-interface. WV and EV supervised the work. PR, LM, WV and EV wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Amoutzias GD, He Y, Lilley KS, Van de Peer Y & Oliver SG (2012) Evaluation and Properties of the Budding Yeast Phosphoproteome. *Mol. Cell. Proteomics* **11:** M111.009555

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Bye-AJee H, Cowley A, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Castro LG, Figueira L, Garmiri P, et al (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45:** D158–D169

Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, Villén J, Lim WA, Fraser JS, Frydman J & Krogan NJ (2012) Systematic functional prioritization of protein posttranslational modifications. *Cell* **150:** 413–425

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2005) The Protein Data Bank. *Struct. Bioinforma.***:** 181–198

Blom N, Gammeltoft S & Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294:** 1351–1362

Boekhorst J, van Breukelen B, Heck AJR & Snel B (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.* **9:**

Bozoky Z, Krzeminski M, Chong PA & Forman-Kay JD (2013) Structural changes of CFTR R region upon phosphorylation: A plastic platform for intramolecular and intermolecular interactions. *FEBS J.* **280:** 4407–4416

Chen SCC, Chen FC & Li WH (2010) Phosphorylated and nonphosphorylated serine and threonine residues evolve at different rates in mammals. *Mol. Biol. Evol.* **27:** 2548–2554

Cilia E, Pancsa R, Tompa P, Lenaerts T & Vranken WF (2013) From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.* **4:**

Cohen P (2002) The origins of protein phosphorylation. *Nat. Cell Biol.* ***4:*** *E127–E130*

Creasy DM & Cottrell JS (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics* **4:** 1534–1536

Deribe YL, Pawson T & Dikic I (2010) Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* **17:** 666–672

Degroeve S, Martens L & Jurisica I (2013) MS2PIP: A tool for MS/MS peak intensity prediction. Bioinformatics 29: 3199–3203

Degroeve S, Maddelein D & Martens L (2015) MS2PIP prediction server: Compute and visualize MS2 peak intensity predictions for CID and HCD fragmentation. Nucleic Acids Res. 43: W326–W330

Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ & Diella F (2011) Phospho.ELM: A database of phosphorylation sites-update 2011. *Nucleic Acids Res.* **39:**

Dou Y, Yao B & Zhang C (2014) PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids* **46:** 1459–1469

Gabriels R, Martens L & Degroeve S (2019) Updated MS2PIP web server delivers fast and accurate MS2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.*

Gao Y, Hao W, Gu J, Liu D, Fan C, Chen Z & Deng L (2016) PredPhos: An ensemble framework for structure-based prediction of phosphorylation sites. *J. Biol. Res.* **23:**

Gnad F, Ren S, Cox J, Olsen J V., Macek B, Oroshi M & Mann M (2007) PHOSIDA (phosphorylation site database): Management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **8:**

Goldenberg O, Erez E, Nimrod G & Ben-Tal N (2009) The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.* **37:**

Gupta R, Birch H, Rapacki K, Brunak S & Hansen JE (1999) O-GLYCBASE version 4.0: A revised database of O-glycosylated proteins. *Nucleic Acids Res.* **27:** 370–372

Holt LJ, Tuch BB, Villen J, Johnson AD, Gygi SP & Morgan DO (2009) Global analysis of cdk1 substrate phosphorylation sites provides insiqhts into evolution. *Science (80-.).* **325:** 1682–1686

Hornbeck P V., Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V & Sullivan M (2012) PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40:**

Hulstaert N, Sachsenberg T, Walzer M, Barsnes H, Martens L & Riverol YP (2019) ThermoRawFileParser: modular, scalable and cross-platform RAW file conversion. *BioRxiv* doi: https://doi.org/10.1101/622852 [PREPRINT]

Humphrey SJ, James DE & Mann M (2015) Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **26:** 676–687

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z & Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32:** 1037–1049

Jin J, Pawson T (2012) Modular evolution of phosphorylation-based signalling systems. *Philos. Trans. R. Soc. Lon. B Biol. Sci.* **367:** 2540– 2555

Johnson C, Crowther S, Stafford MJ, Campbell DG, Toth R & MacKintosh C (2010) Bioinformatic and experimental survey of 14-3-3-binding sites. *Biochem. J.* **427:** 69–78

Johnson LN (1992) Glycogen phosphorylase: control by phosphorylation and allosteric effectors. *FASEB J.* **6:** 2274–2282

Johnson LN (2009) The regulation of protein phosphorylation. *Biochem. Soc. Trans.* **37:** 627– 641

Kabsch W & Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, et al (2009) Human Protein Reference Database - 2009 update. *Nucleic Acids Res.* **37:**

Krissinel E & Henrick K (2007) Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372:** 774–797

Landry CR, Levy ED & Michnick SW (2009) Weak functional constraints on phosphoproteomes. *Trends Genet.* **25:** 193–197

Li J, Bigelow DJ & Squier TC (2003) Phosphorylation by cAMP-dependent protein kinase modulates the structural coupling between the transmembrane and cytosolic domains of phospholamban. *Biochemistry* **42:** 10674–10682

Li T, Li F & Zhang X (2008) Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins Struct. Funct. Genet.* **70:** 404–414

Liang X & Van Doren SR (2008) Mechanistic insights into phosphoprotein-binding FHA domains. *Acc. Chem. Res.* **41:** 991–999

Lim YP (2005) Mining the tumor phosphoproteome for cancer markers. *Clin. Cancer Res.* **11:** 3163–3169

Lu CT, Huang KY, Su MG, Lee TY, Bretaña NA, Chang WC, Chen YJ, Chen YJ & Huang H Da (2013) DbPTM 3.0: An informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.* **41:**

Martens L, Hermjakob H, Jones P, Adamsk M, Taylor C, States D, Gevaert K, Vandekerckhove J & Apweiler R (2005) PRIDE: The proteomics identifications database. *Proteomics* **5:** 3537–3545

Martens L & Vizcaíno JA (2017) A Golden Age for Working with Public Proteomics Data. *Trends Biochem. Sci.* **42:** 333–341

Nguyen Ba AN & Moses AM (2010) Evolution of characterized phosphorylation sites in budding yeast. *Mol. Biol. Evol.* **27:** 2027–2037

Nishi H, Hashimoto K & Panchenko AR (2011) Phosphorylation in protein-protein binding: Effect on stability and function. *Structure* **19:** 1807–1815

Peck SC (2006) Analysis of protein phosphorylation: Methods and strategies for studying kinases and substrates. *Plant J.* **45:** 512–522

Peters JS, Calder B, Gonnelli G, Degroeve S, Rajaonarifara E, Mulder N, Soares NC, Martens L & Blackburn JM (2016) Identification of quantitative proteomic differences between Mycobacterium tuberculosis lineages with altered virulence. *Front. Microbiol.* **7:**

Piovesan D, Walsh I, Minervini G & Tosatto SCE (2017) FELLS: Fast estimator of latent local structure. *Bioinformatics* **33:** 1889–1891

Raimondi D, Orlando G, Pancsa R, Khan T & Vranken WF (2017) Exploring the Sequence-based Prediction of Folding Initiation Sites in Proteins. *Sci. Rep.* **7:**

Rose AS & Hildebrand PW (2015) NGL Viewer: A web application for molecular visualization. *Nucleic Acids Res.* **43:** W576–W579

Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlić A, Quesada M, Quinn GB, Ramos AG, Westbrook JD, Young J, Zardecki C, Berman HM & Bourne PE (2013) The RCSB Protein Data Bank: New resources for research and education. *Nucleic Acids Res.* **41:**

Ryu GM, Song P, Kim KW, Oh KS, Park KJ & Kim JH (2009) Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Res.* **37:** 1297–1307

Santos AL & Lindner AB (2017) Protein Posttranslational Modifications: Roles in Aging and Age-Related Disease. *Oxid. Med. Cell. Longev.* **2017:** 1–19

Scheinin M, Koulu M, Karhuvaara S & Zimmer RH (1990) Evidence that the reversible MAO-A inhibitor moclobemide increases prolactin secretion by a serotonergic mechanism in healthy male volunteers. *Life Sci.* **47:** 1491–1499

Shi Y (2009) Serine/Threonine Phosphatases: Mechanism through Structure. *Cell* **139:** 468–484

Silva ASC, Bouwmeester R, Martens L & Degroeve S (2019) Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics*

Ubersax JA & Ferrell JE (2007) Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **8:** 530–541

Ullah S, Lin S, Xu Y, Deng W, Ma L, Zhang Y, Liu Z & Xue Y (2016) DbPAF: An integrative database of protein phosphorylation in animals and fungi. *Sci. Rep.* **6:**

Vandermarliere E & Martens L (2013) Protein structure as a means to triage proposed PTM sites. *Proteomics* **13:** 1028–1035

Velankar S, Dana JM, Jacobsen J, Van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ & Kleywegt GJ (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41:**

Verheggen K, Volders PJ, Mestdagh P, Menschaert G, Van Damme P, Gevaert K, Martens L & Vandesompele J (2017) Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products. *J. Proteome Res.* **16:** 2508–2515

Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz P-A, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus H-J, Albar JP, Martinez-Bartolomé S, et al (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32:** 223–226 Available at: http://www.nature.com/articles/nbt.2839

Xiao Q, Miao B, Bi J, Wang Z & Li Y (2016) Prioritizing functional phosphorylation sites based on multiple feature integration. *Sci. Rep.* **6:**

Xin F & Radivojac P (2012) Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics* **28:** 2905–2913

Zanzoni A, Carbajo D, Diella F, Gherardini PF, Tramontano A, Helmer-Citterich M & Via A (2011) Phospho3D 2.0: An enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Res.* **39:**

Zhang Y, Xu T, Shan B, Hart J, Aslanian A, Han X, Zong N, Li H, Choi H, Wang D, Acharya L, Du L, Vogt PK, Ping P & Yates JR (2015) ProteinInferencer: Confident protein identification and multiple experiment comparison for large scale proteomics projects. *J. Proteomics* **129:** 25–32

Tables

Table 1. Data sources/tools and derived data integrated into Scop3P

Data obtained	Database	Data statistics
P-sites and evidences Experimental Combined Similarity Human protein variants Variants on P-sites	**UniProtKB/Swiss-Prot**	40034 P-sites in 9775 proteins 3231 26785 10018 78027 single amino acid variants mapped onto 12688 proteins 380
Experimental P-sites from re-analysis	**PRIDE**	31606 (unique 18724) from 14 projects
Protein structures X-ray NMR EM Neutron other	**PDB**	14665 structures from 1963 proteins 14094 378 175 7 11
Solvent accessibility	**PDBePISA**	52198 P-sites with structural information
Secondary structural propensity (experimental)	**DSSP**	52198 P-sites with structures
Residue conservation scores	**Consurf-DB**	52198 P-sites with structures (contains missing values)
Backbone dynamics, disorderd propensity, early folding predictions	**DynaMine** **DisOmine** **EfoldMine**	All Swiss-Prot Human proteins (release-2018_02)

Figures

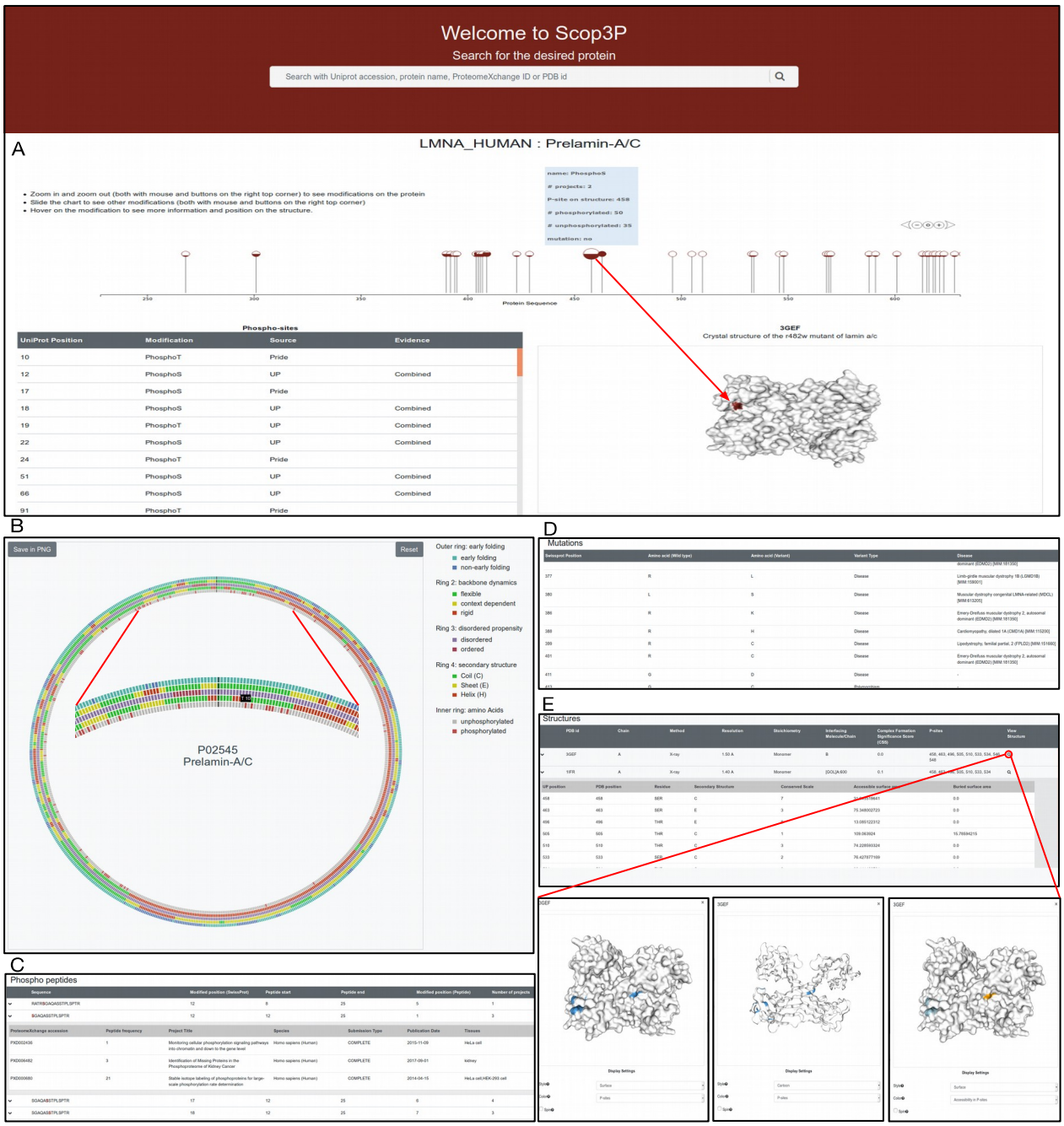

Figure 1. Scop3P web-interface reveals different context information at different levels for all phosphosites in a phosphoprotein

For a given phosphoprotein, Scop3P provides a variety of information in tables and interactive graphs.

A Quick overview of all P-sites mapped onto the protein sequence as ball and stick. The color fill in the ball represents the number of different peptides identified for that particular site across PRIDE projects. This information is also presented as a table below. A tooltip over the ball will display additional information, while a quick preview panel on the right shows where that P-site is located on the protein structure.
B All predicted biophysical properties and secondary structures are presented in an interactive circular plot with different concentric rings. The graph can be zoomed to focus on a region of interest (inner panel).
C Identified phospho peptides, source projects, and frequency information are presented in tabular form. Details about a project can be viewed by clicking the project ID (inner panel).
D Mutation table provides information on amino acid variants, type of polymorphism, and associated diseases (if any).
E Information related to structures, their secondary structure information, and P-sites mapped onto the particular structure are displayed in an overview table. Clicking the view structure icon will launch the structure viewer and shows P-sites mapped onto that particular structure. P-sites can be colored based on different properties.

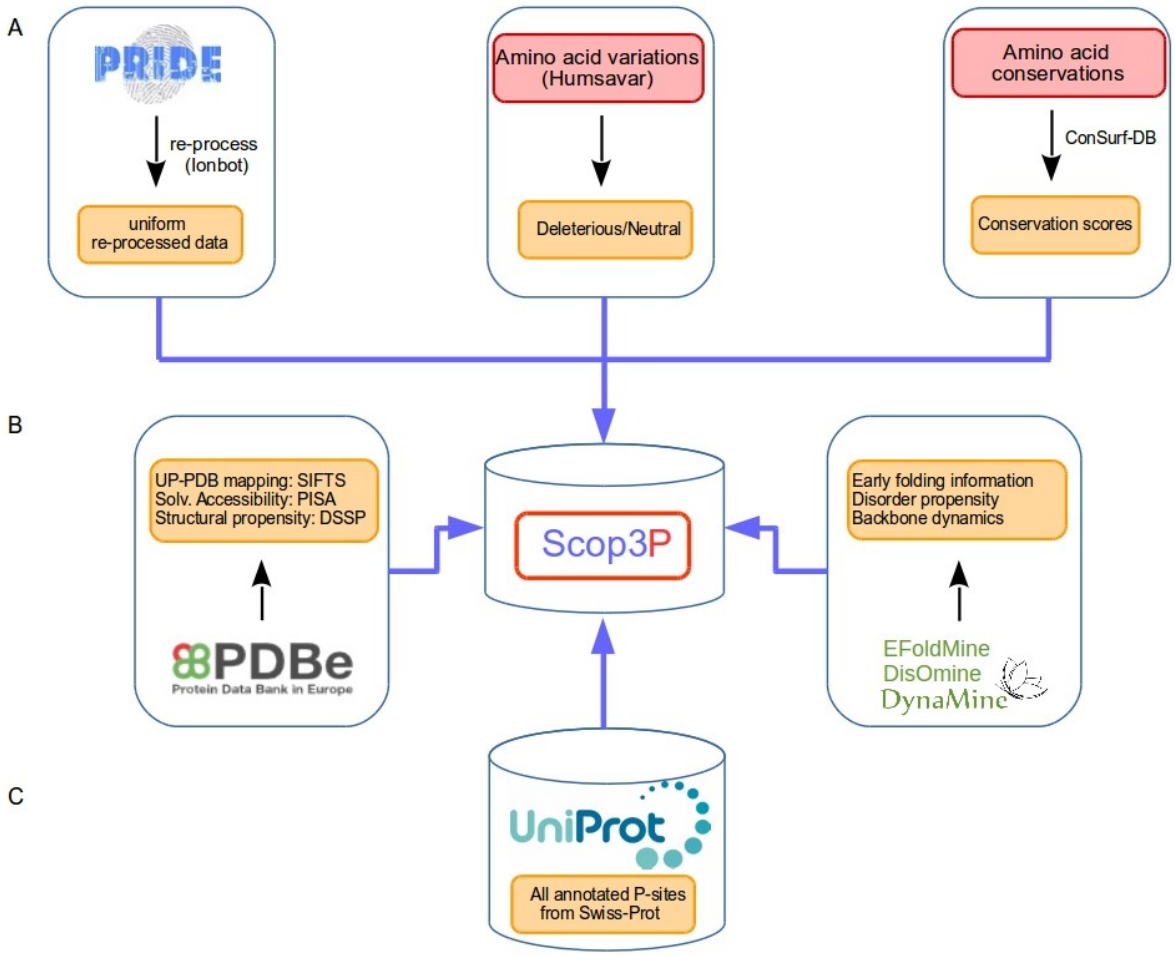

Figure 2. Scop3P data collection and integration flow

A Sequence level information like experimental phosphosites (P-sites) obtained from re-processed public proteomics data from PRIDE (14 projects), single amino acid variations and the associated disease and polymorphism details from Humsavar dataset and the amino acid conservation of P-sites from ConSurf-DB were integrated onto the amino acid sequence of proteins.

B,C All available P-sites from Swiss-Prot along with other structural information (secondary structures from DSSP, solvent accessibility from PISA) for all P-sites (Swiss-Prot+PRIDE) and predicted residue level biophysical properties of phosphoproteins (disorder and folding Propensities and backbone dynamics) were integrated on amino acid sequences. Sequence to structure mapping was done using SIFTS mapping.