Assembling a phosphoproteomic knowledge base using ProtMapper to normalize phosphosite information from databases and text mining

John A. Bachman[1], Benjamin M. Gyori[1], Peter K. Sorger[1]

1 Laboratory of Systems Pharmacology, Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115

Abstract

A major challenge in analyzing large phosphoproteomic datasets is that information on phosphorylating kinases and other upstream regulators is limited to a small fraction of phosphosites. One approach to addressing this problem is to aggregate and normalize information from all available information sources, including both curated databases and large-scale text mining. However, when we attempted to aggregate information on post-translational modifications (PTMs) from six databases and three text mining systems, we found that a substantial proportion of phosphosites were positioned on non-canonical residue positions. These errors were attributable to the use of residue numbers from non-canonical isoforms, mouse or rat proteins, post-translationally processed proteins and also from errors in curation and text mining. Published mass spectrometry datasets from large-scale efforts such as the Clinical Proteomic Tumor Analysis Consortium (CPTAC) also localize many PTMs to non-canonical sequences, precluding their accurate annotation. To address these problems, we developed ProtMapper, an open-source Python tool that automatically normalizes site positions to human protein reference sequences using data from PhosphoSitePlus and Uniprot. ProtMapper identifies valid reference positions with high precision and reasonable recall, making it possible to filter out machine reading errors from text mining and thereby assemble a corpus of 29,400 regulatory annotations for 13,668 sites, a 2.8-fold increase over PhosphoSitePlus, the current gold standard. To our knowledge this corpus represents the most comprehensive source of literature-derived information about phosphosite regulation currently available and its assembly illustrates the importance of sequence normalization. Combining the expanded corpus of annotations with normalization of CPTAC data nearly doubled the number of CPTAC annotated sites and the mean number of annotations per site. ProtMapper is available under an open source BSD 2-clause license at `https://github.com/indralab/protmapper`, and the corpus of phosphosite annotations is available as Supplementary Data with this paper under a CC-BY-NC-SA license. All results from the paper are reproducible from code available at `https://github.com/johnbachman/protmapper_paper`.

Author Summary

Phosphorylation is a type of chemical modification that can affect the activity, interactions, or cellular location of proteins. Experimentally measured patterns of protein phosphorylation can be used to infer the mechanisms of cell behavior and disease, but this type of analysis depends on the availability of functional information about the regulation and effects of individual phosphorylation sites. In this study we show that inconsistent descriptions of the physical locations of phosphorylation sites on proteins present a barrier to the functional analysis of phosphorylation data. These inconsistencies are found in both pathway databases and text mining results and often come from the underlying scientific publications. We describe a method to normalize phosphosite locations to standard human

protein sequences and use this method to robustly aggregate information from many sources. The result is a large body of functional annotations that increases the proportion of phosphosites with known regulators in two large experimental surveys of phosphorylation in cancer.

Introduction

Advances in protein mass spectrometry have made it possible to obtain precise information on large numbers of protein post-translational modifications (PTMs). These modifications, phosphorylation in particular, play a role in cell fate decisions and information processing under a wide variety of conditions. Efforts are currently underway to collect 'omic data on human proteins and their modifications in large numbers of tumors to advance the diagnosis and treatment of disease. The Clinical Proteomic Tumor Analysis Consortium (CPTAC; https://proteomics.cancer.gov/programs/cptac) is typical of these efforts [1,2]. CPTAC is a national (now international) effort to accelerate the understanding of the molecular basis of cancer through the application of large-scale proteome and genome analysis.

Generating mechanistic insight from this type of data requires functional annotations, ideally the identity of the enzymes mediating specific PTMs (e.g., kinases that phosphorylate a specific site) as well as the effects of PTMs on protein function. This information that is typically gleaned from "low throughput" functional studies in the literature. Among PTMs, phosphorylation has long been of interest for its role in cellular dynamics and disease [3]. Unfortunately, the proportion of experimentally observed phosphosites that have one or more functional annotations is very low, estimated at around 3% for regulatory annotations and below 3% for annotation of functional effects [4]. New approaches are needed not only to generate new information about the function and regulation of PTMs, but also to make the most effective use of existing information in literature and databases.

To enable functional analysis of proteomic data on PTMs, measured sites are typically referenced to annotated sites and, when available, to information on the enzymes that add and remove that the PTMs. Such information is currently available from databases such as PhosphoSitePlus [5], SIGNOR [6] and Reactome [7]. These databases were assembled by manual curation but automated text mining has also been used to extract information on PTMs from the literature [8,9]. Ideally, functional analysis would involve the use of information aggregated from as many of these databases and text mining tools as possible. However, we show in this paper that inconsistencies in the way site positions for PTMs are recorded make information aggregation difficult. In many cases, sites of phosphorylation extracted by human curators or by text mining algorithms cannot be matched to protein reference sequences. Inconsistencies can be traced to the use of site positions from non-human species, non-canonical isoforms, processed forms of the protein sequence, and curation or text mining errors. These inconsistencies result in mismatched sites when attempts are made to link mass spectrometry data to regulatory information.

This problem is compounded by the fact that processing pipelines for mass spectrometry data often map peptide sequences present in multiple isoforms of the same protein to the database identifiers for a non-canonical isoform (despite the fact that the peptide in question is found in *both* the canonical sequence and the isoform). As a result, many PTMs in mass spectrometry datasets do not match known site positions in databases or literature. Many known functional annotations are therefore missed, exacerbating the already poor coverage of PTM annotations and potentially affecting the biological interpretation of phosphoproteomic experiments.

Here we present two resources: an open source software tool, ProtMapper, that resolves inconsistencies among PTM annotations and experimentally identified sites; and a large corpus of regulatory phosphosite annotations aggregated from six databases and three text mining tools using this method. Site normalization via ProtMapper resolves inconsistencies

among these sources by mapping non-reference sites to known phosphorylation sites in the Uniprot human reference sequence whenever possible. The goal of the approach is both to find annotations for previously undescribed PTM positions and to correctly assemble annotations for known positions (thereby improving our understanding of the underlying biology). The final assembled corpus contains 29,400 regulatory annotations for 13,668 distinct sites, a 2.8-fold increase over PhosphoSitePlus which is currently the most comprehensive and accurate source of information on protein phosphorylation. By way of a biological application, we show that the use of data aggregated and normalized by ProtMapper nearly doubles the number of annotated sites in CPTAC Breast Cancer and Ovarian Cancer datasets, with roughly one-third of the increase attributable to the mapping of site annotations and experimental peptides to the correct reference sequences.

Results

Pathway databases and literature contain annotations of human PTMs that do not match reference sequences

To construct a corpus of phosphosite annotations we aggregated information from curated databases and available text mining systems by processing them into a standard format using the Integrated Network and Dynamical Reasoning Assembler (INDRA) [10] (Figure 1A). Databases included PhosphoSitePlus [5], SIGNOR [6], HPRD [11], NCI-PID [12], Reactome [7], and the BEL Large Corpus (`http://www.openbel.org`). Text mining was performed using multiple systems having complementary strengths, including REACH [9], Sparser [13], and RLIMS-P [14]. REACH and Sparser were run on a text corpus that included both abstracts and full-text articles; RLIMS-P results were obtained from the iTextMine service [15] (Methods). Each information source was mined for phosphorylation reactions that included a residue and position on a target substrate. Available information on upstream regulators (e.g. the relevant kinase) and downstream effects was also extracted. Information on the functional consequences of PTMs was drawn primarily from pathway databases in the form of a precondition for a protein to participate in a downstream reaction (e.g., MAPK1 *phosphorylated at T185 and Y187* phosphorylates RPS6KA1).

Databases such as Uniprot [16] and NCBI RefSeq Protein [17] catalog the sequences for all protein isoforms arising from alternative splicing of a gene product and they also specify a reference or "canonical" isoform. The designation is based on multiple criteria such as length, prevalence, similarity to orthologues, etc. (see `https://www.uniprot.org/help/canonical_and_isoforms`). In assembling a corpus of PTM annotations, we observed frequent mention of PTMs at sites that did not match the corresponding reference sequence, i.e., the phosphorylatable residue was not present at that position. For example, the positions of the T and Y residues in the T-X-Y activation motif of human MAPK1 (mitogen-activated protein kinase 1, also known as ERK2) are variously recorded as lying at amino acid positions 183/185, 184/186, and 185/187. The TXY motif is actually found between residues 185-187 in the reference sequences listed by the Uniprot [16] and NCBI RefSeq [17] databases (Figure 1B). The resulting erroneous assignment of a single set of PTMs to multiple distinct positions prevents database annotations and literature-based evidence on a phosphorylation event essential for cell proliferation from being accurately aggregated and used (Figure 1A).

To measure the extent of the mis-annotation problem, we extracted phosphorylation site data from multiple sources and compared PTM positions to Uniprot reference sequences [16]; specifically, we asked whether the residue reported to be phosphorylated actually exists at that location. We found that the total number of unique site annotations, and the fraction of annotations that had valid residue positions in the reference sequence, varied substantially between sources (Figures 1C, 1D; Table 1). Among databases, PhosphoSitePlus had by far

the largest number of unique site annotations (16,415) (Figure 1C), and the second-highest proportion of annotations associated with valid sites (99.2%) with matching residues in the reference sequence (Figure 1D, Table 1). This likely reflects the fact that PhosphoSitePlus is actively maintained by a team of curators. Text mining contributed a large number of site annotations, with both REACH and Sparser independently extracting over 12,000 unique site annotations, more than any database except PhosphoSitePlus (1C; Table 1) Over 40% of these annotations were associated with sites that did not match the reference sequence but it was not clear *a priori* whether these represented machine reading errors or legitimate and simply non-canonical protein sequence positions (Figure 1D, Table 1).

Site inconsistencies can be traced to the original literature as well as curation and text mining errors

By reviewing extracted site information we found that many non-canonical site positions were attributable to inconsistency in the numbering of site positions in the original source literature. These non-canonical positions were then propagated by human curators and text mining systems when PTM annotations were generated (Table 2). Four common causes of non-canonical positions in primary research papers arose from the use of residue numbering corresponding to: (i) mouse or rat proteins when describing human PTMs, (ii) non-canonical protein isoforms (generated by alternative splicing), (iii) other members of a multi-gene family, and (iv) proteins processed post-translationally as opposed to the primary protein sequence (Table 2).

We also identified examples of curator and machine-reading errors. Curator errors included misannotation of a serine site as a threonine, or the incorrect annotation of a site from a closely related protein (e.g., curators of NCI-PID annotated JAK2 kinase as having a phosphosite at Y485, whereas Y485 is actually a phosphosite on the erythropoietin receptor precursor - EPOR - a protein to which JAK2 binds; Table 2).

Text mining systems made a variety of errors including incorrect identification of gene and protein names (incorrect grounding) which then led to assignment of PTM sites to the wrong protein. For example, in the sentence "*p120's phosphorylation at S268 was found to contribute to cellular foci formation*" [18], the ambiguous label "p120" was misidentified as gene HNRNPU, whereas it is clear from the context of the paper that the authors were referring to the CTNND1 gene, which also has "p120" as a synonym. In other cases, machine readers linked sites to the wrong protein in a sentence, for example in the phrase "*phosphorylation of IRS-1 and subsequent activation Akt at Thr308*" (*sic*) [19], the site T308 was incorrectly associated with IRS1 rather than AKT. A third type of error resulted from the mis-identification of site-like text, for example the extraction of a (non-existent) phosphoserine site TP53 S1 from the text "*...as determined by p53 stabilization and phosphorylation (see Figure S3 and Text S1)*" [20] (Table 2). While these sources of site inconsistencies are diverse they all interfere with the generation of phosphoproteome networks and the analysis of mass spectrometry data (Figure 1A).

Normalizing phosphosite information by mapping sites to human reference sequences

To unify information from curated databases, text mining, and experimental data we sought to implement a systematic method that addresses the inconsistencies in PTM positions described above by correctly normalizing sites to positions in human reference sequences (Figure 2A). The method was implemented in a new open-source Python tool, ProtMapper, whose source code is available at `https://github.com/indralab/protmapper`. ProtMapper also includes other tools useful for analysis of proteomic data, including a wrapper around PhosphoSitePlus site information, mappings between Uniprot and NCBI

RefSeq Protein identifiers and gene symbols, and sequence lookup for both NCBI RefSeq Protein and Uniprot (Figure 2B).

For site information obtained from curated databases and text mining, the goal of site normalization is to determine the most likely reference position (if one exists) for a reported position that does not match the reference sequence. For 134 frequently occurring non-canonical site positions we manually curated mappings to canonical sequences. When normalizing site positions ProtMapper checks this resource first; if a mapping is not found in this set, it determines whether a non-canonical position is a *known* site of phosphorylation on a closely related protein sequence (Figure 2B). The search of related sequences includes human protein isoforms, mouse and rat proteins, and proteins processed by cleavage of a signal peptide or initiator methionine. The alternative possibilities are tested sequentially in a fall-through fashion with the most conservative options considered first: for example, if a non-canonical site position can be associated with a known phosphorylation site on an alternative human isoform, the corresponding reference position is returned and positions on mouse or rat orthologs are not considered (2B). A site is considered to be a known phosphorylation site if it is listed as such in PhosphoSitePlus, which aggregates mass spectrometry datasets and extensive literature curation ([5]). Mappings are determined by whether two sites for different sequences are included in the same PhosphoSitePlus *site group*. A site group lists corresponding positions among homologous sequences (human and non-human proteins and isoforms), allowing positions of known phosphorylation sites to be mapped between sequences without additional sequence alignment. For example, the (invalid) site human pBAD-S112 can be mapped correctly to human pBAD-S75 because murine pBAD-S112 and human pBAD-S75 are in the same site group (Figure 2C). Note that PhosphoSitePlus site groups do not include corresponding gene positions among paralogs, so mappings between gene family members are not automatically handled by this procedure (e.g., mapping the activating site T308 on the kinase AKT1 to the corresponding T309 on AKT2).

When ProtMapper was applied to the corpus of PTM annotations we assembled from text mining and databases we were able to identify reference human positions for many previously invalid site annotations (Figures 2D and E, Table 1). Among the databases examined, NCI-PID had the largest fraction of initially erroneous PTM site annotations that could be associated with a canonical residue (75% "mappable" site annotations). This likely reflects the fact that the 134 site mappings that we manually curated during the course of ProtMapper development were drawn primarily from this database. In absolute numbers, text mining systems were by far the largest source of both non-canonical, and subsequently mappable, site annotations. For 11,961 invalid site annotations cumulatively extracted by REACH, Sparser, and RLIMS-P, ProtMapper identified reference sequence positions for a total of 3,474 (29%; Table 1). The percentage of mapped sites was very similar among the individual systems, with 27%, 31%, and 29% of annotations mapped for REACH, Sparser, and RLIMS-P respectively. The use of ProtMapper increased the proportion of text mined site annotations having reference sequence positions by 14-21%, depending on the text mining system (valid and mapped annotations vs. valid annotations alone).

Accuracy of automatically inferred mappings for literature-derived sites

The approach that ProtMapper uses to normalize sites has the potential both to miss legitimate mappings to reference positions (false negatives) and to erroneously associate reference positions with unknown sites (false positives). This is a particular problem for normalization of text-mined sites due to the high frequency of technical errors (Table 2. To determine the accuracy of ProtMapper when applied to sites mined from the literature, we generated a frequency-weighted random sample of 100 sites from the total pool of invalid

sites generated by machine reading. We then manually curated these sites and their mappings by reading the underlying papers. Manual curation could only be performed on a sample of sites because it was laborious, requiring careful inspection of the original source literature followed by an examination of the assignment to a canonical position. Criteria for curation of reading and mapping accuracy are described in Methods. The dataset containing the sites, ProtMapper mappings and assessment of correctness is also available at `https://github.com/johnbachman/protmapper_paper`.

Overall, our manual analysis of ProtMapper results yielded estimates of 95% precision and 74% recall, showing that ProtMapper is robust to Type I error (Table 3). Of the 100 invalid, text-mined PTM sites in our sample, 43 were attributable to text mining errors of the types described in Table 2. Mapping of such mis-extracted sites to a reference sequence would represent a false positive from a mapping perspective but only two such erroneous mappings were observed. For example, in the sentence "*phosphorylation of IRS-1 and subsequent activation Akt at Thr308*" (*sic*), the site T308 was incorrectly associated with IRS1 due to a reading error (Table 2; row 52 in curation dataset). There is no threonine at position 308 on IRS1, but there is a threonine at position 309 that is known to be phosphorylated based on PhosphoSitePlus. ProtMapper incorrectly associated "Thr308" with IRS1-T309 under the assumption that it is an off-by-one inconsistency arising from cleavage of the initiator methionine. Only one other reading error resulted in an erroneous mapping to a human reference sequence (row 62 in curation dataset), resulting in a site that did not correspond to the one described in the source text. In the remaining 41 cases, sites incorrectly recovered from text by NLP were not mapped to a canonical sequence by ProtMapper (Table 3).

The remaining 57 of 100 sites were correctly extracted from the source literature by the NLP systems. Among these, 42 sites (74%) were correctly mapped to a reference sequence (true positive mappings), and 15 (26%) were unmapped (false negative mappings) for reasons described in greater detail below. We did not encounter any instances of false positive mappings in which a site that had been extracted correctly by NLP was mapped to an incorrect reference position based on the context of the original article. These data show that ProtMapper achieves good performance from the perspective of both Type I and Type II error.

To determine which mapping rules were used most frequently by ProtMapper and evaluate their accuracy (Figure 2B), we grouped curation results for the 100 sites described above according to the type of mapping (Table 4). Although sample sizes were too low to make statistically rigorous conclusions, we found that the two Type I errors described above resulted from use of the "off-by-one" rule used to correct for residue number inconsistencies arising from initiator methionine cleavage (Figure 2B. Inactivating the "off-by-one" rule increased precision to 1.0 and reduced recall to 0.68. We therefore made use of this rule optional in ProtMapper allowing the precision/recall tradeoff to be tuned for different use cases.

To investigate the causes of false negative mappings (correctly text-mined sites for which ProtMapper identified no reference sequence positions), we manually examined the sentences in which the sites were mentioned. In some cases site positions referred to orthologous proteins in organisms for which PhosphoSitePlus did not contain phosphorylation information (for example site S381 on *Xenopus* protein Msi2 (row 61 in curation dataset). In other cases, the invalid site was due to an error in the primary article itself, for example a reference to tyrosine 828 as "T828" (rather than Y828; row 77). In these and all other cases examined, it was possible to infer a site corresponding to the human reference sequence by manual curation of the source text and related information but we did not find systematic errors that would serve as the basis for further automated mapping functions. We determined, for example, that allowing S, T and Y to be freely substituted for each other increased the rate of Type I error.

Text mining tools identify many uncurated regulators of phosphorylation

We obtained a combined corpus of 29,400 regulatory annotations of 13,668 human sites after normalizing information from all sources with ProtMapper (Figure 3). The full corpus of regulatory annotations along with the underlying sentences for text mined annotations are available as Supplementary Data.

What do text mining and site normalization contribute to *presently uncurated* information about kinases, phosphosites and their regulation? Overall, we found that the combined corpus contains 2.8 times as many regulatory annotations (29,400 vs. 10,366) and 1.9 times as many human sites (13,668 vs. 7,044) as PhosphoSitePlus alone. To investigate the specific contributions of text mining and the various databases, we measured the overlap of regulator-site pairs between (i) PhosphoSitePlus, the largest curated database, (ii) other widely used pathway databases (HPRD, Signor, NCI-PID, Reactome, and the BEL Large Corpus), and (iii) the aggregated output of the REACH, Sparser, and RLIMS-P text mining systems. Overlap was measured following site normalization by ProtMapper and sites that were either invalid in or unmappable to a reference sequence were excluded (thereby emphasizing precision over recall).

Text mining extracted one or more upstream regulators for a total of 8,949 sites, of which 5,311 (59%) had no regulatory annotations in PhosphoSitePlus, and 4,395 (49%) had no regulatory information in any of the curated databases (Figure 3A). Text mining also identified 15,850 unique regulator-site pairs of which 12,470 were absent from curated databases and PhosphoSitePlus (Figure 3B). Text mining extracts information not only about kinases that directly phosphorylate a target site, but also kinases that lie further upstream as well as non-kinase regulators (e.g., growth factors). REACH and Sparser also use FamPlex identifiers, a taxonomy of protein families and complexes for text mining, to extract and normalize phosphorylation events expressed in terms of kinase classes, e.g., "ERK", "AKT", "AMPK" [21]. For an equal comparison between text mining and databases, we therefore repeated the comparison by restricting the regulators to specific human kinase proteins and found that text mining still yielded a substantial body of new information, with 3,792 unique human kinase-site pairs reported by machine readers and not PhosphoSitePlus, and 3,118 that did not appear in any curated database (Fig 3C).

To characterize the contributions made by different reading systems to the final corpus of annotated sites, we measured the overlap among REACH, Sparser, and RLIMS-P for unique annotated sites, regulator-site pairs, and human kinase-site pairs (the same categories analyzed in Figure 3; Supplementary Figures S1A-C). Although there was overlap among readers, which is expected for different reading systems processing the same text corpora, each reading system contributed a substantial number of sites not found by other readers (Supplementary Figures S1A-C). Thus, it is most effective to use the combined output of multiple readers for automated PTM curation.

Many apparently isoform-specific phosphopeptides can be remapped to the reference protein sequence

In the process of using newly aggregated annotations to analyze mass spectrometry data, we identified an additional issue that results in fewer regulatory annotations for phosphopeptides/phosphosites. We found that existing phosphoproteomic datasets often assign phosphopeptides to Uniprot or RefSeq identifiers that correspond to non-canonical protein isoforms, even when the site is also present in the reference sequence for the canonical form. Because phosphosite *annotations* are most often indexed by the site positions in the reference sequence, a phosphosite in an experimental dataset that is associated with a non-canonical isoform will not be correctly associated with a site annotation. As a result many phosphopeptides grounded to isoform-specific identifiers will

be considered as having no annotations, when in fact there exist valid annotations associated with the reference sequence position for the site. As we show below, these irregularities affect approximately 25% of the measured phosphosites in a typical dataset.

A further complication involves the use of protein identifiers from different databases. For example, the breast and ovarian cancer phosphoproteomic datasets released by the NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) [1, 2] identify phosphosites using NCBI RefSeq Protein IDs [17] and corresponding HGNC gene symbols, with site positions based on the NCBI RefSeq Protein sequence. However, all sources of phosphosite annotations (with the exception of the Human Protein Reference Database (HPRD)) use Uniprot IDs and/or HGNC symbols to identify phosphorylated proteins. Sequences in NCBI RefSeq Protein and Uniprot do not match exactly, so changing identifiers from one database to the other can result in a change in sequence and hence in PTM position. To be linked accurately to database annotations, experimentally determined phosphosites identified with NCBI RefSeq Protein IDs and HGNC symbols must be mapped to Uniprot IDs, either via a RefSeq-Uniprot or HGNC-Uniprot mapping table.

To investigate the impact of these two issues we examined two CPTAC datasets, in which phosphosites are linked to NCBI RefSeq IDs and HGNC symbols but not Uniprot IDs. We found that the use of isoform-specific protein IDs and problems with converting IDs between NCBI RefSeq Protein and Uniprot caused many experimental phosphosites to lack canonical (Uniprot) site positions, a necessary prerequisite for finding associated functional annotations. For example, when we used the HGNC symbols identified for the phosphopeptides in the CPTAC breast cancer dataset to obtain reference protein sequences from Uniprot, 19,977, or 22.5%, of the 88,911 phosphosites in the dataset were invalid; that is, they were not present in the Uniprot canonical sequence at the position assigned to them by CPTAC (Table 5, row 3). When we used the NCBI RefSeq IDs in CPTAC to obtain Uniprot IDs via a RefSeq-Uniprot mapping table maintained by Uniprot, a similar fraction of sites (24.8%) were incorrectly mapped to isoform-specific sequences (row 10); an additional 6,895 sites (7.8%) had RefSeq IDs with no matching Uniprot ID (row 6). Results for the CPTAC ovarian cancer dataset were very similar. Thus, depending on the type of source identifiers used, 22-33% of sites in a CPTAC dataset are likely to be excluded from downstream annotations simply due to use of non-canonical site positions, lack of a Uniprot ID, or isoform-specific identifiers.

What fraction of phosphopeptides in CPTAC and similar datasets assigned to non-canonical isoforms actually correspond to a peptide sequence unique to that isoform as opposed to also being present in the reference sequence? To investigate this we implemented a method within ProtMapper (`ProtMapper.map_peptide_to_human_ref`) that relocalizes phosphopeptide sequences to positions within a reference sequence whenever possible. Remapping phosphosites from peptide *sequences* is distinct from, and much more straightforward than, remapping phosphosites identified only by a *residue* and *position*, as the phosphopeptide sequence is itself sufficient to identify a unique location in the target sequence (Figure 2A). We found that the vast majority of phosphopeptides with invalid reference sequence positions could be successfully reassigned to a reference sequence: of 19,977 invalid sites in the CPTAC breast cancer dataset (Table 5, row 3), 18,630, or 93% (row 4), were mappable to alternative positions in a reference sequence. Similarly, of the 22,044 sites with isoform-specific Uniprot IDs obtained via RefSeq-Uniprot mappings (row 10), 21,107 (96%) could be reassigned to a Uniprot reference sequence (row 11). In addition, 94% of sites (6,471 of 6,895; rows 7 and 6) with no Uniprot ID available in the RefSeq-Uniprot mapping table could be remapped to a Uniprot reference sequence by using the HGNC gene name reported by CPTAC. We obtained similar results for the CPTAC ovarian cancer dataset. Overall, after use of ProtMapper, 98% of sites in both CPTAC datasets were either valid in, or mappable to canonical Uniprot protein sequences (Table 5, row 5).

Annotation assembly and site normalization increases functional annotations of experimentally observed phosphosites

To determine the value for phosphoproteomic data analysis of correct site annotation and normalization to canonical sequences, we counted the number of annotated phosphosites in the CPTAC datasets under different analysis conditions (Table 6). Overall, we found that site normalization, and the robust incorporation of text mining output that it allows, substantially increased the proportion of annotated sites. Furthermore, while text mining primarily contributed *new* regulatory information about *previously curated* sites, it also covered a surprisingly high proportion (76%) of the sites annotated in PhosphoSitePlus alone.

Regardless of the information source used to obtain functional annotations, site normalization increased the number of annotated sites and annotations per site. Using all annotation sources, remapping of peptides in the CPTAC dataset to reference positions yielded an increase in annotated sites from 2,284 to 2,754, even without mapping site annotations to canonical positions (Table 6, row 8 vs. row 4). Site mapping for annotations further increased the number of annotated sites, reaching a maximum of 2,860 annotated sites, a 92% increase over the 1,540 sites obtained using PhosphoSitePlus with no mapping (Table 6, row 16 vs row 1). Site mapping had a particularly substantial effect when using only text-mined sources (1,439 vs. 1,228 annotated sites, row 15 vs. row 7), reflecting the large proportion of text mined information associated with non-canonical site positions.

As expected, adding annotations from text mining systems and curated databases other than PhosphoSitePlus increased the number of annotations over PhosphoSitePlus alone, with or without site normalization (Table 6, "All" vs. "PSP-only"). Text mining results contained nearly double the number of annotations per site compared to the aggregated databases (4.11 vs. 2.12, row 15 vs. row 14); the combination of text mining with databases yielded 3.37 annotations per site, a 59% increase over databases alone. Notably, the inclusion of text mining with databases led to a smaller proportional increase of only 14% in the overall number of annotated sites (from 2,508 to 2,860, row 16 vs. row 14). The fact that text mining systems added a proportionally higher number of annotations than sites shows that they largely contributed new regulatory information about sites that were already curated in databases. Despite this fact, using the three text mining systems *alone* (i.e., using no human-curated sources) yielded annotations for as much as 76% of the sites annotated using PhosphoSitePlus, and with a larger number of annotations per site (Table 6, row 15 vs. row 13). This suggests that machines are approaching human curators in their ability to accurately consolidate some types of information on PTMs.

As an illustrative example of how site normalization via ProtMapper adds information to the interpretation of phosphoproteomic data, we selected S560 on IRS2, which is found in the CPTAC ovarian cancer dataset. IRS2 S560 had no functional annotations in any database, but all three reading systems found multiple publications in which PLK1 is identified as a kinase that phosphorylates IRS2 at residue "S556." [22–24] For example, Chen et al. [22] described the potential regulatory significance of the phosphorylation event in human cell lines as follows: *"We show in the study that Plk1-dependent phosphorylation of IRS2-S556 inhibits mitotic exit, partially through reduced AKT activity."*. Although the human IRS2 reference sequence does not have a serine at S556, ProtMapper correctly identifies this as a site on *murine* IRS2 corresponding to S560 on human IRS2 (both human IRS2 S560 and murine IRS2 S556 are known to be phosphorylated in PhosphoSitePlus and are in the same site group). Although Chen et al. used the human HEK239T and HeLa cell lines for functional studies, their materials and methods section reveals that they raised an antibody against a recombinant fusion between the murine IRS2 protein and GST—and then used the antibody to assay human cell extracts [22]. Articles citing Chen et al.

subsequently referenced the non-human site position S556, for example, in the context of human pancreatic cancer cell lines [23, 24]. This example illustrates how a subtle inconsistency in a paper using both murine and human materials can generate an invalid residue assignment that is then propagated to subsequent publications. The error is not insignificant, since IRS2 carries signals from insulin and insulin-like growth factor to the PI3K/AKT signaling pathway, which plays a role in ovarian cancer [25]; both IRS2 [26] and PLK1 [27] have also been reported as relevant to ovarian cancer treatment or prognosis. This example additionally highlights that text mining tools provide the added benefit of linking structured information about phosphorylation to specific passages in the primary literature, allowing researchers to evaluate the provenance and context-specificity of individual PTM annotations relevant to a specific hypothesis or type of analysis.

Taken together, the evaluation of ProtMapper in the context of two CPTAC datasets demonstrates that site normalization for functional annotations and phosphopeptide sequences, in tandem with the integration of text mining tools and curated databases, can substantially increase the proportion of experimentally observed sites with available functional information.

Discussion

In this paper we described a method, implemented as Python software, for increasing the breadth and depth of functional information about human PTMs, with a focus on phosphorylation. Normalizing positions of PTMs to canonical reference sequences greatly facilitates assembly of site information from multiple databases and from text mining tools. We show that ProtMapper can be used to assemble a corpus of functional phosphosite information that is 3-fold larger than the current standard (PhosphoSitePlus) and show that use of the corpus nearly doubles the fraction of sites with known regulators in datasets generated by CPTAC. To our knowledge the corpus we have assembled represents the most comprehensive source of literature-derived information about phosphosite regulation currently available.

Our analysis of PTM information extracted from databases and mined from the literature reveals that inconsistencies in site numbering are common in both sources: database curators and machine reading systems are misled by inconsistent references to sites of PTMs in the literature. These errors appear to originate from historical bias in early functional studies involving a protein isoform that is no longer considered canonical, or to non-human species (particularly mouse or rat) if experimental materials from those species are used in the study (as illustrated above in the example of the use of an antibody against murine IRS2-S556 to study human IRS2-S560). Moreover, antibody manufacturers often list non-reference site positions on their product datasheets and these too are propagated in the literature. Resolving these inconsistencies is challenging and time-consuming for human curators. As a result, some databases, NCI-PID for example, contain a high proportion of erroneous annotations. By automating the process of site normalization ProtMapper has the potential to streamline the maintenance of phosphoproteomic and pathway databases (e.g., PhosphoSitePlus, SIGNOR, and Reactome) and thereby improve their scope and accuracy.

We find that text mining systems available in the public domain are able, in their current forms, to contribute significantly to available information on phosphosite regulation. In a comparison of three such systems, we found that each one extracted slightly different information from the same text corpus. They are therefore better used in combination than individually. In aggregate, the three systems are capable of processing a large corpus of available literature in a period of only a few days and can obtain information about phosphosites not found in any existing databases (Figure 3). These systems yielded 75% as many annotated sites as PhosphoSitePlus, the product of years of human curation effort (Table 6). Such systems could be used to comb literature for new phosphosite information as

it appears, reducing the burden on human curators. The contribution of text mining is certain to increase as methods improve and technical and legal barriers currently preventing access to full text articles are addressed.

We find that a substantial proportion (33-43% depending on the reader) of phosphosite annotations extracted by machine readers are invalid with respect to the reference sequence for the protein they identified in the text (Table 1). A majority of these non-canonical sites represent simple machine reading errors (Table 2). Without the ProtMapper, this information would simply be discarded from further analysis, but with the ProtMapper we found that roughly a third could be "rescued" by mapping them to canonical positions. In our manual evaluation of 100 text mined sites we found that ProtMapper is able to make these mappings with 95% precision, a figure that can be increased still further by disabling the off-by-one rule used to correct for cleavage of the initiator methionine. The low rate of Type I error shows that ProtMapper reliably identifies valid information from text mining that would otherwise be indistinguishable from reading errors, while introducing very few false positives. Automated site normalization therefore plays an essential role in text mined information about PTMs as it can not only identify canonical site positions but also serve as an effective filter for invalid extractions (Table 3).

Information extracted by text mining the primary literature for information on proteins that regulate phosphorylation sties does not necessarily differentiate between direct and indirect effects. Thus, regulatory annotations obtained from text mining include kinases responsible for phosphorylating a site and also receptors, ligands, and other proteins upstream of these kinases. If the goal of a study is to characterize the substrate specificity of kinases, this is a potential weakness, although one that can be mitigated by including only regulatory annotations involving kinases shown to physically interact with the substrate (by cross-referencing with physical interaction databases such as BioGRID [28], for example). For studies in which the goal is to map regulatory pathways, the direct linking of upstream regulators to specific downstream sites is generally a strength. Another differentiating feature is that text-mined information can include protein families and complexes as regulators of phosphorylation as this is how descriptions of phosphorylation often appear in literature. For example, in the sentence "ERK-dependent Serine 383 phosphorylation of Elk-1" [29]), ERK refers to both ERK2/MAPK1 and ERK1/MAPK3. We previously developed the FamPlex resource to unambiguously normalize information about protein families and complexes found in the literature [21]. The REACH and Sparser reading systems both make use of FamPlex and synonyms for named entity recognition and normalization, allowing these annotations to be aligned against gene-level members. While family-level regulatory information might seem less specific than data on specific kinases, it is often a better representation of functional information. For example, MAPK1 and MAPK3 are activated by the same biological ligands and inhibited by the same small molecules drugs, making them difficult to distinguish at the level of function.

One limitation of our approach as currently instantiated is that it uses human proteins as the point of normalization and is therefore not immediately useful for phosphoproteomic studies of distantly related species (e.g., *Drosophila* or yeast). This can be rectified in the future as information on these species becomes available since the ProtMapper approach is in principle species agnostic. It can also be extended to other types of PTMs such as ubiquitination, methylation and acetylation. A second limitation of ProtMapper is that its approach to mapping relies on prioritized set of predefined matching rules (Figure 2B) rather than a probabilistic approach that might, in principle, resolve conflicts between multiple possible mappings (e.g., an invalid site that could be due to erroneous use of a mouse residue number or, alternatively, initiator methionine cleavage, which would result in different mappings). Despite this, our evaluations show that the current implementation of ProtMapper already results in high precision mappings (Table 3).

This paper illustrates the need for new tools to effectively aggregate data on PTMs and

their regulators at proteome scale. Fortunately, relatively simple tools such as ProtMapper can have a substantial positive impact. However, fully resolving inconsistencies and ambiguities in functional descriptions of proteins will require the involvement not only of database curators but also authors, editors and antibody manufacturers. Adoption of standard names for genes and canonical residue numbers will improve reproduciblilty, reusability and machine readability. ProtMapper could be used, for example, to check all phosphoprotein sites in submitted papers prior to their publication. Finally, gaps remain in available software for key tasks in proteomics data analysis; the ProtMapper is aimed at addressing a handful of these gaps. Like genomics, the field will benefit as the ecosystem of open-source software continues to expand.

Materials and Methods

Availability

ProtMapper software is implemented in Python and is available under a BSD 2-clause open-source license from GitHub at `https://github.com/indralab/protmapper`. Documentation is hosted on ReadTheDocs at `https://protmapper.rtfd.io`. Code used to assemble the annotation corpus and generate the results in this paper is on GitHub at `https://github.com/indralab/protmapper_paper`. The assembled corpus of annotations is available under a CC-BY-NC-SA license (per the Share-alike requirements of constituent sources PhosphoSitePlus, RLIMS-P and SIGNOR) and is available as Supplementary Data with this paper.

Information Sources

- PhosphoSitePlus [5]. Downloaded from the PhosphoSitePlus website (`https://www.phosphosite.org`). Kinase-substrate annotations were obtained by processing the `Kinase_substrate.owl` BioPax file with the INDRA BioPax processor into INDRA Statements [10]. The file `Phosphorylation_site_dataset.tsv` was used by the ProtMapper for site mappings. License: CC BY-NC-SA 3.0. See also license information at `https://www.phosphosite.org/staticDownloads`.
- HPRD [11]. Obtained from `http://www.hprd.org/RELEASE9/HPRD_FLAT_FILES_041310.tar.gz` and processed to INDRA Statements by the INDRA HPRD Processor.
- SIGNOR [6]. Interactions obtained from `https://signor.uniroma2.it/` and processed to INDRA Statements by the INDRA SIGNOR Processor. License: CC BY-SA 4.0.
- BEL Large Corpus. Downloaded from `https://arty.scai.fraunhofer.de/artifactory/bel/knowledge/large_corpus/large_corpus-20170611.bel` and processed to INDRA Statements by the INDRA BEL processor.
- Reactome [7]. Obtained from Pathway Commons and processed with the INDRA BioPax processor. License: CC0.
- NCI-PID [12]. Obtained from Pathway Commons and processed with the INDRA BioPax processor.
- REACH [9]. Software obtained from `https://github.com/clulab/reach` and used to process a text corpus including MEDLINE abstracts and full-text articles from the PubMed Central Open Access Subset, the PubMed Central Author's Manuscript

Collection, and others obtained via the Elsevier text and data mining API (`https://dev.elsevier.com/`). REACH reading output processed with the INDRA REACH processor.

- Sparser [13]. Executable software image obtained from the Sparser developers and used to process the same text corpus as REACH. Output processed with the INDRA Sparser processor. The executable Sparser image is available on the INDRA Docker image, which can be obtained from DockerHub at `https://hub.docker.com/r/labsyspharm/indra`.
- RLIMS-P [14]. Text mining results for PubMed Central full-text articles and MEDLINE abstracts obtained via download from the iTextMine service [15] at `https://hershey.dbi.udel.edu/textmining/export/`, and processed to INDRA Statements using the INDRA RLIMS-P Processor. License: CC BY-NC-SA 4.0.
- Uniprot [16]. Protein identifiers, annotations, sequences and mappings to RefSeq identifiers were obtained from the Uniprot website, `https://www.uniprot.org`. Specific download procedures are implemented in `protmapper.resources`.
- RefSeq [17]. Protein sequences obtained from `ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/protein/protein.fa.gz`.

Mass Spectrometry Data Sources

CPTAC phosphoproteomic data were downloaded from `https://cptc-xfer.uis.georgetown.edu/publicData/`. Breast cancer data was downloaded from:

`https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/CPTAC_Breast_Cancer_S039/CPTAC_BCProspective_BI_Phosphoproteome_CDAP_Protein_Report.r1/CPTAC2_Breast_Prospective_Collection_BI_Phosphoproteome.phosphosite.tmt10.tsv`

Ovarian cancer data was downloaded from:

`https://cptc-xfer.uis.georgetown.edu/publicData/Phase_II_Data/CPTAC_Ovarian_Cancer_S038/CPTAC_OVprospective_PNNL_Phosphoproteome_CDAP_Protein_Report.r1/CPTAC2_Ovarian_Prospective_Collection_PNNL_Phosphoproteome.phosphosite.tmt10.tsv`

Manual site curation

Inconsistent site positions in the NCI-PID database were manually examined and matched to site positions in the protein reference sequence. Internet searches of genes and their inconsistent site positions frequently identified both the source of the error (e.g., incorrect site position listed by antibody vendor) and the corresponding sites in the reference sequence. Incorrect sites were prioritized for curation by their frequency of appearance in Biopax reactions. The resulting table listing incorrect sites, their reference positions, and a brief description of the source of the inconsistency is contained in the GitHub repository for the ProtMapper at: `https://github.com/indralab/protmapper/blob/master/protmapper/curated_site_map.csv`.

Curating the accuracy of site normalization

Manual curation of site mappings for machine reading-derived sites were based on two criteria: 1) whether the site extracted by the reader was supported by the corresponding sentence in the source publication, and 2) whether the reference site returned by the

ProtMapper was the correct one based on the context of the original sentence and publication. The criteria for scoring each site are summarized in Table 7. The sample of sites with associated curations is available from GitHub at `https://github.com/johnbachman/protmapper_paper/raw/master/output/literature_site_mapping_analysis.xlsx`.

Acknowledgements

Funding for this work was provided by the Defense Advanced Research Projects Agency under awards W911NF-14-1-0397 (Big Mechanism) and W911NF018-1-0124 (Automated Scientific Discovery Framework) and by NIH grant U24-DK116204 . Data used in this publication were generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH).

References

1. Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J.T., Gatza, M.L., Wilkerson, M., Perou, C.M., Yellapantula, V., Huang, K.L., Lin, C., McLellan, M.D., Yan, P., Davies, S.R., Townsend, R.R., Skates, S.J., Wang, J., Zhang, B., Kinsinger, C.R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A.G., Fenyo, D., Ellis, M.J., Carr, S.A., Carr, S.A., Gillette, M.A., Clauser, K.R., Kuhn, E., Mani, D.R., Mertins, P., Ketchum, K.A., Thangudu, R.R., Cai, S., Oberti, M., Paulovich, A.G., Whiteaker, J.R., Wang, X., Lin, C., Ping, Y., Edwards, N.J., Madhavan, S., McGarvey, P.B., Wang, P., Petralia, F., Tu, Z., Chan, D., Pandey, A., Shih, L.M., Zhang, H., Zhang, Z., Thomas, S., Zhu, H., Whiteley, G.A., Skates, S.J., White, F.M., Levine, D.A., Boja, E.S., Kinsinger, C.R., Hiltke, T., Mesri, M., Rivers, R.C., Rodriguez, H., Shaw, K.M., Stein, S.E., Fenyo, D., Liu, T., McDermott, J.E., Payne, S.H., Rodland, K.D., Smith, R.D., Rudnick, P., Snyder, M., Zhao, Y., Chen, X., Ransohoff, D.F., Hoofnagle, A.N., Liebler, D.C., Sanders, M.E., Shi, Z., Slebos, R.J., Tabb, D.L., Zhang, B., Zimmerman, L.J., Wang, Y., Li, S., Davies, S.R., Ding, L., Maher, C., Townsend, R., Ellis, M.J., Lei, J.T., Luo, J.: Proteogenomics connects somatic mutations to signalling in breast cancer. Nature **534**(7605), 55–62 (2016)

2. Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.Y., Petyuk, V.A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, J., Shah, P., Cha, S.W., Aiyetan, P., Woo, S., Tian, Y., Gritsenko, M.A., Clauss, T.R., Choi, C., Monroe, M.E., Thomas, S., Nie, S., Wu, C., Moore, R.J., Yu, K.H., Tabb, D.L., Fenyo, D., Bafna, V., Wang, Y., Rodriguez, H., Boja, E.S., Hiltke, T., Rivers, R.C., Sokoll, L., Zhu, H., Shih, I.M., Cope, L., Pandey, A., Zhang, B., Snyder, M.P., Levine, D.A., Smith, R.D., Chan, D.W., Rodland, K.D., Carr, S.A., Gillette, M.A., Klauser, K.R., Kuhn, E., Mani, D.R., Mertins, P., Ketchum, K.A., Thangudu, R., Cai, S., Oberti, M., Paulovich, A.G., Whiteaker, J.R., Edwards, N.J., McGarvey, P.B., Madhavan, S., Wang, P., Chan, D.W., Pandey, A., Shih, I.M., Zhang, H., Zhang, Z., Zhu, H., Cope, L., Whiteley, G.A., Skates, S.J., White, F.M., Levine, D.A., Boja, E.S., Kinsinger, C.R., Hiltke, T., Mesri, M., Rivers, R.C., Rodriguez, H., Shaw, K.M., Stein, S.E., Fenyo, D., Liu, T., McDermott, J.E., Payne, S.H., Rodland, K.D., Smith, R.D., Rudnick, P., Snyder, M., Zhao, Y., Chen, X., Ransohoff, D.F., Hoofnagle, A.N., Liebler, D.C., Sanders, M.E., Shi, Z., Slebos, R.J.C., Tabb, D.L., Zhang, B., Zimmerman, L.J., Wang, Y., Davies, S.R., Ding, L., Ellis, M.J.C., Townsend, R.R.: Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell **166**(3), 755–765 (2016)

3. Pawson, T., Scott, J.D.: Protein phosphorylation in signaling–50 years and counting. Trends Biochem. Sci. **30**(6), 286–290 (2005)

4. Needham, E.J., Parker, B.L., Burykin, T., James, D.E., Humphrey, S.J.: Illuminating the dark phosphoproteome. Sci. Signal. **12**(565), 8645 (2019)

5. Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., Sullivan, M.: Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic acids research **40**(D1), 261–270 (2011)

6. Perfetto, L., Briganti, L., Calderone, A., Cerquone Perpetuini, A., Iannuccelli, M., Langone, F., Licata, L., Marinkovic, M., Mattioni, A., Pavlidou, T., *et al.*: Signor: a database of causal relationships between biological entities. Nucleic acids research **44**(D1), 548–554 (2015)

7. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., *et al.*: The reactome pathway knowledgebase. Nucleic acids research **42**(D1), 472–477 (2013)

8. Torii, M., Arighi, C.N., Li, G., Wang, Q., Wu, C.H., Vijay-Shanker, K.: Rlims-p 2.0: a generalizable rule-based information extraction system for literature mining of protein phosphorylation information. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) **12**(1), 17–29 (2015)

9. Valenzuela-Escárcega, M.A., Babur, Ö., Hahn-Powell, G., Bell, D., Hicks, T., Noriega-Atala, E., Wang, X., Surdeanu, M., Demir, E., Morrison, C.T.: Large-scale automated machine reading discovers new cancer-driving mechanisms. Database **2018** (2018)

10. Gyori, B.M., Bachman, J.A., Subramanian, K., Muhlich, J.L., Galescu, L., Sorger, P.K.: From word models to executable models of signaling networks using automated assembly. Mol. Syst. Biol. **13**(11), 954 (2017)

11. Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., *et al.*: Human protein reference database—2006 update. Nucleic acids research **34**(suppl_1), 411–414 (2006)

12. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: Pid: the pathway interaction database. Nucleic acids research **37**(suppl_1), 674–679 (2008)

13. McDonald, D., Friedman, S., Paullada, A., Bobrow, R., Burstein, M.: Extending biology models with deep nlp over scientific articles. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (2016)

14. Torii, M., Arighi, C.N., Li, G., Wang, Q., Wu, C.H., Vijay-Shanker, K.: Rlims-p 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information. IEEE/ACM Transactions on Computational Biology and Bioinformatics **12**(1), 17–29 (2015). doi:10.1109/TCBB.2014.2372765

15. Ren, J., Li, G., Ross, K., Arighi, C., McGarvey, P., Rao, S., Cowart, J., Madhavan, S., Vijay-Shanker, K., Wu, C.H.: iTextMine: integrated text-mining system for large-scale knowledge extraction from the literature. Database (Oxford) **2018** (2018)

16. Consortium, T.U.: UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research **47**(D1), 506–515 (2018). doi:10.1093/nar/gky1049. http://oup.prod.sis.lan/nar/article-pdf/47/D1/D506/27437297/gky1049.pdf

17. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D.: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. **44**(D1), 733–745 (2016)

18. Hong, J.Y., Oh, I.-H., McCrea, P.D.: Phosphorylation and isoform use in p120-catenin during development and tumorigenesis. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1863**(1), 102–114 (2016). doi:10.1016/j.bbamcr.2015.10.008

19. Leclerc, G.M., Leclerc, G.J., Fu, G., Barredo, J.C.: AMPK-induced activation of Akt by AICAR is mediated by IGF-1R dependent and independent mechanisms in acute lymphoblastic leukemia. J Mol Signal **5**, 15 (2010)

20. Coppe, J.P., Patil, C.K., Rodier, F., Sun, Y., Munoz, D.P., Goldstein, J., Nelson, P.S., Desprez, P.Y., Campisi, J.: Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. PLoS Biol. **6**(12), 2853–2868 (2008)

21. Bachman, J.A., Gyori, B.M., Sorger, P.K.: Famplex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. BMC bioinformatics **19**(1), 248 (2018)

22. Chen, L., Li, Z., Ahmad, N., Liu, X.: Plk1 phosphorylation of IRS2 prevents premature mitotic exit via AKT inactivation. Biochemistry **54**(15), 2473–2480 (2015)

23. Mao, Y., Xi, L., Li, Q., Cai, Z., Lai, Y., Zhang, X., Yu, C.: Regulation of cell apoptosis and proliferation in pancreatic cancer through PI3K/Akt pathway via Polo-like kinase 1. Oncol. Rep. **36**(1), 49–56 (2016)

24. Mao, Y., Xi, L., Li, Q., Wang, S., Cai, Z., Zhang, X., Yu, C.: Combination of PI3K/Akt Pathway Inhibition and Plk1 Depletion Can Enhance Chemosensitivity to Gemcitabine in Pancreatic Carcinoma. Transl Oncol **11**(4), 852–863 (2018)

25. Ediriweera, M.K., Tennekoon, K.H., Samarakoon, S.R.: Role of the PI3K/AKT/mTOR signaling pathway in ovarian cancer: Biological and therapeutic significance. Semin. Cancer Biol. (2019)

26. Tan, Y., Cheung, M., Pei, J., Menges, C.W., Godwin, A.K., Testa, J.R.: Upregulation of DLX5 promotes ovarian cancer cell proliferation by enhancing IRS-2-AKT signaling. Cancer Res. **70**(22), 9197–9206 (2010)

27. Zhang, R., Shi, H., Ren, F., Liu, H., Zhang, M., Deng, Y., Li, X.: Misregulation of polo-like protein kinase 1, P53 and P21WAF1 in epithelial ovarian cancer suggests poor prognosis. Oncol. Rep. **33**(3), 1235–1242 (2015)

28. Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., Tyers, M.: The BioGRID interaction database: 2019 update. Nucleic Acids Res. **47**(D1), 529–541 (2019)

29. Besnard, A., Galan-Rodriguez, B., Vanhoutte, P., Caboche, J.: Elk-1 a transcription factor with multiple facets in the brain. Front Neurosci **5**, 35 (2011)

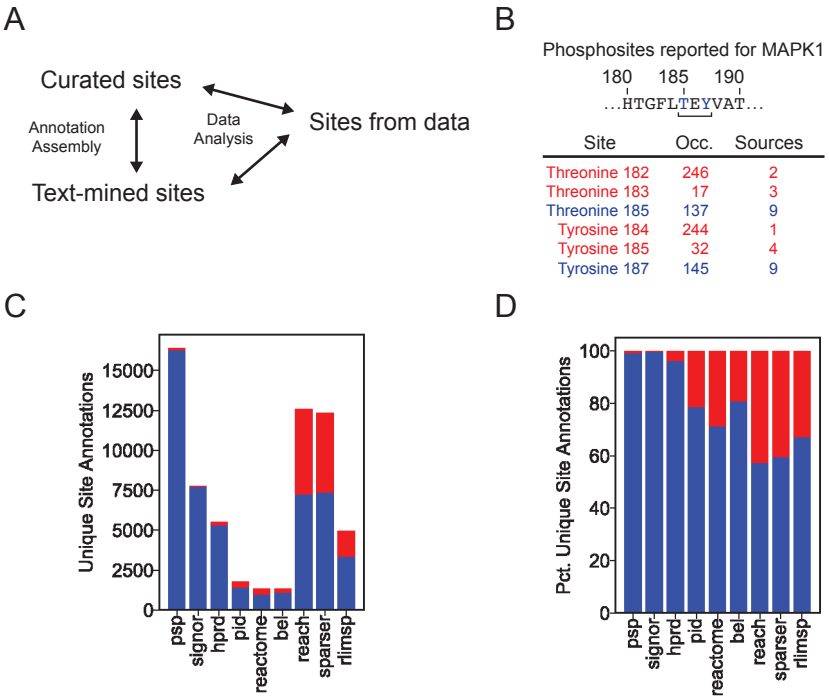

Figure 1. Inconsistencies in reported positions of PTMs. **(A)** Interpreting phosphoproteomic data requires linking experimentally observed phosposites to annotations assembled from both curated databases and literature. **(B)** An example of inconsistent site positions for the T/Y activation motif in human MAPK1. Relevant sequence of the canonical isoform from Uniprot is shown above with T and Y residues shown in blue. "Occ." denotes the total number of occurrences (reactions or sentences) across all sources; "Sources" denotes the number of sources reporting the site at the given position. Non-canonical positions are shown in red; canonical positions in blue. **(C)** Number of unique site annotations found in human reference sequence, by source. A site annotation consists of a unique combination of *regulator, substrate, residue, position*. **(D)** Percentage of unique site annotations matching human reference sequence, by source.

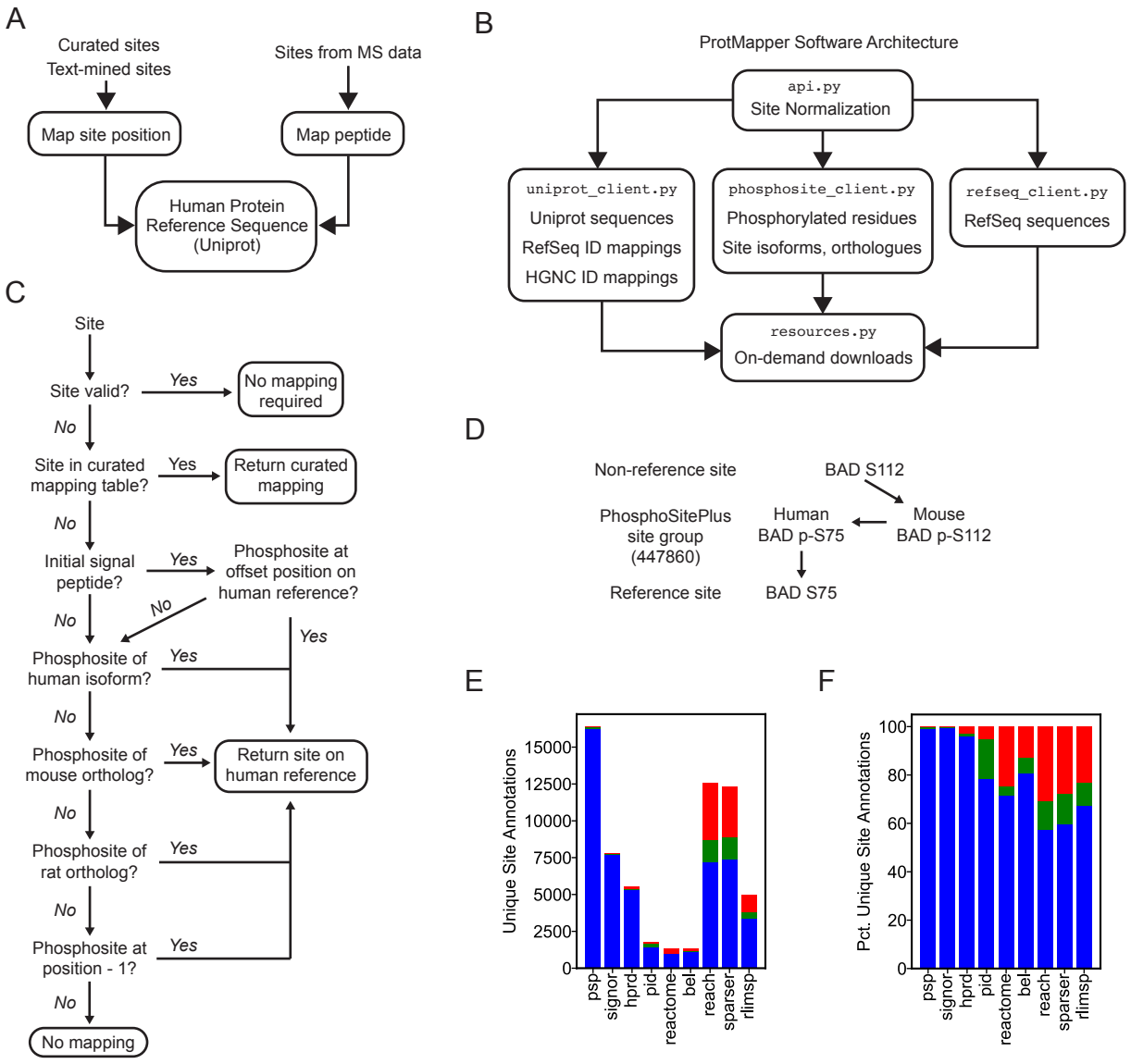

Figure 2. Mapping invalid sites to human reference sequence positions. **(A)** The ProtMapper maps identifiers and site positions for phosphosite annotations experimental phosphopeptides to corresponding positions on the human Uniprot reference sequence. **(B)** ProtMapper Software Architecture. Site normalization is implemented in the `api.py` module, which draws on additional modules for ID normalization, protein sequences, and phosphosite information. Resource files are downloaded as needed at run time using functions in `resources.py`. **(C)** Site normalization procedure. A site, including residue and position, is checked for validity against the Uniprot reference sequence. If invalid, a series of mappings are attempted, starting with a curated mapping table and proceeding through positions for known phosphorylation sites after signal peptide cleavage (if any); known phosphosite positions from human isoforms and mouse and rat orthologues; and known phosphosites at position - 1, a common inconsistency due to the cleavage of the initial methionine of the protein. **(D)** PhosphoSitePlus site groups. If a known phosphorylation site with the given position is found on an alternative sequence (e.g., the mouse ortholog), it can be mapped back to the human reference position via PhosphoSitePlus site groups, which link together corresponding sites from human isoforms and non-human orthologs. **(E)** Counts of valid, mappable, and invalid/unmappable site annotations, by source. **(F)** Percentages of valid, mappable, and invalid/unmappable site annotations, by source.

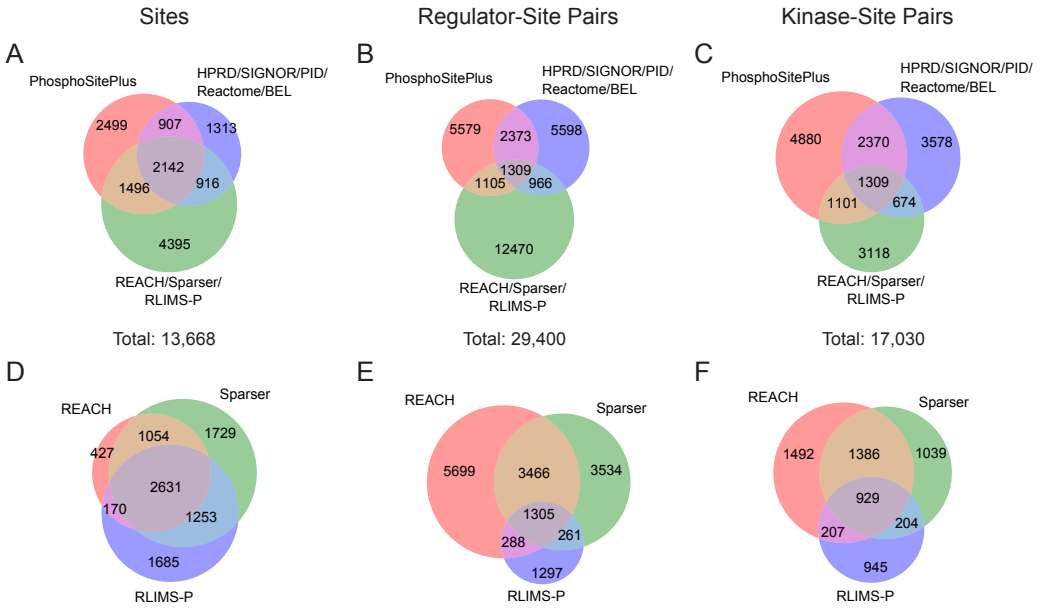

Figure 3. Overlap of phosphosite information between PhosphoSitePlus, the combined output of the HPRD, SIGNOR, PID, Reactome, and BEL databases, and the combined output of the REACH, Sparser and RLIMS-P machine reading systems. **(A)** Venn diagram of unique sites (irrespective of regulator) reported by PSP, other pathway databases, and machine reading systems. **(B)** Venn diagram of unique regulator-site pairs reported by PSP, other pathway databases, and machine reading systems. **(C)** Venn diagram of unique regulator-site pairs filtered to human-only kinases and removing protein families and complexes, reported by PSP, other pathway databases, and machine reading systems. **(D, E, F)** Venn diagrams showing overlap between machine reading systems REACH, Sparser, and RLIMS-P, with categories following panels A, B, and C.

	Source	Total Annot.	Valid	Valid Pct.	Invalid	Invalid Pct.	Mapped	Mapped/Invalid (Pct.)
1	psp	16,415	16,280	99.2	135	0.8	75	55.6
2	signor	7,765	7,735	99.6	30	0.4	17	56.7
3	hprd	5,529	5,321	96.2	208	3.8	43	20.7
4	pid	1,792	1,408	78.6	384	21.4	289	75.3
5	reactome	1,349	963	71.4	386	28.6	51	13.2
6	bel	1,346	1,087	80.8	259	19.2	89	34.4
7	reach	12,575	7,225	57.5	5,350	42.5	1,461	27.3
8	sparser	12,360	7,375	59.7	4,985	40.3	1,538	30.9
9	rlimsp	4,967	3,341	67.3	1,626	32.7	475	29.2

Table 1. Annotations linked to valid, invalid, and mappable sites, by source.

Type of inconsistency	Example Site	Description
Site position from other species	BAD S112	Mouse site, corresponds to human S75
Site position from other isoform	RPS6KB1 T389	Position from isoform Alpha II
Site position from other family member	RPS6KA6 S221	Position from RPS6KA1
Initial methionine cleavage in mature protein	LCK Y393	Y394 in reference sequence
Signal peptide cleavage in mature protein	EGFR Y1173	Position after cleavage of initial 24aa
Curator error: wrong residue	BRAF T151	Should be BRAF S151
Curator error: site from wrong protein	JAK2 Y485	Site on EPOR where JAK2 binds
Text mining error: misidentified protein	HNRNPU S268	Ambiguous "p120" in sentence refers to CTNND1
Text mining error: site from wrong protein	IRS1 T308	Site for AKT1, mentioned in same sentence
Text mining error: misidentified site	TP53 S1	"Text S1" extracted as site

Table 2. Types of inconsistencies for sites in curated databases and text mining results.

Category	Count
True positive	42
True negative	41
False positive	2
False negative	15
Total	100
Precision	0.95
Recall	0.74
F1	0.83

Table 3. Summary of results for mapping invalid sites extracted from the literature using machine reading, with a sample size of 100 sites.

Mapping type	Number (total)	Number (correct)
Inferred mouse site	13	13
Manually curated mapping	10	10
Inferred methionine cleavage	7	5
Reference mismatch between UniProt and PSP	6	6
Inferred alternative isoform	4	4
Inferred signal peptide cleavege	4	4

Table 4. Total and correct site mappings of different types in the curation dataset

Row	Source ID	Site status	BRCA	%	OVCA	%
1	HGNC	No Uniprot ID, invalid gene symbol	67	0.1	23	0.1
2	HGNC	Valid in Uniprot ref sequence for gene	68,867	77.5	16,981	77.0
3	HGNC	Not valid in Uniprot ref sequence for gene	19,977	22.5	5,056	22.9
4	HGNC	Invalid but mappable to Uniprot ref sequence	18,630	21.0	4,783	21.7
5	HGNC	Total mappable to Uniprot ref sequence	87,093	98.0	21,654	98.2
6	RefSeq	No Uniprot ID	6,895	7.8	1,320	6.0
7	RefSeq	No Uniprot ID, mappable to Uniprot seq via HGNC	6,471	7.3	1,258	5.7
8	RefSeq	Valid in Uniprot sequence from RefSeq ID	78,630	88.4	19,849	90.0
9	RefSeq	Not valid in Uniprot sequence from RefSeq ID	3,386	3.8	891	4.0
10	RefSeq	Isoform-specific ID	22,044	24.8	5,920	26.8
11	RefSeq	Isoform-specific ID, mappable to Uniprot ref seq	21,107	23.7	5,688	25.8
12		Total Sites	88,911	100.0	22,060	100.0

Table 5. Results of mapping phosphosites from the CPTAC Breast (BRCA) and Ovarian Cancer (OVCA) datasets to Uniprot sequences via gene symbols or RefSeq IDs.

	Annotations mapped?	Peptides mapped?	Sources	BRCA Ann. Sites	BRCA Mean Ann./Site	OVCA Ann. Sites	OVCA Mean Ann./Site
1	No	No	PSP only	1540	1.67	821	1.69
2			DBs only	2052	2.14	1,053	2.22
3			NLP only	1021	4.01	522	4.66
4			All	2284	3.37	1,165	3.74
5		Yes	PSP only	1,877	1.63	1,021	1.68
6			DBs only	2,479	2.10	1,299	2.20
7			NLP only	1,228	3.88	643	4.53
8			All	2,754	3.28	1,437	3.67
9	Yes	No	PSP only	1,546	1.67	826	1.68
10			DBs only	2,074	2.07	1,063	2.16
11			NLP only	1,181	3.75	603	4.10
12			All	2,365	3.10	1,204	3.34
13		Yes	PSP only	1,885	1.62	1,028	1.67
14			DBs only	2,508	2.03	1,313	2.12
15			NLP only	1,439	3.67	760	4.11
16			All	2,860	3.06	1,493	3.37
17			Total Sites	88,911		22,060	

Table 6. Regulatory annotations of sites in the CPTAC Breast (BRCA) and Ovarian Cancer (OVCA) phosphoproteomic datasets with and without site normalization.

Reading correct	Result of mapping	Classification	Explanation
No	No mapping	True negative	The reader extracted an incorrect site for which no mapping was done.
No	Mapping done	False positive	The reader extracted an incorrect site which was subsequently mapped.
Yes	Correct mapping	True positive	The reader extracted the site correctly, which was then correctly mapped.
Yes	Incorrect mapping	False positive	The reader extracted the site correctly, which was then mapped to the wrong site.
Yes	No mapping	False negative	The reader extracted the site correctly but no mapping was found.

Table 7. Curation categories for invalid sites extracted from the literature by machine reading and then mapped to human reference.