MEDWIN PUBLISHERS
Committed to Create Value for researchers

Journal of Embryology & Stem Cell Research
ISSN: 2640-2637

hPP Corpus: A Tagged Biomedical Corpus for Automatic Extraction of Human Protein Phosphorylation for Understanding Cellular Functions

Raja K, Subramanian D, Abdulkadhar S and Natarajan J*

Data Mining and Text Mining Laboratory, Bharathiar University, Coimbatore, India

***Corresponding author:** Jeyakumar Natarajan, Department of Bioinformatics, Data Mining and Text Mining Laboratory, School of Life Sciences, Bharathiar University, Coimbatore-641 046, India, Fax: +91 422 2422387; Tel: +914222428281; Email: n.jeyakumar@yahoo.co.in

Research Article

Volume 4 Issue 1

Received Date: February 12, 2020

Published Date: April 09, 2020

DOI: 10.23880/jes-16000140

Abstract

Proteins perform their functions by interacting with other proteins. Phosphorylation is a post-transcriptional modification of proteins and plays an important role in cellular functions. Protein interaction and phosphorylation play a critical role in biological functions and indicate disease states including cancer, Alzheimer's disease and Parkinson's disease. Mining protein phosphorylation information from biomedical literature is a topic of interest in biomedical text mining and highly challenging. Text mining researchers apply a variety of algorithms to extract such information. A standard annotated corpus is necessary to evaluate the performance of the text mining algorithms. However, to our best knowledge there is no standard annotated corpus available for evaluating approaches related to the extraction of protein phosphorylation information related to human. The available corpora, iProLink, PTM (Post Transcriptional Modification) phosphorylation extraction corpus and protein phosphorylation corpus from Protein Information Resource (PIR) are not specific to human. In this paper, we present a corpus called 'hPP (human Protein Phosphorylation) corpus' exclusively on human protein phosphorylation information. Current version of hPP corpus contains 2,380 sentences from 1,000 MEDLINE abstracts related to human protein phosphorylation. The corpus is annotated with named entities, event relationship and syntactic dependencies, and freely available at http://www.biominingbu.org/hPPcorpus/hPP_corpus.xml. To our best knowledge hPP corpus is the first and foremost annotated corpus available for evaluating text mining systems on extracting human protein phosphorylation from MEDLINE abstracts.

Keywords: Cellular Function; Protein Phosphorylation; Post-Transcriptional Modification; Text Mining, Information Extraction; Named Entity Recognition

Abbreviations: PTM: Post Transcriptional Modification; PIR: Protein Information Resource; hPP: Human Protein Phosphorylation; PRIDE: PRoteomics Identifications Database; BiGG: Biochemical Genetic and Genomic knowledgebase Database; CGD: Clinical Genomic Database; NER: Named Entity Recognition; PPI: Protein-Protein Interaction; HGNC: Human Gene Nomenclature Committee; NLP: Natural Language Processing; SVM: Support Vector Machines; PK: Phosphorylation Keyword; P: substrate; K: Protein Kinase; S: Phosphorylation Site.

Introduction

Advances in biomedical research with the use of large

scale experimental techniques and bioinformatics tools have greatly accelerated the publication rate of biomedical literature. This exponential growth of experimental data and their publication has promoted the active research in biomedical text mining to facilitate annotation of genes/proteins and to improve the quality of information available in the biological databases. The various curated proteomics databases such as UniProt protein knowledgebase [1], PIR (Protein Information Resource) [2], PRoteomics Identifications DatabasE (PRIDE) [3], MitoMiner [4], etc., and genome databases such as Ensembl [5], Biochemical Genetic and Genomic knowledgebase (BiGG) Database [6], Clinical Genomic Database (CGD) [7] etc. represent annotations derived from the experimentally verified knowledge on genes and proteins from the published literature database PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). Consequently, there has been an increasing interest in applying text mining methods to facilitate the biomedical relation/event extraction from the literature resource [8-10].

One of the major objectives of text mining community is to automatically identity the biological entities and to extract their functional relationships or the related biological processes and cellular functions to produce a structured representation of relevant information. Many systems are being proposed for the automatic extraction of information spread across several publications. However, the demand for more accurate and robust system remains as an open challenge to develop new approaches with improved performance [8]. Evaluation of such text mining systems requires a gold-standard corpus to measure the accuracy as well as to compare their performance with other systems developed for the same purpose. A gold-standard corpus is a collection of texts with enhanced markups specifying both linguistic and domain information such as syntactic structure, entity recognition and their relationships.

A typical text mining system employs three most important steps such as named entity recognition (NER), parsing and domain analysis to extract information from the text. Among these, NER identifies the entities present in the unstructured text, parsing builds the syntactic structure of the text and domain analysis extracts the relationships among the entities. Several biomedical corpora have been developed to facilitate the performance evaluation of individual components of text mining systems. The corpora related to human literature include JNLPBA [11] for NER, and AIMED [12] and HPRD50 [13] for protein-protein interaction (PPI) extraction. The corpora for specific biological process such as protein phosphorylation are very few [14] and require in-depth biological knowledge on post-transcriptional modification of proteins. Phosphorylation is one of the most common post-transcriptional modification of proteins where a serine/threonine/tyrosine residue is phosphorylated by a protein kinase. It is one of the emerging domains of interest among the text mining community for building the regulatory network of biological pathways [15] and requires standardized corpus to understand the syntactic and semantic structure of the text with such information.

In this paper, we first present brief introduction about protein phosphorylation process and the three biological entities related to protein phosphorylation extraction. Next, the literature resource and various external databases and tools used to perform the annotation task is introduced. This is followed by the complete and comprehensive description of the various annotation processes explored in constructing the hPP corpus. Finally, statistics on the annotations made in the corpus and comparison with other available corpora are given.

Materials and Methods

Protein Phosphorylation Process

Protein phosphorylation plays an essential role in cellular function and signal transduction, and receives a significant amount of attention [16]. Phosphorylation is one of the most common post-transcriptional modification of proteins where a serine/threonine/tyrosine residue is phosphorylated by a protein kinase. Phosphorylation occurs on several amino acids within a protein. Apart from the commonly observed serine, threonine, and tyrosine residues, phosphorylation has been reported to occur on phosphoserine residue in eukaryotes, and on histidine / aspartate residues in prokaryotes [17]. The biological entities related to protein phosphorylation are:

(i) Enzyme/kinase (that phosphorylates protein)
(ii) Protein/substrate (that is phosphorylated)
(iii) Site (the residue (serine/threonine/tyrosine) that is phosphorylated

Retrieval of Literature Dataset

PubMed is the main resource for biomedical literature. The database contains 122,250 MEDLINE abstracts related to human protein phosphorylation and literature data sampling for creating the corpus is highly complicated. However, experimentally verified information on phosphorylation is available in UnitProt Knowledgebase [1] as well as in specialized databases such as Phospho.ELM [18], PhosphoBase [19], PhosphoSitePlus [20], PhosphoNET [21] and PhosphoPOINT [22].

The literature data for corpus creation is derived from

PubMed using the MEDLINE abstract references from two popular phosphorylation databases namely Phospho. ELM [18] and PhosphoSitePlus [20]. Phospho.ELM is a database of experimentally verified phosphorylation sites in eukaryotic proteins, providing UniPROT/Ensembl accession number, sequence, phosphorylated residue and its position, PubMed ID, the upstream kinase (when known), source (High- ThroughPut/Low-ThroughPut) and entry date for the various species including human. PhosphoSitePlus is a comprehensive resource on experimentally determined post-transcriptional modification of proteins including acetylation, methylation, and phosphorylation in man and mouse. The database entries include protein name, type and accession number, phosphorylation residue, PubMed ID etc. created by cell-signaling technology.

We adopted the likelihood prediction approach of GeneTag [23] to rank the MEDLINE abstracts containing information on gene/protein names i.e. substrate and kinase. High-scoring abstracts always contain substrate/kinase names when compared to the low-scoring abstracts with few or no substrate/kinase names. We randomly selected 1000 high-scoring MEDLINE abstracts as the basis for hPP corpus, because we required the corpus to contain both substrate and kinase names. A sample sentence from biomedical literature on protein phosphorylation information is illustrated in Example 1, where ERalpha is a substrate protein phosphorylated by a kinase, protein kinase A at the phosphorylation site serine-236.

Example 1: PMID: 9891036

Here we show that ERalpha is phosphorylated by protein kinase A on serine-236 within the DNA binding domain.

Text Preprocessing

Our text preprocessing is carried out in two phases. An initial preprocessing to retrieve article ID, title and abstract from MEDLINE format is carried out with an in-house Java program. We prefer MEDLINE format than XML format for data collection, because the former automatically converts the special characters such as α, β, μ etc. to their corresponding English terms i.e. alpha, beta, mew etc., thus facilitating the automatic text processing by text mining tools. The input text is split into a set of lines, where each line contains only one sentence assigned with PubMed ID. Sentences containing phosphorylation related keywords (Table 1) are more likely to contain phosphorylation information than those containing other words and are filtered out for further processing.

Root word	Keywords
phosphorylation	phosphorylate, phosphorylates, phosphorylating, phosphorylated, phosphorylation, phosphorylations
auto-phosphorylation	auto-phosphorylate, auto-phosphorylates, autophosphorylating, auto-phosphorylated, auto-phosphorylation, auto-phosphorylations
auto phosphorylation	auto-phosphorylate, auto-phosphorylates, autophosphorylating, auto-phosphorylated, auto-phosphorylation, auto-phosphorylations

Table 1: List of phosphorylation keywords.

Corpus Annotation

We describe the corpus annotation with two key steps namely, entity annotation and event annotation. The first stage of corpus annotation is the identification of biomedical entities such as proteins, substrates, protein kinases and phosphorylation sites related to protein phosphorylation. The next stage is an extended annotation of the identified biomedical entities to recognize the pertaining phosphorylation event relationship between the entities. For instance, the annotated entities in Example 1 are not only proteins (ERalpha and protein kinase A) and phosphorylation sites (serine-236), but also the biological process "ERalpha is phosphorylated by protein kinase A on serine-236" that pertain to the proteins (Figure 1).

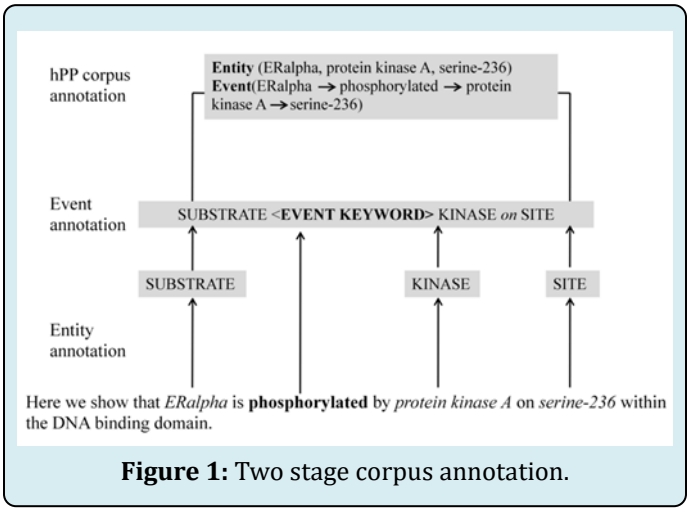

Figure 1: Two stage corpus annotation.

Automatic Annotation of Gene/Protein Names

NER approach using NAGGNER: Two of the three phosphorylation objects are proteins i.e., kinases and substrates. Hence, we used the hybrid protein name tagger called NAGGNER [24] for automatic recognition of protein names. NAGGNER consists of three major components namely (i) machine learning using conditional random fields (CRF) for initial learning and labeling, (ii) rule based tagging to improve the performance of initial tagging process, and (iii) a two stage abbreviation identification algorithm for the recognition of human gene/protein mentions.

Pattern templates for gene/protein name annotation: We introduced a set of pattern templates to tag the protein names (e.g. suppressor of fused (Sufu) protein in PubMed ID: 21317289) that were missed by the hybrid tagger. Our templates are specific patterns of grammatical components within the noun phrase (NP) of the input sentence [25]. We parsed the input sentence with Stanford lexical parser with grammar settings to englishPCFG module [26] to obtain a parse tree structure. We located the head-word within the top level NP. The outputs from the parser were manually reviewed for parsing accuracy. The head-word is always a noun or proper name (possibly with more than one word). We considered all the words preceding the first head-word as pre-modifiers and all the words following the first head word as post-modifiers within NP and apply the pattern templates to extract the phrase with protein names. We used the proteins dictionary from our previous work [27] to confirm whether the phrase corresponds to a protein name. The syntax tags expressed in the pattern matching templates are listed in Table 2.

Syntax Tag	Description
NP	Noun phrase
PP	Prepositional phrase
NN	Noun
DT	Determiner
JJ	Adjective
IN	Preposition
?	Optional
+	One or more words
/	Or

Table 2: List of syntax tags and description.

Pattern 1: NP [?DT/JJ/NN NN+)]

A simple NP with optional pre-modifiers such as determiner (DT), adjective (JJ), or noun (NN) followed by a head-word is the most common pattern for a protein name.

Pattern 2: NP [NP ?PP [IN NP]]

NP that includes another NP followed by an optional PP is tagged as a protein name. PP is made up of a preposition (IN) such as 'for', 'in', 'etc.', and 'of', followed by another NP.

Pattern 3: NP [NP ?acronym]

The pattern template is similar to Pattern 2. NP within a top level NP is followed by an optional acronym. Antiporter, antizyme, complement, exchanger, neuropeptide, oncoprotein, photoreceptor, receptor, symporter, and transporter are few important acronyms which have to be tagged as part of protein names. A set of generic names which may not be tagged individually or with preceding protein names is in given in Table 3.

Single word non-protein acronyms			
Acceptor	Domain	Isoenzyme	Polysaccharide
Activator	Effector	Isoform	Precursor
Adapter	Ehancer	Isolog	Proactivator
Adaptor	Enzyme	Isotype	Product
Antibody	Facilitator	Isozyme	Proenzyme
Biglycan	Factor	Mediator	Propeptide
Binder	Fragment	Modifier	Protein
Carrier	Glycopeptides	Modulator	Proteoglycan
Chain	Heterodimer	Molecule	Proteolipid
Channel	Holoenzyme	Motif	Pump
Coactivator	Homolog	Oligopeptide	Regulator
Coatomer	Homologue	Ortholog	Repressor
Coenzyme	Inducer	Partner	Responder
Complex	Inhibitor	Pentapeptide	Sequence

Component	Initiator	Peptide	Subunit
Cotransporter	Integrator	Peptidoglycan	Subtype
Dipeptide	Interactor	Polypeptide	Suppressor
Multi word non-protein acronyms			
Protein kinase			

Table 3: Generic non-protein acronyms.

Example 2: PMID: 7925415
Original sentence: An epidermal-growth-factor (EGF)-receptor preparation isolated by calmodulin-affinity chromatography from rat liver plasma membranes is able to phosphorylate calmodulin.

Tagging by NAGGNER and pattern: An <KINASE> epidermal-growth-factor (EGF)–receptor </-KINASE> preparation isolated by calmodulin-affinity chromatography from rat liver plasma membranes is able to phosphorylate <PROTEIN> calmodulin </PROTEIN> .

In Example 2, epidermal growth factor receptor is identified and tagged as protein by NAGGNER and two pattern templates 1 and 3 (Figure 2).

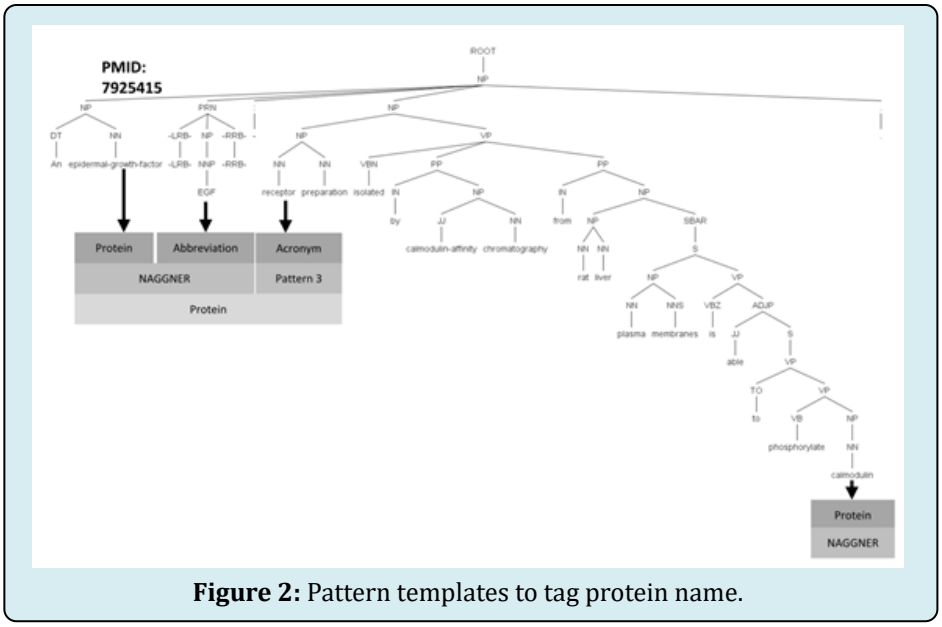

Figure 2: Pattern templates to tag protein name.

Protein kinase recognition: Recognition of kinases among the tagged gene/protein names is a challenging task because each protein kinase has a list of related synonyms [27]. For example, spleen tyrosine kinase has 3 related synonyms (SYK, p72-Syk, tyrosine-protein kinase SYK). We developed a specialized synonyms dictionary to distinguish human protein kinases from other human genes/proteins (Table 4).

Official Symbol	Gene ID	Synonyms	Human Protein / Protein Kinase
.....	.....	.....	.....
BMP5	653	bone morphogenetic protein 5; BMP5; BMP-5	Human Protein
LIMK2	3985	LIM domain kinase 2; LIMK2;	Human Protein Kinase
TMEM26	219623	transmembrane protein 26; TMEM26;	Human Protein
SYK	6850	spleen tyrosine kinase; SYK; p72-Syk; tyrosine-protein kinase SYK	Human Protein Kinase
.....	.....	.....	.....

Table 4: Specialized synonyms dictionary of human proteins and protein kinases.

First, we downloaded all human genes/proteins approved by Human Gene Nomenclature Committee (HGNC) using the keyword (Homo sapiens [Organism]) AND HGNC from EntrezGene (http://www.ncbi.nlm.nih.gov/gene). Next, we mapped the protein names to the synonyms collected from three resources namely (i) the synonym dictionary provided by BioCreAtIvE-II [28] contest, comprising of 32,975 entries (ii) the human protein/gene name dictionary available at NCBI (http://www.ncbi.nlm.nih.gov/gene) covering 47,177 entries and (iii) the human protein synonyms available at UniProt [1,29] covering 14,893 entries. Finally, we generated a list of 518 human protein kinases from two datasets namely (i) human genes/protein kinases approved by HGNC from EntrezGene and (ii) protein kinase entries from two popular databases namely KinBase [30] containing 516 entries and Kinweb [31] containing 518 entries. We mapped the kinase list with the corresponding entries in the specialized synonyms dictionary. Thus, the dictionary contains (i) official symbol, (ii) gene ID (iii) all possible synonyms and (iv) Disambiguation as proteins and kinases for all HGNC approved human genes/proteins.

Automatic Recognition of Phosphorylation Sites

Recognition of phosphorylation residue/site in the unstructured text is equally challenging like protein/gene name recognition. The three amino acid (serine, threonine and tyrosine) residues related to protein phosphorylation are represented in three forms in the biomedical literature: i) full name (e.g. serine), ii) short name (e.g. Ser) and iii) acronym (e.g. S) [15,32]. The residue is commonly represented with the location number (e.g. serine 432). We used Java Regex function with a set of pattern templates to tag all possible forms of phosphorylation site. Table 5 lists the phosphorylation site patterns and the related Regex functions.

Site name	Residue	Regex pattern
Full name	Serine 84 / serine 84	[S|s]erine\\s\\d+
	Threonine 84 / threonine 84	[T|t]hreonine\\s\\d+
	Tyrosine 84 / tyrosine 84	[T|t]yrosine\\s\\d+
	Serine-84 / serine-84	[S|s]erine-\\d+
	Threonine-84 / threonine-84	[T|t]hreonine-\\s\\d+
	Tyrosine-84 / tyrosine-84	[T|t]yrosine-\\s\\d+
Short name	Ser 84 / Thr 84 / Tyr 84 /	[S|T|s|t][e|h|y]r\\s\\d+
	ser 84 / thr 84 / tyr 84	
	Ser-84 / Thr-84 / Tyr-84 /	[S|T|s|t][e|h|y]r-\\d+
	ser-84 / thr-84 / tyr-84	
	Ser84 / Thr84 / Tyr84 /	[S|T|s|t][e|h|y]r\\d+
	ser84 / thr84 / tyr84	
	Ser (84) / Thr (84) / Tyr (84) /	[S|T|s|t][e|h|y]r\\s(\\d+)
	ser (84) / thr (84) / tyr (84)	
	Ser(84) / Thr(84) / Tyr(84) /	[S|T|s|t][e|h|y]r(\\d+)
	ser(84) / thr(84) / tyr(84)	
	Ser (84) / Thr (84) / Tyr (84)	[S|T|s|t][e|h|y]r\\s(\\s\\d+\\s)
	/ ser (84) / thr (84) / tyr (84)	
Acronym	S 84 / T 84 / Y 84	[S|T|Y]\\s\\d+
	S-84 / T-84 / Y-84	[S|T|Y]-\\d+
	S84 / T84 / Y84	[S|T|Y]\\d+
	S(84) / T(84) / Y(84)	[S|T|Y](\\d+)
	S (84) / T (84) / Y (84)	[S|T|Y]\\s(\\s\\d+\\s)

Table 5: Phosphorylation site and the related regex patterns.

Manual Annotation Process

The common practice in the annotation process is to start with the automatically parsed and annotated text. However, since the automatic annotation could not provide 100% accurate results, we chose to perform a second level of manual annotation on the automatically tagged text. We hired three annotators with knowledge in protein phosphorylation to investigate the automatically annotated data:

(i) to tag biological entities missed / incorrectly tagged by automatic tagging

Example 3: PMID: 2457390

Automatic Tagging: The major site of phosphorylation by adenosine cyclic 3', 5' -phosphate dependent protein kinase was on the carboxy-terminal half of the molecule at <SITE> Thr-216 </SITE>.

Manual Tagging: The major site of phosphorylation by <KINASE> adenosine cyclic 3', 5' -phosphate dependent protein kinase </KINASE> was on the carboxy-terminal half of the molecule at <SITE> Thr-216 </SITE>.

(ii) to remove / re-tag the incorrectly assigned tags

Example 4: PMID: 12118371

Automatic Tagging: These results suggest that phosphorylation of the <SUBSTRATE> EGF receptor </SUBSTRATE> at <SITE> Thr669 </SITE> and <SITE> Ser671 </SITE> mediates interaction of the receptor with a specific <PROTEIN> tyrosine kinase </PROTEIN> substrate and is required for efficient ligand-induced receptor internalization.

Manual Tagging: These results suggest that phosphorylation of the <SUBSTRATE> EGF receptor </SUBSTRATE> at <SITE> Thr669 </SITE> and <SITE> Ser671 </SITE> mediates interaction of the receptor with a specific tyrosine kinase substrate and is required for efficient ligand-induced receptor internalization.

(iii) to re-tag the complex phosphorylation site mentions in simple machine readable form

Example 5: : 2019585

Automatic Tagging: Solid phase sequencing revealed phosphorylation at serines 122 , 150 , 212 , 220 , 234 , and 315 and <SITE> threonine 159 </SITE> .

Manual Tagging: Solid phase sequencing revealed phosphorylation at serines <SITE> Ser 122 </SITE>, <SITE> Ser 150 </SITE>, <SITE> Ser 212 </SITE>, <SITE> Ser 220 </SITE>, <SITE> Ser 234 </SITE>, and
<SITE> Ser 315 </SITE> and <SITE> threonine 159 </SITE> .

(iv) to confirm the event relationship between the related entities using a set of annotation rules.

We created a set of annotation rules using the available pattern templates [15,32], experience gained during manual annotation and experts' knowledge for identifying sentences with positive event relationship. The entity and relationship annotations were created in parallel, partially based on the specialized synonyms dictionary for annotating entities and a set of available pattern templates for identifying positive event relationship (Table 6). The annotation of hPP corpus consumed 10 months including time spent on related software modules developments and the design of annotation scheme.

Pattern templates	Annotation rule
<K> to <S> PK <P>	KINASE to SITE phosphorylate SUBSTRATE
<K> <S> PK	KINASE SITE auto-phosphorylate(ion)
<K> PK <S> of ?<P>	KINASE phosphorylate(s)/(ed) SITE of ?SUBSTRATE
<K> PK <P> at ?<S>	KINASE phosphorylate(s)/(ed) SITE of ?SUBSTRATE
<P> PK <S> by ?<K>	SUBSTRATE phosphorylate(s)/(ed) SITE by ?KINASE
<P> PK <K> on ?<S>	SUBSTRATE phosphorylated KINASE on ?SITE
<P> PK on ?<S>	SUBSTRATE phosphorylate(s)/(ed) on ?SITE
<S> PK <P>	SITE phosphorylate(s)/(ed) PROTEIN
<S> PK <K>	SITE phosphorylate(s)/(ed) KINASE
<S> of <P> PK <K>	SITE of SUBSTRATE phosphorylate(s)/(ed) KINASE
<K> PK ?<S>	KINASE phosphorylate(s)/(ed) ?SITE
<P> PK ?<S>	SUBSTRATE phosphorylate(s)/(ed) ?SITE
<P> became <S> PK by <K>	SUBSTRATE became SITE phosphorylated by KINASE
PK <S> by ?<K>	Phosphorylated SITE by ?KINASE
PK <P> by ?<K>	Phosphorylated SUBSTRATE by ?KINASE
PK <K> by <P>	Phosphorylated KINASE by SUBSTRATE

PK-phosphorylation keyword; P-substrate; K-protein kinase; S-phosphorylation site

Table 6: Pattern templates and annotation rules for event relationship.

Results and Discussion

Corpus Dataset

The two popular protein phosphorylation databases namely Phospho.ELM [18] and PhosphoSitePlus [20] are utilized for collecting PubMed abstract IDs related to human protein phosphorylation. Phospho.ELM contains 2,530 PubMed IDs with 5,374 proteins and 37,144 phosphorylation sites (27,421 serine, 6,256 threonine and 3,467 tyrosine residues) related to human protein phosphorylation information. PhosphoSitePlus consists of 4,902 PubMed IDs with 1,725 proteins and 4,820 phosphorylation sites (2,826 serine, 963 threonine and 1,031 tyrosine residues) related to human protein phosphorylation information. The statistics of human related phosphorylation data available in both databases is given in Table 7.

Source	Total entries	Proteins	Phosphorylation sites			PubMed IDs
			Serine	Threonine	Tyrosine	
Phospho.ELM	37,145	5,374	27,421	6,256	3,467	2,530
PhosphoSitePlus	4,822	1,725	2,826	963	1,031	4,902

Table 7: Statistics of human related phosphorylation data.

The combined list of 6,232 PubMed IDs (without duplicate entries) from both databases is ranked with the likelihood function of GeneTag [23]. We randomly selected 1,000 abstract IDs from top 3,000 entries for downloading the abstracts from PubMed database (http://www.ncbi.nlm.nih.gov/pubmed) for corpus creation. The random selection is executed for 100 times to select the most frequently sampled abstracts. The current version of hPP corpus is a sentence-based corpus containing phosphorylation related sentences from the selected list of 1,000 MEDLINE abstracts.

Corpus Annotation

The main objective of hPP corpus is to provide a gold-standard data for developing and evaluating text mining systems to extract the phosphorylation event information existing between substrate, kinase and phosphorylation site. This focus influences the entity annotation to tag the entities which are relevant to the phosphorylation information Tables 8&9. For instance, proteins relevant to phosphorylation event are tagged specifically as substrate or kinase. Additionally, the corpus tags the non-relevant protein entities to provide related information such as PPI (Figure 3).

	Total number	Average sentences/ abstract
Abstract	1,000	-
Total sentences	14,899	14.89
Sentences with phosphorylation keyword	5,870	5.87
Sentences related to phosphorylation event	2,381	2.38

Table 8: Basic statistics of hPP corpus.

Entity annotation	Protein kinase	Substrate	Phosphorylation site
	2,075	2,480	1,730
Event annotation	One entity	Two entities	Three entities
	576	1,605	648

Table 9: Statistics on annotated entities and phosphorylation event.

The event annotation captures a stated relationship between phosphorylation related biomedical entities identified during entity annotation. The entity and event annotation together explicitly reflect the phosphorylation information stated in a sentence in a more structured way. The entity mentions in an event annotation vary diversely across the corpus sentences. Figure 3 presents a complete event annotation where all three entities (substrate, kinase and site) together provide a phosphorylation event. However, the corpus includes a wide range of various combinations of entities representing phosphorylation event. For instance, the kinase responsible for phosphorylation event is not available in Example 4 and the information on phosphorylated site alone is provided in Example 5 without any substrate and kinase information. Likewise, the phosphorylation related keyword is absent and the equivalent keyword indicating the phosphorylation event is 'substrate' as shown in Example 6.

Example 6: PMID: 2843348
<SUBSTRATE> Calmodulin </SUBSTRATE > is a substrate for

the <KINASE> insulin receptor kinase </KINASE>.

Example 7: PMID: 15849194 Tagged sentence: <KINASE> p21 -activated protein kinases </KINASE>
(<KINASE> Paks </KINASE>) are serine/threonine protein kinases that phosphorylate <KINASE> Raf-1 </KINASE> at <SITE> Ser-338 </SITE> and <SITE> Ser-339 </SITE>.

A very closely related biological process appears in combination with protein phosphorylation is protein-protein interaction. We incorporated the annotation on protein-protein interaction information together with phosphorylation event annotation to explore the correlation between protein-protein interaction and phosphorylation using information retrieval systems (Figure 3). On the other hand, gold standard corpora (AIMED and HPRD50) available to evaluate protein-protein interaction extraction systems contain very few annotations on protein phosphorylation.

We considered the synonym and abbreviation related to a protein name are annotated as equal. For instance, the kinase names, casein kinase 1 epsilon and CK1 epsilon are annotated as same entry (Figure 3). Identification of negative event information is an important key to avoid the extraction of false phosphorylation information by a text mining system. We annotated the sentences with negative phosphorylation information within tags <NEGATIVE PHOSPHORYLATION> and </NEGATIVE> as shown in Example 8.

Example 8: PMID: 11855836
<NEGATIVE PHOSPHORYLATION> Our results suggest that <KINASE> Akt </KINASE> is a negative regulator of <SUBSTRATE> FANCA </SUBSTRATE> phosphorylation. </NEGATIVE>.

Corpus Statistics

The basic statistics of hPP corpus is listed in Table 8. The information on the number of annotated entities and phosphorylation information are listed in Table 9. Among 5,264 annotated protein entities, 2,480 are identified as substrates and 2,075 are identified as kinases. The phosphorylation information annotation identified 2,829 phosphorylation information in 2,381 sentences from 1,000 MEDLINE abstracts. Among these, 576 information include any one phosphorylation related entity, 1,605 include two entities and 648 include three entities. Most of the phosphorylation information 1,605 (56.7%) in hPP corpus are contributed by two entities patterns (substrate and kinase, substrate and site, kinase and site). The number of phosphorylation information by one entity pattern 576 (20.4%) and three entities patterns 648 (22.9%) are almost the same.

Comparison with Available Corpora

A public resource of existing biomedical corpora and benchmarks is maintained at http://www2.informatik.hu-berlin.de/~hakenber/links/benchmarks.html. The collection comprises of 39 corpora which are primarily intended for biomedical NLP. While ten corpora among the collection are available for NER evaluation, GENIA corpus [33] and JNLPBA [11] are specifically available for human genes/proteins. However, the protein annotation in the available corpora is general and does not distinguish protein kinases from other proteins. In phosphorylation event extraction, distinguishing protein kinases from other proteins is mandatory.

PPI relation corpora such as AIMED [12] and HPRD50 [13] consists of a fewer number of annotated sentences related to protein phosphorylation. The corpora available for protein phosphorylation information extraction are the feature evidence resource of iProLink [14], PTM phosphorylation extraction corpus (https://research.bioinformatics.udel.edu/iprolink/corpora.php) and another corpus for protein phosphorylation (http://pir.georgetown.edu/pirwww/iprolink/ftcorpora.shtml). All the corpora are from PIR. The iProLink corpus was originally developed for evaluating a phosphorylation information extraction tool called RLIMS-P and consists of 59 MEDLINE abstracts related to protein phosphorylation. Later, PTM phosphorylation extraction corpus was released by PIR for evaluating RLIMS-P v.2 and consists of 150 MEDLINE abstracts and 105 full-length articles. In addition to iProLink, PIR has developed five literature corpora of evidence tagging for protein functional features related to post-transcriptional modification of proteins (acetylation, glycosylation, methylation, phosphorylation and hydroxylation). Each corpus consists of two types of dataset, one with abstracts and the other with full-length articles. The corpus for protein phosphorylation consists of 186 abstracts and 76 full-length articles (http://pir.georgetown.edu/pirwww/iprolink/ftcorpora.shtml) with annotations on feature lines having phosphorylation information. BioNLP shared task corpus is an event extraction corpus that includes nine different events and phosphorylation is one of the events [34,35]. Many text mining systems for extracting phosphorylation event extraction were evaluated using this corpus [36,37]. However, phosphorylation event extraction and phosphorylation information extraction are not the same [38]. The biological processes such as gene expression, transcription, and localization are events and their extraction from published literature is termed as event extraction. The relationship between the entities or the events themselves is defined as information extraction [39]. hPP corpus is meant for phosphorylation information extraction.

There are several aspects that make hPP corpus unique from

the available corpora on NER and protein phosphorylation:

(i) The corpus provides a distinct annotation of biomedical named entities (substrate, kinase and phosphorylation site) related to protein phosphorylation. Such annotation is useful to evaluate the system on identifying and distinguishing kinases and substrates, though both are proteins in general. Identification of phosphorylation related entities such as kinase, substrate and site is a prerequisite for extracting phosphorylation information by a text mining system [38]. The existing corpora do not highlight the entities within the sentences conveying phosphorylation information.

(ii) The corpus further provides annotations of gene/protein names not involved in the phosphorylation information given. Though many corpora are available to evaluate systems on gene or protein mentions, none is available to distinguish kinases and substrates involved in protein phosphorylation from other proteins.

(iii) The current version of the corpus is a dataset of sentences with annotation on phosphorylation information, in addition to NER.

(iv) As the name indicates, hPP corpus is specifically developed for evaluating text mining system related to human protein phosphorylation. The two available corpora from PIR are general corpora (not specific to human).

Application of hPP Corpus

We developed a text mining system for extracting protein phosphorylation information from PubMed articles [38]. Our hybrid approach includes NLP (Natural Language Processing) parsing for identifying phosphorylation information patterns and Support Vector Machines (SVM), a machine algorithm for classifying the extracted phosphorylation information as true or false. We used annotations from 300 PubMed abstracts of hPP corpus to train and test our proposed system: 214 sentences from 200 PubMed abstracts as training data and 207 sentences from 100 PubMed abstracts as test data. We also used the existing corpora PLC and iProLink for validating the performance of our system. hPP corpus provided insights on protein phosphorylation patterns specific to human. Such patterns were absent in the existing corpora. Thus, a corpus specific to human protein phosphorylation is highly preferred.

Phosphorylation in Cellular Functions

Oxidative phosphorylation is the fourth step in cellular respiration that produces most of the energy in the form of adenosine triphosphate [40]. Phosphorylation is an important regulatory mechanism in cell. It regulates protein functions and cell signaling by causing conformational changes in substrate protein. The protein kinase catalyze the conformational changes in substrate protein and induces cellular transduction signaling. The hyperactivity, malfunction, or overexpression of cellular transduction signaling is found in several diseases including cancer [41]. Phosphorylation is necessary for cells to respond to stress or external stimuli that leads to cell death [42]. Thus, phosphorylation is important for various cellular activities and hPP corpus is useful for evaluating text mining systems meant for the extraction of human protein phosphorylation information from PubMed abstracts.

Conclusion

We have developed hPP corpus with linguistically rich entity and event annotations to facilitate text mining researches related to human protein phosphorylation. To our best knowledge this is the first and foremost annotated corpus available for human protein phosphorylation. The entity annotation includes three biomedical entities namely substrate protein, protein kinase and phosphorylation site related to protein phosphorylation. The event annotation provides the phosphorylation relationship between the annotated entities. The corpus is freely available for download and usage. We believe that hPP corpus will be used by many biomedical researchers to explore literature information on human protein phosphorylation for understanding cellular functions.

Competing Interests: None

Funding: No funding agencies to be reported

Authors' contributions

KR and JN designed the study. KR performed automated annotation of proteins and protein kinases. KR developed the text mining approach to annotate sites automatically. KR, DS and SA compiled the corpus after manual annotation by the hired experts. KR made the corpus to be available in BioC format. KR and JN wrote the manuscript. All the authors read and approved the contents of the manuscript.

Acknowledgements: Not applicable

References

1. Apweiler R, Bairoch A, Wu CH, Barker WC, Boechmann B, et al. (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Research 32(S1): D115-D119.

2. Wu CH, Yeh L-SL, Huang H, Arminski L, Castro-Alvear J, et al. (2003) The Protein Information Resource. Nucleic Acid Research. 31(1): 345-347.

3. Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, et al. (2005) PRIDE: The proteomics identifications database. Proteomics 5: 3537-3545.

4. Smith AC, Blackshaw JA, Robinson AL (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. Nucleic Acid Research 40(D1): D1160-D1167.

5. Flicek P, Amode MR, Barrell D, Ensembl (2012) Nucleic Acid Research. 40: D84-D90.

6. Schellenberger J, Park JO, Conrad TC, Palsson BO (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11: 213.

7. Solomona BD, Nguyenb AD, Beara KA, Wolfsberg TG (2013) Clinical Genomic Database. In Proceedings of the National Academy of Sciences of United States of America 110(24): 9851-9855.

8. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH (2002) Accomplishments and challenges in literature data mining for biology. Bioinformatics 18(12): 1553-1561.

9. Shatkay H, Feldman R (2003) Mining the biomedical literature in the genomic era: An overview. Journal of Computational Biology 10(35): 821-855.

10. Cohen KB, Hunter L (2004) Natural language processing and systems biology. Artificial intelligence and systems biology 5: 147-175.

11. Kim JD, Ohta T, Tsuruoka Y, Tateisi Y, Collier N (2004) Introduction to the Bio-Entity Recognition Task at JNLPBA. In Proceeding of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA- 2004), Geneva, Switzerland pp: 70-75.

12. Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, et al. (2005) Comparative experiments on learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine 33: 139-155.

13. Fundel K, Kuffner R, Zimmer R (2007) RelEx–relation extraction using dependency parse trees. Bioinformatics 23(3): 365-371.

14. Hu ZZ, Mani I, Hermoso V, Liu H, Wu CH (2004) iProLINK: an integrated protein resource for literature mining. Computational Biology and Chemistry 28: 409-416.

15. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker C, Wu CH (2005) Literature mining and database annotation of protein phosphorzlation using a rule-based system. Bioinformatics 21(11): 2759-2765.

16. Cohen P (2002) The Origins of Protein Phosphorylation. Nature Cell Biology. 4(5): E127-E130.

17. Thomason P, Kay R (2000) Eukaryotic Signal Transduction via Histidine-Aspartate Phosphorelay. Journal of Cell Science 113(pt 18): 3141-3150.

18. Holger D, Claudia C, Allegra V, Gould CM, Jensen LJ, et al. (2011) Phospho.ELM: a database of phosphorylation sites-update 2011. Nucleic Acids Research 39: 261-267.

19. Kreegipuu A, Blom N, Brunak S (1999) PhosphoBase, a Database of Phosphorylation Sites: Release 2.0. Nucleic Acids Research 27(1): 237-239.

20. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, et al. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined posttranslational modifications in man and mouse. Nucleic Acids Research 40(D1): D261-D270.

21. PhosphoNET.

22. Yang CY, Chang CH, Yu YL, Lin TC, Lee SA, et al. (2008) PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. Bioinformatics 24(16): i14-i20.

23. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 6(1): S3-S9.

24. Raja K, Subramani S, Natarajan J (2014) A hybrid named entity recognition for tagging human proteins/genes. International Journal of Data Mining and Bioinformatics 10(3): 315-328.

25. Hu ZZ (2004) Guidelines for protein name tagging, version 2.0. pp: 2-10.

26. Klein D, Manning CD (2003) Accurate unlexicalized parsing. In Proceedings of the forty-first Meeting of the Association for Computational Linguistics. Morristown, NJ, USA, pp: 423-430.

27. Subramani S, Raja K, Natarajan J (2014) ProNormz--an integrated approach for human proteins and protein kinases normalization. Journal of Biomedical Informatics pp: 47:131-138.

28. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, et al. (2008) Overview of BioCreative II gene normalization. Genome Biology 9(2): S3.

29. UniProt Consortium (2008) The Universal Protein Resource (UniProt). Nucleic Acids Research 36(Database

issue): D190-195.

30. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complements of the human genome. Science 298(5600): 1912-1934.

31. Milanesi L, Petrillo M, Sepe L, Boccia A, D'Agostino N, et al. (2005) Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity. BMC Bioinformatics 6(4): S20.

32. Xu Y, Teng D, Lei Y (2012) MinePhos: A Literature Mining System for Protein Phoshphorylation Information Extraction. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(1): 311-315.

33. Kim JD, Ohta T, Tateisi Y, Tsujii J (2003) GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics 19: i180-182.

34. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J (2006) Overview of BioNLP'09 shared task on event extraction. pp: 1-9.

35. Nedellec C, Kim JD, Ohta T, Zweigenbaum P, Bossy R, et al. (2013) Overview of BioNLP Shared Task 2013. In Proceedings of the BioNLP Shared Task 2013 Workshop pp: 1-7.

36. Bjome J, Ginter F, Pyysalo S, Tsujii J, Salakoski T (2010) Complex event extraction at PubMed scale. Bioinformatics 26(12): i382-i390.

37. Miwa M, Saetre R, Kim JD, Tsujii J (2010) Event extraction with complex event classification using rich features. Journal of Bioinformatics and Computational Biology 8(1): 131-146.

38. Raja K, Natarajan J (2018) Mining Protein Phosphorylation information from Biomedical Literature using NLP Parsing and Support Vector Machines. Computer Methods and Programs in Biomedicine 160: 57-64.

39. Raja K, Subramani S, Natarajan J (2013) Template filling, text mining. In Encyclopedia of Systems Biology, Springer, New York, USA, pp: 2150-2154.

40. Zhang D, Ding G, Ge B, Zhang H, Tang B (2017) Molecular evolution of mitochondrial coding genes in the oxidative phosphorylation pathway in malacostraca: purifying selection or accelerated evolution? Mitochondrial DNA. Part A, DNA mapping, sequencing and analysis 28(4): 593-596.

41. Ardito F, Giuliani M, Perrone D, Troiano G, Muzio LL (2017) The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int J Mol Med 40(2): 271-280.

42. Fulda S, Gorman AM, Hori O, Samali A (2010) Cellular Stress Responses: Cell Survival and Cell Death. Cell Stress and Cell Death 2010: 214074.