

Data and text mining

Literature mining and database annotation of protein phosphorylation using a rule-based system

Z. Z. Hu^{1,*}, M. Narayanaswamy², K. E. Ravikumar², K. Vijay-Shanker³ and C. H. Wu¹¹Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, Washington, DC 20057, USA, ²AU-KBC Research Centre, Anna University, Chennai 600044, India and ³Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716, USA

Received on February 1, 2005; revised on March 10, 2005; accepted on March 11, 2005

Advance Access publication April 6, 2005

ABSTRACT

Motivation: A large volume of experimental data on protein phosphorylation is buried in the fast-growing PubMed literature. While of great value, such information is limited in databases owing to the laborious process of literature-based curation. Computational literature mining holds promise to facilitate database curation.

Results: A rule-based system, RLIMS-P (Rule-based Literature Mining System for Protein Phosphorylation), was used to extract protein phosphorylation information from MEDLINE abstracts. An annotation-tagged literature corpus developed at PIR was used to evaluate the system for finding phosphorylation papers and extracting phosphorylation objects (kinases, substrates and sites) from abstracts. RLIMS-P achieved a precision and recall of 91.4 and 96.4% for paper retrieval, and of 97.9 and 88.0% for extraction of substrates and sites. Coupling the high recall for paper retrieval and high precision for information extraction, RLIMS-P facilitates literature mining and database annotation of protein phosphorylation.

Availability: The program is available on request from the authors. The phosphorylation patterns and datasets used in this study are available at <http://pir.georgetown.edu/iprolink/>

Contact: zh9@georgetown.edu

1 INTRODUCTION

Phosphorylation is one of the most common post-translational modifications (PTMs) for proteins and is involved in numerous biological processes (Cohen, 2002). Detection of the dynamic phosphorylation state of the cellular proteome is essential for understanding the regulatory network of biological pathways. Protein phosphorylation information is provided in several protein databases, including UniProt Knowledgebase—the central database of protein sequence and function (Apweiler *et al.*, 2004), as well as specialized databases, such as Phospho.ELM (Diella *et al.*, 2004) and Phosphorylation Site Database (<http://vigen.biochem.vt.edu/xpd/xpd.htm>). Phospho.ELM contains 556 eukaryotic protein entries, covering 1703 experimental phosphorylation sites manually curated from literature; while Phosphorylation Site Database consists of 97 prokaryotic protein entries compiled from literature. Overall, such experimental data are limited in databases which have not kept up with the fast-growing literature. With an ever-increasing volume of scientific literature now available

electronically, there is both a pressing need and a great opportunity to develop more efficient ways for literature data mining. Indeed, in recent years, natural language processing technologies are being utilized for biological literature mining and information extraction (Hirschman *et al.*, 2002), such as the PreBIND system for mining protein–protein interactions from literature (Donaldson *et al.*, 2003).

Recently, our group has developed a resource for literature mining, iProLINK, which provides a bibliography system and curated data sources for training and benchmarking text mining algorithms (Hu *et al.*, 2004). In particular, it includes literature corpora that are tagged with experimental features annotated in the PIR–PSD database (Wu *et al.*, 2003a). Literature tagging is part of the evidence attribution mechanism at Protein Information Resource (PIR) (Wu *et al.*, 2003b), which is being integrated into the UniProt Knowledgebase. The attribution distinguishes experimental from computational annotations, and provides both citation mapping (finding citations from the Reference section that describe the given experimental feature) and evidence tagging (tagging the sentences providing experimental evidence in an abstract and/or full text article). There are nearly 10 000 experimental features annotated in PIR–PSD, including over 2000 corresponding to five common PTMs—phosphorylation, acetylation, glycosylation, methylation and hydroxylation. Another work that systematically utilizes literature from the curated protein database for benchmarking text mining techniques is the BioMINT (<http://www.biomint.org/>) biological text mining system.

Here we report literature mining for protein phosphorylation using RLIMS-P (Rule-based Literature Mining System for Protein Phosphorylation), a rule-based system developed based on the algorithm of Ravikumar *et al.* (2004). The system utilizes shallow parsing and extracts phosphorylation information by matching text with manually developed patterns. Similar rule/pattern-based approaches have been used in information extraction (Blaschke *et al.*, 1999; Ng and Wong, 1999; Pustejovsky *et al.*, 2002; Rindfleisch *et al.*, 1999; Sekimizu *et al.*, 1998). Other approaches employed for extracting protein–protein interactions have been based on detecting co-occurring proteins (Proux *et al.*, 2000; Stapley and Benoit, 2000; Stephens *et al.*, 2001) in literature or using a parser tailored for the specialized language typically found in the biology literature (Friedman *et al.*, 2001; Yakushiji *et al.*, 2001; Park *et al.*, 2001). The RLIMS-P literature mining system was benchmarked using the iProLINK annotation-tagged corpus as a benchmark standard, and the results were evaluated by PIR curators.

*To whom correspondence should be addressed.

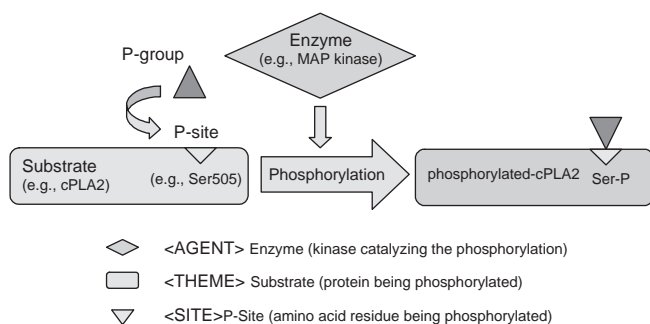


Fig. 1. Objects in the phosphorylation process.

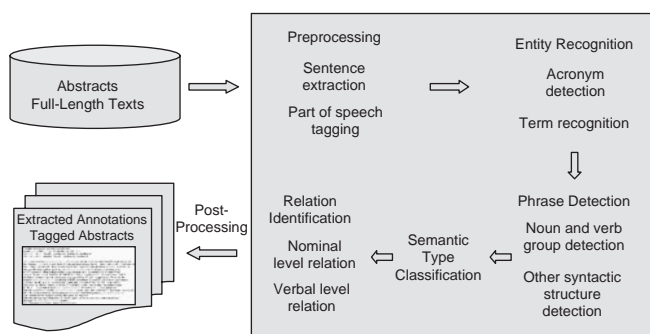


Fig. 2. System architecture of the RLIMS-P literature mining system.

2 SYSTEMS AND METHODS

2.1 Phosphorylation objects

Information extraction and database annotation of protein phosphorylation require the identification of three objects that play key roles in the phosphorylation process (Fig. 1): ‘enzyme’ (kinase that phosphorylates proteins), ‘substrate’ (protein that is phosphorylated) and ‘site’ (phosphorylated residue). Correspondingly, the RLIMS-P system is designed to detect and extract these three types of objects from MEDLINE papers, and assign them to argument roles named <AGENT>, <THEME> and <SITE>, respectively.

Below is an example text from a MEDLINE abstract that describes the three objects (italicized) of the phosphorylation process (Fig. 1):

Full-length *cPLA2*<THEME> was phosphorylated stoichiometrically by *p42 mitogen-activated protein (MAP) kinase*<AGENT> *in vitro* . . . and the major site of phosphorylation was identified by amino acid sequencing as *Ser505*<SITE>. [PMID: 8706669]

2.2 The RLIMS-P architecture

Figure 2 shows the overall architecture and system components of the RLIMS-P literature mining system. During preprocessing, the text is first split into sentences and tokenized into words and punctuation; the words are then assigned part-of-speech (POS) tags (adverbs, verbs, adjectives, etc.). The system currently uses the Brill’s tagger (Brill, 1995). This tagger makes a few types of mistakes in the biology domain that are corrected by our supplemental rules. For example, one common error was to tag words ending with ‘ed’ (e.g. ‘conserved’ in the noun phrase ‘highly conserved regions’) or ‘ing’ (e.g. ‘splicing’ in ‘alternative splicing factor’) as verbs in past-tense or gerundive even though the context suggested the adjectival sense. We added a rule that changes such verb tags to adjectival whenever we find a sequence of a determiner followed by zero or more adjectives/adverbs followed by such a (mis)tagged word ending with ‘ed’ or ‘ing’. In addition, since many words in biology domain are not found in Brill tagger’s training corpus

(of business newspaper articles), it sometimes mistagged as plural nouns or present-tense verbs words ending with ‘s,’ e.g. ‘phosphorylates’ as a plural noun and ‘kinases’ as a present-tense verb.

The system component for named entity recognition includes acronym detection and term recognition, as described previously (Narayanaswamy et al., 2003).

2.3 Phrase detection

The ‘phrase chunker’ identifies various phrases within a sentence including the so-called BaseNP—simple noun phrases that do not include another noun phrase. BaseNP detection involves using the POS tags of words that usually appear at the boundaries. The system also detected other types of phrases to match text with patterns at a higher level of syntactic abstraction than many other pattern-matching methods. For instance, the phrase ‘was found to be able to phosphorylate’ in the following sentence was detected as a sequence of verb group chunks. The sentence can then be matched with a simple pattern ‘<AGENT> phosphorylate <THEME> at <SITE>’ (the three arguments are italicized).

Active *p90Rsk2* was found to be able to phosphorylate *histone H3* at *Ser10*.

RLIMS-P treats all words between ‘p90Rsk2’ and ‘phosphorylate’ as part of a verb group sequence and, hence detects p90Rsk2 as the subject. Similar to the grammar/parser-based approaches, this syntax recognition allows one to succinctly capture the patterns between the verb and its subject. Furthermore, ‘was found,’ ‘to be able’ and ‘to phosphorylate’ were individually detected allowing the detection of the voice of ‘to phosphorylate’—that it is in active form, not in passive form. Therefore, the syntactic subject corresponds to the <AGENT>, rather than the <THEME> (the phosphorylated protein).

Other syntactic constructions related to noun phrases are recognized, including entity coordination and appositives. The detection of these constructs serves mainly to correctly identify the location of the arguments. Consider the following sentence, where identifying appositives helps find the right arguments, as with relative clauses.

In the yeast *Saccharomyces cerevisiae*, *Sic1*, an inhibitor of Clb-Cdc28 kinases, must be phosphorylated and degraded in G 1 for cells to initiate DNA replication, . . .

Here, ‘Sic1’ must be extracted as the <THEME>, which requires the recognition that the noun phrase ‘an inhibitor of Clb-Cdc28 kinases’ is in apposition to ‘Sic1’ and hence must be skipped over when matching with a pattern.

Using such phrase chunking and syntax processing, the system is able to obtain many of the advantages of a full-parsing approach without using a complex grammar. Similar approaches are employed in well-known information extraction systems like FASTUS (Hobbs et al., 1997) to develop general-purpose, domain-independent information extraction tools. As in FASTUS, we use a few standard patterns of POS tags for sequences of tokens in order to detect BaseNP chunks, Verb group chunks and phrases in apposition. These patterns were tested for their effectiveness using several MEDLINE abstracts that we had previously marked up for such chunking.

2.4 Semantic type classification

One of the key components of the RLIMS-P system is the assignment of semantic types to noun phrases. Semantic type assignment which was also employed in information extraction systems as outlined in Rindfleisch et al. (1999, 2000), Pustejovsky et al. (2002), simplifies pattern specification and improves the precision. Consider the following:

- (1) ATR/FRP-1 also phosphorylated *p53* in *Ser 15* . . .
- (2) Active Chk2 phosphorylated *the SQ/TQ sites* in *Ckk2 SCD* . . .
- (3) cdk9/cyclinT2 could phosphorylate *the retinoblastoma gene (pRb)* in *human cell lines*

While all three examples match this same syntactic pattern ‘X phosphorylated Y in Z’, the relation extracted will depend on what matches Y and Z (the italicized phrases). These would correspond to the <THEME>

and <SITE> in the first example, <SITE> and <THEME> in the second example and only <THEME> in the third example. If the patterns were merely syntactic and did not include type information, then the relation extracted would be correct in only one example.

These examples indicate that noun phrases (NP) must be classified as to whether they are of type protein (appropriate for the role of enzyme <AGENT> or substrate <THEME>), amino acid residue (for <SITE>) or cells, tissues, etc. (for source). In the RLIMS-P system, NPs are classified into these types or 'others'. Based on our previous work (Narayanaswamy *et al.*, 2003), the classification uses lexical information in the form of informative words that appear as head words (e.g. 'mitogen activated protein kinase' is classified as a protein because of its head word 'kinase'), suffixes and nearby phrases. The manually selected list of such words/suffixes/phrases (Narayanaswamy *et al.*, 2003) has been augmented by the method described in Torii *et al.* (2004). Additional rules and heuristics are employed based on detecting acronym–full form pairs where the full form contains lexical information pertinent to classification (e.g. 'mitogen activated protein kinase' and 'MAPK' pair), appositives and conjunction of entities. More details on the use of both clues, internal to the name and contextual words for type assignment, can be found in Narayanaswamy *et al.* (2003).

2.5 Rule-based relation identification: pattern templates and argument mapping

Pattern templates were manually created after examining a development text corpus of 300 MEDLINE abstracts and 10 journal articles and observing the different forms used to describe phosphorylation interactions.

2.5.1 Verbal forms The verbal inflected forms 'phosphorylate/phosphorylated/phosphorylating/phosphorylates' are captured in various patterns. Below are example patterns that illustrate the different orders and optionality of the three arguments <AGENT>, <THEME> and <SITE>.

Pattern 1: <AGENT> <VG-active-phosphorylate> <THEME> (in/at <SITE>)? where 'VG' denotes verb group and '?' denotes optional argument.

This pattern is read as requiring an NP of the type appropriate for <AGENT> appearing to the left of verb groups. The head of the main verb group must be in the inflected form 'phosphorylate' or 'phosphorylated' and have an active voice. Furthermore, it requires that an NP of type appropriate for <THEME> appears to the right of this verb group. Finally, an NP with type appropriate for <SITE> in a prepositional phrase with 'in' or 'at' can be 'optionally' matched. Clauses matched by this pattern include 'ATR/FRP-1 also phosphorylated p53 in Ser 15' as well as 'The recombinant protein was shown to phosphorylate Kemptide' but not 'Active Chk2 phosphorylated the SQ/TQ sites in Ckk2 SCD'. The order of <SITE> and <THEME> in the latter is captured by another pattern.

When a clause matches the pattern, the phrases that match the arguments <AGENT>, <THEME> and <SITE> are extracted and assigned to the corresponding argument slots.

Pattern 2: <THEME> <VG-passive-phosphorylated> by <AGENT>

This is a commonly observed pattern, in which 'phosphorylated' appears in passive form.

Several patterns for the passive usage capture the situation where the <THEME> and/or the <SITE> can appear in the syntactic subject position. The inflected form 'phosphorylated' appears quite often as a noun modifier (as in 'the phosphorylated protein', 'the phosphorylated site' or even with the amino-acid as in 'tyrosine-phosphorylated pRb').

2.5.2 Nominal form While many previous information extraction projects have concentrated only on the verbal forms of interactions, patterns for the nominal form in the case of 'phosphorylate' interactions is needed. Indeed, 'phosphorylation' was the most frequent inflected form. Furthermore, the patterns of occurrences of the arguments are most varied for this form.

The <THEME> can appear before 'phosphorylation' as in 'vitronectin phosphorylation by the kinase'. When an argument appears before 'phosphorylation', typically it is the <THEME>. When <AGENT> appears before 'phosphorylation', its role is normally indicated clearly with the <THEME> appearing after 'phosphorylation of' as in:

Pattern 3: [<AGENT> phosphorylation]_{NP} of <THEME>

The <AGENT> and <THEME> can also appear after 'phosphorylation' as captured by the following pattern:

Pattern 4: phosphorylation of <THEME> (by <AGENT>)? (in/at <SITE>)?

Some patterns for phosphorylation are even more complicated, such as:

Pattern 5: <AGENT> <VG-active> <THEME> by/via phosphorylation at (<SITE>)?

The pattern matches with 'Both kinases also inactivate spinach sucrose phosphate synthase via phosphorylation at Ser-15', capturing the fact that 'inactivation' and 'phosphorylation' have the same arguments (i.e. both kinases) because inactivation is the result of phosphorylation. (A simple anaphora resolution program has been implemented that will attempt to resolve anaphoric expressions, such as 'both kinases').

3 IMPLEMENTATION

3.1 Datasets

The datasets for testing and benchmarking the RLIMS-P literature mining system were derived from data sources in iProLINK. Specifically, we used the annotation-tagged literature corpora that were developed for evidence attribution of experimental phosphorylation features annotated in PIR-PSD (Fig. 3). There were two types of data corresponding to the two tasks for evidence attribution—citation mapping and evidence tagging. Citation mapping involves finding the specific paper(s) describing a given phosphorylation feature of a protein entry from a list of papers in the PSD Reference section (Fig. 3A). Evidence tagging involves tagging the sentences providing experimental phosphorylation evidence in the abstract and/or full-text of the paper, which may include information of <THEME>, <SITE> and <AGENT> (Fig. 3B).

Here, the tagged sentence:

'Phosphorylation of bovine brain PLC-beta by PKC in vitro resulted in a stoichiometric incorporation of phosphate at serine 887, without any concomitant effect on PLC-beta activity'.

contains information on <THEME>bovine brain PLC-beta, <AGENT>PKC and <SITE>serine 887, thereby, providing evidence for the experimental feature line where the site is phosphate (Ser) at sequence position 887, the enzyme is protein kinase C (i.e. PKC) and the feature is for the PIR 'ENTRY A28822' for '1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase I' (also known as Phospholipase C-beta-1) from 'bovine' (i.e. brain PLC-beta).

The citation mapping data were used to evaluate the ability of the system to identify papers describing phosphorylation information, i.e. performance for 'information retrieval' (IR). The dataset provides a direct mapping of protein IDs, phosphorylation features and the PubMed ID (PMID) of papers describing experimental phosphorylation features (positive papers). All other papers in the Reference sections of the corresponding PIR entries not describing phosphorylation features were flagged as negative papers.

<pre> ENTRY A28822 #type complete TITLE 1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase (EC 3.1.4.11) I - bovine ALTERNATE_NAMES 1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase beta; phosphoinositide-specific phospholipase C, PLC-154; triphosphoinositide phosphodiesterase, PLC-154 ... REFERENCE A28822 #authors Katan, M.; Kriz, R.W.; Totty, N.; Philp, R.; Meldrum, E.; Aldape, R.A.; Knopf, J.L.; Farker, F.J. #journal Cell (1988) 54:171-177 #title Determination of the primary structure of PLC-154 demonstrates diversity of phosphoinositide-specific phospholipase C activities. #cross-references MUID:88270496; PMID:2455601 #accession A28822 ##molecule_type mRNA ##residues 1-1216 ##label KAT ##cross-references UNIPROT:PI0894; UNIPARC:UPI0000131ABE; GB:J03137; MID:g163521; FIDN:AAA30702.1; FID:g163522 REFERENCE A39236 #authors Ryu, S.H.; Kim, U.H.; Wahi, M.I.; Brown, A.B.; Carpenter, G.; Huang, K.P.; Rhee, S.G. #journal J. Biol. Chem. (1990) 265:17941-17945 #title Feedback regulation of phospholipase C-beta by protein kinase C. #cross-references MUID:91009263; PMID:2211670 #accession A39236 ##molecule_type protein ##residues 879-889 ##label RYU ##cross-references UNIPARC:UPI0000175999 ... FEATURE 318-467 #domain 1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase domain X homology #label FIPX 539-659 #domain 1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase domain Y homology #label FIPY <88> #binding site phosphate (Ser) (covalent) by protein kinase C #status experimental PMID:2211670 </pre>	A
<pre> >A28822 FT - binding site phosphate (Ser) (covalent) (by protein kinase C) 887 (all) TI - Feedback regulation of phospholipase C-beta by protein kinase C. AB - Treatment of a variety of cells and tissues with 12-O-tetradecanoylphorbol-13-acetate (TPA), an activator of protein kinase C (PKC) results in the inhibition of receptor-coupled inositol phospholipid-specific phospholipase C (PLC) activity. To determine whether or not the targets of TPA-activated PKC include one or more isozymes of PLC, studies were carried out with PC12, C6Bu1, and NIH 3T3 cells, which contain at least three PLC isozymes, PLC-beta, PLC-gamma, and PLC-delta. Treatment of the cells with TPA stimulated the phosphorylation of serine residues in PLC-beta, but the phosphorylation state of PLC-gamma and PLC-delta was not changed significantly. Phosphorylation of <i>bovine brain PLC-beta</i> by PKC in vitro resulted in a stoichiometric incorporation of phosphate at <i>serine 887</i>, without any concomitant effect on PLC-beta activity. We propose, therefore, that rather than having a direct effect on enzyme activity, the phosphorylation of PLC-beta by PKC may alter its interaction with a putative guanine nucleotide-binding regulatory protein and thereby prevent its activation. SO - J Biol Chem 1990 Oct 15;265(29):17941-5. PMID- 2211670 </pre>	B

Fig. 3. Evidence attribution of PIR-PSD experimental features. (A) Citation mapping from referenced papers to features; (B) Evidence tagging of features in abstracts.

In this study, we used only MEDLINE abstracts (titles included) obtained using PMIDs in the datasets. Note that positive papers contain information on the phosphorylation process along with at least one of the three arguments (<AGENT>, <THEME> and <SITE>). Therefore, abstracts mentioning kinases or phosphoproteins without associating with the process of phosphorylation did not constitute positive papers. For example, abstract describing ‘growth-associated protein (GAP)-43 is a neuron-specific ‘phosphoprotein’ whose expression is associated with axonal outgrowth’ [PMID: 2153895], was not considered as positive paper. Furthermore, papers containing no relevant phosphorylation information in the abstracts regardless of information in full-text were regarded as negative data for program evaluation (as RLIMS-P was applied only to abstracts in this study).

The evidence-tagged abstracts were used to evaluate the ability of the system to extract specific phosphorylation annotation, i.e. performance for ‘information extraction’ (IE). Specifically, individual phosphorylation site features that have been tagged in abstracts were positive features.

3.2 System implementation and performance measure

The RLIMS-P system is implemented in the PERL programming language and runs in a Linux operating system with a Pentium 3 processor, 40GB disk and 256MB RAM. The program runs fast—for processing 1000 abstracts, the real time was 5 min 3.24 s.

The system was evaluated based on the following performance measures:

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}); \text{recall} = \text{TP}/(\text{TP} + \text{FN})$$

where TP is true positive FP is false positive and FN is false negative.

3.3 IR evaluation: retrieval of protein phosphorylation papers

The RLIMS-P system was evaluated for IR performance in two stages, a preliminary study using a small dataset to refine the system, followed by a benchmarking study using a larger dataset. The preliminary study was conducted using a dataset of 146 abstracts, consisting of 56 positive papers and 90 negative papers. The program identified 44 true positive and 9 false positive papers, giving a precision of 83.0%. It missed 12 papers (false negative), giving a recall of 78.6%.

The false positive and negative cases were analyzed to identify areas for system improvement. Common false positives include detection of phosphorylation of non-proteins (e.g. ‘domain responsible for mannitol phosphorylation’ [PMID: 1946374]) or detection of dephosphorylation (e.g. ‘dephosphorylation of the diphosphorylated peptide on threonine and tyrosine residues’ [PMID: 7876121]). The major false negative pattern was specific phosphorylated residues of a phosphoprotein, such as phosphoserine or phosphothreonine (e.g. ‘a phosphoserine at residue 2’ [PMID: 2229609]). These phospho-residue patterns were later added to the rules.

For the benchmarking study, a larger dataset with 370 abstracts was used, including 110 positive and 260 negative papers. One hundred and sixteen abstracts were detected, with 106 true positives, giving a precision of 91.4% (Table 1). Four positive papers were missed, yielding a recall of 96.4%. The major improvement of system performance over the preliminary study was mainly because of the addition of new patterns, especially those containing phospho-residues.

The analysis of the false positive cases indicates that they often involve texts that describe general consensus sequence or predicted sites of protein phosphorylation. Examples are:

- [PMID:2223773] ‘ADF contains a sequence similar to ... a calcium/calmodulin-dependent protein kinase II phosphorylation consensus sequence’; or
- [PMID:8179334] ‘... Exon 15 is a unique exon for ... and contains the phosphorylation sites for protein kinases A and C’.

These false positives may result from a condition used in the system that focuses on finding all potential phosphorylation site information. The condition allows site information extraction (therefore retrieval of phosphorylation papers), even when the <SITE> argument does not fit any pattern if the site information lies within the same sentence of ‘phosphorylation’.

Table 1. RLIMS-P system performance for retrieving phosphorylation papers (IR) and extracting phosphorylation site information (IE)

	Benchmark Standard		RLIMS-P			Recall (%)	Precision (%)
	Positive data	Negative data	True positive	False positive	False negative		
IR: Paper retrieval (No. of abstracts)	110	260	106	10	4	96.4	91.4
IE: Theme/site extraction (No. of sites)	108	n/a	95	2	13	88.0	97.9

n/a - not assessed.

The improved system missed only four phosphorylation papers, which contained texts with some unusual patterns, such as:

- (1) [PMID:2755948] ‘the appearance of a phosphate group on 75 Ser’,
- (2) [PMID:3944083] ‘digestion with carboxypeptidase A ... where Pse represents phosphoserine’, and,
- (3) [PMID:8647113] ‘The two proteins are targets for Cdc2 kinase in meiotic maturation’.
- (4) [PMID:6311252] ‘an N-tetradecanoyl (myristyl) group blocking the NH2 terminus and phosphate groups at threonine-197 and serine-338’.

However, the fourth false negative above could have been avoided.

3.4 IE evaluation: extraction of phosphorylation information

To better evaluate how the RLIMS-P system can assist database annotation of phosphorylation features, we further analyzed the performance of the program on phosphorylation information extraction using the PIR evidence-tagged abstracts as the benchmark standard. As shown in Figure 3, the tagged sentences provide experimental evidence for the corresponding features—kinases (<AGENT>) and phosphorylated residues and their positions (<SITE>)—in the feature lines of the PIR-PSD protein entries (substrates or <THEME>).

Since positive abstracts may not always provide information on all three phosphorylation objects, the system extracted varying degrees of phosphorylation information from the abstracts. The following are three classes of annotation information extracted by RLIMS-P.

Case 1: Complete phosphorylation information on all three objects. This is the perfect case when information on the kinase, the protein substrate, and the phosphorylation residue and position are all available and extracted from the abstract by RLIMS-P (Fig. 4).

Case 2: Phosphorylation information on <THEME> and <SITE>, but not <AGENT>. Frequently, the kinases responsible for the phosphorylation are not known or not mentioned in the abstracts. Indeed, only 47 (43%) of the 110 positive abstracts from the benchmarking dataset contained the kinase information. From the 47 abstracts, the program correctly extracted the kinase information from 44 abstracts (94%). In the absence of <AGENT>, the <THEME> and <SITE>, information is still sufficient for protein feature annotation.

Case 3: Phosphorylation information on <THEME> and/or <AGENT>, but not <SITE>. When there is no explicit site information in the abstracts, annotators need to examine full-length article

<pre>ENTRY DVHU1 #type complete TITLE multidrug resistance protein 1 - human ALTERNATE_NAMES P-glycoprotein 1 ... FEATURE ... 91,94,99 #binding site carbohydrate (Asn) (covalent) #status predicted\ 433 #binding site ATP (Lys) #status predicted\ 661,667,671 #binding site phosphate (Ser) (covalent) (by protein kinase C) #status experimental\ 667,671,683 #binding site phosphate (Ser) (covalent) (by cAMP-dependent kinase) #status experimental\ 1076 #binding site ATP (Lys) #status predicted</pre>	A
<pre>PMID: 7909431 PIR: DVHU1 Fields - <AGENT> <THEME> <SITE> AN - 3: PKC PG-2 serine-661, serine-667, serine-671 AN - 3: PKA PG-2 serine-667, serine-671, serine-683 TI - Phosphorylation by protein kinase C and cyclic AMP-dependent protein kinase of synthetic peptides derived from the linker region of human P-glycoprotein. AB - Specific sites in the linker region of human P-glycoprotein phosphorylated by protein kinase C (PKC) were identified by means of a synthetic peptide substrate, PG-2, corresponding to residues 656-689 from this region of the molecule. As PG-2 has several sequences of the type recognized by the cyclic AMP-dependent protein kinase (PKA), PG-2 was also tested as a substrate for PKA. PG-2 was phosphorylated by purified PKC in a Ca2+/phospholipid-dependent manner, with a Km of 1.3 microM, and to a maximum stoichiometry of 2.9 +/- 0.1 mol of phosphate/mol of peptide. Sequence analysis of tryptic fragments of PG-2 phosphorylated by PKC identified Ser-661, Ser-667 and Ser-671 as the three sites of phosphorylation. PG-2 was also found to be phosphorylated by purified PKA in a cyclic AMP-dependent manner, with a Km of 21 microM, and to a maximum stoichiometry of 2.6 +/- 0.2 mol of phosphate/mol of peptide. Ser-667, Ser-671 and Ser-683 were phosphorylated by PKA. Truncated peptides of PG-2 were utilized to confirm that Ser-661 was PKC-specific and Ser-683 was PKA-specific. Further studies showed that PG-2 acted as a competitive substrate for the P-glycoprotein kinase present in membranes from multidrug-resistant human KB cells. The membrane kinase phosphorylated PG-2 mainly on Ser-661, Ser-667 and Ser-671. These results show that human P-glycoprotein can be phosphorylated by at least two protein kinases, stimulated by different second-messenger systems, which exhibit both overlapping and unique specificities for phosphorylation of multiple sites in the linker region of the molecule. SO - Biochem J 1994 Apr 1;299 (Pt 1):309-15.</pre>	B

Fig. 4. RLIMS-P extraction of protein phosphorylation information. (A) Curated PIR-PSD experimental features; (B) Automated extraction and tagging of <AGENT>, <THEME> and <SITE>.

or additional papers because <SITE> is needed for position-specific feature annotation.

In this IE performance evaluation, we focused on <SITE> and <THEME>, with information on both amino acid residues and their sequence positions in the context of protein substrates.

Among the 110 positive papers used in the IR benchmarking study, 59 abstracts were tagged for experimental site features, covering site residue and sequence position information for a total of 129 sites. Among the tagged sites, the positions of 21 sites were based on implicit information, such as sequence patterns rather than explicit residue numbers. Examples of derivable residue position information

in tagged sentences include:

- sequence patterns, such as ‘phosphopeptide AT(P)S(P)NVFA-MFDQSIIQEFK’ (which indicates phosphorylation of both threonine and serine residues) or
- N-terminal amino acid, such as ‘phosphorylated on the amino-terminal residue, N-acetyl-serine’.

In such cases, annotators need to verify sequence positions of the phosphorylated residues by mapping the sequence patterns onto the protein sequences in the database. Such implicit phosphorylation positions were excluded for IE evaluation because current version of the program does not provide sequence mapping.

Thus, the benchmark standard to evaluate RLIMS-P for site feature extraction consists of tagged abstracts for 108 sites (i.e. <SITE>–<THEME> pairs), where explicit sequence position is given for each site residue in the substrate. The results showed that 95 of the 108 phosphorylation sites were extracted for site residues and positions as well as the protein substrates, giving a recall of 88.0% (Table 1). The analysis of false negative results showed that the program sometimes missed multiple sites in one sentence. The current program extracted all sites in sentences if the residues are linked by conjunctions as in ‘serine residues 972, 985 and 1007 are phosphorylated by phosphorylase kinase’. But, it missed the second site (Ser-311) in ‘phosphorylation occurs at Ser-315 in the myosin IB heavy chain, Ser-311 in myosin IC’. Other false negatives include cases where correct sites were extracted but the <THEME> was not identified.

The RLIMS-P system had a high precision (97.9%) with only two false positives. The robustness of the system is illustrated below where the protein substrate and all six site residues were correctly extracted in addition to the kinase.

- (1) [PMID:2500966] ‘... we attempted to identify the sites of *vimentin* phosphorylated by each protein kinase. Sequential analysis of the purified phosphopeptides ... revealed that *Ser-8*, *Ser-9*, *Ser-20*, *Ser-25*, *Ser-33*, and *Ser-41* were specifically phosphorylated by *protein kinase C*’.

The two false positive sites occurred in sentence 2 where the text does not indicate phosphorylation of Ser24 and Thr25.

- (2) [PMID:7615564] ‘HeLa cells, transfected with either chick wild-type ADF cDNA or a cDNA mutated to code for Ala in place of Ser24 or Thr25, express and phosphorylate the exogenous ADF’.

4 DISCUSSION

We present here a rule-based system, RLIMS-P, for mining and extracting protein phosphorylation information from MEDLINE abstracts. The system has several special features that result in performance advantages over other text mining systems: it provides semantic type assignment to simplify pattern specification and improve precision; it provides phrase detection for pattern matching at a high level of syntactic abstraction; it uses patterns for both verbal and nominal forms, which are common for describing PTMs; it focuses on the specific interaction of protein phosphorylation and extracts not only the proteins involved but also the target sites. Such a special-purpose, custom-tailored approach allows the system to capture many specialized patterns and words (e.g. ‘phosphoserine’).

The RLIMS-P system takes advantage of the PIR annotated literature dataset for benchmarking of the system’s performance. The system achieves an overall recall of 96% in information retrieval and precision of 98% in information extraction. The high recall of citation mapping will ensure minimal ‘loss’ of phosphorylation papers and result in significant time saving for annotators to find relevant phosphorylation citations from long lists of papers in given protein entries. For example, among the 59 tagged abstracts from the 110 positive papers in the IR benchmarking dataset, 57 were detected by the program during citation mapping. On the other hand, the high precision of annotation extraction from retrieved phosphorylation papers will ensure minimal effort in manual checking to validate the annotation. Indeed, RLIMS-P extracted several site features from non-tagged abstracts that were later validated by annotators as true positives. These include cases where evidence of experimental features is found in more than one abstract but only tagged in one, or where evidence is found in both abstract and full-length text but only tagged in the latter. A few site features detected by RLIMS-P are missed by curators. Given the excellent performance of the system in both IR and IE, we have already applied the program to provide computer-assisted evidence attribution at PIR for experimental features still awaiting retrospective attribution. RLIMS-P will also be employed to extract new phosphorylation information from PubMed literature for site feature annotation in the UniProt Knowledgebase.

Whereas the evaluation study conducted herein focuses on mining MEDLINE abstracts, the program is generally applicable to phosphorylation information extraction from full-length articles. Future enhancements of the RLIMS-P system will include: (1) adding more phospho-residue rule patterns using chemical synonyms for phosphorylated amino acids, such as ‘phosphoserine’, (2) coupling the rule patterns with short sequence patterns to recognize phosphorylated residues from sequence patterns and (3) fusing information from multiple sentences, especially when <THEME> and <SITE> are described in separate sentences. The system can also be adapted to mine other PTMs, such as methylation and acetylation.

ACKNOWLEDGEMENTS

The project is supported in part by grant U01-HG02712 from the National Institutes of Health, USA. M.N. and K.E.R. acknowledge the Council of Scientific and Industrial Research, India for JRF and SRF grants, respectively.

REFERENCES

- Apweiler,R. et al. (2004) UniProt: Universal Protein Knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
- Brill,E. (1995) Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguist.*, **21**, 543–565.
- Blaschke,C. et al. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 60–67.
- Cohen,P. (2002) The origins of protein phosphorylation. *Nat. Cell Biol.*, **4**, E127–E30.
- Collier,N., Nobata,C., and Tsujii,J. (2000) Extracting the names of genes and gene products with a hidden Markov model. In Proceedings of the 18th International Conference on Computational Linguistics, Saarbrücken, Germany, pp. 201–207.
- Diella,F. et al. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
- Donaldson,I. et al. (2003) PreBIND and textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.
- Friedman,C. et al. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17** (Suppl. 1), 74–82.

- Hirschman,L. *et al.* (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.
- Hobbs,J. *et al.* (1997) FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In Roche,E. and Schabes,Y. (eds), *Finite-State Language Processing*. The MIT Press, Cambridge, MA, pp. 383–406.
- Hu,Z.Z. *et al.* (2004) iProLINK: an integrated protein resource for literature mining. *Comput. Biol. Chem.*, **28**, 409–416.
- Narayanaswamy,M. *et al.* (2003) A biological named entity recognizer. *Pac. Symp. Biocomput.*, 427–438.
- Ng,S. and Wong,M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 104–112.
- Park,J.C. *et al.* (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammars. *Pac. Symp. Biocomput.*, 396–407.
- Proux,D. *et al.* (2000) A pragmatic information extraction strategy for gathering data on genetic interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 279–285.
- Pustejovsky,J. *et al.* (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac. Symp. Biocomput.*, 362–373.
- Ravikumar,K.E. *et al.* (2004) Towards building a database of phosphorylate interactions: extracting information from the literature. *Proc. SCI*, 57–63.
- Rindflesch,T.C. *et al.* (1999) Mining molecular binding terminology from biomedical text. *Proc. AMIA Symp.*, 127–131.
- Rindflesch,T.C. *et al.* (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.*, 517–528.
- Stapley,B.J. and Benoit,G. (2000) Bio-bibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.*, 529–540.
- Stephens,M. *et al.* (2001) Detecting gene relations from Medline abstracts. *Pac. Symp. Biocomput.*, pp. 483–495.
- Sekimizu,T. *et al.* (1998) Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 62–71.
- Thomas,J. *et al.* (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.*, pp. 541–552.
- Yakushiji,A. *et al.* (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.*, pp. 408–419.
- Torii,M. *et al.* (2004) Using name-internal and contextual features to classify biological terms. *J. Biomed. Inform.*, **37**, 498–511.
- Wu,C.H. *et al.* (2003a) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
- Wu,C.H. *et al.* (2003b) Protein family classification and functional annotation. *Comput. Biol. Chem.*, **27**, 37–47.