BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i319–i327
doi:10.1093/bioinformatics/bti1011

Beyond the clause: extraction of phosphorylation information from medline abstracts

M. Narayanaswamy[1], K. E. Ravikumar[1] and K. Vijay-Shanker[2,]*

[1]AU-KBC Research Centre, Anna University, Chennai, India and [2]Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Phosphorylation is an important biochemical reaction that plays a critical role in signal transduction pathways and cell-cycle processes. A text mining system to extract the phosphorylation relation from the literature is reported. The focus of this paper is on the new methods developed and implemented to connect and merge pieces of information about phosphorylation mentioned in different sentences in the text. The effectiveness and accuracy of the system as a whole as well as that of the methods for extraction beyond a clause/sentence is evaluated using an independently annotated dataset, the Phospho.ELM database. The new methods developed to merge pieces of information from different sentences are shown to be effective in significantly raising the recall without much difference in precision.

Contact: vijay@cis.udel.edu

INTRODUCTION

Research in biology is flourishing and large amounts of data are being created. While some of the data is stored in structured form in various databases, vast amounts of information are still available only through the literature. With the volume of literature itself growing so rapidly, it is extremely difficult for researchers to keep track of rapid advances in their own specializations, let alone in broader areas. 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. Natural language processing (NLP) technologies are being utilized for biological literature mining and information extraction (Blaschke *et al*., 2002).

The work presented here describes a system to extract phosphorylation information from text. The phosphorylation reaction is catalyzed by a kinase and involves the transfer of a phosphate group to a specific amino acid residue in a substrate. Phosphorylation is perhaps one of the more widely studied protein modifications and phosphorylation cascades play important roles in signal transduction pathways. Many signal transduction pathways can be initiated with the phosphorylation of key signal proteins.

Protein phosphorylation information is available in some protein databases, including UniProt (Apweiler *et al*., 2004), as well as specialized databases such as Phospho.ELM (Diella *et al*., 2004) and the Phosphorylation Site Database (http://vigen.biochem.vt.edu/xpd/xpd.htm). Phospho.ELM, a comprehensive database source of phosphorylation information, is important for this work, as our method is evaluated on it.

The work here reports a rule-based phosphorylation information extraction system—RLIMS-P. It attempts to extract (1) the kinase—henceforth also called the 'agent' (of phosphorylation), (2) the substrate—henceforth also called 'theme' and (3) the residue involved—henceforth also called the 'site' of phosphorylation. This is in contrast to many of the current information extraction systems that extract protein–protein interactions and, hence, are concerned only with the two proteins involved—the kinase and the substrate in the case of phosphorylation. Our system is based on the use of manually developed patterns that express many different ways by which phosphorylation reactions are expressed in text. In other works (Ravikumar *et al*., 2004; Hu *et al*., 2005),[1] we have evaluated the effectiveness of these patterns. However, these patterns are matched against clauses (simple sentence like constituents). The focus of this work is on the extraction of information by looking beyond individual clauses. We will now discuss the reasons for our belief that to improve the effectiveness of pattern-based phosphorylation information extraction, we need to look beyond clauses. We will also discuss the extra-clausal rules that we employ and evaluate their effectiveness.

As we noted above, we are interested not only in extracting the two proteins (the kinase, i.e. agent; and the substrate, i.e. theme) involved in phosphorylation but also the site of phosphorylation. It is uncommon to see all the three objects mentioned in the same sentence let alone the same clause.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[1]The extra-clausal processing, the focus of this paper, was neither discussed nor evaluated in these two papers.

Given that many of the patterns we deploy are matched against constituents of a clause, they cannot be used to extract the three objects. For example, with the verbal use of 'phosphorylate' (e.g. 'phosphorylated', 'phosphorylates', 'to phosphorylate') it is not uncommon to see only two of the objects mentioned in the clause (e.g. the agent and theme the active form). However, with the nominal usage (e.g. using the word 'phosphorylation'), the focus seems to be more on the theme or site and the agent may not be mentioned. It can sometimes even be the case that only one argument is mentioned in a sentence. For example,

EXAMPLE 1. 'The phosphoserine was located in position 2.'

This sentence of course only mentions that a serine residue was phosphorylated and that it is located at the second position in the protein sequence. Clearly, the author assumed that the readers would be able to know the identity of the substrate/theme from the context of the occurrence of the sentence in the text. We do not see how pattern matching or parsing/grammar based approaches can be used to connect the site and theme together. Furthermore, the agent of phosphorylation (the kinase involved) may be specified in yet another sentence. We believe that such situations call for mechanisms that connect such pieces of information together that considers the larger context /discourse. In our case, we consider a few rules to connect together arguments mentioned in different places in the text. We also introduce an operation we call fusion that also merges information extracted from different places in the text. In this paper, we focus on these new rules and the fusion operation and evaluate their effectiveness.

RELATED WORK ON INFORMATION EXTRACTION

Recently, there has been considerable interest in applying language technology in the biomedical domain. There has been much effort put into the extraction of information from biomedical text. Much of this body of work has been devoted to the extraction of protein–protein interactions from the literature. Such extraction is of course most relevant to the topic of this work. There are different approaches that have been employed to extract such information. One line of work (e.g. Marcotte *et al*., 2001; Stapley *et al*., 2000; Stephens *et al*., 2001), has been to use co-occurrence of gene/protein names (e.g. in abstracts or sentences) and sometimes their co-occurrence frequencies to predict their relatedness. These efforts have high recall rates but it can be noted that co-occurrence within a sentence or abstract does not necessarily imply a direct relation. The statistics of the occurrences of terms in abstracts have also been used to predict whether an abstract contains interaction information. In contrast to the co-occurrence approaches, another approach has been to employ more sophisticated NLP to extract protein–protein interactions. Some systems (e.g. Blaschke and Valencia, 2002; Thomas *et al*., 2000; Pustejovsky *et al*., 2002; Humphreys *et al*., 2000; Sekimizu *et al*., 1998; Rindflesch *et al*., 2000) concentrate on few particular verbs (e.g. inhibit, associate, etc.) in their information extraction. Also, in addition to limiting the extraction to certain interactions, like in our case, some of these efforts might be characterized as pattern-based extraction. Finally, some systems (Friedman *et al*., 2001; Yakushiji *et al*., 2001; Park *et al*., 2001) represent efforts to capture much more sophisticated grammar/syntactic structures and do not necessarily limit the verbs under consideration. Information extraction systems, as reported, are at various stages of development.

All of the above cited works on extraction of protein–protein interactions only extract the pairs of interacting proteins and do not recover any additional object (such as site of phosphorylation in our case). Additionally, like our basic system that is based on patterns, they extract information from sentences. The focus of this paper though is on extending our basic system to merge information from multiple sentences and information spread more widely across the text.

METHODS: THE BASIC SYSTEM

The pipeline of processing starts with simple text processing, dividing the text (currently Medline abstracts) into sentences, identifying the title, etc. Next the sentences are input to a part of speech (POS) tagger [currently, the system uses the Brill's tagger (Brill, 1995)], supplemented with a few rules to correct the common errors it makes in the biology domain. As better POS taggers are being developed for the specialized biomedical domain, they can be easily used to replace the current POS tagging system. With each word being tagged with the appropriate POS tags, using simple non-recursive or right-linear grammar rules (i.e. equivalent to regular grammar), certain simple phrases are recognized.

The most crucial type of phrases is the so-called BaseNP—simple noun phrases that do not involve any recursion. BaseNPs play an important role in information extraction as they often refer to named entities. BaseNP detection involves using the POS tags of words that usually appear at the boundaries. For example, verbs (e.g. activates) and prepositions (e.g. of, by, in, etc.) help identify the end of a BaseNP, whereas a determiner (e.g. this, the, etc.) indicates the start of a BaseNP.

Additionally, a few other syntactic constructions related to noun phrases are recognized, including coordination of entity names, and appositives. The detection of constructs such as appositives is mainly to correctly identify the location of the arguments.

We also chunk verb groups and determine the voice of the verb group. The following example illustrates the utility of chunking verb groups.

EXAMPLE 2. 'Active p90Rsk2 was found to be able to phosphorylate histone H3 at Ser10.'

The verb group sequence 'was found to be able to phosphorylate' is recognized and identified as containing three separate verb groups. The last verb group 'to phosphorylate' is detected as having an active form in contrast to 'was found' which is in the passive form. The detection of the active form allows the assignment of an agent role to the subject (while with the passive form the subject noun phrase must be assigned the theme role).

Type classification

One of the key components of the RLIMS-P system is the assignment of semantic types to noun phrases that refer to the phosphorylation objects of interest (i.e. arguments ⟨AGENT⟩ ⟨THEME⟩ and ⟨SITE⟩). Semantic type assignment (also employed in information extraction systems such as those outlined in Rindflesch *et al*., 2000; Pustejovsky *et al*., 2002) plays a fundamental role in pattern specifications. It simplifies the pattern specifications and improves the precision. Consider the following examples:

(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

Clearly, the theme and site cannot be disambiguated simply by using a pattern that is based only on syntax/surface form and roughly corresponds to 'X phosphorylated Y in Z,' without additional constraints on the types of X, Y and Z. While all three examples match this same syntactic pattern, the relation extracted will depend on what matches Y and Z (the underlined 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 the first example above.

In RLIMS-P system, NPs are classified into the types 'protein', 'protein sites', 'chemical compounds' or 'others.' The classification method is based on the ones that we employed in our name entity extraction system (Narayanaswamy *et al*., 2003) and further developed in Torii (2004).

Patterns

The three key phosphorylation objects are next extracted by matching text with manually designed pattern templates. The patterns cover the verbal inflected forms such as 'phosphorylated/phosphorylating/phosphorylates' as well as nominal forms. An example of a pattern for the verbal form is:

Pattern 1. ⟨AGENT⟩ ⟨VG-active-phosphorylate⟩ ⟨THEME⟩ (in/at ⟨SITE⟩)?

where 'VG-active-phosphorylate' denotes a verb group in active form, which is headed by an inflected form of the verb phosphorylate and '?' denotes optional argument.

This pattern is read as requiring an NP of the type appropriate for ⟨AGENT⟩ of phosphorylation appearing to the left of a sequence of verb groups. The head (or one of the heads in case of conjunction) of the main verb group must be in the inflected form 'phosphorylate' or 'phosphorylated' and have an active voice. Furthermore, it requires an NP of the type appropriate for ⟨THEME⟩ appearing to the right of this verb group. Finally, NP with type appropriate for ⟨SITE⟩ in a prepositional phrase with 'in' or 'at' can be optionally (indicated by ? in the pattern) 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 sentence is captured by another pattern. The patterns for the passive form capture the fact that the subject of the clause gives the theme and/or site. For example, one of the patterns is:

Pattern 2. ⟨SITE⟩ (in/of ⟨THEME⟩)? ⟨VG-passive-phosphorylate⟩ (by ⟨AGENT⟩)?

that is matched by 'all three sites in c-jun were phosphorylated'. Of course, the subject in the passive case could only specify the theme with the site appearing optionally after the verb as mentioned in the pattern

Pattern 3. ⟨THEME⟩ ⟨VG-passive-phosphorylate⟩ (by ⟨AGENT⟩)? (at ⟨SITE⟩)?

While many previous information extraction projects have concentrated only on the verbal forms of interactions, patterns to extract information from the nominal form are needed in the case of 'phosphorylate' interactions. Indeed, 'phosphorylation' is the most frequent inflected form, appearing about as often as other inflected forms put together in the development text corpus for RLIMS-P. 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; such as with the theme appearing after 'phosphorylation' as in:

Pattern 4. $[\langle\text{AGENT}\rangle \text{ phosphorylation}]_{\text{NP}}$ of ⟨THEME⟩

In contrast, when a protein name appears before 'phosphorylation site', this protein is likely to be the agent. The agent and theme can also appear after 'phosphorylation' as captured by the pattern:

Pattern 5. phosphorylation of ⟨THEME⟩ (by ⟨AGENT⟩)? (in /at ⟨SITE⟩)?

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

Pattern 6. ⟨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'.)

METHODS: EXTRA-CLAUSAL PROCESSING

The processing discussed above involves extraction of phosphorylation objects by matching the patterns against the text. Given the nature of these patterns, these patterns will typically be matched by phrases that appear in a clause. Now, we will consider some rules that allow us to extract arguments not mentioned in the same clause or sentence.

First, we begin by considering some rules that allow us to extract the theme.

Theme Rule 1 (TR-1). Consider the case where a sentence/clause contains information about the site of phosphorylation but not the theme, as for example in:

EXAMPLE 3. 'The phoshoserine is located at position 2.'

The use of the term phosphoserine (more generally, an amino acid following the prefix phospho) provides the information that a serine residue is phosphorylated. The position of the serine residue in the protein sequence is also stated in this sentence. But this substrate protein (the theme) is not mentioned in the clause/sentence.

Clearly, the site alone could not have been given without the theme being known. We believe that the author(s) assumed that the theme would be clear to the reader. Therefore, in all likelihood the theme is mentioned in the text. Standard English discourse processing assumptions would suggest that this information was probably mentioned at a prior point. Instead of simply taking any protein as a candidate for the theme, in developing TR-1, we feel that the author would have made the protein, that is the substrate, abundantly clear. Therefore, we assume that the substrate protein is not only mentioned earlier, but is also mentioned as a substrate of phosphorylation. If several mentions of phosphorylation and substrates are made in the text, we take the one closest to the clause, which triggers the use of TR-1, i.e. clause from which the site but not the theme is extracted.

This heuristic is employed quite often and has turned out to be almost always correct when employed. An example of its usage is shown in EXAMPLE 9.

Theme Rule 2 (TR-2). This rule also applies when the use of patterns allows the extraction of site but not the theme. In some cases, there may be no mention of the substrate of the phosphorylation in text prior to the abstract, we look at the title. If the processing of the title allows the extraction of the substrate of phosphorylation, then we use this substrate as the theme for the current clause. Even if the title does not mention phosphorylation and, hence, the substrate, a protein name occurring in the title is taken as the theme for the current clause. This heuristic is based on the assumption that a single protein name occurring in the title is probably the focus of the paper (and the abstract). Further, what is phosphorylated (i.e. the theme) is probably more central and more important information than what does the phosphorylation (agent). Thus, with this heuristic, we assume that the protein mentioned in the title is more likely to be reported as a substrate rather than a kinase, and that when we encounter a sentence about the phosphorylation site where no theme has been mentioned thus far, this protein (the one in the title) is the missing theme.

As an example, the title of an article (PMID-11861906) is

EXAMPLE 4. 'TI—Phosphorylation of mammalian translation initiation factor 5 (eIF5) *in vitro* and *in vivo*'

Here, of course, the title mentions a protein that is also specified as a substrate explicitly.

Theme Rule 3 (TR-3). In some cases, the patterns might allow us to extract only the agent without the theme or the site. TR-3 is used in such cases and works identical to TR-1 in that it takes the previous but closest mention of a substrate protein as a theme.

Theme Rule 4 (TR-4). The final theme rule, employed as a last resort, seeks to find the theme within the sentence (but perhaps beyond the clause) itself. This rule will assign the theme role to a protein name found[2] in the same sentence, provided it has not been assigned as the agent role already. We assume of course that this protein was not picked as theme because of the lack of patterns or inadequate syntax processing.

This rule without further modifications tends to be very imprecise. To improve the precision of extraction, we assign the theme role to a protein name in the sentence provided it meets some additional conditions. These conditions stipulate, for example, that the noun phrase containing the protein name must not immediately follow the word 'by' (as in such cases the protein tends to be an agent of phosphorylation or another interaction rather than the theme). Also, if the protein name is known to be the name of a kinase (we maintain a large list of known kinase names and abbreviations), it does not get assigned the theme role.

The next rule we consider allows us to fill in the site slot.

Site Rule 1 (SR-1). This rule is used to identify site information and is triggered by the use of a pattern that does not

[2]In case more than one protein name can be identified in the sentence then the one closest to the 'phosphorylation' word is chosen.

provide the site information. The rule is applied to see if the site information can be located within the same sentence. In the following example, note that the site information is found in the same sentence but not the same clause.

EXAMPLE 5. 'skp1 is also phosphorylated at a serine which has been identified as S-335.'

Even with co-reference resolution of 'which', the fact still remains that the theme and position are in two separate clauses. SR-1 will correctly identify the serine residue at position 335 as the site of phosphorylation.

However, there are cases where the site information is extracted from the same clause. An example for such a case is '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.'

The extraction of theme and agent is straightforward with the matching of the appropriate pattern. But these patterns do not match with all intervening material and also simultaneously allow for extraction of the site (serine 887). SR-1 allows the site information to be extracted.

Unlike the TR rules, SR-1 is less motivated by discourse principles but is more reminiscent of the co-occurrence-based approaches to information extraction (where a pair of protein names mentioned in a sentence is taken as an indication of interaction between them and hence extracted).

SR-1 can also apply because we did not have any appropriate pattern (although conceivably a general pattern may exist): Consider the following (PMID 11500516):

EXAMPLE 6. 'We have previously demonstrated that four (Tyr(1144), Tyr(1201), Tyr 1226/1227, or Tyr(1253)) of the five known Neu/ErbB-2 autophosphorylation sites can independently mediate transforming signals.'

Here the four specific positions of phosphorylation are extracted using SR-1 only and not through the use of any pattern.

Fusion. We also developed a fusion/merging operation to combine information extracted from different sentences. Stated simply, fusion combines the information when there is some overlapping information extracted from two sentences and if there is nothing incompatible about them then the information extracted from the two can be merged. This operation is similar to the notion of (most general) unifier of first-order terms where 'agent', 'theme', etc. are treated as function symbols; names of proteins, amino acids and positions are treated as constant symbols; and unknown values are treated as variables. Therefore, for example, the agent argument resulting from the fusion of two individual records will have a protein name if both original records have the same agent argument or if one has it as unspecified/unknown. For example, the fusion of the record ⟨agent = casein kinase II, theme = skeletal muscle calsequestrins, site = UNK⟩ with the record ⟨agent = UNK, theme = skeletal muscle calsequestrins, site = threonine, 363⟩, which are extracted from different sentences (PMID 1985907), yields the record: ⟨agent = casein kinase II, theme = skeletal muscle calsequestrins, site = threonine, 363⟩.

Note when the original records contain incompatible information (say for example different proteins playing the agent role), fusion is not possible. This is, of course, the case with first-order term unification as well.

Anaphora Resolution. We have implemented a rudimentary anaphora resolver in order to extract the correct phosphorylation arguments. Our resolution method almost exclusively relies upon the semantic type and number agreement between the anaphoric phrase and a candidate antecedent phrase. For example, given the anaphoric phrase 'both sites' in the following fragment.

EXAMPLE 7. 'Dephosphorylated hsp 90 is phosphorylated at both sites by casein kinase II . . .'

the candidate antecedent that our method would consider are those phrases referring to two objects, where the objects themselves are determined to be of the type that is given by the head word 'site'. In this case, our example correctly identified the anaphor with 'serine 231 and serine 263' which appeared in a preceding sentence.

'For the alpha protein , these sites correspond to serine 231 and serine 263.'

However, the simple anaphora resolution method fails to resolve in some cases because it cannot meet the above conditions or resolves incorrectly because we always take the closest antecedent that meets these conditions as the antecedent. It fails to resolve in the anaphoric expression 'the entire protein' in the following sentence.

EXAMPLE 8. 'The entire protein was phosphorylated by rEGFR at eight tyrosine residues (Tyr285, Tyr373, Tyr406, Tyr447, Tyr472, Tyr619, Tyr657 and Tyr689).'

This lack of resolution had some significant consequences. Because the patterns matched to pick the phrase for the theme, although it did not get resolved, the system did not consider the application of any of the Theme rules (TRs). The previous sentence not only mentioned the correct antecedent but also mentions it as a substrate. Hence while TR-1 could have correctly identified the theme, it did not apply. Further, because of the theme phrase not being resolved, even fusion with information from the previous sentence was not allowed (because there was no overlapping information in the information extracted from the two sentences).

Order of application. First, RLIMS-P applies its patterns to extract the basic information. At this time, if any of the extracted slots correspond to an anaphoric expression, we attempt to resolve it. Then the rule SR-1 is applied prior to

TR-1. This is because TR-1 is applied only when the extracted information contains the site/position information, TR-1 can be applied more often (Example 9 below) since SR-1 might have been applied to extract the site. Then TR-2, followed by TR-3 and TR-4 are applied. We note that there are several cases where both TR-1 or TR-2 (triggered by extraction of site) and TR-3 (triggered by extraction of agent) can be applied. For example,

EXAMPLE 9. 'The primary site of phosphorylation by protein kinase C was also near the amino terminus at Ser-27'

SR-1 recovers the site and patterns extract the agent from this sentence. Since TR-1 is applied before TR-3, it is applied correctly and recovers the theme from a preceding sentence: 'Lipocortin I was phosphorylated near the amino terminus at Tyr-21 by recombinant c-src'. It is interesting to note that the theme is picked as same for the two, despite the fact that they are two distinct relations with different agents.

After TR-3, TR-4 is applied. This is the last rule applied for filling the theme slot and hence gets used only if none of the patterns nor TR-1 through TR-3 help extract the theme. Finally fusion is applied to the (partial) relations extracted thus far.

EVALUATION

In order to evaluate RLIMS-P and in particular the effectiveness of the extra-clausal processing it performs, we used annotated data from Phospho.ELM database. This database contains a collection of experimentally verified Serine, Threonine and Tyrosine phosphorylation sites in eukaryotic proteins. The entries are manually annotated and based on scientific literature. Each entry contains the phosphorylated amino acid and position (site), the substrate (theme), and the kinase responsible for the modification (agent) and links to bibliographic references. More details regarding this database and the entire dataset can be found at http://phospho.elm.eu.org/

Our evaluation is based on extraction from 386 abstracts from the Phospho.ELM dataset. This set constitutes an unseen test dataset, which is distinct from the abstracts we had used for the development of our system, its patterns and the extra-clausal rules we evaluate here. In contrast, our development was based on a couple of hundred abstracts and 10 journal articles. Further improvements were made based on an additional subset of abstracts from PIR's iProLINK, a feature annotated literature corpora (http://pir.georgetown.edu/iprolink).

Some notes about the test dataset

Although we downloaded the entire dataset, the evaluation was not completely automated. The following list discusses some important issues regarding the nature of the dataset and the factors we took into consideration during the evaluation.

- We considered only a subset of the data from the Phospho.ELM dataset. In particular, our evaluation set corresponds to information extracted from 386 abstracts.
- In some cases, we had to make a systematic mapping between the sites annotation in the dataset and the site mentioned in the text. For example, in the abstract of PMID 12387894, we can find the sentence: 'We found that cdk5 phosphorylated tau (441) at Thr-181, Ser-199, Ser-202, Thr-205, Thr-212, Ser-214, Thr-217, Thr-231, Ser-235, Ser-396 and Ser-404, but not at Ser-262, Ser-400, Thr-403, Ser-409, Ser-413 or Ser-422'.

  Here, a systematic difference of 316 residues was found between the sites Ser-214, Thr-217, Thr-231, and Ser-235 found in the text and the sites Ser-530, Thr-533, Thr-547 and Ser-551 in the Phospho.ELM annotation. Such systematic shifts in the position numbers occurred in a few other abstracts as well. In such cases, we took the RLIMS-P extraction of Ser-214, Thr-231 and Ser-235 as correct.
- In some cases, some of the sites mentioned in the text do not appear in the annotation. An example can be found in the same sentence we discussed in the above sentence. The annotation in Phospho.ELM contains entries only for the four sites given above and there are no entries for the other possibilities, i.e. Thr 181, Ser 199, Ser 202, Thr 205, Thr 212, Ser 396 and Ser 404. Our program extracted these sites as well. After manual inspection, we considered this extraction as correct.
- In some cases, while the dataset annotation might refer to some site position of phosphorylation, the text in the abstract might bear no information about this position. As an example, one of sites annotated in Phospho.ELM is Threonine 689 but the abstract of PMID 12387894, even taking into account any systematic mapping between annotated sites and sites mentioned in text, does not mention it. We presume that the human annotation of Phospho.ELM set took the information from the full length paper and not just the abstract in order to annotate the site in such cases. In our evaluation, since we ran RLIMS-P only on the abstracts from the PMID literature reference link in Phospho.ElM, we did not consider this as a case of a site missed by our program. However, we still evaluate the ability of RLIMS-P in extracting the agent and theme in these cases, if they are indeed present in such abstracts.
- Sometimes the abstract might contain no information about the agent (as an example, see PMID 2570779). Like above, in such cases, the abstract might still be used for evaluation of RLIMS-P's ability to extract the other two arguments, i.e. theme and site.

Table 1. Precision and recall of the RLIMS-P system(s)

System	ATS			AT			TS		
	Pre	Rec	F-mes	Pre	Rec	F-mes	Pre	Rec	F-mes
PAT	95.1	24.2	38.6	98.1	64.9	78.1	97.7	25.1	39.9
PAT + SR-1	93.4	40.7	56.7	98.1	64.9	78.1	96.8	61.5	75.2
PAT + SR-1 + TRs	93.2	57.2	70.9	97.2	82.6	89.3	97.1	82.4	89.2
PAT + SR1 + Fusion	94.3	50.5	65.8	98.1	64.9	78.1	96.8	61.5	75.2
PAT + SR-1 + TRs + Fusion	95.2	77.4	85.4	97.2	82.6	89.3	97.1	82.4	89.2

PAT = extraction using only patterns; PAT + SR-1 = extraction using only patterns plus application of SR-1 (i.e. no use of TRs and fusion); Pre = Precision; Rec = recall; F-mes = F-measure. ATS is the ternary relation between agent, theme and site. AT stands for the binary relation between agent and theme. TS is the binary relation between theme and site.

Table 2. Precision and recall for individual argument slots

System	Agent			Theme			Site		
	Pre	Rec	F-mes	Pre	Rec	F-mes	Pre	Rec	F-mes
PAT	97.1	84.9	90.6	99.6	85.2	91.8	97.1	73.2	83.5
PAT + SR-1 + TRs + Fusion	97.1	84.9	90.6	99.3	98.9	99.1	97.4	97.5	97.3

- Another reason we could not automate the evaluation against the Phospho.ELM dataset was because RLIMS-P currently only picks the agent and theme by the name as it appears in the text. However, in the annotation of Phospho.ELM, they might be represented by some canonical name. For example, (PMID 9733784) the theme extracted by the program is Vitronectin, whereas the Phospho.ELM annotation mentions its synonym S-protein. In such cases we manually verified whether there was a synonym relation between the name as it appears in the text and the annotation by looking up Swiss-Prot.

RESULTS

We evaluate five separate systems. The first is the one where only patterns are applied but no extra-clausal processing is applied and the last is the complete system. The difference in the evaluation results between these two systems shows the utility of the extra-clausal processing. These correspond to the first and last row in Table 1. The second and third rows correspond to the application of SR-1 and SR-1 and the TRs (respectively) to the basic patterns based system. Finally, Row 4 corresponds to the system that uses fusion instead of TRs in the system corresponding to Row 3.

Let us first consider the first set of numbers under the column with the heading ATS. All five systems show high precision, showing that we do not sacrifice precision in order to boost recall by adding the extra-clausal processing. The utility of the extra-clausal processing for our system can be immediately seen by noting how poor the recall is for the patterns-only system and how it is boosted with the use of the different rules or fusion. Further, by comparing Rows 3 and 4, we note that fusion is not an adequate replacement for the TRs.

The heading ATS (for the first column of numbers) refers to the fact that we are evaluating the extraction of the entire relation (i.e. the ternary relationship between the agent, theme and site). So, even the incorrect extraction of a single argument (with other arguments being correctly extracted) would mean that the relation extracted is considered wrong. To evaluate the extraction of each individual argument, we need to consider Table 2.

In Table 1, we have two other sets of numbers under the headings AT and TS. These two refer to the extraction of the binary relations between agent and theme and between theme and site, respectively. The agent–theme (AT) relation is similar to extraction of protein–protein interactions and might be of interest to those who are only interested in the interactions between pairs of proteins. The theme–site relation might be of interest to those researchers who are more interested in the biological consequence of phosphorylation (where the identity of the kinase catalyzing the modification is not as important). Note that the recall for both relations is significantly better than that for the composite ternary relation. The slight increase in precision might be attributed to the fact that for the two binary relations, many more instances are considered. Note, as discussed above (in notes about the dataset), several abstracts do not provide information about the agent or the site. As noted, in such cases, these were not used for evaluation of the ATS relation but are included in the evaluation of AT, TS as well as evaluation of the extraction of the individual arguments (Table 2). Table 2 shows the utility of the extra-clausal

processing by showing the results for individual arguments with or without them.

Note that SR and TRs augment the patterns in allowing us to extract arguments. Hence, we can see that extraction of theme and site improve with use. In contrast, there is nothing that can augment patterns for the extraction of agent (fusion only serves to merge together extracted arguments), and hence the performance on agent extraction, when viewed in isolation, does not show any change.

EXAMPLES

We begin with an example that illustrates the utility of our extra-clausal processing, showing that we can extract some relations that might not be otherwise possible using just our parsing/chunking and pattern application.

From the following sentence (PMID 7925415), neither the theme nor the site is extracted by applying the patterns.

EXAMPLE 10. Partial proteolysis of phosphocalmodulin by thrombin identifies Tyr99, located in the third calcium-binding domain of calmodulin , as the phosphorylated residue.

However, the appearance of the phrase 'phosphorylated residue' triggers SR-1 that enables the site ⟨99, tyrosine⟩ to be picked. This in turn, causes TR-1 to be triggered. Although Calmodulin was not picked from the sentence, since it was identified as theme in a preceding sentence, TR-1 connects the site with the theme. Now, this allows the fusion (because of overlap in theme) with information from the title ('Phosphorylation of calmodulin by the epidermal-growth-factor receptor tyrosine kinase'), which contains both agent and theme.

Without SR-1, we would not have been able to relate the three objects together. SR-1's application allows the extraction of site but also allows application of TR-1. Without application of TR-1, there would not have been enough overlap to consider the fusion.

Now, we will turn our focus on errors made by RLIMS-P. An incorrect result was obtained owing to the application of SR-1 for the following sentence:

EXAMPLE 11. 'Substitution of Tyr-525 and Tyr-526 at the autophosphorylation site of Syk in mCD8-Syk substantially reduced the kinase activity and the binding of this variant chimera to PLC-gamma1 SH2(C) *in vitro*; it also failed to induce tyrosyl phosphorylation of PLC-gamma1 *in vivo*.'

Two relations are recovered from this sentence. The first was correctly extracted corresponding to Syk being both theme and agent with the two sites being identified. However, an error was made in the extraction of the second relation. The theme 'PLC-gamma1' was correctly identified. However, SR-1 was incorrectly applied and the same two sites (of autophosphorylation of Syk) were also extracted erroneously as the site of this second phosphorylation relation.

The following gives an example of missed site information despite the application of SR-1 (PMID 1689310)

EXAMPLE 12. 'The rate of phosphorylation was fastest with the sites at 771 and 783, then at 1254, and slowest at 472.'

The problem arises because the conjunction of 771 and 783 is recognized and these two sites are extracted. However, the remaining two are not extracted, because recognition of their conjunction with other sites is missed.

A rare error owing to the application of TR-1 resulted from a missed theme from the sentence itself (PMID 10978177)

EXAMPLE 13. 'Recently, we identified a similar cohort of tyrosine phosphorylation sites for the epidermal growth factor receptor (EGFR) with a predominant phosphorylation of tyrosine residue Y657 and binding of Syp.'

Here for the second occurrence of 'phosphorylation', the theme is annotated to be 'Syp'. However this was not picked and TR-1 was applied, retrieving a theme from a preceding sentence. That sentence referred to another phosphorylation and hence an incorrect theme was chosen for the relation extracted for this sentence.

Some errors are not owing to aspects related to extra-clausal processing methods, the focus of this paper. For example, because RLIMS-P currently does not consider conjunction of prepositional phrases (but only conjunction of noun phrases), only one agent (polyoma middle c-src complex) was picked from the following sentence (PMID 2457390).

EXAMPLE 14. 'The same tyrosine residue was phosphorylated by polyoma middle c-src complex, by recombinant v-abl, and with A431 cell membranes by the EGF receptor/kinase.'

Another example of a relation that was missed by RLIMS-P (not owing to the failure of SR-1, TRs or fusion) for the following sentence (PMID 10608806):

EXAMPLE 15. 'Putative ATM in vitro targets include p95/nibrin, Mre11, Brca1, Rad17, PTS, WRN and ATM (S440) itself.'

In fact, looking at the sentence alone it would be difficult to infer the autophosphorylation of ATM at Serine 440. (Perhaps this autophosphorylation relation is spelled out more clearly in the full length paper.) However RLIMS-P was applied on abstracts only. The assumption that the targets mentioned in the sentence are targets of phosphorylation becomes clearer when we consider the previous sentence: 'We determined a general phosphorylation consensus sequence for ATM and identified putative *in vitro* targets by using glutathione S-transferase peptides as substrates.'

We conclude by noting that the new extra-clausal processing methods we have introduced make a considerable difference in boosting the recall without any noticeable drop in precision. The high precision and recall of the resulting system

makes it a candidate system to be employed for the development of a curated large-scale database of phosphorylation modification.

ACKNOWLEDGEMENTS

M.N. acknowledges The University Grants Commission, India for the JRF Grant. K.E.R. acknowledges the Council of Scientific and Industrial Research, India for the SRF Grant. The authors wish to thank S.V. Ramanan, Carl Schmidt and Manabu Torii for various discussions and many valuable suggestions. The authors are grateful to Zhangzhi Hu and Cathy Wu for their many suggestions and collaboration which have significantly improved the performance of the RLIMS-P system. Finally, we are indebted to the developers of Phospho.ELM for making their database available to us.

REFERENCES

Apweiler,R., Bairoch,A., Wu,C.H., Barker,WC., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M., Martin,M.J., Natale,D.A., O'Donovan,C., Redaschi,N., Yeh,L.S. *et al*. (2004) UniProt: the Universal Protein Knowledgebase, *Nucleic Acids Res.*, **32**, D115–D119.

Blaschke,C., Hirschman,L. and Valencia,A. (2002) Information extraction in molecular biology, *Brief Bioinformatics*, **3**, 154–165.

Blaschke,C. and Valencia,A. (2002) The frame-based module of the Suiseki information extraction system. *IEEE Intell. Syst.*, **17**, 14–20.

Brill,E. (1995) Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguistics*, **21**, 543–565.

Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.

Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17** (suppl. 1), 74–82.

Hu,Z.Z., Narayanaswamy,M., Ravikumar,K.E., Vijay-Shanker,K. and Wu,C.H., (2005), Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*.

Humphreys,K., Demetriou,G. and Gaizauskas,R. (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 502–513.

Marcotte,E.M., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.

Narayanaswamy,M., Ravikumar,K.E. and Vijay-Shanker,K. (2003) A biological named entity recognizer. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 427–438.

Park,J.C., Kim,H.S. and Kim,J.J. (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 396–407.

Pustejovsky,J., Castaño,J., Zhang,J., Kotecki,M. and Cochran,B. (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 362–373.

Ravikumar,K.E., Narayanaswamy,M. and Vijay-Shanker,K. (2004) Towards building a database of phosphorylate interactions: extracting information from the literature. *Proceedings of the 8th Systemics, Cybernetics, and Informatics Conference*. pp. 57–63.

Rindflesch,T.C., Rajan,J.V. and Hunter,L. (2000) Extracting molecular binding relationships from biomedical text. *Proceedings of the 6th Applied Natural Language Processing Conference*, pp. 188–195.

Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 529–540.

Stephens,M., Palakal,M., Mukhopadhyay,S., Raje,R. and Mostafa,J. (2001) Detecting gene relations from Medline abstracts. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 483–495.

Sekimizu,T., Park,H. and Tsujii,J. (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., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 541–552.

Torii,M., Kamboj,S. and Vijay-Shanker,K. (2004) Using name-internal and contextual features to classify biological terms. *J. Biomed. Inform.*, **37**, 498–511.

Yakushiji,A., Tateisi,Y., Miyao,Y. and Tsujii,J. (2001) Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.*, World Scientific Publishing, Singapore, pp. 408–419.