

MPTM: A tool for mining protein post-translational modifications from literature

Dongdong Sun, Minghui Wang* and Ao Li
*School of Information Science and Technology
University of Science and Technology of China
Hefei, Anhui 230027, P. R. China
mhwang@ustc.edu.cn

Received 8 September 2017
Accepted 8 September 2017
Published 5 October 2017

Due to the importance of post-translational modifications (PTMs) in human health and diseases, PTMs are regularly reported in the biomedical literature. However, the continuing and rapid pace of expansion of this literature brings a huge challenge for researchers and database curators. Therefore, there is a pressing need to aid them in identifying relevant PTM information more efficiently by using a text mining system. So far, only a few web servers are available for mining information of a very limited number of PTMs, which are based on simple pattern matching or pre-defined rules. In our work, in order to help researchers and database curators easily find and retrieve PTM information from available text, we have developed a text mining tool called MPTM, which extracts and organizes valuable knowledge about 11 common PTMs from abstracts in PubMed by using relations extracted from dependency parse trees and a heuristic algorithm. It is the first web server that provides literature mining service for hydroxylation, myristoylation and GPI-anchor. The tool is also used to find new publications on PTMs from PubMed and uncovers potential PTM information by large-scale text analysis. MPTM analyzes text sentences to identify protein names including substrates and protein-interacting enzymes, and automatically associates them with the UniProtKB protein entry. To facilitate further investigation, it also retrieves PTM-related information, such as human diseases, Gene Ontology terms and organisms from the input text and related databases. In addition, an online database (MPTMDB) with extracted PTM information and a local MPTM Lite package are provided on the MPTM website. MPTM is freely available online at <http://bioinformatics.ustc.edu.cn/mptm/> and the source codes are hosted on GitHub: <https://github.com/USTC-HILAB/MPTM>.

Keywords: Post-translational modifications; text mining; dependency parse trees.

1. Introduction

Post-translational modifications (PTMs) are various biological processing events that control many aspects of protein behaviors such as cell signaling, growth, interactions, division and expression level.^{1,2} Previous studies have revealed that

* Corresponding author.

PTMs play an important role in human health and diseases.³ For instance, it is indicated that the phosphorylation of pRB1 is associated with tumor genesis through cell division.⁴ S-nitrosylation of parkin leads to protein accumulation in sporadic Parkinson disease by regulating its E3 ligase activity in protein ubiquitination.⁵ Moreover, hydroxylation is an important PTM and is closely related to various diseases.⁶ A better understanding of the role of hydroxylation can help inform the development of drug for diseases.⁷ At the same time, the latest study shows that myristoylation plays an important role in human immune response⁸ and abnormal or irregular myristoylation on proteins can lead to several pathological changes in the cell.⁹ GPI-anchored proteins are the major form of cell-surface proteins in humans. They may function in a variety of roles serving as receptors, adhesion molecules, etc.^{10,11} Because of the important role of PTMs, a large amount of literature reports novel findings about various PTMs. Currently, the UniProt knowledge base (UniProtKB) database¹² includes PTM information manually collected from the scientific literature. However, with the continuing and fast growth of this literature and the complex expression for PTMs, it is a huge challenge for database curators and researchers who want to track the latest knowledge about PTMs and their functional associations.

Employing various techniques from different fields such as machine learning, biological information retrieval, and natural language processing, text mining can help researchers automatically extract valuable knowledge and information from a large amount of literature. And various applications in biomedical research fields (cf. BioNLP'09¹³ and BioNLP'11¹⁴) adopt text mining techniques. For example, TEES (Turku Event Extraction System)¹⁵ which is based on support vector machine (SVM) and a unified graph representation, aims to generalize event extraction techniques. As a result of its good generalizability and performance across the BioNLP 2013 task corpora,¹⁶ it achieved first place in four out of eight tasks. EVEX¹⁷ extends the TEES event extraction system and is used to extract additional features for event extraction of GENIA Event Extraction task. Other text-mining advances include biological relationship extraction such as gene-disease interactions,¹⁸ named entity recognition (gene/protein name,¹⁹ organisms,²⁰ and diseases,²¹ etc.).

Current web-based text mining techniques for PTM information extraction include RLIMS-P,²² RLIMS-P 2.0,²³ eFIP²⁴ and MinePhos,²⁵ RLIMS-P²² is the first tool for PTM information extraction. It is a rule-based system which utilizes shallow parsing technique and manually developed patterns to extract phosphorylation information (substrates, kinases and sites) from abstracts. However, RLIMS-P²² only extracts phosphorylation information, and its complex and restricted patterns limit its overall performance. eFIP²⁴ integrates various tools to extract protein phosphorylation information from abstracts, and aids scientists to quickly find phosphorylation information from abstracts with a given protein (including site and kinase). MinePhos²⁵ combines several methods including enhanced RLIMS-P patterns and SVM to extract protein phosphorylation information from published literature. Veuthey *et al.* proposed a PTM information extraction tool to assist the curation of UniProtKB database,²⁶ it can annotate modification sites and related

protein information for 7 PTMs. In addition, iPTMnet^{27,28} is a resource that integrates PTM information from text mining results (RLIMS-P, eFIP), curated databases and ontologies. iPTMnet also provides visualization tools for exploring PTM networks, PTM crosstalk and conservation across species. PRO²⁹ is a very useful database to formally define protein entities and explicitly represents their major forms and interrelations. A latest version called PRO 2.0³⁰ extends the PRO database to facilitate the comparison of taxon-specific and taxon-neutral protein-related entities described in different resources. PRO 2.0 develops a standardized representation of proteoforms using UniProtKB as sequence reference and PSI-MOD as a post-translational modification reference.

Here, we propose an efficient online tool called Mining protein Post-Translational Modifications (MPTM) to extract and organize comprehensive PTM information from literature in PubMed. We have selected 11 PTMs according to the iProLINK³¹ literature corpora and the BioMinT project³² (available at <http://biomint.isb-sib.ch>), and eight PTMs among them have already been used in existing methods^{26,33}. Also, 10 these PTMs are included in the PTM Statistics Curator ranking list.³⁴ MPTM has the following advantages: (1) it covers 11 common PTMs (phosphorylation, methylation, glycosylation, acetylation, amidation, hydroxylation, myristoylation, sulfation, GPI-anchor, disulfide and ubiquitination). MPTM is the first web server that provides literature mining service for hydroxylation, myristoylation and GPI-anchor due to their important biological significance⁶⁻¹¹; (2) by using an efficient relation extraction method based on dependency parse trees³⁵ and a heuristic algorithm, it has good performance in both substrate and modification site detection; (3) it extracts not only basic PTM information including substrates and modification sites, but also comprehensive PTM-related information such as enzymes, Gene Ontology (GO) terms, organisms, diseases and crosstalk from literature. The web server is also accompanied by a precompiled collection of PTM information named MPTMDB with 435,433 PTM records, which enables versatile visualization of PTMs, substrates and interacting enzymes.

2. Material and Method

2.1. Data resources

The dataset in our work is divided into two parts: training data and testing data. The training data for phosphorylation is randomly chosen from PubMed based on the PubMed query “phosphorylation [Title/Abstract] AND phosphorylate [Title/Abstract]” and contains 150 abstracts (available at MPTM website). The training data is used to generate efficient relations for MPTM tool. The testing data contains 633 abstracts selected from iProLINK.³¹ literature corpora and BioMinT project³² (available at <http://biomint.isb-sib.ch>). It should be noticed that the two datasets are totally different from each other. Existing benchmarks such as iProLINK literature corpora mark the evidence but do not generate more detailed annotations for substrates, enzymes and sites.³⁶ BioMinT provides PMIDs of abstracts. As the

detailed annotations of existing benchmarks for multiple types of PTMs are incomplete or unavailable, we re-analyzed these existing benchmarks to generate more detailed annotations and to provide a manually labeled dataset (including 150 abstracts for training data) for measuring the performance of our tool. This re-analysis was carried out by four researchers in biology and biomedical text mining. All curators are experienced in annotating PTM information. Similar to RLIMS-P,³⁷ each abstract was annotated by at least two curators and the phosphorylation abstracts annotated were gathered to calculate inter-annotator agreement rate using F-score. Specifically, annotations by one curator were regarded as gold-standard, and the performance of the other curator was measured using F-score.³⁷ The inter-annotator agreement rate achieved is 0.86, and 0.92 for substrate and site, respectively. All manually curated datasets can be downloaded from the MPTM website (<http://bioinformatics.ustc.edu.cn/mptm/download.html>).

2.2. A workflow of the MPTM tool

MPTM is an information tool that extracts 11 common PTMs and their PTM-related information from the published literature. The approach in MPTM consists of five steps: (1) triggers filtering; (2) named entity recognition (NER) and substitution; (3) identification of substrates, enzymes and modification sites; (4) identification of GO terms, diseases, organisms, crosstalk and (5) identification of diseases from OMIM.³⁸ We describe each step of this workflow further in the following subsections. In addition, Fig. 1 illustrates the workflow of the MPTM tool for a given MEDLINE abstract (PMID: 12501250).

2.2.1. Triggers filtering

First, the PTM type of the input texts from PubMed containing paper titles and abstracts is identified. If the input texts contain useful information, it always contains several typical tokens (also known as triggers). Table 1 shows the regular expressions of filtering triggers for different PTM types. For example, a trigger for phosphorylation is defined as “phospho”. When an abstract has more than one PTM type, we use a frequency statistics method. Specifically, we count the number of each trigger in the text and choose the largest one to be the text type. For example, abstract [PMID: 164350] has 10 phospho-triggers and 1 glyco-trigger. Then the PTM type of this abstract is identified as Phosphorylation.

2.2.2. Named entity recognition (NER) and substitution

When a PTM type is identified for an abstract, the abstract is split into sentences and named entity recognition by BANNER.³⁹ In detail, we downloaded the source codes of BANNER from <http://banner.cvs.sourceforge.net/banner/> and compiled the codes into a jar package by Apache Ant tool. Also, we used the default training model in BANNER to identify protein names.

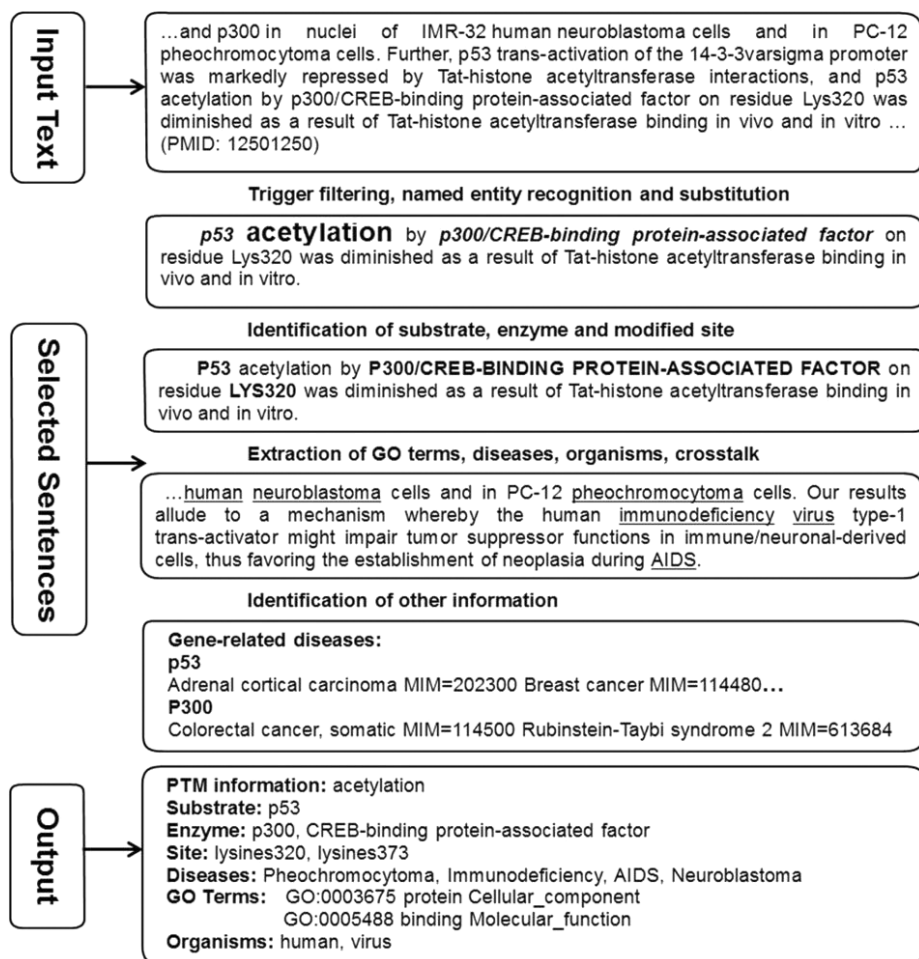


Fig. 1. An example workflow of the MPTM for MEDLINE abstract (PMID: 12501250).

Table 1. Triggers of different PTMs.

PTMs	Filtering triggers
Phosphorylation	Phospho
Methylation	Methyl, methyl[amino acid], e.g. "methyllysine"
Glycosylation	Glycol[o]an], O-GlcNAC, [N O C]-link
Acetylation	Acety
Amidation	Amid, carboxyamid
Hydroxylation	Hydroxy
Myristoylation	Myrist
Sulfation	Sulfa, sulph
GPI-Anchor	Anchor, GPI, glycosylphosphatidylinositol
Disulfide	Disulf, disulphide
Ubiquitination	Ubiquit, Ub, E1, E2, E3, E4, E6

In addition, in order to increase the accuracy of identifying protein names, name substitution is used for processing sentences with complex and long protein names by following existing studies.²² In detail, we employed semantic tokenization to replace multi-word protein names with single token. An abbreviated notation PRO_i (i is the index of the corresponding protein name in the dictionary) is used to replace protein names in each sentence. Thus, a hash table is generated including the replaced names and original protein names; therefore we can quickly map the original protein names to the replaced names for further analysis.

2.2.3. Identification of substrates, enzymes and modification sites

After Sec. 2.2.2, several sentences are generated corresponding to each input abstract. Each sentence is processed using the Stanford Parser.⁴⁰ to generate a dependency parse tree to represent structured grammatical relations between words in the sentence. Dependency parse tree has been widely applied since they are designed to be easily understood and effectively used for people who want to extract grammatical relations.⁴¹ For Stanford Parser, we use the release prepared by John Bauer. This release contains three parsers: a high-accuracy unlexicalized PCFG parser, a lexicalized dependency parser, and a factored model with a lexicalized PCFG parser. Here, we use the high-accuracy unlexicalized PCFG parser, which is quick to load and run, and quite accurate. This version of Stanford Parser contains approximately 50 grammatical relations in Stanford Dependencies Format.⁴² And some grammatical relations such as “*nsubj*”, “*prep*”, “*prep-of*”, “*prep-by*”, “*prep-with*”, “*prep-including*”, “*agent*”, “*conj-and*”, “*dobj*”, and “*num*” are common in PTM text mining task. Table 2 is a summary of their abbreviation, meaning, and usage in MPTM. For example, “*nsubj*” represents a nominal subject which is the syntactic subject of a clause. “*prep*” means a prepositional modifier of a verb, adjective, or noun. “*prep-of*”, “*prep-by*”, “*prep-with*” and “*prep-including*” mean a prepositional modifier such as “*of*”, “*by*”, “*with*” and “*including*”, respectively. “*agent*” is the complement of a passive verb which is introduced by the preposition “*by*” and can replace “*prep-by*”. “*conj-and*” represents “*conjunction-and*” with noun or verb. “*dobj*” is a direct object of a verb phrase which is the noun phrase. “*num*” means a numeric of a noun which is any number phrase that serves to modify the meaning of the noun with a quantity. More detailed descriptions about Stanford Dependencies Format can be found in Stanford typed dependencies manual⁴³ and more relations extraction in MPTM can be found in our source codes by downloading the local package.

In our study, dependency parse trees are generated by Stanford Parser in Stanford Dependencies Format.⁴² For instance, the dependency parse tree for the complicated sentence “We show that PRO1 interacts with PRO2 and directly phosphorylates PRO2 at residue T203” [PMID: 21952048] is shown in Fig. 2. It can be seen that the predicate verb “phosphorylates” is the sub-root of the tree and the dependency relationship “*nsubj*” (trigger, PRO1) [trigger = phosphorylates] located in the left of the sub-root. Further, a tree structure is formed by informative words and their

Table 2. A summary of common grammatical relations in MPTM with stanford dependencies format.

Grammatical relations	Abbreviations means	Example in Stanford Parser	Relations extraction in MPTM
nsubj	“nsubj” represents a nominal subject which is the syntactic subject of a clause.	“Clinton defeated Dole” nsubj(defeated, Clinton)	nsubj (trigger, PRO _i)
prep, prep_of, prep_by, prep_with, prep_including	“prep” means a prepositional modifier of a verb, adjective, or noun. “prep_of”, “prep_by”, “prep_with” and “prep_including” mean a prepositional modifier such as “of”, “by”, “with” and “including”, respectively.	“I saw a cat in a hat” prep (cat, in) “Bell, a company which is based in LA, makes and distributes computer products” prep_in (based, LA)	prep_of (substrate relation) prep_including (PRO1, PRO2)
agent	“agent” is the complement of a passive verb which is introduced by the preposition “by” and can replace “prep_by”	“The man has been killed by the police” agent(killed, police)	prep_by (enzyme relation)
conj_and	“conj_and” represents “conjunction_and” with noun or verb	“Bell, a company which is based in LA, makes and distributes computer products.” conj_and (makes, distributes)	conj_and (substrate1/enzyme1, substrate2/enzyme2)
dobj	“dobj” is a direct object of a verb phrase which is the noun phrase	“They win the lottery” dobj (win, lottery)	dobj(trigger, PRO _i)
num	“num” means a numeric of a noun which is any number phrase that serves to modify the meaning of the noun with a quantity	“Sam spent forty dollars” num(dollars, 40)	num(trigger, digits)

corresponding grammatical relations. By refining grammatical relations with those informative words, a complex sentence structure is represented in a concise dependency relation, which is beneficial for extracting PTM information. In this study, using the training data (150 abstracts) we have substantially extended the original relations extracted from the dependency trees in our previous study³⁶ and evaluated the performance of extracting phosphorylation substrate and site information using our testing data (99 abstracts for phosphorylation). For example, a new relation “*prep_including* (PRO1, PRO2)” is proposed for the identification of multiple substrates or enzymes. Another example is a new relation “*num* (trigger, digits)”, which is used for the identification of modification sites. These relations are combined together and represented by regular expression models. Furthermore, we have applied the extended relations (Table 2) to efficiently extract substrate and site

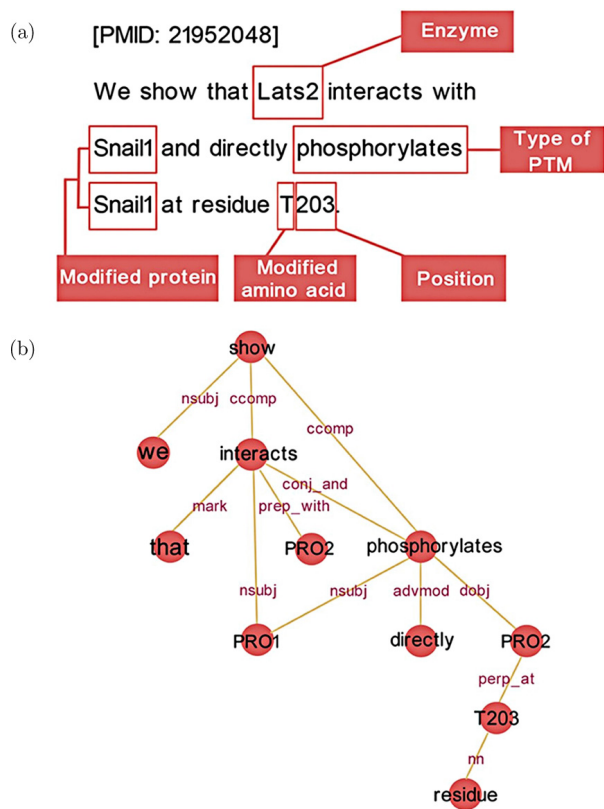


Fig. 2. The annotation and dependency relationships of a sample sentence. (a) is a sentence with protein phosphorylation information (b) is a dependency parse tree in stanford dependencies format.

information for other types of PTM and evaluated their performance using our testing data (534 abstracts for the other 10 PTMs).

To further accurately identify substrates, modification sites and enzymes, we use a new approach based on a heuristic algorithm to search dependency trees parsing from the input abstracts. In detail, we choose a *PROi* as a target protein and detect the grammatical relations directly linked with it, and call these relations as the first layer relation. For example, the relation with the node “PRO1” and node “phosphorylates” in Fig. 2 is the first layer when we choose the “PRO1” as a target protein. If the other node of the first layer relation is a trigger and the relation is “*prep_of*” (substrate relation), “*prep_by*” (enzyme relation) etc., the target protein is recognized as a substrate or an interactional enzyme. If the first layer relation node is not a trigger, we choose the next node directly linking with the first layer relation. Then if the other node of the first layer relation node is a trigger and the relation is “*prep_of*” or “*prep_by*”, the target protein is also recognized as a substrate or an interactional enzyme. Another case is that DNA-methylation is very common in PTM corpus; this results in false positive cases in our tool. We use a new approach

based on a heuristic algorithm (tree-based search algorithm) covering three layer grammatical relations to correct this. For example, when MPTM captures a relation such as “*prep_of*” (trigger, PRO1) [trigger = methylation] and the previous or next layer is “*nn*” (trigger, DNA) [trigger = methylation], it may be a false positive case since it describes DNA-methylation. However, site relations are more diverse than substrate and enzyme relations and therefore we further add a few rules to obtain good performance.³⁶ For example, in the sentence fragment “Gln-Ser-Gly-Ser(P)-Val-Ile-Tyr-Pro-Leu-Lys” [PMID: 164350], the fourth residue “Ser” in the peptide is a phosphorylation site that can be correctly identified by regular expression “/appos \(((?: [A - Z][a - z]{2}) -) * ? [A - Z][a - z]{2}) - \d+ ' ? , P - \d+ ' ? \) /”.

2.2.4. Identification of GO terms, diseases, organisms, crosstalk

In order to identify GO terms in a sentence with PTM information by using MPTM tool, we generated a dictionary comprising 974 GO terms that describe functions or processes of PTMs. With this generated dictionary, MPTM identifies GO terms by using the longest match. MPTM also extracts diseases and organisms from text using a dictionary comprising about 1446 common diseases and eight species (human, mouse, rat, yeast, bovine, zebrafish, fly, virus). We also use patterns such as “crosstalk between PTM1 and PTM2” to match PTM crosstalk; in this study we found more than 300 abstracts mentioning crosstalk about PTMs.

2.2.5. Identification of diseases from OMIM

OMIM database is a comprehensive, authoritative compendium of human genes and genetic diseases. It is freely available and updated daily.³⁸ We obtained 5165 human disease records from OMIM database (March 21, 2015) and integrated them into MPTM tool. MPTM searches for human disease names based on a dictionary look up in the text from OMIM. Unfortunately, we cannot directly evaluate the performance of extracting relations between PTMs and diseases. However, it is still very useful to help users to find relations between PTMs and diseases by screening these potential diseases.

3. Result

3.1. Implementation and web interface

MPTM runs on a Linux machine with Apache 2.2.3 HTTP Server. The web server is implemented using the Perl language. Asynchronous JavaScript and XML/Document Object Model (AJAX/DOM) is used to obtain data from MPTMDB along with MySQL 5.1.73 and Cytoscape Web 1.0.4. Figure 3 shows the system architecture of MPTM tool. Through a web interface, inputs for MPTM are PMIDs or journal abstract texts. The web server provides three retrieval modes: PMID search, text search and keyword search. By employing PubMed’s API (E-Utilities), the keyword search enables users to easily retrieve interesting PTM information from PubMed. Given a valid input, MPTM produces structured output about PTM

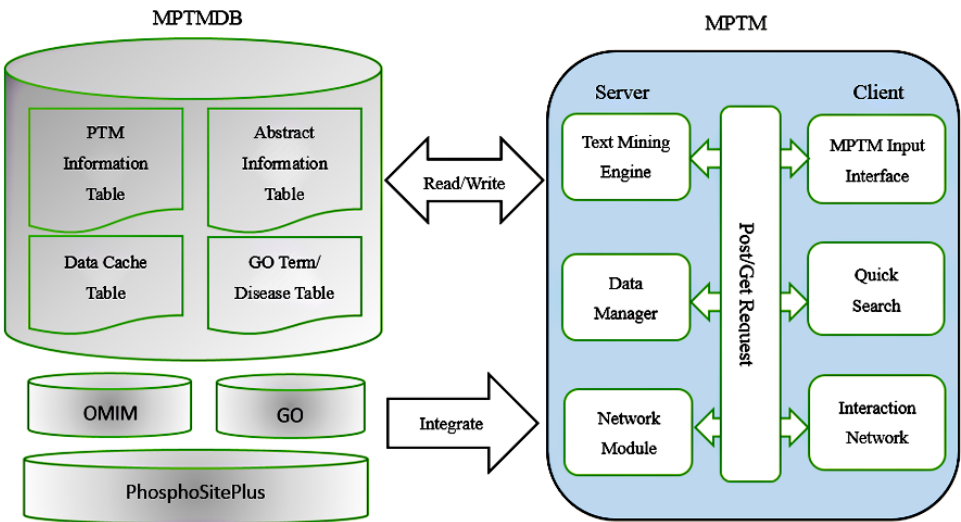


Fig. 3. The system architecture of MPTM contains three parts: Client, Server and MPTMDB. The Client provides an interactive interface of search functions. the Server includes a text mining engine, data manager and a network module which implements with Cytoscape Web 1.0.4. The pre-compiled PTM data called MPTMDB enables visual exploration and statistical summaries of substrate-enzyme interaction network of multiple PTMs.

information from the input text of abstracts. A typical structured output contains relevant PTM information from the input text of abstracts and relevant databases, including substrates, enzymes, sites, diseases, GO terms and referent information. First, each substrate, enzyme and site is extracted from text by the described approach. Next, disease names from the text are extracted and the potential disease names related to substrates and enzymes were imported from OMIM with their cross-links to the OMIM database. Then, the GO terms are extracted from the input text. MPTM also finds the crosstalk among PTMs by a sample matching rule. The output also includes references related to the identified terms. Finally, in order to give a summary of result data, a detailed section will show below the result by clicking the “Details” button. More details of the output from MPTM are available online at (<http://bioinformatics.ustc.edu.cn/mptm/tutorial.html>).

To further facilitate analysis of PTM data retrieved from literature, an online database named MPTMDB was built with 435,433 PTM records from more than 170,000 abstracts (Table 3), and is fully accessible for users via quick search for substrates and enzymes on the MPTM website. For example, the “p53” is one of the most important phosphorylation proteins. We extracted 2399 abstracts by using quick search with a query “p53” as a substrate in MPTM. It also provides sorting and filtering options for further view. A distinct feature of MPTMDB is that it enables visual exploration and statistical summaries of the substrate-enzyme interaction network of multiple PTMs. Since in various biological networks, substrate-enzyme

Table 3. Statistics of abstracts in MPTMDB.

Retrieved abstracts	178,446
Phosphorylation	103,633
Methylation	19,278
Glycosylation	18,600
Acetylation	8693
Amidation	1151
Hydroxylation	2412
Myristoylation	536
Sulfation	1396
GPI-anchor	2194
Disulfide	12,401
Ubiquitination	8152

interaction network construction and analysis is regarded as one of the most popular research fields, some important findings were made in this respect.^{44–46}

For users' convenience, a local package of MPTM is also provided, which can be integrated into new applications. The package was tested on two platforms, Windows 2007 and Linux (Ubuntu 12.04), and should work on similar systems that support Perl in a command-line environment. The package and user manual can be obtained from MPTM.

3.2. Evaluation

In order to provide a comprehensive evaluation, each sentence was used as a basic unit instead of the whole abstract. Sentences containing information of any type of 11 PTMs were defined as positive sentences and others were treated as negative sentences.³⁶ Given the fact that complete PTM information including substrate and site may not be always provided in one sentence, in this study we evaluated the identification performance for PTM substrate and site independently. Specifically, an element (substrate or site) was defined as true positive (TP) if it was correctly extracted in a positive sentence and otherwise was defined as false negative (FN). On the other hand, an element (substrate or site) in a positive sentence was defined as false positive (FP) if it was not related to PTM. We adopted four measurements commonly used in text-mining studies,⁴⁷ i.e. precision, recall, F-measure and Fowlkes–Mallows Index (FMI),⁴⁸ to assess the performance of MPTM. Precision is the fraction of true positives over the sum of true positives and false positives. Recall is the fraction of true positives over the sum of true positives and false negatives and the F-measure can be interpreted as a weighted average of the precision and recall. FMI is defined as the geometric mean between of the precision and recall. These formulas of performance measurements are given below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

$$\text{FMI} = \sqrt{\text{Precision} * \text{Recall}}. \tag{4}$$

In order to measure the MPTM tool performance, we compared the performance of MPTM with two state-of-the-art methods, RLIMS-P 2.0²³ and EAGL²⁶ (<http://eagl.unige.ch/PTM/>), using the same testing data which was described in detail in “Data resources” section. Here, due to the fact that site information was extracted by EAGL based on PTM type specified by user, for fair comparison we assessed the ability of extracting PTM-related information based on a given PTM type and the results are shown in Tables 4 and 5. MPTM obtains higher overall performance than EAGL for most of the PTM types and compares favorably to RLIMS-P 2.0 in Table 4. For example, for glycosylation site MPTM obtains the precision value of 0.84, while the precision value of EAGL is 0.69. In terms of sulfation site, the corresponding precision, recall and F-measure values of MPTM are 0.95, 0.80, 0.87, respectively, which are 4%, 14% and 10% higher than those of EAGL, respectively. At the same time, for phosphorylation site, the precision, recall and F-measure values of MPTM are 16%, 25% and 21% higher than those of EAGL, respectively. In addition, for the F-measure of phosphorylation substrate, MPTM achieves a comparable performance with RLIMS-P 2.0 while it is 1% lower than RLIMS-P 2.0 in term of recall of phosphorylation substrate. Meanwhile, the F-measure of phosphorylation site for MPTM is 0.93, which is slightly better than RLIMS-P 2.0. As to FMI in Table 5, MPTM achieves the best values and is consistently better than EAGL for most of the PTM types. Taken together, these results clearly demonstrate

Table 4. Performance evaluation for testing dataset with existing tools using F-measure.

	MPTM						EAGL			RLIMS-P2.0					
	Substrate			Site			Site			Substrate			Site		
	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F	Pre	Rec	F
Phosphorylation	0.91	0.88	0.90	0.96	0.90	0.93	0.80	0.65	0.72	0.91	0.89	0.90	0.92	0.90	0.91
Methylation	0.91	0.68	0.78	0.96	0.73	0.83	0.61	0.38	0.47	—	—	—	—	—	—
Glycosylation	0.88	0.68	0.77	0.84	0.85	0.84	0.69	0.68	0.69	—	—	—	—	—	—
Acetylation	0.85	0.75	0.80	0.96	0.87	0.91	0.76	0.43	0.55	—	—	—	—	—	—
Amidation	1.00	0.71	0.83	1.00	0.22	0.36	0.25	0.11	0.15	—	—	—	—	—	—
Disulfide	0.91	0.78	0.84	0.94	0.93	0.93	0.94	0.68	0.79	—	—	—	—	—	—
Sulfation	0.89	0.53	0.67	0.95	0.80	0.87	0.91	0.66	0.77	—	—	—	—	—	—
Hydroxylation	0.45	0.50	0.47	0.87	0.47	0.61	—	—	—	—	—	—	—	—	—
Myristoylation	0.92	0.71	0.80	0.75	1.00	0.86	—	—	—	—	—	—	—	—	—
GPI — anchor	0.83	0.70	0.76	1.00	0.85	0.92	—	—	—	—	—	—	—	—	—
Ubiquitination	0.67	0.71	0.70	0.87	0.39	0.54	—	—	—	—	—	—	—	—	—

Notes: Precision; Rec: Recall; F: F-measure. —: not applicable.

Table 5. Performance evaluation for testing dataset with existing tools using FMI.

	MPTM		EAGL	RLIMS-P2.0	
	Substrate FMI	Site FMI	Site FMI	Substrate FMI	Site FMI
Phosphorylation	0.89	0.93	0.72	0.90	0.91
Methylation	0.79	0.84	0.48	—	—
Glycosylation	0.77	0.84	0.68	—	—
Acetylation	0.80	0.91	0.57	—	—
Amidation	0.84	0.47	0.17	—	—
Disulfide	0.84	0.93	0.80	—	—
Sulfation	0.69	0.87	0.77	—	—
Hydroxylation	0.47	0.64	—	—	—
Myristoylation	0.81	0.87	—	—	—
GPI-anchor	0.76	0.92	—	—	—
Ubiquitination	0.69	0.58	—	—	—

Note: FMI: Fowlkes–Mallows Index. —: not applicable.

that MPTM is effective in extracting PTM-related information from literature. Moreover, we further assess the ability of our tool for identifying corresponding PTM type for each abstract, Table 6 shows the detailed performance of identifying PTM types. Recall varies according to PTMs from 96% for both phosphorylation and myristoylation to 71% for acetylation.

3.3. Substrate-enzyme interaction network

MPTM is designed to extract PTM information and assist the annotation of information on protein PTMs. Another potential application of MPTM is to provide a function for helping researchers to better study substrate-enzyme interactions in an interactive manner: quick search. In detail, at the “Network” module, the quick search allows users to obtain an interaction network between substrates and enzymes by entering one or more proteins as substrates separated by comma. Here an example

Table 6. The performance of MPTM to identify PTM types.

PTMs	Abstracts	True positive	False positive	False negative	Precision	Recall	F-measure	FMI
Phosphorylation	99	95	0	4	1.00	0.96	0.98	0.98
Methylation	46	39	2	7	0.95	0.85	0.90	0.90
Glycosylation	51	39	4	12	0.91	0.76	0.83	0.83
Acetylation	28	20	3	8	0.87	0.71	0.78	0.79
Amidation	46	39	5	7	0.89	0.85	0.87	0.87
Disulfide	73	56	6	17	0.89	0.77	0.83	0.83
Sulfation	32	24	1	8	0.96	0.75	0.84	0.85
Hydroxylation	67	62	2	5	0.97	0.93	0.95	0.95
Myristoylation	32	31	0	1	1.00	0.96	0.98	0.98
GPI-anchor	65	59	2	6	0.97	0.91	0.94	0.94
Ubiquitination	94	85	3	9	0.97	0.91	0.94	0.94

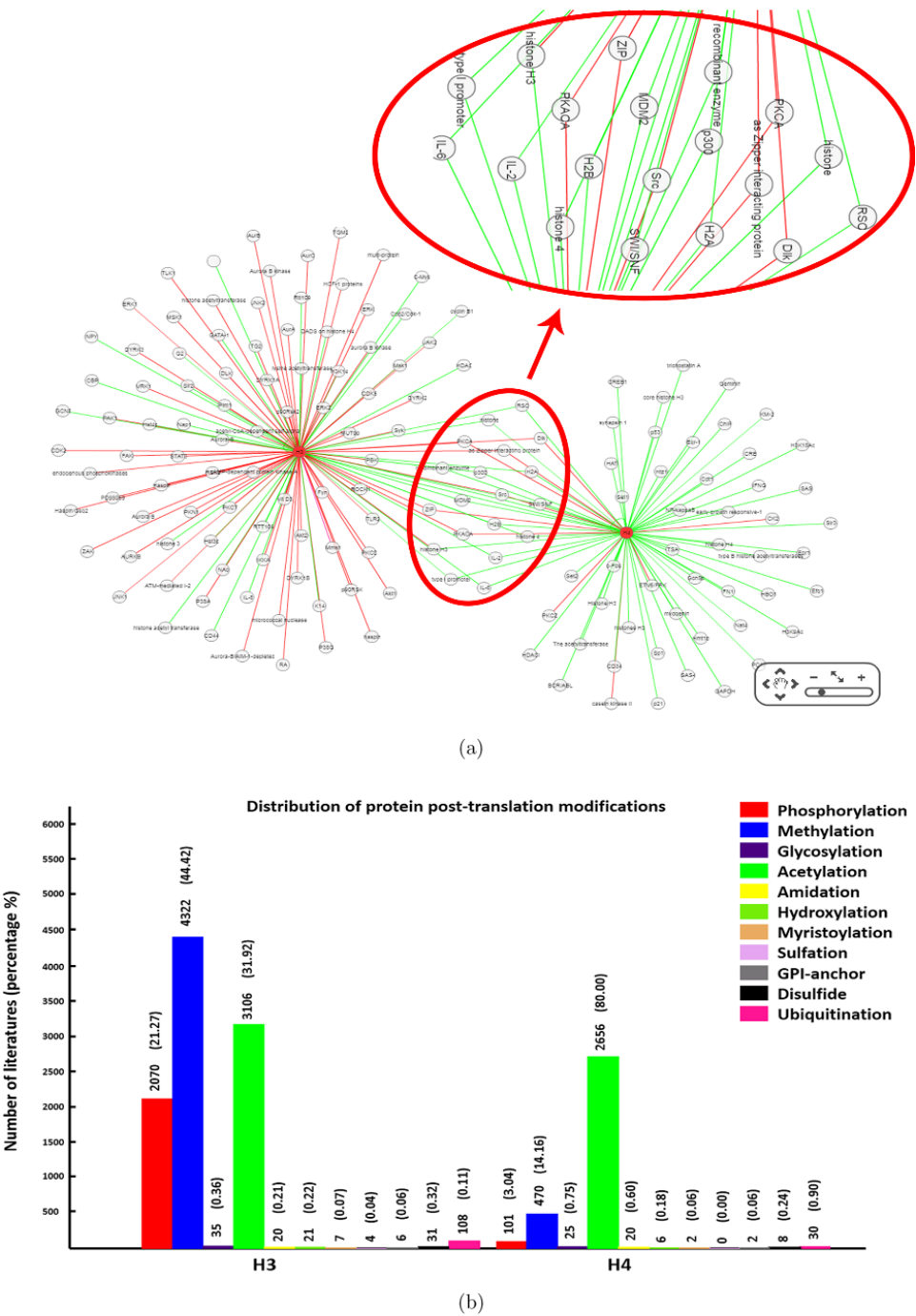


Fig. 4. An interaction network between two proteins. (a) An interaction network of substrate “H13” and substrate “H4”. (b) A distribution of protein post-translational modification with substrates “H3” and “H4”.

MPTM: A tool for mining protein post-translational modifications from literature

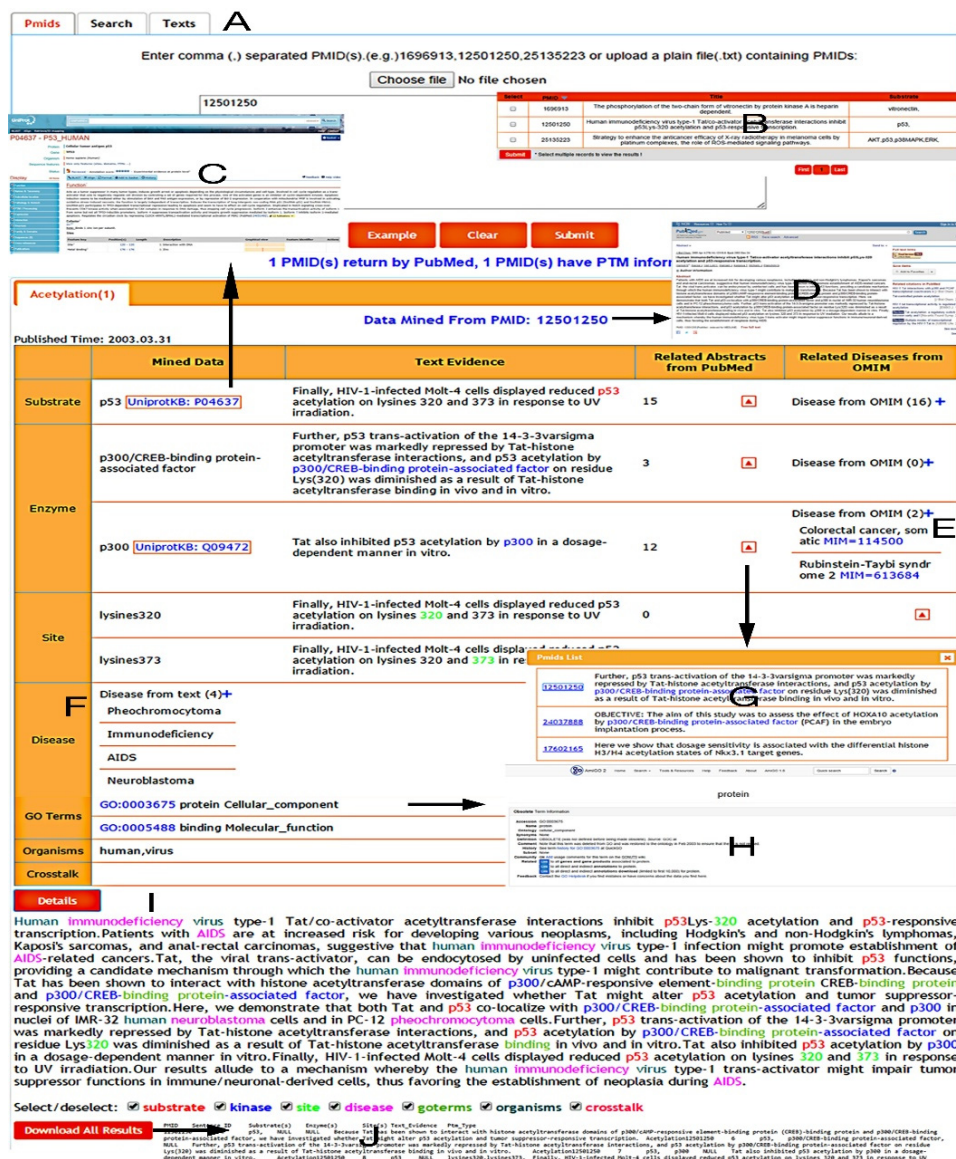


Fig. 5. A query example of MPTM for abstract from PubMed (PMID: 12501250). (A) User can select one of the three retrieval modes: PMID search, text search and keywords search. (B) An overview page includes some basic information in abstracts. (C) MPTM automatically associates the PTM protein with the UniProtKB protein entry by using GNAT tool. (D) A link to abstract in PubMed and full text article PubMed Central (PMC). (E) Visualization of the related diseases from OMIM database. (F) Visualization of the related diseases from the input text. (G) Visualization of the related abstracts from PubMed. (H) A list of GO terms related to the mined protein and each of them link to AminGo2. (I) A detail section with an option is provided for turning on/off the color-tagging for each type of object in the abstracts for visual inspection. (J) All extracted results can be directly downloaded from MPTM).

is also given to further help explain this function. First we entered protein name “H3” and “H4” separated by comma, then we can notice that there were many common enzymes that can interact with both proteins in Fig. 4(a). These common enzymes may give us some evidence in the research of the PTM-level association between different proteins. In addition, at the “Network” module, MPTM also gives a distribution diagram (Fig. 4(b)) of PTMs about entered proteins. From the distribution diagram, some meaningful and interesting information can be found. For example, in the diagram it can be seen that there were more than 4000 abstracts which had methylation information with protein “H3”, and 2000 abstracts that contained acetylation information with protein “H4”. Through these distribution statistics of PTMs about proteins, it may imply that there is a hot point in the current research of PTMs.

3.4. MPTM query example

Here, we show a query example of MPTM with an abstract from PubMed (PMID: 12501250) in Fig. 5, MPTM first identifies the PTM type of the abstract by triggers filtering, and then finds substrates, enzymes and sites from the abstract. In this case, MPTM successfully recognizes p53 as substrate, p300 and CREB-binding protein-associated factor as enzyme, and lysine 320 and 373 as site. MPTM also associates the p53 and p300 with the UniProtKB protein entry P04637 and Q09472, respectively (Fig. 5). In addition, MPTM extracts related diseases, such as “Adrenal cortical carcinoma” (MIM = 202300) and “Breast cancer” (MIM = 114480) related to p53, “Colorectal cancer, somatic” (MIM = 114500), “Rubinstein-Taybi syndrome 2” (MIM = 613684) related to p300, and “Pheochromocytoma”, “Immunodeficiency”, “AIDS” and “Neuroblastoma” from texts. It also finds GO terms related to PTMs. In the output (Fig. 5), the three columns: “Mined Data”, “Related Abstracts from PubMed” and “Text Evidence” show the PTM data identified from texts, the related abstracts from PubMed and the text evidences for the mined data, respectively. By clicking the “Details” button, a section with detailed information will appear below the results. All the results can be downloaded by clicking the “Download All Results” button.

4. Discussion

MPTM is designed to facilitate the extraction of protein PTM information from biological scientific literature in PubMed. The goal of the MPTM tool is to accelerate the work of the database curators, and its precision and recall are not necessary to be perfect but it should extract PTM information at an acceptable level. The key challenges for mining multiple type of PTMs are (1) how to efficiently extract comprehensive information including both substrates and sites from multiple type of PTMs; and (2) there were few available manually labeled dataset for hydroxylation, myristoylation and GPI-anchor. We have successfully used relations extracted from

dependency parse trees and a heuristic algorithm to handle multiple PTMs and extracted comprehensive information including both substrates and sites simultaneously. Meanwhile, since the detailed annotations of existing benchmarks for multiple types of PTMs were incomplete or unavailable, we have provided a manually labeled dataset including hydroxylation, myristoylation and GPI-anchor and developed an effective online tool for extracting and analyzing valuable information from multiple types of PTMs. MPTM provides three retrieval modes, PMID search, text search and keyword search, as well as a database called MPTMDB which facilitates visualization of the extracted information. It also integrates useful resources from other databases. The process is simple and can be integrated in many text mining workflows.

The performance of MPTM highly depends on the usefulness of relations and therefore efficient relation extraction method plays an important role in obtaining good performance for MPTM. The results of Tables 4 and 5 show that MPTM compares favorably with other exiting PTM-related text mining tools, indicating the efficiency of the relations as well as the relation extraction method in MPTM in detecting substrate and site information from the literature. In addition, we also perform analysis of two examples in our testing data. One example is the sentence “Dimethylation or trimethylation of Lys4 of histone H3 (H3 Lys4) has been found in expressible euchromatin of yeasts and mammals” [PMID: 12679815], for which MPTM successfully discriminates both the site “Lys4” and substrate “H3” by using two relations including “*appos*” (trigger, Lys4) [trigger = methylation] and “*prep-of*” (Lys4, PRO37) [PRO37 = H3]. Unfortunately, this methylation site cannot be recognized by the existing tool investigated in this study. Another example is the sentence “In histone H4 of eukaryotes ranging from fungi to man, acetylation is restricted to four lysines (residues 5, 8, 12, and 16) possibly caused by a quantitative methylation of lysine-20” [PMID: 1627562]. The existing tool investigated in this study can correctly detect the methylation site “lysine-20” but not the acetylation sites “(residues 5, 8, 12, and 16)”. Instead, MPTM recognizes all sites by relations extracted from dependency parse trees including “*prep-of*” (trigger, lysine-20) [trigger = methylation], “*num*” (residues, 5), “*num*” (residues, 8), “*num*” (residues, 12) and “*conj_and*” (residues, 16). These results further demonstrate the contribution of efficient relations for good performance of MPTM.

In fact, for some particular PTMs, it is difficult to improve precision except by the time-consuming design of very complex patterns and exclusion rules. For example, protein acetylation or methylation result in some false positive sentences, due to the frequent occurrence of the “acetyl” and “methy” tokens in the chemical compound names. Another type of errors occurs due to the detection of sentences including non-specific information, such as “Phosphorylation, a factor controlling the synthesis of L-erythrodihydrobiopterin (BH2)” (PMID: 697844). As to the extraction of modification sites, in order to indirectly identify the terms related to protein modification, we have designed some specific expressions to reduce the false negative sites, such as the amino acid N-terminal (PMID: 8645219) or C-terminal of the protein

(often occur at the PTMs such as methylation, acetylation and amidation). At the same time, it is also very difficult to identify a component consisting of a protein name and a modification site, since the component is recognized as a protein name in the NER (Named entity recognition) step and replaced by an abbreviated notation *PRO_i*, which results in missing site information. For example, the sentence like “histone H4 modifications with hyperacetylation of H4K5/H4K8, hypoacetylation of H4K12/H4K16” (PMID: 18974389), where “H4” is a protein name and “K5” is a site. In MPTM, the “H4K5” is wrongly recognized as a protein name in the NER step.

In our future work, we will consider optimizing our method to enhance the procedure for the identification of PTM information. The records in MPTM contain different alias or ambiguous names of substrates and enzymes automatically extracted from literature. The issue of named entity normalization will rise in converting different alias or ambiguous names to manually curated official names. In our study, we utilize GNAT⁴⁹ for converting different alias or ambiguous names to UniProtKB entry which also causes some inevitable mistakes. In the future, we plan to develop a feedback system: each record extracted from abstracts in MPTM can be corrected by users by an “Agree” and a “Reject” button, which will improve the accuracy of records in MPTMDB. We also consider adding an additional retrieval mode by integrating PDF interface for our MPTM in future work.

5. Conclusion

In our study, we have developed a user-friendly and efficient online tool for automatically extracting and organizing valuable knowledge about common PTMs from abstracts in PubMed by using relations extracted from dependency parse trees and a heuristic algorithm. MPTM is a worthy supplementary tool for other existing systems and an effective tool to help researchers and database curators easily find and retrieve PTM information. The development and expansion of MPTM will keep on going and more valuable features will be integrated into the web server in the future. As an efficient bioinformatics text mining tool, MPTM will assist in the discovery of novel substrate, modification site and enzyme in a broad spectrum of PTM research.

Acknowledgments

First and second authors contributed equally to the paper. This work was supported by National Natural Science Foundation of China (61471331, 61571414 and 61101061).

References

1. Mann M, Jensen ON, Proteomic analysis of post-translational modifications, *Nature Biotechnol* **21**(3):255–261, 2003.
2. Mínguez P, Letunic I, Parca L, Bork P, PTMcode: A database of known and predicted functional associations between post-translational modifications in proteins, *Nucleic Acids Res* **41**(D1):D306–D11, 2013.

3. Gong C-X, Liu F, Grundke-Iqbal I, Iqbal K, Post-translational modifications of tau protein in Alzheimer's disease, *J Neural Transm* **112**(6):813–838, 2005.
4. Zarkowska T, Mitnacht S, Differential phosphorylation of the retinoblastoma protein by G1/S cyclin-dependent kinases, *J Biol Chem* **272**(19):12738–12746, 1997.
5. Ardley HC, Robinson PA, The role of ubiquitin-protein ligases in neurodegenerative disease, *Neuro-Degener Dis* **1**(2–3):71–87, 2003.
6. Hu L-L, Niu S, Huang T, Wang K, Shi X-H, Cai Y-D, Prediction and analysis of protein hydroxyproline and hydroxylysine, *PLoS One* **5**(12):e15917, 2010.
7. Ismail HD, Newman RH, RF-hydroxysite: A random forest based predictor for hydroxylation sites, *Mol Biosyst* **12**(8):2427–2435, 2016.
8. Udenwobele DI, Su R-C, Good SV, Ball TB, Shrivastav SV, Shrivastav A, Myristoylation: An important protein modification in the immune response, *Front Immunol* **8**:751, 2017.
9. Wang S, Zhang Y-H, Cai Y-D, Huang G, Chen L, Analysis and prediction of myristoylation sites using the mRMR method, the IFS method and an extreme learning machine algorithm, *Comb Chem High Throughput Scr* **20**(2):96–106, 2017.
10. Kinoshita T, Glycosylphosphatidylinositol (GPI) anchors: Biochemistry and cell biology: Introduction to a thematic review series, *J Lipid Res* **57**(1):4–5, 2016.
11. Maeda Y, Murakami Y, Kinoshita T, Synthesis, genetics, and congenital diseases of GPI-anchored proteins, *Paroxysmal Nocturnal Hemoglobinuria*, Springer, p. 11–54, 2017.
12. UniProt C, Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Res* **40**(Database issue):D71–75, 2012, doi: 10.1093/nar/gkr981.
13. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii Ji (eds.), Overview of BioNLP'09 shared task on event extraction, *Proc Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pp. 1–19, 2009.
14. Kim J-D, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii Ji (eds.), Overview of BioNLP shared task 2011, *Proc BioNLP Shared Task 2011 Workshop*, pp. 1–6, 2011.
15. Björne J, Salakoski T (eds.), TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 shared task, *Proc BioNLP Shared Task 2013 Workshop*, pp. 16–25, 2013.
16. Nédellec C, Bossy R, Kim J-D, Kim J-J, Ohta T, Pyysalo S, Zweigenbaum P (eds.), Overview of BioNLP shared task 2013, *Proc BioNLP Shared Task 2013 Workshop*, pp. 1–7, 2013.
17. Van Landeghem S, Ginter F, Van de Peer Y, Salakoski T (eds.), EVEX: A PubMed-scale resource for homology-based generalization of text mining predictions, *Proc. BioNLP 2011 Workshop*, pp. 28–37, 2011.
18. Pletscher-Frankild S, Pallegà A, Tsafou K, Binder JX, Jensen LJ, DISEASES: Text mining and data integration of disease–gene associations, *Methods* **74**:83–89, 2015.
19. Cohen AM, editor Unsupervised gene/protein named entity normalization using automatically extracted dictionaries, *Proc. acl-ismb Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pp. 17–24, 2005.
20. Gerner M, Nenadic G, Bergman CM, LINNAEUS: A species name identification system for biomedical literature, *BMC Bioinform* **11**(1):85, 2010.
21. Leaman R, Doğan RI, Lu Z, DNorm: Disease name normalization with pairwise learning to rank, *Bioinformatics* **29**:474, 2013.
22. Hu Z-Z, Narayanaswamy M, Ravikumar K, Vijay-Shanker K, Wu CH, Literature mining and database annotation of protein phosphorylation using a rule-based system, *Bioinformatics* **21**(11):2759–2765, 2005.
23. Torii M, Arighi CN, Li G, Wang Q, Wu CH, Vijay-Shanker K, RLIMS-P 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information, *IEEE/ACM Trans Comput Bioinform* **12**(1):17–29, 2015.

24. Arighi CN, Siu AY, Tudor CO, Nchoutmboube JA, Wu CH, Shanker VK, eFIP: A tool for mining functional impact of phosphorylation from literature, *Bioinformatics for Comparative Proteomics*, Springer, pp. 63–75, 2011.
25. Xu Y, Teng D, Lei Y, MinePhos: A literature mining system for protein phosphorylation information extraction, *IEEE/ACM Trans Comput Biol Bioinform* **9**(1):311–315, 2012.
26. Veuthey A-L, Bridge A, Gobeill J, Ruch P, McEntyre JR, Bougueleret L, Xenarios I, Application of text-mining for updating protein post-translational modification annotation in UniProtKB, *BMC Bioinform* **14**(1):104, 2013.
27. Ross KE, Huang H, Ren J, Arighi CN, Li G, Tudor CO, Lv M, Lee J-Y, Chen S-C, Vijay-Shanker K, iPTMnet: Integrative bioinformatics for studying PTM networks, *Protein Bioinform* **1558**:333–353, 2017.
28. Ross KE, Natale DA, Arighi C, Chen S-C, Huang H, Li G, Ren J, Wang M, Vijay-Shanker K, Wu CH, Scalable text mining assisted curation of post-translationally modified proteoforms in the protein ontology, *CEUR Workshop Proc.*, 2016.
29. Natale DA, Arighi CN, Blake JA, Bult CJ, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Drabkin HJ, Helfer O, Protein Ontology: A controlled structured network of protein entities, *Nucleic Acids Res.* **42**(D1):D415–D21, 2013.
30. Natale DA, Arighi CN, Blake JA, Bona J, Chen C, Chen S-C, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Protein ontology (PRO): Enhancing and scaling up the representation of protein entities, *Nucleic Acids Res.* **45**(D1):D339–D46, 2016.
31. Hu Z-Z, Mani I, Hermoso V, Liu H, Wu CH, iProLINK: An integrated protein resource for literature mining, *Comput Biol Chem* **28**(5):409–416, 2004.
32. Bradley P, Daelemans W, Dehaspe L, Durant F, Hilario M, Kim J-H, Mitchell A, Petrak J, Pillet V, Seewald A, Van BK, Veuthey A, Zehnder M, BioMinT: A database curator's assistant for biomedical text processing, 2004.
33. Lee H, Yi G-S, Park JC, E3Miner: A text mining tool for ubiquitin-protein ligases, *Nucleic Acids Res* **36**(suppl.2):W416–W22, 2008.
34. Khoury GA, Baliban RC, Floudas CA, Proteome-wide post-translational modification statistics: Frequency analysis and curation of the swiss-prot database, *Sci. Rep.* **1**:90, 2011.
35. Fundel K, Küffner R, Zimmer R, RelEx — Relation extraction using dependency parse trees, *Bioinformatics* **23**(3):365–371, 2007.
36. Wang M, Xia H, Sun D, Chen Z, Wang M, Li A, Literature mining of protein phosphorylation using dependency parse trees, *Methods* **67**(3):386–393, 2014.
37. Torii M, Li G, Li Z, Oughtred R, Diella F, Çelen I, Arighi CN, Huang H, Vijay-Shanker K, Wu CH, RLIMS-P: An online text-mining tool for literature-based extraction of protein phosphorylation information, *Database (Oxford)* **2014**:bau081, 2014.
38. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA, Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders, *Nucleic Acids Res* **33** (suppl 1):D514–D517, 2005.
39. Leaman R, Gonzalez G (eds.) BANNER: An executable survey of advances in biomedical named entity recognition, *Pacific Symp. Biocomput*, Vol. 13, pp. 652–663, 2008.
40. De Marneffe M-C, MacCartney B, Manning CD (eds.), Generating typed dependency parses from phrase structure parses, *Proc. LREC*, 2006.
41. Carroll J, Minnen G, Briscoe T, Corpus annotation for parser evaluation, arXiv preprint cs/9907013, 1999.
42. De Marneffe M-C, Manning CD (eds.), The Stanford typed dependencies representation. Coling 2008, *Proc Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8, 2008.
43. De Marneffe M-C, Manning CD, Stanford typed dependencies manual, Technical report, Stanford University, 2008.

44. Yamada T, Bork P, Evolution of biomolecular networks — lessons from metabolic and protein interactions, *Nat Rev Mol Cell Biol* **10**(11):791–803, 2009.
45. Ingham RJ, Colwill K, Howard C, Dettwiler S, Lim CS, Yu J, Hersi K, Raaijmakers J, Gish G, Mbamalu G, WW domains provide a platform for the assembly of multiprotein networks, *Mol Cell Biol* **25**(16):7092–7106, 2005.
46. Xu X, Wang M, Inferring disease associated phosphorylation sites via random walk on multi-layer heterogeneous network, *IEEE/ACM Trans Comput Biol Bioinform* **13**(5):836–844, 2016.
47. Larsen B, Aone C (eds.), Fast and effective text mining using linear-time document clustering, *Proc Fifth ACM SIGKDD Int Conf Knowledge Discovery and Data Mining*, ACM, pp. 16–22, 1999.
48. Fowlkes EB, Mallows CL, A method for comparing two hierarchical clusterings, *J Amer Statist Assoc* **78**(383):553–569, 1983.
49. Hakenberg J, Gerner M, Haeussler M, Solt I, Plake C, Schroeder M, Gonzalez G, Nenadic G, Bergman CM, The GNAT library for local and remote gene mention normalization, *Bioinformatics* **27**(19):2769–2771, 2011.



Dongdong Sun received his BS degree in Electronic Information Engineering from Auhui University, China in 2013. He is currently pursuing the Ph.D. degree in Biomedical Engineering. He is a part of the Centers for Biomedical Engineering, University of Science and Technology of China (USTC). His research interests include the text mining, and machine learning, bioinformatics and biostatistics.



Minghui Wang received the BS degree from School of Gifted Youth, University of Science and Technology of China (USTC), and the Ph.D. degree in Biomedical Engineering from School of Information Science and Technology, USTC, in 2006. She is an associated professor in School of Information Science and Technology and Centers for Biomedical Engineering, USTC. Her research interests include bioinformatics, biostatistics and machine learning.



Ao Li received the BS degree in Biophysics from School of Life Science, University of Science and Technology of China (USTC), in 2000 and the Ph.D. degree in Biomedical Engineering from School of Information Science and Technology, USTC, in 2005. Currently, he is an associated professor in School of Information Science and Technology and Centers for Biomedical Engineering, USTC. His research contributions are in computational cancer genomics and bioinformatics with a focus on issues concerning systematic identification and evaluation of genome- wide variants in cancer.