PubMed-Scale Event Extraction for Post-Translational Modifications, Epigenetics and Protein Structural Relations

Jari Björne [1,2], **Sofie Van Landeghem** [3,4], **Sampo Pyysalo** [5], **Tomoko Ohta** [5], **Filip Ginter** [2], **Yves Van de Peer** [3,4], **Sophia Ananiadou** [5] and **Tapio Salakoski** [1,2]

[1]Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3-5B, 20520 Turku, Finland
[2]Department of Information Technology, 20014 University of Turku, Finland
[3]Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium
[4]Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium
[5]National Centre for Text Mining and University of Manchester,
Manchester Interdisciplinary Biocentre,131 Princess Street, Manchester, UK

Abstract

Recent efforts in biomolecular event extraction have mainly focused on core event types involving genes and proteins, such as gene expression, protein-protein interactions, and protein catabolism. The BioNLP'11 Shared Task extended the event extraction approach to sub-protein events and relations in the Epigenetics and Post-translational Modifications (EPI) and Protein Relations (REL) tasks. In this study, we apply the Turku Event Extraction System, the best-performing system for these tasks, to all PubMed abstracts and all available PMC full-text articles, extracting 1.4M EPI events and 2.2M REL relations from 21M abstracts and 372K articles. We introduce several entity normalization algorithms for genes, proteins, protein complexes and protein components, aiming to uniquely identify these biological entities. This normalization effort allows direct mapping of the extracted events and relations with post-translational modifications from UniProt, epigenetics from PubMeth, functional domains from InterPro and macromolecular structures from PDB. The extraction of such detailed protein information provides a unique text mining dataset, offering the opportunity to further deepen the information provided by existing PubMed-scale event extraction efforts. The methods and data introduced in this study are freely available from `bionlp.utu.fi`.

1 Introduction

Biomedical domain information extraction has in recent years seen a shift from focus on the extraction of simple pairwise relations (Pyysalo et al., 2008; Tikk et al., 2010) towards the extraction of *events*, represented as structured associations of arbitrary numbers of participants in specific roles (Ananiadou et al., 2010). Domain event extraction has been popularized in particular by the BioNLP Shared Task (ST) challenges in 2009 and 2011 (Kim et al., 2009; Kim et al., 2011). While the BioNLP ST'09 emphasized protein interactions and regulatory relationships, the expressive event formalism can also be applied to the extraction of statements regarding the properties of individual proteins. Accordingly, the EPI (Epigenetics and Post-Translational Modifications) subchallenge of the BioNLP ST'11 provided corpora and competitive evaluations for the detection of epigenetics and post-translational modification (PTM) events, while the REL (Entity Relations) subchallenge covers structural and complex membership relations of proteins (Ohta et al., 2011b; Pyysalo et al., 2011). The complex memberships and domains define the physical nature of an individual protein, which is closely linked to its function and biological activity. Post-translational modifications alter and regulate this activity via structural or chemical changes induced by the covalent attachment of small molecules to the protein. In epigenetic regulation, gene expression is controlled by the chemical modification of DNA and the histone proteins supporting chromosomal DNA. All of these aspects are important for defining the biological role of a protein, and thus the EPI and REL tasks enable the development of text mining systems that can extract a more complete picture of the biomolecular reactions and relations than previously possible (cf. Table 1). Furthermore, previous work has shown promising results for improving event extraction by

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 82–90,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

integration of "static" entity relations (Pyysalo et al., 2009), in particular for the previously only available PTM event, phosphorylation (Van Landeghem et al., 2010).

Information on protein modifications is available in general-purpose protein databases such as UniProt, and there are also a number of dedicated database resources covering such protein modifications (Wu and others, 2003; Lee et al., 2006; Li et al., 2009). While the automatic extraction of PTMs from text has also been considered in a number of earlier studies, these have primarily involved single PTM reactions extracted with special-purpose methods (Hu et al., 2005; Yuan et al., 2006; Lee et al., 2008). The EPI task and associated work (Ohta et al., 2010) were the first to target numerous PTM reactions in a general framework using retrainable extraction methods. The automatic detection of modification statements using keyword matching-based methods has been applied also in support of DNA methylation DB curation (Ongenaert et al., 2008; Fang et al., 2011). However, as for PTM, the EPI task and its preparatory efforts (Ohta et al., 2011a) were the first to consider DNA methylation using the general event extraction approach. To the best of our knowledge, the present study is the first to extend the event extraction approach to PTM and DNA methylation event extraction to the scale of the entire available literature.

The Turku Event Extraction System (TEES), first introduced for the BioNLP ST'09 (Björne et al., 2009), was updated and generalized for participation in the BioNLP ST'11, in which it had the best performance on both the EPI and REL challenges (Björne and Salakoski, 2011). With an F-score of 53.33% for the EPI and 57.7% for the REL task, it performed over 16 pp better than the next best systems, making it well suited for our study. We apply this system to the extraction of EPI events and REL relations from all PubMed abstracts and all PMC open access articles, using a pipeline of open source text mining tools introduced in Björne et al. (2010).

We further process the result using a recently created bibliome-scale gene normalization dataset[1]. This normalization effort connects protein and gene mentions in text to their database IDs, a prerequisite for effective use of text mining results for most bioinformatics applications. In addition to protein names, the EPI and REL challenges refer to the protein substructures, modifications and complexes, which we also need to normalize in order to determine the biological context of these events. In this work, we develop a number of rule-based algorithms for the normalization of such non-protein entities.

With both proteins and other entities normalized, we can align the set of events extracted from the literature with biological databases containing annotations on protein features, such as UniProt. We can determine how many known and unknown features we have extracted from text, and what percentage of various protein feature annotations our text mining results cover. This association naturally also works in the other direction, as we can take a gene or protein and find yet unannotated post-translational modifications, domains, or other features from scientific articles, a promising use case for supporting biomedical database curation.

2 Methods

2.1 PMC preprocessing

PMC full texts are distributed in an XML format that TEES cannot use directly for event extraction. We convert this XML into a flat ASCII text format with a pipeline built on top of BioNLP ST'11 supporting resource tools (Stenetorp et al., 2011). This processing resolves embedded LaTeX expressions, separates blocks of text content (titles, sections, etc.) from others, maps non-ASCII characters to corresponding ASCII sequences, and normalizes whitespace. Resolving non-ASCII characters avoids increased error rates from NLP tools trained on ASCII-only data.

2.2 Event Extraction

We use the Turku Event Extraction System for extracting both REL relations and EPI events. TEES is a modular event extraction pipeline, that has recently been extended for all the subtasks of the BioNLP'11 ST, including EPI and REL (Björne and Salakoski, 2011). TEES performs all supported tasks using a shared graph scheme, which can represent both events and relations (Figure 1 D). The system also provides confidence scores enabling selection of the most likely correct predictions. Before event extrac-

[1] Data currently under review.

Event/relation type	Example
Hydroxylation	*HIF-alpha* proline **hydroxylation**
Phosphorylation	(D) siRNA-mediated ATM depletion blocks *p53* Serine-15 **phosphorylation**.
Ubiquitination	K5 **ubiquitinates** *BMPR-II* on a Membrane-proximal Lysine
DNA methylation	*RUNX3* is frequently inactivated by P2 **methylation** in solid tumors.
Glycosylation	Also, two asparagine residues in *alpha-hCG* were **glycosylated**.
Acetylation	This interaction was regulated by *Tat* **acetylation** at lysine 50.
Methylation	**Methylation** of lysine 37 of *histone H2B* is conserved.
Catalysis	*GRK2* **catalyzed** modest phosphorylation of BAC1.
Protein-Component	Three enhancer elements are located in the 40 kb intron of the *GDEP gene*.
Subunit-Complex	The most common form is a heterodimer composed of the *p65*/p50 subunits.

Table 1: Sentences with examples of the eight EPI event and two REL relation types, with highlighted **triggers**, and *protein* and site arguments. Relations have no trigger and Catalysis takes as an argument another event.

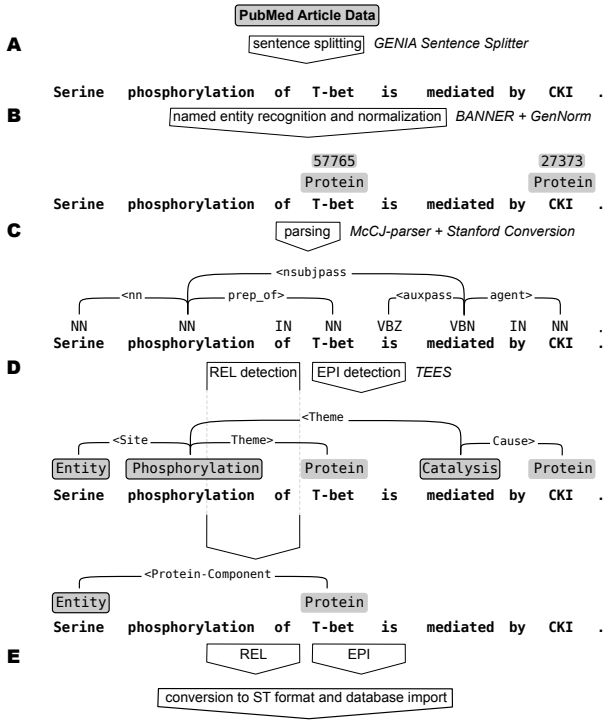

Figure 1: Event and relation extraction. Article text is split into sentences (A), where gene/protein entities are detected and normalized to their Entrez Gene IDs (B). Each sentence with at least one entity is then parsed (C). EPI events and REL relations are extracted from the parsed sentences (D) and following conversion to the BioNLP ST format are imported into a database (E). (Adapted from Björne and Salakoski (2011)).

tion, protein/gene names are detected and sentences are parsed. TEES handles all these preprocessing steps via a pipeline of tool wrappers for the GENIA Sentence Splitter (Kazama and Tsujii, 2003), the BANNER named entity recognizer (Leaman and Gonzalez, 2008), the McClosky-Charniak-Johnson (McCCJ) parser (Charniak and Johnson, 2005; McClosky, 2010) and the Stanford tools (de Marneffe et al., 2006). For a detailed description of TEES we refer to Björne and Salakoski (2011) and for the computational requirements of PubMed-scale event extraction to Björne et al. (2010).

2.3 Entity normalization

The extraction of events and relations as described in the previous sections is purely text-based and does not rely on any domain information from external resources. This ensures generalizability of the methods to new articles possibly describing novel interactions. However, practical use cases often require integration of text mining results with external resources. To enable such an integration, it is crucial to link the retrieved information to known gene/protein identifiers. In this section, we describe how we link text mining data to biomolecular databases by providing integration with Entrez Gene, UniProt, InterPro and the Protein Data Bank.

2.3.1 Protein annotations

A crucial step for integrating statements in domain text with data records is gene name normalization As part of a recent PubMed-scale effort,[2] gene

[2]Data currently under review.

normalizations were produced by the GenNorm system (Wei and Kao, 2011), assigning unique Entrez Gene identifiers (Sayers and others, 2010) to ambiguous gene/protein symbols. The GenNorm system represents the state-of-the-art in gene normalization, having achieved first rank by several evaluation criteria in the BioCreative III Challenge (Lu and others, 2011).

For practical applications, the Entrez Gene identifiers have been mapped to UniProt (The UniProt Consortium, 2011) through conversion tables provided by the NCBI. As Entrez Gene and UniProt are two of the most authoritative resources for gene and protein identification, these annotations ensure straightforward integration with other databases.

2.3.2 Complex annotations

The REL task Subunit-Complex relations all involve exactly one protein complex and one of its subunits, but the same complex may be involved in many different Subunit-Complex relations (Pyysalo et al., 2011). A key challenge for making use of these relations thus involves retrieving a unique identification of the correct complex. To identify protein complexes, we use the Protein Data Bank (PDB), an archive of structural data of biological macromolecules (Berman et al., 2000). This resource currently contains more than 80,000 3-D structures, and each polymer of a structure is annotated with its respective UniProt ID.

To assign a unique PDB ID to an entity involved in one or more Subunit-Complex relations, there is usually no other lexical context than the protein names in the sentence, e.g. "*the Rad9-Hus1-Rad1 complex*". Consequently, we rely on the normalized protein names (Section 2.3.1) to retrieve a list of plausible complexes, using data downloaded from UniProt to link proteins to PDB entries. Ambiguity is resolved by selecting the complex with the highest number of normalized proteins and giving preference to so-called representative chains. A list of representative chains is available at the PDB website, and they are determined by clustering similar protein chains[3] and taking the most confident ones based on resolution quality.

Each assignment of a PDB identifier is annotated with a confidence value between 0 and 1, expressing the percentage of proteins in the complex that could be retrieved and normalized in text. For example, even if one out of three UniProt identifiers is wrongly assigned for a mention, the correct complex might still be assigned with 0.66 confidence.

2.3.3 Domain annotations

Protein-Component relations define a relation between a gene/protein and one of its components, such as a gene promoter or a protein domain. To identify at least a substantial subset of these diverse relations, we have integrated domain knowledge extracted from InterPro. InterPro is a rich resource on protein families, domains and functional sites, integrating data from databases like PROSITE, PANTHER, Pfam, ProDom, SMART and TIGRFAMs (Hunter and others, 2012). Over 23,000 distinct InterPro entries were retrieved, linking to more than 16.5 million protein identifiers.

To assign an InterPro ID to an entity involved in one or more Protein-Component relations, a set of candidates is generated by inspecting the InterPro associations of each of the proteins annotated with that domain in text. For each such candidate, the description of the InterPro entry is matched against the lexical context around the entity by comparing the number of overlapping tokens, excluding general words, such as *domain*, and prepositions. The amount of overlap is normalized against the length of the InterPro description and expressed as a percentage, creating confidence values between 0 and 1.

Additionally, a simple pattern matching algorithm recognizes statements expressing an amino acid interval, e.g. "*aristaless domain (aa 527-542)*". When such expressions are found, the intervals as annotated in InterPro are matched against the retrieved interval from text, and the confidence values express the amount of overlap between the two intervals.

2.3.4 PTM site normalization

Six of the eight[4] EPI event types refer to post-translational modification of proteins. These events are *Hydroxylation*, *Phosphorylation*, *Ubiquitination*, *Glycosylation*, *Acetylation* and *(Protein) Methylation*. To evaluate the events predicted

[3] Requiring at least 40% sequence similarity.

[4] As we are interested in PTM sites, we make no distinction between "additive" PTMs such as *Acetylation* and their "reverse" reactions such as *Deacetylation*.

from text, we compare these to annotated post-translational modifications in UniProt. UniProt is one of the largest manually curated databases for protein knowledge, and contains annotations corresponding to each of the EPI PTM event types.

We use the reviewed and manually annotated UniProtKB/Swiss-Prot dataset (release 2012_02) in XML format. We take for each protein all *feature* elements of types *modified residue*, *cross-link* and *glycosylation site*. Each of these feature elements defines the site of the modification, either a single amino acid, or a sequence of amino acids. We select only annotations based on experimental findings, that is, features that do not have a non-experimental status (*potential*, *probable* or *by similarity*) to avoid e.g. features only inferred from the sequence.

The *modified residue* feature type covers the event types *Hydroxylation*, *Phosphorylation*, *Acetylation* and *Methylation*. We determine the class of the modification with the UniProt controlled vocabulary of post-translational modifications[5]. The *description* attribute is the ID attribute of an entry in the vocabulary, through which we can determine the more general keyword (KW) for that description, if defined. These keywords can then be connected to the corresponding event types in the case of *Hydroxylation*, *Phosphorylation*, *Acetylation* and *Methylation*. For *Ubiquitination* events, we look for the presence of the string "ubiquitin" in the *description* attribute of *cross-link* features. Finally, features corresponding to *Glycosylation* events are determined by their feature element having the type *glycosylation site*.

The result of this selection process is a list of individual modification features, which contain a type corresponding to one of the EPI PTM event types, the UniProt ID of the protein, and the position and amino acid(s) of the modification site. This data can be compared with extracted events, using their type, normalized protein arguments and modification site arguments. However, we also need to normalize the modification site arguments.

PTM sites are defined with a modification type and the numbered target amino acid residue. In EPI events, these residues are defined in the *site* argument target entities. To convert these into a form that can be aligned with UniProt, we apply a set of rules that try to determine whether a site is an amino acid. We start from the main site token, and check whether it is of the form *AA#*, where *AA* is an amino acid name, or a one or three letter code, and # an optional site number, which can also be in a token following the amino acid. For cases where the *site* entity is the word "residue" or "residues", we look for the amino acid definition in the preceding and following tokens. All strings are canonicalized with removal of punctuation, hyphens and parenthesis before applying the rules. In total, of the 177,994 events with a site argument, 75,131 could be normalized to an amino acid, and 60,622 of these to a specific residue number.

[5] http://www.uniprot.org/docs/ptmlist/

Event Type	Extracted	PMC (%)
Hydroxylation	14,555	34.17
Phosphorylation	726,757	44.00
Ubiquitination	74,027	70.46
DNA methylation	140,531	52.27
Glycosylation	154,523	42.31
Acetylation	114,585	69.40
Methylation	122,015	74.86
Catalysis	45,763	67.86
Total EPI	1,392,756	51.53
Protein-Component	1,613,170	52.59
Subunit-Complex	537,577	51.18
Total REL	2,150,747	52.23

Table 2: Total number of EPI events and REL relations extracted from PubMed abstracts and PMC full-text articles, with the fractions extracted from PMC.

3 Results

The source for extraction in this work is the set of 21 million PubMed abstracts and 372 thousand PMC open-access full-text articles. From this dataset, 1.4M EPI events and 2.2M REL relations were extracted (Table 2). For both tasks, about half of the results were extracted from PMC, confirming that full-text articles are an important source of information for these extraction targets. The total numbers of events and relations are considerably lower than e.g. the 21.3M events extracted for the GENIA task from PubMed abstracts (Björne et al., 2010; Van Landeghem et al., 2012), likely relating to the comparatively low frequency with which EPI and REL extraction targets are discussed with respect to the basic GENIA biomolecular reactions.

Event type	UniProt	Events	Match	Coverage	Events (site)	Match	Coverage
Hydroxylation	1,587	14,555	1,526	19	4,298	130	5
Phosphorylation	57,059	726,757	286,978	4,795	86,974	9,732	748
Ubiquitination	792	74,027	4,994	143	10,562	54	20
Glycosylation	6,708	154,523	18,592	897	22,846	68	31
Acetylation	6,522	114,585	15,470	764	25,689	158	30
Methylation	1,135	122,015	2,178	113	27,625	36	10
Total	73,803	1,206,462	329,738	6,731	177,994	10,178	844

Table 3: PTM events. PTMs that are not marked with non-experimental qualifiers are taken from UniProt. The *Events* column lists the total number of predicted events, and the *Events (site)* the number of events that also have a predicted site-argument. For these groups, *Match* is the number of events that matches a known PTM from UniProt, and *Coverage* the number of UniProt PTMs for which at least one match exists. For *Events* matching takes into account the PTM type and protein id, for *Events (site)* also the amino acid and position of the modified residue.

Event type	AA	UP	#	Highest confidence event	Article ID
Phosphorylation	S9	•	2	*p53* isolated from ML1, HCT116 and RKO cells, after short term genotoxic stress, were **phosphorylated** on Ser 6, Ser 9	PMC:2777442
Acetylation	S15		4	phosphorylated (Ser15), **acetylated** *p53*(Lys382)	PMC:2557062
Methylation	S15		1	**phosphorylation** of *p53* at serine 15 and acetylation	PM:10749144
Phosphorylation	S15	•	238	Chk2, as well as *p53* Ser(15) **phosphorylation** and its	PM:16731759
Phosphorylation	T18	•	12	*p53* stabilization and its **phosphorylation** in Thr18	PMC:3046209
Phosphorylation	S20	•	45	that **phosphorylation** of *p53* at Ser20 leads to	PMC:3050855
Phosphorylation	S33	•	14	**phosphorylation** of *p53* at serine 33 may be part of	PMC:35361
Phosphorylation	S37	•	20	serine 33 of *p53* in vitro when serine 37 is already	PMC:35361
Phosphorylation	S46	•	55	**phosphorylation** of *p53*, especially at Serine 46 by	PMC:2634840
Phosphorylation	T55	•	7	that **phosphorylation** of *p53* at Thr55 inhibits its	PMC:3050855
Phosphorylation	S99	•	0		
Phosphorylation	S183	•	0		
Phosphorylation	S269	•	0		
Phosphorylation	T284	•	0		
Ubiquitination	K291	•	0		
Acetylation	K292	•	0		
Ubiquitination	K292	•	0		
Acetylation	K305	•	0		
Phosphorylation	S313	•	1	**hyperphosphorylation** of *p53*, particularly of Ser313	PM:8649812
Phosphorylation	S314	•	0		
Phosphorylation	S315	•	6	to require **phosphorylation** of *p53* at serine 315 (35)	PMC:2532731
Methylation	K370	•	6	by **methylating** lysine 370 of *p53*	PMC:1636665
Acetylation	K372		1	for lysine 372 and 383 **acetylated** *p53* (Upstate,	PMC:1315280
Methylation	K372	•	5	**methylation** of *p53* by the KMT7(SET7/9) methyltransferase enzyme on Lys372	PMC:2794343
Acetylation	K373	•	16	p53 and **acetylated** *p53* (lysine-373 and lysine-382)	PMC:1208859
Methylation	K373	•	4	EHMT1-mediated *p53* **methylation** at K373	PM:20588255
Acetylation	K381	•	0		
Acetylation	K382	•	82	*p53* **acetylation** at lysine 382 was found not	PM:17898049
Methylation	K382	•	6	SET8 specifically **monomethylates** *p53* at lysine 382	PM:17707234
Methylation	K386	•	1	that **sumoylation** of *p53* at K386 blocks subsequent	PM:19339993
Phosphorylation	S392	•	35	and **phosphorylation** of *p53* at S392	PM:17237827

Table 4: Extracted and known PTM sites of *p53*. The type and site of the modification are in the first two columns. *UP* indicates whether the PTM is present in the UniProt annotation for *p53*. Column # shows the number of extracted events, followed by the event with the highest confidence score and the PubMed abstract or PMC full-text article it has been extracted from.

3.1 Extracted PTMs compared to UniProt

The EPI PTM events were compared to annotated PTMs from UniProt (Table 3). The majority of extracted PTM events (85%) have only a protein argument, and no information about the modification site, so these can only be compared by the protein id and PTM type. For the subset of proteins that also have a site, which can be normalized to an amino acid position, we can make a detailed comparison with UniProt. Finding a match for these normalized amino acids is more difficult, and for both categories, only a small fraction of proteins from UniProt is covered. In part this may be due to the limitations of the gene name normalization, as finding the exact species-specific protein ID remains a challenging task (Lu and others, 2011). However, even if the overall coverage is limited, well-known protein modifications can be assigned to specific residues, as we show in the next section.

3.2 Extracted PTMs for a single protein

For an in-depth example of PTM modifications, we study the protein *p53*, a central tumor suppressor protein that is the subject of many studies. *p53* is also among the proteins with the most UniProt PTM sites for which EPI events were predicted, making it a good example for a case study (see Table 4).

We take from UniProt all known *p53* PTMs corresponding to our EPI event types and list the number of predicted events for them (see Table 4). When the number of predicted events is high, the most confident prediction is usually a correctly extracted, clear statement about the PTM. All events for PTMs known in UniProt are correct except for the type of K386. For events not in UniProt, the two S15 ones are false positives, and K372 acetylation, while correctly extracted, is most likely a typo in the article. For the PTMs for which no event was extracted, we checked the reference article from UniProt annotation. K291, K292 ubiquitination, and K305 are from abstracts, and thus missed events. S183, S269 and T284 are from a non-open access PMC article, while S99, K292 acetylation, K305, S314 and K381 are from Excel or PDF format supplementary tables, sources outside our extraction input.

In total, we have extracted 561 PTM events related to *p53*, 554 of which correspond to a PTM annotated in UniProt. Of the 28 EPI-relevant PTMs on *p53*, 17 have at least one predicted event. The highest confidence events are about equally often from abstracts as from full texts.

Item	PubMeth	Extracted	Recall
PMID+UPID	2776	1698	61.2%
UPID	392	363	92.6%
PMID	1163	1120	96.3%

Table 5: Evaluation of DNA methylation event extraction recall against PubMeth.

3.3 DNA methylation analysis

Two recently introduced databases, PubMeth (Ongenaert et al., 2008) and MeInfoText (Fang et al., 2011) provide manually curated information on DNA methylation, primarily as it relates to cancer. To evaluate the coverage of DNA methylation event extraction, we focus here on PubMeth, as the full content of this database could be directly used. Each PubMeth DB record provides the primary name of the methylated gene and the PMID of the publication supporting the curation of the record. We used these two pieces of information to evaluate the recall [6] of DNA methylation event extraction.

We mapped PubMeth entries to UniProt identifiers (UPIDs), and extracted all unique (PMID, UPID) pairs from both PubMeth and the automatically extracted DNA methylation/demethylation events. The results of comparison of these sets of ID pairs are given in Table 5. We find that for over 60% of PubMeth entries, the system is able to recover the specific (document, gene) pair. This result is broadly in line with the recall of the system as evaluated in the BioNLP ST. However, if the matching constraint is relaxed, asking either 1) can the system extract the methylation of each gene in PubMeth *somewhere* in the literature or, inversely, 2) can the system detect *some* DNA methylation event in each document included in PubMeth as evidence, recall is over 90%. In particular, the evaluation indicates that the system shows very high recall for identifying documents discussing DNA methylation.

[6] As PubMeth does not aim for exhaustive coverage, precision cannot be directly estimated in this way. For example, PubMeth covers fewer than 2,000 documents and DNA methylation events were extracted from over 20,000, but due to differences in scope, this does not suggest precision is below 10%.

REL Type	Extracted	Match (p)	Match (e)
Prot-Cmp	1613.1K	561.8K	150.7K
SU-Cmplx	537.6K	226.5K	99.6K

Table 6: Numbers of extracted entity relations, with the protein (p) or both protein and entity (e) identified.

3.4 REL statistics

Table 6 presents the amount of extracted entity relations and the coverage of the normalization algorithms assigning protein, domain and complex identifiers. From a total of 537.6K Subunit-Complex relations, 226.5K (42%) involve a protein that could be unambiguously identified (Section 2.3.1). From this subset, 99.6K relations (44%) could be assigned to a PDB complex identifier (Section 2.3.2), accounting for 3800 representative 3D protein structures.

The Protein-Component relations are much more frequent in the data (1.6M relations) and here 35% of the relations (561.8K) involve a normalized protein mention. The assignment of InterPro domains to these Protein-Component relations (Section 2.3.3) further covers 150.7K relations in this subset (27%), identifying 5500 distinct functional domains. The vast majority of these annotations (99%) are produced by matching the lexical context against the InterPro descriptions, and only a few cases (200) matched against the amino-acid pattern.

4 Conclusions

We have combined state-of-the-art methods for gene/protein name normalization together with the best available methods for event-based extraction of protein post-translational modifications, reactions relating to the epigenetic control of gene expression, and part-of relations between genes/proteins, their components, and complexes. These methods were jointly applied to the entire available literature, both PubMed abstracts and PMC full-text documents, creating a text mining dataset unique in both scope and breadth of analysis. We further performed a comprehensive analysis of the results of this automatic extraction process against major biological database resources covering various aspects of the extracted information. This analysis indicated that text mining results for protein complexes, substructures and epigenetic DNA methylation provides already quite extensive coverage of relevant proteins. For post-translational modifications, we note that coverage still needs to be improved, but conclude that the extracted events already provide a valuable link to PTM related literature. In future work we hope to further extend the event types extracted by our PubMed-scale approach. The extraction methods as well as all data introduced in this study are freely available from `bionlp.utu.fi`.

Acknowledgments

We thank the Academy of Finland, the Research Foundation Flanders (FWO) and the UK BBSRC (reference number: BB/G013160/1) for funding, and CSC – IT Center for Science Ltd for computational resources.

References

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 183–191.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop*, pages 10–18.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the BioNLP 2010 Workshop*, pages 28–36.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of ACL*, pages 173–180.

Y.C. Fang, P.T. Lai, H.J. Dai, and W.L. Hsu. 2011. Meinfotext 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC bioinformatics*, 12(1):471.

Z. Z. Hu, M. Narayanaswamy, K. E. Ravikumar, K. Vijay-Shanker, and C. H. Wu. 2005. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, 21(11):2759–2765.

Sarah Hunter et al. 2012. Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, 40(D1):D306–D312.

Jun'ichi Kazama and Jun'ichi Tsujii. 2003. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of EMNLP 2003*, pages 137–144.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of BioNLP 2009*, pages 1–9.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011*, pages 1–6.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

Tzong-Yi Lee, Hsien-Da Huang, Jui-Hung Hung, Hsi-Yuan Huang, Yuh-Shyong Yang, and Tzu-Hao Wang. 2006. dbPTM: an information repository of protein post-translational modification. *Nucleic acids research*, 34(suppl 1):D622–D627.

Hodong Lee, Gwan-Su Yi, and Jong C. Park. 2008. E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucl. Acids Res.*, 36(suppl.2):W416–422.

Hong Li, Xiaobin Xing, Guohui Ding, Qingrun Li, Chuan Wang, Lu Xie, Rong Zeng, and Yixue Li. 2009. SysPTM: A Systematic Resource for Proteomic Research on Post-translational Modifications. *Molecular & Cellular Proteomics*, 8(8):1839–1849.

Zhiyong Lu et al. 2011. The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12(Suppl 8):S2+.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*, pages 449–454.

David McClosky. 2010. *Any domain parsing: automatic domain adaptation for natural language parsing*. Ph.D. thesis, Department of Computer Science, Brown University.

Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction for post-translational modifications. In *Proceedings of BioNLP'10*, pages 19–27.

Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, and Jun'ichi Tsujii. 2011a. Event extraction for DNA methylation. *Journal of Biomedical Semantics*, 2(Suppl 5):S2.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011b. Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 16–25.

Maté Ongenaert, Leander Van Neste, Tim De Meyer, Gerben Menschaert, Sofie Bekaert, and Wim Van Criekinge. 2008. PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucl. Acids Res.*, 36(suppl_1):D842–846.

Sampo Pyysalo, Antti Airola, Juho Heimonen, and Jari Björne. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl. 3):S6.

Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the BioNLP 2009 Workshop*, pages 1–9.

Sampo Pyysalo, Tomoko Ohta, and Jun'ichi Tsujii. 2011. Overview of the entity relations (REL) supporting task of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 83–88.

Eric W. Sayers et al. 2010. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 38(suppl 1):D5–D16.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp shared task 2011: Supporting resources. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 112–120.

The UniProt Consortium. 2011. Ongoing and future developments at the universal protein resource. *Nucleic Acids Research*, 39(suppl 1):D214–D219.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837, 07.

Sofie Van Landeghem, Sampo Pyysalo, Tomoko Ohta, and Yves Van de Peer. 2010. Integration of static relations to enhance event extraction from text. In *Proceedings of BioNLP'10*, pages 144–152.

Sofie Van Landeghem, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2012. Exploring biomolecular literature with EVEX: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics*.

Chih-Hsuan Wei and Hung-Yu Kao. 2011. Cross-species gene normalization by species inference. *BMC bioinformatics*, 12(Suppl 8):S5.

Cathy H. Wu et al. 2003. The Protein Information Resource. *Nucl. Acids Res.*, 31(1):345–347.

X. Yuan, ZZ Hu, HT Wu, M. Torii, M. Narayanaswamy, KE Ravikumar, K. Vijay-Shanker, and CH Wu. 2006. An online literature mining tool for protein phosphorylation. *Bioinformatics*, 22(13):1668.