Event Extraction for DNA Methylation

Tomoko Ohta[*] **Sampo Pyysalo**[*] **Makoto Miwa**[*] **Jun'ichi Tsujii**[*†‡]
[*]Department of Computer Science, University of Tokyo, Tokyo, Japan
[†]School of Computer Science, University of Manchester, Manchester, UK
[‡]National Centre for Text Mining, University of Manchester, Manchester, UK
{okap,smp,mmiwa,tsujii}@is.s.u-tokyo.ac.jp

Abstract

We consider the task of automatically extracting DNA methylation events from the biomedical domain literature. DNA methylation is a key mechanism of epigenetic control of gene expression and implicated in many cancers, but there has been little study of automatic information extraction for DNA methylation. We present an annotation scheme following the representation of the recent BioNLP'09 shared task on event extraction, select a set of 200 abstracts including a balanced sample of all PubMed citations relevant to DNA methylation, and introduce manual annotation for this corpus marking nearly 3000 gene/protein mentions and 1500 DNA methylation and demethylation events. We retrain a state-of-the-art event extraction system on the corpus and find that automatic extraction can be performed at 78% precision and 76% recall. The introduced resources are freely available for use in research from the GENIA project homepage.[1]

1 Introduction

During the previous decade of concentrated study of biomedical information extraction (IE), most efforts have focused on the foundational task of detecting mentions of entities of interest and the extraction of simple relations between these entities, typically represented as undifferentiated binary associations (Pyysalo et al., 2008). However, in recent years there has been increased interest in biomolecular event extraction using representations that capture typed, structured n-ary associations of entities in specific roles, such as *regulation* of the *phosphorylation* of a specific *domain* of a particular *protein* (Ananiadou et al., 2010). The state of the art in such extraction methods was evaluated in the BioNLP'09 Shared Task on Event Extraction (below, BioNLP ST) (Kim et al., 2009), and event extraction following the BioNLP ST model has continued to draw interest also after the task, with recent work including advances in extraction methods (Miwa et al., 2010a; Poon and Vanderwende, 2010), the release of extraction system software and large-scale automatically annotated data (Björne et al., 2010) and the development of additional annotated resources following the event representation (Ohta et al., 2010).

Of the findings of the BioNLP ST evaluation, it is of particular interest to us that the highest-performing methods include many that are purely machine-learning based (Kim et al., 2009), learning what to extract directly from a corpus annotated with examples of the events of interest. This implies that state-of-the-art extraction methods for new types of events can be created by providing annotated resources to an existing system, without the need for direct development of natural language processing or IE methods. Here, we apply this approach to DNA methylation, a specific and biologically highly relevant entity type not considered in previous event extraction studies.

In the following, we first outline the biological significance of DNA methylation and discuss existing resources. We then introduce the event extraction approach applied, present the new annotated corpus created in this study, and event extraction results using a method trained on the corpus.

2 DNA Methylation

The term *epigenetics* refers to a set of molecular mechanisms "beyond genetics" – i.e. without change in DNA sequence – that are today understood to play an important role in several biological processes, including genetic program for development, cell differentiation and tissue specific

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

gene expression. DNA methylation was first suggested as an epigenetic mechanism for the control of gene activity during development in 1975 (Riggs, 1975; Holliday and Pugh, 1975), and the role of DNA methylation in cancer was first reported in 1987 (Holliday, 1987). DNA methylation of CpG islands in promoter regions is now understood to be one of the most consistent genetic alterations in cancer, and DNA methylation is a prominent area of study.

Chemically, DNA methylation is a simple reaction adding a methyl group to a specific position of cytosine pyrimidine ring or adenine purine ring. While a single nucleotide can only be either methylated or unmethylated, in text the overall degree of promoter methylation is often reported as hypo- and hyper-methylation, with hyper-methylation implying that the expression of a gene is silenced. Because of the precise definition of the phenomenon and the relatively specific terms in which it is typically discussed in publications, we expected it to provide a well-defined target for annotation and automatic extraction.

2.1 DNA Methylation in PubMed

We follow common practice in biomedical IE in drawing texts for our corpus from PubMed abstracts. Currently containing more than 20 million citations for biomedical literature (over 11M with abstracts) and growing exponentially (Hunter and Cohen, 2006), the literature database provides a rich resource for IE and text mining.

To facilitate access to documents relevant to specific topics, each PubMed citation is manually assigned terms that identify its primary topics using MeSH, a controlled vocabulary of over 25,000 terms. MeSH contains also a *DNA Methylation* term, allowing specific searches for citations on the topic. Figure 1 shows the number of citations per year of publication matching this term contrasted with overall citations, illustrating explosive growth of interest in DNA methylation, outstripping the overall growth of the literature. Particular increases can be seen after the introduction of DNA microarrays for monitoring gene expression (Schena et al., 1995) and the introduction of high-throughput screening methods (Kononen et al., 1998; MacBeath and Schreiber, 2000). The total number of PubMed citations tagged with *DNA Methylation* at the time of this writing is 15456 (14350 of which have an abstract). The large number of documents tagged for the DNA methylation MeSH term and the human judgments assuring their relevance make querying for this term a natural choice for selecting text. However, direct PubMed query as the only selection strategy would ignore significant existing resources, discussed in the following.

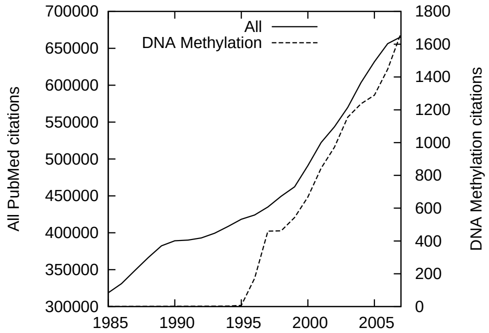

Figure 1: Citations tagged with the MeSH term *DNA Methylation* compared to all citations in PubMed by publication year. Note different scales.

2.2 DNA Methylation Databases

A growing number of databases collating information on DNA methylation are becoming available. The first such database, MethDB (Amoreira et al., 2003), was introduced in 2001 and remains actively developed. MethDB contains PubMed citation references as evidence for contained entries, but no more specific identification of the expressions stating DNA methylation events. The methPrimerDB (Pattyn et al., 2006) database provides additional information on PCR primers on top of MethDB, but does not add further specification of the methylated gene or text-bound annotation. PubMeth (Ongenaert et al., 2008) is a database of DNA methylation in cancer with evidence sentences from the literature. This database stores information on cancer types and subtypes, methylated genes and the experimental method used to identify methylation, as well as evidence sentences. MeInfoText, (Fang et al., 2008) is a database of DNA methylation and cancer information automatically extracted from PubMed documents matching the query terms *human, methylation and cancer* using term co-occurrence statistics. Of the DNA Methylation resources, only PubMeth and MeInfoText contain text-bound annotation identifying specific spans of characters containing the gene mention and ex-

a) MS-PCR revealed the [methylation] of the [*p16*] gene in 10(34%)of 29 [**NSCLCs**]

b) 30% (27 of 91) of [**lung tumors**] showed [hypermethylation] of the 5'CpG region of the [*p14ARF gene*]

c) [Promotor hypermethylations] were detected in [*O6-methylguanine-DNA methyltransferase (MGMT), RB1, estrogen receptor, p73, p16INK4a, death-associated protein kinase, p15INK4b, and p14ARF*]

d) The promoter region of the [*p16INK4*] gene was [hypermethylated] in the tumor samples of the primary or metastatic site

Table 1: Examples of PubMeth evidence sentence annotation. Annotated spans delimited by brackets and statements expressing methylation underlined, gene mentions shown in italics, and cancer mentions in bold.

pressing DNA methylation in evidence sentences supporting database entries. In this study, we consider specifically PubMeth as a source of reference text-bound annotations due to availability and the ability to redistribute derived data.

Initial text-bound annotations in PubMeth were generated using keyword lookup, but the database annotations are manually reviewed. Table 1 shows example evidence sentences from PubMeth and their annotated spans. While the PubMeth annotation differs from the BioNLP ST representation in a number of ways, such as not separating coordinated entities (Table 1c) and not annotating methylation sites (Table 1d), it provides both a reference identifying annotation targets from a biologically motivated perspective and a potential starting point for full event annotation.

3 Annotation

For annotation, we adapted the representation applied in the BioNLP ST on event extraction with minimal changes in order to allow systems developed for the task to be applied also for the newly annotated corpus. Documents were selected following the basic motivation presented above, with reference to the requirements specified by the annotation scheme, and some automatic preprocessing was applied as annotator support. This section details the annotation approach.

3.1 Entity and Event Representation

For the core named entity annotation, we thus primarily follow the gene/gene product (GGP) annotation criteria applied for the shared task data (Ohta et al., 2009). In brief, the guidelines specify annotation of minimal contiguous spans containing mentions of specific gene or gene product (RNA/protein) names, where *specific name* is understood to be one allowing a biologist to identify the corresponding entry in a gene/protein database such as Uniprot or Entrez Gene. The annotation thus excludes e.g. names of families and complexes. A single annotation type, *Gene or gene product*, is applied without distinction between genes and their products. In addition to the identification of the modified gene, it is important to identify the site of the modification. We marked mentions of sites relevant to the events as *DNA domain or region* terms following the original GENIA term corpus annotation guidelines (Ohta et al., 2002).

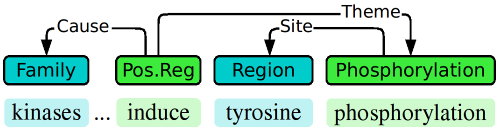

Figure 2: Event annotation for phosphorylation.

For representing DNA methylation events, the annotation applied to capture protein phosphorylation events in the BioNLP ST task 2 closely matched the needs for DNA methylation (Figure 2). While the Site arguments of the ST Phosphorylation events are protein domains, machine-learning based extraction methods should be able to associate this role with DNA domains given training data. We thus adopted a representation where DNA methylation events are associated with a gene/gene product as their Theme and a DNA domain or region as Site. Each event is also associated with a particular span of text expressing it, termed the *event trigger*.[2] We further initially marked catalysts using Positive regulation events following the BioNLP ST model, but dropped this class of annotation as a sufficient number of examples was not found in the corpus.

The event types of the BioNLP ST are drawn from the GENIA Event ontology (Kim et al., 2008), which in turn draws its type definitions from the community-standard Gene Ontology (GO) (The Gene Ontology Consortium, 2000). To maintain compatibility with these resources, we opted to follow the GO also for the definition of

[2] Annotators were instructed to always mark some trigger expression. We note that while we do not here specifically distinguish hypo- and hyper-methylation, the trigger annotations are expected to facilitate adding these distinctions if necessary.

the new event type considered here. GO defines DNA methylation as

> The covalent transfer of a methyl group to either N-6 of adenine or C-5 or N-4 of cytosine.

We note that while the definition may appear restrictive, methylation of adenine N-6 or cytosine C-5/N-4 encompasses the entire set of ways in which DNA can be methylated. This definition could thus be adopted without limitation to the scope of the annotation.

3.2 Document Selection

The selection of source documents for an annotated corpus is critical for assuring that the corpus provides relevant and representative material for studying the phenomena of interest. Domain corpora frequently consist of documents from a particular subdomain of interest: for example, the GENIA corpus focuses on documents concerning transcription factors in human blood cells (Ohta et al., 2002). Methods trained and evaluated on such focused resources will not necessarily generalize well to broader domains. However, there has been little study of the effect of document selection on event extraction performance. Here, we applied two distinct strategies to get a representative sample of the full scope of DNA methylation events in the literature and to assure that our annotations are relevant to the interests of biologists.

In the first strategy, we aimed in particular to select a representative sample of documents relevant to the targeted event types. For this purpose, we directly searched the PubMed literature database. We further decided not to include any text-based query in the search to avoid biasing the selection toward particular entities or forms of event expression. Instead, we only queried for the single MeSH term *DNA Methylation*. While this search is expected to provide high-prevision results for the full topic, not all such documents necessarily discuss events where specific genes are methylated. In initial efforts to annotate a random sample of these documents, we found that many did not mention specific gene names. To reduce wasted effort in examining documents that contain no markable events, we added a filter requiring a minimum number of (likely) gene mentions. We first tagged all 14350 citations tagged with *DNA Methylation* that have an abstract in PubMed using the BANNER tagger (Leaman and Gonzalez, 2008). We found that while the overwhelmingly most frequent number of tagged mentions per document is zero, a substantial mass of abstracts have large mention counts (Figure 3).[3] We decided after brief preliminary experiments to filter the initial selection of documents to include only those in which at least 5 gene/protein mentions were marked by an automatic tagger. This excludes most documents without markable events without introducing obvious other biases.

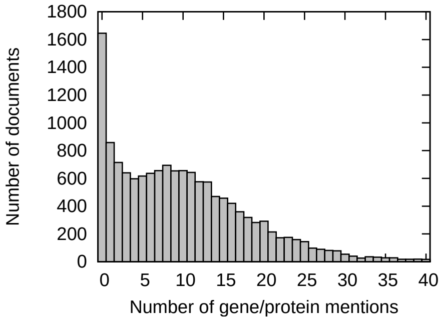

Figure 3: Number of citations with given number of automatically tagged gene/protein mentions.

In the second strategy, we extended and completed the annotation of a random selection of PubMeth evidence sentences, aiming to leverage existing resources and to select documents that had been previously judged relevant to the interests of biologists studying the topic. This provides an external definition of document relevance and allows us to estimate to what extent the applied annotation strategy can capture biologically relevant statements. This strategy is also expected to select a concentrated, event-rich set of documents. However, the selection may also necessarily carry over biases toward particular subsets of relevant documents from the original selection and will not be a representative sample of the overall distribution of such documents in the literature.

For producing the largest number of event annotations with the least effort, the most efficient way to use the PubMeth data would have been to simply extract the evidence sentences and complete the annotation for these. However, viewing the context in which event statements occur as centrally important, we opted to annotate complete abstracts, with initial annotations from PubMeth evidence sentences automatically transferred into the abstracts. We note that not all PubMeth

[3] The tagger has been evaluated at 86% F-score on a broad-coverage corpus, suggesting this is unlikely to severely misestimate the true distribution.

evidence spans were drawn from abstracts, and not all that were matched a contiguous span of text. We could align PubMeth evidence annotations into 667 PubMed abstracts (approximately 57% of the referenced PMID number in PubMeth) and completed event annotation for a random sample of these.

3.3 Document Preprocessing

To reduce annotation effort, we applied automatic systems to produce initial candidate sentence boundaries and GGP annotations for the corpus. For sentence splitting, we applied the GENIA sentence splitter[4], and for gene/protein tagging, we applied the BANNER NER system (Leaman and Gonzalez, 2008) trained on the GENETAG corpus (Tanabe et al., 2005). The GENETAG guidelines and gene/protein entity annotation coverage are known to differ from those applied for GGP annotation here (Wang et al., 2009). However, the broad coverage of PubMed provided by the GENETAG suggests taggers trained on the corpus are likely to generalize to new subdomains such as that considered here. By contrast, all annotations following GGP guidelines that we are aware of are subdomain-specific.

We note that all annotations in the produced corpus are at a minimum confirmed by a human annotator and that events are annotated without performing initial automatic tagging to assure that no bias toward particular extraction methods or approaches is introduced.

4 Results

4.1 Corpus Statistics

Corpus statistics are given in Table 2. There are some notable differences between the subcorpora created using the different selection strategies. While the subcorpora are similar in size, the PubMeth GGP count is 1.4 times that of the PubMed subcorpus[5], yet roughly equal numbers of methylation sites are annotated in the two. This difference is even more pronounced in the statistics for event arguments, where two thirds of PubMeth subcorpus events contain only a Theme argument identifying the GGP, while events where both Theme and Site are identified are more frequent in the other subcorpus.[6] As the extraction of events specifying also sites is known to be particularly challenging (Kim et al., 2009), these statistics suggest the PubMed subcorpus may represent a more difficult extraction task. Only very few DNA demethylation events are found in either subcorpus. Overall, the PubMeth subcorpus contains nearly twice as many event annotations as the PubMed one, indicating that the focused document selection strategy was successful in identifying particularly event-rich abstracts.

	PubMeth	PubMed	Total
Abstracts	100	100	200
Sentences	1118	1009	2127
Entities			
GGP	1695	1195	2890
Site	240	234	474
Total	1935	1429	3364
Events			
Theme only	660	214	874
Theme and Site	323	297	620
DNA methylation	977	485	1462
DNA demethyl.	6	26	38
Total	983	511	1494

Table 2: Corpus statistics.

4.2 Annotation Quality

To measure the consistency of the produced annotation, we performed independent double annotation for a sample of 40% of the abstracts selected from the PubMed subcorpus; 20% of all abstracts. As the PubMed subcorpus event annotation is created without initial human annotation as reference (unlike the PubMeth subcorpus), agreement is expected to be lower on this subcorpus. This experiment should thus provide a lower bound on the overall consistency of the corpus.

We first measured agreement on the gene/gene product (GGP) entity annotation, and found very high agreement among 935 entities marked in total by the two annotators: 91% F-score using exact match criteria and 97% F-score using the relaxed "overlap" criterion where any two overlapping annotations are considered to match.[7] We then separately measured agreement on event annotations

[4] http://www-tsujii.is.s.u-tokyo.ac.jp/∼y-matsu/geniass/

[5] The differences in the number of GGP annotations may be affected by the PubMeth entity annotation criteria.

[6] The number of annotated sites is less than the number of events with a Site argument as the annotation criteria only call for annotating a site entity when it is referred to from an event, and multiple events can refer to the same site entity.

[7] The high agreement is not due to annotators simply agreeing with the automatic initial annotation: the F-score of the automatic tagger against the two sets of human annotations was 65%/66% for exact and 85%/86% for overlap match.

for those events that involved GGPs on which the annotators agreed, using the standard evaluation criteria described in Section 4.4. Agreement on event annotations was also high: 84% F-score overall (85% for DNA methylation and 75% for DNA demethylation) over a total of 442 annotated events.

The overall consistency of the annotation depends on joint annotator agreement on the GGP and event annotations. However, in experimental settings such as that of the BioNLP ST where gold GGP annotation is assumed as the starting point for event extraction, measured performance is not affected by agreement on GGPs and thus arguably only the latter factor applies. As this setting is adopted also in the present study, annotation consistency suggests a human upper bound no lower than 84% F-score on extraction performance.

Estimates of the annotation consistency of biomedical domain corpora are regrettably seldom provided, and to the best of our knowledge ours is the first estimate of inter-annotator agreement for a corpus following the event representation of the BioNLP ST. Given the complexity of the annotation – typed associations of event trigger, theme and site – the agreement compares favorably to e.g. the reported 67% inter-annotator F-score reported for protein-protein interactions on the ITI TXM corpora (Alex et al., 2008) and the full event agreement on the GREC corpus (Thompson et al., 2009).

4.3 Event Extraction Method

To estimate the feasibility of automatic extraction of DNA methylation events and the suitability of presently available event extraction methods to this task, we performed experiments using the EventMine event extraction system of (Miwa et al., 2010b). On the task 2 of the BioNLP ST dataset, the benchmark most relevant to our task setting, the applied version of EventMine was recently evaluated at 55% F-score (Miwa et al., 2010a), outperforming the best task 2 system in the original shared task (Riedel et al., 2009) by more than 10% points. To the best of our knowledge, this system represents the state of the art for this event extraction task.

EventMine is an SVM-based machine learning system following the pipeline design of the best system in the BioNLP ST (Björne et al., 2009), extending it with refinements to the feature set, the use of a machine learning module for complex event construction, and the use of two parsers for syntactic analysis (Miwa et al., 2010b). We follow Miwa et al. in applying the HPSG-based deep parser Enju (Miyao and Tsujii, 2008) using the high-speed parsing setting ("mogura") and the GDep (Sagae and Tsujii, 2007) native dependency parser, both with biomedical domain models based on the GENIA treebank data (Tateisi et al., 2006).

For evaluation, we applied a version of the BioNLP'09 ST evaluation tools[8] modified to recognize the novel DNA_methylation event type.

4.4 Evaluation Criteria

We followed the basic task setup and primary evaluation criteria of the BioNLP'09 ST. Specifically, we followed task 2 ("event enrichment") criteria, requiring for correct extraction of a DNA methylation event both the identification of the modified gene (GGP entity) and the identification of the modification site (*DNA domain or region* entity) when stated. As in the shared task, human annotation for GGP entities was provided as part of the system input but other entities were not, so that the system was required to identify the spans of the mentioned modification sites.

The performance of the system was evaluated using the standard precision, recall and F-score metrics for the recovery of events, with event equality defined following the "Approximate span" matching criterion applied in the primary evaluation for the BioNLP'09 ST. This criterion relaxes strict matching requirements so that a detected event trigger or entity is considered to match a gold trigger/entity if its span is entirely contained within the span of the gold trigger, extended by one word both to the left and to the right.

4.5 Experimental Setup

We divided the corpus into three parts, first setting one third of the abstracts aside as a held-out test set and then splitting the remaining two thirds in a roughly 1:3 ratio into a training set and a development test set, giving 100 abstracts for training, 34 for development, and 66 for final test. The splits were performed randomly, but sampling so that each set has an equal number of abstracts drawn from the PubMeth and PubMed subcorpora.

The EventMine system has a single tunable threshold parameter that controls the tradeoff be-

[8] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/downloads.shtml

Event type	prec.	recall	F-score
DNA methylation	77.6%	77.2%	77.4%
DNA demethylation	100.0%	11.1%	20.0%
Total	77.7%	76.0%	76.8%

Table 3: Overall extraction performance.

	Test set		
Training set	PubMed	PubMeth	Both
PubMed	64.9%	71.2%	71.6%
PubMeth	62.9%	80.0%	74.0%
Both	**66.2%**	**82.5%**	**76.8%**

Table 4: F-score by subcorpus.

tween system precision and recall. We first set the tradeoff using a sparse search of the parameter space [0:1], evaluating the performance of the system by training on the training set and evaluating on the development set. As these experiments did not indicate any other parameter setting could provide significantly better performance, we chose the default threshold setting of 0.5. To study the effect of training data size on performance, we performed extraction experiments randomly downsampling the training data on the document level with testing on the development set. In final experiments EventMine was trained on the combined training and development data and performance evaluated on the held-out test data.

4.6 Extraction performance

Table 3 shows extraction results on the held-out test data. While DNA methylation events could be extracted quite reliably, the system performed poorly for DNA demethylation events. The latter result is perhaps not surprising given their small number – only 38 in total in the corpus – and indicates that a separate selection strategy is necessary to provide resources for learning the reverse reaction. Overall performance shows a small preference for precision over recall at 77% F-score. We view this level of performance very good as a first result.

To evaluate the relative difficulty of the extraction tasks that the two subcorpora represent and their merits as training material, we performed tests separating the two (Table 4). As predicted from corpus statistics (Section 4.1), the PubMed subcorpus represents the more challenging extraction task. When testing on a single subcorpus, results are, unsurprisingly, better when training data is drawn from the same subcorpus; however, training on the combined data gives the best performance for all three test sets, indicating that the subcorpora are compatible.

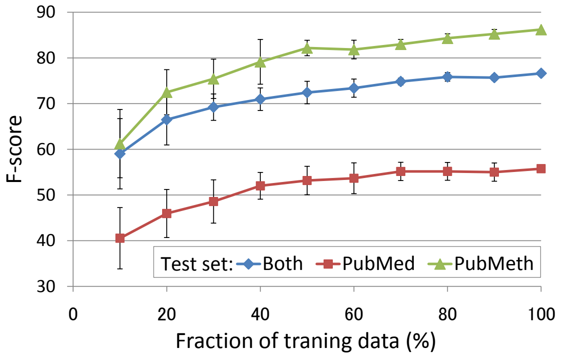

Figure 4: Learning curve for the two subcorpora and their combination. Both subcorpora used for training. Average and error bars calculated by 10 repetitions of random subsampling of training data, testing on the development set.

The learning curve (Figure 4) shows relatively high performance and rapid improvement for modest amounts of data, but performance improvement with additional data levels out relatively fast, nearly flattening as use of the training data approaches 100%. This suggests that extraction performance for this task is not primarily limited by training data size and that additional annotation following the same protocol is unlikely to yield notable improvement in F-score without a substantial investment of resources. As performance for the PubMed subcorpus (for which inter-annotator agreement was measured) is not yet approaching the limit implied by the corpus annotation consistency (Section 4.2), the results suggest further need for the development of event extraction methods to improve DNA methylation event extraction.

5 Related Work

DNA methylation and related epigenetic mechanisms of gene expression control have been a focus of considerable recent research in biomedicine. There are many excellent reviews of this broad field; we refer the interested reader to (Jaenisch and Bird, 2003; Suzuki and Bird, 2008).

There is a wealth of recent related work also on event extraction. In the BioNLP’09 shared task, 24 teams participated in the primary task and six teams in Task 2 which mostly resembles our setup in that it also required the detection of modified gene/protein and modification site. The top-performing system in Task 2 (Riedel et al., 2009) achieved 44% F-score, and the highest performance reported since that we are aware of is 55% F-score for EventMine (Miwa et al., 2010b). The performance we achieved for DNA methylation is considerably better than this overall result, essentially matching the best reported performance for Phosphorylation events, which we previously argued to be the closest shared task analogue to the new event category studied here. Nevertheless, direct comparison of these results may not be meaningful due to confounding factors. The only text mining effort specifically targeting DNA methylation that we are aware of is that performed for the initial annotation of the PubMeth and MeInfoText databases (Ongenaert et al., 2008; Fang et al., 2008), both applying approaches based on keyword matching. However, neither of these studies report results for instance-level extraction of methylation statements.

The present study is in many aspects similar to our previous work targeting protein post-translational modification events (Ohta et al., 2010). In this work, we annotated 422 events of 7 different types and showed that retraining an existing event extraction system allowed these to be extracted at 42% F-score. Our approach here clearly differs from this previous work in its larger scale and concentrated focus on a particular event type of high interest, reflected also in results: while extraction performance in our previous work was limited by training data size, in the present study notably higher extraction performance was achieved and a plateau in performance with increasing data reached.

6 Discussion and Future Work

We have presented a study of the automatic extraction of DNA methylation events from literature following the BioNLP'09 shared task event representation and a state-of-the-art event extraction system. We created an corpus of 200 publication abstracts selected to include a representative sample of DNA methylation statements from all of PubMed and manually annotated for nearly 3000 mentions of genes and gene products, 500 DNA domain or region mentions and 1500 DNA methylation and demethylation events. Evaluation using the EventMine system showed that DNA methylation events can be extracted simply by retraining an off-the-shelf event extraction system at 78% precision and 76% recall. The learning curve suggested that the corpus size is sufficient and that in future efforts in DNA methylation event extraction should focus on extraction method development.

One natural direction for future work is to apply event extraction systems trained on the newly introduced data to abstracts available in PubMed and full texts available at PMC to create a detailed, up-to-date repository of DNA methylation events at full literature scale. Such an effort would require gene name normalization and event extraction at PubMed scale, both of which have recently been shown to be technically feasible (Gerner et al., 2010; Björne et al., 2010). Further combining the extracted events with cancer mention detection could provide a valuable resource for epigenetics research.

The newly annotated corpus, the first resource annotated for DNA methylation using the event representation, is freely available for use in research from from the GENIA project homepage `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA`.

Acknowledgments

We would like to thank Maté Ongenaert and other creators of PubMeth for their generosity in allowing the release of resources building on their work and the anonymous reviewers for their many insightful comments. This work was supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

Bea Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. 2008. The ITI TXM corpora: Tissue expressions and protein-protein interactions. In *Proceedings of LREC'08*.

Celine Amoreira, Winfried Hindermann, and Christoph Grunau. 2003. An improved version of the DNA methylation database (MethDB). *Nucl. Acids Res.*, 31(1):75–77.

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.

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 BioNLP'09 Shared Task*, 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 BioNLP'10*, pages 28–36.

Yu-Ching Fang, Hsuan-Cheng Huang, and Hsueh-Fen Juan. 2008. Meinfotext: associated gene methylation and cancer information from text mining. *BMC Bioinformatics*, 9(1):22.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In *Proceedings of BioNLP 2010*, pages 72–80.

Robin Holliday and JE Pugh. 1975. Dna modification mechanisms and gene activity during development. *Science*, 187:226–232.

Robin Holliday. 1987. The inheritance of epigenetic defects. *Science*, 238:163–170.

Lawrenece Hunter and K. Bretonnel Cohen. 2006. Biomedical language processing: What's beyond PubMed? *Molecular Cell*, 21(5):589–594.

Rudolf Jaenisch and Adrian Bird. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33:245–254.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

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'09*.

Juha Kononen, Lukas Bubendorf, Anne Kallionimeni, Maarit Barlund, Peter Schraml, Stephen Leighton, Joachim Torhorst, Michael J Mihatsch, Guido Sauter, and Olli-P. Kallionimeni. 1998. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med*, 4(7):844–847.

R. Leaman and G. Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Proceedings of PSB'08*, pages 652–663.

Gavin MacBeath and Stuart L. Schreiber. 2000. Printing Proteins as Microarrays for High-Throughput Function Determination. *Science*, 289(5485):1760–1763.

Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara, and Jun'ichi Tsujii. 2010a. A comparative study of syntactic parsers for event extraction. In *Proceedings of BioNLP'10*, pages 37–45.

Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010b. Event extraction with complex event classification using rich features. *Journal of Bioinformatics and Computational Biology (JBCB)*, 8(1):131–146.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):35–80.

Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2002. GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT'02*, pages 73–77.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of BioNLP'09*, pages 106–107.

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.

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.

Filip Pattyn, Jasmien Hoebeeck, Piet Robbrecht, Evi Michels, Anne De Paepe, Guy Bottu, David Coornaert, Robert Herzog, Frank Speleman, and Jo Vandesompele. 2006. methblast and methprimerdb: web-tools for pcr based methylation analysis. *BMC Bioinformatics*, 7(1):496.

Hoifung Poon and Lucy Vanderwende. 2010. Joint inference for knowledge extraction from biomedical literature. In *Proceedings of NAACL/HLT'10*, pages 813–821.

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

Sebastian Riedel, Hong-Woo Chun, Toshihisa Takagi, and Jun'ichi Tsujii. 2009. A markov logic approach to biomolecular event extraction. In *Proceedings of BioNLP'09 Shared Task*, pages 41–49.

A.D. Riggs. 1975. X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14:9–25.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of EMNLP-CoNLL 2007*, pages 1044–1050.

Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470.

Miho M. Suzuki and Adrian Bird. 2008. Dna methylation landscapes: provocative insights from epigenomics. *Nature Review Genetics*, 9:465–476.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl. 1):S3.

Yuka Tateisi, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. 2006. Subdomain adaptation of a pos tagger with a small corpus. In *Proceedings of BioNLP'06*, page 136137, New York, USA, June.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Paul Thompson, Syed Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10(1):349.

Yue Wang, Jin-Dong Kim, Rune Sætre, Sampo Pyysalo, and Jun'ichi Tsujii. 2009. Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC Bioinformatics*, 10(403).