HistoNer: Histone modification extraction from text

Philippe Thomas **Ulf Leser**
Humboldt-Universität zu Berlin
Knowledge Management in Bioinformatics
Unter den Linden 6
Berlin, 10099, Germany
{thomas,leser}@informatik.hu-berlin.de

Abstract

Systematic recognition of histone modifications in text is an important task to cope with the fast increase of biomedical literature. The high variability of phrases to express histone modifications renders keyword based search as insufficient for information retrieval. We present HistoNer, a rule based system for the recognition of histone modifications from text. Patterns are collected semi-automatically and manually corrected. With 305 distinct patterns the system achieves an F_1 measure of 93.6 % on an unseen test set of 1,000 annotated documents.

HistoNer is licensed under GNU General Public License Version 3 and available at http://code.google.com/p/histoner/. The repository contains corpora, evaluation scripts, and intermediate files generated during pattern development.

1 Introduction

In eukariotic cells, DNA is densely packed around proteins. These proteins are referred to as histones. Winding DNA around histones greatly reduces the amount of space required for the DNA. The status of histones primarily determines the availability of DNA to binding proteins like transcription factors. Thus, histone modifications are known to directly affect transcription and regulation of other genes. Some of these modifications are involved in disease progression and are an a promising target for novel drug therapies (Kelly et al., 2010).

The vast majority of novel research findings is initially presented in scientific literature. Over the years, the amount of accumulated text has grown enormously and has reached a point where finding specific information becomes troublesome. Altough the Brno nomenclature for the description of histone modification has been developed (Turner, 2005), a high variety of natural language expressions describing histone modifications has been observed (Kolářik et al., 2009). The high number of possible phrases describing histone modifications hinders systematic retrieval of relevant articles describing such modifications.

1.1 Related work

Extraction of histone modification terms has been previously tackled by Kolářik et al. (2009). They use conditional random fields to detect histone modification mentions. Recognized modification mentions are subsequently term normalized to a self-developed ontology, which is inspired by the Brno nomenclature. Their system achieves an F_1 measure of 81 % on a test corpus of 1,000 documents.

Rule based systems often achieve good precision but lack a high recall due to the high variability of free text. More importantly, the development of hand crafted patterns is a time intensive and laborious task. Several approaches have been proposed to rapidly engineer such patterns with little or without any human intervention. For instance, Hakenberg et al. (2008) automatically derive phrase–motifs describing protein–protein interactions from scientific abstracts using a knowledge base with known protein–protein interactions. This workflow is usually referred to as distant supervision (Mintz et al., 2009). Rinaldi et al. (2010) also

follow the same rationale to generate potential patterns describing protein–protein interactions. But in difference to Hakenberg et al. (2008) the automatically generated patterns are manually evaluated and removed if too unspecific.

In this publication we generate patterns in a semi-automatic fashion by following the approach from Caporaso et al. (2007) which was proposed for the generation of mutation patterns. The approach uses background knowledge about mutations to generate potential patterns. Potential patterns are subsequently refined by the authors.

2 Methods

In this section we describe the corpus generation strategy, followed by a brief discussion of the pattern generation process and the evaluation strategy.

2.1 Histone modification definition

We define a histone modification following the specifications from the Brno nomenclature, which characterizes a histone modification as having the following four arguments:

1. Histone name (H1, H2a, H2b, H3, or H4)
2. Modification type (*e.g.* phosphorylation, dimethylation, acetylation, ...)
3. Modified amino acid (*e.g.* lysine, arginine, ...)
4. The amino acid position where the modifications occurs on the histone polypeptide (*e.g.* 7, 23, ...)

Hence, histone modification recognition can be regarded as quaternary relationship extraction. In this work we regard all four arguments as mandatory.

2.2 Corpus

For development and evaluation of our tool we use the 1,187 abstracts originally annotated by Kolářik et al. (2009). The corpus originally contained underspecified annotations, like "acetylated histones". We re-annotated the corpora and retained only histone modification terms mentioning all four modification arguments. The corpus has been split into training and test corpus consisting of 187 and 1,000 documents respectively by Kolářik et al. (2009). In this work we used the same corpus splits. The smaller training corpus contains 603 and the testing corpus 224 histone modification mentions. Differences in ratios between mentions of histone modifications to number of articles are due to the corpus selection strategy. The 187 training documents are manually selected by Kolářik et al. (2009) to cover a large variety of different histone modifications. The evaluation corpus has been randomly sampled from 24,635 articles annotated with the MeSH term "epigenetics".

2.3 Generation of patterns

Patterns are generated in a semi-automatic fashion by the following strategy originally proposed by Caporaso et al. (2007). Citations from MEDLINE and fulltext articles from PMC open access have been separated into sentences by using a segmentation model trained for biomedical publications (Buyko et al., 2006). These sentences have been searched for mentions of amino acids, modification terms, and numbers. For recognizing amino-acids we generated a list of different amino-acid terms, where we used long-forms, three letter abbreviations, and one letter abbreviations (*e.g.* Lysine, Lys, K, ...). For numbers we used a regular expression matching all number mentions between 1 to 999. Longer numbers are ignored as all histone proteins are shorter than 999 amino-acids. The Brno nomenclature currently lists seven different types of histone modifications (*e.g.* acetylation, methylation, ribosylation, ...). For all these modification terms we generated regular expressions matching verb, noun and adjective forms, like acetylation, acetylates, acetylating. We further build possible word inflections and active/passive word forms.

Mentions of amino-acids, numbers, and modifications are searched in the unannotated sentences. Detected mentions are replaced by a generic symbol. For instance, amino-acids are replaced by *<aa>*, modification terms by *<mod>*, and numbers by *<number>*. Patterns are derived from sentences containing all four required arguments. In other words, from sentences which contain at least two numbers, one amino-acid, one modification mention. To build these surface patterns we selected the shortest span between all relevant mentions and the words between them. Potential patterns are sorted

by their occurrence in MEDLINE and PMC. Patterns occurring at least twice are manually evaluated. This pattern generation strategy is also exemplified in Figure 1.

Input Sentence	"The major function of MYST is acetylation of H4 at the K16 residue."
Term recognition	"The major function of protein MYST is <mod> of H<number> at the <aa><number> residue."
Potential Pattern	"<mod> of H<number> at the <aa><number>"
Annotation	"<mod> of H<hnumber> at the <aa><aanumber>"

Figure 1: Example of the different steps for pattern generation. First, relevant terms are replaced by the respective class (*i.e.* <aa>, <number>, <mod>). Second, surface patterns are generated. Third, patterns are evaluated and manually refined.

This procedure results in a set of patterns which can later be used to find histone modifications. During the search phase generic symbols are replaced by regular expressions. For instance the symbol *aa* is replaced by all possible amino acids "(lysine|lys|K|, ...)". Recall, that for pattern generation the system did not use any information contained in the training corpus. All patterns are learned from the sentences provided in MEDLINE and PMC.

2.3.1 Pattern refinement

Manually annotated patterns are automatically refined by the following steps:

1. Mentions of conjunctions (and/or) are replaced by the regular expression "and|or"
2. Prepositions are replaced by the regular expression "to|of|at|on|in"
3. For the collocation "histone H',' two additional patterns are induced, containing only "histone" and "H"
4. For patterns containing only "histone" or "H" an additional pattern with "histone H" is produced

3 Results

3.1 Pattern generation

We retrieve potentially important articles by using the prefix query "histon*" on a set of ∼20 million articles. This query leads to a set of 52,113 documents with 3,656,587 sentences. 81,526 sentences contained all four required elements and are transformed into potential patterns. Patterns occurring at least twice over all MEDLINE are subsequently annotated by the authors. This leads to 268 manually annotated pattern, which are refined into 305 different patterns by the steps described in Section 2.3.1.

3.2 Evaluation

Both sets of patterns (original and refined) are used to find histone modification mentions on the two corpora. Results are shown in Table 1. Using the linguistically refined patterns improves recall by approximately 5 percentage points on both corpora. On training and test corpus we observe similar results in terms of precision, recall and F_1 measure.

Pattern	Training			Evaluation		
	P	R	F_1	P	R	F_1
Original	98.8	81.5	89.3	**99.0**	84.3	91.1
Refined	**98.9**	**87.2**	**92.7**	98.1	**89.6**	**93.6**

Table 1: Performance of HistoNer on the training and testing corpus. Original refers to the unrefined pattern, whereas refined refers to the modified pattern.

4 Discussion

For named entity recognition, the approach of Kolářik et al. (2009) achieves an F_1 of 81 % on the test corpus. Due to the slightly different scope of the two tools (HistoNer extracts only histone modifications normalizable to Brno), the results serve only as indicator for the high quality of HistoNer and can not be directly compared. An advantage of our pattern based strategy is that term normalization to the Brno nomenclature is implicitly performed by the usage of regular expressions.

Finally, we applied HistoNer on a local repository of more than 21 million PubMed citations, where our system detects 97,563 histone modifications. An overview of the five most frequently used patterns is shown in Table 2.

Pattern	Occurrence
H|histone\p?<hnumber>\p?<aa>\p?<aanumber>\p?<mod>	62,818
<mod>\s(to|of|at|on|in)(H|histone)<hnumber>\p?<aa>\p?<aanumber>	2,831
H|histone\s?<hnumber>\p?<mod>\s(to|of|at|on|in)\s<aa>\p?<aanumber>	1,678
<mod> H|histone<hnumber>\p?<aa>\p?<aanumber>	1,343
H|histone\p?<hnumber>\p?<mod>\p?<aa>\p?<aanumber>	1,268

Table 2: Overview of the five most frequently matching patterns. Terms in brackets are replaced by the corresponding regular expression as described in Methods. To simplify the regular expression we introduced the symbol "\p" matching an arbitrary punctuation mark or whitespace.

5 Conclusion

HistoNer is a stand alone tool capable of recognizing histone modification mentions in text. Detected mentions are normalized to the Brno nomenclature. For recognition it uses a set of 305 patterns, which have been automatically generated and subsequently manually corrected. The automatic refinement strategy is capable of improving recall by about 5 percentage points with unchanged precision. HistoNer achieves a remarkable performance of roughly 93 % F_1 on two unseen data sets.

Recognized histone modifications are integrated in our web service GeneView[1] (Thomas et al., 2012). The tool, including the set of regular expressions, evaluation scripts, intermediate files generated during pattern engineering, and documentation are freely available at `http://code.google.com/p/histoner/`.

We have shown that the bootstrapping strategy introduced by Caporaso et al. (2007) can be extended to another NER task, namely histone modification recognition.

References

Ekaterina Buyko, Joachim Wermter, Michael Poprat, and Udo Hahn. 2006. Automatically adapting an nlp core engine to the biology domain. In *Proceedings of Joint Linking Literature, Information and Knowledge for Biology and the 9th Bio-Ontologies Meeting*.

J. Gregory Caporaso, William A Baumgartner, David A Randolph, K. Bretonnel Cohen, and Lawrence Hunter. 2007. Rapid pattern development for concept recognition systems: application to point mutations. *J Bioinform Comput Biol*, 5(6):1233–1259.

Jörg Hakenberg, Conrad Plake, Loic Royer, Hendrik Strobelt, Ulf Leser, and Michael Schroeder. 2008. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol*, 9 Suppl 2:S14.

Theresa K Kelly, Daniel D De Carvalho, and Peter A Jones. 2010. Epigenetic modifications as therapeutic targets. *Nat Biotechnol*, 28(10):1069–1078.

Corinna Kolářik, Roman Klinger, and Martin Hofmann-Apitius. 2009. Identification of histone modifications in biomedical text for supporting epigenomic research. *BMC Bioinformatics*, 10 Suppl 1:S28.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 1003–1011.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. 2010. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7:472–480.

Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. 2012. Geneview: a comprehensive semantic search engine for pubmed. *Nucleic Acids Res*, 40(Web Server issue):W585–W591.

Bryan M Turner. 2005. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*, 12(2):110–112.

[1] `http://bc3.informatik.hu-berlin.de/`