Text Mining for Epigenomic Research

Corinna Kolářik, Roman Klinger, and Martin Hofmann-Apitius
Department of Bioinformatics

Fraunhofer Institute Algorithms and Scientific Computing

Background

Epigenomic research focuses on DNA methylation, **histone modifications**, its regulations and functional effects

Functions of posttranslational histone modifications:

- Influence the structure of chromatine
- Take part in the regulation of gene expression
- Are connected to cell and tissue differentiation
- Support adaption of organisms to their environment

Abnormal changes influence the development of disease states, like:

- Cancer, Mental disorders, Late onset diseases, etc.

Motivation:

1. Regulation mechanisms for modifying histones and its functions are still not completely understood
2. Text provides a rich resource of knowledge on modifications of histones

Scientific Challenge

- The way how histone modifications are described in text is not uniform and consistent, typical spelling variants of a histone modification example:

  H3K9me3; (41)* – term corresponds to the official Brno nomenclature,
  Me3-K9 H3; (1)*,
  H3 Lys9 trimethylation; (11)*,
  H3 tri-methylated at lysine 9; (14)*,
  histone H3 trimethylated at lysine (K) 9; (3)*,
  K9 trimethylation at histone H3; (36)*,

* The numbers in brackets provide the quantity of abstracts obtained with a PubMed search in July 2008

Goal

- Identification of histone modifications in text
- Standardization of the identified entities and embedding them into a hierarchy
- Usage for semantic text search
- Support of epigenomic research

Recognition of Histone Modifications

Corpus Annotation

Text example: `In contrast, K36- and K79-methylated H3 tails,... In unstimulated cells, continuous turnover of H3K9 acetylation occurs on all K4-trimethylated histone H3 tails...`

Training corpus*: 187 Articles from Medline (414 entities)

Test corpus*: 1000 Articles sampled from Medline (123 entities)

NER Approach: Conditional Random Fields

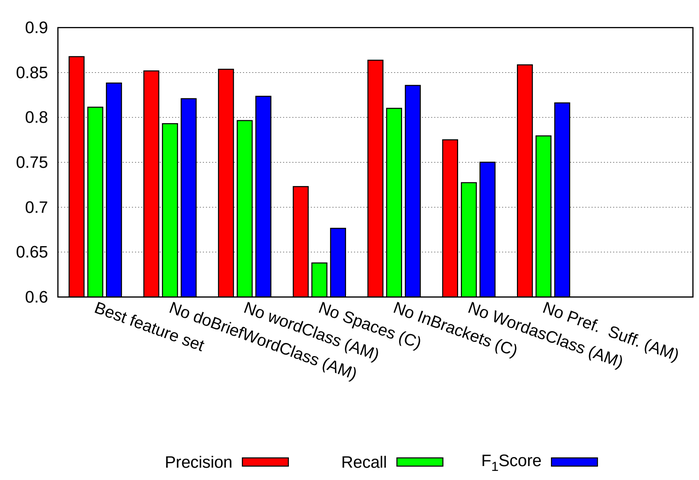
- Rich set of features
- Implementation based on Mallet
- 10 fold cross-validation

Identification Results:

	Training Corpus	Test Corpus
Recall	0.81 (± 0.05)	0.76
Precision	0.87 (± 0.05)	0.87
F1 measure	0.84 (± 0.05)	0.81

Usage of an optimized feature set

Feature Analysis



False positive terms – Error classes:

1. Modification descriptions without histone mentions: '`acetylation and methylation`'
2. Enzymes introducing or removing histone modifications: '`H3K9 methyltransferase`'
3. Boundary problems: '`H3 – K9 ) with no sign of histone H2AX phosphorylation`'
4. Terms with other meaning: '`phosphorylation of IRS`'

Term Standardization and Histone Modification Hierarchy

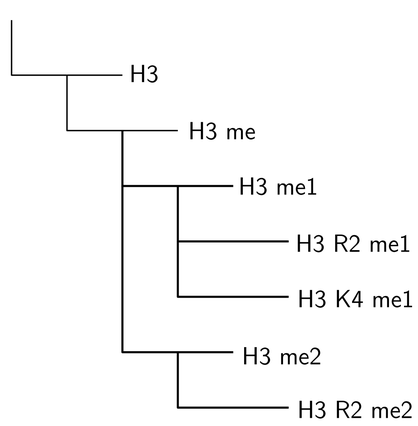
Postprocessing Procedure

- Filtering of frequent false positives
- Transformation to Brno nomenclature (Train c.: 95.89 %, Test c.: 98.37 %)
- Mapping to histone modification hierarchy*
  - Contains 462 concepts

Term Standardization Examples:

Identified Terms	Standardized Terms
dimethylated lysine 20 of histone H4	H4 K 20 me 2
di- and trimethylation of lysine 4 at histone 3	H3 K 4 me 2 H3 K 4 me 3

Histone Modification Hierarchy (Section)



Conclusion

- Approach supports an **easy retrieval** of all texts from Medline corresponding to a certain histone modification,
- Inclusion of the hierarchy in SCAIView allows for **semantic search**,
- Is a basis for supporting **epigenomic research** using text about histone modifications (e.g. further check for co-occurring diseases, genes, proteins, drugs)

* The corpora and the hierarchy are available from: `http://www.scai.fraunhofer.de/histone-corpora.html`

This work is funded by the Bonn-Aachen International Center for Information Technologies (`http://www.b-it-center.de`) and Fraunhofer-Max-Planck Machine Learning Cooperation (`http://lip.fml.tuebingen.mpg.de`)

Contact
Corinna Kolářik
Schloss Birlinghoven
53754 Sankt Augustin
corinna.kolarik@scai.fraunhofer.de
`http://www.scai.fraunhofer.de`