RESEARCH

An overview of the MiNCor: annotation guidelines and text corpora for MicroRNA mentions in the scientific literature

José C Sammartino[1]*, Martin Krallinger[2] and Alfonso Valencia[2]

Abstract

Micrornas are small endogenous molecules of non-coding RNA that act as post-transcriptional regulators of gene expression in a wide spectrum of physiological and pathological states. In recent years, these molecules have been the focus of the biomedical research, with a substantial increase of the number of publications. Therefore, without the proper specific text mining supervised tools, the retrieval of the information embedded in unstructured data can prove to be a labouring process. In this perspective, our aim is to provide a comprehensive granular annotation protocol for the annotation of non-coding RNA molecules, as well as two corpora and a large dictionary for the training and testing of text mining tools. The corpora and guidelines are freely downloadable at: http://zope.bioinfo.cnio.es/mincor/minacor.tar.gz.

Keywords: Text Mining; MicroRNA; Entities Annotation

MiNCor

The MiNCor takes place in a larger project which objective is the design of an automatic system for the extraction and analysis of microRNA mentions in relation with other entities (Diseases, Mutations, Chemicals ...) in the scientific literature. Here we briefly describe a new annotation protocol for labelling microRNA mentions that encompasses all microRNA mentions regardless of the species, the maturing step or the classification, including also the class of non-coding RNA and miRNA clusters. This annotation protocol has been iteratively refined and was then used for the annotation of the MiNCor corpora, which was used for the evaluation of several microRNA mentions recognition approaches.

Guidelines

The guidelines for the MiNCor annotation protocol integrates information from previous miRNA corpora, revision of multiple different resources and is based on the model of the Manual for annotation of chemical entities of the CHEMDNER corpus [1]. Is structured into rule types (General rules, Positive rules, Negative rules, Class rules) together with example and exception cases.

MiNCor Lexycon

Is a large dictionary of microRNA mentions derived from multiple microRNA databases as well as microRNA mentions detected by GNormplus [2]. A dictionary expansion step was carried out taking into account the nomenclature guidelines of microRNAs by considering additional core terms, prefixes and suffixes for a total of 788784 different entities (e.g.: 'mirna'; 'microrna'; 'non-coding RNA'; 'lin-4'; 'let-7'; 'antagomir'; 'oncomir'; 'mir-1'; 'has-miR-12').

Corpora

In this section are described the annotation protocols for the MiNCor corpora. The abstracts used for the construction of the corpora were retrieved from Pubmed using different MeSH queries for microRNAs and restricting the search to the abstracts published in 2016.

Annotation Protocol of the MiNCor Gold

102 abstracts were randomly selected and manually labelled using the customised AnnotateIt web-interface, similar to the one used for the annotation of the CHEMDNER-Patents Corpus [1]. The labelling was based on our MiNCor Guidelines for the annotation protocol and normalized with the pubmed ID. Table 1 provides the statistics of MiNCor Gold.

*Correspondence: j.sammartino.88@gmail.com
[1]Department of Molecular Medicine and Medical Biotechnology, University of Naples 'Federico II', Naples, Italy
Full list of author information is available at the end of the article

Annotation Protocol of the MiNCor Silver

All the retrieved files, excluding the ones used for the MiNCor Gold, were segmented into sentences, tokenized, lemmatized, part-of-speech and chunk tagged and then labelled using the MiNCor lexicon. A dictionary pruning step was carried out to remove highly ambiguous mentions. After applying the dictionary look-up a cascade of rules was used to adjust the mention boundaries.Table 2 provides the statistics of MiNCor Silver.

Conclusion and future work

Our aim is to expand the annotation Guidelines to other classes of non-coding RNAs to increase the accuracy of the annotation process. Furthermore, we are going to use such guidelines to enlarge our MiNCor Gold corpus so to have more variability in the manually labelled entities. We believe that the release of the MiNCor corpora and guidelines might be useful as an annotation template for the corpus construction of other biomedical entities as well as for the testing of Named Entity Recognition models.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Molecular Medicine and Medical Biotechnology, University of Naples 'Federico II', Naples, Italy. [2]Centro Nacional de Investigaciones Oncológicas, Madrid, Spain.

References

1. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., *et al.*: The chemdner corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics **7**(S1), 1–17 (2015)
2. Wei, C.-H., Kao, H.-Y., Lu, Z.: Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. BioMed research international **2015** (2015)

Tables

Table 1 statistics of MiNCor Gold

Abstracts	102
Sentences	1063
Total Microrna mentions	1154
Total unique mentions	232
General Microrna mentions	607
Specific Microrna mentions	501
Multiple Microrna mentions	14
Nested Microrna mentions	2
Cluster Microrna mentions	1
ncRNA Microrna mentions	29

Table 2 statistics of MiNCor Silver

Sentences	302560
Tokens	over 3000000
Total Microrna mentions	175367
Total unique mentions	8506