Annotation process, guidelines and text corpus of small non-coding RNA molecules: the MiNCor for microRNA annotations

José Camilla Sammartino
Department of Molecular Medicine
and Medical Biotechnology.
University of Naples Federico II, Italy
j.sammartino.88@gmail.com

Martin Krallinger
Centro Nacional de
Investigaciones Oncológicas.
Madrid, Spain
mkrallinger@cnio.es

Alfonso Valencia
Centro Nacional de
Investigaciones Oncológicas.
Madrid, Spain
avalencia@cnio.es

Abstract

MicroRNA are small non-coding molecules that act as post-transcriptional regulators of gene expression in a wide spectrum of biological states. Mostly, the information about microRNA is embedded in unstructured data (text files) which needs specific text mining techniques for its retrieval and analysis. These are generally based on supervised (or semi-supervised) learning methods, which require collections of neatly annotated and categorised training data. In this study we propose a comprehensive granular annotation protocol for the annotation of non-coding RNA molecules, focusing primarily on microRNA mentions. This annotation protocol was used to construct a manually annotated corpus (MiNCor Gold) for microRNA mentions as well as a large semi-automatically generated microRNA mentions silver standard corpus (MiNCor Silver) and a large microRNA name dictionary. Therefore, the efficiency of these standards was evaluated using a named entity recognition (NER) system in comparison with another microRNA mentions standard freely available online. The NER system trained with our silver corpus showed a better performance, with higher precision (96,67% vs. 94,00%) and recall (97,57% vs. 95,00%) on their test data and on our (precision 89,26% vs. 88,97% and recall 90,03% vs. 86,74%). The corpora and guidelines are freely downloadable at http://zope.bioinfo.cnio.es/mincor/minacor.tar.gz.

1 Introduction

MicroRNAs are small non-coding RNA molecules involved in the post-transcriptional regulation of gene expression. In the last decade they have been linked to a wide spectrum of biological/developmental processes and diseases including cancer, metabolic disorders or infectious diseases (Bayoumi et al., 2016; Smith et al., 2015; Pogue et al., 2014; Ohtsuka et al., 2015; Pileczki et al., 2016). MicroRNAs are post-transcriptional regulators of gene expression acting on the messenger RNA target. The maturation of microRNAs is a double step-and-area process, starting in the Nucleus of the cell, where is cleaved then exported in the Cytoplasm where is subjected to another cleavage which produce a double-strands microRNA of 22 nucleotides. This dsmicroRNA is recognised by the RNA-Induced Silencing Complex (RISC) (Stroynowska-Czerwinska et al., 2014). Even though the exact mechanism of action of RISC is yet fully understood, there are evidences that RISC is lead to the messenger RNA (mRNA) target by the microRNA, which has a homologous sequence to the 3' - UnTranslated Region (3' - UTR) of the target. The binding to this region allows the regulation process, that can happen before or during the translation in protein of the messenger, which means that is possible to have or don?t have protein products (Morozova et al., 2012). Temporal and spatial expression of these molecules is important as much as their expression levels, a modification in one of these can lead to a dysregulation of the biological processes in which they are involved, with effects that can expand to entire biological pathways. Different studies show the importance of a correct microRNA post-transcriptional regulation to prevent the development of pathological states and development defects (Bhaskaran and Mohan, 2014), but their importance is also enlightened by their possible application as fast, specific and non-invasive biomarkers in a large spectrum of harmful states (Rubio et al., 2016; Benz et al., 2016; Larrea et al., 2016). Furthermore, these molecules can be used as target in pharmacological therapies and clinical application (for the diagnosis and the follow-up) (Lin et al., 2014; Du et al., 2014; Mao et al., 2013). This promoted the publication of an increasing number of publications especially devoted to the study of microRNA biology as well as predictive bioinformatics analysis methodologies tailored to the characterisation of miRNA expression and target prediction.

Biomedical Natural Language Processing (BioNLP) techniques and text mining strategies can be applied for the retrieval, filtering and analysis of knowledge from unstructured data such the scientific literature. One of the main hurdles for the implementation of text mining building block components is the construction of manually annotated text-bound corpora, as they require usually a considerable human workload together with annotators with deep domain knowledge and basic linguistic expertise. The development of corpora is a time-consuming, tedious and very so much needed process for Text Mining and BLP methods (Neves, 2014). To promote advances in BLP, different competitive evaluations have been held (Hersh et al., 2004; Kim et al., 2004; Hirschman et al., 2005), in which distinct groups participated in different tasks, ranking from document retrieval, NER to complex relation/event extraction tasks (Hunter and Cohen, 2006). Those challenges resulted in valuable text corpora that have been re-used by the biomedical text mining community.

Despite the release of several manually annotated text corpora devoted to biological entities, there isn't one and only manual, work or reference that can be considered as a general guide to build specific guidelines which are usually written based on the background knowledge of the authors or don't include all the possible characteristics. Furthermore, the annotation process can be very variable and complex, due to the interconnection of different disciplines (medical/biological and linguistic) and the different aims of the annotation (chemical compounds, disease, connection between mutated proteins and disease, case reports).

The assembling of a corpus requires specific documents that describe the annotation process and define its guidelines.

As for microRNAs, several attempts have been made to facilitate the extraction of information directly from the literature (Bagewadi et al., 2014; Griffiths-Jones et al., 2006; Li et al., 2015; Naeem et al., 2010; Xie et al., 2013). To our knowledge there are three freely-available corpora for microRNA (miRNA) mentions, two of them, MirBase and MirTex (Griffiths-Jones et al., 2006; Li et al., 2015), do provide very short annotation guidelines (Ambros et al., 2003; Meyers et al., 2008; Griffiths-Jones, 2004) for the annotation of microRNA mentions which mostly focus on the identification of single mentions, without considering more granular annotation types. The third corpus (SCAI corpus) (Bagewadi et al., 2014), does provide additional details and a set of annotation rules, but we believe that it underspecified some of the relevant annotation criteria and it primarily focuses only on human microRNA Mentions. For instance it covers the annotation of species specific prefixes, e.g. hsa for human miRNAs, but does not annotated terms such as human preceding miRNA mentions. Moreover, general prefixes (anti-, onco-, pre-, pri-), specific miRNA class names (angiomir, antagomir, isomir), as well as non-coding RNA names are not included in the annotation process.

Here we propose a comprehensive annotation protocol for labelling microRNA mentions in biomedical literature. It encompasses all microRNA mentions regardless of the species or origin, the maturing step or the classification and includes also a class of non-coding RNA names and miRNA clusters. This annotation protocol has been iteratively refined and was then used for the annotation of the MiNCor corpus, which as used for the evaluation of several microRNA mentions recognition approaches. We believe that the release of this MiNCor corpus guidelines might be useful as an annotation template for the corpus construction of other biomedical entities.

We tested our corpus in comparison with the SCAI corpus, which is to our knowledge, the one whose guidelines are the most comprehensive so far. Therefore, to test the efficiency of our corpus we trained and tested a named entity recognition (NER) system with it and evaluated the results in comparison with SCAI, whose trainer and tester for the NER system were the only ones with characteristics that could be compared to ours.

2 Annotation protocol and guidelines

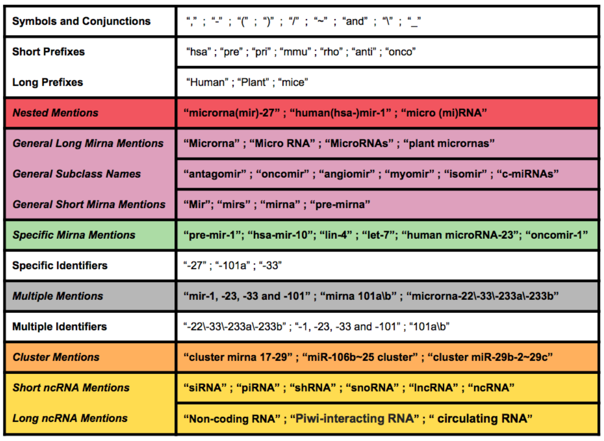
The guidelines for the MiNCor annotation protocol is composed of a 14 pages written manual defined by a biotechnologist with extensive biological knowledge, integrating information from previous miRNA corpora, revision of multiple different resources (NCBI, MeSH terms, miRNA review articles) and the model of the Manual for annotation of chemical entities of the CHEMDNER corpus (Krallinger et al., 2015). The annotation protocol is structured into rule types together with example cases, which we call the GPNCE annotation system, standing for general rules, positive rules, negative rules, class rules and examples. We believe that structuring the annotation protocol into such rules, makes it easier to follow the annotation criteria by the human annotators during the labelling of the mentions.

2.1 The GPCNE annotation protocol

We based our guidelines on a three phase annotation protocol that we called GPNCE (General, Positive, Negative, Class and Examples).

We firstly describe the different classes that can be identified in literature. We cover in detail six different classes of microRNA mentions: (1) general microRNA names, (2) specific microRNA names, (3) multiple microRNA mentions, (4) nested microRNA mentions, (5) microRNA cluster mentions and (6) other/non-coding RNA mentions. Figure 1 provides different examples for the classes.

In the second phase we propose three types of rules for the annotation: General, Positive and Negative. The General Rules describe the decisions the annotator should take into account during the annotation process (what constitutes at a general level a miRNA mention and how to deal with cases of uncertainty). The Positive Rules describe how to annotate correct miRNA mentions, what to include in the mentions (positive word-boundaries, prefixes, suffixes, symbols) and illustrate the criteria through different examples of correctly annotated mentions. The Negative Rules, cover what needs to be excluded from the mentions during the annotation (negative-word-boundaries, prefixes, part-of-speech entities, other entities, wrong mentions). All the rules described included examples with positive cases (specified by the check mark symbol) and negative ones (specified by the cross mark symbol). The definition of these rules was based on the Manual for annotation of chemical entities of the CHEMDNER corpus (Krallinger et al., 2015).



Symbols and Conjunctions	“,” ; “-” ; “(” ; “)” ; “/” ; “~” ; “and” ; “\” ; “_”
Short Prefixes	“hsa” ; “pre” ; “pri” ; “mmu” ; “rho” ; “anti” ; “onco”
Long Prefixes	“Human” ; “Plant” ; “mice”
Nested Mentions	**“microrna(mir)-27” ; “human(hsa-)mir-1” ; “micro (mi)RNA”**
General Long Mirna Mentions	**“Microrna” ; “Micro RNA” ; “MicroRNAs” ; “plant micrornas”**
General Subclass Names	**“antagomir” ; “oncomir” ; “angiomir” ; “myomir” ; “isomir” ; “c-miRNAs”**
General Short Mirna Mentions	**“Mir”; “mirs” ; “mirna” ; “pre-mirna”**
Specific Mirna Mentions	**“pre-mir-1”; “hsa-mir-10”; “lin-4” ; “let-7”; “human microRNA-23”; “oncomir-1”**
Specific Identifiers	“-27” ; “-101a” ; “-33”
Multiple Mentions	**“mir-1, -23, -33 and -101” ; “mirna 101a\b” ; “microrna-22\-33\-233a\-233b”**
Multiple Identifiers	“-22\-33\-233a\-233b” ; “-1, -23, -33 and -101” ; “101a\b”
Cluster Mentions	**“cluster mirna 17-29” ; “miR-106b~25 cluster” ; “cluster miR-29b-2~29c”**
Short ncRNA Mentions	**“siRNA” ; “piRNA” ; “shRNA” ; “snoRNA” ; “lncRNA” ; “ncRNA”**
Long ncRNA Mentions	**“Non-coding RNA” ; “Piwi-interacting RNA” ; “ circulating RNA”**

Figure 1: Here are shown the classes of microRNA mention and the elements that are part of them. In white are the different components of the mentions that can help discriminate the different classes. The examples of different classes are highlighted with different colours (one for each class).

The last phase, Examples, consist in two appendix at the end of the manual in which are represented different examples of annotated mentions in sentences extracted from abstracts and possible errors to avoid. To help visualising the correct labels, those were highlighted with a specific colour coding system in regard of the class to which them belong.

3 MiNCor Corpora

We decided to build two different corpus for microRNA mention following the GPCNE annotation protocol. The first one is a manually labeled corpus of 102 abstracts called MiNCor Gold, the second is a semi-automatically generated corpus of 302K sentences called MiNCor Silver. The corpora and the guidelines can be downloaded at `http://zope.bioinfo.cnio.es/mincor/minacor.tar.gz`. The directory contains six files with a ’README.txt’ file which contains all the descriptions of the other files.

3.1 MiNCor Gold

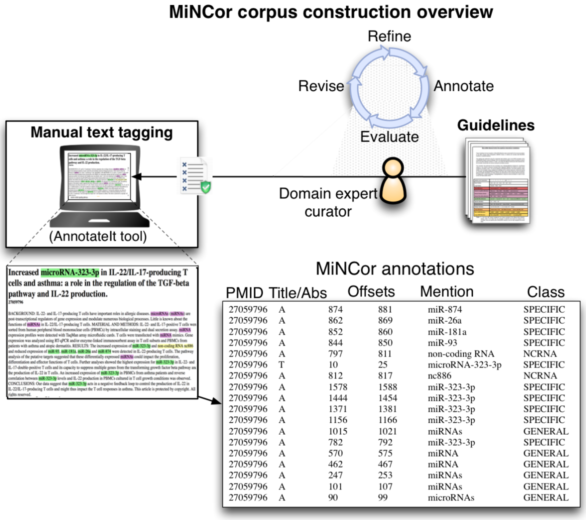
Using the MiNCor annotation guidelines a domain expert annotator performed a manual labelling of 102 abstracts with at least one microRNA mention. The abstracts were randomly selected from 3869 abstracts retrieved using the MeSH query ”mirna” on Pubmed but restricting the search only to papers published in 2016. The labelling was performed manually using the customised AnnotateIt web-interface `http://ubio.bioinfo.cnio.es/people/fleitner/mirnaner_test_250.html`, similar to the one used for the annotation of the CHEMDNER-Patents Corpus (Krallinger et al., 2015), but adjusting the different classes and labels to the microRNA Mention Classes, a schematic overviw of the annotation protocol is summarised in figure 2. Out of the 102 abstracts we extracted a total of 1154 mentions. Table 1 provides an overview of the distribution of the microRNA class types.

Type of Mention	Total mentions
General microRNA	607
Specific microRNA	501
Multiple microRNA	14
Nested microRNA	2
Cluster microRNA	1
ncRNA microRNA	29

Table 1: MicroRNA class types number of mentions.

To validate the annotation process 20 of these abstracts were randomly selected and de novo annotated using the same annotation guidelines by a second annotator. The results obtained were then compared with the first annotation considering only perfect mention matches. The annotator agreement scores resulted in: Precision: 99,00, Recall: 99,45 and F1: 99,22. All the non-overlapping labeled entities were analysed and modified only when the annotators could agree, if there still was uncertainty the mentions were left unlabelled. The errors in the annotation mostly concerned the non-coring RNA class mentions :

- [...the long intergenic non-coding RNAs (lincRNAs) expressed in...]
- Should we consider ’non-coding RNAs (lincRNAs)’ as a unique mention or as a two separate mention ’non-coding RNAs’, ’lincRNAs’?
- Should ’long intergenic’ be included in the mention?



Increased microRNA-323-3p in IL-22/IL-17-producing T cells and asthma: a role in the regulation of the TGF-beta pathway and IL-22 production.
27059796

BACKGROUND: IL-22- and IL-17-producing T cells have important roles in allergic diseases. microRNAs (miRNAs) are post-transcriptional regulators of gene expression and modulate numerous biological processes. Little is known about the functions of miRNAs in IL-22/IL-17-producing T cells. MATERIAL AND METHODS: IL-22- and IL-17-positive T cells were sorted from human peripheral blood mononuclear cells (PBMCs) by intracellular staining and dual secretion assay. miRNA expression profiles were detected with TaqMan array microfluidic cards. T cells were transfected with miRNA mimics. Gene expression was analyzed using RT-qPCR and/or enzyme-linked immunosorbent assay in T cell subsets and PBMCs from patients with asthma and atopic dermatitis. RESULTS: The increased expression of miR-323-3p and non-coding RNA nc886 and reduced expression of miR-93, miR-181a, miR-26a and miR-874 were detected in IL-22-producing T cells. The pathway analysis of the putative targets suggested that these differentially expressed miRNAs could impact the proliferation, differentiation and effector functions of T cells. Further analyses showed the highest expression for miR-323-3p in IL-22- and IL-17-double-positive T cells and its capacity to suppress multiple genes from the transforming growth factor beta pathway and the production of IL-22 in T cells. An increased expression of miR-323-3p in PBMCs from asthma patients and reverse correlation between miR-323-3p levels and IL-22 production in PBMCs cultured in T cell growth conditions was observed. CONCLUSIONS: Our data suggest that miR-323-3p acts in a negative feedback loop to control the production of IL-22 in IL-22/IL-17-producing T cells and might thus impact the T cell responses in asthma. This article is protected by copyright. All rights reserved.

PMID	Title/Abs	Offsets		Mention	Class
27059796	A	874	881	miR-874	SPECIFIC
27059796	A	862	869	miR-26a	SPECIFIC
27059796	A	852	860	miR-181a	SPECIFIC
27059796	A	844	850	miR-93	SPECIFIC
27059796	A	797	811	non-coding RNA	NCRNA
27059796	T	10	25	microRNA-323-3p	SPECIFIC
27059796	A	812	817	nc886	NCRNA
27059796	A	1578	1588	miR-323-3p	SPECIFIC
27059796	A	1444	1454	miR-323-3p	SPECIFIC
27059796	A	1371	1381	miR-323-3p	SPECIFIC
27059796	A	1156	1166	miR-323-3p	SPECIFIC
27059796	A	1015	1021	miRNAs	GENERAL
27059796	A	782	792	miR-323-3p	SPECIFIC
27059796	A	570	575	miRNA	GENERAL
27059796	A	462	467	miRNA	GENERAL
27059796	A	247	253	miRNAs	GENERAL
27059796	A	101	107	miRNAs	GENERAL
27059796	A	90	99	microRNAs	GENERAL

Figure 2: This figure shows a schematic overview of the process used for the construction of the MiNCor corpus.

In this example, if we follow the guidelines, the correct labelling should be ’long intergenic non-coding RNAs’ and ’lincRNAs’, but in this type of cases the annotators labelled differently (one did the label as a single mention and the other did as two separate mentions) showing uncertainty. Therefore, we suggest to not label the mention when there are doubts.

3.2 MiNCor Silver

The MiNCor Silver was obtained from sentences that were derived from PubMed results containing the MeSH term microRNA as well as additional manually defined microRNA search query terms (”mirna”; ”microrna”; ”non-coding RNA”; ”lin-4”; ”let-7”, ”antagomir”; ”oncomir”). The research was limited to all the abstracts ad full text papers published starting from 2016. All the resulting files were segmented into sentences and then tagged using a large dictionary of microRNA names (MiNCor lexicon). This dictionary contained names derived from multiple microRNA databases as well as microRNA mentions detected by GNormplus. We carried out a dictionary expansion step taking into account the nomenclature guidelines of microRNAs by considering core terms (e.g. miRNA, microRNA), prefixes (e.g. hsa, mmu) and suffixes (e.g. -101; -23a/b). A dictionary pruning step was carried out to remove highly ambiguous mentions (e.g. ’MIR’ was found to be referred to other entities as for example the 'Space Station Mir' (Johannes et al., 2016)). After applying the dictionary look-up we additionally used a cascade of rules to adjust the mention boundaries, to cover for instance mentions of co-ordinated microRNAs or lists of microRNAs. In the end our training silver corpus had a total of 302'560 sentences, over 3'000'000 of tokens and over 175K labeled microRNA mentions, the sum of the entities used in the different dictionaries for the post-processing amount to 788'784 different terms in total.

4 Comparison with other corpus

We decided to use both the MiNCor corpus and the SCAI miRNA corpus (Bagewadi et al., 2014) as a test set to evaluate the performance of a CRF-based miRNA entity tagger. We chose the SCAI corpus, because it did provide annotation criteria, and thus allowed some interpretation of differences in the annotation process. We constructed a miRNA entity tagger using the NERsuite toolkit (Cho et al., 2010). The NERsuite toolkit is freely available at `http://nersuite.nlplab.org`. It is based on the CRFsuite (`http://www.chokkan.org/software/crfsuite/` an implementation of the Conditional Random Field) (Okazaki, 2007) and includes different feature types commonly used for biomedical NER tasks, including aspects covering lemmatization, Part-Of-Speech and word morphology.

We trained two different NER models, one using the miRNA SCAI training set (201 manually labeled abstracts) and another based on a large silver standard MiNCor training dataset comprising 302K sentences. We generated a CRF model both using the SCAI training set and the MiNCor silver standard training set.

Both the two corpora used for the train of the models were segmented in sentences and tokenised. At token level the two corpora were lemmatised, labelled with Part-Of-Speech and chunking tags, and labelled following the I.O.B. format. The results were defined in terms of Precision, Recall and F1. The obtained results with the two models are shown in table 2, for the SCAI test set, and in table 3, using the MiNCor gold standard test set. Table 4 shows the overall statistics of the two test corpora (MiNCor Gold and SCAI).

Score	SCAI-Test	MiNCor-Test
Precision	94,00	88,97
Recall	95,00	86,74
F-score	94,49	87,84

Table 2: Results for the CRF model trained on the SCAI training set.

Score	SCAI-Test	MiNCor-Test
Precision	96,67	89,26
Recall	97,57	90,03
F-score	97,11	89,64

Table 3: Results for the CRF model trained on the MiNCor silver standard training set

5 Results and Discussion

Using our guidelines we manually annotated 102 abstracts retrieved form Pubmed using the MeSH query "mirna" and filtering the results including only the recent ones (2016). At the same time, with a more refined search on Pubmed (including different MeSH queries) we extracted 302K sentences that were semi-automatically labeled following our guidelines and subsequently pruned with dictionary look-up and a cascade of rules to adjust the mention boundaries. We then tested our corpora in comparison with the SCAI manually labelled corpus using the NERsuite toolkits to perform the named entity recognition task for microRNA mention in literature. We used the MiNCor Gold as our test and the MiNCor Silver as the trainer to build our model. To obtain the SCAI model we trained the NERsuite with their trainer downloadable at `http://www.scai.fraunhofer.de/mirna-corpora.html`. As shown in Table 2 and Table 3, the microRNA tagger models, trained using the SCAI training set (Table 2) and our dictionary/rule-based Silver Standard training set (Table 3), report lower scores when using our corpus as gold standard (second column of the two tables). This is due to the more granular definition of the microRNA mentions and by including for instance also other ncRNA types that were not labelled in the used training collections. On the other end, our model had a better performance in comparison with SCAI on both test sets, this is due to our model, even though not being manually curated, covers more possible mentions, including microRNA mentions for all different species and

Statistic	SCAI-Test	MiNCor-test
Abstracts	100	102
Sentences	780	1063
Total Mentions	712	1154
Unique Mentions	130	232

Table 4: Statistics of the two microRNA test corpora.

biosynthesis steps, furthermore, it includes more classes of mentions, leading to a more comprehensive identification.

Even if our model had a better performance, the resulting score wasn't perfect. Some of the main sources of errors related to the microRNA mention recognition was due to mention of lists of microRNAs, where microRNA mentions are expressed as multiple overlapping entity mentions (mir-1, -23, -33 and -101). Other errors occurred in the labelling of non-coding RNAs.

Non-coding RNA mentions are hard to define because there isn't a specific nomenclature to which the researcher can refer. Nevertheless , there are resources online (NCBI, MeSH terms, miRNA review articles, books) that can help in the definition of this class. What we tried to do was to give rules for the identification of non-coding RNA mentions, where the most important was that in case of uncertainty the mention shouldn't be labelled, which results in a lower accuracy for the model.

6 Conclusion and future works

Here we have presented the MiNCor corpora and the Guidelines for the Annotation for microRNA and non-coding RNA mentions in scientific literature. The aim of this work was to provide annotation guidelines that are comprehensive and explicative, using different examples for the annotation and rules to help the annotator during the process. The availability of exhaustive guidelines for the annotation of biomedical entities is a very important contribution for Biomedical Natural Language Processing tasks, because gives the researcher the possibility to have a standardised tool that can help in the definition of a line of research even without extensive knowledge of the field. Furthermore, the possibility to use predefined guidelines for the construction of corpora can reduce the time needed for the process.

We also constructed two corpora (gold and silver) using our guidelines and tested them with a named entity recognition task using the NERsuite toolkit and comparing the results with another microRNA tagger already available. Manually curated corpora are considered a gold standard in Natural Language Processing because they can generally reach higher level of accuracy. In our case that is not true, which provide an example of a good surrogate for manually annotated gold standard corpora. At the moment there aren't very large gold standard for microRNA mention that encompass all the possible characteristic and types of mention, which is why our MiNCor Silver can be considered a better option, even though not being manually curated, as shown by the results we obtained.

In the future, our intent is to enlarge our guidelines with other types of non-coding RNAs (e.g. ribosomial RNAs, transfer RNAs) that are not included at the moment, provide a larger corpus of microRNAs derived from full text and patent abstract sentences and describe additional rules to help defying the relations of these molecules with other biological entities (e.g. chemical compounds, genes, proteins).

References

Victor Ambros, Bonnie Bartel, David P Bartel, Christopher B Burge, James C Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R Eddy, SAM Griffiths-Jones, Mhairi Marshall, et al. 2003. A uniform system for microrna annotation. *Rna*, 9(3):277–279.

Shweta Bagewadi, Tamara Bobić, Martin Hofmann-Apitius, Juliane Fluck, and Roman Klinger. 2014. Detecting mirna mentions and relations in biomedical literature. *F1000Research*, 3.

Ahmed S Bayoumi, Amer Sayed, Zuzana Broskova, Jian-Peng Teoh, James Wilson, Huabo Su, Yao-Liang Tang, and Il-man Kim. 2016. Crosstalk between long noncoding rnas and micrornas in health and disease. *International journal of molecular sciences*, 17(3):356.

Fabian Benz, Sanchari Roy, Christian Trautwein, Christoph Roderburg, and Tom Luedde. 2016. Circulating micrornas as biomarkers for sepsis. *International journal of molecular sciences*, 17(1):78.

M Bhaskaran and M Mohan. 2014. Micrornas history, biogenesis, and their evolving role in animal development and disease. *Veterinary Pathology Online*, 51(4):759–774.

HC Cho, N Okazaki, M Miwa, and J Tsujii. 2010. Nersuite: a named entity recognition toolkit. *Tsujii Laboratory, Department of Information Science, University of Tokyo, Tokyo, Japan.*

Bowen Du, Zhe Wang, Xin Zhang, Shipeng Feng, Guoxin Wang, Jianxing He, and Biliang Zhang. 2014. Microrna-545 suppresses cell proliferation by targeting cyclin d1 and cdk4 in lung cancer cells. *PloS one*, 9(2):e88022.

Sam Griffiths-Jones, Russell J Grocock, Stijn Van Dongen, Alex Bateman, and Anton J Enright. 2006. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic acids research*, 34(suppl 1):D140–D144.

Sam Griffiths-Jones. 2004. The microrna registry. *Nucleic acids research*, 32(suppl 1):D109–D111.

William Hersh, Ravi Teja Bhupatiraju, and Sarah Corley. 2004. Enhancing access to the bibliome: the trec genomics track. *Medinfo*, 11(Pt 2):773–777.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of biocreative: critical assessment of information extraction for biology. *BMC bioinformatics*, 6(Suppl 1):S1.

Lawrence Hunter and K Bretonnel Cohen. 2006. Biomedical language processing: what's beyond pubmed? *Molecular cell*, 21(5):589–594.

Bernd Johannes, Vyacheslav Salnitski, Alexander Dudukin, Lev Shevchenko, and Sergey Bronnikov. 2016. Performance assessment in the pilot experiment on board space stations mir and iss. *Aerospace medicine and human performance*, 87(6):534–544.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(S1):1–17.

Erika Larrea, Carla Sole, Lorea Manterola, Ibai Goicoechea, María Armesto, María Arestin, María M Caffarel, Angela M Araujo, María Araiz, Marta Fernandez-Mercado, et al. 2016. New concepts in cancer biomarkers: Circulating mirnas in liquid biopsies. *International journal of molecular sciences*, 17(5):627.

Gang Li, Karen E Ross, Cecilia N Arighi, Yifan Peng, Cathy H Wu, and K Vijay-Shanker. 2015. mirtex: A text mining system for mirna-gene relation extraction. *PLoS Comput Biol*, 11(9):e1004391.

Regina Lin, Ling Chen, Gang Chen, Chunyan Hu, Shan Jiang, Jose Sevilla, Ying Wan, John H Sampson, Bo Zhu, and Qi-Jing Li. 2014. Targeting mir-23a in cd8+ cytotoxic t lymphocytes prevents tumor-dependent immunosuppression. *The Journal of clinical investigation*, 124(12):5352–5367.

Yiping Mao, Ramkumar Mohan, Shungang Zhang, and Xiaoqing Tang. 2013. Micrornas as pharmacological targets in diabetes. *Pharmacological research*, 75:37–47.

Blake C Meyers, Michael J Axtell, Bonnie Bartel, David P Bartel, David Baulcombe, John L Bowman, Xiaofeng Cao, James C Carrington, Xuemei Chen, Pamela J Green, et al. 2008. Criteria for annotation of plant micrornas. *The Plant Cell*, 20(12):3186–3190.

Nadya Morozova, Andrei Zinovyev, Nora Nonne, Linda-Louise Pritchard, Alexander N Gorban, and Annick Harel-Bellan. 2012. Kinetic signatures of microrna modes of action. *Rna*, 18(9):1635–1655.

Haroon Naeem, Robert Küffner, Gergely Csaba, and Ralf Zimmer. 2010. mirsel: automated extraction of associations between micrornas and genes from the biomedical literature. *BMC bioinformatics*, 11(1):135.

Mariana Neves. 2014. An analysis on the entity annotations in biological corpora. *F1000Research*, 3.

Masahisa Ohtsuka, Hui Ling, Yuichiro Doki, Masaki Mori, and George Adrian Calin. 2015. Microrna processing and human cancer. *Journal of clinical medicine*, 4(8):1651–1667.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Valentina Pileczki, Roxana Cojocneanu-Petric, Mahafarin Maralani, Ioana Berindan Neagoe, and Robert Sandulescu. 2016. Micrornas as regulators of apoptosis mechanisms in cancer. *Clujul Medical*, 89(1):50.

Aileen I Pogue, James M Hill, and Walter J Lukiw. 2014. Microrna (mirna): sequence and stability, viroid-like properties, and disease association in the cns. *Brain research*, 1584:73–79.

Mercedes Rubio, Quique Bassat, Xavier Estivill, and Alfredo Mayor. 2016. Tying malaria and micrornas: from the biology to future diagnostic perspectives. *Malaria journal*, 15(1):1.

Tanya Smith, Cha Rajakaruna, Massimo Caputo, and Costanza Emanueli. 2015. Micrornas in congenital heart disease. *Annals of translational medicine*, 3(21).

Anna Stroynowska-Czerwinska, Agnieszka Fiszer, and Wlodzimierz J Krzyzosiak. 2014. The panorama of mirna-mediated mechanisms in mammalian cells. *Cellular and Molecular Life Sciences*, 71(12):2253–2270.

Boya Xie, Qin Ding, Hongjin Han, and Di Wu. 2013. mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, page btt014.