METHOD ARTICLE

REVISED Detecting miRNA Mentions and Relations in Biomedical Literature [version 3; peer review: 2 approved, 1 approved with reservations]

Shweta Bagewadi[1,2], Tamara Bobić[3], Martin Hofmann-Apitius[1,2], Juliane Fluck[1], Roman Klinger[4]

[1]Fraunhofer SCAI, Bioinformatics, Schloss Birlinghoven, 53754, Sankt Augustin, Germany
[2]University of Bonn, B-IT, Dahlmannstr. 2, 53113 Bonn, Germany
[3]Hasso Plattner Institute Potsdam, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Potsdam, Germany
[4]Semantic Computing Group, CIT-EC, Bielefeld University, 33615 Bielefeld, Germany

v3 **First published:** 28 Aug 2014, **3**:205
https://doi.org/10.12688/f1000research.4591.1
Second version: 23 Dec 2014, **3**:205
https://doi.org/10.12688/f1000research.4591.2
Latest published: 01 Oct 2015, **3**:205
https://doi.org/10.12688/f1000research.4591.3

Abstract

Introduction: MicroRNAs (miRNAs) have demonstrated their potential as post-transcriptional gene expression regulators, participating in a wide spectrum of regulatory events such as apoptosis, differentiation, and stress response. Apart from the role of miRNAs in normal physiology, their dysregulation is implicated in a vast array of diseases. Dissection of miRNA-related associations are valuable for contemplating their mechanism in diseases, leading to the discovery of novel miRNAs for disease prognosis, diagnosis, and therapy.

Motivation: Apart from databases and prediction tools, miRNA-related information is largely available as unstructured text. Manual retrieval of these associations can be labor-intensive due to steadily growing number of publications. Additionally, most of the published miRNA entity recognition methods are keyword based, further subjected to manual inspection for retrieval of relations. Despite the fact that several databases host miRNA-associations derived from text, lower sensitivity and lack of published details for miRNA entity recognition and associated relations identification has motivated the need for developing comprehensive methods that are freely available for the scientific community. Additionally, the lack of a standard corpus for miRNA-relations has caused difficulty in evaluating the available systems.

We propose methods to automatically extract mentions of miRNAs, species, genes/proteins, disease, and relations from scientific literature. Our generated corpora, along with dictionaries, and miRNA

Open Peer Review

Approval Status ✓ ? ✓

	1	2	3
version 3 (revision) 01 Oct 2015			
version 2 (revision) 23 Dec 2014	✓ view		✓ view
version 1 28 Aug 2014	? view	? view	

1. **Sofie Van Landeghem**, Ghent University, Ghent, Belgium, Belgium
2. **Filip Ginter**, University of Turku, Turku, Finland
3. **Robert Leaman**, National Institutes of Health, Bethesda, USA

Any reports and responses or comments on the article can be found at the end of the article.

regular expression are freely available for academic purposes. To our knowledge, these resources are the most comprehensive developed so far.

Results: The identification of specific miRNA mentions reaches a recall of 0.94 and precision of 0.93. Extraction of miRNA-disease and miRNA-gene relations lead to an F_1 score of up to 0.76. A comparison of the information extracted by our approach to the databases *miR2Disease* and *miRSel* for the extraction of Alzheimer's disease related relations shows the capability of our proposed methods in identifying correct relations with improved sensitivity. The published resources and described methods can help the researchers for maximal retrieval of miRNA-relations and generation of miRNA-regulatory networks.

Availability: The training and test corpora, annotation guidelines, developed dictionaries, and supplementary files are available at http://www.scai.fraunhofer.de/mirna-corpora.html

Keywords
MicroRNAs, prediction algorithms, corpus

This article is included in the Machine learning: life sciences collection.

Corresponding author: Shweta Bagewadi (shweta.bagewadi@scai.fraunhofer.de)

Competing interests: No competing interests were disclosed.

Grant information: Shweta Bagewadi was supported by University of Bonn. Tamara Bobić was partially funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School during her contribution to this work at Fraunhofer SCAI.

How to cite this article: Bagewadi S, Bobić T, Hofmann-Apitius M *et al.* **Detecting miRNA Mentions and Relations in Biomedical Literature [version 3; peer review: 2 approved, 1 approved with reservations]** F1000Research 2015, **3**:205 https://doi.org/10.12688/f1000research.4591.3

First published: 28 Aug 2014, **3**:205 https://doi.org/10.12688/f1000research.4591.1

REVISED **Amendments from Version 2**

The final revised version of the manuscript includes changes as per the reviewers' recommendation. We have mainly modified text in the "Corpus selection, annotation and properties" section to simplify the ambiguous texts, as pointed out by Robert Leaman. Additionally, grammatical errors pointed out by the reviewers have also been corrected. Please read the response provided to reviewers' comments for detailed information of the changes.

See referee reports

Introduction

Functionally important non-coding RNAs (ncRNAs) are now better understood with the progress of high-throughput technologies. Discovery of the major class of ncRNAs, microRNAs (miRNAs[1]) has further facilitated the molecular aspects of biomedical research.

MicroRNAs are a large group of small endogenous single-stranded non-coding RNAs (17–22nt long) found in eukaryotic cells. They post-transcriptionally regulate gene expression of specific mRNAs by degradation, translational inhibition, or destabilization of the targets (transcripts of protein-coding genes)[2]. Esquela-Kerscher *et al.* have reported on miRNAs involvement in almost every regulation aspect of biological processes such as apoptosis, and stress response[3]. Wubin *et al.* demonstrated that miR-29a regulatory circuitry plays an important role in epididymal development and its functions[4]. Additionally, tissue-specificity of miRNAs has been shown to provide a better clue of their fundamental roles in normal physiology[5].

Dysregulation of miRNAs and their ability to regulate repertoires of genes (as well as co-ordinate multiple biological pathways) has been linked to several diseases[6,7]. One example is chronic lymphocytic leukemia where (in about 68% of the cases) miRNA genes (*miR15* and *miR16*) are missing or down-regulated[8]. Thus, uncovering the relations between miRNAs and diseases as well as genes/proteins is crucial for our understanding of miRNA regulatory mechanisms for diagnosis and therapy[9,10].

Several databases, prediction algorithms and tools are available, providing insight into miRNA-disease and miRNA-mRNA associations. Although the detailed target recognition mechanism is still elusive, several algorithms attempt to predict miRNA targets. However, a limited precision of 0.50 and recall of 0.12 has been reported when evaluated against proteomics supported miRNA targets[11]. Despite the fact that these resources provide insight into miRNA-associated relationships, the majority of relations are scattered as unstructured text in scientific publications[12]. Figure 1 shows the growth of publications in MEDLINE and in addition depicts the normalized growth of publications that reference the keyword "microRNA".

Some databases such as *miR2Disease* and *PhenomiR* store manually extracted relations from literature. The *miR2Disease* database[13] contains information about miRNA-disease relationships with 3273 entries (as of the last update on March 14, 2011). *PhenomiR*[14] is a database on miRNA-related phenotypes extracted from published experiments. It consists of 675 unique miRNAs, 145 diseases, and 98 bioprocesses from 365 articles (Version 2.0, last updated on February 2011). *TarBase*[11] hosts more than 6500 experimentally validated miRNA targets extracted from literature.

However, manual retrieval of relevant articles and extraction of relation mentions from them is labor-intensive. A solution is to use text-mining techniques. Moreover, the vast majority of the research in this direction is mainly focused around extraction of

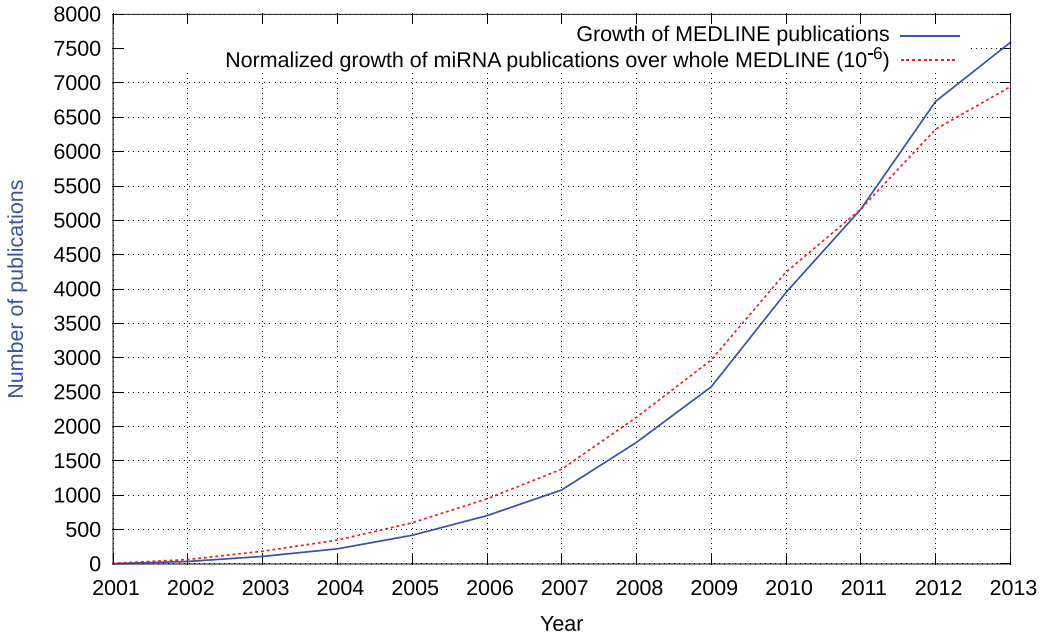


Figure 1. Growth of miRNA-related publications in comparison with the growth of MEDLINE. The dotted line points out the relative increase of miRNA-related publications per year in comparison to the growth of MEDLINE (as of 31 December, 2013).

protein-protein interactions[15]. On the contrary, miRNA relation extraction is still naive. The shift of focus towards identification of miRNA-relations is slowly establishing with the rise in systems approaches to investigate complex diseases. The manually curated database *miRTarbase*[16] incorporates such text-mining techniques to retrieve miRNA-related articles. Recently, the *miRCancer* database has been constructed using a rule-based approach to extract miRNA-cancer associations from text[17]. As of June 14, 2014, this database contains 2271 associations between 38562 miRNAs and 161 human cancers from 1478 articles.

Related work

Text-mining technologies are established for a variety of applications. For instance, the BioCreative competition[18,19] and BioNLP Shared Task[20–22] series have been conducted to benchmark text mining techniques for gene mention identification, protein-protein relation extraction and event extraction, among others.

To our knowledge, only limited work has been carried out in the area of miRNA-related text-mining. Murray *et al.* considered miRNA-gene associations from PubMed database using semantic search techniques[23]. For their analysis, experimentally derived datasets were examined, combined with network analysis and ontological enrichment. Regular expressions were used to detect miRNA mentions. The authors claim to have optimized the approach to reach 100% accuracy and recall for detecting miRNAs mentions as in *miRBase*. Relations were identified based on a manually curated rule set. The authors extracted 1165 associations between 270 miRNAs and 581 genes from the whole MEDLINE.

The freely available *miRSel*[12] database integrates automatically extracted miRNA-target relationships from **PubMed Abstract**s. A set of regular expressions is used for miRNA recognition that matches all *miRBase* synonyms and generic occurrences. The authors reach a recall of 0.96 and precision of 1.0 on 50 manually annotated abstracts for miRNA mention identification. Further, the relations between miRNA and genes were extracted at sentence level employing a rule-based approach. They evaluated on 89 sentences from 50 abstracts resulting in a recall of 0.90 and precision of 0.65. Currently, it hosts 3690 miRNA-gene interactions[11].

Since the miRNA naming convention has been formalized very early in comparison to other biological entities such as genes and proteins, applying text-mining approaches is relatively simple[17]. Thus, most of the previously applied text mining approaches for miRNA detection has been based on keywords. *miRCancer* uses keywords to obtain abstracts from PubMed, further miRNA entities have been identified using regular expressions based on prefix and suffix variations. Similarly, *miRWalk* database uses keyword search approach to download abstracts and applies a curated dictionary (compiled from six databases) for miRNA identification of human, rat, and mouse species[24]. *TarBase*, *miR2Disease*, *miRTarBase*, and several others have followed related search strategies. However, several authors still tend to use naming variations for acronyms, abbreviations, nested representations, *etc.* for listing miRNAs. Additionally, in contrast to the previous text-mining approaches focusing purely on miRNA gene relations, we extend the information extraction approach additionally to retrieve miRNA-disease relations. Furthermore, we evaluate our approach using a larger corpus to achieve robustness. We differentiate between actual miRNA mentions (refered to as Specific miRNAs) and co-referencing miRNAs (Non-Specific miRNAs), which could in addition enhance keyword search. We evaluated three different relation extraction approaches, namely co-occurrence, tri-occurrence and machine learning based methods.

To support further research, our corpora are made publicly available in an established XML format as proposed by Pyysalo *et al.*[25], as well as the regular expressions used for miRNAs named entity recognition. In addition, our dictionary for trigger term detection and general miRNA mention identification are made available. To our knowledge, the annotated corpora as well as the information extraction resources are the most comprehensive developed so far.

Methods

Data curation and corpus selection

Named entities annotation. Mentions of miRNAs consisting of keywords (case-insensitive and not containing any suffixed numerical identifier) such as *"Micro-RNAs"* or *"miRs"* are annotated as Non-Specific miRNA. Names of particular miRNAs such as *miRNA-101*, suffixed with numerical identifiers are labeled as Specific miRNA. Numerical identifiers (separated by delimiters such as ",", "/", and "and") occurring as part of specific miRNA mentions are annotated as a single entity. Box 1 depicts the annotation of specific miRNA mentions (including an example for part mentions). In addition, Disease, Gene/Protein, Species, and Relation Trigger are annotated. The detailed annotation guideline for annotating specific miRNA mentions is available as a supplementary file.

Box 1. Example of miRNAs annotations. Here "*-181b*", and "*-181c*" are the part mentions annotated as a single entity along with "*miR-181a*" in box. A non-specific miRNA mention is shown in italics.

Interesting results were obtained from miR-181a, -181b, and -181c. These set of brain-enriched *miRNAs* are down-regulated in glioblastoma. However, miR-222, and miR-128 are strongly up-regulated.

Mentions of disease names, disease abbreviations, signs, deficiencies, physiological dysfunction, disease symptoms, disorders, abnormalities, or organ damages are annotated as Disease. Only disease nouns were considered, adjective terms such as "*Diabetic patients*" are not marked; however, specific adjectives that can be treated as nouns were marked, e.g. "Parkinson's disease patients". Mentions referring to proteins/genes which are either single word (*e.g. "trypsin"*), multi-word, gene symbols (*e.g. "SMN"*), or complex names (including of hyphens, slashes, Greek letters, Roman or Arabic numerals) are annotated as Gene/Protein. Only those organisms that are having published miRNA sequences and annotations represented in *miRBase* database are labeled as Species. Any verb, noun, verb phrase, or noun phrase associating miRNA mention to either labeled disease or gene/protein term is annotated as Relation Trigger.

Relations annotation. We restrict the relationship extraction to sentence level and four different interacting entity pairs: Specific miRNA-Disease (SpMiR-D), Specific miRNA-Gene/Protein (SpMiR-GP),

Non-Specific miRNA-Disease (NonSpMiR-D), and Non-Specific miRNA-Gene/Protein (NonSpMiR-GP). Relevant triples, an interacting pair (from one of the above-mentioned) co-occurring with a Relation Trigger in a sentence are defined to form a relation and can belong to one of the four above-mentioned Relation classes. On the contrary, if an interacting pair does not co-occur with any Relation Trigger then we do not tag such pair as a relation.

The annotation has been performed using Knowtator[26] integrated within the Protégé framework[27].

Corpus selection, annotation and properties. We develop a new corpus based on MEDLINE, annotated with miRNA mentions and relations. Shah *et al.*[28] showed that abstracts provide a comprehensive description of key results obtained from a study, whereas full text is a better source for biological relevant data. Thus, we choose to build the corpus for abstracts only. Out of 27001 abstracts retrieved using the keyword "miRNA", 201 were randomly selected as training and 100 as test corpus. Two annotators performed the annotation. The first annotator annotated the training corpus iteratively to develop guidelines and built the consensus annotation. The second annotator followed these guidelines and annotated the same corpus. Disagreeing instances were harmonized by both the annotators through manual inspection for correctness and its adherence to the guidelines. Any changes to the guidelines were made if needed. During the harmonization process only the non-overlapping instances between the two annotators were investigated. Decisions were based on the rule that only noun forms were to be marked (specific adjectives that can be treated as nouns were also considered). In case of partial matches, where conflicting parts could be interpreted as an adjective were not resolved. For example, in "chronic inflammation", marking either "chronic inflammation" or just "inflammation" were considered correct. Table 1 provides the inter-annotator agreement (measured as F_1, for both exact and boundary match, and Cohen's κ) for the test corpus. Exact string match occurs only when both the annotators annotate identical strings, whereas in partial match fraction of the string has been annotated by either of the annotators. It is evident (*cf.* Table 1) that in almost all cases partial match performs better than exact string match, indicating variations in span of mentioned entities. An example annotation is shown in Box 1.

Table 2 shows the number of annotated concepts in the training and test corpora for each entity class and the count for manually extracted relations (triplets), categorized for different interacting entity pairs. Table 3 provides the overall statistics of the published corpora (additional information about the corpus is given in the README supplementary file).

Table 1. Inter-annotator agreement scores for the test corpus.

Annotation Class	F_1 (Exact Match)	F_1 (Partial Match)	κ
Non-specific MiRNAs	0.9985	0.9985	0.996
Specific MiRNAs	0.9545	0.9779	0.916
Genes/Proteins	0.8343	0.8705	0.752
Diseases	0.8270	0.9575	0.853
Species	0.9329	0.9437	0.875
Relation Triggers	0.8441	0.9543	0.798

Automated named entity recognition

For identification of specific miRNA mentions in text (*cf.* Table 4), we developed regular expression patterns using manual annotations of miRNA mentions as the basis. Similarly, a dictionary has been generated for general miRNA recognition. The regular expression patterns are represented in the format as defined by Oualline *et al.*[29]. For simplicity and reusability, several aliases are defined (*cf.* Table 5) to be used in the final regular expression patterns for specific miRNA identification, given in Table 4. Detected entities are resolved to a unique miRNA name and disambiguated to adhere to standard naming conventions as authors use several morphological variants to report the same miRNA term. For example, miR-107 can be represented as miRNA-107, Micro RNA-107, MicroRNA 107, has-mir-107, mir-107/108, micro RNA 107 and 108, micro RNA (miR) 107 and so on. Thus, the identified miRNA entity has been resolved to its base form (*e. g. hsa-microRNA-21* to *hsa-mir-21* and *microRNA 101* to *mir-101*) following the miRBase naming convention. Manual inspection of the test corpus for species distribution revealed that 71% of the documents belonged to human, followed by mouse (15%), rat (8%). Pig has 2 abstracts, zebrafish, HIV-1, HSV-1, and *Caenorhabditis elegans* 1 each (*cf.* Supplementary Figure A for the distribution). Thus, we assumed that most of the abstracts belonged to human and resolved the identified miRNA entities to human identifier in miRBase. Unique miRNA terms are mapped to human *miRBase* database identifiers through the mirMaid Restful web service. For those names where we do not retrieve any database identifiers, we fall back to another organism mention found in the abstract (if any), using the NCBI taxonomy dictionary (see below)

Table 2. Manually annotated entities statistics. Counts of manually annotated entities in the training and the test corpora as well as annotated sentences describing relations.

Annotation Class	Corpus	
	Training	Test
Non-specific MiRNAs	1170	336
Specific MiRNAs	529	376
Genes/Proteins	734	324
Diseases	1522	640
Species	546	182
Relation Triggers	1335	625
SpMiR-D	171	127
SpMiR-GP	195	123
NonSpMiR-D	124	54
NonSpMiR-GP	77	16

Table 3. Statistics of the published miRNA corpora.

Occurrences in the corpus	Training	Test
Sentences	1864	780
Entities	5836	2483
Entity pairs	2001	868
Positive entity pairs	567	320
Negative entity pairs	1434	548

Table 4. Regular expression patterns used for miRNAs identification. Aliases used to form the final regular expression, see Table 5, are highlighted in bold.

Regular expression patterns	Description	Example of identified text
(Pref+(Lin,Let))	Detection of *Lin* and *Let* variations of miRNAs	lin-4; hsa-let-7a-1
(Pref+(miRNA, Onco)(S*Tail)(Sep Tail)*)	MiRNAs mentions for different separators	hsa-mir-21/22; Oncomir-17~92
(Pref+(miRNA, Onco) S*(D(Z([/]Z)*)+) ([\,] S*? (Pref+(miRNA, Onco) S*(D(Z([/]Z)*)+)*)))	Multiple miRNA mentions occurring progressively	miR-17b, -1a; hsa-miR-21,22, and hsa-miR-17

Table 5. MiRNAs regex aliases. Aliases used in regular expression patterns for miRNAs identification (highlighted in bold).

Description	Alias	Regular Expression Pattern
Digit sequences	**D**	(\d?\d*)
Admissible hypens with a trailing space	**Z**	([\-]?[\-]*)
Admissible hypens with a leading space	**S**	([\-]?[\-]*)
3-letter prefix for human followed by a hyphen	**Pref**	([hH][sS][aA][\-])
Non-specific miRNA mentions	**miRNA**	([mM][iI]([cC][rR][oO])+[rR]([nN][aA]s+)+)
Let-7 miRNA mention	**Let**	([lL][eE][tT]**S***[7]?\l+)
Lin-4 miRNA mention	**Lin**	([lL][iI][nN]**S***[4]?\l+)
Oncomir miRNA mention	**Onco**	([oO][nN][cC][oO][mM][iI][rR])
Admissible tilde and word boundaries	**Cluster**	(~[\b]-[\b]-*)
Admissible hyphen and separator *and* and *comma*	**Sep**	(**S***((and?,**S**,\/,)? **S***)+)
Admissible combination of upper and lower case alphabets	**UL**	(?\l?\l+,?\u?\u+)
Admissible alpha-numerical identifiers in specific miRNA mentions	**AN**	(**UL**((/, *and*,**D**+)? **UL**)+)
Admissible alpha-numerical identifiers in oncomir mentions	**Tail**	(**D**(**AN Cluster**+,\-**D AN**+)+)

(*cf.* Supplementary Figure B), otherwise we retain the unique normalized name (*cf.* Box 2).

Box 2. Un-normalized and normalized entities that are mapped to miRBase identifiers. Here MIR0000007, MIR0000008, and MIR0000005 are internal identifiers used by ProMiner.

MIR0000007:MIMAT0015092@MIRBASE|MI0000002@MIRBASE|cel-lin-4|lin-4
MIR0000008: miR-171|microRNA 171
MIR0000005:MIMAT0000416@MIRBASE|has-miR-1|miRNA-1

We detect SPECIES with a dictionary-based approach. The built dictionary consists of all the concepts from the NCBI taxonomy corresponding to only those organisms mentioned in *miRBase*.

Similarly, for identification of DISEASE and GENE/PROTEIN mentions in text we adapted a dictionary-based approach. To detect DISEASE, we apply three dictionaries: MeSH, MedDRA[30] and *Allie*. For GENE/PROTEIN, a dictionary[31] based on SwissProt, EntrezGene, and HGNC is included. Gene synonyms which could be potentially tagged as miRNAs are removed to overcome redundancy. For example, genes encoding microRNA, *hsa-mir-21* are named as *miR-21*, *miRNA21* and *hsa-mir-21*, the gene symbol of *MIR16 membrane interacting protein of RGS16* is *MIR16*, which can represent a miRNA mention.

The RELATION TRIGGER dictionary comprises of all interaction terms from the training corpus. After reviewing the training corpus for relation trigger terms, we retrieved not one but many variants of the same RELATION TRIGGER occurring in alternative verb-phrase groups. For example, "change in expression" can be represented as one of the following verb-phrases: Change MicroRNA-21 Expression, Expression of caveolin-1 was changed, Change in high levels of high-mobility group A2 expression, change of the let-7e and miR-23a/b expression, expression of miR-199b-5p in the non-metastatic cases was significantly changed, etc. To allow flexibility for capturing RELATION TRIGGER along with its variants spanning over different phrase length, we first manually represented all the relations in its root form, such as "regulate expression" to "regulate" (*cf.* Relation_Dictionary.txt file in Dataset 1). The base form has been extended manually to different spelling variants, *e.g.* regulate to regulatory, regulation, etc., the detailed listing of variants is provided in Word_variations.txt in Dataset 1. Not all combinations of

the root forms are logical; target and up-regulation terms cannot be combined to form a relation trigger. Thus, we additionally defined a set of relation combinations that are allowed (see Permutation_terms.txt in Dataset 1 for all combinations).

For all named entity recognition performed, the dictionary-based system ProMiner[31] is used. Supplementary Table A (Dataset 1) provides a quantitative estimate of the entities available in the dictionaries used in this work.

Relation extraction

We consider three approaches for addressing automatic extraction of interacting entity pairs from free text, described in the following.

The co-occurrence approach serves as a baseline. Assuming all interactions to be present in isolated sentences, this approach is complete but may be limited in precision. Reducing the number of false positives can be achieved by filtering with the dictionary of relation triggers occurring in the same sentence. The rationale behind this filter is that the interaction is more likely to be described if such a term is present (we refer to this as tri-occurrence).

To increase the precision, we use a machine learning-based approach formulating the relation detection as a binary classification problem: each instance (consisting of a pair of entities) is classified either as not-containing a relation or belonging to one of the four-relation classes. Our system uses lexical and dependency parsing features. We evaluate linear support vector machines (SVM)[32] as implemented in the LibSVM library, as well as LibLINEAR, a specialized implementation for processing large data sets[33], and naive Bayes classifiers[34]. For more details, we refer to Bobić *et al.*[35].

Lexical features capture characteristics of tokens around the inspected pair of entities. The sentence text can roughly be divided into three parts: text between the entities, text before the entities, and text after the entities. Stemming[36] and entity blinding is performed to improve generalization. Features are bag-of-words and bi, tri, and quadri-gram based. This feature setting follows Yu *et al.* and Yang *et al.*[37,38]. The presence of relation triggers is also taken into account, using the previously described manually generated list. Next to lexical features, dependency parsing (created using Stanford parser) provides an insight into the entire grammatical structure of the sentence[39] and was performed using the Stanford CoreNLP library (http://nlp.stanford.edu/software/corenlp.shtml). Deep parsing follows the shortest dependency path hypothesis[40]. We analyzed the vertices *v* (tokens from the sentence) in the dependency tree from a lexical (text of the token) and syntactical (POS tag) perspective. Edges *e* in the tree correspond to the information about the grammatical relations between the vertices. Extracting relevant information from the dependency parse tree is usually done following the shortest dependency path hypothesis[40]. Lexical and syntactical *e-walks* and *v-walks* on the shortest path are created by alternating sequence of vertices and edges, with the length of 3. We capture the information about the common ancestor vertex, in addition to checking whether the ancestor node represents a verb form (*e.g.* POS tag could be VB, VBZ, VBD, etc.). Finally, the length of the shortest path (number of edges) between the entities is considered as a numerical feature.

Results and discussion

Dataset 1. Version 2. Manually annotated miRNA-disease and miRNA-gene interaction corpora

http://dx.doi.org/10.5256/f1000research.4591.d40643

Please see README.txt in the zip file for precise details about the corpus and supplementary files. The updated zip file contains new files (Permutation_terms.txt, Non-Specific_miRNAs_Dictionary.txt and Word_variations.txt) and Table A has been updated.

In the following, we present results for named entity recognition and relation extraction. This section concludes with two use-case analyses.

Performance evaluation of named entity recognition

Among the 201 abstracts present in the training corpus, 82% contained general miRNA mentions, in comparison to specific miRNAs with 45%. In Table 6, results for miRNA entity recognition are reported. Non-specific miRNA recognition is close to perfect. Specific miRNA mention recognition has an F_1 measure of 0.94.

For disease mention recognition, combined dictionaries, based on three established resources, resulted in 0.79 and 0.69 F_1 score for the training and test corpus respectively. The low score for disease identification could be due to the variation in disease mentions, such as multi-word, synonym combination, nested names, etc. However, the partial matches result for diseases reported 0.88 of F_1, providing the possibility for detection of similar text strings for better recall (*cf.* Supplementary Table B in Dataset 1). Genes/proteins dictionary showed a performance of 0.84 and 0.85 of F_1 in training and test corpus respectively.

The evaluation of the relation trigger dictionary (*cf.* Table 6) suggests that it covers a substantial part of the vocabulary with recall of 0.86 for the training and 0.79 for the test corpus.

Relation extraction

We queried MEDLINE for "miRNA and Epilepsy" documents, among which 16 documents containing miRNA-related relations were manually selected (*cf.* Supplementary Figure C for the detailed distribution statistics). To avoid any biased approach we choose Epilepsy disease domain. Manual inspection of these articles revealed

Table 6. Evaluation results for miRNA entity classes. Here only complete match results are presented. The performance of named entity recognition is evaluated using recall (*R*), precision (*P*) and F_1 score.

Entity Class	*R*	*P*	F_1	*R*	*P*	F_1
	Training Corpus			Test Corpus		
Non-specific MiRNAs	1.000	0.995	0.997	1.000	0.997	0.999
Specific MiRNAs	0.921	0.928	0.924	0.936	0.934	0.935
Relation Triggers	0.864	0.885	0.874	0.790	0.842	0.815

11.5% of miRNA-related associations occur outside the sentence level. Thus, our work focused on relations at sentence level. Sentences in which co-occurring entity pairs do not participate in any relation are tagged as *false*. A comparison of the different relation extraction approaches is shown in Figure 2. Supplementary Table D in Dataset 1 provides statistical details of the applied approaches given in Figure 2. If all the entities are correctly identified then co-occurrence based approach leads to 100% recall for relation extraction. The recall is not diminished using the tri-occurrence approach, as the true entity pairs remain constant, while the precision increases between 4pp (percentage points) and 17pp when compared to the co-occurrence based approach, reducing false positives (*cf.* Figure 2). However, overall the precision reaches less than 60%. In our work, we assume that all the entities have been identified giving a recall of 100% for both co-occurrence and tri-occurrence based approaches. Using the machine-learning based classification, precision is increased up to 76% for specific miRNA-gene relations for both LibLINEAR and LibSVM methods, although Naïve Bayes is not far behind. Similarly, these two methods performed nearly the same for specific miRNAs-disease relations, the F_1 measure is not substantially different but a trade-off between precision and recall can be observed. An increase in F_1 measure is observed for non-specific miRNA relations when Naïve Bayes method is applied, out performing other strategies. Nevertheless, preference of the method highly depends on the compromise one chooses, whether better recall or precision. Overall, better recall and acceptable precision can be achieved with tri-occurrence method.

Most relation extraction approaches are dependent on the performance of named entity recognition. The impact of error propagation coming from automated entity recognizers is evaluated by applying the tri-occurrence method on the automatically annotated training and test corpus, here termed as "NERTri". Compared to the results on the gold standard entity annotation a drop of 13 pp for NonSpMiR-D, 7pp for NonSpMiR-GP, 22pp for SpMiR-D, and 30pp for SpMiR-GP in F_1 is observed for the test corpus. Overall performance of the NERTri approach on training and test corpus is detailed in Supplementary Table C in Dataset 1.

Use case analysis

For the impact analysis of the proposed approach, we compare the extracted information with two databases, namely *miR2Disease* and *miRSel*. We focus on relations and articles concerning Alzheimer's disease.

Alzheimer's disease (AD) is ranked sixth for causing deaths in major developed countries[41]. It affects not only individuals but also incurs a high cost to the society. Recently, miRNAs have shown close associations with AD pathophysiology[42,43]. Increasing the need to identify new therapeutic targets for AD, after major set backs due to failed drugs, motivates the need to look in this direction. *In silico* methods, such as the one proposed in this work, can aid in building miRNA-regulatory networks specific to AD, for further analysis such as identifying the mechanisms, sub-networks, and key targets.

Extracting miRNA-Alzheimer's disease relations from full MEDLINE

The database *miR2Disease* is queried to return all miRNA-disease relations occurring in Alzheimer's disease. For comparison, we retrieved miRNA-disease relations from MEDLINE using NERTri approach, resulting in 41 abstracts containing 159 relations. Obtained triplets have been manually curated to remove 51 false positives. False negatives have not been accounted, which may result in loss of information (*cf.* Relation extraction section). Comparison between the relations obtained from *miR2Disease* and NERTri are summarized in Table 7. The *miR2Disease* database returns 28 evidential statements from 9 articles. Among these, only 14 evidences are present in abstracts. Moreover, 16 evidences are extracted from one full text document[44]. Only two evidences are identified at abstract level among these 16 evidences. Overall, 26 miRNAs identified by *miR2Disease* refer to Alzheimer's disease.

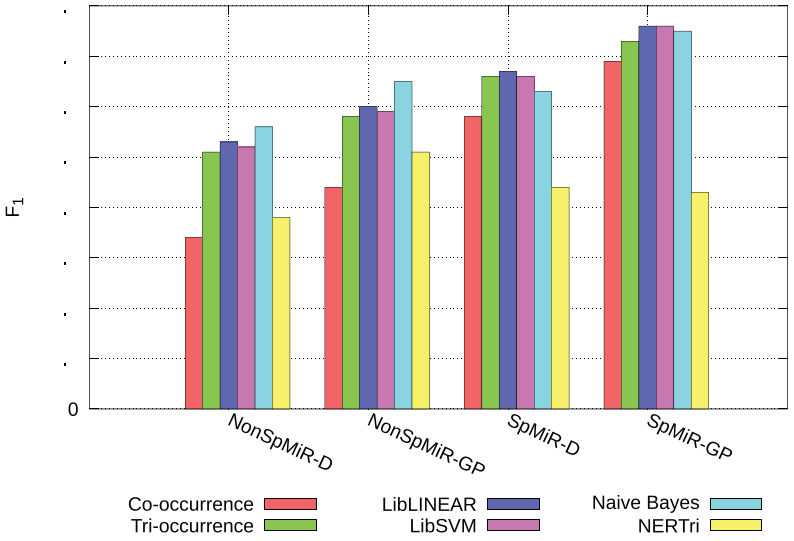


Figure 2. Comparison of different relation extraction approaches. On the x-axis, different entity pair relations are represented as SpMiR-D for Specific miRNA-Disease, SpMiR-GP for Specific miRNA-Gene/Protein, NonSpMiR-D for Non-Specific miRNA-Disease, and NonSpMiR-GP for Non-Specific miRNA-Gene/Protein.

Table 7. miR2Disease database comparison. MiRNA-Alzheimer's disease relation retrieved from MEDLINE and in *miR2Disease* database.

	miR2Disease	NERTri	True Positives in NERTri	NERTri and miR2Disease Overlap
Publications	9	41	36	8
Relations	28	159	108	11
Evidences (abstracts)	14	159	108	10
Unique miRNAs	26	46	40	16

Therefore, our text-based extraction proposes approximately three times more relations than the database provides.

The analysis of 17 false negative relations which are in the database but not found by our approach shows that most of the relations could be found only in full text and that the automatic system misses four miRNA-Alzheimer's disease relations from abstracts. Manual inspection reveals that in three out of these missing four evidences the disease name is not mentioned in the sentence (relation occurred at co-reference level).

Extraction of miRNA-gene relations for Alzheimer's diseases from full MEDLINE

Here we compare the performance of our relation detection NERTri with another text-mining database, *miRSel*. For comparison, 100 abstracts from PubMed were retrieved using the query *"alzheimer disease"[MeSHTerms] OR ("alzheimer disease"[All Fields] OR "alzheimer"[All Fields]) AND ("micrornas"[MeSH Terms] OR "micrornas"[All Fields] OR "microrna"[All Fields]) AND ("2001/01/01"[PDAT]:"2013/7/4"[PDAT])*. Manual inspection of these articles leads to 184 miRNA-gene relations, at sentence level, (Table 8) in 37 abstracts.

NERTri approach was able to identify 140 of these found relations in 28 abstracts. Among the 37 abstracts from the PubMed query, *miRSel* contained only 12 abstracts with 56 miRNA-gene relations (cf. Table 8). False negatives in our approach when compared with *miRSel* could not be directly identified as the database is not downloadable and searchable for disease specific relations. However, low intersection between *miRSel* and NERTri can be observed.

Table 8. miRSel database comparison. Comparison of miRNA-gene relations retrieval for Alzheimer's disease in MEDLINE.

Approach	Articles	Relations
PubMed Query ("Alzheimer AND miRNA")	100	NA
PubMed Query with relations at sentence level	37	184
PubMed Query ∩ NERTri	28	140
PubMed Query ∩ miRSel	12	56
NERTri ∩ miRSel	14	22

In summary, our approach provides AD related gene-microRNA relations from PubMed which have not been available in the database before.

Overall, the results are promising when compared with the *miR2Disease* and *miRSel* databases and indicate that we can extend the databases to a large extent with new relations. Such an approach makes it much easier to keep databases up to date. Nevertheless full text processing would most certainly increase the recall of automatic processing.

Conclusion and future work

In this work, we proposed approaches for identification of relations between miRNAs and other named entities such as diseases, and genes/proteins from biomedical literature. In addition, details of named entity recognition for all the above entity classes have been described. We distinguished two types of miRNA mentions, namely Specific (with numerical identifiers) and Non-Specific (without numerical identifiers). Non-specific miRNAs entity recognition has enabled us to achieve better recall and precision in document retrieval. Three different relation extraction approaches are compared, showing that the tri-occurrence based approach should be the first reliable choice among all others. The tri-occurrence based approach is comparable to a machine learning-based method but considerably faster. In comparison to two well-established databases, we have shown that additional useful information can be extracted from MEDLINE using our proposed methods.

To best of our knowledge, this is the first work where manually annotated corpora containing information about miRNAs and miRNA-relations are published. Moreover, the corpora and methods provided represent useful basis and tools for extracting the information about miRNAs-associations from literature. This work serves as an important benchmark for current and future approaches in automatic identification of miRNA relations. It provides the basis for building a knowledge-based approach to model regulatory networks for identification of deregulated miRNAs and genes/proteins.

The proposed methods encourage the extension of this work to full-text articles, to elucidate many more relations from Biomedical literature. Non-specific miRNA mention identification could prove highly beneficial for co-reference resolution in full-text articles, in addition to abstracts. Proposed machine-learning approaches could be applied to only tri-occurrence based instances for reducing the false positive rates. Extending the current approach to other model organisms such as mouse, and rat could be helpful in revealing

important relations for translational research. Inclusion of additional named entities such as drugs, pathways, etc. could lead to an interesting approach for detection of putative therapeutic or diagnostic drug targets through a gene-regulatory network generated from identified relations.

Data availability

Corpora availability: http://www.scai.fraunhofer.de/mirnacorpora.html

Archived corpora at time of publication: F1000Research: Dataset 1. Version 2. Manually annotated miRNA-disease and miRNA-gene interaction corpora, 10.5256/f1000research.4591.d40643[45]

Author contributions

SB, RK, JF, and MHA conceived and designed the overall research strategy. SB carried out all the development work and performed the analysis. She is the major contributor of manuscript preparation and principal annotator. TB developed the machine learning-based workflow for relation extraction, transformed corpora into the standard format, and contributed to manuscript writing. JF supported in use-case analysis and paper writing. MHA is the scientific supervisor for this work. RK contributed to critical discussions, analysed the results and a major contributor in correcting and writing manuscript. All authors read and approved the final version of the manuscript.

Competing interests

No competing interests were disclosed.

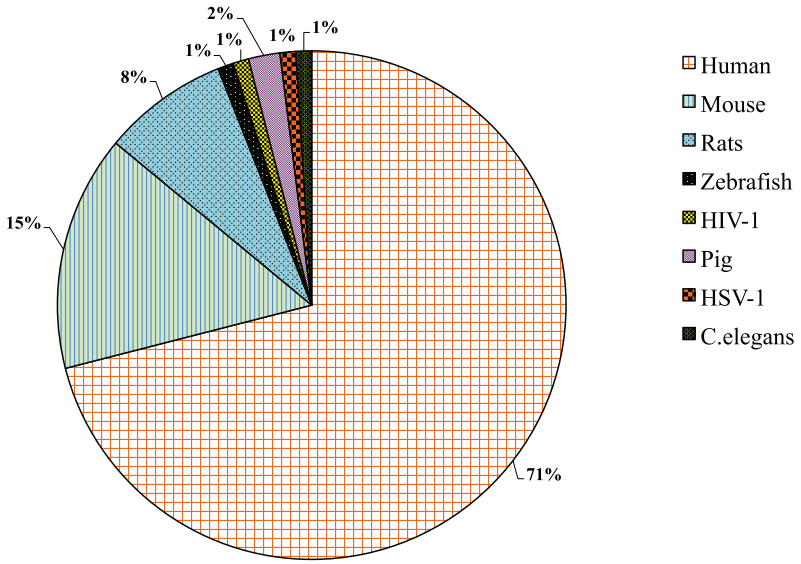
Grant information

Shweta Bagewadi was supported by University of Bonn. Tamara Bobić was partially funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School during her contribution to this work at Fraunhofer SCAI.

Acknowledgements

We would like to thank Heinz-Theo Mevissen for all the support during implementation of the dictionaries and regular expressions in ProMiner. We acknowledge Anandhi Iyappan for her contribution as the second annotator. We are also grateful to Harsha Gurulingappa for all his support and fruitful discussions during this work. We would like to thank Ashutosh Malhotra for proof reading the manuscript.

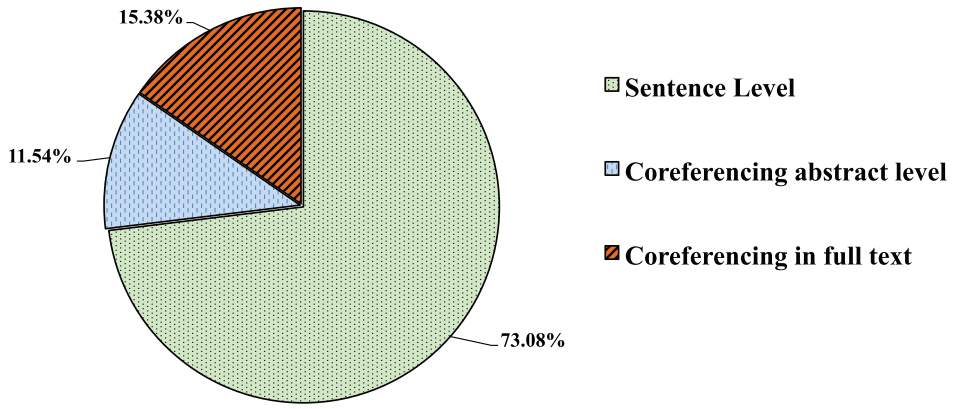
Supplementary figures



Supplementary Figure A. Distribution of organism mentions in training corpus.

Title: Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity.
Sentence : Interestingly, mir-125 is a putative homologue of lin-4.
Normalized miRNA name: dme-mir-125

Supplementary Figure B. A screenshot example of how we handle other organism miRNA normalization. There is no miR-125 entry related to human in miRBASE. Since the abstract mentions *Drosophila melanogaster* in the title, the miRNA is normalized to dme-mir-125.



Supplementary Figure C. Coverage of relations occurring in Epilepsy Documents.

References

1. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell.* 1993; **75**(5): 843–54.
PubMed Abstract | Publisher Full Text
2. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell.* 2004; **116**(2): 281–297.
PubMed Abstract | Publisher Full Text
3. Esquela-Kerscher A, Slack FJ: **Oncomirs microRNAs with a role in cancer.** *Nat Rev Cancer.* 2006; **6**(4): 259–69.
PubMed Abstract | Publisher Full Text
4. Ma W, Hu S, Yao G, *et al.*: **An androgen receptor-microrna-29a regulatory circuitry in mouse epididymis.** *J Biol Chem.* 2013; **288**(41): 29369–81.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Babak T, Zhang W, Morris Q, *et al.*: **Probing microRNAs with microarrays: tissue specificity and functional inference.** *RNA.* 2004; **10**(11): 1813–1819.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Bottoni A, Zatelli MC, Ferracin M, *et al.*: **Identification of differentially expressed microRNAs by microarray: a possible role for microRNA genes in pituitary adenomas.** *J Cell Physiol.* 2007; **210**(2): 370–377.
PubMed Abstract | Publisher Full Text
7. Wu X, Song Y: **Preferential regulation of miRNA targets by environmental chemicals in the human genome.** *BMC Genomics.* 2011; **12**(1): 244.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Calin GA, Dumitru CD, Shimizu M, *et al.*: **Frequent deletions and downregulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.** *Proc Natl Acad Sci U S A.* 2002; **99**(24): 15524–9.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Banno K, Yanokura M, Iida M, *et al.*: **Application of microRNA in diagnosis and treatment of ovarian cancer.** *BioMed Res Int.* 2014; **2014**: 232817.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell.* 2009; **136**(2): 215–33.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Vergoulis T, Vlachos IS, Alexiou P, *et al.*: **TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support.** *Nucleic Acids Res.* 2011; **40**(Database issue): D222–229.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Naeem H, Küffner R, Csaba G, *et al.*: **miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature.** *BMC Bioinformatics.* 2010; **11**: 135.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Jiang Q, Wang Y, Hao Y, *et al.*: **miR2Disease: a manually curated database for microRNA deregulation in human disease.** *Nucleic acids Res.* 2009; **37**(Database issue): D98–104.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Ruepp A, Kowarsch A, Schmidl D, *et al.*: **PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes.** *Genome Biol.* 2010; **11**(1): R6.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Czarnecki J, Nobeli I, Smith A, *et al.*: **A text-mining system for extracting metabolic reactions from full-text articles.** *BMC Bioinformatics.* 2012; **13**(1): 172.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Hsu SD, Lin FM, Wu WY, *et al.*: **miRTarBase: a database curates experimentally validated microRNA-target interactions.** *Nucleic acids Res.* 2011; **39**(Database issue): D163–9.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Xie B, Ding Q, Han H, *et al.*: **miRCancer: a microRNA-cancer association database constructed by text mining on literature.** *Bioinformatics.* 2013; **29**(5): 639–44.
PubMed Abstract | Publisher Full Text
18. Smith L, Tanabe LK, nee Ando RJ, *et al.*: **Overview of BioCreative II gene mention recognition.** *Genome Biol.* 2008; **9**(Suppl 2): S2.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Arighi CN, Lu Z, Krallinger M, *et al.*: **Overview of the BioCreative III Workshop.** *BMC Bioinformatics.* 2011; **12**(Suppl 8): S1.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Nedellec C, Bossy R, Kim JD, *et al.*: **Proceedings of the BioNLP Shared Task 2013 Workshop**. Association for Computational Linguistics, Sofia, Bulgaria, 2013.
Reference Source
21. Tsujii J, Kim JD, Pyysalo S: **Proceedings of BioNLP Shared Task 2011 Workshop**. Association for Computational Linguistics, Portland, Oregon, USA, 2011.
Reference Source
22. Tsujii J: **Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task**. Association for Computational Linguistics, Boulder, Colorado, 2009.
Reference Source
23. Murray BS, Choe SE, Woods M, *et al.*: **An *in silico* analysis of microRNAs: mining the miRNAome.** *Mol Biosyst.* 2010; **6**(10): 1853–62.
PubMed Abstract | Publisher Full Text
24. Dweep H, Sticht C, Pandey P, *et al.*: **miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes.** *J Biomed Inform.* 2011; **44**(5): 839–47.
PubMed Abstract | Publisher Full Text
25. Pyysalo S, Airola A, Heimonen J, *et al.*: **Comparative analysis of five protein-protein interaction corpora.** *BMC Bioinformatics.* 2008; **9**(Suppl 3): S6.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Ogren PV: **Knowtator: A Protégé plug-in for annotated corpus construction**. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations.* New York, Association for Computational Linguistics. 2006; 273–275.
Publisher Full Text
27. Gennari JH, Musen MA, Fergerson RW, *et al.*: **The evolution of Protégé: an environment for knowledge-based systems development.** *Int J Hum Comput Stud.* 2003; **58**(1): 89–123.
Publisher Full Text
28. Shah PK, Perez-Iratxeta C, Bork P, *et al.*: **Information extraction from full text scientific articles: where are the keywords?** *BMC Bioinformatics.* **4**: 20.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Oualline S: **Vi iMproved**. New Riders Publishing, Thousand Oaks, CA, USA, 2001.
Reference Source
30. Brown EG, Wood L, Wood S: **The medical dictionary for regulatory activities**

(MedDRA). *Drug Saf.* 1999; **20**(2): 109–17.
PubMed Abstract | Publisher Full Text

31. Fluck J, Mevissen HT, Oster M, *et al.*: **ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries**. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, Spain. 2007; 149–151.
Reference Source

32. Cortes C, Vapnik V: **Support-vector networks**. In *Machine Learning*, 1995; **20**(3): 273–297.
Publisher Full Text

33. Fan E, Chang K, Hsieh C, *et al.*: **LIBLINEAR: A Library for Large Linear Classification.** *Machine Learning Research.* 2008; **9**: 1871–1874.
Reference Source

34. John GH, Langley P: **Estimating continuous distributions in Bayesian classifiers**. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, UAI'95, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. 1995; 338–345.
Reference Source

35. Bobić T, Klinger R, Thomas P, *et al.*: **Improving distantly supervised extraction of drug-drug and protein-protein interactions**. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, Avignon, France, Association for Computational Linguistics. 2012; 35–43.
Reference Source

36. Porter M: **An algorithm for suffix stripping.** *Program.* 1980; **14**(3): 130–137.
Publisher Full Text

37. Yu H, Qian L, Zhou G, *et al.*: **Extracting protein-protein interaction from biomedical text using additional shallow parsing information**. In *Biomedical Engineering and Informatics, 2009. BMEI '09. 2nd International Conference on*, 2009; 1–5.
Publisher Full Text

38. Yang Z, Lin H, Li Y: **BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using svm and rich feature sets.** *J Biomed Inform.* 2010; **43**(1): 88–96.
PubMed Abstract | Publisher Full Text

39. De Marneffe MC, Manning CD: **Stanford typed dependencies manual**. 2010.
Reference Source

40. Bunescu RC, Mooney RJ: **A shortest path dependency kernel for relation extraction**. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics. HLT '05, Stroudsburg, PA, USA. 2005; 724–731.
Publisher Full Text

41. Thies W, Bleiler L, Alzheimer's Association: **2011 Alzheimer's disease facts and figures.** *Alzheimers Dement.* 2011; **7**(2): 208–244.
PubMed Abstract | Publisher Full Text

42. Cheng L, Quek C, Sun X, *et al.*: **Deep-sequencing of microRNA associated with Alzheimer's disease in biological fluids: From biomarker discovery to diagnostic practice.** *Frontiers in Genetics.* 2013; **4**(150).

43. Wang WX, Rajeev BW, Stromberg AJ, *et al.*: **The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1.** *J Neurosci.* 2008; **28**(5): 1213–23.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Hébert SS, Horré K, Nicolaï L, *et al.*: **Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's disease correlates with increased BACE1/beta-secretase expression.** *Proc Nat Acad Sci U S A.* 2008; **105**(17): 6415–6420.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Bagewadi S, Bobi T, Hofmann-Apitius M, *et al.*: **Dataset, 1 version 2 in: Detecting miRNA Mentions and Relations in Biomedical Literature.** *F1000Research.*
Data Source

Open Peer Review

Current Peer Review Status: ✔ ? ✔

Version 2

Reviewer Report 22 July 2015

https://doi.org/10.5256/f1000research.6352.r5979

✔ **Robert Leaman**
National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD, USA

In this manuscript the authors motivate and describe the creation of a corpus of miRNA mentions. As noted by the authors, miRNA has important biological functions and is not yet well studied from a text mining standpoint. The corpus described represents a substantial undertaking. The authors demonstrate that the corpus is sufficient to create a named entity recognizer with good performance, and also that a relation extraction technique based on the method finds relations not present in existing databases. Overall, the authors have been quite transparent and thorough in documenting their approach.

Two of the largest concerns with any annotation effort are whether the corpus is representative and whether it is consistent. The authors describe their method of corpus selection, which is straightforward and reasonable. The authors have used best practices for annotation - their guidelines are provided, the corpus is annotated by two annotators, and the inter-annotator agreement is high. This shows in their use of the corpus to train an NER system. It would be good to mention how annotator disagreements were harmonized and whether any effort was made to make the annotation consistent across the corpus. For example, the word "chronic" is included in the annotation "chronic inflammation" but not included in "chronic neurological disorder" - there is probably a good reason for this, but the process of ensuring this sort of consistency is not mentioned.

Some of the wording in the manuscript seems to communicate a different annotation result than what is seen in the corpus. For example, page 4 asserts "Possessive terms such as 'Diabetic patients' are not marked." It seems likely that not annotating disease mentions used as a modifier would cause a consistency issue that will be a problem for training an NER system, but the corpus itself seems to have these terms annotated and I was unable to manually locate a phrase where the annotation was not provided. For example, document 21295623 mentions "Parkinson's disease subjects," and "Parkinson's disease" is annotated as a disease. Document 19703993 contains another example with "human meningioma samples." Is the rule that adjective forms of diseases (diabetic, hypertensive, malarial) are not annotated but nouns (diabetes, hypertension, malaria) are? In other words "diabetic patient" is indeed not annotated (the corpus does not

contain the word "diabetic") but the phrase "diabetes patient" would contain a disease annotation? I would like to reiterate that the actual annotations in the corpus look good, the only problem is this statement about possessive terms.

It should be noted that not annotating species names which are not present in a database could also lead to inconsistency. Since the species distribution is so skewed, however, this does not appear to be a concern to me.

How was sentence breaking performed? Some documents (e.g. 19703993) seem to lack sentence breaks where they would be expected.

Please provide a citation for the assertion that miRNA is associated with AD pathophysiology (Page 8).

Minor comments:

P.5: "201 are randomly selected as test corpus" --> "201 were randomly selected as the test corpus"
P.5: "Two annotators have been involved in the annotation." --> "Two annotators performed the annotation."
P.6: "Not all combination of the root forms is logical" --> "Not all combinations of the root forms are logical"
P.7: "However, partial matches result for the same" --> "However, the partial matches result for the same" or "However, the partial matches result for diseases"
P.8: Using "evidences" as in "The miR2Disease database returns 28 evidences" seems odd, would "evidential statement" or "assertion" or something similar be acceptable?
P.9: "miR2IDsease are in relation with Alzheimer's disease" --> "miR2IDsease refer to Alzheimer's disease"
P.9: "The proposed methods encourage future work of implementing the same for full-text articles" --> "The proposed methods encourage the extension of this work to full-text articles"
P.9: "Non-specific miRNA mentions identification" --> "Non-specific miRNA mention identification"

As noted by the authors, the naming convention for miRNA was standardized soon after their discovery, and thus the performance of the NER methods is higher than for many other entity types (notably genes and proteins). It might be interesting to note that the performance was not perfect even with a well supported naming convention; this suggests that naming conventions are probably not sufficient for mining: NLP methods would remain important even with well supported naming conventions.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Sep 2015

Shweta Bagewadi, Fraunhofer-Gesellschaft, Germany

We thank the reviewer for his comments. All the comments have been addressed in the newer verion of the manuscript, details below:

1. **COMMENT:** Two of the largest concerns with any annotation effort are whether the corpus is representative and whether it is consistent. The authors describe their method of corpus selection, which is straightforward and reasonable. The authors have used best practices for annotation - their guidelines are provided, the corpus is annotated by two annotators, and the inter-annotator agreement is high. This shows in their use of the corpus to train an NER system. It would be good to mention how annotator disagreements were harmonized and whether any effort was made to make the annotation consistent across the corpus. For example, the word "chronic" is included in the annotation "chronic inflammation" but not included in "chronic neurological disorder" - there is probably a good reason for this, but the process of ensuring this sort of consistency is not mentioned.

- **RESPONSE:** Considering the reviewer's concern we have now added a few sentences in "Corpus selection, annotation and properties" section describing the harmonization of disagreeing annotations, as below:

Disagreeing instances were harmonized by both the annotators through manual inspection for correctness and its adherence to the guidelines. Any changes to the guidelines were made if needed. During the harmonization process only the non-overlapping instances between the two annotators were investigated. Decisions were based on the rule that only noun forms were to be marked (specific adjectives that can be treated as nouns were also considered). In case of partial matches, where conflicting parts could be interpreted as an adjective were not resolved. For example in "chronic inflammation", marking "chronic inflammation" or just "inflammation" were considered as correct.

2. COMMENT: Some of the wording in the manuscript seems to communicate a different annotation result than what is seen in the corpus. For example, page 4 asserts "Possessive terms such as 'Diabetic patients' are not marked." It seems likely that not annotating disease mentions used as a modifier would cause a consistency issue that will be a problem for training an NER system, but the corpus itself seems to have these terms annotated and I was unable to manually locate a phrase where the annotation was not provided. For example, document 21295623 mentions "Parkinson's disease subjects," and "Parkinson's disease" is annotated as a disease. Document 19703993 contains another example with "human meningioma samples." Is the rule that adjective forms of diseases (diabetic, hypertensive, malarial) are not annotated but nouns (diabetes, hypertension, malaria) are? In other words "diabetic patient" is indeed not annotated (the corpus does not contain the word "diabetic") but the phrase "diabetes patient" would contain a disease annotation? I would like to reiterate that the actual annotations in the corpus look good, the only problem is this statement about possessive terms.

- **RESPONSE:** We thank the reviewer for pointing out the ambiguous nature of the statement. To clarify this, we have modified the below text:

Only disease nouns were considered, adjective terms such as " Diabetic patients" are not marked; however, specific adjectives that can be treated as nouns were marked, e.g. "Parkinson's disease patients".

3. **COMMENT:** It should be noted that not annotating species names which are not present in a database could also lead to inconsistency. Since the species distribution is so skewed,

however, this does not appear to be a concern to me.

- **RESPONSE:** We agree with the reviewer that by annotating the species names we would be able to add organism context to the miRNA-gene or miRNA-disease relations. However, currently there are only 32 species in which miRNAs have been identified, for which miRBase database provides official identifier and symbols. Thus, we cannot normalize the miRNA mentions to additional species.

4. **COMMENT:** How was sentence breaking performed? Some documents (e.g. 19703993) seem to lack sentence breaks where they would be expected.

- **RESPONSE:** The sentences are split based on the boundary annotations as described by Tomanek *et al.* These rules are implemented in our in-house NER tool, ProMiner. However, we are aware that it may not function perfectly for all cases, 19703993 is one such case.

K. Tomanek, J. Wermter, and U. Hahn, “Sentence and token splitting based on conditional random fields,” in Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pp. 49–57, 2007.

5. **COMMENT:** Please provide a citation for the assertion that miRNA is associated with AD pathophysiology (Page 8).

- **RESPONSE:** As per reviewer’s suggestion, we have now provided citation to an article where miRNAs are involved in AD pathogenesis.

Minor comments:

6. COMMENTS:

P.5: "201 are randomly selected as test corpus" --> "201 were randomly selected as the test corpus"

P.5: "Two annotators have been involved in the annotation." --> "Two annotators performed the annotation."

P.6: "Not all combination of the root forms is logical" --> "Not all combinations of the root forms are logical"

P.7: "However, partial matches result for the same" --> "However, the partial matches result for the same" or "However, the partial matches result for diseases"

P.9: "miR2IDsease are in relation with Alzheimer's disease" --> "miR2IDsease refer to Alzheimer's disease"

P.9: "The proposed methods encourage future work of implementing the same for full-text articles" --> "The proposed methods encourage the extension of this work to full-text articles"

P.9: "Non-specific miRNA mentions identification" --> "Non-specific miRNA mention identification"

RESPONSE: All the above comments have been addressed as per reviewer's suggestion.

7. COMMENT: P.8: Using "evidences" as in "The miR2Disease database returns 28 evidences" seems odd, would "evidential statement" or "assertion" or something similar be acceptable?

- **RESPONSE:** Considering the reviewer's suggestion we have replaced "evidences" to "evidential statements".

Competing Interests: None

Reviewer Report 31 December 2014

https://doi.org/10.5256/f1000research.6352.r7132

Sofie Van Landeghem
Bioinformatics and Evolutionary Genomics, Ghent University, Ghent, Belgium, Belgium

I thank the authors for the detailed response to my earlier comments. I am also happy to see more results in Table 1 and Figure 2, and to see that the manuscript has been updated with more background and details.

Considering my remark of the performance of the ML method with tri-occurrence-based candidate instances: the definition of candidate instances was unclear to me before. Now I understand that relations always need to have a trigger term. This comment is thus not relevant anymore.

Regarding my remark about "an automated entity recognizers", I merely meant to point out the typo "an". I noticed this was corrected in the new version. I apologize for the confusion. And even though I didn't mean to ask about a change of title, I think that adding the word "Automated" to the NER title is in fact better.

Remaining (small) comments

I am afraid that I am still a bit confused by the numbers in Table 8. Is it correct that it (still) reads 37 articles for the PubMed Query with relations at sentence level, and 39 articles with 184 relations for NERTri while the surrounding text now reads "Manual inspection leads to 184 miRNA-gene relations in 37 abstracts."?

In the author's response to construction of the training and test set, I think a typo may have crept into the numbers in this sentence, as currently it seems to state that you selected 301 abstracts

from an initial set of 300: "We first randomly retrieved 300 abstracts from the PubMed using the keyword “miRNA”. From these we manually selected 301 abstracts that contain gene/proteins or disease terms without looking in detail for any relation term."

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Sep 2015

Shweta Bagewadi, Fraunhofer-Gesellschaft, Germany

We thank the reviewer for her comments and providing clarifications of the previous comments. We have addressed these comments in the newer version of the manuscript.

Remaining (small) comments

1. COMMENT: I am afraid that I am still a bit confused by the numbers in Table 8. Is it correct that it (still) reads 37 articles for the PubMed Query with relations at sentence level, and 39 articles with 184 relations for NERTri while the surrounding text now reads "Manual inspection leads to 184 miRNA-gene relations in 37 abstracts."?

- **RESPONSE:** We understand the concern of the reviewer, here we meant that through manual inspection we detected 184 relations in 37 abstracts (retrieved using the PubMed Query). NERTri identified 2 additional abstracts (hence the total 39) that were false positives. To avoid any confusion for the reader, we have now simplified the text and deleted the NERTri statistics in the table.

2. COMMENT: In the author's response to construction of the training and test set, I think a typo may have crept into the numbers in this sentence, as currently it seems to state that you selected 301 abstracts from an initial set of 300: "We first randomly retrieved 300 abstracts from the PubMed using the keyword “miRNA”. From these we manually selected 301 abstracts that contain gene/proteins or disease terms without looking in detail for any relation term."

- **RESPONSE:** We apologize to the reviewer for the typo error. Here we meant to say that: We first randomly retrieved 27001 abstracts from the PubMed using the keyword “miRNA”. From these we manually selected 301 abstracts that contain gene/proteins or disease terms without looking in detail for any relation term.

Competing Interests: None

Version 1

Reviewer Report 17 October 2014

https://doi.org/10.5256/f1000research.4912.r5975

?

Filip Ginter
Department of Information Technology, University of Turku, Turku, Finland

This paper presents an annotated corpus of miRNA, gene/protein, disease, and species mentions, together with their miRNA-specific relations. Further, the authors implement a simple dictionary-based method for their extraction from text. This is a little studied, but highly relevant text mining target. Numerically, the results look rather promising. The paper is relatively easy to follow, but could be expanded somewhat to be more self-contained. More about that later.

The main problem I have when reading the paper is that it does not give me a good intuitive insight into how difficult this problem actually is. This relates to several individual passages in the text that left me wondering whether I understood correctly:

1. As for miRNA detection: On page 4, second paragraph, the authors mention that a prior study achieved 100% accuracy on miRNA detection task. Not being told more details, this either means the task is trivial, or that experiment is flawed.

2. As for relation detection: On page 7, first paragraph, the authors mention that moving from occurrence to tri-occurrence *"does not diminish recall"* which in that context is 100%. The way I understand this is that each and every positive sentence in the test data does have a relation trigger which is present also present in the training data. Is that really possible? That would seem to disagree with Table 6. How can you add a trigger filter but keep 100% recall?

I think the reader would benefit from more discussion on the variance of miRNA names and trigger expressions, to gain an intuitive grasp of the difficulty of the task. Which leads me to the regular expressions described in tables 4 and 5. Specifically, I do not understand the alias *Let* and *Lin.* Are these individual miRNAs? Why would you want to define a re-useable alias for individual miRNAs? Alternatively, I am misunderstanding something here, in which case I would appreciate clarification: what are these alias symbols, and how do I know they generalize? The impression I get from the regular expressions is that miRNA naming is highly regular and very simple. Is that the case really?

In the results and discussion section, I am perplexed by the 0.79 vs 0.69 F-score train/test difference for disease mentions. Do you have an insight as to why specifically disease mentions would have such a major difference when the other entities do not? This is especially puzzling since Table 2 shows disease as the largest class, i.e. it should exhibit least noise in the results.

Some other not so major points:

- I don't know what is meant by "improvised framework" in the abstract.
- I am not sure what is the status of relations that do not have a trigger. Or are there such?
- Where does the number 11.5% of cross-sentence relations come from? Counting in the corpus? (just checking)
- The previous point about 11.5% of cross-sentence relations makes me then wonder how can

the co-occurrence approach reach 100% recall? Have these cross-sentence relations been simply deleted from the data?
- Towards the end of the methods section, the paper describes the use of "deep parsing". Through the citation I'm guessing this relates to Stanford Dependencies. I think the paper really should give some detail about how this parsing was done.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Nov 2014

Roman Klinger, University of Stuttgart, Germany

We thank the referee for the detailed comments. We will address them and then submit a new version together with detailed comments.

Competing Interests: No competing interests were disclosed.

Author Response 07 Dec 2014

Shweta Bagewadi, Fraunhofer-Gesellschaft, Germany

COMMENT: "The main problem I have when reading the paper is that it does not give me a good intuitive insight into how difficult this problem actually is. This relates to several individual passages in the text that left me wondering whether I understood correctly:"
- RESPONSE: Considering reviewer's concern, we have now included the below text in Related Work sub-section in Introduction:

  Since the miRNA naming convention has been formalized very early in comparison to other biological entities such as genes and proteins, applying text-mining approaches is relatively simple[17]. Thus, most of the previously applied text mining approaches for miRNA detection has been based on keywords. miRCancer uses keywords to obtain abstracts from PubMed, further miRNA entities have been identified using regular expressions based on prefix and suffix variations. Similarly, miRWalk database uses keyword search approach to download abstracts and applies curated dictionary (compiled from six databases) for miRNA identification of human, rat, and mouse species[24]. TarBase, miR2Disease, miRTarBase, and several others have followed related search strategies. However, several authors still tend to use naming variations for acronyms, abbreviations, nested representations, etc. for listing miRNAs. Additionally, in contrast to the previous text-mining approaches focusing purely on miRNA gene relations, we extend the information extraction approach additionally to retrieve miRNA-disease relations.

COMMENT: "As for miRNA detection: On page 4, second paragraph, the authors mention that a prior study achieved 100% accuracy on miRNA detection task. Not being told more details, this either means the task is trivial, or that experiment is flawed."

- RESPONSE: As described by Murray *et al.*, they developed a regular expression for miRNA identification using the stems identified by term frequency analysis ("miR", "mirn", "mirna", and "microRNA") and later optimized it to attain 100% accuracy and recall against miRBase. However, the procedure is not fully clear to us. Below we provide the exact text as mentioned in Murray *et al.*

  "Identifying miRNAs Using the stems identified by term-frequency analysis ("miR", "mirn", ' 'mirna" and "microRNA"), we developed regular expression patterns to identify novel miRNA terms. Regular expressions were optimized to achieve 100% accuracy and recall against miRBase.[9–11] Regular expressions were designed to identify novel miRNas with, and without, species identifiers (e.g. hsa-miR-1 and mir-1). Regular expressions were used to identify novel miRNAs by mining the entire National Library of Medicine's PubMed abstract collection. All identified miRNas were curated into preferred terms to encompass synonymic variants; false positive hits were identified and filtered out during this process."

  The authors's do not however validate their regular expression against a larger corpus.

  Identification of miRNA mentions could be trivial if the variants are captured in a single regular expression. This issue we have addressed in our manuscript. The more challenging task is normalizing the miRNA mentions to correct database identifier.

  We have modified the pointed out text to the following:

  The authors claim to have optimized the approach to reach 100% accuracy and recall for detecting miRNAs mentions as in miRBase.

COMMENT: "As for relation detection: On page 7, first paragraph, the authors mention that moving from occurrence to tri-occurrence "does not diminish recall" which in that context is 100%. The way I understand this is that each and every positive sentence in the test data does have a relation trigger which is present also present in the training data. Is that really possible? That would seem to disagree with Table 6. How can you add a trigger filter but keep 100% recall?"

- RESPONSE: We describe a relation as a tri-occurrence (non-specific miRNAs- relation trigger -gene/proteins or disease). As rightly pointed out by the reviewer, the relation is true only when a relation trigger is present. We agree with the reviewer that this is not completely correct and the sentence could be mis-leading. The tri-occurrence approach cannot be 100% as there could be some entities that could be missed out during the automated named entity recognition. On the other hand, we assume that the entity recognition reaches 100% and all the relations (entity pair with a relation trigger) are retained in tri-occurrence. Thus, the assumption that there is no diminish in recall. We have corrected the corresponding text in out revised manuscript for clarity, as given below:

  The recall is not diminished using the tri-occurrence approach as the true entity pairs remain constant, while the precision increases between 4pp (percentage points) and 17pp when compared to the co-occurrence based approach, reducing false positives (cf. Figure

2).

In our work, we assume that all the entities have been identified giving a recall of 100% for both co-occurrence and tri-occurrence based approaches.

In Relation to the extraction sub-section of Results and Discussion we have already tried to address this issue in the following text:

"Most relation extraction approaches are dependent on the performance of named entity recognition. The impact of error propagation coming from automated entity recognizers is evaluated by applying the tri-occurrence method on the automatically annotated training and test corpus, here termed as "NERTri". Compared to the results on the gold standard entity annotation a drop of 13 pp for NonSpMiR-D, 7pp for NonSpMiR-GP, 22pp for SpMiR-D, and 30pp for SpMiR-GP in F 1 is observed for the test corpus."

COMMENT: "I think the reader would benefit from more discussion on the variance of miRNA names and trigger expressions, to gain an intuitive grasp of the difficulty of the task. Which leads me to the regular expressions described in tables 4 and 5. Specifically, I do not understand the alias Let and Lin. Are these individual miRNAs? Why would you want to define a re-useable alias for individual miRNAs? Alternatively, I am misunderstanding something here, in which case I would appreciate clarification: what are these alias symbols, and how do I know they generalize? The impression I get from the regular expressions is that miRNA naming is highly regular and very simple. Is that the case really?"

- RESPONSE: Considering the reviewer's comment, we have now modified the description in Automated entity recognizers sub-section in methods to:

Detected entities are resolved to a unique miRNA name and disambiguated to adhere to standard naming conventions as authors use several morphological variants to report the same miRNA term. For example, miR-107 can be represented as miRNA-107, MicroRNA-107, MicroRNA 107, has-mir-107, mir-107/108, micro RNA 107 and 108, micro RNA (miR) 107 and so on. Thus, the identified miRNA entity has been resolved to its base f form (e. g. hsa-microRNA-21 to hsa-mir-21 and microRNA 101 to mir-101) following the miRBase naming convention.

The relation trigger dictionary comprises of all interaction terms from the training corpus. After reviewing the training corpus for relation trigger terms, we retrieved not one but many variants of the same relation trigger occurring in alternative verb-phrase groups. For example, "change in expression" can be represented in one of the following verb-phrases: Change MicroRNA-21 Expression, Expression of caveolin-1 was changed, Change in high levels of high-mobility group A2 expression, change of the let-7e and miR-23a/b expression, expression of miR-199b-5p in the non-metastatic cases was significantly changed, etc. To allow flexibility for capturing relation trigger along with its variants spanning over different phrase length, we first manually represented all the relations in its root form, such as "regulate expression" to "regulate" (cf. Relation_Dictionary.txt file in supplementary). The base form has been extended manually to different spelling variants, e.g. regulate to regulatory, regulation, etc., the detailed listing of variants is provided in Supplementary file Word_variations.txt. Not all combination of the root forms is logical; target and up-regulation terms cannot be combined to form a relation trigger. Thus, we

additionally defined a set of relation combinations that are allowed (see Permutation_terms.txt in supplementary for all combinations).

Lin and Let were the first known miRNAs, identified in nematode. Naming convention for miRNAs were later developed after their identification leading to recognition of microRNAs as a class of small regulatory molecules. The original names of "lin" and "let" are still used as is. Yes, Let and Lin are individual miRNA types/families.

The regular expression can get very complicated when one wants to re-use. Thus, we tried to simplify it and split it into several small regular expressions to be used in bigger complex one. We have defined the aliases following a simple representation; the user can redefine the aliases if needed. Since the regular expression follows a standard representation (Oualline, 2001) we assume that it should be flexible to be adapted to different representations for implementations in other frameworks or programming languages. We partially agree with the reviewers that the miRNA naming is regular. However, the naming can have many variants in several combinations of our developed regular expressions capturing the variant descriptions in publications. We have tried our best to capture as many variants as possible. Also, we have improved the regular expression aliases for better understanding.

COMMENT: "In the results and discussion section, I am perplexed by the 0.79 vs 0.69 F-score train/test difference for disease mentions. Do you have an insight as to why specifically disease mentions would have such a major difference when the other entities do not? This is especially puzzling since Table 2 shows disease as the largest class, i.e. it should exhibit least noise in the results."

- RESPONSE: Disease mentions vary in the way they are represented. A disease entity can occur as multi-word with case variation and synonym combination, such as "Chronic Lymphocytic Leukemia" could be also represented as "Chronic Leukemia (lymphocytic)". Resolving the acronyms for diseases can be tricky as well, for example "AD" could be resolved to "Alzheimer's Disease" or "Atopic Dermatitis". This leads to ambiguity during tagging of the disease entities. Nested disease names are common where abbreviations are represented within the disease name itself, *e.g.* "Alzheimer's (AD) Disease". Thus, there could be large difference in the way disease names are tagged in training and test corpus by our NERTri approach in comparison to manual annotation. The performance varied between these two corpora when considering the exact match. However, we report a performance of 0.88 of F_1 for both train and test corpus, showing that partial matches perform better in disease entity identification.

  Considering reviewer's concern we have included a sentence in Performance evaluation of named entity recognition sub-section in results and Discussion:

  T*he low score for disease identification could be due to the variation in disease mentions, such as multi-word, synonym combination, nested names, etc.*

Some other not so major points:

COMMENT: I don't know what is meant by "improvised framework" in the abstract.

- RESPONSE: We have improved the text in Abstract for clarity as shown below:

  Additionally, most of the published miRNA entity recognition methods are keyword based, further subjected to manual inspection for retrieval of relations. Despite the fact that several databases host miRNA-associations derived from text, lower sensitivity and lack published details for miRNA entity recognition and associated relations identification has motivated the need for developing comprehensive methods that are freely available for the scientific community.

COMMENT: "I am not sure what is the status of relations that do not have a trigger. Or are there such?"

- RESPONSE: The relations are defined as a tri-occurrence, where two entities (in our case miRNA-genes/proteins or miRNA-disease) co-occur along with a relation trigger term in a single sentence. During manual annotations we considered only those sentences where two entities co-occur with a relation trigger as relation. If there occurred a sentence where miRNA and gene/proteins entity appeared but without a relation trigger, we did not tag them as relations. Thus, relations without trigger term never occur.

  Here is an example of relation trigger (target), which does not participate in any relation (from PubMED ID: 21346322):

  *"As single **miRNAs** are often predicted to **target** up to hundreds of individual transcripts, **miRNAs** are able to broadly affect the overall protein expression state of the cell."*

  In the above sentence there are two Non-specific miRNA entities along with one Relation trigger term. Since we do not have the second entity mention, such as genes/proteins or diseases, we do not tag this sentence as a relation instance. However, the above-mentioned named entities are tagged to their respective classes.

  Considering the reviewer's comment we have modified the text in Relations annotation sub-section of Methods as follows:

  Relevant triples, an interacting pair (from one of the above-mentioned) co-occurring with a RELATION TRIGGER in a sentence is defined to form a relation and can belong to one of the four above-mentioned Relation classes. On the contrary, if an interacting pair does not co-occur with any RELATION TRIGGER then we do not tag such pair as a relation.

COMMENT: Where does the number 11.5% of cross-sentence relations come from? Counting in the corpus? (just checking)

- RESPONSE: We have now included text describing the approach through which we obtained the statistic, shown below:

  We queried MEDLINE for "miRNA and Epilepsy" documents, among which 16 documents containing miRNA-related relations were manually selected. Manual inspection of these articles revealed 11.5% of miRNA-related associations occur outside the sentence level. T

hus, our work focused on relations at sentence level.

COMMENT: The previous point about 11.5% of cross-sentence relations makes me then wonder how can the co-occurrence approach reach 100% recall? Have these cross-sentence relations been simply deleted from the data?

- RESPONSE: As pointed out by the reviewer, we have modified the pointed text as follows:

  If all the entities are correctly identified then co-occurrence based approach leads to 100% recall for relation extraction. The recall is not diminished using the tri-occurrence approach, as the true entity pairs remain constant, approach while the precision increases between 4pp (percentage points) and 17pp when compared to the co-occurrence based approach, reducing false positives (cf. Figure 2). However, overall the precision reaches less than 60%. In our work, we assume that all the entities have been identified giving a recall of 100% for both co-occurrence and tri-occurrence based approaches.

  In our work, we assume that all the entities have been identified giving a recall of 100% for both co-occurrence and tri-occurrence based approaches.

  The cross-sentence has not been deleted from data and is freely available for researcher's to use the corpus and build a co-reference approach for the same.

COMMENT: "Towards the end of the methods section, the paper describes the use of "deep parsing". Through the citation I'm guessing this relates to Stanford Dependencies. I think the paper really should give some detail about how this parsing was done."

- RESPONSE: Yes, the "deep parsing" in the manuscript refers to Stanford Dependencies. The following pointed text has been modified in the manuscript:

  Next to lexical features, dependency parsing (created using Stanford parser) provides an insight into the entire grammatical structure of the sentence 37 and was performed using the Stanford CoreNLP library (http://nlp.stanford.edu/software/corenlp.shtml). Dependency parsing follows the shortest dependency path hypothesis[38].

Competing Interests: No competing interests disclosed.

Reviewer Report 29 August 2014

https://doi.org/10.5256/f1000research.4912.r5973

? **Sofie Van Landeghem**

Bioinformatics and Evolutionary Genomics, Ghent University, Ghent, Belgium, Belgium

This manuscript presents a manually annotated corpus of miRNA entities, genes/proteins and diseases, as well as the relations between them. The authors have used this dataset to develop NER and relation detection tools, which perform quite well in terms of precision, recall and F-score. The resulting miRNA relations detected automatically in Medline can be useful to extend existing databases with reliable literature information. All datasets are made freely available.

This research domain is highly relevant and it is great seeing text mining efforts focus specifically on miRNAs and their relation with diseases. The creation of a manually annotated training set certainly helps to advance this field. Two unfortunate design choices are the limitation to abstracts, even though so much data is readily available in PMC OA, and the missing support for cross-sentence relations, but these issues would definitely make for interesting future work.

Major questions/remarks

How exactly are candidate instances, relations and relation triggers defined? During manual annotation, can their be relations annotated without relation triggers or does this never occur? And what do relation triggers look like if they're not part of a relation? (cf. Table 2). Finally, how would the ML method perform if you would give it only tri-occurrence-based candidate instances?

I was hoping to see more implementation/evaluation details, specifically concerning the differences between LibSVM, LibLINEAR and Naive Bayes. It is not even stated which of these methods performed best.

How exactly are detected entities resolved to unique miRNA names? Because no details are given, can we assume that this step is not as complicated/ambiguous as gene name normalization for instance? Can database identifiers be retrieved for non-human cases? Do the evaluation results pertain to the textual symbols, or is the normalization to unique IDs also taken into account?

Could the authors clarify how they have ensured that data from the test set was not used in any way to develop the regular expressions for NER? I find it striking that the F-scores to detect miRNA are higher on the test set, or is it simply the case that miRNA entities are expressed in a homogeneous fashion throughout literature?

The evaluation set presented in Table 7 seems rather limited. Would it not be feasible to compare the newly presented methods on bigger datasets, such as those discussed in related work (miRCancer DB, Murray et al), or expand the scope of the evaluation beyond Alzheimer's disease? Additionally, why are only 100 abstracts retrieved for Alzheimer? Is this because the evaluation is done (partly) manually?

Minor questions

- Why are species mentions restricted to those occurring in miRBase, and why are only human-specific prefixes defined in the regular expression?
- Is the gene name dictionary built for human genes only, for the miRBase species only, or all?
- How useful are the non-specific miRNA mentions? I could see their value in trying to resolve co-reference relations across sentences, but this does not seem to be the aim in this study.

In this sense, I find this statement puzzling: "Distinguishing between two types of miRNA mentions has enabled us to achieve better recall and precision in document retrieval and relations identification".
- Was Table 1 constructed using exact string matching? It would be interesting to see both numbers for stringent criteria as well as those for allowing partial mis-matches (e.g. slightly different entity span).
- How can there be 39 articles with relations if the query only returned 37? (Table 8 + surrounding text)
- I was expecting the four last rows of Table 2 to add up to the same number as the "positive entity pairs" number in Table 3?
- How were the 41 abstracts in the second section of "use case analysis" selected? Was there not more information to be found in Medline?

Minor writing comments

- I don't see the need to "normalize" the number of miRNA publications, multiple Y-axis tend to complicate data plots. Personally, I would use the same (logarithmic) scale
- I would make a more obvious distinction between the manual curation efforts and the development of the NER and relation detection tools, for instance by placing the first 3 sections of Methods in a different "Data curation" section.
- Second to last sentence of the "Motivation" paragraph in the abstract: "regular expression" should be plural
- "Relation extraction paragraph": "an automated entity recognizers"
- "classes diseases" in Conclusion paragraph
- I had trouble reading/understanding this sentence: "Boundary matches result for the same reported 0.88 of F_1"
- The claim that 11.5% of miRNA relations are across sentences should be justified by a citation. Further, I personally think this is a significant portion, and wouldn't use the phrase "only 11.5%".

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 13 Nov 2014

Roman Klinger, University of Stuttgart, Germany

We thank the referee for the detailed comments. We will address them and then submit a new version together with detailed comments.

Competing Interests: No competing interests were disclosed.

Author Response 07 Dec 2014

Shweta Bagewadi, Fraunhofer-Gesellschaft, Germany

COMMENT: "This research domain is highly relevant and it is great seeing text mining efforts focus specifically on miRNAs and their relation with diseases. The creation of a manually annotated training set certainly helps to advance this field. Two unfortunate design choices are the limitation to abstracts, even though so much data is readily available in PMC OA, and the missing support for cross-sentence relations, but these issues would definitely make for interesting future work."

- RESPONSE: As rightly pointed out by the reviewers, we agree with the limitation of our design choice. We also agree that it makes an interesting future work. However, as stated by Shah *et al.*(Shah, Perez-Iratxeta, Bork, & Andrade, 2003) abstracts provides the best proportion of keywords for extracting biological information. We have included the below text in Corpus selection, annotation and properties sub-section of Methods explaining the choice:

  Shah et al.[27] showed that abstracts provide a comprehensive description of key results obtained from a study, whereas full text is a better source for biological relevant data. Thus, we choose to build the corpus for abstracts only.

  Additional text has been included in Conclusion and future work section:

  The proposed methods encourage future work of implementing the same for full-text articles to elucidate many more relations from Biomedical literature. Non-specific miRNA mentions identification could prove highly beneficial for co-reference resolution in full-text articles, in addition to abstracts.

  We are currently working on a more robust miRNA-relation identification workflow implemented in UIMA framework and JAVA for full text. We will include the co-reference resolution approach once the complete workflow has been validated. However, this manuscript is our first attempt to detect miRNA relations and to use them in a bigger integrative modeling approach, NeuroRDF, we recently developed, please refer to Iyappan *et al.* (Iyappan, Bagewadi, Page, Hofmann-Apitius, & Senger, 2014).

Major questions/remarks

COMMENT: "How exactly are candidate instances, relations and relation triggers defined?"

- RESPONSE: Description for candidate instances, and relation trigger is provided in Named entities annotation sub-section in Methods. The describing text is provided below:

  “Mentions of miRNAs consisting of keywords (case-insensitive and not containing any suffixed numerical identifier) such as “Micro-RNAs” or “miRs” are annotated as N on-S pecific miRNA. Names of particular miRNAs such as miRNA-101, suffixed with numerical identifiers are labeled as S pecific miRNA. Numerical identifiers (separated by delimiters such as “,”, “/”, and “and”) occurring as part of specific miRNA mentions are annotated as a single entity.

  Mentions of disease names, disease abbreviations, signs, deficiencies, physiological dysfunction, disease symptoms, disorders, abnormalities, or organ damages are

annotated as Disease. Possessive terms such as " Diabetic patients" are not marked. Mentions referring to proteins/genes which are either single word (e.g. "trypsin"), multi-word, gene symbols (e.g. "SMN"), or complex names (including of hyphens, slashes, Greek letters, Roman or Arabic numerals) are annotated as Gene/Protein. Only those organisms that are having published miRNA sequences and annotations represented in miRBase database are labeled as Species. Any verb, noun, verb phrase, or noun phrase associating miRNA mention to either labeled disease or gene/protein term is annotated as Relation Trigger."

For description of relations, please look into Relations annotation sub-section of Methods. The text available in manuscript is provided below:

"We restrict the relationship extraction to sentence level and four different interacting entity pairs: S pecific miRNA-D isease (SpMiR-D), S pecific miRNA-G ene/P rotein (SpMiR-GP), N on-S pecific miRNA-D isease (NonSpMiR-D), and N on-S pecific miRNA-G ene/P rotein (NonSpMiR-GP). Relevant triples, an interacting pair co-occurring with a R elation T rigger are defined to form a relation and can belong to one of the four above mentioned Relation classes."

COMMENT: "During manual annotation, can their be relations annotated without relation triggers or does this never occur? And what do relation triggers look like if they're not part of a relation? (cf. Table 2)."

- RESPONSE: The relations are defined as a tri-occurrence, where two entities (in our case miRNA-genes/proteins or miRNA-disease) co-occur along with a relation trigger term in a single sentence. During manual annotations we considered only those sentences where two entities co-occur with a relation trigger as relation. If there occurred a sentence where miRNA and gene/proteins entity appeared but without a relation trigger, we did not tag them as relations. Thus, relations without trigger term never occur.

  Here is an example of relation trigger (target), which does not participate in any relation (from PubMED ID: 21346322):

  "As single **miRNAs** are often predicted to **target** up to hundreds of individual transcripts, **miRNAs** are able to broadly affect the overall protein expression state of the cell."

  In the above sentence there are two Non-specific miRNA entities along with one Relation trigger term. Since we do not have the second entity mention, such as genes/proteins or diseases, we do not tag this sentence as a relation instance. However, the above-mentioned named entities are tagged to their respective classes.

  Considering the reviewer's comment we have modified the text in Relations annotation sub-section of Methods as follows:

  Relevant triples, an interacting pair (from one of the above-mentioned) co-occurring with a RELATION TRIGGER in a sentence is defined to form a relation and can belong to one of the four above-mentioned Relation classes. On the contrary, if an interacting pair does not co-

occur with any RELATION TRIGGER then we do not tag such pair as a relation.

COMMENT: "Finally, how would the ML method perform if you would give it only tri-occurrence-based candidate instances?"

- RESPONSE: We are not fully sure what is meant with this comment. Does the reviewer mean to only provide positive instances for training? In that case, a one-class SVM could be used, but we do not see a reason to believe that this could provide improved performance over a two-class SVM. If the reviewer means to only use features based on the tri-occurrence, this would provide another baseline for the machine learning approach.

COMMENT: "I was hoping to see more implementation/evaluation details, specifically concerning the differences between LibSVM, LibLINEAR and Naive Bayes. It is not even stated which of these methods performed best."

- RESPONSE: We would like to thank the reviewer for interest in more detailed results. We have now included the result of the other machine learning methods in Figure 2. Additionally, text discussing these results have been included in Relation extraction sub-section in Results and Discussion as below:

  Using the machine-learning based classification, precision is increased up to 76% for specific miRNA-gene relations for both LibLINEAR and LibSVM methods, although Naïve Bayes is not far behind. Similarly, these two methods performed nearly the same for specific miRNAs-disease relations, the F 1 measure is not substantially different but a trade-off between precision and recall can be observed. An increase in F 1 measure is observed for non-specific miRNA relations when Naïve Bayes method is applied, out performing other strategies. Nevertheless, preference of the method highly depends on the compromise one chooses, whether better recall or precision. Overall, better recall and acceptable precision can be achieved with tri-occurrence method.

COMMENT: "How exactly are detected entities resolved to unique miRNA names? Because no details are given, can we assume that this step is not as complicated/ambiguous as gene name normalization for instance? Can database identifiers be retrieved for non-human cases? Do the evaluation results pertain to the textual symbols, or is the normalization to unique IDs also taken into account?"

- RESPONSE: The detected miRNA entities have been resolved using the mirMaid REST service (described in “Named Entity Recognition Section”). Normalization of miRNA names is not as complicated as gene names since the authors follow naming scheme that can be captured with good regular expressions. However, the major challenge what we see is resolving the miRNA names to the right species. We have used species dictionary to support non-human miRNA normalization. The evaluation results are done using the normalized names as given in miRBase. In addition, we can make the miRBase IDs also available. Considering the reviewer’s comment we have included text which explains the normalization step in detail, given below:

  Detected entities are resolved to a unique miRNA name and disambiguated to adhere to standard naming conventions. Each identified miRNA entity has been resolved to its base form (e. g. hsa-microRNA-21 to hsa-mir-21 and microRNA 101 to mir-101). Manual inspection of the test corpus for species distribution revealed that 71% of the documents

belonged to human, followed by mouse (15%), rat (8%). Pig has 2 abstracts, zebrafish, HIV-1, HSV-1, and Caenorhabditis elegans 1 each (cf. Supplementary Figure A for the distribution). Thus, we assumed that most of the abstracts belonged to human and resolved the identified miRNA entities to human identifier in miRBase. Unique miRNA terms are mapped to human miRBase database identifiers through the mirMaid Restful web service. For those names where we do not retrieve any database identifiers, we fall back to another organism mention found in the abstract (if any), using the NCBI taxonomy dictionary (see below) (cf. Supplementary Figure B), otherwise we retain the unique normalized name (cf. Box 2).

MIR0000007:MIMAT0015092@MIRBASE|MI0000002@MIRBASE|cel-lin-4|lin-4
MIR0000008: miR-171|microRNA 171
MIR0000005:MIMAT0000416@MIRBASE|has-miR-1|miRNA-1

Box 2: Represents the un-normalized and normalized entities that are mapped to miRBase identiers. Here MIR0000007, MIR0000008, and MIR0000005 are internal identifiers used by ProMiner.

COMMENT: "Could the authors clarify how they have ensured that data from the test set was not used in any way to develop the regular expressions for NER? I find it striking that the F-scores to detect miRNA are higher on the test set, or is it simply the case that miRNA entities are expressed in a homogeneous fashion throughout literature?"

- RESPONSE: We first randomly retrieved 300 abstracts from the PubMed using the keyword “miRNA”. From these we manually selected 301 abstracts that contain gene/proteins or disease terms without looking in detail for any relation term. We randomly split these in train (201) and test (100) set (described in Corpus selection, annotation and properties section) before the annotation step. Thus, we are confident that we have not used test set for development of regular expression. The miRNA naming convention has been in effect quite early after discovery of miRNAs. This has led to uniform naming usage relative to genes/proteins. In our experience, from manual annotation, we can conclude that most authors follow one of the patterns, among a set of naming schemes, to mention miRNAs in publications. Our regular expression pattern has been developed to be robust enough to capture all these patterns. Since, the test corpus is smaller than the training set (we built our regular expression on training set) we expect the performance to be better when a set of common naming schemes is followed for mentioning miRNAs.

COMMENT: "The presented in Table 7 seems rather limited. Would it not be feasible to compare the newly presented methods on bigger datasets, such as those discussed in related work (miRCancer DB, Murray *et al*), or expand the scope of the evaluation beyond Alzheimer's disease?"

- RESPONSE: We agree with the reviewers that the evaluation set in Table 7 is limited, but currently this is the only database where manually miRNA-disease relations are available. Also, another reason could be that the database has not been updated since April 2008.

  We appreciate the suggestion of the reviewer for extending the validation to other datasets, but our department is heavily working within the field of

Neurodegeneration, we would like to limit ourselves to diseases that fall into this domain.

COMMENT: Additionally, why are only 100 abstracts retrieved for Alzheimer? Is this because the evaluation is done (partly) manually?

- RESPONSE: Using the keyword search (described in "Extraction of miRNA-gene relations for Alzheimer's diseases from full MEDLINE" section) we retrieved 124 abstracts. Yes, since we manually selected relevant abstracts among these, we retained only 100 abstracts.

Minor questions

COMMENT: Why are species mentions restricted to those occurring in miRBase, and why are only human-specific prefixes defined in the regular expression?

- RESPONSE: miRBase is the primary database that publishes miRNA sequences and annotations. miRBase registry assigns unique names to all miRNAs for publication (just like HUGO for gene names). This database has published a list of species for which miRNA sequences have been identified, which means only for these organisms miRNAs have been discovered as of today. Thus, we restrict ourselves to miRBase-listed organisms.

  We currently developed the proposed workflow to primarily capture human miRNA relations since we have curated diseases and gene/proteins dictionary only for humans. However, even if the miRNA mentions are given for other organism such as "cel-lin-4" our method captures "lin-4" which is later resolved to other organisms during the normalization process. We used human-specific prefixes in our regular expression for simplicity and instant capture of human related miRNAs. Also, we included other unique prefixes "lin-4" and "let-7" since these were the first miRNAs identified in nematodes even before the naming convention was in place. However this can be easily extended to other organisms using the three-letter code provided by miRBase.

COMMENT: "Is the gene name dictionary built for human genes only, for the miRBase species only, or all?"

- RESPONSE: For the current manuscript preparation we have used gene/proteins dictionary that has been built for human only. However, we plan to extend this to other miRBase species in future.

COMMENT: "How useful are the non-specific miRNA mentions? I could see their value in trying to resolve co-reference relations across sentences, but this does not seem to be the aim in this study. In this sense, I find this statement puzzling: "Distinguishing between two types of miRNA mentions has enabled us to achieve better recall and precision in document retrieval and relations identification"."

- RESPONSE: The reviewer has aptly pointed out the interesting future work we planned to implement. Considering the reviewer's comment, we have improved the description in the pointed text to follows:

  In this work, we proposed approaches for identification of relations between miRNAs and other named entities such as diseases, and genes/proteins from biomedical literature. In

addition, details of named entity recognition for all the above entity classes have been described. We distinguished two types of miRNA mentions, namely Specific (with numerical identifiers) and Non-Specific (without numerical identifiers). Non-specific miRNAs entity recognition has enabled us to achieve better recall and precision in document retrieval.

The proposed methods encourage future work of implementing the same for full-text articles to elucidate many more relations from Biomedical literature. Non-specific miRNA mentions identification could prove highly beneficial for co-reference resolution in full-text articles, in addition to abstracts. Extending the current approach to other model organisms such as mouse, and rat can help in revealing important relations for translational research. Inclusion of additional named entities such as drugs, pathways, etc. could lead to an interesting approach for detection of putative therapeutic or diagnostic drug targets through a gene-regulatory network generated from identified relations.

COMMENT: "Was Table 1 constructed using exact string matching? It would be interesting to see both numbers for stringent criteria as well as those for allowing partial mis-matches (e.g. slightly different entity span)."

- RESPONSE: Yes, Table 1 represents the exact string match results. As requested by the reviewer, we have now included results for partial mis-matches (called as partial match) in Table 1. Additionally, we have included discussion text to the pointed section as follows:

  Table 1 provides the inter-annotator agreement (measured as F 1, for both exact and boundary match, and Cohen's κ) for the test corpus. Exact string match occurs only when both the annotators annotate identical strings, whereas in partial match fraction of the string has been annotated by either of the annotators.It is evident (cf. Table 1) that in almost all cases partial match performs better than exact string match, indicating variations in span of mentioned entities.

COMMENT: "How can there be 39 articles with relations if the query only returned 37? (Table 8 + surrounding text)"

- RESPONSE: We thank the reviewer for pointing out the error. The typo error has been corrected in the revised manuscript. The correct number of articles has been re-checked for 37 articles.

COMMENT: "I was expecting the four last rows of Table 2 to add up to the same number as the "positive entity pairs" number in Table 3?"

- RESPONSE: We thank the reviewer for pointing out the mistake in the statistics. We have now re-checked the statistics using the published corpus and have updated the manuscript with correct statistics along with the README file in our website (http://www.scai.fraunhofer.de/mirna-corpora.html).

COMMENT: "How were the 41 abstracts in the second section of "use case analysis" selected? Was there not more information to be found in Medline?"

- RESPONSE: We applied NERTri approach (tri-occurrence based approach applied on the entities identified by ProMiner system) to retrieve 41 abstracts, which have miRNA-Alzheimer's disease relations at sentence level. We did not identify any more relations than the provided (as of 4th July, 2013). However, we assume that there could be some false negatives occurring due to the error propagation from automated entity recognizers (*cf.* Relation extraction sub-section in Results for

details). Also, relations that occur at document level could have been missed out since our current focus is more on sentence level relations.

From another point of view, for last 25 years Alzheimer's disease research has mainly focused on amyloid-beta deposits that lead to neuronal death and tangle formation. However, the amyloid hypothesis has not been successful in late-phase drug trials (Delay & Hébert, 2011; Golde, Schneider, & Koo, 2011). The focus of miRNAs research on Alzheimer's is relatively new, there are in total 187 articles in PubMed in comparison to 1713 articles on miRNA research in Breast cancer (as of 25th Nov 2014). Thus, we can conclude that the number of articles related to miRNA-Alzheimer's disease research is rather limited and relatively new.

We have modified the pointed out text in the manuscript to the following for clarity:

For comparison, we retrieved miRNA-disease relations from MEDLINE using NERTri approach, resulting in 41 abstracts containing 159 relations. Obtained triplets have been manually curated to remove 51 false positives. False negatives have not been accounted, which may result in loss of information (cf. Relation extraction section). Comparison between the relations obtained from miR2Disease and NERTri are summarized in Table 7.

Minor writing comments

COMMENT: "I don't see the need to "normalize" the number of miRNA publications, multiple Y-axis tend to complicate data plots. Personally, I would use the same (logarithmic) scale"

- RESPONSE: The growth of miRNA publications has been normalized using the number of articles published in PubMed for the given year. The decision to normalize values has been taken due to low number of articles for miRNA, e.g. 37 articles related to miRNA were published in comparison to 561169 articles in the whole MEDLINE for the year 2002. Thus, we would prefer to keep the scale as it is. We addressed that comment and are now using only one axis in the depiction.

COMMENT: "I would make a more obvious distinction between the manual curation efforts and the development of the NER and relation detection tools, for instance by placing the first 3 sections of Methods in a different "Data curation" section."

- RESPONSE: We thank the reviewer for her suggestion. We have now moved the first three sections of Methods under "Data curation and corpus selection" subsection. Hope this is inline with what the reviewer had in mind during the suggestion.

COMMENT: "Second to last sentence of the "Motivation" paragraph in the abstract: "regular expression" should be plural"

- RESPONSE: As rightly pointed out by the reviewer, we have corrected the above text in the manuscript.

COMMENT: "Relation extraction paragraph": "an automated entity recognizers"

- RESPONSE: We thank the reviewer for the suggestion. However, "relation extraction" has an entirely different meaning when compared to "an automated entity recognizers". We assume that the reviewer meant to change the "Named entity recognition" paragraph title. We appreciate the suggestion and have now changed the "Named entity recognition" paragraph title to "Automated named entity recognition". We hope this is in agreement with the reviewer.

COMMENT: ""classes diseases" in Conclusion paragraph"

- RESPONSE: As rightly pointed out by the reviewer, we have corrected the above text in the manuscript.

COMMENT: "I had trouble reading/understanding this sentence: "Boundary matches result for the same reported 0.88 of F1""

- RESPONSE: The “boundary matches” refers to “partial matches”. For clarity we have now modified the text and included reference to the provided supplementary file as below:

  Partial matches result for the same reported 0.88 of F 1, providing the possibility for detection of similar text strings for better recall (cf. Supplementary Table B).

COMMENT: "The claim that 11.5% of miRNA relations are across sentences should be justified by a citation. Further, I personally think this is a significant portion, and wouldn't use the phrase "only 11.5%"."

- RESPONSE: We have now included text describing the approach through which we obtained the statistic, shown below:

  We queried MEDLINE for “miRNA and Epilepsy” documents, among which 16 documents containing miRNA-related relations were manually selected. To avoid any biased approach we choose Epilepsy disease domain. Manual inspection of these articles revealed 11.5% of miRNA-related associations occur outside the sentence level. Thus, our work focused on relations at sentence level.

REFERENCES:

Delay, C., & Hébert, S. S. (2011). MicroRNAs and Alzheimer’s Disease Mouse Models: Current Insights and Future Research Avenues. *International journal of Alzheimer’s disease*, *2011*, 894938. doi:10.4061/2011/894938

Golde, T. E., Schneider, L. S., & Koo, E. H. (2011). Anti-Aβ therapeutics in Alzheimer’s disease: The Need for a Paradigm Shift. *Neuron*, *69*(2), 203–213. doi:10.1016/j.neuron.2011.01.002.Anti-A

Iyappan, A., Bagewadi, S., Page, M., Hofmann-Apitius, M., & Senger, P. (2014). NeuroRDF : Semantic Data Integration Strategies for Modeling Neurodegenerative Diseases. *Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM2014)* (pp. 11–18). Aveiro, Portugal.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics*, *4*, 20. doi:10.1186/1471-2105-4-20

Competing Interests: No competing interests were disclosed.