Automatically Identifying Biological Functions of microRNAs from the Literature

Samir Gupta[1], **Catalina O Tudor**[1,2], **Cathy H Wu**[1,2], **Carl J Schmidt**[3], **K Vijay-Shanker**[1]
[1] Department of Computer and Information Sciences
[2] Center for Bioinformatics and Computational Biology
[3] Department of Animal and Food Sciences
University of Delaware, Newark, DE 19716
sgupta@udel.edu, oanat@udel.edu, wuc@dbi.udel.edu,
schmidtc@udel.edu, vijay@udel.edu

Abstract

Currently, miRNA databases contain limited information about the functional properties of miRNAs and their roles in biological processes. Automatically extracting and filtering such information from the literature would help biologists to quickly get an overall idea about the miRNA's functions and involvement in biological processes. We employ a combination of frequency-based and pattern-matching techniques to extract highly relevant function-like and process-like terms for a miRNA, as well as their exact relations to the miRNA. The results of this study show that our approach can successfully select functions and processes that could be used in both short summaries and somewhat longer summaries for miRNAs.

1 Introduction

MicroRNAs (also miRNAs or miRs) are small molecules involved in many biological processes and affecting gene regulation. In the early 2000s, miRs have been recognized as a distinct class of biological regulators with conserved functions. Since then, the research on miRs has been skyrocketing, with an increase in publications of over thirty to sixty percent year after year.

Many miR databases have been created, containing manually curated information. Some databases benefit from text mining assistance (e.g., ranking documents or identifying miR-target associations in text for presentation to the curator). However, nearly all of these resources focus on miR-gene associations (e.g., Hsu et al., 2014) and miR-disease associations (e.g., Xie et al., 2013). Although these are valuable sources of information about miR targets, they contain limited information about miR functional properties and roles in biological processes.

Our project aims to augment the information found in miR databases with information automatically extracted from the literature. This is motivated by trying to create resources for miRs similar to those for genes/proteins, such as EntrezGene, which includes short summaries, GeneRIF sentences, and annotations of functions and processes. The goal of this study is to identify processes/functions that are highly related to miRs.

For genes, there have been many frequency-based approaches that automatically identify and rank functional information from the scientific literature (e.g., Andrade and Valencia, 1998, Tudor et al., 2010). Although useful in providing an initial ranking for terms co-occurring with a miR in the literature, a frequency-based approach does not ensure a direct miR-term association. Such direct associations are necessary if the terms are to be included in summaries for miRs. To detect such associations, we develop an approach that selects function-like and process-like terms for which a strong relationship can be identified in the sentences in which they co-occur.

Our approach combines frequency-based methods with rule-based syntactic relation extraction to identify terms for functional annotation that can also be used to create abstractive summaries for miRs. We call our system **STEM** (**S**ummary **T**erm **E**xtractor for **M**iRNA). The evaluation indicates that STEM can identify such terms with high precision and recall.

2 Approach

The basic steps of the algorithm are explained in the next three subsections: (1) identify *FPterms* (function-like and process-like terms) using a frequency-based approach; (2) detect syntactic relations for each miR-*FPterm* pair; and (3) assign relevancy based on additional features.

2.1 Identifying *FPterms*

We used the frequency-based approach of eGIFT (Tudor et al., 2010), which extracts *i*Terms, to identify *FPterms* for miRs. In this approach, the terms are stemmed and grouped as lexemes (groups of words having the same root). As in eGIFT, a lexeme corresponding to an *i*Term is associated with a specific miR by assigning a score. The score is based on the frequency of the *i*Term in abstracts containing the specific miR, as compared to the frequency of the *i*Term in a background set of abstracts mentioning any miRs. Using the method of eGIFT, the *i*Terms are grouped into various categories, based on dictionary look-ups, surrounding words (e.g., gene, protein, complex, domain, analysis) or morphological derivatives (e.g., -tion, -sion, -sis). The *FPterms* considered in this study are those subsets of *i*Terms, which are identified as functions or processes by this categorization. Examples of *FPterms* identified for miR-1 are "tumor suppressor", "myogenesis", "cell proliferation", and "gene expression".

2.2 Detecting Syntactic Relations

Although many *FPterms* describe properties of miRs, not all are important enough to be included in a summary. For example "gene expression" for miR-1 is probably not appropriate, as any miR will affect gene expression. Also the eGIFT-based method might pick some terms because of their frequencies even though there is no important relationship with the miR. Therefore, to find the relationship for a miR-*FPterm* pair, STEM starts by extracting sentences mentioning any variations of the miR name (e.g., miR-1, microRNA1, miRNA-1), and any variation of the given *FPterm* (e.g., suppressor, suppression, suppresses). Then, chunks of syntactically related tokens are formed, and simple constructs, such as appositives, relative clauses and reduced relative clauses, conjunctions, and parenthetical elements are detected using iSimp (Peng et. al., 2012). This helps to break down a complex sentence into multiple simple sentences. While looking at gene summaries, we observed certain specific types of relationships between important terms and the genes. These relationships are motivated and described below.

Regulation/Involvement Relations: A term has a strong relationship with a miR, if the miR is thought to "regulate" or "be involved in" that process. We identify such relations using simple lexical/syntactic rules, that follow agent-theme patterns, where the miR is in agent position and the term is in theme position. The verb-based trigger connecting the miR and the *FPterm* is used to differentiate between involvement (e.g., "is involved in", "plays a role in", "implicated in", etc.) and regulation ("regulates", "mediates", "activates", etc.) relations. Lexical and syntactic information, simplification constructs and the voice of the connecting verb all play a crucial role in detecting if a relationship exists. For example, "regulation of cell proliferation by miR-9" and "miR-9, a ... which regulates cell proliferation" will both extract the relation [mir-9, regulates, cell proliferation]. Note that in the second fragment, the nominalized form of the verb is used. Certain sentences like "miR-10b actively participates in cancer formation by promoting cell migration" indicate causal relations, where miR-10b first promotes cell migration and then participates in cancer formation. In this case, we extend iSimp simplification to generate two sentences: "miR-10b, actively participates in cancer formation" and "miR-10b promotes cell migration", from which the relationships are then easily extracted.

Is_a Relations. Certain terms describe attributes/properties of miRs, which are important to be included in short descriptive summaries. Terms ending in "or/er" (e.g. tumor suppressor, biomarker) fall into this category. We use patterns similar to the above to extract such relations. The verb-based triggers connecting the miR and the *FPterm* include "is a", "functions as", "acts as", etc. Simple sentences constructed from appositives more often fall in this category than others. For example, "miR-1, a putative tumor suppressor" is rephrased to "miR-1 is a putative tumor suppressor", hence, extracting the relationship [mir-1, is_a, putative tumor suppressor].

State Relations. A process term might not be a direct biological process, however, it might convey important functional information about the miR.

For example, "Hsa-miR-9 methylation status is associated with cancer development" indicates a state (i.e., methylation) of a miR (i.e., miR-9) that is associated with a disease (i.e., cancer development). We use pattern-based rules like "state of miR" and "miR state", where the state is the nominalized form of a non-regulatory verb. In the above sentence, STEM extracts the relationship [mir-9, state, methylation].

2.3 Assigning Relevancy to *FPterms*

Consider the fragment "**miR-9a** is involved in the regulation of **neurogenesis**". The involvement/ regulation patterns, as described in the previous section, will extract the relation between agent "miR-9" and theme "regulation of neurogenesis". However, the *FPterm* that we are interested in is simply "neurogenesis". Therefore, we extend our relation extraction process to allow for larger noun phrases (NP) containing the miR or the *FPterm*. However, by doing so, we also introduce weak miR-*FPterm* relations (e.g., "**MicroRNA-9** regulates the migratory mechanism in human **neural progenitor** cells). Hence, we use the type of relation and the positions of the miR and the *FPterm* in the corresponding noun phrases to assign relevance scores. Head positions (e.g., "neurogenesis" in the first example) or nouns attached to head positions by preposition "of" (i.e., nominalized forms of regulation triggers, e.g., "regulation" in the first example) are considered important and assigned a high relevance score.

The presence of one of the three relation types does not automatically imply a strong relationship between the miR and the *FPterm*. Considering *FPterms* are lexemes and will contain different morphological variations, the relations extracted for a miR-*FPterm* pair may be many. For example, the *FPterm* "suppression" will have variations like "suppress", "suppressing", and "suppression". Thus, one relation may indicate [mir-9, isa, suppressor] while another may indicate [mir-9, suppresses, some other process]. Relations where the *FPterm* occurs in a verbal form are considered not important and assigned a low relevance score. Attributes of miR detected via is_a relationships are assigned a high relevance score only if the miR and the *FPterm* occur in head positions within their corresponding NPs.

State relations derive their importance from the fact that they act as facilitating associations for the miR with some other important entity (e.g., disease). Thus, for a state relation to be considered important, we look at the type of verb phrase preceding or following the miR-state occurrence. Biologically meaningful verb phrases, as well as triggers like "associated with" or "correlated with", indicate that the miR's state is important, and so we assign a high relevance score.

For a given miR-*FPterm* pair, multiple relations will be extracted from multiple sentences. STEM uses a set of heuristics to assign a single relevance score by looking at the frequency as well as the types of the relationships. For example, an *FPterm* frequently occurring in a verbal position will be assigned a "not relevant" score.

3 Results and Discussion

3.1 Experimental Design

Four biology researchers, who were not involved in the design and development of STEM, were asked to provide annotations to be used in the evaluation. For this, 12 miRs were chosen randomly from all the miRs mentioned in 100 or more Medline abstracts, and top *FPterms* were chosen from these miRs for a total of 210 miR-*FPterm*. Each annotator was given a different set of around 50 miR-*FPterm* pairs. For each pair, we provided the annotators with all the sentences in which the miR and the term co-occur. We then asked them to annotate the pair with one of three options: (1) *relevant (R)*, if they would pick the *FPterm* to include in a short summary about the miR; (2) *somewhat relevant (SR)*, if they would only pick the *FPterm* to include in a longer summary about the miR; or (3) *not particularly interesting (NI)*, if they would not include this *FPterm* in a descriptive summary of the miR. No other information and examples were provided in the guidelines.

3.2 Results

The judges annotated the 210 miR-*FPterm* pairs with *R* in 88 cases (42%), *SR* in 30 cases (14%), and *NI* in 92 cases (44%). Our system, in turn, grouped the 210 miR-*FPterm* pairs into 75 *R* (36%), 47 *SR* (22%), and 88 *NI* (42%). Because the overall goal of this project is to identify *FPterms* for generating summaries about microRNAs, we show the results in terms of Precision, Recall, F-measure, and Accuracy for each type of summary, as explained in our guidelines.

The results are shown in Table 1. The first four columns contain the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) numbers of the system's outcome as compared to the annotations. Precision (P), Recall (R), F-measure (F), and Accuracy (A) are provided in the last three columns.

Summaries	TP	TN	FP	FN	P	R	F	A
Short (*R*)	63	110	12	25	**84.0**	71.6	77.3	**82.4**
Long (*R+SR*)	94	64	28	24	77.1	**79.7**	**78.4**	75.2
None (*NI*)	64	94	28	24	69.6	72.7	71.1	75.2

Table 1. Precision, Recall, F-measure, and Accuracy for the different types of summaries

3.3 Discussion

Our system performs well when identifying *FPterms* for short summaries (marked with *R*), as shown in the first row of Table 1. It appears that our approach is conservative, preferring precision over recall. Among the cases that influenced the precision, only 4 were R-NI errors in comparison to 8 R-SR errors. An example of R-SR disagreements is "colony formation" for miR-21, which matches a strong pattern of STEM but was marked as somewhat relevant by the annotators. Examples of R-NI disagreements include common terms, such as "translation" for miR-30. Meanwhile, recall errors were evenly distributed, with 13 SR-R errors and 12 NI-R errors. An example of SR-R disagreements is "cell migration" for miR-210, and an example of NI-R disagreements is "methylation" for miR-124. These were not captured by any of the strong patterns used in STEM, indicating that our relation extraction patterns need to be improved to detect such cases.

We noticed a drop in precision when identifying *FPterms* for long summaries (marked with *R* and *SR*), as shown in the second row of Table 1. In total, there were 24 SR-NI disagreements influencing the precision. An example is "secretion" for miR-15, which was assigned somewhat relevant by our heuristics, but marked as not interesting by the annotators. In total, there were 12 NI-SR disagreements influencing the recall of long summaries. An example is "exocytosis" for miR-124, which was due to complex sentences not clearly indicating to a direct relationship.

In conclusion, our analysis revealed that a majority of these errors can be attributed to missing patterns and incompleteness of the heuristics rules used in assigning relevancy scores.

4 Conclusion & Future Work

We described an approach for identifying functions, processes, and descriptive terms highly associated with microRNAs. Our method achieves high precision (84%) in selecting terms for short summaries, and good F-measure (79.7%) in selecting terms for somewhat longer summaries. The results of this preliminary study, presented in Table 1, showed that our approach has a great potential for selecting *FPterms* and corresponding sentences for the generation of microRNA summaries. We believe that results can be improved if machine learning would be used in the assignment of summary types to miR-*FPterm* pairs.

Acknowledgment

This work has been supported in part by the National Science Foundation under Grants No. 1147029 and 1062520. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Andrade, M. A., & Valencia, A. (1998). Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, *14*(7): 600-607.

Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., ... & Huang, H. D. 2014. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic acids research*,*42*(D1): D78-D85.

Tudor, C. O., Schmidt, C. J., & Vijay-Shanker, K. 2010. eGIFT: mining gene information from the literature. *BMC Bioinformatics*, *11*(1): 418.

Xie, B., Ding, Q., Han, H., & Wu, D. 2013. miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics*, btt014.

Peng Y, Tudor CO, Torii M, et al. iSimp: A Sentence Simplification System for Biomedical Text. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM2012) 2012. pp. 211–216.