Micro-RNA Recognition in Patents in BioCreative V.5

Chen-Kai Wang[1], Hong-Jie Dai[2,3*], Nai-Wun Chang[4]

[1]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan, R.O.C., [2]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan, R.O.C., [3]Interdisciplinary Program of Green and Information Technology, National Taitung University, Taitung, Taiwan, R.O.C., [4]Institute of Information Science, Academia Sinica, Taipei, Taiwan

dennisckwang@gmail.com; hjdai@nttu.edu.tw; naiwun@gmail.com;

Abstract. MicroRNAs (miRNAs) have been considered as good candidates for early detection or prognosis biomarkers for various diseases. Patents related to methods of identifying, isolating and amplifying miRNAs and potential use of miRNAs as biomarkers for cancers are increasing rapidly. In this work, we extend our miRNA recognition method based on the statistical principle-based approach and develop a web service followed the communication protocol defined by the biomedical annotation meta-server (BeCalm) platform to provide a service of miRNA recognition. The method can achieve an F-score of 0.988 for miRNA recognition on a manually annotated test dataset. During the participation of the BioCreative V.5 Technical Interoperability and Performance of Annotation Servers (TIPS) task, we set up our web service successfully and it can exchange the status message with the BeCalm platform and process the requests from BeCalm. Unfortunately, we met technical problems to send back the annotation results to the BaCalm platform.

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs of approximately 23 nucleotides, which negatively regulate the gene expression at the post-transcriptional level. Recently miRNAs have been considered as good candidates for early detection or prognosis biomarkers for various diseases. Patents related to methods of identifying, isolating and amplifying miRNAs and potential use of miRNAs as biomarkers for cancers

[*] Corresponding author

are therefore increasing rapidly. To facilitate the understanding of the state-of-the-art researches and applications of miRNAs, we present a REST (Representational State Transfer) web service for miRNA recognition.

2 Method

Our RESTful service contains three main components. The first is the data retrieval component which retrieves patents from remote data sources. In our current implementation, four sources are supported. The first and two data sources are PubMed Central (PMC) and PubMed. We used NCBI E-utilities to fetch requested data from the two data sources. The third and fourth sources are the pattern server and the abstract server released by the TIPS task.

The core of the web service is a miRNA recognition component based on our statistical principle-based approach (SPBA) [1]. The component integrate several natural language process and information extraction modules to process downloaded patents. Given a patent, MedPost [2] is used to split text into sentences and generate tokens for each sentence. We then employed our SPBA-based miRNA recognition method to recognize miRNAs in the preprocessed sentences. Our miRNA recognizer is developed based on the corpus released by S Bagewadi, T Bobic, M Hofmann-Apitius, J Fluck and R Klinger [3].

The training phase of SPBA consists of three main steps. The first is knowledge construction. In this step, we represent the knowledge related to miRNA terms through semantic slots and principles manually or semi-automatically with Information Map [4]. Figure 1 illustrates the hierarchical knowledge structure constructed for representing a miRNA in our approach. The root node is “miRNA” indicating that the structure represents the knowledge for miRNA names. The first child node of the root node is the “SLOT” node, under which we define the fundamental semantic unit (i.e. slot) for the root node. Consider the miRNA name as an example. Similar contents can be found among descriptions about miRNAs, which form the backbone of miRNA’s slots. For instance, both the miRNA “cel-miR-123-5p” and “hsa-microRNA-24-3P” consists of a species (*cel* and *hsa*), the indicating word “miRNA” and a hairpin that possess unique feature in representing a miRNA.

After constructing the knowledge, the principle generation step was applied. In this step, slots are assembled and summarized by observing the arrangement of principle slots which can accomplish the miRNA recognition task. For example, both the miRNA "cel-miR-123-5p" and "hsa-microRNA-24-3P" can be designated using the following combination of slots "[Species][miRNA][order][Hair-pin]". Here we use brackets to enclose a slot name for representing a slot. For example, "[Species]" is a slot that encodes the species in which the miRNA appears. "[miRNA]" is the slot representing the word indicating an occurrence of a miRNA name.

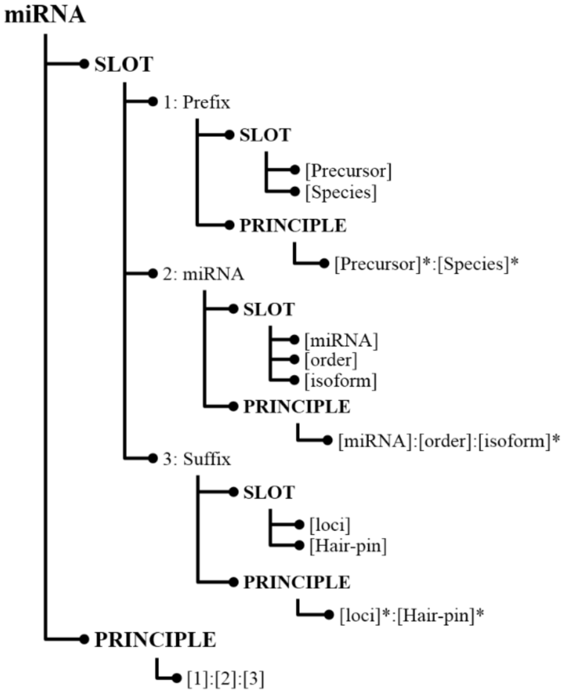
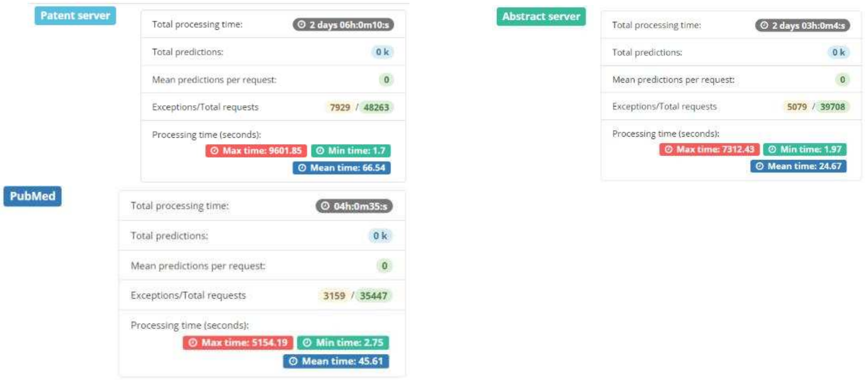


Figure 1. Semantic slots and principles defined for miRNA.

Lastly, a flexible principle matching algorithm allowing insertion, deletion, and substitution is applied to extract miRNAs represented by the compiled principles in the given text. Unlike traditional template matching that involves rigid left-right relation of slots in a sentence, a scoring criteria during principle alignment was used in SPBA, in which the collocation and bigram statistics is incorporated to estimate matching scores. During the principle matching procedure, we score those possible candidate principles based on matched slots, slot relations and insertions. Each exactly matched slot gets a score of 4. If there are insertion/deletion/substitution in the string, the scoring mechanism will assign scores accordingly. We calculate the score of an insertion by gathering its left (resp. right) bigram statistics with its neighboring left (resp. right) slots in the training set. A substitution is either a partial match or a category match of the slot, which is assigned a score of 1. The final score of a principle is the sum of all the scores of this principle. The length of a principle is used as the threshold to determine whether this principle is matched or not. Finally, the longest principle or a principle which contains the most slots will be considered as matched.

The last component is the BeCalm communication module. The module listens requests from BeCalm platform [5], check the correctness of the authentication key provided in each request, authorized the requests and then respond to BeCalm with an acknowledge message. All approved requests are sent to the first component for downloading patents from remote data sources. The download patents are then processed by the core of our service for miRNA recognition. Finally, the recognized miRNAs are encoded in the JSON format defined by the TIPS task and send to BeCalm through the saveAnnotations method provided by the BeCalm platform.

3 Results and Discussion



Patent server

Total processing time: 2 days 06h:0m10:s
Total predictions: 0 k
Mean predictions per request: 0
Exceptions/Total requests 7929 / 48263
Processing time (seconds):
Max time: 9601.85 Min time: 1.7
Mean time: 66.54

Abstract server

Total processing time: 2 days 03h:0m4:s
Total predictions: 0 k
Mean predictions per request: 0
Exceptions/Total requests 5079 / 39708
Processing time (seconds):
Max time: 7312.43 Min time: 1.97
Mean time: 24.67

PubMed

Total processing time: 04h:0m35:s
Total predictions: 0 k
Mean predictions per request: 0
Exceptions/Total requests 3159 / 35447
Processing time (seconds):
Max time: 5154.19 Min time: 2.75
Mean time: 45.61

Figure 2. Statistical results of our web service.

Figure 2 shows the official statistical results of our web services. As one can see that our server can successfully receive the request from BeCalm. Most requests were requested for data from the patent server. The results illustrate that our server did not generate any predictions for all of the three data sources. We tried to debug our server by store the downloaded data from the remote sources. In total of 818 patents were saved for our analysis. We then applied an offline processing for all the downloaded data. Table 1 shows the statistical results on the dataset. We can observed that our miRNA recognizer can recognize miRNAs from the patent documents. The zero predictions shown in Figure 2 seems owing to the failure of our communication module in replying the annotation results to the BeCalm platform.

Table 1. Offline recognition results.

# of Processed Patents	# of Predictions	# of Patents Containing Predictions
818	34	15

4 Conclusion

In the paper, we introduce our miRNA recognition web service developed for the BioCreative V.5 TIPS task. Although our miRNA recognizer can recognize miRNAs mentioned in patents, our service fails to response processing results to BeCalm. In the future, we will fix the defeats and manually annotate miRNAs observed in collected patents to evaluate the performance of our miRNA recognition in patent documents. We will also study the distributions of recognized miRNA mentions among PubMed/PMC articles and chemical patents.

5 Acknowledgment

The study was supported by Ministry of Science and Technology (MOST) 105-2221-E-143-003 and MOST-106-2922-I-038-014.