A Study on Identification of Organism and micro-RNA Mentions in Figure Captions

Nai-Wen Chang[1,2], Jitendra Jonnagaddala [3,4], Feng-Duo Wang[5], Hong-Jie Dai[5,6*]
[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3]School of Public Health and Community Medicine, UNSW Sydney, Australia
[4]Prince of Wales Clinical School, UNSW Sydney, Australia
[5]Department of Computer Science and Information Engineering, National Taitung University, Taitung, Taiwan
[6]Interdisciplinary Program of Green and Information Technology, National Taitung University, Taitung, Taiwan

***Abstract*—**The detection of organism mentions in scientific literature facilitates researchers with the ability to find relevant subsets of papers based on species-specific queries. Furthermore, most biological articles will describe pathways or regulation information in figure captions to enhance the understanding of experimental results. The extraction of miRNA and organism from figure captions is useful in characterizing the research studies. In this study, we adopted openly available organism recognition tools and our statistical principle-based miRNA recognizer for identifying organism and miRNA mentions in figure captions of an article. The miRNA recognizer is extended by generating scores for matched slots and indexes for matched terms to normalize recognized miRNAs to identifiers in the Rfam database. We study the performance of the existing tools in recognizing terms in figure captions and the challenges remained to address by evaluating them on the BioCreative VI Bio-ID dataset. We believe the Bio-ID corpus provide a nice starting point for evaluating the performance of miRNA normalization system. In the future, we would like to undertake more comprehensive evaluation of existing tools for extraction of organism/species and would like to enhance the consistency and comprehensiveness of miRNA annotations in the dataset.

Keywords*—biomedical text mining; organism; micro-RNA; statistical principle-based approach*

I. Introduction

The term organism is used to represent an important class of entities frequently mentioned in biomedical literature. Normalizing organism terms mentioned in literature to database records, such as NCBI taxonomy can be used for disambiguating biomedical entities such as mutations, proteins or genes [1]. Primarily, organism names are based on a well established hierarchical nomenclature conventions developed from the 18th century. However, the recognition of taxonomic groups in biomedical texts such as PubMed articles present a number of issues and challenges. Especially, there is a great degree of ambiguity in the way taxonomic information is expressed in biomedical literature. Large numbers of abbreviations of species names and use of common English names instead of Latin names are the primary reasons for this ambiguity. The use of acronyms, which can be both species specific and species independent, also pose a challenging problem for information extraction tasks. Lastly, incorrect spellings are often an issue with biomedical texts.

MicroRNAs (miRNAs) have become one of the hottest subjects in science and medicine recently. The first formal recognition of miRNAs was ten years ago. Since then miRNAs have been found to have a critical role in regulating many physiological processes and pathological processes. Numerous miRNAs and their associated targets have been identified by bioinformatics tools [2-4] and high-throughput sequencing [5-7]. Therefore, the demand of recognizing miRNAs mentioned in literature is increasing. miRNAs are evolutionary-conserved small non-coding RNA molecules that post-transcriptionally regulate gene expression by base-pairing to messenger RNAs (mRNAs). Many freely available, web-based miRNA-related database systems have been developed for researchers to retrieve miRNAs and their target genes. For instance, miR2Disease is a manually curated database, providing a comprehensive resource for miRNA deregulation in various human diseases [8]. miR2Disease provides researchers with information such as miRNA-disease relationships and experimentally verified miRNA-target genes, as well as references to the relevant biomedical literature. Similarly, the miRWalk database provides predicted and validated miRNA binding site information related to miRNAs in humans, mice and rats [9].

Information extraction methods can be employed to extract organism and miRNA related information. The identification of these two entities can facilitate taxonomy-aware information extraction in construction of valuable knowledge bases such as miRWalk and miR2Disease. In addition, these methods can also be used to enhance the index created by search tools for retrieving more relevant literature using species-specific and miRNA-related keywords. Furthermore, most biological articles will describe pathways or regulation information in figure captions to enhance the experimental results. The advantages of extracting miRNA and organism from figure captions can capture the most important and real data. With this in mind, we assessed the current automated information extraction methods available to extract organism information from figure captions

[*] Corresponding author

We also extended our previous miRNA recognition method [10] by developing a method for normalizing recognized terms to the Rfam database.

II. Methods

A. Dataset

We used the dataset released by Bio-ID track for our research purpose. The dataset was prepared as part of the EMBO SourceData project[2], which consisted documents in BioC [11] format with number of figure captions from full-length articles along with the annotations for multiple bio-entities. Organisms and miRNAs are two of the entity types annotated in this dataset and the annotations includes their spans in figure captions and their corresponding database IDs. This dataset was used to assess the performance of several current openly available tools. We employed these tools for extract organism/species information from figure captions to understand the complexity, issues and challenges involved in that.

B. Species Entity Recognition and Normalization

For recognition and normalization of organism/species entities we consider SR4GN [1], ORGANISM/SPECIES tool [12], and NCBO Annotator [13]. After studying their performance on the training set, we decided to use the SPECIES tool for processing the test set of the Bio-ID track.

The SPECIES tool identifies and normalizes species entities using dictionary look-up approach. This tool employs NCBI Taxonomy for dictionary look-up. A minor post-processing enhancement was developed in this work. The enhancement primarily involved selection of top 10 entities with highest frequency observed in the training set. Once these entities are identified, they are checked against the output (on the test set) from SPECIES tool where an entity is observed with no NCBI taxonomy ID assigned. Should there be any such entity, NCBI taxonomy ID is assigned based on the top 10 entities list prepared earlier. This process was mainly employed to improve the performance but also at the same time limit the number of false positives that may creep in because of this post processing enhancement.

C. miRNA entity Recognition and Normalization

Rule-based and machine learning-based approaches are two popular methods used in the task of miRNA recognition, which have been applied to public miRNA databases, such as miRCancer, miRSel [13, 14] and TarBase [15, 16]. However, rule-based approaches require explicit rules developed by domain experts, which are not flexible enough to cover all variations, such as the insertion, deletion or substitution (IDS) of words appearing in the entities, phrases or sentences. On the other hand, machine learning models can learn the implicit patterns automatically, but the model cannot be easily interpreted by humans. We have proposed the statistical principle-based approach (SPBA) for miRNA recognition to deal with the drawbacks of the above approaches [10]. In general, SPBA can automatically extract labeled sequences, combine them into more representative principles through the observation of dominated principles, and employ a partial matching algorithm to harness the advantages of both rule- and machine learning-based approaches while surpassing their limitations. The performance of the SPBA-based miRNA recognition is evaluated on the corpus annotated by Bagewadi, Bobić, Hofmann-Apitius, Fluck and Klinger [14]. The evaluation achieved a 0.988 F-score, 0.986 precision (P) and 0.991 recall (R), which outperformed the traditional rule-based methods. However, this method only considers the recognition task. In this study, we extended the existing method to support the normalization process that can link recognized miRNA mentions to identifiers in the Rfam database, a database of non-coding RNA families and other structured RNA elements. The details of SPBA is described in the following subsection.

D. Knowledge Construction for miRNA Recognition

Our SPBA-based method used a collection of principles generated from the training phase to match the content. If the content can be matched with a compiled labeled sequence, the corresponding entity is determined. The training phase of SPBA consists of three main steps. The first is knowledge construction. In case of miRNA recognition, we represent the knowledge related to miRNA through semantic slots and principles semi-automatically. Following is the principle generation step, in which slots are assembled and summarized by observing the arrangement of principle slots which can accomplish the target task. Lastly, a flexible principle matching algorithm allowing insertion/deletion/substitution is applied to extract the information represented by the compiled principles in unstructured text. Fig. 1 illustrates a simplified example of how the knowledge was constructed for representing a miRNA in SPBA under the principle-slot scheme. More precisely, the knowledge is constructed in a hierarchical structure.

In the knowledge representation of SPBA, the root node is usually the name of a domain or a subject. In Fig 1. the root node is "miRNA" indicating that the structure represents the knowledge for miRNA names. The first child node of a root node is usually the "SLOT" node, under which we can define the fundamental slot for miRNA. Albeit the heterogeneous writing styles, some common contents can be found among 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. Hence, they 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.

The last two slots can be further generalized into one slot, "[Suffix]", which can be used to differentiate distinctive types between miRNAs (e.g., has-let-7a-2-3p). Therefore, they are organized in a hierarchical structure as showed in Fig 1 (3: [Suffix] → SLOT → [Hair-pin]). For each slot, a list of terms that could be written in literature are collected and listed under that slot. For instance, the instances of the "[Species]" slot are

[2] http://sourcedata.embo.org/

tri-grams, such as “hsa” and “cel”. The indicating words for [miRNA] include “mir”, “let”, “lsy”, “micro RNA”, etc.

E. *Principle Matching*

During the principle alignment 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. The bigram frequency gives a way to assign the insertion scores, which usually fall in the ranges, -(∞), -2, -1, 0, 1, and 2. Deletions are assigned a score of -(∞), -2, -1, or 0. A substitution is either a partial match or a category match of the slot, which is usually assigned a score of 1 or 2. The final score of a principle is the sum of all the scores of this principle. The length of a principle, which means the number of slots 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. In other words, the principles will be ignored if the final score closes to -(∞).

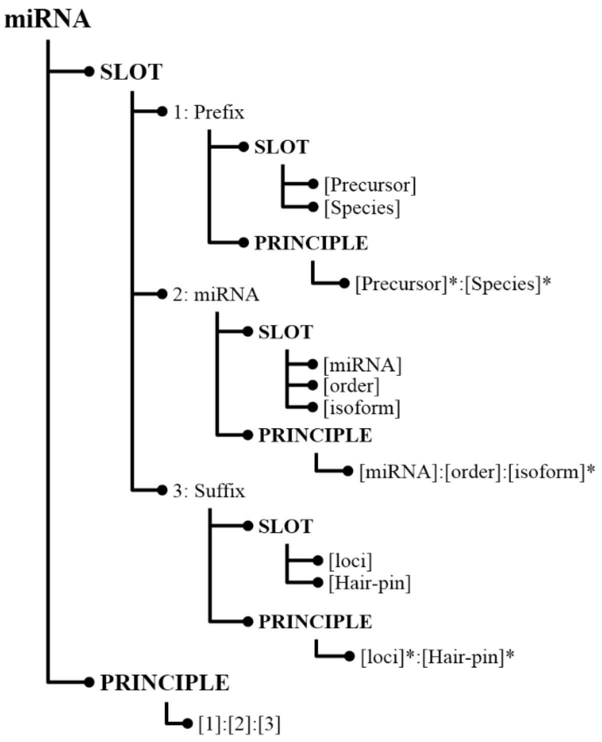

Fig. 1. Knowledge represented for miRNA in SPBA.

F. *Principle-based Normalization*

For normalization, we first downloaded the family file from ftp://ftp.ebi.ac.uk/pub/databases/Rfam/CURRENT/database_files. We extracted the following columns to compile the lexicon for normalization:

- The first column: the family accession number (e.g. RF00994).
- The second column: the family id (e.g. mir-1255)
- The fourth column: the family description (e.g. microRNA mir-1255)

We then used the generated principles to match all columns contained miRNA names. During the principle matching process, we scored the matched slots based on the matched principles over all entries in the compiled lexicon and built indexes for each slot. Therefore, each slot will associate with all possible corresponding grounding entries in our lexicon along with a matching score.

After the principle matching step, we were able to recognize possible miRNA mentions. For a recognized miRNA mention, the indexes of its matched slots were used to retrieve all possible grounding in the Rfam database. We then used the associated score to select the normalization ID.

III. Results

A. *Entiy Recognition Performance on the miRNA Interaction Corpus*

We first report the performance of the developed miRNA recognizer on the corpus annotated by Bagewadi, Bobic, Hofmann-Apitius, Fluck and Klinger [15]. This corpus only contains annotations for the spans of miRNAs that appeared in literature. Therefore, we can only estimate the entity recognition performance. They distinguished their annotations for miRNA into two types: specific-miRNA (*e.g.* has-miR-124b), and non-specific-miRNA (*e.g.* microRNAs or miRNAs). In our evaluation, we only focused on the specific-miRNA type.

TABLE I. Entiy Recognition Performance on the miRNA corpus

	Training set	**Test set**
Precision	0.994	0.986
Recall	0.9902	0.991
F-score	0.992	0.988

As shown in Table I, our tool can achieve a precision (P) of 0.994, recall (R) of 0.9902, and F-score (F) of 0.992 in the training set. Moreover, the performances in the test corpus are a precision of 0.986, recall of 0.991, and F-score of 0.988.

B. *Entity Normalization Performance on the Bio-ID corpus*

Table II shows the performance of normalization on the training set of the Bio-ID track. We report the performance in terms of micro-Precision, Recall and F-measure under the overlapping mode. We observed that the extended methods achieved recall of 0.865 on the training set but a very lower precision of 0.253 resulting in a frustrated F-score of 0.373 on the Bio-ID dataset. Although miRNA was annotated in the Bio-ID corpus, the test set was not. Thus, we cannot report our performance on the test set. After analyzing the errors on the training set, we observed that the majority of the errors are due to inconsistent annotations. For instance, U2 (Rfam:RF00004) mentioned several times in the Figure 4 of the article (PMC4801943) was not annotated in the corpus. However, our method recognized and normalized that entity.

Table II also shows the performance of the three off-the-shelf organism/species identification tools on the training and test sets. The performance was evaluated under overlap setting only for normalization component of the tools. We can observe that NCBO annotator had the best R but a very low P. The SPECIES tool achieves the best F-score. Therefore we selected SPECIES tool as a baseline system to study its performance on the test set. The SPECIES tool was employed under two different settings (Run 1 and Run 2). Under the Run 1 setting, the tool was executed on the Bio-ID corpus with default configuration. In Run 2, the post-processing enhancement described in the Methods section was applied. The performance of SPECIES tool in Run 1, with default configuration is better than Run 2, where the post-processing enhancement didn't impact R but decreased the overall F-score.

TABLE II. NORMALIZATION PERFORMANCE

Configuration	**Train set**			**Test set**		
	P	**R**	**F**	**P**	**R**	**F**
miRNA	0.253	0.865	0.373	n/a	n/a	n/a
NCBO Annotator	0.061	**0.920**	0.118	n/a	n/a	n/a
SR4GN	0.468	0.382	0.419	n/a	n/a	n/a
SPECIES-Run1	**0.670**	0.476	**0.557**	**0.677**	**0.580**	**0.625**
SPECIES-Run2	0.460	**0.481**	0.471	0.432	**0.580**	0.495

IV. CONCLUSION

In this study, we presented performance assessment of miRNA and organism information extraction tools with focus on normalization aspect, using two different datasets. We believe the results presented in this study provide a good starting point for evaluating the performance of miRNA and organism entity recognition and normalization system. In future, we would like to improve the manual annotations in these two datasets. Specifically, we would like to enhance the consistency and comprehensiveness of the miRNA annotations. Additionally, we would also like to undertake more comprehensive evaluation of existing tools for extraction of organism/species and miRNA related information.

ACKNOWLEDGMENT

This study was conducted as part of the electronic Practice Based Research Network (ePBRN) and Translational Cancer research network (TCRN) research programs. ePBRN is funded in part by the School of Public Health & Community Medicine, Ingham Institute for Applied Medical Research, UNSW Medicine and South West Sydney Local Health District. TCRN is funded by Cancer Institute of New South Wales and Prince of Wales Clinical School, UNSW Medicine. This study was also supported by the Ministry of Science and Technology of Taiwan [grant numbers MOST-105-2221-E-143-003 and MOST-106-2221-E-143-007-MY3].

REFERENCES

[1] C.-H. Wei, H.-Y. Kao, and Z. Lu, "SR4GN: A Species Recognition Software Tool for Gene Normalization," *PLoS ONE,* vol. 7, no. 6, pp. e38460, 2012.

[2] D. M. Garcia, D. Baek, C. Shin, G. W. Bell, A. Grimson, and D. P. Bartel, "Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs," *Nat Struct Mol Biol,* vol. 18, no. 10, pp. 1139-46, Oct, 2011.

[3] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks, "MicroRNA targets in Drosophila," *Genome Biol,* vol. 5, no. 1, pp. R1, 2003.

[4] A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky, "Combinatorial microRNA target predictions," *Nat Genet,* vol. 37, no. 5, pp. 495-500, May, 2005.

[5] S. Baskerville, and D. P. Bartel, "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes," *RNA,* vol. 11, no. 3, pp. 241-7, Mar, 2005.

[6] M. V. Iorio, M. Ferracin, C. G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, S. Menard, J. P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G. A. Calin, P. Querzoli, M. Negrini, and C. M. Croce, "MicroRNA gene expression deregulation in human breast cancer," *Cancer Research,* vol. 65, no. 16, pp. 7065-7070, Aug 15, 2005.

[7] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey, "Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding," *Cell,* vol. 153, no. 3, pp. 654-665, Apr 25, 2013.

[8] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Res,* vol. 37, no. Database issue, pp. D98-104, Jan, 2009.

[9] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "miRWalk--database: prediction of possible miRNA binding sites by "walking" the genes of three genomes," *J Biomed Inform,* vol. 44, no. 5, pp. 839-47, Oct, 2011.

[10] N.-W. Chang, H.-J. Dai, Y.-L. Hsieh, and W.-L. Hsu, "Statistical Principle-based Approach for Detecting miRNA-target Gene Interaction Articles," in Proceeding of the IEEE 16th International Conference on BioInformatics and BioEngineering (BIBE), Taichung, Taiwan, 2016.

[11] D. C. Comeau, R. Islamaj Dogan, P. Ciccarese, K. B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, A. Valencia, K. Verspoor, T. C. Wiegers, C. H. Wu, and W. J. Wilbur, "BioC: a minimalist approach to interoperability for biomedical text processing," *Database (Oxford),* vol. 2013, pp. bat064, 2013.

[12] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen, "The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text," *PLoS ONE,* vol. 8, no. 6, pp. e65390, 2013.

[13] C. Jonquet, N. H. Shah, and M. A. Musen, "The Open Biomedical Annotator," in AMIA Summit on Translational Bioinformatics, San Francisco, CA, USA, 2009, pp. 56-60.

[14] S. Bagewadi, T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger, "Detecting miRNA Mentions and Relations in Biomedical Literature," *F1000Research,* vol. 3, no. 205, 2014.

[15] S. Bagewadi, T. Bobic, M. Hofmann-Apitius, J. Fluck, and R. Klinger, "Detecting miRNA Mentions and Relations in Biomedical Literature," *F1000Res,* vol. 3, pp. 205, 2014.