NTTMUNSW BioC Modules for Recognizing and Normalizing Species and Gene/Protein Mentions in Full Text Articles

Onkar Singh[*1], Jitendra Jonnagaddala[2], Hong Jie Dai[*3], Emily Chia-Yu Su[*1]

[1]Graduate Institute of Biomedical Informatics,
College of Medical Science and Technology, Taipei Medical University, Taiwan.
[2]School of Public Health and Community Medicine, The University of New South Wales Australia
[3]Department of Computer Science and Information Engineering, National Taitung University, Taiwan.

[1]{m610103011, emilysu}@tmu.edu.tw
[2]z3339253@unsw.edu.au
[3]hjdai@nttu.edu.tw

Abstract. In recent years, the intensity of published biomedical literatures is increasing due to vast majority of researches being focused on biological domain to investigate the functions of biological objects, such as genes and their products. However, the ambiguous nature of genes and proteins makes literature more complex for readers and curators for molecular interaction databases such as BioGRID and IntAct. To address the ambiguity problem, biomedical researchers are inclined towards applying the gene normalization (GN) technique of text mining. In this work, we developed three BioC-compatible modules with the purpose to assist the curation process of BioGRID curators. First, a species recognition module was developed, which can identify species names and normalize them to NCBI Taxonomy IDs. The recognition results of the module include the full mentioned species terms and the designated symbol of a gene name that refers to an abbreviation of a species name. For GN, two modules were implemented based on our multi-stage normalization system developed for processing full-text articles in BioCreative II.5. The first module recognizes gene mentions and the second module normalizes them to their Entrez Gene IDs by utilizing the characteristics of different paper sections. All of the developed BioC-compatible modules were implemented using Microsoft .NET and will be openly available at NuGet gallery.

* Corresponding author

1 Introduction

At present, the great majority of researchers are interested in life science with aims of exploring biological processes, cause and their associated objects. Most of these biological processes are genes or proteins dependent, which propel researchers to collect information related to genes and gene products that can assist researchers in gaining advanced perception of the complex mechanisms behind biological phenomena. As a result, a large number of biomedical literature based on gene/protein functions are published every year. Therefore, the ability to acquire the timely and update-to-date information of genes/proteins cited in the large volume of biomedical literature attract the attention of biologists. To this end, data mining researchers are developing text mining techniques to extract high quality information from biomedical literature. The gene/protein normalization (GN) technique [1] facilitates the above process by first automatically recognizing genes and proteins mentioned in biomedical literature and then determining their database identifiers, such as the Entrez Gene IDs, to create a linkage between literature-recorded genes/protein and their corresponding biological databases. One of the major challenge encountered in GN is the disambiguation of candidate gene IDs because of the presence of orthologous genes in cross species. Therefore, the accurate recognition of species could provide important information that is helpful for GN as well as many downstream tasks such as protein-protein interaction [2].

On the other hand, text mining techniques like species recognition or GN relies on a varieties of different corpora and natural language processing (NLP) modules. To facilitate the interoperability among those resources, the BioC format [3] was proposed to establish a simple XML-based format to represent, store and exchange the data among different text mining systems. The collaborative biocurator assistant task of the BioCreative V workshop makes an attempt to integrate all state-of-the-art text mining modules into one annotation tool by providing full text datasets encoded in the BioC format and several useful BioC-related NLP utilities and resources. This work complements the task by developing three BioC-compatible modules for processing the full text articles represented in the BioC format, and generate the annotation results for species and gene/protein names along with their NCBI Taxonomy IDs and Entrez Gene IDs. In contrast to most of the previous released species recognition tools [4, 5] which only recognize the fully mentioned species terms, such as human in the gene name "human brain 25 kDa lysophospholipid-specific lysophospholipase", the developed species recognition module can recognize and normalize the designated symbol in the prefixes of a gene name that refers to an abbreviation of a species name. For example, the "h" symbol in "hLysoPLA" will be normalized to the taxonomy ID 9606. For recognizing and normalizing gene mentions, our multistage GN system [6] developed for processing full-text articles in BioCreative II.5 was ported to support the processing of the BioC format and normalize gene mentions to Entrez Gene IDs. All of the modules developed in this work are implemented using the .NET framework and will be made openly on the NuGet Gallery.

2 Material &Methods

2.1 Overview of the Developed Modules

The workflow of the developed modules is shown in Fig. 1. At first, the BioC C# implementation ported from the BioC-Java implementation[†] was used to read the given full-length article encoded in BioC format. The BioC-C# developed by our group is available at https://www.nuget.org/packages/BioC/. For each passage, the following NLP pipeline was employed: sentence breaking, tokenization, part-of-speech (PoS) tagging [7] and abbreviation recognition [8]. Based on the generated linguistic information, the gene/protein mention recognizer was employed to recognize gene/protein mentions. The recognized gene/protein mentions along with the linguistic information were set as the input for the developed species recognizer for identifying species terms. Finally, all the above information was processed by the multistage GN module to normalize all of the recognized gene/protein mentions to their corresponding Entrez Gene IDs.

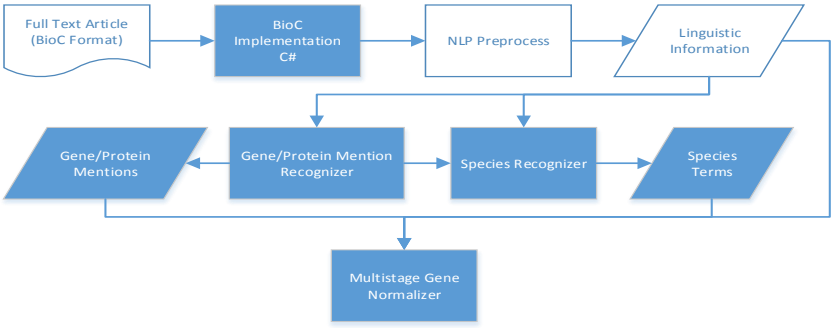

Fig. 1. The workflow of the developed modules.

2.2 Gene/protein Mention Recognizer

This module recognizes gene mentions from the input source and matches the recognized gene mentions against a lexicon consisting of gene name variations generated by several rules and the corresponding Entrez Gene IDs. For recognizing gene mentions, our NERBio system [9] developed based on the BioCreative II gene mention recognition corpus [10] was used. For gene mention matching, we implemented an exact matching strategy to get a set candidate Entrez Gene IDs against all of the variations of the lexicon.

After matching, the following refinement process is executed. First, all of the recognized genes which cannot be matched with at least one ID were filtered out form the output. The names of all successfully matched gene mentions are then collected to generate a refinement dictionary. Finally, refinement is performed by using the exact

[†] http://sourceforge.net/projects/bioc/files/BioC_Java_1.0.1.tar.gz/download

matching algorithm to search the whole article for mentions in this dictionary, which were not recognized by the recognizer.

2.3 Species Recognizer

This module extracts organism/species information from the input source by employing a flexible string searching algorithm. The species dictionary created by Pafilis et al. [5] was used for the partial matching. All of the records in the dictionary were preprocessed to generate variations. After identified all possible species terms, the PoS information is used to filter out false positive cases. For example, the term whose PoS information indicates it as a verb is removed. In order to recognize the symbol within a gene name that refers to a species name, two results associated for the given full text article generated by the preceding modules were used. One is the recognition results of the gene/protein mention recognizer. The other is the full name-abbreviation results generated in the NLP pipeline. First of all, a list of 1-and 2-letter organism codes was collected. For example, "h" and "Zm" stands for the taxonomy ID 9606 and 381124, respectively. The 3-letter codes were collected from the KEGG organism website. The species recognizer was then check the prefixes of each recognized gene mention with the compiled *n*-letter organism codes. If a match is found, the species recognizer checks whether or not a full name was defined for the current gene mention. If there is no full name observed, the matched prefixes and the corresponding ID are output. Otherwise, the observed full name is further matched with the organism name represented by the prefixes. If the full name matches with the organism name, the prefix symbol and its corresponding ID are output, otherwise the prefix symbol is filtered.

2.4 Multistage Gene Normalization

The multistage GN module uses the algorithm specifically developed for full text article processing to normalize gene mentions to their Entrez Gene IDs. The multistage normalization algorithm was developed based on the idea that the research article structure, in the majority of cases, follows basic principles to assign information accurately to each section. Each section of the paper has different characteristics which can be used to guide normalization algorithm. For example, the Introduction often contains information that repeatedly appears throughout the article (key genes), while the Results section presents new scientific findings. Normalizing a gene mention from the Results section may require resolving an acronym whose full name or its associated species has only been mentioned in the Introduction.

In order to use the characteristics of difference sections recorded in the BioC-encoded full text article, a utility C# class, BioC AsciiKeyReader, was developed to transform the BioC-XML file into C# object that enables the multistage GN module can access the full text content in an arbitrary sequence based on the section headings. The module then perform the normalization process from the Introduction section, which usually has the richest context information, to those with the poorest, including Figures, Tables and footnotes. IDs normalized in earlier parts are therefore can be used to help GN in later parts. The entire algorithm is divided into the following three stages.

In the first stage, GN is executed in the following order: Introduction, Discussion, and Abstract. Successfully normalized IDs are kept in memory for use in subsequent sections. Following the above order, certain disambiguation rules, such as the history and the full-name/acronym rules, can be more effective. Take the history rule for example. The rule aggregates all successfully normalized IDs before processing the target gene mention. For all of the ambiguous IDs of the target gene mention, the rule select the one occurred most frequently in the aggregated IDs as the disambiguated ID. If we process the article in a linear order of the full text (Title→Abstract →Introduction), the History rule cannot disambiguate the first gene mention when processing the Title. Because of the lack of context information, other rules also tend to fail.

In the second stage, the successfully normalized gene mentions and corresponding IDs are extracted to generate a dictionary. A dictionary-based tagger is executed to search the whole article for mentions in this dictionary. The tagger also checks the species information generated by the species recognizer. If species are found and matched with the corresponding ID's species, the ID is assigned. Otherwise, all of matched ambiguous IDs are assigned to the gene mention. This assignment can boost the processing speed of the multistage GN because in next time when the algorithm tries to normalize the gene mention, the algorithm doesn't need to match the gene mention to generate its IDs. However, the dictionary-based tagger may generate a boundary that is inconsistent with the one recognized by the gene/protein mention recognizer. In this case, if the normalized ID is determined, the normalized ID is set to the recognized gene mention. Otherwise, both boundaries are reserved.

The remaining paper sections are processed by the multistage GN module in the final stage. Because the process in stage 2 or the input BioC-encode articles may contain overlapped gene mentions, all gene mentions that have not been associated with candidate IDs were normalized and compared with the overlapped ones.

Disambiguation. As there can be multiple IDs for a given gene/organism pair, the role of disambiguation is used to resolve the ambiguity based on contextual relevance. For example, Entrez search for gene "DDX39B" and organism "Homo Sapiens" give 74 results (Fig. 2).

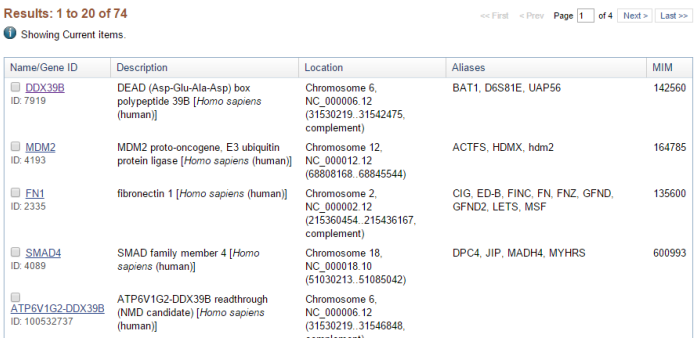

Fig. 2. Example of Ambiguous Genes

This ambiguity can be resolved by adding contextual information surrounding the gene mentions like chromosome number, loci information, molecular function, disease caused by mutation of the gene etc. The multistage GN module developed in this work used some of the rule proposed in our previous work [6] for disambiguation. However, a gene mention may still not be able to disambiguate because of the lack of contextual information. Therefore, the source of the matched record in the Entrez Gene record was also considered in this work. In this case, the ID whose official symbols or names matches with the target gene name is selected. For example, if the current ambiguous gene mention has the name “DDX39B”, we can observe that “DDX39B” is the official symbol for Entrez Gene ID 7919. The ID is therefore selected.

3 Results and Future Work

3.1 Datasets

Because the collaborative biocurator assistant task have not released the gold annotations for the processed full text dataset. This work used the instance-level GN (IGN) corpus [11] released in our previous work to give a preliminary evaluation result of the developed species recognizer modules. The original IGN corpus contains instance-level annotations for human genes. In order to develop and evaluate the developed modules, the corpus was updated to include annotations of cross-species genes and species mentions. More specifically, the annotations contain the exact occurrence information of all described gene/species mentions and their Entrez Gene/NCBI Taxonomy IDs. The designated symbol of a gene name that refers to an abbreviation of a species name is also annotated as a species term. For instance, the organism symbols “h”, “Hs”, “Sc” and “Ca” for the gene mentions “hLysoPLA”, “HsUap1p”, “ScUAP1”, “CaUap1p” were annotated, which represented “Homo Sapience”, “Saccharomyces Cerevisiae”, and “C. albicans” respectively. Recognizing the above symbols appeared in gene mentions could improve the performance of GN systems. Fig. 3 shows an annotated example.

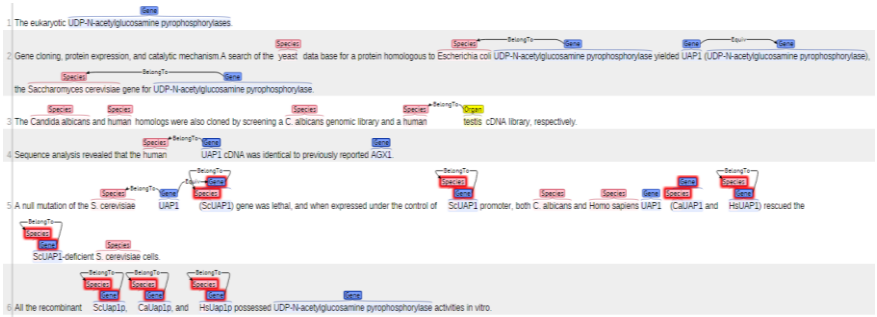

Fig. 3. The annotations for the article PMID: 9603950.

The developed species recognizer achieved a satisfied precision/recall/F-measure of 0.980/0.917/0.948 and 0.852/0.864/0.858 on the IGN training and test set, respectively. Through the error analysis, we observed that some prefix characters of gene mentions

were frequently incorrectly identified as a species term, such as "Ca" stands for "calcium" in "CaMKII" and CaBP1". Some false positives come from the results of our species term variation generation process. For example, "kcat" was recognized for taxonomy ID 9685, which refers to "Korat cats" because the variation "kcat" was generated by removing white space from the abbreviated term "K cat". We also observed that the developed module failed to recognize some abbreviated terms defined by authors. For example, the term "AMV" in the sentence "The myb gene is the transforming oncogene of the **avian myeloblastosis virus** (**AMV**)".

The performance of the developed gene/protein mention recognizer and multistage GN modules have been reported in our previous works [6, 9]. In the future work, we would like to use the extend IGN corpus and the full text dataset annotated by the BioGRID curators to evaluate the impact of the newly proposed rules implemented in this work and study the distribution of the different contextual information appeared in the sections of full text article to adjust and provide statistical support to our multistage algorithm.

4 Acknowledgment

This work was supported by the Ministry of Science and Technology of Taiwan (MOST-104-2221-E-143-005)

REFERENCES

1. Hirschman L, Colosimo M, Morgan A, Yeh A: **Overview of BioCreAtIvE task 1B: normalized gene lists**. *BMC Bioinformatics* 2005, **6**(Suppl 1):S11.
2. Wei CH, Kao HY, Lu Z: **SR4GN: a species recognition software tool for gene normalization**. *PloS one* 2012, **7**(6):e38460.
3. Comeau DC, Batista-Navarro RT, Dai H-J, Islamaj Doğan R, Jimeno Yepes A, Khare R, Lu Z, Marques H, Mattingly CJ, Neves M *et al*: **BioC interoperability track overview**. *Database* 2014, **2014**.
4. Gerner M, Nenadic G, Bergman CM: **LINNAEUS: A species name identification system for biomedical literature**. *BMC Bioinformatics* 2010, **11**:85.
5. Pafilis E, Frankild SP, Fanini L, Faulwetter S, Pavloudi C, Vasileiadou A, Arvanitidis C, Jensen LJ: **The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text**. *PLoS ONE* 2013, **8**(6):e65390.
6. Dai H-J, Lai P-T, Tsai RT-H: **Multistage Gene Normalization and SVM-Based Ranking for Protein Interactor Extraction in Full-Text Articles**. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS* 2010, **7**(3):412-420.
7. Tsuruoka Y, Tateishi Y, Kim J-D, Ohta T, McNaught J, Ananiadou S, Tsujii Ji: **Developing a robust part-of-speech tagger for biomedical text**. *Lecture notes in computer science* 2005:382-392.
8. Schwartz A, Hearst M: **A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text**. *Proceedings of the Pacific Symposium on Biocomputing (PSB)* 2003:451 - 462.

9. Tsai RT-H, Sung C-L, Dai H-J, Hung H-C, Sung T-Y, Hsu W-L: **NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition**. *BMC Bioinformatics* 2006, **7**(Suppl 5):S11.
10. Smith L, Tanabe LK, Ando RJn, Kuo C-J, Chung I-F, Hsu C-N, Lin Y-S, Klinger R, Friedrich CM, Ganchev K *et al*: **Overview of BioCreative II gene mention recognition**. *Genome Biology* 2008, **9**(Suppl 2):S2.
11. Dai H-J, Wu JC-Y, Tsai RT-H: **Collective Instance-level Gene Normalization on the IGN corpus**. *PLOS ONE* 2013.