Database, 2022, 1–7
DOI: https://doi.org/10.1093/database/baac090
Original article

Assigning species information to corresponding genes by a sequence labeling framework

Ling Luo[†], Chih-Hsuan Wei[†], Po-Ting Lai, Qingyu Chen, Rezarta Islamaj and Zhiyong Lu[*]

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), 8600 Rockville Pike, Bethesda, MD 20894, USA

[*]Corresponding author: Tel: +301 594 7089; Fax: +301 480 2288; Email: zhiyong.lu@nih.gov
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

Citation details: Luo, L., Wei, C., Lai, P. *et al.* Assigning species information to corresponding genes by a sequence labeling framework. *Database* (2022) Vol. 2022: article ID baac090; DOI: https://doi.org/10.1093/database/baac090

Abstract

The automatic assignment of species information to the corresponding genes in a research article is a critically important step in the gene normalization task, whereby a gene mention is normalized and linked to a database record or an identifier by a text-mining algorithm. Existing methods typically rely on heuristic rules based on gene and species co-occurrence in the article, but their accuracy is suboptimal. We therefore developed a high-performance method, using a novel deep learning-based framework, to identify whether there is a relation between a gene and a species. Instead of the traditional binary classification framework in which all possible pairs of genes and species in the same article are evaluated, we treat the problem as a sequence labeling task such that only a fraction of the pairs needs to be considered. Our benchmarking results show that our approach obtains significantly higher performance compared to that of the rule-based baseline method for the species assignment task (from 65.8–81.3% in accuracy). The source code and data for species assignment are freely available.

Database URL: https://github.com/ncbi/SpeciesAssignment

Introduction

With the rapid growth of biomedical literature, automatically extracting and summarizing the knowledge in the literature becomes increasingly important to biomedical research in areas such as biocuration assistance (1, 2) and COVID-19 research (3). The gene is one of the most important key concepts in biomedical research and is relevant to genetic variation, pharmacogenomics, cancer research and precision medicine. Many text-mining studies (4–6) rely highly on the automatic extraction of gene names in the text. Because multiple genes may share the same name, mapping gene names to unique concept identifiers is a very important step. National Center for Biotechnology Information (NCBI) Gene is a major database for gene records. Many studies (7–9) focus on mapping the gene names to the gene identifiers, and this task is widely known as 'gene name normalization (or gene linking)'.

The ortholog gene in different species, however, is associated with different gene identifiers, which exacerbates the difficulty of the gene normalization task. One means to decrease ambiguity is to identify the corresponding species of each gene mention (termed 'species assignment') and aim to narrow down the candidates for the possible gene identifiers. Few existing tools (10–12), however, have sufficient accuracy in regard to the species name recognition task, and none was designed to disambiguate the genes to the corresponding species. Furthermore, few studies have developed rule-based methods (13–15) to disambiguate the corresponding species of the gene, based on co-occurrence in the same sentence or paragraph. The most popular strategies to determine the corresponding species include the use of (i) the most nearby species of the gene in the same sentence, (ii) the most frequent species in the same paragraph, (iii) the corresponding species of the gene prefix that is represented (e.g. 'hCB1R' to human) and (iv) the species in the title. One of our previous works, SR4GN (15), was designed to recognize the corresponding species by leveraging the rules noted above to obtain good performance. SR4GN was successfully embedded in a previous gene tagger, GNormPlus (7), and applied to the entire PubMed and PMC (6) databases for gene recognition and normalization. SR4GN, however, frequently failed to find the correct corresponding species of the gene mentions when no species was mentioned within the same sentence. In addition, SR4GN only utilized a subset of gene-related species, accounting for a small portion (<10%) of the entire NCBI taxonomy repository.

As we learned from previous studies, two main challenges remained. (i) Most of the existing species recognition methods were not designed for gene normalization. In particular, some species-sensitive concepts (e.g. cell line and species strain) that are also helpful to species assignment and gene normalization were ignored. (ii) Although the rule-based method used to assign the species to gene mentions is straightforward, it does not work when the corresponding species is not in proximity

Received 10 July 2022; Revised 26 August 2022; Accepted 11 October 2022

of the gene. This is especially the case when an article mentions multiple species as it increases the difficulty of the task. For example, when a study uses multiple animal models to observe the expression of the human gene under defined criteria, this can cause difficulties even for manual assignment. To address the problem, we propose a novel species assignment approach based on deep learning via a sequence labeling framework. To the best of our knowledge, this work is the first that explores deep learning methods to assign species to gene mentions. The main contributions of our work can be summarized as follows:

- We develop a dictionary-based species tagger with state-of-the-art performance (94.3% in *F*-measure).
- We explore machine learning-based methods for the species assignment task. Instead of the traditional binary classification framework, we propose a novel species assignment approach based on a sequence labeling framework. We apply cutting-edge biomedical pretrained language models (PLMs) (i.e. PubMedBERT (16) and Bioformer (17)) for both frameworks and improve the performance from 65.8% to 81.3% in accuracy.
- We comprehensively compare the binary classification framework-based deep learning method, sequence labeling framework-based deep learning method and existing rule-based method in the species assignment task.

Methods

The workflow of the automatic species recognition and assignment

To address the two challenges, we proposed a new species tagger and a deep learning framework to optimize the performance of the species assignment. Figure 1 shows the architecture of our method, with end-to-end steps including species recognition and assignment. The input data require free text with highlighted gene mentions. To precisely evaluate the performance of the species assignment but not the effect of another component, we provide the manually annotated gene mentions in the input. The species recognition first identifies the species mentions with the corresponding identifiers (NCBI taxonomy ID); then, the species assignment links the corresponding species to gene mentions. Additional details are presented in the following section.

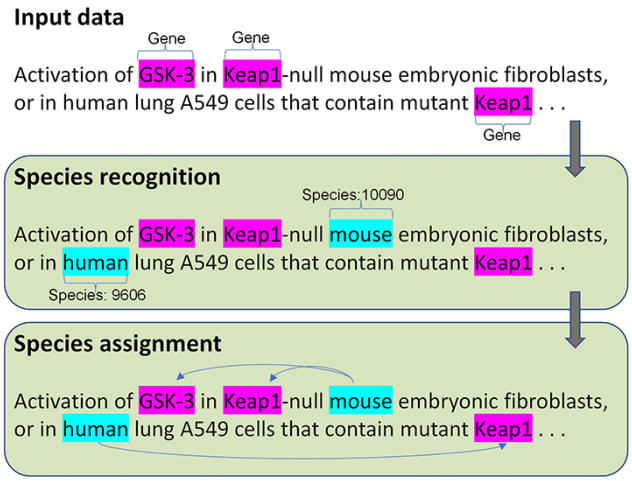

Figure 1. The architecture of the species recognition and assignment.

Species recognition

Machine learning-based named entity recognition methods always achieve much better performance than the dictionary-based methods on most bioconcepts (i.e. gene, disease and chemical). Nevertheless, the performance of species recognition, using dictionary-based methods, is competitive with machine learning-based methods (18). This is because the main challenge of species recognition is not term variation or ambiguity, as most of the species names in the text follow the nomenclature of the Carl Linnaeus standard system (19) and the species names in the text are standardized. Rather, the volume of the species taxonomy is critical. More than 2 million unique species (>16 million species names) are recorded in NCBI taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) as of 2022. The number of species names is larger than that of other popular bioconcepts (e.g. disease and chemicals). Such a supervised learning method may not be able to maintain coverage of a large-scale data set without support from a species lexicon. In addition, although species names are required to be linked to concept identifiers (NCBI taxonomy IDs), none of the existing supervised learning-based species taggers can map the species names to the specific concept identifiers.

Thus, we generated a dictionary-based species tagger using the species names recorded in the NCBI taxonomy repository. Based on the hierarchical structure of the taxonomy system, our tagger can better handle the large size of the species lexicon in an efficient way. More specifically, our species tagger was implemented by adopting a prefix tree to reorganize the species names within a highly efficient structure for a string search. In addition, such a structure maximizes the capacity to recognize name variations and the strain prefixes of the primary lesser ranks (e.g. 'str' and 'substr' in 'E. coli str. K-12 substr. MG1655' presented the 'strain' and 'sub-strain'). Every node in the prefix tree is a token of the species name. Every species name is recorded in the prefix tree as a path. The corresponding taxonomy id is stored in the terminal node of the path. For example, Tax:562 is stored in 'coli', which is the terminal node of 'E. coli'. The children of a node are the next words in the species name. Thus, the same token in different species names shares the same node. For example, 'Escherichia' and 'coli' are the shared prefix nodes of 'Escherichia coli K-12' and the 'Escherichia coli BL21'. 'K-12' and 'BL21' are two individual child nodes under the node 'coli'. As shown in the species name recognition module in Figure 2, the structure of the prefix tree-based dictionary perfectly presents a corresponding structure to the taxonomy hierarchical system.

Due to the flexible design of the dictionary, our tagger can also solve several common term variations. (i) The strain prefixes (e.g. 'str.' and 'substr.') can be simply recognized and skipped. Thus, there is no need to have a separate branch for the same species name with strain prefix. This not only saves space on the tree but also decreases search complexity. (ii) The longer species name can be more easily recognized than that of another species with a shorter name. For example, 'E. coli strain O157:H7' would be retrieved from the text, 'Using a similar approach, we show that E. coli strain O157:H7

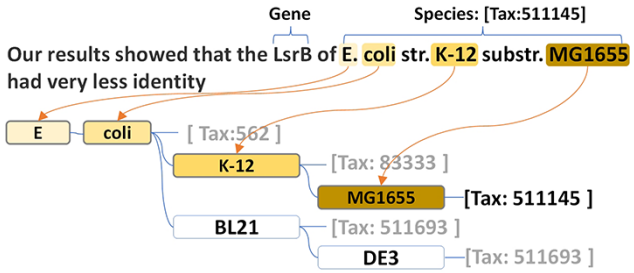

Figure 2. The species name prefix tree and the name recognition in the text.

Stx prophage or prophage remnants invariably include paralogs of nanS often located downstream of the Shiga-like toxin genes' (PMID: 27481927), rather than 'E. coli'. In general, the biomedical literature uses abbreviations to represent the concept name, with multiple tokens, but this is rare for species. Instead, genus names and strain codes are frequently used to represent parts of the species names. The genus name 'Arabidopsis' represents 'Arabidopsis thaliana' in PMID: 31279220, and the species strain code (e.g. 'MG1655') can represent the specific strain of the species (e.g. 'E. coli str. k-12 substr. MG1655'). To handle the case, we built a mapping from the genus names and strain codes to their corresponding species. Once a species (e.g. 'E. coli') is found in the text, the tagger searches the strain codes and the genus names via the species hierarchical system.

In addition, a frequent case was observed in our previous work (6). Shortening the genus name by using the first capital letter (e.g. 'Escherichia coli' to 'E. coli') to represent the species occurs frequently. Sometimes, however, the genus is abbreviated by the first two letters (e.g. 'Aedes aegypti' to 'Ae. aegypti'). We expanded the species lexicon to cover such cases. We also recognize the cell line in the text as it usually represents the animal model *in vitro* and *in vivo*.

Species assignment

The traditional rule-based method of species assignment relies highly on the species that are mentioned together in the surrounding context. The species, however, may not be in proximity of the gene. In addition, the case of multiple surrounding species is confusing in terms of the detection of the corresponding species of the gene name.

We deal with this task as a relation extraction between gene and species and establish a supervised machine learning-based method using biomedical transformer-based PLMs (e.g. PubMedBERT (16) and Bioformer (17)) for this task. As an encoder to represent the input text, PLMs can measure the relevance between tokens (e.g. gene and species), which is then applied to various biomedical text-mining tasks and can significantly surpass state-of-the-art performance. A straightforward framework that can recognize the corresponding species to the gene spans is the binary classification, which classifies each pair of gene and species. A positive outcome means that the gene corresponds to the species, and a negative one means that it does not. The binary classification method, however, is required to process all the pairs between species and gene names, one by one, which is time-consuming, and it is difficult to handle large-scale data using advanced deep learning techniques. Moreover, these methods ignore the dependency between entity pairs, as it deconstructs the task into multiple independent entity pair classification subtasks.

Inspired by several previous works on relation extraction (20–22), we proposed a novel species assignment method based on the sequence labeling framework. As shown in Figure 3, we converted the task to a sequence labeling problem. Given an input text with the target entity (e.g. species entity of 'mouse'), the goal of the model is to recognize all corresponding genes (e.g. 'GSK-3', 'Keap1' and 'phosphoinositide 3-kinase-protein kinase B') at once. Therefore, the speed of the sequence labeling framework is significantly faster than that of the binary classification framework.

Two strategies to predict the corresponding species for gene mentions include (i) targeting the species to reach its belonging genes (S→G; species to gene) and (ii) targeting the genes to reach the corresponding species (G→S; gene to species). S→G is much more efficient than is G→S, about seven times faster, as the number of species is usually less than the gene mentions in the input text. In addition, S→G is slightly more accurate than is G→S. Next, we use the strategy of S→G to present the details of our sequence labeling framework. Specifically, to distinguish the gene and species from other tokens in the text, we inserted a pair of tags '<GENE>' and '</GENE>', in front of and at the end of the genes, and '<SPEC>' and '</SPEC>' in the same way for the species. In each iteration, we sequentially selected a species and assigned a pair of tags ('<ARG>' and '</ARG>') to distinguish the target species from the others. We further translated the tokens of the input text into a sequence within two statuses: 'I' (inside) and 'O' (outside), as the predicted sequence. The example in Figure 3 shows the architecture of our model. The genes that correspond to the target species (including the surrounding tags) are in 'I' status, and other tokens and the genes that do not correspond to the target species are in 'O' status. In the example, the input document contains two species (i.e. 'mouse' and 'human') and four gene mentions. The predicted sequence labels a gene to 'I' status, indicating that the gene corresponds to the target species ('mouse'). At the end of the model, we used the SoftMax classification layer to summarize the probability scores of the labels of each token. We applied PubMedBERT (16) and Bioformer (17) as the PLMs models. PubMedBERT is the biomedical version of BERT and was recently created using only a biomedical vocabulary and data sets without transfer learning. Bioformer is a lightweight version of the traditional BERT model, which has been successfully applied in the biomedical domain. In most cases, each gene should be assigned to one species. Thus, we assigned the species with the highest predicted score to the genes, unless the two species with the highest score were in conjunction with each other (e.g. 'human and mouse cDNAs of ABCB9'). We defined a simple regular expression to detect the species within the conjunction and assign both taxonomy IDs to the gene name. In addition, we assigned the most frequent species to the genes when no corresponding species could be reached by the model.

Results

Evaluation of the species recognition

To better understand the performance of our species tagger, we first compared it with two other species taggers [i.e. Linnaeus (10) and SPECIES (12)] by their proposed corpora. Although many other species taggers (18, 23–25) obtained

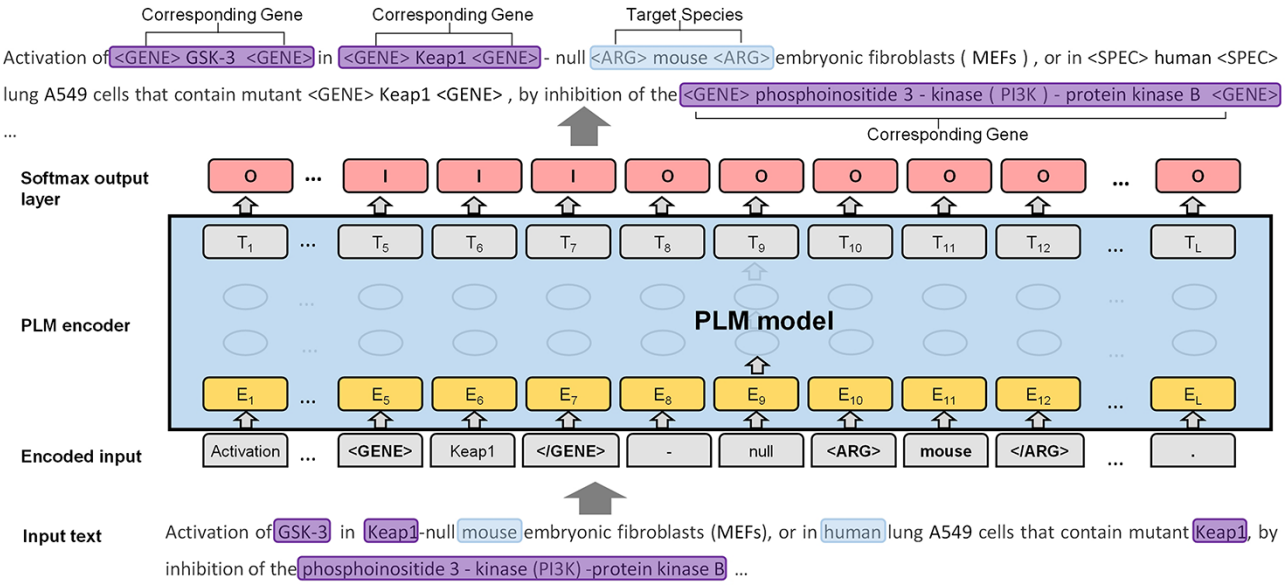

Figure 3. The formulated labeled sequence and the PLM model.

Table 1. Species normalization performance on Linnaeus and SPECIES corpora

Corpus	No. of articles	No. of species	Our tagger	Linnaeus	SPECIES
Linnaeus	100 full texts	2851	94.3%	90.3%	92.0%
SPECIES	800 abstracts	3708	82.7%	79.6%	77.8%

The performances are in *F*-measures (the harmonic mean of the precision and recall)

good performance with regard to the recognition of species boundary mentions, they did not address the normalization of the species. In fact, the normalization of species mentions for the species concepts (i.e. NCBI taxonomy IDs) is important for the species assignment task. Table 1 shows the performance of recognizing the species concept identifiers of individual taggers on Linnaeus and SPECIES corpora. The evaluation is at the document level, which means that, when one species appears multiple times, it should be counted only once. Our tagger attained the best performance as compared to the other two taggers on both corpora.

Evaluation of the species assignment

For the benchmarking of the species assignment, we chose two well-known gene-rich corpora [i.e. GNormPlus (7) and National Library of Medicine (NLM)-Gene (26)]. Some articles, however, mentioned only one species (or no species). In such cases, all genes in the articles should link to the same species, without ambiguity (human is the species at default). To determine the extent of improvement with the new method, we need to focus on the articles' ambiguity issues. In that regard, the articles qualified for benchmarking require more than one species candidate for the gene mentions in the text. In addition, the two corpora were annotated in abstracts, but not full texts. However, the annotation guideline allows the annotators to use the full text if the corresponding species of a gene is not asserted in the abstract. We excluded those abstracts that contained genes, for which the corresponding species is in the full text but not in the abstract. Table 2 shows the number of abstracts in GNormPlus and NLM-Gene corpora that are qualified for benchmarking. Based on the criteria, fewer than half of the articles are eligible. The GNormPlus corpus is primarily focused on human genes, such that if a gene in an article corresponds to two or more species, only human genes are annotated in the corpus. To ensure that the evaluation can reflect the actual species diversity of the genes, we thus randomly selected the articles from the NLM-Gene corpus only for testing. The remaining eligible articles in the NLM-Gene and GNormPlus corpora were used for model development. In total, we collected 403 abstracts for training and 75 abstracts for evaluation.

Table 2. The corpora for benchmarking

Corpus	No. of abstracts	Training	Test
GNormPlus	262 (694)	262	0
NLM-Gene	216 (550)	141	75
Total	478 (1245)	403	75

The numbers in parentheses are the original numbers of articles in individual corpus

In our experiments, we downloaded two biomedical PLMs (i.e. PubMedBERT and Bioformer) and evaluated them in both frameworks. The title and abstract are concatenated as an input instance. The models were trained using the Adam (27) optimizer to minimize categorical cross-entropy loss. For the parameter setting, we used PLMs with the default parameter settings and set the other hyper-parameters as follows: a learning rate of 5×10^{-6}, a batch size of 16 and a max input length of 512 tokens (truncated if the length of text is over the maximum length). Due to the small size of the training data, we did not split a validation subset for early stopping. Instead,

Table 3. The performance of the species assignment

PLM model	Framework	Strict-ACC	Relax-ACC	Processing time (seconds per 100 abstracts)	
				GPU	CPU
PubMedBERT	Sequence labeling	**81.3%**	**85.4%**	16	50
Bioformer	(S→G)	80.7%	83.7%	10	25
PubMedBERT	Sequence labeling	79.7%	83.2%	90	740
Bioformer	(G→S)	78.1%	82.5%	60	320
PubMedBERT	Binary classification	79.8%	83.6%	140	1580
Bioformer		78.1%	82.7%	80	674
GNormPlus	Rule-based	65.8%	71.7%	4	4

Note that all models were trained and evaluated on the same GPU (Tesla V100-SXM2-32GB) and CPU (Intel(R) Xeon(R) Gold 6226 CPU at 2.70 GHz, 24 Cores). Some genes may correspond to more than one species. Strict-ACC denotes strict accuracy, which requires that all the corresponding species of the gene should be extracted. Relax-ACC denotes relax accuracy, which accepts that only one corresponding species is extracted.

we set up the patience parameter (patience = 5) to stop the training if no improvement within five epochs is observed. We focused on evaluating the performance on species assignment, using the resampled corpus, and the manually curated gene mentions are given.

To explore the effectiveness of our sequence labeling framework, we compared the performance of PLMs in classification and sequence labeling frameworks on the test. We also used the species assignment module (a rule-based method) of GNormPlus as a baseline. Table 3 shows the results of different methods in terms of accuracy and processing time. As expected, deep learning methods provide better performance than does the rule-based method. Compared with the binary classification framework, our sequence labeling framework achieves similar or better performance. For processing time, our sequence labeling framework is more efficient. Specifically, it is 10–20 times faster than the binary classification framework when tested on graphics processing unit (GPU) and central processing unit (CPU), respectively. Furthermore, when comparing the two sequence labeling strategies (S→G and G→S), we found that S→G is more efficient and accurate than is G→S for both PLM models (PubMedBERT (16) and Bioformer (17)). The highest performance (sequence labeling framework with the PubMedBERT model, using the S→G search strategy) increased about 16% compared to the baseline (GNormPlus), from 65.8% to 81.3%. Even though PubMedBERT achieved a slightly higher performance than did Bioformer, Bioformer is more efficient than is PubMedBERT in terms of both GPU and CPU environments (around two times faster) and may be a better option for processing the large-scale data sets (e.g. entire PubMed abstracts or PMC full texts).

Discussion

Despite our best efforts, there are still errors in the results of the species assignment. We examined all the errors of the S→G sequence labeling framework using the Bioformer model (i.e. 83.7% accuracy) in the test set and grouped them into several major categories, as shown in Figure 4. In most cases, the nearest species have the highest probability to be the corresponding species of the gene spans. Thus, it is confusing to the machine if the surrounding species does not correspond to the gene. This situation causes 44% of our errors and is the major category of errors for our results. As the example in PMID: 25277705, the species names of the respiratory syncytial virus (RSV) are glutted in the article, but the article

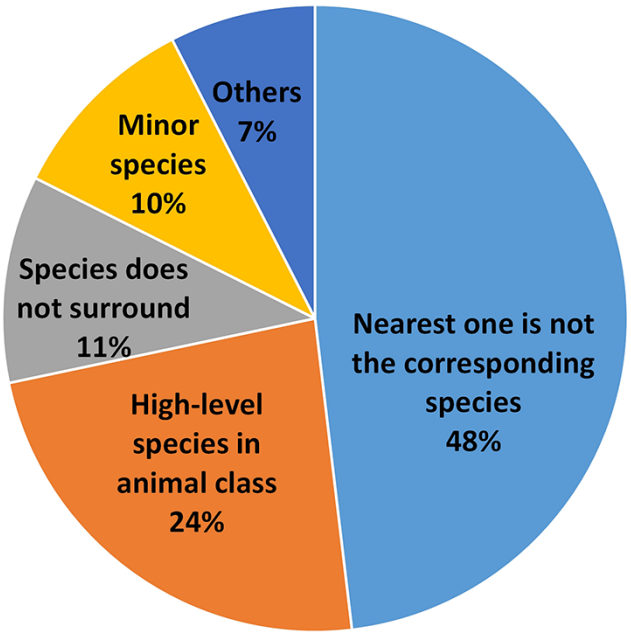

Figure 4. The species assignment error types.

concerns RSV infection, not the genetics of the virus. Thus, the corresponding species of the genes is human but not RSV. As we learned from this type of error, the genes of humans and viruses can be confused when there are two different research topics. In the first topic, the human gene function is relevant to the virus infection. In PMID: 35238065, the human CLIC3 gene is a potential indicator of poor prognosis of hepatitis B virus-related acute-on-chronic liver failure. The other article, however, focuses on the variants of the virus sequence. In PMID: 35416390, SARS-CoV-2 with E484K mutation in the spike gene is expressed in lower expected inhibitory activity of antibodies. To better address the issue, it is necessary to recognize the topic of the research.

The second error category is caused by a confusion of the species names within the higher level of the animal class. As an example, in PMID: 23195221, the 'mammalian' represents human, mouse and other mammals. The NLM-Gene corpus annotated the genes to mouse, however, as the full text analyzed the genes in the mouse model. Such an ortholog gene, which represents all the belonging genes, exists widely in various literature studies. No existing NCBI gene identifier, however, can represent those ortholog genes.

The third error type occurs when the corresponding species is far from the gene, and even the species is not in the same (or nearby) sentence(s). Sometimes, a closed species does not correspond to the gene. Such an example is seen in PMID: 20644716, the experiment *in vitro*/*vivo* that used a mouse model (cell line RAW 264.7) to understand the expression of the human ortholog gene. Unfortunately, our method cannot always handle the cases well. The other error type occurs when the number of the corresponding species in the article is much less than that of the other species. In such a case, our method may incorrectly assign the most frequently occurring species to the genes instead. For instance, the species of *Caenorhabditis elegans* appears four times, which is significantly higher than that of the human (which appears only once) in PMID: 18627611. The case led some human genes to be wrongly assigned to *C. elegans*. The remainder of the errors are relatively small and are caused by various factors, e.g. one gene mention for multiple species.

Furthermore, as shown in Table 2, 767 (1245 – 478 = 767) documents in GNormPlus and NLM-Gene corpora were not used to benchmark, as the deep learning approach is not applicable. These articles can be grouped into three types: (i) no species can be found in the article, and thus, the straightforward method is to assign all the genes to humans; (ii) only one species is found in the article, and thus, all the genes can be assigned to the species without using the deep learning method and (iii) the corresponding species of the genes are not in abstract, and thus, we can assign the remaining species in the article only to the genes, but none is correct. When we apply simple rules, we attain a comparable accuracy of 85.4% on these documents.

Conclusion

In this article, we first proposed a species tagger with state-of-the-art performance and further presented a novel idea to address species assignment, which is the biggest challenge of the gene normalization task. The task of recognizing the corresponding species from various candidates for gene mentions is more relevant to information-retrieval or relation-extraction tasks, but we rephrased the problem into a sequence labeling task, which is normally applied to a named-entity recognition task. The new method raised the performance accuracy of the species assignment (from 65.8% to 81.3%) within an acceptable process speed for large-scale data processing. Based on these promising results, we believe that the sequence labeling framework of species assignment can work with other relevant topics as well (e.g. the corresponding genes/species of variants and the corresponding variants of the phenotypes). Nevertheless, the tool is currently being developed and evaluated only on abstracts. Because more detailed information is recorded in the full text, in the future, we would like to further improve our methods to be able to handle full-length articles with multiple passages as well as the highly unstructured parts (e.g. tables) of the text.

Funding

This research was supported by the Intramural Research Program of the NLM, National Institutes of Health.

Data availability

The dataset used in this study is publicly available at https://github.com/ncbi/SpeciesAssignment/tree/main/data.

Conflict of interest

None declared.

References

1. Poux,S., Arighi,C.N., Magrane,M. *et al.* (2017) On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*, **33**, 3454–3460.
2. Wu,C.H., Arighi,C.N., Cohen,K.B. *et al.* (2012) BioCreative-2012 virtual issue. *Database*, **2012**, bas049.
3. Chen,Q., Allot,A. and Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, **49**, D1534–D1540.
4. Allot,A., Peng,Y., Wei,C.-H. *et al.* (2018) LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.*, **46**, W530–W536.
5. Lee,K., Famiglietti,M.L., McMahon,A. *et al.* (2018) Scaling up data curation using deep learning: an application to literature triage in genomic variation resources. *PLoS Comput. Biol.*, **14**, e1006390.
6. Wei,C.-H., Allot,A., Leaman,R. *et al.* (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.
7. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2015) GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *Biomed. Res. Int.*, **2015**, 918710.
8. Lu,Z., Kao,H.-Y., Wei,C.-H. *et al.* (2011) The gene normalization task in BioCreative III. *BMC Bioinform.*, **12**, S2.
9. Hakenberg,J., Gerner,M., Haeussler,M. *et al.* (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.
10. Gerner,M., Nenadic,G. and Bergman,C.M. (2010) LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinform.*, **11**, 1–17.
11. Naderi,N., Kappler,T., Baker,C.J. *et al.* (2011) OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, **27**, 2721–2729.
12. Pafilis,E., Frankild,S.P., Fanini,L. *et al.* (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One*, **8**, e65390.
13. Verspoor,K., Roeder,C., Johnson,H.L. *et al.* (2010) Exploring species-based strategies for gene normalization. *IEEE/ACM Trans. Comput. Biol. Bioinform. Biol. Insights*, **7**, 462–471.
14. Huang,M., Liu,J. and Zhu,X. (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033.
15. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2012) SR4GN: a species recognition software tool for gene normalization. *PLoS One*, **7**, e38460.
16. Gu,Y., Tinn,R., Cheng,H. *et al.* (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, **3**, 1–23.
17. Fang,L. and Wang,K. (2021) Team bioformer at BioCreative VII LitCovid Track: multic-label topic classification for COVID-19 literature with a compact BERT model. In: *BioCreative VII Workshop*. BioCreative, Online, pp. 272–274.
18. Weber,L., Sänger,M., Münchmeyer,J. *et al.* (2021) HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, **37**, 2792–2794.
19. Linnaeus,C. (1735) Systema naturae; sive, Regna tria naturae: systematice proposita per classes, ordines, genera & species. Haak.

20. Li,Z., Yang,Z., Xiang,Y. *et al.* (2020) Exploiting sequence labeling framework to extract document-level relations from biomedical texts. *BMC Bioinform.*, **21**, 1–14.
21. Luo,L., Yang,Z., Cao,M. *et al.* (2020) A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *J. Biomed. Inform.*, **103**, 103384.
22. Luo,L., Lai,P.-T., Wei,C.-H. *et al.* (2021) Extracting drug-protein interaction using an ensemble of biomedical pre-trained language models through sequence labeling and text classification techniques. In: *Proceedings of the BioCreative VII Challenge Evaluation Workshop*. BioCreative, Online, pp. 26–30.
23. Giorgi,J.M. and Bader,G.D. (2018) Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, **34**, 4087–4094.
24. Lee,J., Yoon,W., Kim,S. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
25. Weber,L., Münchmeyer,J., Rocktäschel,T. *et al.* (2020) HUNER: improving biomedical NER with pretraining. *Bioinformatics*, **36**, 295–302.
26. Islamaj,R., Wei,C.-H., Cissel,D. *et al.* (2021) NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *J. Biomed. Inform.*, **118**, 103779.
27. Kingma,D.P. and Ba,J. (2015) Adam: a method for stochastic optimization, In: *The International Conference on Learning Representations (ICLR)*. ICLR, San Diego, CA, USA, pp. 1–15.