Computational and Structural Biotechnology Journal 23 (2024) 1469–1476

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

RIscoper 2.0: A deep learning tool to extract RNA biomedical relation sentences from literature

Hailong Zheng[a,1], Linfu Xu[a,1], Hailong Xie[a,1], Jiajing Xie[b], Yapeng Ma[a], Yongfei Hu[a], Le Wu[a], Jia Chen[a], Meiyi Wang[a], Ying Yi[a], Yan Huang[a], Dong Wang[a,c,*,2]

[a] *Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, 510515 Guangzhou, China*
[b] *National Institute for Data Science in Health and Medicine, Xiamen University, 361102 Xiamen, China*
[c] *Guangdong Province Key Laboratory of Molecular Tumor Pathology, 510515, Guangzhou, China*

ARTICLE INFO

Keywords:
Non-coding RNA
Deep learning
Artificial intelligence

ABSTRACT

RNA plays an extensive role in a multi-dimensional regulatory system, and its biomedical relationships are scattered across numerous biological studies. However, text mining works dedicated to the extraction of RNA biomedical relations remain limited. In this study, we established a comprehensive and reliable corpus of RNA biomedical relations, recruiting over 30,000 sentences manually curated from more than 15,000 biomedical literature. We also updated RIscoper 2.0, a BERT-based deep learning tool to extract RNA biomedical relation sentences from literature. Benefiting from approximately 100,000 annotated named entities, we integrated the text classification and named entity recognition tasks in this tool. Additionally, RIscoper 2.0 outperformed the original tool in both tasks and can discover new RNA biomedical relations. Additionally, we provided a user-friendly online search tool that enables rapid scanning of RNA biomedical relationships using local and online resources. Both the online tools and data resources of RIscoper 2.0 are available at http://www.rnainter.org/riscoper.

1. Introduction

Biomedical relations are essential for systematically elucidating molecular networks and facilitating novel discoveries [1,2]. Recent studies on biomedical relations mainly focus on proteins, including protein-protein interactions [3,4] and protein-chemical interactions [5]. However, the protein biomedical relations only represent a part of the life activities. Diverse RNA biomedical relations may play a more extensive role in multi-dimensional regulatory systems and act as a necessary complement to the biological network [6–8]. Especially, the biomedical relations of non-coding RNA (ncRNA), including microRNA and long ncRNA (lncRNA), are far more intricate and dynamic and have attracted more attention [9,10]. Moreover, the prior knowledge of RNA contributed to the analysis of biological mechanisms, such as the endogenous competition between RNA transcripts [11] and the "coherent" and "incoherent" feedforward loops between miRNAs and transcription factors [12].

Biomedical relations involving RNA are usually scattered across numerous biological studies and prediction methods, which are mainly collected in databases [13,14]. Although several tools are available for exploring biomedical relations, none of them are specifically designed for RNA. For example, PubTator Central only annotated genes/proteins [15], while LPInsider only focused on the lncRNA-protein interactions [16]. For another example, Luo et al. used a deep neural network to extract miRNA-target interactions [17]. However, this tool is limited to miRNA and cannot be applied for RNA entity recognition. In our previous study, we continuously updated RNA-RNA interactions (RR) [7,18], which encompass both direct and indirect interactions, and RNA-Disease relations (RD) [6,19], which described the RNAs whose disorder can lead to the diseases. Furthermore, we provided the RIscoper [20], the only tool for extracting RNA-RNA interactions from literature, to address the delay in database updates and the limited prediction methods for certain types of RNA. However, RIscoper had two limitations that need to be addressed to make it more practical. Firstly, RIscoper detected relations only for RR, which is just one of the RNA biomedical relations. Secondly, RIscoper is a word frequency statistical

[*] Corresponding author at: Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, 510515 Guangzhou, China.
E-mail address: wangdong79@smu.edu.cn (D. Wang).

[1] These authors contributed equally to this work.

[2] Postal address: 510515

https://doi.org/10.1016/j.csbj.2024.03.017
Received 24 January 2024; Received in revised form 15 March 2024; Accepted 21 March 2024
Available online 24 March 2024

method that recognizes RNA through matching rather than predicting the semantic information of words. Therefore, it is unsuitable for RIscoper to capture the biomedical relations involving new RNAs. With the development of deep learning, particularly the emergence of pre-trained models, text mining tools significantly advanced in their capacity to grasp semantic information and predict biological entities [21,22]. Several deep learning-based text mining tools have been developed to extract and organize the inside textual information related to various biological entities [23–25], such as proteins, drugs and diseases [26–28]. However, there is a lack of tools designed for comprehensive scanning of RNA biomedical relations.

In this study, we manually curated and established a comprehensive and reliable corpus about RNA biomedical relations. In addition, we updated the text mining tool RIscoper 2.0 to extract RNA biomedical relation sentences from the literature and identify the entities in sentences. By leveraging the pretrained models in deep learning, RIscoper 2.0 obtained a new capacity to quickly recognize RNA biomedical relations covering novel entities. Furthermore, a user-friendly website was provided for database curators, experimental biologists and bioinformaticians.

2. Material and methods

2.1. Description of the RNA biomedical corpus

The RNA biomedical relations discussed in this article include RNA-RNA interactions (RR) and RNA-Disease relations (RD), which were consistently recorded in our database works (RR, from RAID [18] to RNAInter 4.0 [7], RD, from MNRD [19] to RNADisease 4.0 [6]). The RR and RD records contain multiple types of information, including the names of both entities in the entity pairs, corresponding sentences and PubMed IDs. Besides, the extent of manually extracted sentences is consistent with the description in the original database. For RR, literature within PubMed (mainly from 2000 to 2020) was screened with the following keyword combinations: (RNA molecule) AND (other RNA molecule) AND (interaction keywords). For RD, literature within the PubMed database (mainly from 2000 to 2021) was screened with the following keyword combinations: (ncRNA symbols or ncRNA category names) and (disease names). In this study, we re-verified the corpus of these RNA biomedical relations and further annotated RNAs and diseases with the corresponding named entities (Table 1). Notably, the RNA biomedical relations in this corpus are supported by experiments or prediction tools, which were unified and recorded in our database.

2.2. Annotation methodology

For the sentences in the above records, we re-verified the RNA biomedical relations and used Colabeler software (http://www.colabeler.com/) to annotate RNAs and diseases with BIO (Begin, Inside, Outside) tags, such as breast (B-Disease) cancer (I-Disease). A detailed annotation guideline was provided to educate annotators when creating the corpus (Supplementary Note 1). The entity annotation process consists of two steps to ensure the quality of annotations: (1) We referred to the raw records and the referenced entity databases (Table S1), such as NCBI Gene database and Disease Ontology [29]. If the entity can be found in the references (in the form of name, symbol or aliases), the entire entity was labeled with the corresponding tag. (2) For uncertain entities, particularly abbreviations, we searched the corresponding article using the PMID information in the record. We then confirmed whether the word is an RNA or disease entity based on the information where the entity first appears. In this study, positive sentences are defined as sentences that contain at least one RR or RD. Entity pairs in positive sentences are combined with interaction keywords, such as target and bind.

To build a sentence classification model that distinguishes the positive sentences from others, the negative sentences in this study were defined as sentences that do not contain any RD or RR. Thus, negative sentences included both sentences with entity pair and without entity pair. Negative sentences without entity pair describe the sentences with no RNA-RNA or RNA-disease pair. Negative sentences with entity pair match one of the following two conditions: (1) Sentences do not contain keywords of interaction. (2) Sentences contain keywords of interaction but negative words that affect the judgment of negative samples, such as the word “unchanged” in the sentence “Phosphorylation of SMAD2 was inhibited, while that of SMAD1 remained unchanged”. Ensuring the coverage of negative sentences is important to develop a model that can be applied at the article level. We hypothesized that the vast majority of sentences in the article do not contain RNA biomedical relations. Further, we adopted a three-step strategy to ensure the coverage and reliability of the negative sentences. Firstly, we searched the keywords “RNA-RNA interaction” and “RNA-Disease interaction” on PubMed and downloaded the full text of 15,500 articles published on PMC from 2000 to 2022. Then, we generated 46,500 negative sentences by randomly selecting three sentences from each article. Secondly, we used the dictionary matching method to identify entities in the above sentences. The names of RNAs and diseases were extracted from multiple databases, including HGNC [30], GENCODE [31], RNAInter [7], RNADisease[6], MeSH terms, and HumanDO [32] (Table S1). We further manually confirmed that out of 1376 sentences containing at least one pair of entities, 803 sentences contained RR or RD, while 573 sentences did not contain either RR or RD. The 803 sentences were removed, resulting in a total of 45,697 negative sentences. Thirdly, we randomly selected 100 sentences containing less than two entities and verified them with all three annotators. As shown in Table S2, all annotators confirmed that none of these sentences contained RR or RD, which supported negative sentences for model construction.

2.3. Flowchart of RIscoper 2.0

Fig. 1 illustrates the flowchart of the RIscoper 2.0 model, which comprises three layers: the initial feature extraction layer, text classification (TC) layer and named entity recognition (NER) layer. Notably, the TC and NER tasks are merged in the loss function, which benefited from the sentences with both entity annotations and sentence labels. All three parts were fine-tuned during training.

Firstly, in the initial feature extraction layer, we leveraged PubMedBERT [22] to tokenize the input sentences and generate the 768-dimensional contextual embeddings for each token. PubMedBERT is a specialized variant of the Bidirectional Encoder Representations from Transformers (BERT) model, fine-tuned specifically for biomedical text and information extraction from the PubMed database. Upon receiving input text, the PubMedBERT tokenizer first matches each word in the text against the terms in the PubMedBERT vocabulary, which includes common words, subwords, and characters. If a word is not in

Table 1
RNA biomedical corpus.

	Relation type		Entity type				Number of literatures
	RNA-RNA	RNA-disease	RNA	RNA entry	Disease	Disease entry	
RIscoper	12,728	-	-	-	-		4897
RIscoper 2.0	16,693	14,292	75,661	13,431	17,567	1569	15,581

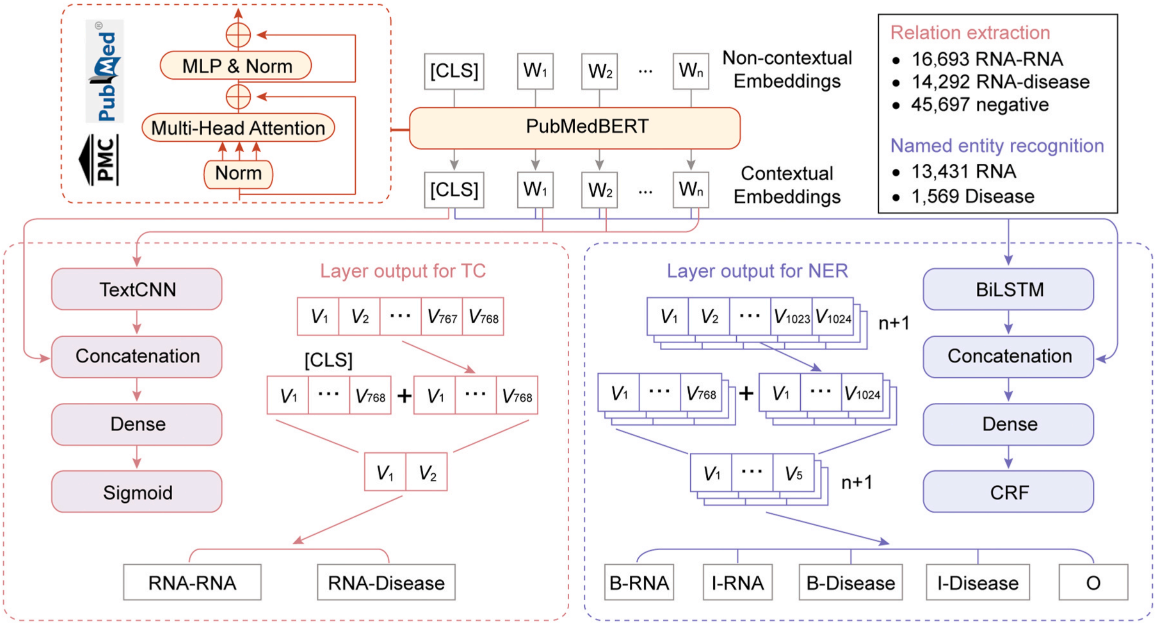

Fig. 1. **Flowchart of RIscoper 2.0.**

the vocabulary, the tokenizer then segments the word into smaller subwords. For example, the lncRNA entity “lnc-XLEC1″, which is manually tagged with “B-RNA”, would be tokenized into “ln”, “##c”, “-”, “XL”, “##EC”, and “##1″. Furthermore, the first token of an entity is designated with a “B-” prefix. For example, “ln” is tagged with “B-RNA” when “##c”, “-”, “XL”, “##EC”, and “##1″ are tagged with “I-RNA”. Then, PubMedBERT pre-trained model maps the tokens to 768-dimensional contextual embeddings, which were further utilized for both the TC and NER tasks.

Secondly, in the TC layer, we used Convolutional Neural Network for Text Classification (TextCNN) [33] on all the contextual embeddings except the CLS (Classification) token. TextCNN uses multiple convolution kernels of different sizes to extract key information in sentences, which can better capture local features. In this study, we deployed three types of convolutional kernels with different sliding windows (3, 4 and 5), each consisting of 256 individual convolutional kernels. This process yielded a 768-dimensional vector for each input token, consistent with the length of the CLS token generated by PubMedBERT. Then, the output of TextCNN was concatenated with the contextual embedding of the CLS token. Finally, the model incorporated a fully connected layer and a sigmoid layer to determine whether the input sentence contains RR or RD.

Thirdly, in the NER layer, we used one of the state-of-the-art NER models, Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF), to process all the contextual embeddings. The input tokens were first encoded into a 1024-dimensional hidden state vector sequence via BiLSTM [34]. Subsequently, the BiLSTM output was concatenated with the initial input contextual embeddings and further fused through a fully connected layer to predict labels for each token in the sentence. Finally, we utilized CRF [35] to decode the optimal tag sequence among all possible sequences.

The final loss function of the model was a sum of the Cross-Entropy and Negative Log-Likelihood functions from both the TC and NER layers. Detailed hyperparameter settings of model training were shown in Table 2.

Table 2
Hyperparameter setting.

Global		Local		
Parameters	Values	Layer	Parameters	Values
Learning rate	5 × 10-5	TextCNN	Kernels number	768
Epochs	10		Kernels size	3, 4, 5
Batch size	16		Loss function	Cross entropy
Optimization function	Adam		Pooling	1-max pooling
Activation function	ReLU	BiLSTM-CRF	Loss function	Negative Log-Likelihood
Dropout	0.5		LSTM units	512

2.4. Evaluation metrics

Fleiss’ kappa was used to assess the consistency among annotators. The formula for calculating Fleiss’ kappa is as follows:

$$\text{Fleiss}' \text{ kappa} = \frac{P_o - P_e}{1 - P_e}$$

When P_e is the assumed probability of consistency and P_o is the observed probability of consistency. P_e and P_o are calculated as follows:

$$P_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij}\right)^2$$

$$P_o = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right)$$

When n is the number of annotators, N is the number of annotated samples, k is the number of categories of samples, and n_{ij} is the number of annotators who classify sample i as category j. For the TC task, samples were defined as all input sentences, and the categories included RR, RD and None. For the NER task, samples were defined as all words in sentences, and the categories included RNA, disease and None.

Further, to evaluate the performance of text mining tools in TC and NER tasks, we calculated three metrics: Precision, Recall, and F1 score. These metrics are defined as follows:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

$$\text{F1 score} = 2 \times P \times R / (P + R)$$

Where TP, FP, and FN represent true positive, false positive and false negative, respectively. Here, we used a strict full length matching strategy to determine named entities.

3. Results

3.1. Summary of the RNA biomedical corpus

Before data collection, we assessed the inter-annotator agreement among three individuals who participated in the annotation task. Two of the annotators are master's students, and one is a Ph.D. candidate. All the annotators possess a background in Biomedicine and have substantial experience working in the RNA-related biomedical text mining domain. A trial collection was created to determine the inter-annotator agreement between the annotators. For both the TC and NER tasks, we calculated Fleiss' kappa to calculate the consistency among three annotators. Notably, the Fleiss' kappa of the NER task was calculated in the exact match and partial match patterns. Exact match is a situation where the annotations should completely overlap, whereas partial match is a situation where the annotations may partially or completely overlap.

The trial collection was created using a random sample of 100 sentences from RR, RD and negative corpora (Table S2-S4). For the TC task, three annotators used the original record as a reference for their judgment, resulting in a high Fleiss' kappa of 0.990. For the NER task, the evaluators achieved a Fleiss' kappa of 0.824 in the exact matching mode and 0.927 in the partial matching mode (Table 3).

Based on the high inter-annotator agreement, we constructed a corpus composed of 16,693 RR and 14,292 RD sentences, which were manually annotated with about 100,000 named entities (Table 1). The sentence lengths range predominantly from 1 to 80 (Fig. 2A, B), and the maximum lengths of RR and RD sentences are 163 and 256, respectively. Further, all these sentences have at least one entity pair, and many sentences contain more than two entities (Fig. 2C-D). When compared to the original RIscoper, the corpus within RIscoper 2.0 exhibits two advancements. Firstly, both the number and type of the corpus were improved. Secondly, we added about 100,000 annotations for RNA and disease entities in the corresponding corpora and further confirmed the accuracy of both sentence and entity annotations.

To demonstrate the uniqueness of the RNA biomedical corpus, we compared the semantic features between RNA-RNA and Protein-Protein corpus (Intact database [36]), as well as those between RNA-Disease and protein-Disease corpus (Genetic Association Database [37]). For each sentence, we used the original PubMedBERT model to compute the contextual embeddings of CLS as the sentence features. We then used principal component analysis (PCA) to reduce the dimensionality of these features. The results showed a clear separation between the semantic features of the RNA biomedical corpus and those of the protein biomedical corpus (Fig. 2E, F), which demonstrated the necessity to develop a TC and NER method for the RNA biomedical corpus.

3.2. Development and evaluation of RIscoper 2.0

Benefiting from the RNA biomedical corpus annotated with both sentence labels and named entities, we developed RIscoper 2.0, a BERT-based deep learning model that merged both NER and TC tasks. In the model, PubMedBERT was used for the initial feature extraction, and TextCNN and BiLSTM-CRF were employed for TC and NER tasks. Further, the loss functions for both tasks were added in the final loss function (Materials and Methods, Fig. 1). All three layers (initial feature extraction layer, TC layer and NER layer) were fine-tuned during training. In this study, RIscoper 2.0 was evaluated in the 10-fold cross-validation framework. We divided the RR, RD and negative corpora into ten parts, with nine parts being the training set and the remaining part being the test set. For each fold, we first used the RR and RD corpora for model training. Then, we froze all layers except the TC layer and continued to train the model using negative datasets that only contained sentence label information. Therefore, the NER function can only be used for positive sentences, and all subsequent NER performance evaluations were carried out on positive sentences. The average values of the evaluation metrics (F1 score, Precision, Recall) across the ten models were presented in Table 4. RIscoper 2.0 achieved F1 scores of 0.925 and 0.916 in TC and NER tasks, which highlighted its strong performance within RNA biomedical corpora.

The model's feature extraction capabilities were evaluated to demonstrate the enhancements resulting from our strategic modifications in the three layers. Following feature extraction at these layers, the input sentences generated high-dimensional vectors at both entity and sentence levels. We applied PCA to these high-dimensional vectors and projected them into a 2-dimensional space to facilitate visual comparison. In this study, we assessed the feature extraction capabilities of different layers by evaluating the separation of features at entity and sentence levels. Firstly, for the initial feature extraction layer, we used the vector of the CLS token as the sentence feature and found that the fine-tuned PubMedBERT model was able to distinguish between different RNA biomedical relations (Fig. S1A, B). Similarly, for entity tokens labeled with the BIO tags, the fine-tuned model could differentiate between RNA and disease entities (Fig. S1C-F). Secondly, For the TC layer, in contrast to previous approaches that directly used the outputs of TextCNN for classification, we concatenated the output vectors of both TextCNN and Fine-tuned PubMedBERT for classification. Fig. 3A-C demonstrated the feature extraction capabilities of three different strategies: (1) The first strategy involved only fine-tuned PubMedBERT, which resulted in the close distance among the same positive sentences. (2) The second strategy involved only TextCNN, which resulted in a clearer boundary between features of different sentence types (3) The third strategy, the merged strategy, successfully inherited the strengths of the above two layers. Compared to the two strategies mentioned above, the merged strategy resulted in a closer clustering of similar sentence types and a clearer boundary between different sentence types. Thirdly, for the NER layer, we concatenated the output vectors of both BiLSTM and Fine-tuned PubMedBERT for each token. As depicted in Fig. 3D-I, this strategy showed a more concentrated clustering of entities from the same class and a clearer demarcation between entities of different types, highlighting the advantages of the combined use of these two output vectors in both RR and RD sentences. The study also evaluated the impact of different hyperparameters on the model, such as different pretrained models and the integration or separation of the TC and NER tasks (Fig. S2). The results indicated that these hyperparameters have almost no impact on the results.

Table 3
Inter-annotator agreement among three annotators.

Annotators	Exact match (entity)			Partial match (entity)			Text classification
	RNA (RR)	RNA (RD)	Disease (RD)	RNA (RR)	RNA (RD)	Disease (RD)	
1 and 2	0.853	0.821	0.760	0.915	0.952	0.915	0.985
1 and 3	0.901	0.825	0.752	0.951	0.952	0.850	1.000
2 and 3	0.864	0.858	0.780	0.914	0.956	0.935	0.985
All	0.873	0.834	0.764	0.927	0.953	0.900	0.990

Note: RNA-RNA, RR; RNA-Disease, RD

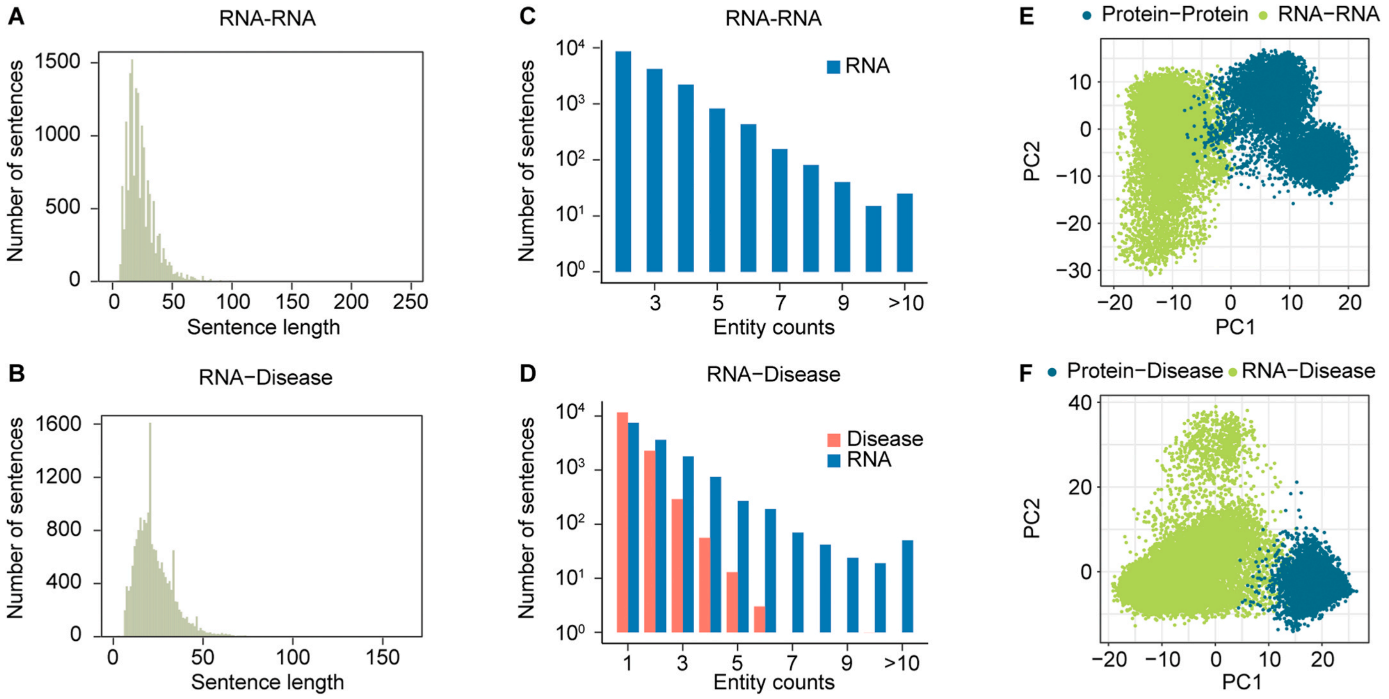

Fig. 2. Characteristics and Distinctiveness of the RNA biomedical relation corpus. The distribution of sentence length within the (A) RNA-RNA and (B) RNA-Disease corpora. The distribution of entity counts within the (C) RNA-RNA and (D) RNA-Disease corpora. Comparison of RNA and protein biomedical corpora in terms of (E) protein-protein interactions and (F) protein-disease associations.

Table 4
Evaluation of the RIscoper 2.0 in 10-fold cross-validation.

	Text classification			Named entity recognition		
	RNA-RNA	RNA-Disease	All	RNA	Disease	All
F1 score	0.937	0.912	0.925	0.933	0.845	0.916
Precision	0.936	0.919	0.928	0.915	0.864	0.905
Recall	0.938	0.906	0.923	0.953	0.828	0.928

3.3. Performance comparison with similar tools

In this section, we compared the performance of our tool with other state-of-art methods, including RIscoper 1.0 [20], MTI extractor [17], LPInsider [16], and GPT [38]. The detailed usage process and reproduction results of these tools were provided in Supplementary Note 2. Firstly, for a fair comparison with RIscoper 1.0, MTI extractor and LPInsider, we retrained our model with the 12,727 RR and 9418 RD sentences from the RIscoper 1.0 and MNDR databases. All the methods were tested in the independent dataset composed of 3966 RR and 4874 RD sentences that were newly integrated into the RNAInter and RNA-Disease databases. Of note, the test set included 1208 new entities for RR and 3495 new entities for RD (Fig. 4A, B). As presented in Table 5, RIscoper 2.0 achieved F1 scores of 0.887 and 0.916 in TC and NER tasks, while RIscoper 1.0 only achieved F1 scores of 0.541 and 0.695. When focusing on the sentences that contain new entities, RIscoper 2.0 still achieved higher F1 scores than RIscoper 1.0 in both TC (0.815 versus 0.466) and NER tasks (0.957 versus 0.761). Further, our tool overlaps with the MTI extractor in extracting microRNA-target interaction sentences. The results showed that the MTI extractor also achieved a lower F1 score of 0.860 in miRNA-related sentence classification. Similarly, the NER for lncRNAs is the overlapping function of our tool with LPInsider. The function of LPInsider was tested on the test dataset, resulting in a higher precision of 1 but a lower recall of 0.029, which may be due to the small size of its original training data.

PubMedBERT was pre-trained from scratch on PubMed articles up to 2020. To ensure fair performance comparison, we divided our corpus into training and testing sets based on the time point of 2020 and retrained the model. All corpus was divided into two parts based on the publication year of corresponding references: 16,671 RR and 8550 RD sentences published before 2020 for model training, and 22 RR and 5742 RD sentences published after 2020 for testing (Fig. 4C, D). Notably, the test set added 17 and 3581 new entities in RR and RD corpora, respectively, compared to the training set (Fig. 4E, F). Due to the limited number of RR sentences, we used recall values to evaluate the text classification results. As shown in Table S5, RIscoper 2.0 still achieved a recall value of 0.955 in the TC task and an F1 score of 0.885 in the NER task, which performed better in fair comparison with other methods.

GPT is a large generative language model widely used in natural language text processing. For comparison, we utilized chatGPT (version: GPT3.5, https://chat.openai.com/) provided by OpenAI to perform TC and NER tasks. GPT3.5 was trained on PubMed text data up to January 2022, which includes our corpus. Therefore, we directly applied GPT3.5 to all sentences in our corpus for comparison. The corpus was divided into 1000 equal parts based on the original sentence order, and the first sentence from each part was extracted for testing. In the end, we obtained 1000 sentences for testing and lowercased these sentences to keep the input data consistent with RIscoper 2.0. As shown in Table S6, RIscoper 2.0 outperformed zero-shot GPT (0.925 vs 0.882) in the TC task but was outperformed by few-shot GPT (0.939). In the NER task, RIscoper 2.0 achieved a higher F1-score (0.916) than both zero-shot (0.681) and few-shot GPT (0.759), which may be the reason that most genes and disease entities were abbreviations lacking useful information.

Overall, compared to the original RIscoper, RIscoper 2.0 has made advancements in three key aspects: (1) a more extensive and comprehensive RNA corpus, (2) improved recognition accuracy, and (3) the ability to explore novel RNA biomedical relations.

3.4. Web tool for practical applications

To facilitate easy access and application of RIscoper 2.0 to extract RNA biomedical relation sentences, we developed a publicly available

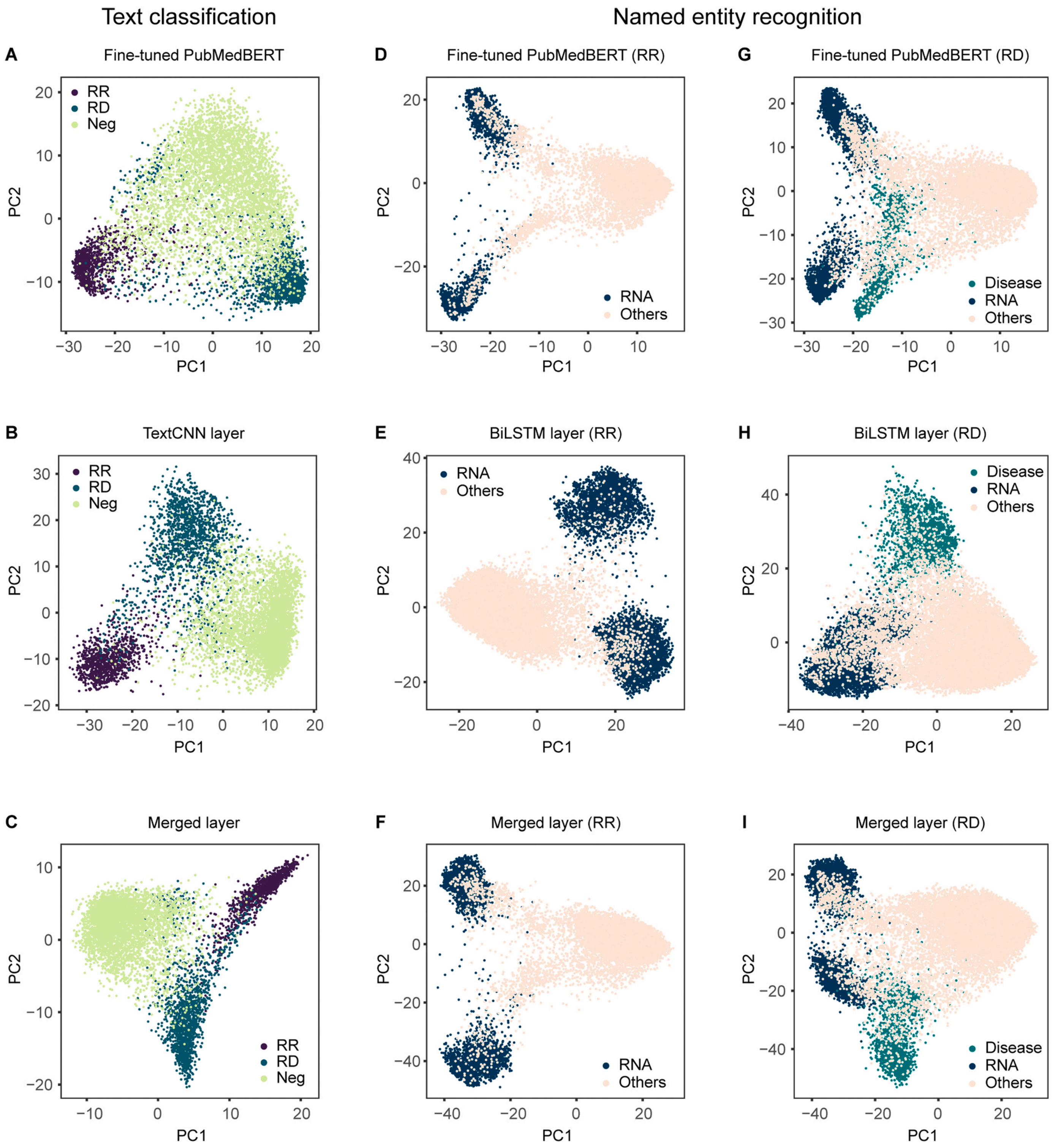

Fig. 3. PCA plots of feature vectors for different layers. (A-C) Distribution of sentence features output from different layers in the text classification task. (D-F) Distribution of entity features output from different layers in the RNA-RNA interactions (RR) corpus. (G-I) Distribution of entity features output from different layers in the RNA-Disease relations (RD) corpus.

online tool at http://www.rnainter.org/riscoper. This tool accepts both local and online resources, including Text, PDF, PubMed ID and Keyword. Notably, the Keyword function can be especially valuable as a supplementary resource when dealing with newly discovered RNAs that may not be included in the database timely, or in scenarios where the available biomedical relations are insufficient.

Table S7 presented three illustrative cases to assess the practicality of Keyword function, including one miRNA (symbol: MIR367), one lncRNA (MIR17HG), and one mRNA (symbol: BAZ2A). For example, by querying all symbols of miRNA MIR367 (MIR367, MIRN367, hsa-mir-367) in PubMed, we retrieved 83 relevant articles (abstracts) containing 820 sentences. RIscoper 2.0 recognized 127 positive RR and 115 positive RD, out of which 101 unique RR pairs and 103 unique RD pairs were manually validated (Table S8). When compared with the experimentally

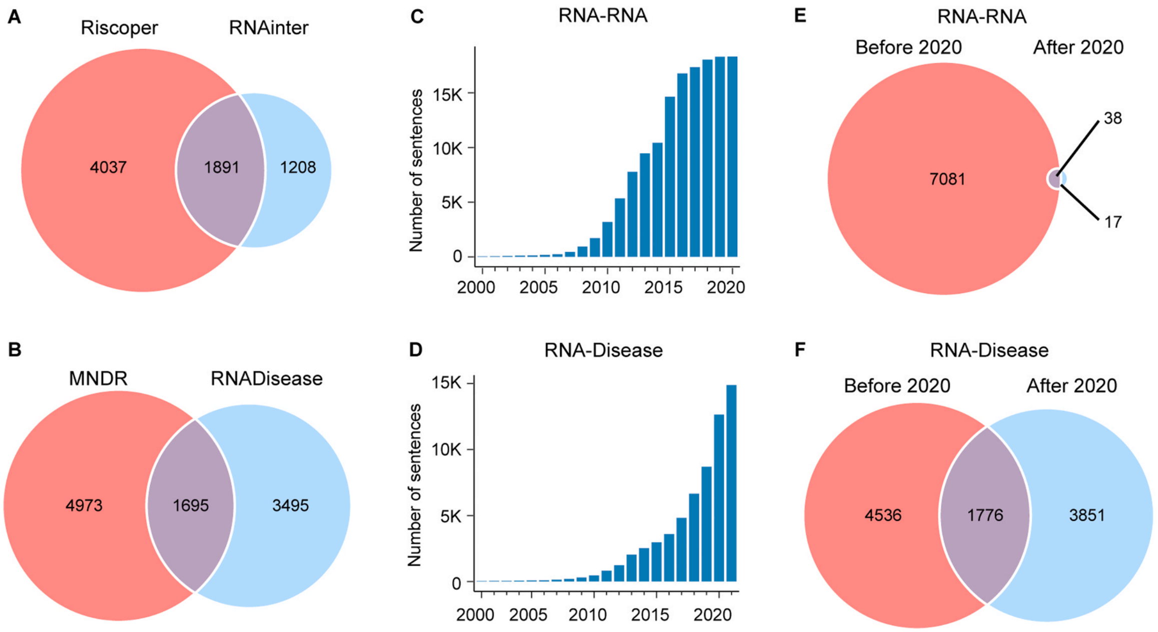

Fig. 4. Novel RNA biomedical relations and entities. (A-B) Overlapping of entities between different databases in both the RNA-RNA and RNA-Disease corpora. (C-D) The number of RNA-RNA and RNA-Disease sentences identified over time. (E-F) Overlapping of entities between sentences published before 2020 and after 2020 in both the RNA-RNA and RNA-Disease corpora.

Table 5
Evaluation of the RIscoper 2.0 in extracting new relations.

Test dataset	Method	Text classification			Named entity recognition		
		RNA-RNA	RNA-Disease	All	RNA	Disease	All
All	Riscoper 2.0	0.889	0.886	0.887	0.941	0.828	0.916
	Riscoper 1.0	0.541	-	0.541	0.695	-	0.695
miRNA	Riscoper 2.0	0.954	-	0.954	-	-	-
	MTI extractor	0.860	-	0.860	-	-	-
lncRNA	Riscoper 2.0	-	-	-	0.983	-	0.983
	LPInsider	-	-	-	0.056	-	0.056

validated data from RNAInter database, the 103 RR sentences involved 36 new literature sources and 23 novel RNA entities (Table S9). Similarly, the 99 RD sentences added 33 new literature sources, 24 novel RNA entities, and five new disease entities to the RNADisease database. Furthermore, upon further comparison with the data supported by strong experiments, RIscoper 2.0 demonstrated an enhanced capacity to discover additional interaction relationships. As shown in Table S8 and S9, the other two case studies similarly demonstrated the ability of RIscoper 2.0 to extract most of the true unique RNA biomedical relations and provide valuable supplements to the existing databases. Further, we tested our tool's ability to discover intra-sentence relationships where one of the entity pairs is replaced by a pronoun or general word, such as the words "Potential targets" in the sentence "Potential targets of miR-367 were screened by miRWalk software and luciferase reporter assays". Out of the 37 predicted RR and 32 predicted RD sentences that contained only one entity, 16 and 30 sentences were manually reconfirmed as RR and RD sentences where another entity was replaced by a pronoun or general word (Table S7). These results indicated that RIscoper 2.0 is suitable for mining intra-sentence relationships.

4. Discussion and conclusion

In this study, we established a comprehensive and reliable corpus focusing on RNA biomedical relations. The corpus exhibits unique semantic features compared to protein biomedical corpora, indicating its necessary role in the biomedical domain. Benefiting from the corpus annotated with both sentence labels and named entities, we developed RIscoper 2.0, a BERT-based deep learning model to extract the RNA biomedical relations. When compared to the original RIscoper and other text mining tools, RIscoper 2.0 demonstrated superior performance in both TC and NER tasks. Additionally, RIscoper 2.0 obtained the ability to extract the biomedical relations involving new entities. We also provided a user-friendly website that accepts local and online resources as input, which facilitates the exploration of RNA biomedical relations and serves as a valuable supplement to existing databases.

Despite these improvements, RIscoper 2.0 still has some limitations. Firstly, during the development of RIscoper 2.0, the RNA biomedical relations were limited to RR and RD due to the scarcity of other suitable corpus types for training deep learning models. For example, the RNA-Localization corpus contains only about 1000 sentences, which is significantly less than the RR and RD corpora. Similarly, most corpora are related to ncRNA, such as miRNA and lncRNA, making RIscoper 2.0 more suitable for ncRNA corpora. Secondly, the TC and NER tasks related to disease performed worse than those related to RNA, which can be attributed to two potential factors: (1) The unique entities of diseases were significantly fewer than RNA. (2) While RNA entities predominantly comprise a single word, many disease entities consist of multiple words. To enhance model performance, we could add additional disease-related sentences and annotations. Fourth, RIscoper 2.0 currently lacks the capability to distinguish RNA from other gene products. Our findings indicated an overlap of semantic features between RNA and protein biomedical corpora, which suggests that it is not feasible to distinguish

RNA and protein at the sentence level. A potential approach to distinguish RNA and protein may include extending analysis from the sentence level to paragraph or article level, facilitating more accurate identification of RNA and proteins.

In our future work, we will focus on the continued collection and annotation of RNA biomedical corpus to provide a favorable resource for ongoing text-mining studies of RNA biomedical relations. Our aim is to create a benchmark dataset for future machine learning works within the field. Additionally, we are committed to developing new models capable of comprehending the RNA biomedical corpus and useful web tools to promote the development of RNA-related research.

Funding

This work was supported by grants from the National Key Research and Development Project of China [2020YFA0113300, 2022YFA0806303, 2021YFC2500300]; National Natural Science Foundation of China [82370106, 82070109]; Guangdong Basic and Applied Basic Research Foundation [2024A1515011769, 2022A1515011253].

Author statement

AI and AI-assisted technologies were only applied to the use of basic tools for checking grammar, spelling or polishing.

CRediT authorship contribution statement

Ying Yi: Visualization. **Huang Yan:** Funding acquisition. **Hailong Zheng:** Data curation, Investigation, Software, Writing – original draft, Writing – review & editing. **Dong Wang:** Conceptualization, Funding acquisition, Writing – review & editing. **Linfu Xu:** Data curation, Investigation. **Hailong Xie:** Data curation, Software. **Jiajing Xie:** Investigation. **Yapeng Ma:** Visualization. **Yongfei Hu:** Writing – review & editing. **Le Wu:** Writing – review & editing. **Jia Chen:** Writing – review & editing. **Meiyi Wang:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

The annotated data and model code provided by this work were available at the website http://www.rnainter.org/riscoper/download and https://github.com/Wanglabsmu/RIscoper.

Acknowledgements

We thank members of the Wang lab for their discussions and feedback on the manuscript. We also thank Mei Wang for her valuable suggestions in response to the reviewers.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.03.017.

References

[1] Zhou H, Beltran JF, Brito IL. Host-microbiome protein-protein interactions capture disease-relevant pathways. Genome Biol 2022;23:72.

[2] Kovacs IA, Luck K, Spirohn K, Wang Y, Pollis C, et al. Network-based prediction of protein interactions. Nat Commun 2019;10:1240.

[3] Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell 2021;184:3022–40. e3028.

[4] Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8.

[5] Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 2021;30:187–200.

[6] Chen J, Lin J, Hu Y, Ye M, Yao L, et al. RNADisease v4.0: an updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. Nucleic Acids Res 2023;51:D1397–404.

[7] Kang J, Tang Q, He J, Li L, Yang N, et al. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. Nucleic Acids Res 2022;50:D326–32.

[8] Konig J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet 2012;13:77–83.

[9] Ghafouri-Fard S, Shoorei H, Mohaqiq M, Majidpoor J, Moosavi MA, et al. Exploring the role of non-coding RNAs in autophagy. Autophagy 2022;18:949–70.

[10] Slack FJ, Chinnaiyan AM. The role of non-coding RNAs in oncology. Cell 2019;179:1033–55.

[11] Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature 2014;505:344–52.

[12] Hausser J, Zavolan M. Identification and consequences of miRNA-target interactions–beyond repression of gene expression. Nat Rev Genet 2014;15:599–612.

[13] Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife 2015;4.

[14] Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.

[15] Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 2019;47:W587–93.

[16] Li Y, Wei L, Wang C, Zhao J, Han S, et al. LPInsider: a webserver for lncRNA-protein interaction extraction from the literature. BMC Bioinforma 2022;23:135.

[17] Luo M, Li S, Pang Y, Yao L, Ma R, et al. Extraction of microRNA-target interaction sentences from biomedical literature by deep learning approach. Brief Bioinform 2023;24.

[18] Yi Y, Zhao Y, Li C, Zhang L, Huang H, et al. RAID v2.0: an updated resource of RNA-associated interactions across organisms. Nucleic Acids Res 2017;45:D115–8.

[19] Ning L, Cui T, Zheng B, Wang N, Luo J, et al. MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Res 2021;49:D160–4.

[20] Zhang Y, Liu T, Chen L, Yang J, Yin J, et al. RIscoper: a tool for RNA-RNA interaction extraction from the literature. Bioinformatics 2019;35:3199–202.

[21] Lee J, Yoon W, Kim S, Kim D, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40.

[22] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, et al. Domain-specific language model pretraining for biomedical natural language processing 2021;3:1–23.

[23] Chikka VR, Karlapalem KJapa. A Hybrid Deep Learn Approach Med Relat Extr 2018.

[24] Guo S, Huang L, Yao G, Wang Y, Guan H, et al. Extracting biomedical entity relations using biological interaction knowledge. Inter Sci 2021;13:312–20.

[25] Lee K, Kim B, Choi Y, Kim S, Shin W, et al. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinforma 2018;19:21.

[26] Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 2022;38:2102–10.

[27] Zhang Y, Lin H, Yang Z, Wang J, Sun Y, et al. Neural network-based approaches for biomedical relation classification: a review. J Biomed Inf 2019;99:103294.

[28] Zhou H, Liu Z, Ning S, Lang C, Lin Y, et al. Knowledge-aware attention network for protein-protein interaction extraction. J Biomed Inf 2019;96:103234.

[29] Schriml LM, Munro JB, Schor M, Olley D, McCracken C, et al. The human disease ontology 2022 update. Nucleic Acids Res 2022;50:D1255–61.

[30] Wain HM, Lush M, Ducluzeau F, Povey S. Genew: the human gene nomenclature database. Nucleic Acids Res 2002;30:169–71.

[31] Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res 2023;51:D942–9.

[32] Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 2019;47:D955–62.

[33] Y.Japa Kim, Convolutional neural networks for sentence classification, (2014).

[34] Graves A. Long Short-term Memory. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Heidelberg: Springer; 2012. p. 37–45.

[35] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, (2001).

[36] Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, et al. The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res 2022;50:D648–53.

[37] Bravo A, Pinero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinforma 2015;16:55.

[38] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, et al., Language models are few-shot learners, 33 (2020) 1877–1901.