W616–W622 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 10 May 2022*
https://doi.org/10.1093/nar/gkac310

pubmedKB: an interactive web server for exploring biomedical entity relations in the biomedical literature

Peng-Hsuan Li[1], Ting-Fu Chen[1], Jheng-Ying Yu[1], Shang-Hung Shih[1], Chan-Hung Su[1], Yin-Hung Lin[1], Huai-Kuang Tsai[1,2], Hsueh-Fen Juan[1,3,4], Chien-Yu Chen[1,4,5] and Jia-Hsin Huang[1,*]

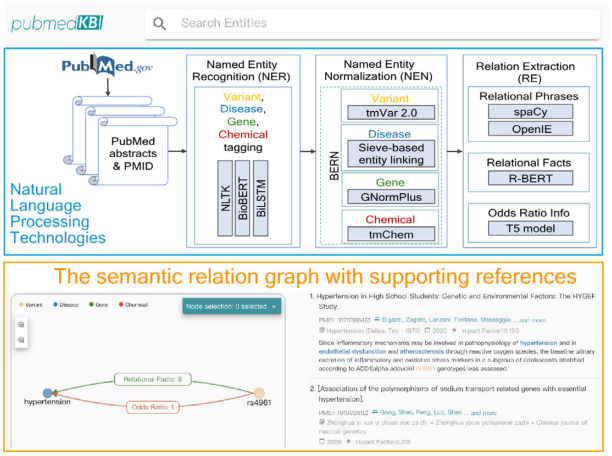
[1]Taiwan AI Labs, Taipei 10351, Taiwan, [2]Institute of Information Science, Academia Sinica, Taipei, 11529, Taiwan, [3]Department of Life Science, National Taiwan University, Taipei 10617, Taiwan, [4]Center for Computational and Systems Biology, National Taiwan University, Taipei 10617, Taiwan and [5]Department of Biomechatronics Engineering, National Taiwan University, Taipei, 10617, Taiwan

Received February 27, 2022; Revised April 06, 2022; Editorial Decision April 16, 2022; Accepted April 18, 2022

ABSTRACT

With the proliferation of genomic sequence data for biomedical research, the exploration of human genetic information by domain experts requires a comprehensive interrogation of large numbers of scientific publications in PubMed. However, a query in PubMed essentially provides search results sorted only by the date of publication. A search engine for retrieving and interpreting complex relations between biomedical concepts in scientific publications remains lacking. Here, we present pubmedKB, a web server designed to extract and visualize semantic relationships between four biomedical entity types: variants, genes, diseases, and chemicals. pubmedKB uses state-of-the-art natural language processing techniques to extract semantic relations from the large number of PubMed abstracts. Currently, over 2 million semantic relations between biomedical entity pairs are extracted from over 33 million PubMed abstracts in pubmedKB. pubmedKB has a user-friendly interface with an interactive semantic graph, enabling the user to easily query entities and explore entity relations. Supporting sentences with the highlighted snippets allow to easily navigate the publications. Combined with a new explorative approach to literature mining and an interactive interface for researchers, pubmedKB thus enables rapid, intelligent searching of the large biomedical literature to provide useful knowledge and insights. pubmedKB is available at https://www.pubmedkb.cc/.

GRAPHICAL ABSTRACT



INTRODUCTION

Following recent advances in genomic sequencing technologies, a large amount of genomic data has been generated in biomedical research. To interpret genomic data properly, most scientists and healthcare professionals need to use the biomedical literature effectively to identify relationships between biomedical terms and entities. Although a large collection of biomedical literature is indexed in PubMed, the diverse biomedical terminology used in the literature and the massive amount of data presented make the task of information retrieval repetitive and tedious. Researchers usually access the PubMed database through simple keyword searches or with some operational filters, and the search engine essentially displays the results that contain the exact search targets and ranks them by date of publication. Hence, finding comprehensive and contextual information in the large volume of scholarly publications, which is pre-

*To whom correspondence should be addressed. Tel: +886 7729 5753; Email: jiahsin.huang@gmail.com

sented as unstructured textual data, remains a significant challenge.

To date, a handful of knowledge bases have attempted to facilitate the extraction of relevant information regarding, and relationships between, medical concepts (1). Some knowledge databases, such as ClinVar (2), PharmGKB (3), and UniProtKB (4), have been manually constructed by domain-expert curators who read a large amount of publications, to disseminate knowledge about genetic entities, such as variants, genes, and diseases. However, this manual curation often suffers from scalability problems due to the recent vast increase in the number of publications (5–7). Indeed, it is nearly impossible for human curators to keep up with all the scientific publications. Also, researchers usually use PubMed queries to identify journal articles relevant to their specific interests. A recent survey of six knowledge bases on cancer genomic variants revealed a highly disparate curation: fewer than 20% of the publications were recorded in at least two of these databases (8). This situation suggests that a literature search engine that employs proper text-mining techniques is needed to obtain a comprehensive coverage of the relevant information.

In the past decade, the field of biomedical literature mining has yielded great progress in developing text mining and natural language processing (NLP) techniques for automated extraction from the biomedical literature (9). For example, LitVar uses named-entity recognition (NER) and named-entity normalization (NEN) methods to match the user input query with the indexed literature for searching publications concerning variants (10). In our previous work, we developed variant2literature, which uses the same NER and NEN methods to identify variants from not only the full-text parts of biomedical publications but also the attached supplementary files (11). Notably, variant2literature yields more than double the number of search results provided by a LitVar search of our manually curated dataset. Despite LitVar and variant2literature providing results containing the co-occurrence of different entity types highlighted in the texts, the use of the semantic relations between the entities (which is very important information in the biomedical domain) for further filtering the literature has not yet been explored.

Here, we present pubmedKB, a novel literature search engine that combines a large number of state-of-the-art text-mining tools optimized to automatically identify the complex relationships between biomedical entities—variants, genes, diseases, and chemicals—in PubMed abstracts. We focus on the extraction of biomedical entities that co-occurred in the same sentences in the abstracts. We consider such co-occurrences to be plausible indications of links between entities. Based on these four biomedical entity types (variants, genes, diseases, and chemicals), pubmedKB extracts more than 2 million relationships in the distinct pairs of entities, using its various relation extraction modules. pubmedKB also provides an interactive semantic graph to facilitate visualization and selection of the related entities and singles out of the publications that are relevant to the user query.

SYSTEM DESCRIPTION

Literature text annotation—biomedical entity extraction

Figure 1 provides a schematic system overview of pubmedKB. The data in PubMed Baseline 2022, which was released on 12 December 2021, was retrieved from the NCBI's FTP server to construct pubmedKB at the time of writing on January 2022. The PubMed Baseline 2022 contains about 33 million abstracts. Its entire set of abstracts from journal and book articles is used for text mining (Figure 1A). The pubmedKB database is currently being updated monthly.

Recognition of entities related to biomedical topics is an essential first step in extracting knowledge from medical papers. The NER task seeks to locate and classify entities mentioned in unstructured text. pubmedKB uses several NLP and deep-learning technologies to extract four biomedical entity types—variants, genes, diseases, and chemicals—from the PubMed abstracts (Figure 1B). The NER task can be treated as a sequenced labelling problem, in which we first tokenize sentences and then assign a label to each word. We use Natural Language Toolkit (NLTK) (12) for tokenization and a combination of the pretrained Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) (13) and a bidirectional long short-term memory (BiLSTM) network, based on a previous study (14), to produce a context that provides information on samples surrounding each token. In short, we use NLTK for tokenization and stack a BiLSTM on top of pre-trained BioBERT. The output of the BiLSTM is later fed to a fully connected network, which makes final predictions. The beginning, inside, outside (BIO) format is used for tagging tokens. Detailed descriptions of our NER module are provided in the Supplementary Materials.

After the NER task, we also need to disambiguate entity mentions recognized by the NER model, since different constructions can represent the same entity (Figure 1C). For example, 'coronary heart disease' and its abbreviation 'CHD' should both correspond to the disease with the Medical Subject Headings (MeSH) identifier (ID) 68003327. Specifically, we use multitype normalization models from BERN (15) as our NEN module to link mentions to the established databases; and we use tmVar 2.0 (16), GNormPlus (17), sieve-based entity linking (18), and tmChem (19) in BERN to normalize variants, genes, diseases, and chemicals, respectively. With NEN, we aim to give all entities a unique ID, facilitating matching of all the alias names in the literature and in user search terms.

Relation extraction modules

The pubmedKB Relation Extraction (RE) module comprises three submodules (Figure 1D): the Relational Phrases submodule extracts an open set of relational phrases from free text; the Relational Facts submodule performs sentence classification for a well-defined closed set of relational facts; and the Odds Ratio Info submodule is dedicated to extracting odds ratio statistics for genetic variants and their associated disease phenotypes from clinical trials and meta-analyses.

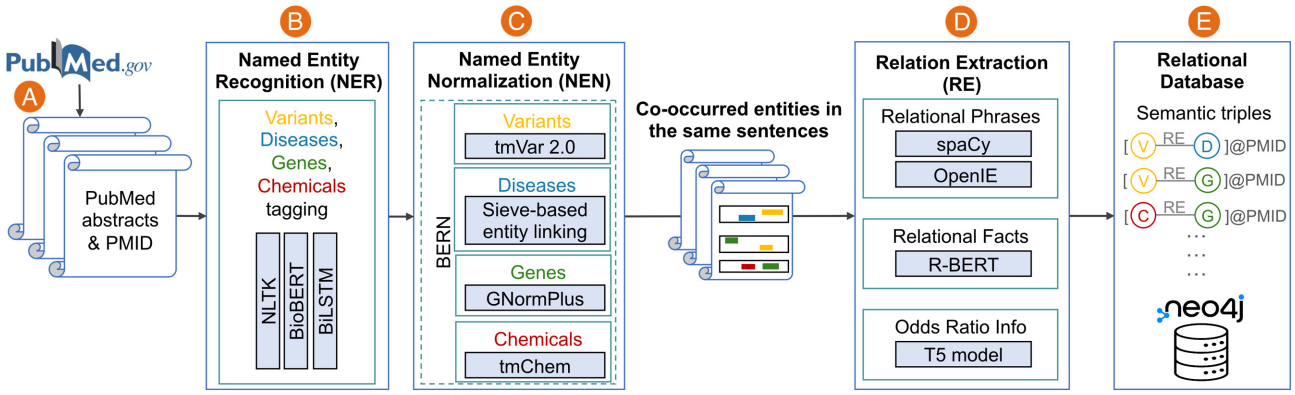


Figure 1. An overview of the pubmedKB processing workflow for text mining (**A–E**). The PubMed abstracts are retrieved form the NCBI FTP server. The journal abstracts are parsed individually to extract relationships between biomedical entities, which are stored in the Neo4j database.

Relational phrases submodule. We implement two algorithms to identify relational phrases that connect biomedical entities. The first algorithm is based on spaCy (https://spacy.io/), which is an accurate syntactic-dependency syntax parser. Notably, spaCy considers the text's grammatical structure to gain insight into how the words are related to each other. In brief, spaCy generates the dependency tree of each sentence. A triplet (subject, verb, object) is the direct result of spaCy dependency parse. Based on the extracted dependencies, we specifically extract verbs as relations connecting their subject and object entities. For example,

> '*TCF7L2* ***promotes*** *pathological retinal neovascularization via ER stress-dependent upregulation of VEGFA.*'

will be taken as an evidence sentence, based on the '***promotes***' relation between TCF7L2 and retinal neovascularization. However, the writing style used in many scientific articles often contains complex constructions forming long clauses in a sentence. We therefore also use Stanford OpenIE (20), which can extract relational triplets across large clauses. OpenIE splits each sentence into its entailed clauses and then divides these into relational triplets. We select those triplets whose entities correspond to entities extracted by our NER module and filter out redundant short relations. For example,

> '*ChIP-seq analysis revealed that TCF7L2* ***co-localizes with*** *HNF4alpha and FOXA2 in HepG2 cells and with GATA3 in MCF7 cells.*'

will be taken as an evidence sentence for the '***co-localizes with***' relation between TCF7L2 and GATA3.

Relational facts submodule. Understanding the role of genetic variants (V) in their associated diseases (D) is of great important in clinical practice (21). We therefore deploy a sentence-based relation classifier for a closed set of relations in sentences with variant and disease co-occurrence. We base our model on the R-BERT relation classification framework (22) and use a pretrained BioBERT (13) to initialize its encoder. The R-BERT classifier predicts which relational fact is between each variant–disease pair in a sentence.

Three types of relational facts are specifically defined as Cause-associated, Appositive, and In-patient. A triplet of (V, Cause-associated, D) is extracted for evidence sentences that state that V is causally related to D or is D-associated. Second, the apposition of genetic variants and diseases in a sentence could provide some useful information for clinical interpretations. The triplet (V, Appositive, D) applies to sentences containing expressions such as '… the D variant V …', '… V, a D variant, …', or '… novel D mutation V …'. Finally, the triplet (V, In-patient, D) defines sentences reporting D patients that carry V. We provide samples of sentences containing the specific relations for better understanding. For example,

> '*A comparison of the allele frequencies or genotype distributions by the* $\chi 2$ *test revealed that rs6929846 of BTN2A1 was significantly associated with dyslipidemia (*$P < 0.05$*).*'

is taken as an evidence sentence for the '*Cause-associated*' relation between rs6929846 and dyslipidemia. The second example of the 'Appositive' relation,

> '*Single-marker analysis revealed significant associations between 2 recently identified candidate schizophrenia susceptibility variants rs1344706 (and rs7597593) and ….*'

is taken as an evidence sentence for the 'Appositive' relation between schizophrenia *and* rs1344706. Finally, as an example of the In-patient relation,

> '*This is a second missense G564T mutation in another VHL patient from Kuwait that will help expand our knowledge of the VHL gene mutation spectrum in this region of the world.*'

is taken as an evidence sentence for the '*In-patient*' relation between G564T and VHL, to emphasize the fact that this discovery was made using patient information.

Odds ratio info submodule. In addition, we use a standalone model that specializes in extracting odds ratio statistics between genetic variants and diseases. We use the pretrained general-purpose sequence-to-sequence model T5 (23) as our base extractor. For example,

'*The adjusted risk of having CAD was more evident for rs1799782 (OR = 1.53; 95% CI: 1.16–2.02; P = 0.003), rs1801133 (OR = 1.54; 95% CI: 1.22–1.94; P < 0.001), and rs4846049 (OR = 1.74; 95% CI: 1.13–2.69; P = 0.013) under the recessive model.*'

is taken as an evidence sentence for three odds ratio relations: (rs1799782, CAD, 1.53, 1.16–2.02, 0.003), (rs1801133, CAD, 1.54, 1.22–1.94, 0.001), and (rs4846049, CAD, 1.74, 1.13–2.69, 0.013).

System implementation

The pubmedKB web server is implemented in NodeJS (https://nodejs.org/en/) using the ExpressJS web development framework. We use the NGINX HTTP server (https://www.nginx.com/) as a gateway, representational state transfer (REST) application programming interface (API) clients, and the Neo4j graph database server to store the semantic triplets of entity pairs and their relations. Neo4j (https://neo4j.com/) provides efficient mechanisms for retrieving relations between two entities in a relational database (Figure 1E). We have built an external entity-mapping API to map user inputs to corresponding BERN IDs and use these IDs to extract the relations and evidence in a sentence from the Neo4j database. The web interface is based on ReactJS (https://reactjs.org/), a modern JavaScript library, to ensure rapid rendering and high performance. We perform page rendering and reference sorting using a JavaScript object notation (JSON) front-end for better efficiency. The semantic graph is built using the antv graphin library (https://graphin.antv.vision/) for better visualization. The pubmedKB web server supports all commonly used internet browsers, although we recommend using it with Chrome for the best experience.

Search bar processing

For the search bar on the pubmedKB website, we use the same NER and NEN modules to disambiguate alias names in user queries as in journal texts. Search terms, separated by commas, are treated as entity mentions. The NER module classifies their types, and then the corresponding NEN submodules are used to retrieve unambiguous BERN IDs. Finally, we match the query terms and their BERN IDs with entities in pubmedKB to retrieve the relational knowledge and evidence sentences mined from the PubMed abstracts. It should be noted that the search bar is case sensitive, but it includes an auto-suggestion function to help users input the correct names of entities. By default, the search will look for relations with entities of all four types (genes, variants, diseases, chemicals), but the user can specify entity types to search for.

RESULTS

Relational knowledge extracted from the PubMed abstracts

The core concept of pubmedKB is that it is an alternative search engine for exploring PubMed abstracts via the extraction of relations mentioned in the literature. Drawing on the data in PubMed Baseline 2022, there are 422 922 unique normalized IDs with relations in pubmedKB, of which about a quarter are single-text mentions, while the rest involve the normalization of several synonymous names by the NEN module. With respect to the semantic relations, there are 1 514 508 OpenIE triplets, 214 870 spaCy triplets, 517 284 R-BERT triplets, and 39 657 odds ratio statistics in pubmedKB. Notably, OpenIE extracts more than 10 times the number of relations that spaCy does. The distribution of the top 50 relational phrases suggests that the two algorithms prefer different sets of relational words (Supplementary Figure 1; see complete dataset in Supplementary Table S1). Although OpenIE extracts a variety of general words that may express nonrelevant relations, it provides a quality index that indicates to what extent the two entities are written in a well-formed sentence (24). In contrast, the relational phrases extracted by spaCy are mostly verb terms that refer to a biomedical action or mechanism. The spaCy results, therefore, are more likely to identify potentially relevant biological relations between entity pairs. Disease type has the most relations with other types of entities in the PubMed abstracts (Table 1). Next, we compared the entity pairs in pubmedKB to the corresponding curated PharmGKB knowledge base, which continues to manually curate relevant knowledge concerning the four types of entities. Overall, pubmedKB achieved an averaged coverage rate of 74.91% of the identified associations in the manually curated entity pairs reported by PharmGKB (Supplementary Figure S2). It is notable that more than 90% of disease-disease, gene-disease, and gene-gene pairs curated in the PharmGKB are presented in the pubmedKB with the supporting sentences and corresponding publications.

Table 1. Distinct entities and distinct entity pairs (unordered) in the pubmedKB relational database

	Entities (n)	Entity pairs (n)			
		Variants	Diseases	Genes	Chemicals
Variants	125 745	3 558	315 452	6 673	1 926
Diseases	71 787	315 452	429 723	40 070	148 156
Genes	45 315	6 673	40 070	33 470	37 074
Chemicals	180 075	1 926	148 156	37 074	360 931

Improvement of R-BERT and odds ratio information extraction for variant–disease relations

Since genetic variant information is becoming increasingly important for precision medicine, pubmedKB uses two classifiers to extract additional information from sentences in which genetic variants and diseases co-occur (V–D pairs).

To train the R-BERT model, we manually labelled about 1 500 samples (2:1 train–test split) from ClinVar (2) abstracts and achieved an F_1 score of 84.1%. We then leveraged iterative teacher–student self-learning with nearly 13 000 unlabelled samples to improve the performance to 91.0%. Our implementation is similar to the techniques described by Xie *et al.* (25). In each retraining iteration, a new model is trained on both the gold and the silver data labelled in the previous iteration by the old model. The main difference from the original image-classification study is that we do not noise the student.

Since the odds ratio and statistical *P*-value information are two of the preferred methods for estimating the degree of association between genetic variants and their respective diseases (26), we used the pre-trained T5 model to extract odds ratio information from the sentences in which variants and diseases co-occurred. Notably, to fine-tune T5 for odds ratio statistics, we manually labelled around 3 000 abstracts from ClinVar (2) and DisGeNET (27) articles and achieved rouge-1, rouge-2, and rouge-L scores of 93.9%, 93.3%, and 94.0%, respectively.

User interface and system features

pubmedKB can be accessed via a simple search and interactive graphical interface (Figure 2). After the user inputs a plain text query comprising a single biomedical entity, or two entities separated by a comma, in the search bar (Figure 2A), pubmedKB normalizes the query to identify an unambiguous BERN ID to use to retrieve semantic triplets from the pubmedKB Neo4j database. pubmedKB returns two main features. First, for each result, it provides a semantic graph showing the relevant entities and the relational evidence associated with the user query (Figure 2B). To ensure a good user experience when navigating the semantic graph, only the top 50 relevant entities with the most relations are shown. The four entity types are represented by different colours. The user can click a particular entity to see the relation information specific to that entity. pubmedKB also provides node selection (Figure 2C) to allow the selection of entities, which are ranked by the number of relations found for them. At the top of the semantic graph, three relation-filtering options enable the user to easily select the linking entities based on specific relations (Figure 2D). This is particularly useful for detecting specific types of relations (e.g. Cause-associated relations for variant–disease pairs). Second, pubmedKB provides the list of supporting publications, with their associated relational sentences, in the right panel (Figure 2E). When the user selects specific linking entities, the publication list will be automatically updated to show only the relevant papers for that linking entity. Of note, all search results (not limited to top 50 entities) including the evidence sentences and PMIDs can be obtained through the Download to explore all the relevant information (Figure 2F). The user can also sort the publications by relevance, publication date, and impact factor (Figure 2G). For each publication, pubmedKB provides the citation information details and a PMID link, which redirects the user to PubMed (Figure 2H). In addition, pubmedKB displays one or more snippets highlighting the relational entities as supporting evidence (Figure 2i).

USE CASES

Below we demonstrate two scenarios of how pubmedKB may be used by scientists.

First, the user can query a single entity to find related entities and corresponding publications. For example, if they start by searching 'rs4961', pubmedKB returns a semantic graph that displays many related entities extracted from the PubMed abstracts (Figure 3A). The design of this semantic graph enables the user to navigate related entities easily and select the publications they wish to investigate further. The user can narrow down the entity nodes via the node selection function, by typing in a free-text search box or marking checkboxes (Figure 3B). They can then click a particular entity node to see the detailed relationships with that entity and the corresponding publications (Figure 3C). Alternatively, the user can choose specific relations by selecting relational options in the top panel of the semantic graph (Figure 3D), and select a particular node to see its relation information (Figure 3E). After selecting a specific related entity, the user can simply click in the blank area to go back to the previous node selection status and easily navigate to another entity of interest.

Alternatively, users can search for two entities, which must be separated by a comma in the search bar, to query for any relationships between them that are represented in the PubMed abstracts. If the relationships are recorded in the pubmedKB database, pubmedKB will return the relationship details as shown in Figure 3C and E.

DISCUSSION

The pubmedKB web server is designed to allow swift exploration of the biomedical concepts in the massive collection of publications in PubMed. For example, when 'rs4588' is used as a query input, PubMed and LitVar return 187 and 400 publications, respectively, while pubmedKB returns 55. Although pubmedKB returns fewer results, all 55 publications contain at least one sentence with supporting evidence concerning the relations of rs4588 with other entities.

For relation extraction, other web servers, such as STRING (28), GeneView (29), and LitVar (10), rely on sentence co-occurrence to infer relations between entities, so their results may include many false positives. Indeed, the NER module in pubmedKB does extract sentence co-occurrences of entity pairs. However, because we have applied advanced text-mining techniques to extract specific types of relations, simple sentence co-occurrences are not included in the results output by pubmedKB, ensuring the high quality of the final results.

There are several limitations to pubmedKB. First, the semantic evidence used for extracting entity information is bound to the performance of the NLP algorithms. As mentioned previously, we have self-annotated a small batch of publications to improve the performance of the R-BERT and T5 models. Second, inconsistent use of biomedical terms and writing styles may result not only in poor document annotation but also in search queries being matched to the wrong existing entities in the pubmedKB database. Third, pubmedKB endeavours to annotate the large volume of PubMed abstracts, but a search query may nevertheless return no results, either because there are no relations in the PubMed literature or because it is not mentioned in the abstract texts.

In the future, we hope to extend the current PubMed abstracts dataset to include the PubMed Central full-text articles, although this will require improving the search performance in speed and accuracy to deal with the substantial increase in the size of the dataset.

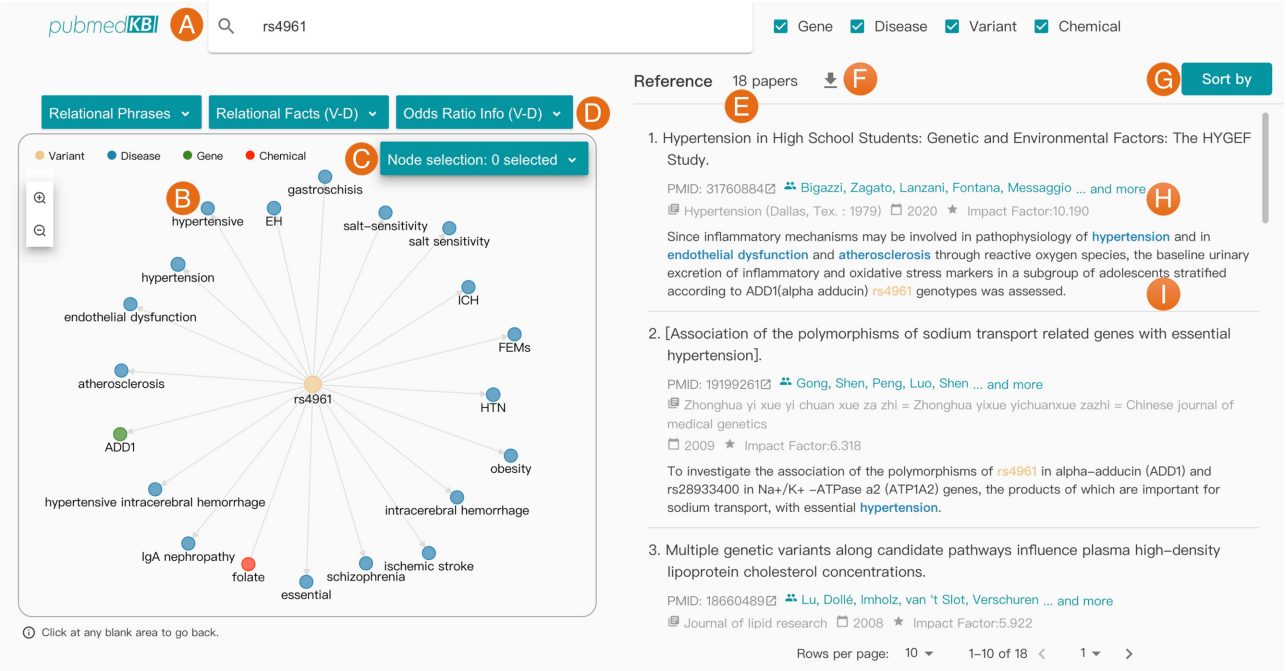


Figure 2. pubmedKB user interface. (**A**) Search bar for entering queries; (**B**) a semantic graph; (**C**) node selection; (**D**) relationships filters; (**E**) number of supporting publications containing relational evidence; (**F**) a download function to retrieve all results as a .csv file; (**G**) filter options with publication information; (**H**) article information, including the PMID link to take the user directly to PubMed; (**I**) supporting-evidence sentences with the relevant entities highlighted in colour.

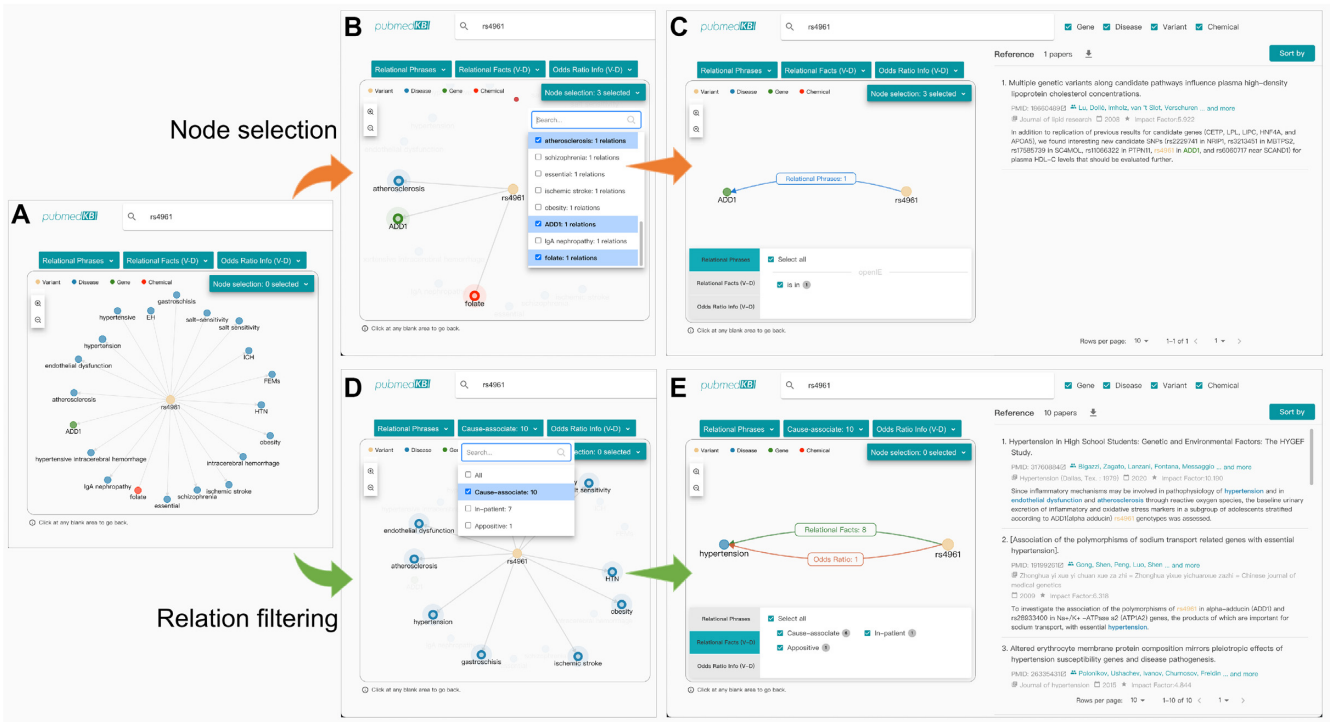


Figure 3. Interactive snapshots of pubmedKB filters. (**A**) If 'rs4961' is queried, many related entities are shown. (**B**) Node selection is used to select one or more node entities. (**C**) If the 'ADD1' entity is clicked, the relevant relationship details and corresponding publications are shown. (**D**) The relation filters at the top of the graph panel provide edge filters for selecting node entities. (**E**) If 'hypertension' is selected by clicking the node, the relationship details and corresponding publications are shown.

DATA AVAILABILITY

pubmedKB is an open-access resource and is publicly available at https://www.pubmedkb.cc/. The scripts of core text-mining modules in the pubmedKB and the datasets of in-house labelling sentences are freely available on GitHub at https://github.com/jacobvsdanniel/pubmedkb_core.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank three domain experts, Yi-Chieh Chen, Yi-Wei Cheng, and Ko-Han Li, for contributing annotations of the biomedical papers to improve our NER models.

FUNDING

This work was supported by the Ministry of Science and Technology, Taiwan (MOST 109-2221-E-002-161-MY3).
Conflict of interest statement. None declared.

REFERENCES

1. Borchert,F., Mock,A., Tomczak,A., Hügel,J., Alkarkoukly,S., Knurr,A., Volckmar,A.-L., Stenzinger,A., Schirmacher,P., Debus,J. *et al.* (2021) Knowledge bases and software support for variant interpretation in precision oncology. *Briefings Bioinf.*, **22**, bbab134.
2. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
3. Thorn,C.F., Klein,T.E. and Altman,R.B. (2013) PharmGKB: the pharmacogenomics knowledge base. *Methods Mol. Biol.*, **1015**, 311.
4. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.
5. Poux,S., Arighi,C.N., Magrane,M., Bateman,A., Wei,C.-H., Lu,Z., Boutet,E., Bye-A-Jee,H., Famiglietti,M.L., Roechert,B. *et al.* (2017) On expert curation and scalability: uniprotkb/Swiss-Prot as a case study. *Bioinformatics*, **33**, 3454–3460.
6. Baumgartner,W.A. Jr, Cohen,K.B., Fox,L.M., Acquaah-Mensah,G. and Hunter,L. (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, **23**, i41–i48.
7. Lee,K., Wei,C.-H. and Lu,Z. (2021) Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. *Briefings Bioinf.*, **22**, bbaa142.
8. Wagner,A.H., Walsh,B., Mayfield,G., Tamborero,D., Sonkin,D., Krysiak,K., Deu-Pons,J., Duren,R.P., Gao,J., McMurry,J. *et al.* (2020) A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat. Genet.*, **52**, 448–457.
9. Zhao,S., Su,C., Lu,Z. and Wang,F. (2021) Recent advances in biomedical literature mining. *Briefings Bioinf.*, **22**, bbaa057.
10. Allot,A., Peng,Y., Wei,C.-H., Lee,K., Phan,L. and Lu,Z. (2018) LitVar: a semantic search engine for linking genomic variant data in pubmed and PMC. *Nucleic Acids Res.*, **46**, W530–W536.
11. Lin,Y.-H., Lu,Y.-C., Chen,T.-F., Hsu,J.S., Lee,K.-H., Cheng,Y.-W., Chen,Y.-C., Fan,J.-S., Tu,C.-T., Hsu,C.-M. *et al.* (2019) variant2literature: full text literature search for genetic variants. bioRxiv doi: https://doi.org/10.1101/583450, 04 June 2019, preprint: not peer reviewed.
12. Loper,E. and Bird,S. (2002) NLTK: the natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1, ETMTNLP '02*. Association for Computational Linguistics, USA, pp. 63–70.
13. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.
14. Yu,X., Hu,W., Lu,S., Sun,X. and Yuan,Z. (2019) BioBERT based named entity recognition in electronic medical record. In: *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*. pp. 49–52.
15. Kim,D., Lee,J., So,C.H., Jeon,H., Jeong,M., Choi,Y., Yoon,W., Sung,M. and Kang,J. (2019) A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, **7**, 73729–73740.
16. Wei,C.-H., Phan,L., Feltz,J., Maiti,R., Hefferon,T. and Lu,Z. (2018) tmVar 2.0: integrating genomic variant information from literature with dbSNP and clinvar for precision medicine. *Bioinformatics*, **34**, 80–87.
17. 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**, e918710.
18. D'Souza,J. and Ng,V. (2015) Sieve-Based entity linking for the biomedical domain. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pp. 297–302.
19. Leaman,R., Wei,C.-H. and Lu,Z. (2015) tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, **7**, S3.
20. Angeli,G., Johnson Premkumar,M.J. and Manning,C.D. (2015) Leveraging linguistic structure for open domain information extraction. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pp. 344–354.
21. Eilbeck,K., Quinlan,A. and Yandell,M. (2017) Settling the score: variant prioritization and mendelian disease. *Nat. Rev. Genet.*, **18**, 599–612.
22. Wu,S. and He,Y. (2019) Enriching Pre-trained language model with entity information for relation classification. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*. Association for Computing Machinery, NY, pp. 2361–2364.
23. Raffel,C., Shazeer,N., Roberts,A., Lee,K., Narang,S., Matena,M., Zhou,Y., Li,W. and Liu,P.J. (2020) Exploring the limits of transfer learning with a unified Text-to-Text transformer. *Journal of Machine Learning Research*, **21**, 1–67.
24. Kadry,A. and Dietz,L. (2017) Open relation extraction for support passage retrieval: merit and open issues. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Shinjuku Tokyo Japan, pp. 1149–1152.
25. Xie,Q., Luong,M.-T., Hovy,E. and Le,Q.V. (2020) Self-training with noisy student improves imagenet classification. arXiv doi: https://arxiv.org/abs/1911.04252, 19 June 2020, preprint: not peer reviewed.
26. Uffelmann,E., Huang,Q.Q., Munung,N.S., de Vries,J., Okada,Y., Martin,A.R., Martin,H.C., Lappalainen,T. and Posthuma,D. (2021) Genome-wide association studies. *Nat Rev Methods Primers*, **1**, 1–21.
27. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
28. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
29. Thomas,P., Starlinger,J., Vowinkel,A., Arzt,S. and Leser,U. (2012) GeneView: a comprehensive semantic search engine for pubmed. *Nucleic Acids Res.*, **40**, W585–W591.