OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–9

https://doi.org/10.1093/bib/bbac465

Advance access publication date 8 November 2022

Problem Solving Protocol

e-TSN: an interactive visual exploration platform for target–disease knowledge mapping from literature

Ziyan Feng, Zihao Shen, Honglin Li and Shiliang Li

Corresponding author: Shiliang Li, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China; Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200062, China. E-mail: shiliangli@ecust.edu.cn

Abstract

Target discovery and identification processes are driven by the increasing amount of biomedical data. The vast numbers of unstructured texts of biomedical publications provide a rich source of knowledge for drug target discovery research and demand the development of specific algorithms or tools to facilitate finding disease genes and proteins. Text mining is a method that can automatically mine helpful information related to drug target discovery from massive biomedical literature. However, there is a substantial lag between biomedical publications and the subsequent abstraction of information extracted by text mining to databases. The knowledge graph is introduced to integrate heterogeneous biomedical data. Here, we describe e-TSN (Target significance and novelty explorer, http://www.lilab-ecust.cn/etsn/), a knowledge visualization web server integrating the largest database of associations between targets and diseases from the full scientific literature by constructing significance and novelty scoring methods based on bibliometric statistics. The platform aims to visualize target–disease knowledge graphs to assist in prioritizing candidate disease-related proteins. Approved drugs and associated bioactivities for each interested target are also provided to facilitate the visualization of drug–target relationships. In summary, e-TSN is a fast and customizable visualization resource for investigating and analyzing the intricate target–disease networks, which could help researchers understand the mechanisms underlying complex disease phenotypes and improve the drug discovery and development efficiency, especially for the unexpected outbreak of infectious disease pandemics like COVID-19.

Keywords: target discovery, text mining, knowledge graphs, visualization

Introduction

Drug discovery and development is one of the most important translational science activities for promoting human health and well-being. The new drug discovery is a highly complicated, expensive and long process that costs $2.6 billion and takes an average of 12 years [1]. Target discovery and identification is the starting point of drug development and the most critical step to the mechanism-based drug discovery campaign's success in diagnosing and fighting human diseases [2]. The discovery of macromolecular targets for bioactive drugs is a bottleneck in the design of chemical probes and drug leads [3]. The record shows that the high failure rate of drug development is primarily due to inappropriate drug targets in the early preclinical stages [4, 5]. Therefore, research and knowledge of therapeutic targets related to specific diseases need to be identified, which can be advantageous in speeding up the drug discovery process. This requires bioinformatics identification, cell and genetic target assessment and genomic and proteome analysis.

The rapid accumulation of biomedical data has led to a paradigm shift in the pharmaceutical industry from phenotype-based discovery to target-based approaches. In the target identification phase, many techniques and open-access databases have provided complementary and comprehensive data, which are used to find and isolate a target, learn more about its functions and prove how these functions influence diseases. In biomedical science, the target in the drug discovery process mainly refers to a molecule such as a gene, protein or miRNA in the body that is intrinsically associated with a particular disease procession and could be actioned by drugs to produce a desired therapeutic effect [6]. Because the pathological and biological mechanisms of human diseases are quite complex, the key task in target discovery is not only to identify, preferentially select reliable 'druggable' targets but also to search for disease-target potential interactions to understand the fundamental molecular mechanism underlying most human diseases [7]. This still requires manual gathering and exploration of multiple data sources, which

Ziyan Feng is a Ph.D. candidate from the School of Pharmacy at East China University of Science and Technology. Her research interests include bioinformatics, computational biomedicine and deep learning.

Zihao Shen is a Ph.D. candidate from the School of Pharmacy at East China University of Science and Technology. His research interests are artificial intelligence, computer vision and deep learning.

Honglin Li is a professor of medicinal chemistry and computational chemistry. He is the director of the Shanghai Key Laboratory of New Drug Design at East China University of Science and Technology and the director of the Innovation Center for AI and Drug Discovery at East China Normal University. His group focuses on artificial intelligence, target discovery and drug design, computational biology and cheminformatics.

Shiliang Li is an associate professor of pharmacy at East China University of Science and Technology and a young investigator of the Innovation Center for AI and Drug Discovery at East China Normal University. Her research interests are bioinformatics, cheminformatics, target discovery and drug design.

Received: July 7, 2022. **Revised:** September 20, 2022. **Accepted:** September 27, 2022

requires significant time and expertise in the biomedical domain. Even with dedicated effort in capturing such information in biomedical databases, much of the potential information remains 'locked' in the unstructured text of biomedical publications [8, 9].

Nevertheless, with the explosion of biological data and information, the overwhelming amount of biomedical knowledge recorded in the scientific literature is increasing rapidly. For example, MEDLINE/PubMed (https://pubmed.ncbi.nlm.nih.gov), the most popular biomedical literature database in the world, now has more than 33 000 000 citations and abstracts of biomedical literature. PubMed Central (https://www.ncbi.nlm.nih.gov/pmc), the full-text database of biomedicine, currently contains more than 7 800 000 full-text records and is growing at a more than 10% compounded annual growth rate [9, 10]. As a result, thousands of articles can be found for the most well-studied diseases or targets. This wealth of biological data and information provides immense new opportunities for target discovery and supports the drug discovery pipeline. However, the phenomenal growth of the biomedical literature has made it harder than ever for scientists to find and assimilate all the publications relevant to their research since the traditional information retrieval modes cannot address this information overload [11], and increasingly more demand of biomedical scientists for assistance in assimilating the high rate of new publications.

With the rapid growth of biological databases, the prosperity of bioinformatics, especially data mining approaches, the combination of biological ideas with in silico approaches or statistical methods to extract or predict valuable targets has changed the way of target discovery [7]. Data-driven methods can generate hypotheses and construct knowledge maps in the context of human disease from high-throughput data to improve understanding of biological processes and functional relationships [12]. Currently, automatic text processing and analysis (referred to as text mining [TM]) [13] is a prevailing approach that can assist researchers in evaluating biomedical literature. Mining unstructured text consisting of entities such as genes, proteins, diseases, symptoms and drugs can speed up critical prediction tasks such as new treatment discovery or drug repurposing [14]. Nowadays, biomedical TM has been extensively applied to identify biological entities such as protein, drug and microbe to understand their roles in diseases by finding information automatically from the scientific literature [15–21]. Many text-mining tools such as PubTator [22], PolySearch [23], Chemotext [24], DISEASES [25], miRCancer [26] and BIONDA [27] have been developed to mine and organize the relevant textual information.

Biomedical text mining has made significant progress, but it is still not enough to be applied in the extensive and profound biomedical field. Several challenges remain to be solved: (a) Such text-mining systems make use of information collected from abstracts that comprise shorter sentences and very concise text presenting only the most significant findings rather than the full-text body, which contains more information. With full-text articles becoming more accessible, there is a growing interest in the text mining of complete articles [28]. (b) Current text-mining tools specialize in specific tasks such as identifying certain types of entities and relations from a single literature. However, there is a substantial lag between biomedical publications and the subsequent abstraction of information extracted by text mining to databases. It is of great importance to view text mining not only as an isolated problem but also as a means to integrate the literature information with other relevant data in particularly established resources that have a broad user base [25]. Big data analysis should be developed due to its ability to turn huge amounts of data into other information to aid subsequent analysis and knowledge discovery [29]. (c) The data generated by text mining are mostly stored in the database, making it difficult to intuitively read the information. Knowledge graphs (KG) and visualization help get a complete view of data values [29] and provide heterogeneous data, including multiple types of entities (e.g. diseases, targets and drugs) and relations (e.g. disease-target pairs) [30–32]. Text-mining analytics and knowledge graph visualization should be integrated seamlessly so that they work best in drug-target discovery applications.

Our current work focuses on understanding the proteins and related mechanisms associated with disease from the published literature. Herein, we developed a new and versatile web-based tool utilizing text-mining, big data, knowledge graph and visualization technologies, termed e-TSN, which aims to be the most comprehensive and complementary web server to visualize disease and target associations. By using named entity recognition (NER) [33] and relation extraction (RE) [34] to extract the relationship between targets and diseases from numerous full-text literature, we presented two new scoring schemes based on bibliometric indices: significance and novelty, which may play a pivotal role in the pathological and biological mechanisms behind human diseases. The significance score represents the relative strength of the association between targets and diseases, which is an important feature of the two entities. The novelty score represents the potential and value of the given entity (target or disease). A novel target means a new biological response mechanism and a new drug discovery pipeline. The scoring scheme simultaneously considers co-occurrences and semantic analysis at the level of individual sentences. After data integration is implemented and valuable information is extracted, we further constructed a web tool e-TSN, for prioritizing the diseases-targets associations aimed at helping users who are interested in individual diseases or targets by visualizing novelty and significance on two-dimensional scatter plots. It would be extremely valuable to construct a knowledge graph of disease–target–drug relationships for identifying and prioritizing under-researched genes and proteins as potentially novel drug targets for further investigation and validation. Moreover, e-TSN can also help find new alternative therapeutic applications for existing drugs.

Materials and methods

Text mining, also known as text data mining, aims to obtain tacit knowledge by uncovering interesting and often hidden relationships in unstructured text to facilitate new drug target discovery and drug repurposing. The system architecture includes components for scientific documents download, preprocessing, named entity recognition, relation extraction, knowledge discovery and visualization. Figure 1 depicts the simplified architecture behind the application. The following sections briefly describe the details of each component.

Data selection and processing

In this section, the datasets used in this study are briefly introduced.

The biomedical graph used in this study is constructed based on text-mining knowledge extracted from PubMed's full literature. The full texts were gathered from the PubMed Central database, an available publication archive of biomedical and life sciences journal literature comprising more than 7.8 million full-text records and spanning several centuries of biomedical and life science research [9]. To date, more than 2.8 million PMC articles

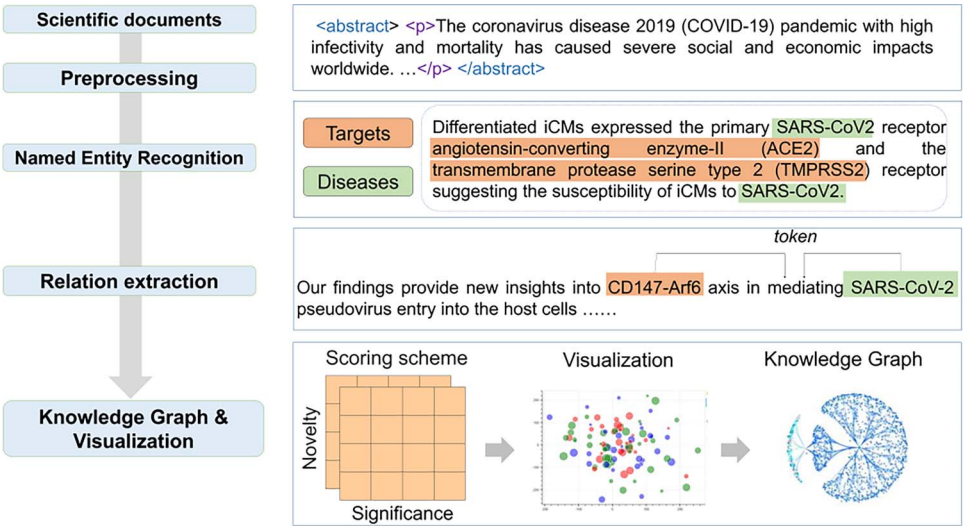

Figure 1. Architecture of the e-TSN web application. The workflow involves several stages of scientific documents download, preprocessing, named entity recognition, relation extraction, knowledge discovery and visualization.

are available for free text mining. The totality of PubMed Central Open Access (PMC OA) full texts was downloaded from NCBI's FTP site (http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/) in XML format for the following text-mining analysis. In short, XML documents consist of elements (root elements and child elements), which are textual data structured by tags. The element composition includes start/end tag pairs, optional attributes defined as key/value pairs and data between the two tags. The article's publication information, such as title, abstract, full text, DOI, is stored in the tag pairs (Figure 1). Parsing XML is a process designed to read XML and create a way for programs to use XML. We extracted the paragraphs from the full literature by parsing the XML file using a Python ElementTree module to transform the entire document into a DOM tree and utilizing XPath syntax to find the values in tag pairs.

In addition, we integrated an extensive collection of biomedical dictionaries for target and disease entities. When performing text mining, dictionary methods have limited recall due to the large number of entities that dictionaries need to cover. In biomedicine, however, the entity's scope is generally better defined and limited. Therefore, it is feasible and straightforward to make the dictionary cover all terms of a particular type comprehensively. To construct a dictionary of targets for NER, the human protein and gene names were mainly gathered from several public databases, including Target Central Resource Database (TCRD) [35], UniProt [36] and ChEMBL [37]. The TCRD collates abundant heterogeneous gene/protein datasets, and the targets were categorized as following protein families: enzymes, epigenetics, G-protein coupled receptors (GPCRs), orphan G-protein coupled receptors (oGPCRs), ion channels, kinases, nuclear receptors (NRs), transcription factors and transporters and non-IDG. TCRD builds a high-level classification scheme called the Target Development Level (TDL) based on the data collected for each target to objectively describe the target availability level in human proteins. TDL represents the extent to which they are studied or not studied, as evidenced by publications, small-molecule interventions, disease relationships and other characteristics [35].

For disease, disorders and human clinical phenotypes, Disease Ontology (DO) is a database for unifying disease annotations across species that contain clinical vocabularies (e.g. OMIM, ORDO, MeSH) [38]. We collected all non-redundant diseases' names and their synonyms, which represent common key concepts found in medical literature from the DO [38] and Orphanet, including infectious disease, disease of anatomical entity, disease of cellular proliferation, disease of metabolism, disease of mental health, genetic disease, physical disorder, syndrome and rare disease.

Entity identification

Recognizing named entities (NER) [33] from literature is the basis for most biomedical text-mining applications. NER is used to identify biomedical entities which have important meanings for scientific discovery, such as disease names, symptoms and target names. Current biomedical named entity recognition technology is mainly divided into dictionary-based, rule-based and machine learning methods. We used the dictionary-based methods that rely on matching a dictionary of names against texts and have the crucial advantage of being able to normalize names [25, 39]. Dictionary entries were specified by a normalized named entity and one or more variants (synonyms). To match a document against the conducted dictionary, we have developed a highly exact match algorithm based on RegExp implemented in Python to map target and disease entities from biomedical full-text. We use a custom hash table to store the dictionary for quick look-up when dealing with changes in case and spaces and hyphens for multiple words. The hashtable is case-insensitive, ignores the hyphen in the name and removes other punctuation characters at the beginning and end of the name, such as quotes and parentheses. We then calculated each paper's target and disease frequency for the subsequent evaluation score.

Relation extraction

Once entities are located in the text, the next step is to find the relationships between them. The relation extraction task includes identifying the relationships between the identified entities in the unstructured text of interest. Due to the complexity of sentences, relation extraction is one of the most challenging tasks in text mining of biomedical literature. According to preview studies, the conventional relationship extraction was most estimated based on the co-occurrence assumption [40], which is the fundamental approach in biomedical text mining and can be used to explore the relationships between two entities. However, this method has

difficulty in distinguishing direct and indirect associations. Natural language processing (NLP) techniques typically mine textual knowledge from existing literature to extract meaningful semantic associations between two biomedical concepts from selected knowledge sources. To implement the extraction of more complex associations at the level of sentences between targets and diseases, we further introduced the Natural Language Processing (NLP) [41] algorithm to extract knowledge from an unstructured text by using Lexical-syntactic patterns defined manually or generated automatically (Figure 1). To extract interaction associations from millions of full-text publications, we prioritize relationship extraction speed over complexity. After utilizing the Stanford CoreNLP's application for sentence splitting, we identified the relationships between targets and diseases by using a list of controlled verbs and nominalization terms that are used to describe relations of interest. In our case, if two entities simply co-occur in a sentence, we detected the verb that contains relationship between two entities in a sentence and can reasonably assume that they tend to be related.

The significance and novelty evaluation schemes

Knowledge discovery is a critical and effective part of text mining and a process of creating new discoveries from a large amount of structured or unstructured data, such as identifying new drug targets or new biomarkers for cancer diagnosis [13, 41]. This makes it a challenge to determine appropriate statistical criteria for target–disease associations. In order to facilitate new target discovery from multiple biomedical texts, we score associations between targets and diseases using the significance ($S(i,j)$) and novelty ($N(i)$) scoring scheme based on biometrics. The significance represents the relative strength of the association between targets and diseases. Almost all co-occurrence methods adopt a frequency-based scoring scheme, but a pair of entities or concepts may occur several times at the same time without any connection. Therefore, we introduced a scoring scheme that simultaneously considers co-occurrences and semantic relationships between two entities at the level of individual sentences. While the novelty represents the potential and value of the given entity (target or disease), a novel target means a new cellular mechanism or phenotypes of human diseases. All full-text articles from PubMed Central were text mined and ranked according to our scoring scheme. The scores are pre-computed using the following formulae:

$$S(i,j)_{=}\sum_{k=1}^{n}\frac{3C(i,j)+5R(i,j)+T(i)+D(j)}{\sum 3C+\sum 5R+\sum T+\sum D}$$

For literature (k), where $C(i, j)$ is the number of sentences that target (i) and disease (j) co-occur, $R(i, j)$ stands for the number of sentences existing verbs to indicate the relationship between the target (i) and disease (j), $T(i)$ and $D(j)$ signify the frequency of target (i) and disease (j) respectively of paper (k). The normalizing factor C is the sum over all pairs of targets and diseases and R is the sum over all pairs of related proteins and diseases

$$N(i)=1/\sum_{k=1}^{n}\frac{m*F(i)}{\sum F}$$

where $F(i)$ represents the frequency of target (i) or disease (j) in literature (k), m stands for the number of targets or diseases and n signifies the summation over all publications, including target (i) or disease (j).

Web server development

The e-TSN web server (http://www.lilab-ecust.cn/etsn/) involves two main functions: (1) building the knowledge graph that outlines causal associations between targets and disease, and providing useful searchable archives which allow users to either query for a target name to find associated diseases or query for a disease to discover new potential targets according to the significance and novelty distribution in the scatter plot. (2) supporting information, including approved drugs and associated bioactivities, to facilitate visual exploration of disease–drug–target relationships.

The web interfaces are implemented by using Java on Apache Tomcat Server. In addition to the most recent web standards, such as HTML 5 and CSS 3, we implemented the front-end web interface mainly using Bootstrap web framework (https://getbootstrap.com/). The visualization is implemented by utilizing the ECharts (https://echarts.apache.org/zh/index.html), which is a declarative framework for the rapid construction of web-based visualization. Several Javascript libraries, such as JQuery, were used explicitly for drag-and-drop functions.

The back-end was developed in Java programming language and used the open-source Spring application framework. Data access is provided by a Java Database Connectivity interface, with MySQL as the database system for storing the score data and the task details. A representational state transfer interface is provided to allow the front end to query information from the database process. The website requires a browser that supports HTML5 and will work well on most major browsers such as Chrome, Opera, Firefox, Edge and IE11.

Results

Characteristics of datasets

This work intends to provide an interactive KG platform based on literature to facilitate target discovery and indication expansion. The relationships between diseases and targets support the vital structure of the e-TSN database. The current version database contains 315 810 563 target–disease associations classified into the curated and scientific literature, between 20 257 targets and 17 043 diseases (Figure 2A). As far as we know, this is the largest disease–target relationship database. The target–disease associations are mainly extracted based on the exploitation of semantic information from full literature. In the current release, 2 220 721 full publications were obtained from NCBI PubMed Central. Moreover, we also integrated the gene/protein and disease pairs from DisGeNET [42] database, which is a collection of genes–human associations from expertly curated repositories and text mining using Befree system. Based on these collections, we calculated the significance and novelty score to assist in the prioritization of target–disease relationships.

As shown in Figure 2B, the targets were mainly obtained from TCRD [35], UniProt [36] and ChEMBL [37] and were classified into 10 categories: enzymes, epigenetics, GPCRs, oGPCRs, ion channels, kinases, nuclear receptors, transcription factors transporters and non-IDG. The target development level (TDL) classification can help better characterize the role and selection priority of new targets in disease, and it contains four categories [43]: those already targeted by approved drugs are classified as Tclin; those have reported small molecular modulators with suitable activities that are not approved yet are classified as Tchem; those are identified to be associated with OMIM disease or annotated by the Gene Ontology (GO) but still have no known drug or small molecular modulators are classified as Tbio; and the remaining

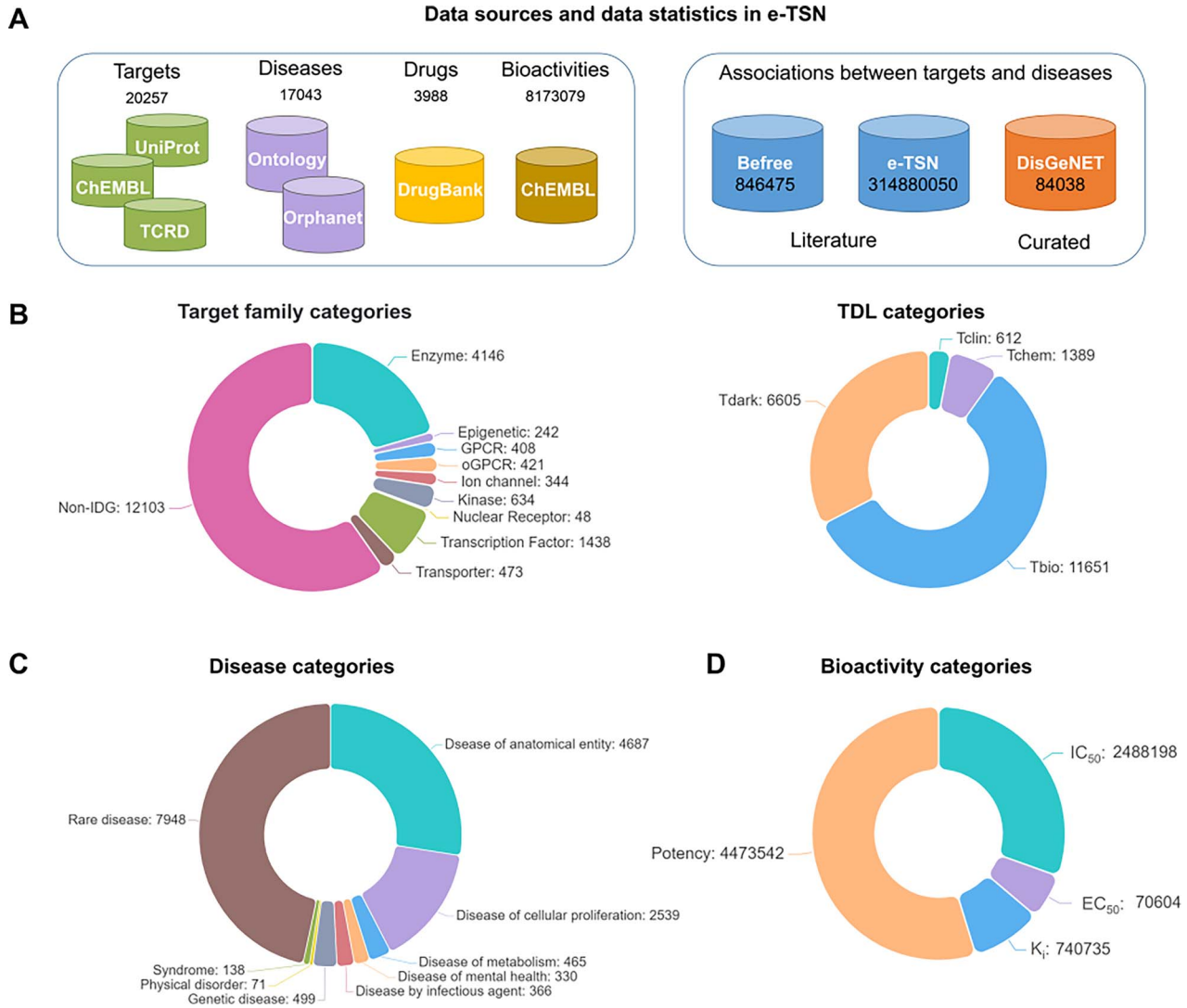

Figure 2. Statistical analysis of the entries in e-TSN. **(A)** A summary of the e-TSN entities and quantities. The data sources of target–disease associations are classified as Literature and Curated. **(B)** Distribution of the entries across ten target families. **(C)** Distribution of the entries across nine disease categories. **(D)** Distribution of the entries across four bioactivity categories.

Figure 3. The e-TSN interface of targets associated with SARS-CoV-2 infection. The figure shows how the disease–target associations are presented in the scatter plot, exemplified by the 'SARS-CoV-2 infection' disease.

proteins are classified as Tdark, for which nothing is known. The diseases in e-TSN were mapped to the Disease Ontology database [38] and Orphanet, including nine categories (Figure 2C): infectious disease, disease of anatomical entity, disease of cellular proliferation, disease of metabolism, disease of mental health, genetic disease, physical disorder, syndrome and rare disease.

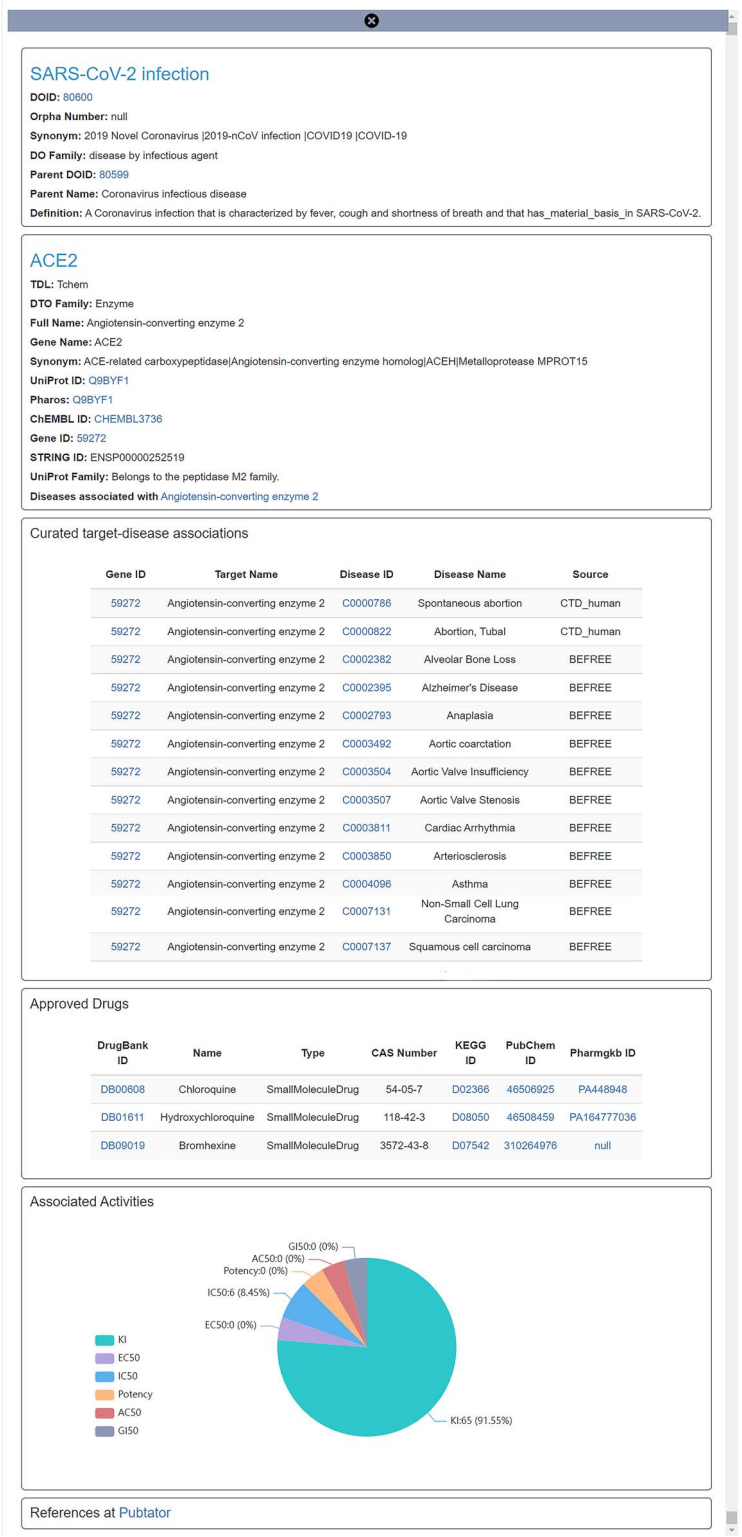

Figure 4. e-TSN provides comprehensive information for related targets and diseases. Click on the selected point shows more details including approved drugs and associated activities, exemplified by the 'angiotensin-converting enzyme 2' target.

Meanwhile, we also integrated approval drugs data from DrugBank [44] and bioactivities data from the ChEMBL database [37] for each target to give a more comprehensive understanding of the protein, covering 3988 approved drugs and 8 173 079 bioactivities (Figure 2D). In addition, we will also carry out the maintenance and update of the database, which is expected to be updated regularly once a year.

The web application

e-TSN' s web application provides quick access to data through a direct search field on the home page. Searching by disease name or target name (with auto-suggest) is supported. Users can obtain all related targets' significance and novelty distribution by retrieving keywords of the interested disease. The novel coronavirus SARS-CoV-2 outbreak at the tail end of 2019 has swept the globe,

dramatically changing every aspect of our lives [45]. There was an urgent need to find relevant targets and effective drugs, which could provide practical examples of how e-TSN can help search the new targets from literature. When inputting in the search bar for the keyword 'SARS-CoV-2 infection,' targets associated with the selected disease are plotted in a scatter plot with log–log significance–novelty axes (Figure 3).

In the scatter layout, the targets on the right side of the plot may have a stronger correlation with the given disease, while the targets in the upper part of the plot may not have been widely studied. In general, the more interesting associations will appear in the upper-right corner of the plot. They could be given top priority when choosing candidates, which can accelerate target discovery by assisting in identifying and prioritizing under-researched genes or proteins as potentially novel drug targets for further research. Targets can be filtered based on protein family (e.g. GPCRs, kinases) and target development level (TDL) such as Tclin, Tchem, Tbio and Tdark.

In addition, users can click on the target name of any records on the scatter plot panel to search for detailed annotation information, including approved drugs and associated bioactivities to facilitate adequate consideration of multiple targets for drug development. Hyperlinks point out to Pharos [35], Uniprot [36], Gene, ChEMBL [37] and Disease Ontology [38]. Figure 4 illustrates the result of mouse click actions for the target 'Angiotensin-converting enzyme 2' from Figure 3. It has been demonstrated that SARS-COV-2 enters the human body through the Angiotensin-converting enzyme 2(ACE2) receptor, and ACE2 is an effective target for novel 2019-nCoV infection treatment [46, 47].

Similarly, users can also explore the significance and novelty distribution of the diseases related to the specific target, with the aim of searching and demonstrating alternative indications for target genes of interest. Figure 5 shows diseases that are associated with 'Angiotensin-converting enzyme 2,' the preferred candidates appear in the upper right corner of the scatter diagram, which can assist researchers in quickly grasping which diseases are associated with the target of interest, and the disease of 'severe acute respiratory syndrome' can be found in the scatter plot. As Figure 6 shows, detailed annotation information about Angiotensin-converting enzyme 2 and severe acute respiratory syndrome is also provided. It can help users promote a better understanding of the complex crosstalk within protein-disease and conduct research on more novel disease fields to find new indications for drug repurposing.

As noted above, e-TSN implements rapid retrieval and visualization throughout the interface. An important visualization of the web data is by using the scatter diagram. The interface allows one to rank targets using two parameters, including the significance and novelty scores based on biometrics in the scatter

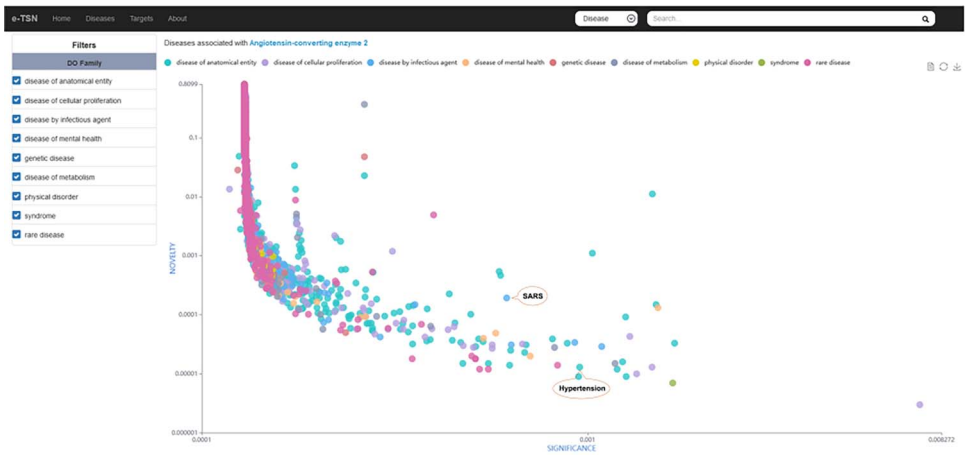

Figure 5. The e-TSN interface of diseases associated with angiotensin-converting enzyme 2. The figure shows how the disease–target associations are presented in the scatter plot, exemplified by the 'angiotensin-converting enzyme 2' target.

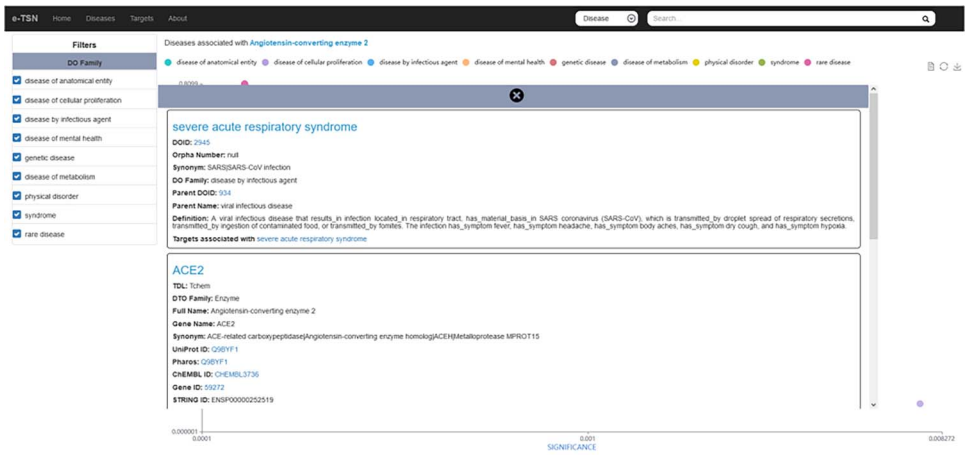

Figure 6. The e-TSN interface of relationship between angiotensin-converting enzyme 2 and severe acute respiratory syndrome. Click on the selected point shows more details of the target and disease, exemplified by the 'severe acute respiratory syndrome' disease.

plot to derive novel biological insight easily. The visualization is interactive, allowing users to zoom, filter and select. Users can filter targets by protein families and TDL in order to quickly identify interested targets that lack specific family in clinical practice. The significance and novelty distribution scatter plot of specific searches carried out through the web interface is available for download in PNG format. This work is being used to guide the discovery and prioritize the thousands of genes/proteins to contribute positively to the early drug discovery process.

Conclusion and future work

Currently, the volume of biomedical texts is enormous, and its rapid growth makes it impossible for researchers to process the information manually. In this paper, a novel method of knowledge synthesis and discovery based on literature is proposed to build a knowledge graph to capture and represent multiple causal relationships between diseases and targets. Using automatic text mining, big data and visualization method, we developed the target significance and novelty scoring schemes based on bibliometrics from full publications and constructed the visualization platform to promote the efficiency of data-driven target discovery. The database can provide comprehensive data resources covering more than 300 million potential relationships between >17 000 diseases and >20 000 genes/proteins. The overarching goal of e-TSN is to provide an interactive visualization platform for fast exploring the target–disease associations to understand the relationship between diseases and targets and facilitate researchers in discovering potential novel drug targets. Data-driven and graded papers using text-mining techniques can lead researchers to knowledge they might not otherwise have been aware of. e-TSN may suggest new associations between the target protein and disease, facilitating computational target discovery and leading to novel perspectives and new applications to the old drugs. The novel potential proteins are more interesting from a target recognition point of view at the same time, require more validation.

Like all current text-mining tools, e-TSN cannot wholly replace manually checking papers and expert human curators. However, with the rapid growth of biomedical literature, the text-mining tools like e-TSN will play an increasingly important role in future biomedical research. e-TSN will be further developed by constantly optimizing the text-mining pipeline, such as using modern deep learning-based architectures to improve the precision of the identified target–disease pairs. Moreover, although text mining has great potential in biomedical applications, it still needs further development. The future development direction of this project includes improving the text-mining algorithm to obtain high semantic correlation, while improving the knowledge graphs framework for meaningful associations between diseases, targets and drugs. To achieve the goal, we plan to extend the data sources of diseases and targets, mining potential associations between them on more full texts and abstracts, and integrate other different data mining approaches to overcome the drawbacks of a single text-mining method. Biomedical text mining, along with knowledge graph and visualization methods, should produce consistent, measurable and testable results so that new drug development can achieve greater results and greater progress in an era of rapid information growth.

Authors' Contributions

Z.F. collected and curated the database and participated in developing the web server and writing the manuscript. Z.S. contributed to the development of the web server. H.L. took part in the discussion of the data. S.L. conceived the study, coordinated the work and contributed to writing the manuscript. All authors are involved in the discussion and finalization of the manuscript.

Key Points

- A new versatile web server for visualizing target–disease knowledge mapping based on text mining from full biomedical literature.
- Target significance and novelty scoring methods were constructed using bibliometric to facilitate target discovery and prioritization.
- The database provides big data resources, covering over 300 million potential disease–target relationships between >17 000 diseases and >20 000 genes/proteins.
- The web server is uniquely designed to improve the efficiency of data-driven target identification and drug repurposing.

Data availability

The data are available online at http://www.lilab-ecust.cn/etsn.

Funding

This work was supported in part by the National Natural Science Foundation of China (grants 82173690 to S.L., 81825020 and 82150208 to H.L.,); the Lingang Laboratory (grant LG-QS-202206-02 to S.L.); the Fundamental Research Funds for the Central Universities; Honglin Li was also sponsored by the National Program for Special Supports of Eminent Professionals and National Program for Support of Top-Notch Young Professionals.

References

1. Chan HCS, Shan H, Dahoun T, *et al.* Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci* 2019;**40**:592–604.
2. Paul SM, Mytelka DS, Dunwiddie CT, *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010;**9**:203–14.
3. Rodrigues T, Bernardes GJL. Machine learning for target discovery in drug development. *Curr Opin Chem Biol* 2020;**56**:16–22.
4. Sams-Dodd F. Target-based drug discovery: is something wrong? *Drug Discov Today* 2005;**10**:139–47.
5. Butcher SP. Target discovery and validation in the post-genomic era. *Neurochem Res* 2003;**28**:367–71.
6. Chen Y-PP, Chen F. Identifying targets for drug discovery using bioinformatics. *Expert Opin Ther. Tar* 2008;**12**:383–9.
7. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. *Drug Discov Today* 2012;**17**:S16–23.
8. Ravikumar KE, Wagholikar KB, Li D, *et al.* Text mining facilitates database curation - extraction of mutation-disease associations from bio-medical literature. *BMC Bioinform* 2015;**16**:185.
9. Agarwala R, Barrett T, Beck J, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2018;**46**:D8–13.
10. Comeau DC, Wei CH, Dogan RI, *et al.* PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics* 2019;**35**:3533–5.
11. Ananiadou S, Kell DB, Tsujii J-I. Text mining and its potential applications in systems biology. *Trends Biotechnol* 2006;**24**:571–9.

12. Greene CS, Troyanskaya OG. Integrative systems biology for data-driven knowledge discovery. *Semin Nephrol* 2010;**30**:443–54.
13. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**:57–71.
14. McCoy K, Gudapati S, He L, *et al.* Biomedical text link prediction for drug discovery: a case study with COVID-19. *Pharmaceutics* 2021;**13**:794.
15. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015;**74**:97–106.
16. Rahaman T. Discovering new trends & connections: current applications of biomedical text mining. *Med Ref Serv Q* 2021;**40**: 329–36.
17. Xiao W, Jing L, Xu Y, *et al.* Different data mining approaches based medical text data. *J Healthc Eng* 2021;**2021**:1–11.
18. Hansson LK, Hansen RB, Pletscher-Frankild S, *et al.* Semantic text mining in early drug discovery for type 2 diabetes. *PLoS One* 2020;**15**:e0233956.
19. Conceicao SIR, Couto FM. Text mining for building biomedical networks using cancer as a case study. *Biomolecules* 2021;**11**:1340.
20. Bao WZ, Cui QY, Chen BT, *et al.* Phage_UniR_LGBM: phage virion proteins classification with UniRep features and lightGBM model. *Comput Math Methods Med* 2022;**2022**:1–8.
21. Bao WZ, Yang B, Chen BT. 2-hydr_Ensemble: lysine 2-hydroxyisobutyrylation identification with ensemble method. *Chemom Intel Lab Syst* 2021;**215**:104351.
22. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* 2013;**41**:W518–22.
23. Cheng D, Knox C, Young N, *et al.* PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 2008;**36**:W399–405.
24. Capuzzi SJ, Thornton TE, Liu K, *et al.* Chemotext: a publicly available web server for mining drug-target-disease relationships in PubMed. *J Chem Inf Model* 2018;**58**:212–8.
25. Pletscher-Frankild S, Palleja A, Tsafou K, *et al.* DISEASES: text mining and data integration of disease-gene associations. *Methods* 2015;**74**:83–9.
26. Li L, Hu X, Yang Z, *et al.* Establishing reliable miRNA-cancer association network based on text-mining method. *Comput Math Methods Med* 2014;**2014**:1–8.
27. Turewicz M, Frericks-Zipper A, Stepath M, *et al.* BIONDA: A free database for a fast information on published biomarkers. *Bioinform adv*. 2021;**1**:vbab015.
28. Westergaard D, Staerfeldt H-H, Tonsberg C, *et al.* A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput Biol* 2018;**14**:e1005962.
29. Leung CK. Data science for big data applications and services: data lake management, data analytics and visualization. In: Lee W, Leung CK, Nasridinov A (eds). *Big Data Analyses, Services, and Smart Data*, Vol. 899. Singapore: Springer Singapore, 2021, 28–44.
30. MacLean F. Knowledge graphs and their applications in drug discovery. *Expert Opin Drug Discovery* 2021;**16**:1057–69.
31. Gurbuz O, Alanis-Lobato G, Picart-Armada S, *et al.* Knowledge graphs for indication expansion: an explainable target-disease prediction method. *Front Genet* 2022;**13**:814093.
32. Zeng XX, Tu XQ, Liu YS, *et al.* Toward better drug discovery with knowledge graph. *Curr Opin Struct Biol* 2022;**72**:114–26.
33. Yang Z, Lin H, Li Y. Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature. *Comput Biol Chem* 2008;**32**:287–91.
34. Auger A, Barriere C. Pattern-based approaches to semantic relation extraction: a state-of-the-art. *Terminology* 2008;**14**:1–19.
35. Dac-Trung N, Mathias S, Bologa C, *et al.* Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res* 2017;**45**:D995–1002.
36. Bateman A, Martin M-J, Orchard S, *et al.* UniProt: a world-wide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**: D506–15.
37. Mendez D, Gaulton A, Bento AP, *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;**47**:D930–40.
38. Schriml LM, Arze C, Nadendla S, *et al.* Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2012;**40**:D940–6.
39. Cook HV, Jensen LJ. A guide to dictionary-based text mining. *Methods Mol Biol (Clifton, NJ)* 2019;**1939**:73–89.
40. Leroy G, Chen HC. Genescene: an ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *J Am Soc Inf Sci Technol* 2005;**56**:457–68.
41. Cohen KB, Hunter L. Getting started in text mining. *PLoS Comput Biol* 2008;**4**:e20.
42. Pinero J, Manuel Ramirez-Anguita J, Sauch-Pitarch J, *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 2020;**48**:D845–55.
43. Oprea TI, Bologa CG, Brunak S, *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov* 2018;**17**:317–32.
44. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
45. Wang LL, Lo K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Brief Bioinform* 2021;**22**:781–99.
46. Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;**583**:459–68.
47. Kruse RL. Therapeutic strategies in an outbreak scenario to treat the novel coronavirus originating in Wuhan, China. *F1000Research* 2020;**9**:72.