www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Europe PMC annotated full-text corpus for gene/proteins, diseases and organisms

Xiao Yang[1,3], Shyamasree Saha[1,2,3], Aravind Venkatesan[1], Santosh Tirunagari[1,2] ✉, Vid Vartak[1] & Johanna McEntyre[1]

Named entity recognition (NER) is a widely used text-mining and natural language processing (NLP) subtask. In recent years, deep learning methods have superseded traditional dictionary- and rule-based NER approaches. A high-quality dataset is essential to fully leverage recent deep learning advancements. While several gold-standard corpora for biomedical entities in abstracts exist, only a few are based on full-text research articles. The Europe PMC literature database routinely annotates Gene/Proteins, Diseases, and Organisms entities. To transition this pipeline from a dictionary-based to a machine learning-based approach, we have developed a human-annotated full-text corpus for these entities, comprising 300 full-text open-access research articles. Over 72,000 mentions of biomedical concepts have been identified within approximately 114,000 sentences. This article describes the corpus and details how to access and reuse this open community resource.

Background & Summary

Europe PubMed Central (Europe PMC)[1] is a repository of life science research articles, which includes peer-reviewed full-text research articles, abstracts, and preprints–all freely available for use via the website (https://europepmc.org). Europe PMC houses over 33.3 million abstracts and 8.7 million full-text articles. Since 2020, it has added over 1.7 million new articles annually. The rapid growth in the number of publications within the biological research space makes it challenging and time-consuming to track research trends and assimilate knowledge. Thanks to the digitization of large portions of biological literature and advancements in natural language processing (NLP) and machine learning (ML), it is now possible to build sophisticated tools and the necessary infrastructure to process research articles. This allows for the extraction of biological entities, concepts, and relationships in a scalable manner.

Harnessing the NLP techniques, tools such as LitSuggest[2] and PubTator[3] are being used in biomedical literature curation[4,5], recommending relevant biomedical literature, or automatically annotating biomedical concepts[6], such as genes and mutations, in PubMed abstracts and PubMed Central (PMC) full-text articles. Furthermore, in a step towards FAIRification[7–9] and sharing text-mined outputs across the scientific community, Europe PMC has established a community platform to capitalise on the advances made. Annotations from various text-mining groups are consolidated and made available via open APIs and a web application called SciLite[10], which highlights the annotations on the Europe PMC's website. Several other biological resources including STRING[11] and neXtProt[12] have embedded NLP processes in their data workflows to serve their user community better. Developement of such NLP tools require the availability of open data (full-text corpora). Thanks to the biomedical text mining community, which has endorsed open data, resources such as PubMed, PubMed Central and Europe PMC provide open access abstracts and full-text for researchers to download. The COVID-19 Open Research Dataset Challenge (CORD-19 dataset)[13] is a recent example of using text-mining to tackle specific scientific questions. This dataset consists of full-text scientific articles about COVID-19 and related coronaviruses. Additionally, BioC[14] provides a subset of those full-text articles in a simple BioC format, which can reduce the efforts of text processing. Biomedical datasets, such as those from BioASQ[15] and BioNLP[16] shared tasks, enable the development and testing of novel ideas, including deep learning methodologies. With the development of such biomedical datasets, great improvements in biomedical text mining systems have been

[1]Literature Services, EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, UK. [2]Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [3]These authors contributed equally: Xiao Yang, Shyamasree Saha. ✉e-mail: stirunag@ebi.ac.uk

made. From the results of recent BioASQ challenges (2013 to 2019), the performance of cutting-edge systems keep advancing for tasks such as large-scale semantic indexing and question answering (QA)[17]. While corpora without annotations are good for learning semantics, text-mining tools trained on human-annotated corpora outperform those trained on non-annotated ones. Therefore, open-source gold-standard datasets are crucial for improving biomedical text mining systems. In particular, transformer-based deep learning models, such as BERT[18] and GPT[19], have show that pre-training language models with large text corpora improves performance on downstream applications. However, compared to the text corpora, gold-standard biomedical datasets with human annotations are expensive to obtain, because they require domain experts to spend significant amounts of time creating accurate annotations. Therefore, generating human-annotated biomedical datasets is valuable for biomedical text mining, because once they are available, machine learning algorithms have an accurate starting point to learn from.

There have been multiple projects that have produced gold standard corpora, such as BioCreative V CDR corpus (BC5CDR)[20], BC2GM[21], Bioinfer[22], S800[23], GAD[24], EUADR[25], miRNA-test corpus[26], NCBI-disease corpus[27], and BioASQ[15]. In addition to these, other efforts have generated gold standard corpora from full-text articles, such as Linnaeus[28], AnatEM[29], and the Colorado Richly Annotated Full-Text Corpus (CRAFT)[30]. The Europe PMC Annotations (EPMCAs) corpus is also a full-text-based corpus, similar to CRAFT, Linnaeus, and AnatEM.

Specifically, the CRAFT Corpus is a human-annotated biomedical dataset that is widely used by researchers to develop and evaluate novel text mining algorithms. It comprises 97 full-text, open-access biomedical journal articles that include both semantic and syntactic annotations, as well as coreference annotations and 10 biomedical concepts. This establishes it as an important gold-standard dataset in the biomedical domain.

Recent publications[31,32] have demonstrated that sophisticated systems can be developed using annotated biomedical datasets. Notably, as pre-trained models like BERT[18] have gained traction in the biomedical field, many systems have been created by training models on multiple biomedical datasets. For example, the BioBert model[33] has been trained and evaluated on multiple datasets for downstream tasks such as Named Entity Recognition, Relation Extraction, and Question Answering.

This study presents the Europe PMC Annotated Full-text Corpus (EPMCA), a collection of 300 research articles from the Europe PMC Open Access subset. The selected articles have been human annotated to indicate mentions of three biomedical concepts; Gene/Protein, Disease, and Organism. Since all annotations are created based on guidelines, this helped the human annotators select the correct text span and type of annotation. Three additional articles that were used in a pilot study are also published with this study. The size of the EPMCA (in terms of the number of full-text articles annotated)is among the largest human-annotated biomedical corpora. We believe that the high-quality gold-standard annotations of the EPMCA corpus will be an important addition to other existing datasets and provide significant benefits for biomedical text mining..

Methods

The overall strategy for the Full-text annotation workflow is presented in Fig. 1. Out of a million Open Access (OA) full-text articles archived on the 31st of August 2018 in Europe PMC, a subset of 300 articles was selected as the gold standard for curation. This section presents the methods we employed to stratify those articles and select the representative gold-standard set, followed by the annotation guidelines and article annotation.

The open access article set in europe pmc and cc-by-licenced articles. Because a primary outcome of this work was to create a training set for anyone to use, the first constraint applied was to use Open Access articles that have a parsable/machine-readable (available in the JATS XML standard, information on which can be obtained at https://jats.nlm.nih.gov under *CC-BY* licence. We used the archived open access set from 31st August 2018 (v.2018.09) [Available at http://europepmc.org/ftp/archive] as a basis, which consists of 2,113,557 articles, of which 991,529 articles had a parsable *CC-BY* licence.

Body size. Using the 991,529 *CC-BY* articles as a starting point, we measured the size of the full-text article $<BODY>$ section and grouped them into bins of 10 *KB* size to find the most representative articles. More than 50% of articles were in the range of 25–5 *KB* (Fig. 2) that were, rich in entities. Using this size range further constrained the pool to 503,950 articles. Constraining the article size range also meant that the annotators would be provided with a more consistent article set as presumably articles falling outside this range are likely to not be research articles.

Entity frequency distribution. The pool of 503,950 "standard-sized" articles were further stratified based on the term frequency of the three entities of interest, namely; Gene/Proteins, Diseases, and Organisms. Using the current Europe PMC dictionary-based annotation pipeline to annotate the articles, we established the range of entity frequencies in the articles (Fig. 3) and created high (H), medium (M), and low (L) frequency tertiles by splitting them at the 33 and the 66 percentiles (Table 1). This resulted in 27 bins of articles from these tertiles of three entities (33) (Fig. 4). All the articles in the Low-Low-Low bin contain a small number or no mentions of any of the entities but represent the largest number of articles (42,261 articles, more than 8% of total articles). Because these would add little value to the training dataset, this bin was excluded from the article selection process. There were 46,1689 articles in the remaining 26 bins. We then randomly selected 300 articles in total across all 26 bins in proportion to the number of articles in each bin (2–20 articles from a bin in real terms, Fig. 5). For example, only two articles were selected from the Low Disease, High Gene/Protein, Low Organism bin.

Ontology/terminology selection. The Europe PMC annotation pipeline currently uses a dictionary-based approach to tag Gene/Proteins, Diseases, and Organisms[1]. The term dictionaries are created from UniProt[5],

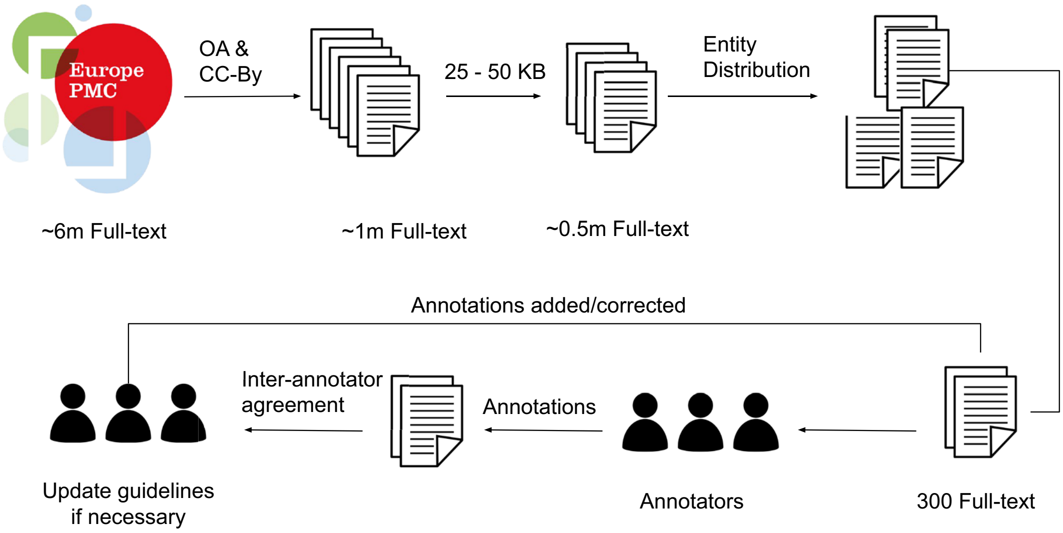


Fig. 1 The illustration of the full-text annotation workflow. There were approximately six million full-text articles in the Europe PMC repository archived on the 31st of August, 2018 (v2018.09) of which approximately one million were Open Access (OA) with a CC-BY licence. Thereafter, to have articles specific to research, size between 25 and 50 KB were selected, which resulted in a collection of approximately 0.5 million articles. This was followed by sorting the articles with the entity mentions into low, medium, and high bins for each entity type, i.e. Gene/Protein, Disease, and Organisms. Finally, 300 articles were selected that represented the aforementioned entity types for each article. The workflow included working with the annotators iteratively to improve the annotation guidelines.

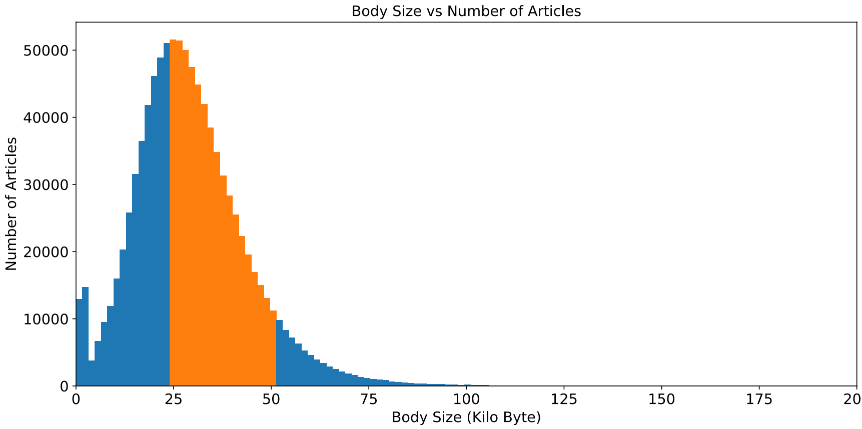


Fig. 2 Distribution of body sizes of full-text articles with a CC-BY licence on the 31st August 2018 (v.2018.09) frozen set.

UMLS[34], and the NCBI taxonomy[35] for the Gene/Proteins, Diseases and Organisms, respectively. The pipeline annotates articles using predefined patterns and regular expressions to accommodate term variations from the dictionaries.

Gene/Protein. The Gene/Proteins dictionary is periodically generated from the SwissProt[36] knowledgebase from the 2014 release. SwissProt is a manually reviewed resource of proteins and genes, and the knowledgebase is released in multiple formats. The entries in the Uniprot knowledgebase are structured to make it both human and machine-readable (for more details please follow https://www.uniprot.org/docs/userman.htm#convent). For tagging Gene/Proteins in the Europe PMC annotations workflow, the DAT file of the knowledgebase release is parsed, generating a Gene/Proteins dictionary from the gene name lines and their aliases (the gene name lines are denoted by starting the line with GN tag according to the knowledgebase data structure). The UniProt knowledgebase release, dated 2014, was used to generate the Gene/Proteins dictionary. In addition, a list of common English words (we call it a common-stop list) is used to avoid predominantly false-positive identifications, for example, 'CAN' as a gene name.

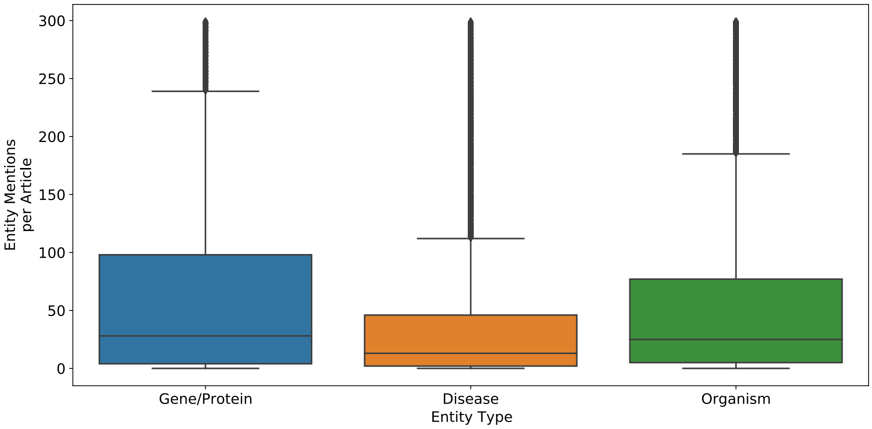


Fig. 3 Distribution of entity mentions (Gene, Disease and Organism) per full-text article from the candidate pool. For the convenience of the display, we have used a threshold of a maximum of 300 mentions per article per entity type for this figure, although the maximum was 2408 for Gene/Protein, 678 for Disease, and 3108 for Organism. This figure shows that, on average, Disease mentions are almost half of Gene/Protein mentions per article. This distribution helped us to set entity count boundaries for the article stratification required to select the final corpus. The horizontal lines within the coloured boxes typically represent the median of the data, also known as Q2 or the 50th percentile. The heights of the boxes indicate the Interquartile Range (IQR), which is the difference between the third quartile (Q3, or the 75th percentile) and the first quartile (Q1, or the 25th percentile). The horizontal lines outside of the boxes are "whiskers," which indicate the range of the data. Specifically, the lower whisker usually extends to the smallest data value within 1.5 * IQR from Q1, and the upper whisker extends to the largest data value within 1.5 * IQR from Q3. The values outside the whiskers are those individual data points that fall outside of the range defined by 1.5 * Interquartile Range (IQR) above the third quartile (Q3) or below the first quartile (Q1). These are outliers, that are significantly different from the majority of the data.

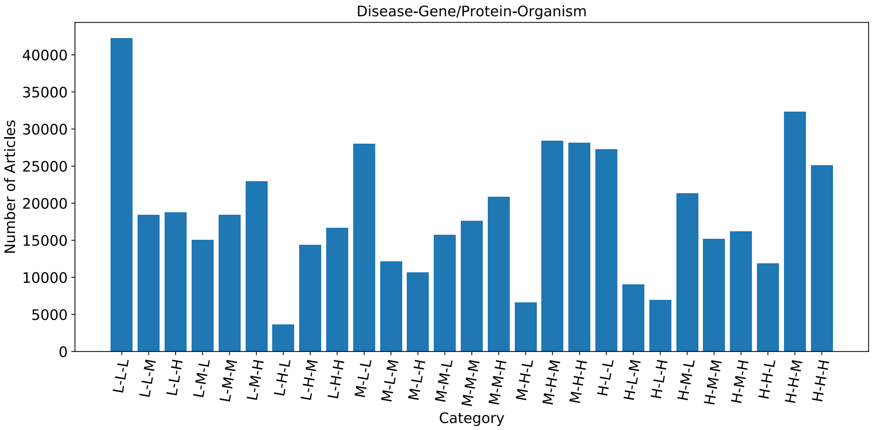


Fig. 4 Distribution of articles based on the entity frequency. Here L, M, and H represent low frequency, medium frequency, and high-frequency tertile. The order of the label is Disease, Gene/Protein and Organism. For example, H-L-H represents articles that are high frequency for Disease and Organism and low frequency for Gene/Protein.

	Low frequency of occurrence count (L)		Medium frequency of occurrence count (M)		High frequency of occurrence count (H)	
Entity	**Lower**	**Upper**	**Lower**	**Upper**	**Lower**	**Upper**
Genes/Proteins	0	11	12	80	81	2408
Organisms	0	9	10	57	58	3108
Diseases	0	4	5	32	33	678

Table 1. The abundance of key entities is used to establish tertile boundaries.

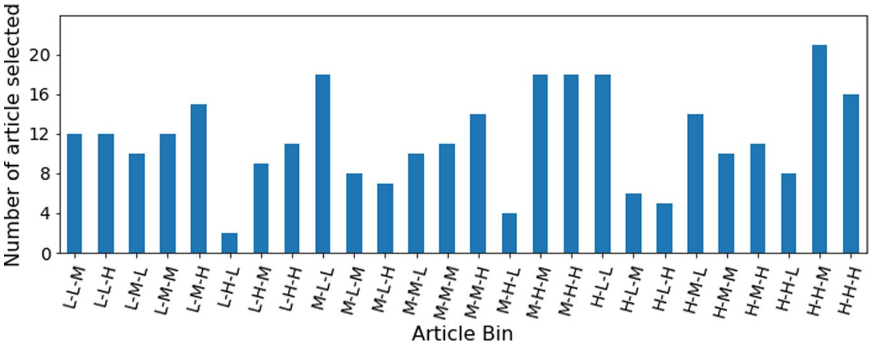


Fig. 5 Number of articles selected from each bin for inclusion in the gold-standard corpus of 300 articles. L, M, and H represent low frequency, medium frequency, and high-frequency tertile.

	Annotator1	Annotator2	Annotator3	Total	PMC count
Batch1	1583	1587	1587	4757	30
Batch2	4745	4727	4733	14205	70
Batch3	5604	5610	5611	16825	80
Batch4	11932	11924	11931	36787	180

Table 2. Batch-wise annotation breakdown of articles and annotations.

Disease. UMLS Diseases terms are used to create the Diseases dictionary. In UMLS, there are twelve different diseases/disorders (DISO) groups; four generate the Diseases dictionary because the other groups mainly comprise phenotypes and symptoms. The four DISO groups used are Disease or Syndrome (T047), Mental or Behavioural Dysfunction (T048), Neoplastic Process (T191), and Pathologic Function (T046). The ULMS version, dated 2015, was used to generate the Diseases dictionary.

Organism. The Organisms dictionary is based on the NCBI Taxonomy. Specific fields, such as acronym, BLAST name, GenBank common name and GenBank synonym, are used to populate the dictionary. The NCBI taxonomy version dated 2015 was used to generate the Organisms dictionary.

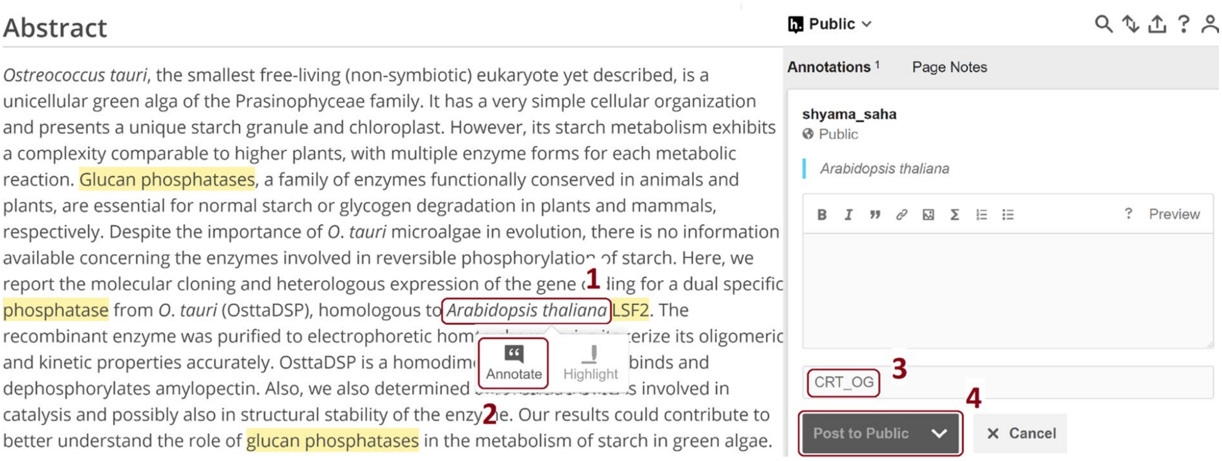
Creation of annotation guidelines. A detailed concept annotation guideline is essential for developing a good corpus and resolving annotation disputes (Supplementary information file: Europe PMC Annotation Guidelines). The CRAFT corpus provides comprehensive annotation guidelines[37], explaining both the text spans to be annotated and the assignment of entity types. We based our annotation guidelines on those of the CRAFT corpus and expanded them to meet our specific requirements. A list of examples was included in the guidelines to assist curators. Before the commencement of the annotation work, a pilot study was conducted, focusing on the annotation of three articles. The outcomes of the pilot study were fourfold:

1. The pilot study helped curators estimate the workload, thereby setting project timelines;
2. Initial feedback was used to improve the annotation guidelines;
3. The curators familiarized themselves with both the task and the annotation tools;
4. The pilot study established the communication channels required to manage the project.

Article annotation. We worked with Molecular Connections (https://molecularconnections.com), India, to employ three PhD-level domain experts to annotate the corpus. We used a triple-anonymous approach to annotation; three annotators annotated the same articles independently to ensure annotation quality and validate inter-annotation agreement. Annotation discrepancies were resolved by the majority vote to achieve/ensure the best quality annotation. That is, at least two annotators must agree on the annotation boundary and the entity type of the entity terms to pass the acceptance threshold. This maximised the total number of annotations. For example, if one annotator misses a term, it will likely be picked by the two other annotators. The triple-anonymous method made it possible to conveniently assess the inter-annotator agreements to ensure the annotation quality.

We sent the articles to the annotators in four batches. Between each batch, annotation quality and inter-annotator agreement were evaluated, and any confusion or quality issues were addressed. If necessary, updates to the annotation guidelines were made after each batch. To assess the quality of the annotations, the first batch consisted of only 30 articles, after which the number of articles per batch increased. This approach allowed us to resolve annotation discrepancies along the way and refine the annotator guidelines. Table 2 shows a detailed breakdown of these batches.

Annotators were instructed to view the articles on the Europe PMC website, where the existing dictionary-based annotations from Europe PMC text-mining pipeline are displayed using Scilite. The Hypothes.is



Abstract

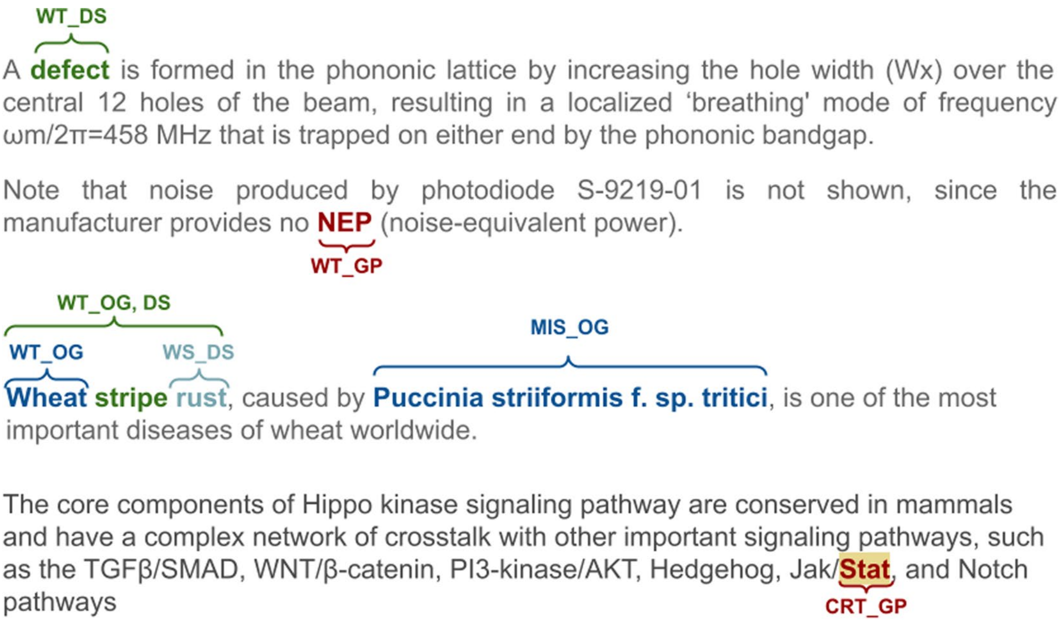
Ostreococcus tauri, the smallest free-living (non-symbiotic) eukaryote yet described, is a unicellular green alga of the Prasinophyceae family. It has a very simple cellular organization and presents a unique starch granule and chloroplast. However, its starch metabolism exhibits a complexity comparable to higher plants, with multiple enzyme forms for each metabolic reaction. Glucan phosphatases, a family of enzymes functionally conserved in animals and plants, are essential for normal starch or glycogen degradation in plants and mammals, respectively. Despite the importance of *O. tauri* microalgae in evolution, there is no information available concerning the enzymes involved in reversible phosphorylation of starch. Here, we report the molecular cloning and heterologous expression of the gene coding for a dual specific phosphatase from *O. tauri* (OsttaDSP), homologous to *Arabidopsis thaliana* LSF2. The recombinant enzyme was purified to electrophoretic hom erize its oligomeric and kinetic properties accurately. OsttaDSP is a homodim binds and dephosphorylates amylopectin. Also, we also determined s involved in catalysis and possibly also in structural stability of the enzyme. Our results could contribute to better understand the role of glucan phosphatases in the metabolism of starch in green algae.

shyama_saha
Public

Arabidopsis thaliana

CRT_OG

Fig. 6 A screenshot of the Hypothes.is annotation platform overlayed on top of the Europe PMC website. Highlighted in yellow are existing dictionary-based text-mined terms. After selecting a term (1), users need to click the 'Annotate' button (2) to annotate the term. It will pop up the Hypothes.is annotation window on the right-hand side, allowing the annotators to add the annotation (3) and then save it using the 'Post to Public' button (4). Please refer to the supplementary information (Section 'How to use the interface' under "demo to molecular connections") and Hypothes.is website for a detailed user manual.



A **defect** is formed in the phononic lattice by increasing the hole width (Wx) over the central 12 holes of the beam, resulting in a localized 'breathing' mode of frequency ωm/2π=458 MHz that is trapped on either end by the phononic bandgap.

Note that noise produced by photodiode S-9219-01 is not shown, since the manufacturer provides no **NEP** (noise-equivalent power).

Wheat stripe rust, caused by **Puccinia striiformis f. sp. tritici**, is one of the most important diseases of wheat worldwide.

The core components of Hippo kinase signaling pathway are conserved in mammals and have a complex network of crosstalk with other important signaling pathways, such as the TGFβ/SMAD, WNT/β-catenin, PI3-kinase/AKT, Hedgehog, Jak/**Stat**, and Notch pathways

Fig. 7 Example of human annotation correcting dictionary-based Europe PMC annotation using the tag set defined for this annotation task. Disease takes higher priority over organism type, while gene/protein tags take precedence over disease tags. In this figure, WT_OG is incorrectly labeled as the organism type for the entity "wheat." Additionally, "rus" is inaccurately spanned for the disease tag (WS_DS). Therefore, the annotators have labeled "Wheat stripe rust" as 'WT_OG, DS' to indicate that the correct tag should be DS, not OG. In another scenario, "*Puccinia striiformis f. sp. tritici*" is identified as MIS_OG indicating a missing organism tag from the the Europe PMC's pre-annotations system.

annotation tool works as a layer on top of the Europe PMC website, allowing the curators/annotators to visualise and curate existing annotations and newly identified entity terms (Fig. 6). We used Hypothes.is platform for annotations over other platforms such as BRAT[38] and GATE[39] as they require pre-processing of articles, for example, converting them to text files. Moreover, Hypothes.is provided easy access to Europe PMC website. We developed a set of standard schemes of tags for the curators to use and therefore classify the existing SciLite annotations.

The standard terms/tags were used as follows (Fig. 7 shows an example of the use of these tags):

1. Correctness of annotation. Allows the annotators to verify existing Europe PMC annotations as Wrong Type (WT), Wrong Span (WS), Missing (MIS), or Correct (CRT).

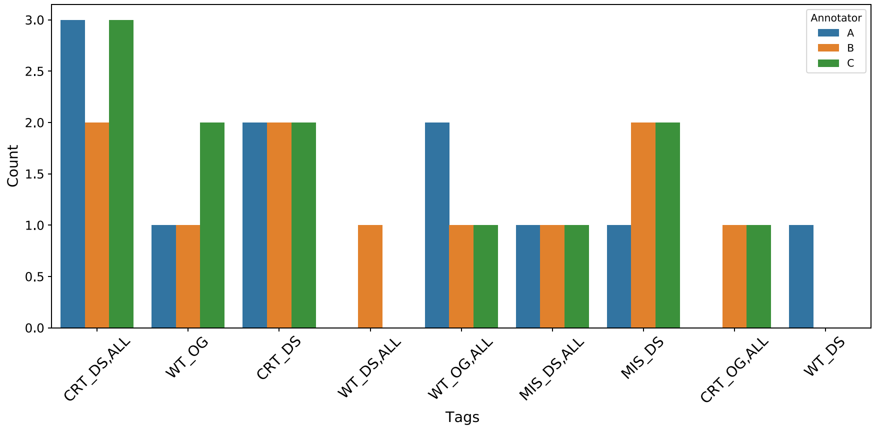


Fig. 8 An example of the tag distributions from batch 1 showing the discrepancies between the annotators. Annotators used the 'ALL' tag to mark all mentions of the entity as correct (CRT) or wrong type (WT), missing (MIS), and so on. The DS and OG represent the Diseases and the Organisms entities respectively.

2. Entity type. Three symbols were used to represent the entity types, GP for Gene/Proteins, DS for Diseases, and OG for Organisms.
3. A special tag 'ALL' allowed the annotators to apply the annotation of the current term to all occurrences of it across the article. This was useful in the case of reducing workload for the annotators and annotation cost but required additional work to find all the occurrences of a concept with an "ALL" tag in the post-processing phase.

These tags were used in combination to fully curate the annotations generated by the existing Europe PMC pipeline. For example,

- A correctly annotated Gene/Proteins (both entity type and annotation boundary) would be marked CRT_GP.
- A wrong Diseases annotation would be marked WT_DS; and if it had been for an organism; that would be marked as: [WT_DS][OG].

Figure 8 presents an example of the differences in curation among the annotators from batch 1 of the annotations.

Gene-disease associations. While the primary objective of this initiative was entity annotation, annotators were additionally instructed to tag sentences that feature co-occurrences of Gene/Protein and Disease mentions. This was done to identify associations between them, leading to the development of a separate annotation scheme for these associations.

Annotators used the tags YGD, NGD, and AMB, where YGD indicates the presence of a gene-disease association in the sentence, NGD signifies the absence of such an association, and AMB denotes ambiguity in the relationship. Examples of each type of tag can be found in the supplementary information under "Demo to Molecular Connections (Tag schema for annotations)." The first 1,000 sentences featuring co-occurrences of a gene/protein and disease were annotated. The inter-rater agreement for classifying the type of association was very high, as illustrated in Fig. 11.

Annotation extraction and processing. Hypothes.is (https://web.hypothes.is) is a free, open and user-friendly platform enabling annotation of web content. The annotators used Hypothes.is to highlight the span of the entity terms, add notes, and tag them with one of the available tags. They reviewed and marked pre-annotated terms as correct or incorrect and saved them using the Hypothes.is platform.

At Europe PMC, sentence boundaries are added to the article XML files using an in-house sentence segmenter prior to entity recognition. The Europe PMC text-mining pipeline annotates the bio-entities using a dictionary-based approach and displays them on the front-end HTML version via the web application (SciLite, which requires further processing of the annotated XML file). The Hypothes.is platform works on the front-end HTML version of the article. Each annotator set up a Hypothes.is account and thus their annotations were saved to the Hypothes.is server (Please refer to Section 'How to use the interface' in the supplementary information "Demo to Molecular Connections" for detailed instructions). We retrieved the annotations using the Hypothe.is API in JSON format and it was converted to a CSV format using in-house tools. The Hypothes.is JSON reported the annotated terms and their locations with respect to the HTML version of the article.

The annotations from the JSON file were extracted or tagged in the sentence-segmented XML file using regular expressions. However, due to the inconsistency between the HTML article page and the XML file, a small number of annotations could not be successfully extracted using regular expressions. We have identified that failure often occurs when an annotation is in a table. We post-processed the Hypothe.is JSON files for presenting the corpus to the wider community in multiple formats. More details are in the following sections. Figure 9 shows an overview of the process.

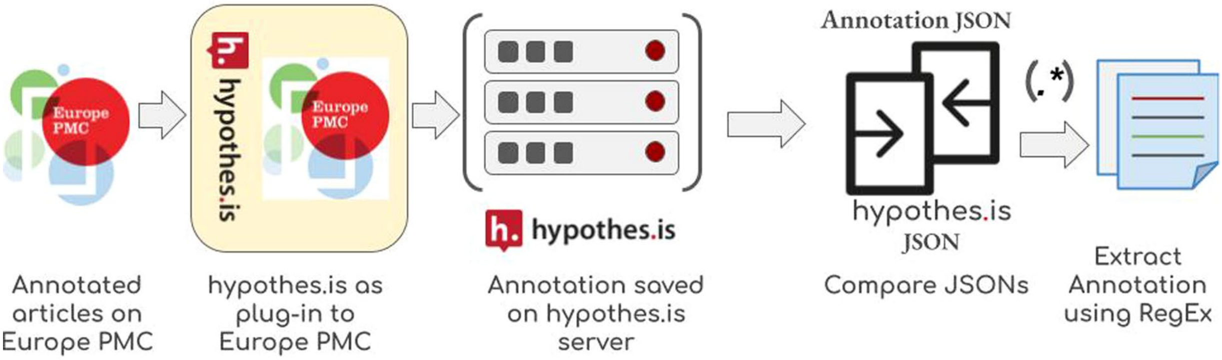

Annotated articles on Europe PMC
hypothes.is as plug-in to Europe PMC
Annotation saved on hypothes.is server
Annotation JSON
hypothes.is JSON
Compare JSONs
(.*)
Extract Annotation using RegEx
Europe PMC
hypothes.is

Fig. 9 Annotation extraction workflow. Hypothes.is was added onto Europe PMC as a plug-in for the annotation work. Annotators saved their annotations to the Hypothes.is server in JSON format and it was retrieved and converted to CSV format using in-house tools. Europe PMC parses the XML version of the articles for sentence tagging and annotating named entities and displays an HTML version on the front end. We compared the hypothe.is annotation JSON files against the XML version and extracted the annotations using regular expressions.

Data Records

The dataset is available at Figshare[40]: https://figshare.com/articles/dataset/Europe_PMC_Full_Text_Corpus/22848380.

To fit the diverse needs of the annotation users, the corpus provides multiple formats of annotations from the raw annotations of Hypothes.is platform (in CSV format) to the standard and ready-to-use IOB format. In addition to the annotations, original full-text articles are released in XML format without the tags.

1. Stand-alone curator annotations.
   (a). CSV
   (b). JSON
   (c). Inside-outside-beginning (IOB)
2. Full-text XML files (without EPMC annotations)
3. Full-text XMLs with sentence boundary (we add <SENT> tag to annotate the sentence boundary)
4. Europe PMC annotation in JSON format.

With the raw annotations in CSV format and full-text XML files, researchers can apply their own text-mining tools to extract the annotations. The comma-separated values (CSV) raw annotation files contain three fields (exact, prefix, and suffix) that are critical to locating the human annotations. "exact" is the annotation itself while "prefix" and "suffix" are characters before and after the annotation, respectively. By combining "prefix", "exact", and "suffix", the snippet can locate the annotation using regular expressions. Raw annotations from all three human annotators are available on Figshare[40], which are helpful for studies of agreement between annotators. Annotations in JavaScript Object Notation (JSON) and IOB formats are provided in addition to raw annotations. Both JSON and IOB format annotations are preprocessed so that only annotations agreed on by at least two annotators are included. The IOB format provides sentences with IOB tags and follows the CoNLL NER corpus standards[41]. While the IOB format is widely used in named entity recognition (NER), researchers may prefer other tagging formats so the JSON format provides sentences and annotations for researchers that are interested in transforming annotations into other tagging formats. Full-text articles are also available in the format that articles are split into sentences by the Europe PMC text mining pipeline.

Technical Validation

This paper presents a corpus of 300 full-text open access articles from the biomedical domain, human-curated with the entities Gene/Proteins, Diseases, and Organisms. Eight articles from the corpus do not contain any entity annotations because the human annotators removed existing dictionary-based annotations as false positives. These articles came from 5 different bins. Tables 3, 4 show an overview of the human-annotated terms and compares these to the existing Europe PMC dictionary-based approach. To evaluate the dictionary-based approach, we applied majority voting acceptance criteria on the granular level annotation tags, that is, entity type tags (GP, DS, OG) along with the correctness tags (CRT, MIS, WT, WS). The annotations were tagged without direct reliance on the ontologies. The terms we annotated were subsequently mapped to the databases and resources detailed in the "Ontologies/Terminologie" of Section 0. This mapping process is responsible for the statistics presented in Table 3 under the category "Normalized to a DB entry".

The triple-anonymous annotation approach had an overall inter-annotator agreement of 0.99. At this level, we assigned granular tags to appropriate entity types. For example, CRT_GP and WS_GP tags were mapped to the GP tag and used the strict evaluation rule for the inter-annotator agreement. The strict evaluation is defined

		Europe PMC dictionary-based				Gold-standard human annotation			
		Gene/ Protein	Disease	Organism	Total	Gene/ Protein	Disease	Organism	Total
Annotations	Total	28,869	10,515	18,040	57,425	36,369	14,518	21,491	72,378
	Unique	3,419	1,752	1,700	6,871	5,600	2,037	2,347	9,970
Normalised to a DB entry	Total	—	—	—	—	21,664	8,476	16,021	46,161
Median per article	Total	53.5	19.5	34	170	54.5	16	30	192
	Unique	12	8	8	36	13	6.5	8	44.5
Max annotation per article	Total	722	219	407	955	795	478	456	940
	Unique	113	78	111	156	178	76	170	201

Table 3. Overall annotation statistics comparing the existing Europe PMC dictionary-based text mining approach to the human curation for the selected 300 gold-standard articles. Overall we have gained around 11k term annotations, with the highest gain existing for the Gene/Protein category. We report unique term count based on the string match and how many normalise to a database identifier of the databases mentioned above rather than unique database identifier counts.

	Gene/Protein		Disease		Organism		
	Unique	Total	Unique	Total	Unique	Total	Overall
Correct	2,551	20,832	1,309	7,518	1,351	15,353	43,703
Added	2,671	13,718	575	5,836	820	5,307	24,230
Removed	697	6,172	447	1,991	207	982	8,514
Modified	561	1,819	269	1,164	311	831	4,445
Precision	0.72		0.70		0.89		0.77
Recall	0.60		0.56		0.74		0.64
F1-score	0.65		0.62		0.80		0.70

Table 4. Evaluation of current Europe PMC dictionary workflow against the human annotation. This table shows the number of dictionary-based Europe PMC annotations updated by human annotators. A large proportion of the Europe PMC annotations are confirmed as correct by the human annotators, although they also added/annotated a significant number of previously unidentified/unannotated terms. The Europe PMC pipeline misses a proportion of these terms due to outdated dictionaries. The removed terms are often common English words or short acronyms. Gene/Protein terms (GP) are more likely to be removed than other entity types, that is, Diseases (DS) and Organisms (OG), due to the frequency of occurrence and the false positive rate for three-letter gene-protein acronyms. This row also counts the annotation where the dictionary-based approach wrongly assigned the type (WT), e.g. Diseases entities wrongly tagged as Gene/Proteins (WT_GP) by the Europe PMC dictionary-based approach (annotators used WT_GP, DS tag) will be added to the 'removed' cell count for the Gene/Proteins and 'added' cell for the Diseases. The "Modified" row shows the number of entities that were modified/split into multiple entities (WS). The overall column is the summation of correctness tags (CRT), i.e. CRT, Missing (MIS) and Wrong Span (WS), going under the Correct, Added and Modified rows. For the WT tag, they were split into two, one under the Removed column and the rest under the Modified row. When an annotation is assigned WT_GP, it means that it is a wrong Gene/Proteins annotation and removed from the annotation set, whereas the [WT_GP, DS] tag means the annotation was not removed from the annotation set, but the entity type is modified.

Agreed by	Gene/Protein	Disease	Organism	Overall
1 annotator	270	178	319	767
2 annotators	480	309	216	1005
3 annotators	35934	14237	21298	71469

Table 5. Inter-annotator agreement statistics. We evaluated annotation agreement using SemEval-2013 Task9.1 strict rule. According to the strict evaluation rule, an annotation agreement is reached only when two annotators agree on the term span and the annotation type. We achieved an overall agreement of 0.99. The first row of this table shows the entity-level breakdown of annotations that were rejected due to the voting system, i.e. at least two annotators must agree on the annotation term, boundary and entity type. Some of these entities were annotated by the other annotators with different entity boundaries.

in the SemEval 2013 Task 9.1[41] where an entity is considered correct only if both its boundary and type match. High inter-annotator agreement with the strictest methods shows that most of the annotations were agreed upon by all three annotators (Table 5). A total of 767 annotations were discarded because just one annotator annotated them. Among these discarded annotations, 289 annotations had overlapping text spans, with the 1,005

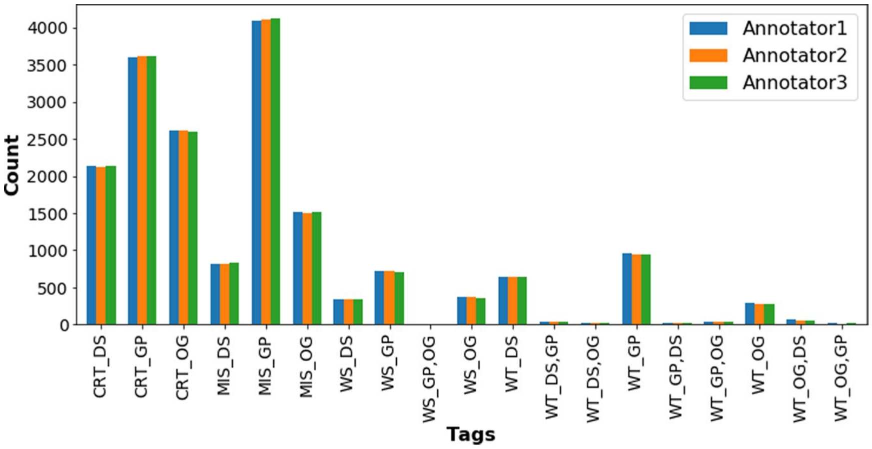


Fig. 10 Entity tags distribution of the corpus and the comparison among the annotators. A large number of Gene/Proteins terms are missed by the dictionary annotation. This figure demonstrates high inter-annotator agreement; correct (CRT), missed (MIS), wrong span (WS), and wrong type (WT). The latter part of the tag represents the entity type namely, Disease (DS), Gene/Protein (GP), and Organisms (OG). Annotators use the WT keyword to remove an annotation and to change the entity type of annotation. They submit the correct entity type by adding the correct entity type keyword after the WT tag, e.g. WT_OG, DS.

	human annotation			
		Gene/Protein	**Disease**	**Organism**
Europe PMC Annotation	Gene/Protein	—	324	113
	Disease	47	—	18
	Organism	19	110	—

Table 6. Europe PMC dictionary-based entity annotation follows a sequential manner to annotate the entities. For example, we apply the Gene/Proteins dictionary before the Disease dictionary, making the Gene/Protein terms unavailable for the disease tagger. We minimise the false positive identifications through this approach. This table shows the number of wrong entity type assignments by the Europe PMC approach corrected by the manual annotators. Europe PMC misses a small percentage of the Disease and Organism entities due to the sequential approach. We are showing Europe PMC annotation in the rows and the manually corrected ones in the columns.

annotations agreed upon by two annotators. For example, two annotators annotated “Welsh Mountain sheep”. However, the third annotator only annotated “sheep” from “Welsh Mountain sheep”. Both of them are correct in terms of the definition of species. Only 478 annotations were truly discarded, accounting for 0.7% of total annotations. Further inspection of the discarded annotations may validate some and help keep the correct ones, but we did not consider this to be a major blocking task.

Our analysis of the distribution of tags set (Fig. 10) shows the highest number of missing terms by the dictionary-based approach is from the Gene/Proteins type (MIS_GP tag). This might be due to the fact that our Gene/Proteins dictionary was last updated in 2014. Updating an entity dictionary involves a number of manual human edits, making it difficult to maintain. Although we were aware of the limitations of the common-stop list approach to limiting false positives, human annotation showed only a small number of these terms (1.6% tagged as MIS_GP) were inappropriately excluded. Using this gold-standard data to train the state-of-the-art machine learning/deep learning models for entity recognition eliminates these challenges. We observe the same trend for the false-positive identifications, i.e. WT_[GP|DS|OG]. The highest number of false positives are from the Gene/Proteins type followed by the Diseases and Organisms terms, respectively. The wrong-type annotation counts are quite low; annotators only correct the entity type for a small number of annotations. This perhaps reflects the way the Europe PMC annotation pipeline works. This pipeline applies dictionaries sequentially, first the Gene/Proteins dictionary, followed by the Diseases dictionary, and then the Organisms dictionary. Once an entity is tagged, it becomes unavailable to tag with subsequent dictionaries, likely reducing false-positive Diseases and Organisms entity identifications. Our analysis shows only a few terms were assigned to the wrong entity type due to this approach, proving our sequential method works. Table 6 shows how many term annotations were updated to reassign the entity type.

The special ‘ALL’ tag was used to indicate that the annotation of a term applies to all occurrences of the term within the article. This was a significant time-saver for articles that mention a particular entity tens or hundreds of times. A total of 23,281 (7,336 unique) terms were tagged ‘ALL’.

Because Hypothes.is allows free text in the tag field, we identified a small number of errors in the tag names; for example, ten annotations from annotators 1 and 2 use ‘DIS’ instead of ‘DS’; one annotation uses ‘CRt’ instead of ‘CRT’. We corrected these errors for downstream analysis.

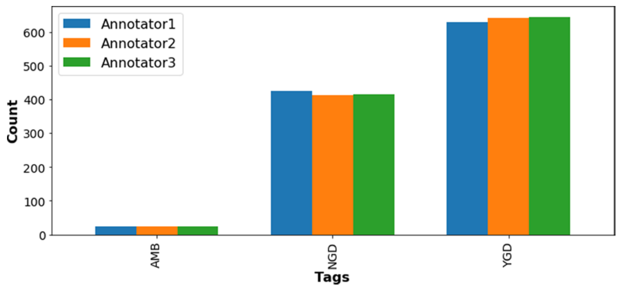


Fig. 11 Association tags distribution of the corpus and the comparison among the annotators, demonstrating high inter-annotator agreement among the annotators across the tags ambiguous (AMB), no gene-disease (NGD) and yes gene-disease (YGD) associations.

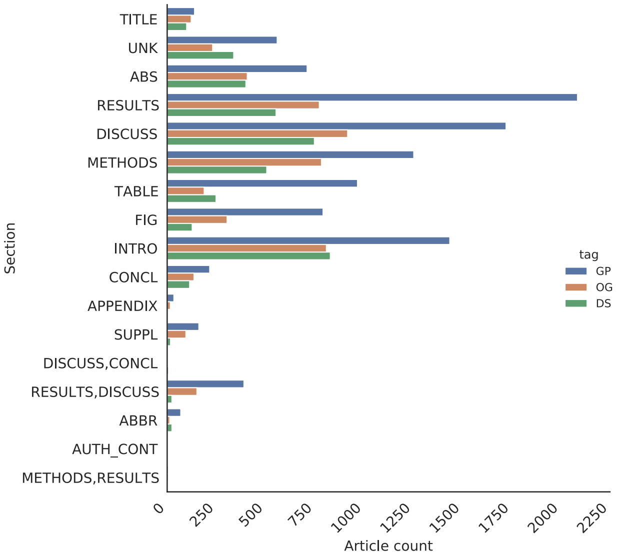


Fig. 12 Term frequency distribution across different sections. The result, discussion, method, and introduction sections contain the highest number of entity mentions. Gene/Proteins mentions in tables and figure titles are significantly higher than Diseases and Organisms mentions.

The titles of sections within a research article can vary widely but typically fall into a small number of categories. For example, "Methods" and "Methods and Reagents" are both classed as Methods sections. In the Europe PMC annotation pipeline, section titles are normalised to a set of 17 titles[42]. Fig. 12 shows the entity distribution across these sections. As anticipated, we found a high frequency of entity mentions in an article's main sections, which demonstrates the value of full-text annotation versus using only abstracts[43]. This entity distribution may help design a targeted annotation approach when resources are limited.

Code availability

The code is available at the repository https://gitlab.ebi.ac.uk/literature-services/public-projects/europepmc-corpus/ and also on Figshare[40]. The scripts include cleaning and formatting the annotations from Hypothes.is platform and generates the dataset in IOB format for input to deep learning algorithms.

Received: 26 May 2023; Accepted: 3 October 2023;
Published online: 19 October 2023

References

1. Ferguson, C. *et al*. Europe pmc in 2020. *Nucleic acids research* **49**, D1507–D1514 (2021).
2. Allot, A., Lee, K., Chen, Q., Luo, L. & Lu, Z. Litsuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Research* **49**, W352–W358 (2021).
3. Wei, C.-H., Kao, H.-Y. & Lu, Z. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research* **41**, W518–W522 (2013).
4. Coudert, E. *et al*. Annotation of biologically relevant ligands in uniprotkb using chebi. *Bioinformatics* **39**, btac793 (2023).
5. Consortium, T. U. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).
6. Wei, C.-H., Allot, A., Leaman, R. & Lu, Z. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research* **47**, W587–W593 (2019).

7. Fairification process. https://www.go-fair.org/fair-principles/fairification-process/. (Accessed on 27/01/2022).
8. Jacobsen, A. *et al*. A generic workflow for the data fairification process. *Data Intelligence* **2**, 56–65 (2020).
9. Sinaci, A. A. *et al*. From raw data to fair data: the fairification workflow for health research. *Methods of information in medicine* **59**, e21–e32 (2020).
10. Venkatesan, A. *et al*. Scilite: a platform for displaying text-mined annotations as a means to link research articles with biological data. *Wellcome open research* **1**, 25 (2017).
11. Szklarczyk, D. *et al*. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605–D612, https://doi.org/10.1093/nar/gkaa1074 (2020).
12. Zahn-Zabal, M. *et al*. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research* **48**, D328–D334, https://doi.org/10.1093/nar/gkz995 (2019).
13. Wang, L. L. *et al*. Cord-19: The covid-19 open research dataset. *ArXiv* (2020).
14. Comeau, D. C., Wei, C.-H., Islamaj Doğan, R. & Lu, Z. Pmc text mining subset in bioc: about three million full-text articles and growing. *Bioinformatics* **35**, 3533–3535 (2019).
15. Tsatsaronis, G. *et al*. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text* (Arlington, VA: Citeseer, 2012).
16. Bionlp workshop. https://aclweb.org/aclwiki/BioNLP_Workshop. (Accessed on 27/01/2022).
17. Nentidis, A., Bougiatiotis, K., Krithara, A. & Paliouras, G. Results of the seventh edition of the bioasq challenge. In *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, 553–568 (Springer, 2020).
18. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
19. Radford, A. *et al*. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
20. Li, J. *et al*. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database* **2016** (2016).
21. Smith, L. *et al*. Overview of biocreative ii gene mention recognition. *Genome biology* **9**, 1–19 (2008).
22. Pyysalo, S. *et al*. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics* **8**, 50, https://doi.org/10.1186/1471-2105-8-50 (2007).
23. Pafilis, E. *et al*. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one* **8**, e65390 (2013).
24. Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M. & Furlong, L. I. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics* **16**, 1–17 (2015).
25. Van Mulligen, E. M. *et al*. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics* **45**, 879–884 (2012).
26. Bagewadi, S., Bobić, T., Hofmann-Apitius, M., Fluck, J. & Klinger, R. Detecting mirna mentions and relations in biomedical literature, https://doi.org/10.12688/f1000research.4591.3 (2015).
27. Doğan, R. I., Leaman, R. & Lu, Z. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics* **47**, 1–10, https://doi.org/10.1016/j.jbi.2013.12.006 (2014).
28. Gerner, M., Nenadic, G. & Bergman, C. M. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics* **11**, 1–17 (2010).
29. Pyysalo, S. & Ananiadou, S. Anatomical entity mention recognition at literature scale. *Bioinformatics* **30**, 868–875, https://academic.oup.com/bioinformatics/article-pdf/30/6/868/48919422/bioinformatics_30_6_868.pdf (2013). 10.1093/bioinformatics/btt580.
30. Bada, M. *et al*. Concept annotation in the craft corpus. *BMC bioinformatics* **13**, 1–20 (2012).
31. Furrer, L., Jancso, A., Colic, N. & Rinaldi, F. Oger++: hybrid multi-type entity recognition. *Journal of cheminformatics* **11**, 1–10 (2019).
32. Ochoa, D. *et al*. The next-generation open targets platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51**, D1353–D1359 (2023).
33. Lee, J. *et al*. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
34. Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* **32**, D267–D270 (2004).
35. Schoch, C. L. *et al*. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020** (2020).
36. Bairoch, A. & Apweiler, R. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research* **28**, 45–48 (2000).
37. Bada, M., Eckert, M., Palmer, M. & Hunter, L. An overview of the craft concept annotation guidelines. In *Proceedings of the Fourth Linguistic Annotation Workshop*, 207–211 (2010).
38. Stenetorp, P. *et al*. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107 (2012).
39. Cunningham, D. M. H. & Bontcheva, K. *Text Processing with GATE (Version 6)*. (University of Sheffield D, 2011).
40. Tirunagari, S. *et al*. Europe PMC Full Text Corpus. *figshare* https://doi.org/10.6084/m9.figshare.22848380.v2 (2023).
41. Segura-Bedmar, I., Martnez, P. & Herrero-Zazo, M. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 341–350 (Association for Computational Linguistics, Atlanta, Georgia, USA, 2013).
42. Kafkas, S. *et al*. Section level search functionality in europe pmc. *Journal of biomedical semantics* **6**, 1–5 (2015).
43. Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J. & Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS computational biology* **14**, e1005962 (2018).

Acknowledgements

We thank the curators at Molecular Connections for their biocuration efforts. We would also extend our thanks to Heidi J. Imker from University of Illinois at Urbana-Champaign and Melissa Harrison Matt Jeffryes, and Mariia Levchenko from team Europe PMC for their valuable suggestions. Last but not least, we would like to thank Europe PMC and Open Targets for strategic guidance on the development of the Europe PMC corpus. This work was supported by: the European Molecular Biology Laboratory-European Bioinformatics Institute (X.Y, A.V); Europe PMC grant, provided by 37 funders of life science research (https://europepmc.org/Funders/) under Wellcome Trust Grant 221523 (A.V, V.V) and the OpenTargets grant 2056 (S.T, S.S).

Author contributions

X.Y. designed and developed annotations guidelines, conceived the experiment(s), analysed the results and wrote the initial draft of the manuscript. S.S. contributed to the writing and revision of the manuscript and analysis of the dataset. A.V. contributed towards annotator guidelines, annotator tags and data structure in Hypothes.is and wrote and revised sections of the manuscripts. S.T. developed scripts for generating machine-trainable IOB formatted datasets, contributed to proofreading, writing and revision of the manuscript, analysed the data and regenerated all the figures for reproducibility. V.V collected and sampled data from Europe PMC database. J.M. conceived the idea and supervised the development of the dataset. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-02617-x.

Correspondence and requests for materials should be addressed to S.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.