A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts

Westergaard, David; Stærfeldt, Hans Henrik; Tønsberg, Christian; Jensen, Lars Juhl; Brunak, Søren; Rzhetsky, Andrey

Published in:
P L o S Computational Biology (Online)

Link to article, DOI:
10.1371/journal.pcbi.1005962

Publication date:
2018

Document Version
Peer reviewed version

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts

David Westergaard[1,2], Hans-Henrik Stærfeldt[1], Christian Tønsberg[3], Lars Juhl Jensen[2]*, Søren Brunak[1]*

1 Center for Biological Sequence Analysis, Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark, **2** Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, **3** Office for Innovation and Sector Services, Technical Information Center of Denmark, Technical University of Denmark, Lyngby, Denmark

* brunak@cbs.dtu.dk (SB); lars.juhl.jensen@cpr.ku.dk (LJJ)

Citation: Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S (2018) A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol 14(2): e1005962. https://doi.org/10.1371/journal.pcbi.1005962

Editor: Andrey Rzhetsky, University of Chicago, UNITED STATES

Received: July 5, 2017

Accepted: January 5, 2018

Published: February 15, 2018

Data Availability Statement: Due to copyright and legal agreements the full text articles cannot be made available. The dictionaries used for named entity recognition can be found at https://doi.org/10.6084/m9.figshare.5827494. The DOIs for the articles can be found at https://doi.org/10.6084/m9.figshare.5419300. The Z-scores used for benchmarking can be found at https://doi.org/10.6084/m9.figshare.5340514. The entities mentioned in articles used for benchmarking can be found at https://doi.org/10.6084/m9.figshare.5620417.

Abstract

Across academia and industry, text mining has become a popular strategy for keeping up with the rapid growth of the scientific literature. Text mining of the scientific literature has mostly been carried out on collections of abstracts, due to their availability. Here we present an analysis of 15 million English scientific full-text articles published during the period 1823–2016. We describe the development in article length and publication sub-topics during these nearly 250 years. We showcase the potential of text mining by extracting published protein–protein, disease–gene, and protein subcellular associations using a named entity recognition system, and quantitatively report on their accuracy using gold standard benchmark data sets. We subsequently compare the findings to corresponding results obtained on 16.5 million abstracts included in MEDLINE and show that text mining of full-text articles consistently outperforms using abstracts only.

Author summary

Text mining has become an integral part of all fields in science. Owing to the large number of articles published every day, it is necessary to employ automated systems to assist in curation, knowledge management and discovery. To date, most systems make use of information collected from abstracts only. Moreover, studies on smaller collections of abstracts and full-text articles have demonstrated some information is available only in the full-text body. Nonetheless, to date there has been no large-scale comprehensive comparison of abstracts and full-text articles. In this work, we analyze a hitherto unprecedented collection of 15 million full-text articles. Through quantitative benchmarks we assess the difference between full-text articles and abstracts. Our findings confirm what has long been discussed, namely that access to the full-text body improved text mining greatly.

Funding: This work was funded by a grant from the Danish e-Infrastructure Cooperation (ActionableBiomarkersDK, https://www.deic.dk/ (SB), and by the Novo Nordisk Foundation (grant agreement NNF14CC0001, http://novonordiskfonden.dk/) (SB, LJJ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: SB and LJJ are on the scientific advisory board and have been among the founders of Intomics A/S with equity in the company.

Introduction

Text mining has become a widespread approach to identify and extract information from unstructured text. Text mining is used to extract facts and relationships in a structured form that can be used to annotate specialized databases, to transfer knowledge between domains and more generally within business intelligence to support operational and strategic decision-making [1–3]. Biomedical text mining is concerned with the extraction of information regarding biological entities, such as genes and proteins, phenotypes, or even more broadly biological pathways (reviewed extensively in [3–9]) from sources like scientific literature, electronic patient records, and most recently patents [10–13]. Furthermore, the extracted information has been used as annotation of specialized databases and tools (reviewed in [3,14]). In addition, text mining is routinely used to support manual curation of biological databases [15,16]. Thus, text mining has become an integral part of many resources serving a wide audience of scientists. The main text source for scientific literature has been the MEDLINE corpus of abstracts, essentially due to the restricted availability of full-text articles. However, full-text articles are becoming more accessible and there is a growing interest in text mining of complete articles. Nevertheless, to date no studies have presented a systematic comparison of the performance comparing a very large number of abstracts and full-texts in corpora that are similar in size to MEDLINE.

Full-text articles and abstracts are structurally different [17]. Abstracts are comprised of shorter sentences and very succinct text presenting only the most important findings. By comparison, full-text articles contain complex tables, display items and references. Moreover, they present existing and generally accepted knowledge in the introduction (often presented in the context of summaries of the findings), and move on to reporting more in-depth results, while discussion sections put the results in perspective and mention limitations and concerns. The latter is often considered to be more speculative compared to the abstract [3].

While text-mining results from accessible full-text articles have already become an integral part of some databases (reviewed recently for protein-protein interactions [18]), very few studies to date have compared text mining of abstracts and full-text articles. Using a corpus consisting of ~20,000 articles from the PubMed Central (PMC) open-access subset and Directory of Open Access Journals (DOAJ), it was found that many explicit protein–protein interactions only are mentioned in the full text [19]. Additionally, in a corpus of 1,025 full-text articles it was noticed that some pharmacogenomics associations are only found in the full text [20]. One study using a corpus of 3,800 articles with focus on *Caenorhabditis elegans* noted an increase in recall from 45% to 95% when including the full text [21]. Other studies have worked with even smaller corpora [17,22,23]. One study have even noted that the majority of claims within an article is not reported in the abstract [24]. Whilst these studies have been of significant interest, the number of full-text articles and abstracts used for comparison are nowhere near the magnitude of the actual number of scientific articles published to date, and it is thus unclear if the results can be generalized to the scientific literature as a whole. The earlier studies have mostly used articles retrieved from PMC in a structured XML file. However, full-text articles received or downloaded directly from the publishers often come in the PDF format, which must be converted to a raw unformatted text file. This presents a challenge, as the quality of the text mining will depend on the proper extraction and filtering of the unformatted text. A previous study dealt with this by writing custom software taking into account the structure and font of each journal at that time [21]. More recent studies typically provide algorithms that automatically determines the layout of the articles [25–27].

In this work, we describe a corpus of 15 million full-text scientific articles from Elsevier, Springer, and the open-access subset of PMC. The articles were published during the period

1823–2016. We highlight the possibilities by extracting protein–protein associations, disease–gene associations, and protein subcellular localization from the large collection of full-text articles using a Named Entity Recognition (NER) system combined with a scoring of co-mentions. We quantitatively report the accuracy and performance using gold standard benchmark data sets. Lastly, we compare the findings to corresponding results obtained on the matching set of abstracts included in MEDLINE as well as the full set of 16.5 million MEDLINE abstracts.

Methods

MEDLINE corpus

The MEDLINE corpus consists of 26,385,631 citations. We removed empty citations, corrections and duplicate PubMed IDs. For duplicate PubMed IDs we kept only the newest entry. This led to a total of 16,544,511 abstracts for text mining.

PMC corpus

The PubMed Central corpus comprises 1,488,927 freely available scientific articles (downloaded 27th January 2017). Each article was retrieved in XML format. The XML file contains the article divided into paragraphs, article category and meta-information such as journal, year published, etc. Articles that had a category matching Addendum, Corrigendum, Erratum or Retraction were discarded. A total of 5,807 documents were discarded due to this, yielding a total of 1,483,120 articles for text mining. The article paragraphs were extracted for text mining. No further pre-processing of the text was done. The journals were categorized according to categories (described in the following section) by matching the ISSN number. The number of pages for each article was also extracted from the XML, if possible.

TDM corpus

The Technical Information Center of Denmark (DTU Library) TDM corpus is a collection of full-text articles from the publishers Springer and Elsevier. The corpus covers the period from 1823 to 2016. The corpus comprises 3,335,400 and 11,697,096 full-text articles in PDF format, respectively. An XML file containing meta-data such as publication date, journal, etc. accompanies each full-text article. PDF to TXT conversion was done using pdftotext v0.47.0, part of the Poppler suite (poppler.freedesktop.org). 192 articles could not be converted to text due to errors in the PDF file. The article length, counted as the number of pages, was extracted from the XML file. If not recorded in the XML file we counted the number of pages in the PDF file using the Unix tool pdfinfo v0.26.5. Articles were grouped into four bins, determined from the 25%, 50%, and 75% quantiles, respectively. These were found to be 1–4 pages (0–25%), 5–7 pages (25–50%), 8–10 pages (50–75%) and 11+ pages (75%-100%). Each article was, based on the journal where it was published, assigned to one or more of the following seventeen categories: Health Sciences, Chemistry, Life Sciences, Engineering, Physics, Agriculture Sciences, Material Science and Metallurgy, Earth Sciences, Mathematical Sciences, Environmental Sciences, Information Technology, Social Sciences, Business and Economy and Management, Arts and Humanities, Law, Telecommunications Technology, Library and Information Sciences. Due to the large number of categories, we condensed anything not in the top-6 into the category "Other". The top-six categories *health science*, *chemistry*, *life sciences*, *engineering*, *physics* and *agricultural sciences* make up 74.8% of the data (S1 Fig). The assignment of categories used in this study was taken from the existing index for the journal made by the librarians at the DTU Library. For the temporal statistics, the years 1823–1900 were condensed into one.

Pre-processing of PDF-to-text converted documents

Following the PDF-to-text conversion of the Springer and Elsevier articles we ran a language detection algorithm implemented in the python package langdetect v1.0.7 (https://pypi.python.org/pypi/langdetect). We discarded 902,415 articles that were not identified as English. We pre-processed the remaining raw text from the articles as follows:

1. Non-printable characters were removed using the POSIX filter [[:^print:]].
2. A line of text was removed if digits make up more than 10% of the text, or symbols make up more than 10% of the text, or lowercase text was less than 50%. Symbols are anything not matching [0-9A-Za-z].
3. Removal of acknowledgments and reference- or bibliography-lists using a rule-based system explained below.
4. Text was split into sentences and paragraphs using a rule-based system described below.

We assumed that acknowledgments and reference lists are always at the end of the article. Upon encountering either of the terms: "acknowledgment", "bibliography", "literature cited", "literature", "references", and the following misspellings thereof: "refirences", "literatur", "références", "refereses". In some cases the articles had no heading indicating the start of a bibliography. We tried to take these cases into account by constructing a RegEx that matches the typical way of listing references (e.g. [1] Westergaard, . . .). Such a pattern can be matched by the RegEx "^\[\d+\]\s[A-Za-z]". The other commonly used pattern, "1. Westergaard, . . .", was avoided since it may also indicate a new heading. Keywords were identified based on several rounds of manual inspection. In each round, 100 articles in which the reference list had not been found was randomly selected and inspected. We were unable to find references in 286,287 and 2,896,144 Springer and Elsevier articles, respectively. Manual inspection of 100 randomly selected articles revealed that these articles indeed did not have a reference list or that the pattern was not easily describable with simple metrics, such as keywords and RegEx. Articles without references were not discarded.

The PDF to text conversion often breaks up paragraphs and sentences, due to new page, new column, etc. Paragraph and sentence splitting was performed using a ruled-based system. If the previous line of text does not end with a ".!?", and the current line does not start with a lower-case letter, it is assumed that the line is part of the previous sentence. Otherwise, the line of text is assumed to be a new paragraph.

Text article filtering

A number of Springer and Elsevier documents were removed due to technical issues post pre-processing. An article was removed if:

1. Article contained no text post-preprocessing (51,399 documents).
2. Average word length was below the 2% quantile (263,902 documents).
3. Article contained specific keywords, described below (286,958 documents).

Some PDF files without texts are scans of the original article (point 1). We did not attempt to make an optical character recognition conversion (OCR) as the old typesetting fonts often are less compatible with present day OCR programs, and this can lead to text recognition errors [28,29]. For any discarded document, we still used the meta-data to calculate summary statistics. In some cases the PDF to text conversion failed, and produced non-sense data with a white space between the characters of a majority of the words (point 2). To empirically

determine a cutoff we gradually increased the cutoff and repeatedly inspected 100 randomly selected articles. At the 2% quantile we saw no evidence of broken text.

Articles with the following keywords in the article were discarded: Author Index, Key Word Index, Erratum, Editorial Board, Corrigendum, Announcement, Books received, Product news, and Business news (point 3). These keywords were found as part of the process of identifying acknowledgments and reference lists. Further, any article that was available through PubMed Central was preferentially selected by matching doi identifiers. This left a total of 14,549,483 full-text articles for further analysis.

Some articles were not separable, or were subsets of others. For instance, conference proceedings may contain many individual articles in the same PDF. We found 1,911,365 articles in which this was the case. In these cases we removed the duplicates, or the shorter texts, but kept one copy for text mining. In total, we removed 898,048 duplicate text files.

The majority of articles had a separate abstract. We matched articles from PubMed Central to their respective MEDLINE abstract using the PMCID to PubMed ID conversion file available from PMC. Articles from Springer and Elsevier typically had a separate abstract in the meta-data. Any abstract from an article that was part of the 1,911,365 articles that could not be separated was removed. This led to a total of 10,376,626 abstracts for which the corresponding full text was also included downstream, facilitating a comparative analysis.

References for the full text articles analyzed can be found at 10.6084/m9.figshare.5419300. An article is preferentially referenced by its Digital Object Identifier (DOI) (98.8%). However, if that was not available, we used the PubMed Central ID for PMC articles (0.005%), or the list of authors, article title, journal name, and year. (0.006%)

Text mining of articles

We performed text mining of the articles using a Named Entity Recognition (NER) system, described earlier [30–33]. The software is open source and can be downloaded from https://bitbucket.org/larsjuhljensen/tagger. The NER approach is dictionary based, and thus depends on well-constructed dictionaries and stop word lists. We used the gene names from the STRING dictionary v10.0 [30], disease names from the Disease Ontology (DO) [34] and compartment names from the Gene Ontology branch cellular component [35]. Stop word lists were all created and maintained in-house. Pure NER based approaches often struggles with ambiguity of words. Therefore, we included additional dictionaries that we do not report the results from. If any identified term was found in multiple dictionaries, it was discarded due to ambiguity. The additional dictionaries include small molecule names from STITCH [36], tissue names from the Brenda Tissue Ontology [37], Gene Ontology biological process and molecular function [35], and the mammalian phenotype ontology [38]. The latter is a modified version made to avoid clashes with the disease ontology. The dictionaries can be downloaded from https://doi.org/10.6084/m9.figshare.5827494.

In the cases where the dictionary was constructed from an ontology co-occurrences were backtracked through all parents. E.g. the term type 1 diabetes mellitus from the Disease Ontology is backtracked to its parent, diabetes mellitus, then to glucose metabolism disease, etc.

Co-occurrences were scored using the scoring system described in [39]. In short, a weighted count for each pair of entities (e.g. disease-gene) was calculated using the formula,

$$C(i,j) = \sum_{k=1}^{n} w_d \delta_{dk}(i,j) + w_p \delta_{pk}(i,j) + w_s \delta_{sk}(i,j) \quad (1)$$

where δ is an indicator function taking into account whether the terms i,j co-occur within the same document (d), paragraph (p), or sentence (s). w is the co-occurrence weight here set to

1.0, 2.0, and 0.2, respectively. Based on the weighted count, the score *S(i,j)* was calculated as,

$$S(i,j) = C_{ij}^{\alpha}\left(\frac{C_{ij}C_{..}}{C_{i.}C_{.j}}\right)^{1-\alpha} \quad (2)$$

where α is set to 0.6. All weights were optimized using the KEGG pathway maps as benchmark (described further below). The *S* scores were converted to *Z* scores, as described earlier [40].

Benchmarking of associations

PPIs were benchmarked using pathway maps from the KEGG database [41–43]. Any two proteins in the same pathway were set to be a positive example, and any two proteins present in at least one pathway, but not the same, were set as a negative example. This approach assumes that the pathways are near complete and includes all relevant proteins. The same approach has been used for the STRING database [39]. The disease–gene benchmarking set was created by setting the disease-gene associations from UniProt [44] and Genetics Home Reference (https://ghr.nlm.nih.gov/, accessed 23th March 2017) as positive examples. The positive examples were then shuffled, and the shuffled examples were set as negative examples. Shuffled examples that ended up overlapping with the positive examples were discarded. This approach has previously been described [31]. The protein–compartment benchmark set was created by extracting the compartment information for each protein from UniProt and counting these as positive examples. For every protein found in at least one compartment, all compartments where it was not found were set as negative examples. The same approach has been used previously [33].

Receiver Operating Characteristic (ROC) curves were created by gradually increasing the *Z*-score and calculating the True Positive Rate (TPR) and False Positive Rate (FPR), as described in eqs (3) and (4).

$$TPR = \frac{True\ Positives}{Positives} \quad (3)$$

$$FPR = \frac{True\ Negatives}{Negatives} \quad (4)$$

We compare the ROC curves by the Area Under the Curve (AUC), a metric ranging from 0 to 1. ROC-AUC curves provide a quantitative way of comparing benchmarks of classifiers, and is often used in machine learning and text mining. A perfect classifier will have an AUC = 1, and a classifier that performs equal to or worse than random will have an AUC $\leq$ 0.5.

Individual mentions of entities used for the benchmark in each article can be downloaded from 10.6084/m9.figshare.5620417.

Results

We analyzed and compared four different corpora comprising all full-text articles (14,549,483 articles, All Full-texts), full-text articles that had a separate abstract (10,376,626 articles, Core Full-texts), the abstract from the full-text articles (10,376,626 abstracts, Core Abstracts), and the MEDLINE corpus (16,544,511 abstracts, MEDLINE) (see Methods for a detailed description of the pre-processing).

Growth and temporal development in full text corpora

The growth of the data set over time is of general interest in itself, however, it is also important to secure that the concepts used in the benchmarks are likely to be present in a large part of the

corpus. We found that the number of full-text articles has grown exponentially over a long period (Fig 1A, a log-transformed version is provided in S2 Fig). We also observed that the growth represents a mixture of two components: one from 1823–1944, and another from 1945–2016. Through linear regression of the $\log_2$-transformed counts for the period 1945–2016 we found that the growth rate is 0.103 ($p < 2 * 10^{-16}$, $R^2 = 0.95$). Thus, the doubling time for the full-text corpus is 9.7 years. In comparison, MEDLINE had a growth rate of 0.195 ($p < 2 * 10^{-16}$, $R^2 = 0.91$) and a doubling time of 5.1 years. We noticed that there was a drop in the number of full-text publications around the years 1914–1918 and 1940–1945. Likewise, we see a decrease in the number of publications indexed by MEDLINE in the entire period 1930–1948.

In the full-text corpora we found a total of 12,781 unique journal titles. The most prevalent journals are tied to health or life sciences, such as *The Lancet*, *Tetrahedron Letters*, and *Biochemical and Biophysical Research Communications*, or the more broad journals such as *PLoS ONE* (see S1 Table for the top-15 journals). *The Lancet* publishes only very few articles per issue, it was established in 1823 and has been active in publishing since then, thus explaining why it so far has nearly published 400,000 articles. In contrast, *PLoS ONE* was launched in 2006, and has published more than 172,000 articles. Of the 12,781 journal titles, 6,900 had one or more category labels assigned by librarians at the Technical University of Denmark. The vast majority of the full-texts, 13,343,040, were published in journals with one or more category labels. The frequency of each category within the corpus can be seen in S1 Fig. We observed that before the 1950's health science dominated and made up almost 75% of all publications (Fig 1B). At the start of the 1950's the fraction started to decrease, and to date health science makes up approximately 25% of all publications in the full-text corpus. Inspecting the remaining eleven categories in a separate plot we found that there was no single category that was responsible for the growth (S3 Fig).

We binned the full-text articles into four categories based on the number of pages (see Methods). The average length of articles has increased considerably during the almost 250 years studied (Fig 1C). Whereas 75% of the articles were 1–3 pages long at the end of the 20th century, less than 25% of the articles published after year 2000 are that short. Conversely, articles with ten or more pages only made up between 0.7%-7% in the 19th century, a level that had grown to 20% by the start of the 21st century. We also observed that the average number of mentioned entities changed over time (S4 Fig). Mentions of genes and compartments were nearly non-existing prior to 1950, and has been increasing at an exponential rate since year 2000. Moreover, disease mentions dropped around year 1950, which correlates well with the decreasing proportion of published articles from health science journals in our corpus (Fig 1C).

Evaluating information extraction across corpora

We ran the textmining pipeline on the two full-text and two abstract corpora. In all cases we found that the AUC-value was far greater than 0.5, from which we conclude that the results were substantially better than random (Fig 2) (see Methods for a definition of the AUC). The biggest gain in performance when using full-text was seen in finding associations between diseases and genes (AUC increase from 0.85 to 0.91) (Table 1). Compared to MEDLINE, the traditional corpus used for biomedical text mining, there was an increase in the AUC from 0.85 to 0.91. The smallest gain was associations between proteins, which increased from 0.70 to 0.73. Likewise, the Core Full-texts always performed better than Core Abstracts, signifying that some associations are only reported in the main body of the text. Consequently, traditional text mining of abstracts will never be able to find this information. All Z-scores used for benchmarking can be downloaded from 10.6084/m9.figshare.5340514

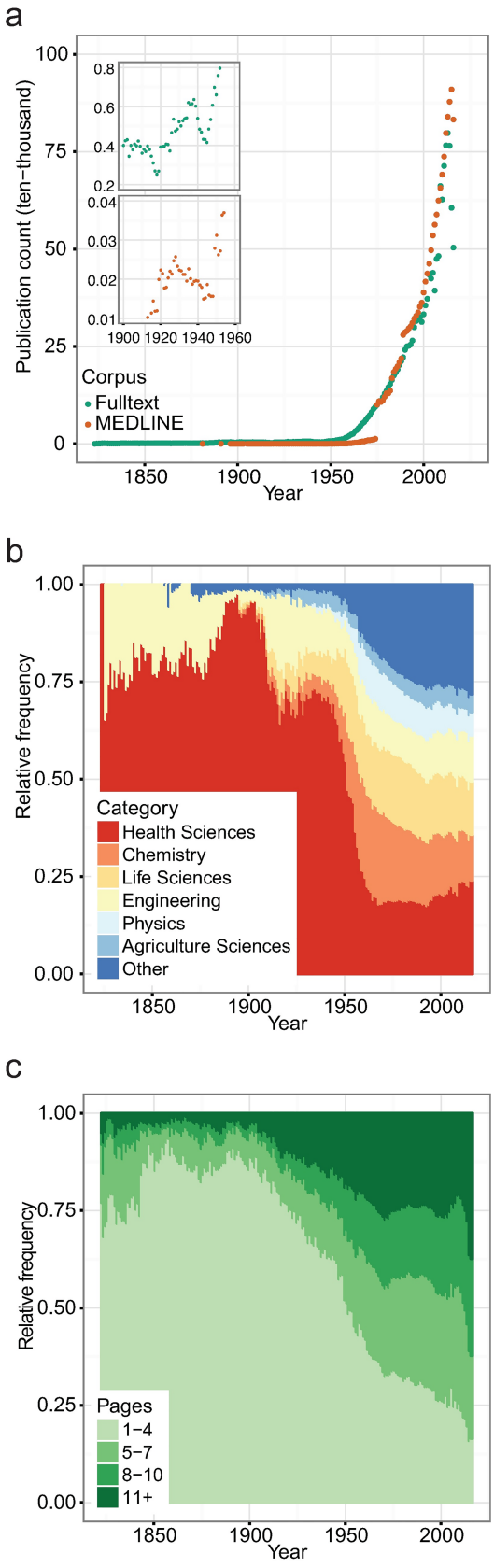
a
Publication count (ten−thousand)
100
75
50
25
0
0.8
0.6
0.4
0.2
0.04
0.03
0.02
0.01
1900 1920 1940 1960
Corpus
Fulltext
MEDLINE
1850
1900
1950
2000
Year
b
Relative frequency
1.00
0.75
0.50
0.25
0.00
Category
Health Sciences
Chemistry
Life Sciences
Engineering
Physics
Agriculture Sciences
Other
1850
1900
1950
2000
Year
c
Relative frequency
1.00
0.75
0.50
0.25
0.00
Pages
1−4
5−7
8−10
11+
1850
1900
1950
2000
Year

Fig 1. Temporal corpus statistics derived from articles passing the pre-processing. (a) Number of publications per year in the period 1823–2016. The full-text corpus encompasses both the PMC and TDM corpus. The growth in publications was found to fit an exponential model. **(b)** Temporal development in the distribution of six different topical categories in the period 1823–2016. Publications from health science journals made up nearly 75% of all publications until 1950, at which point it started to decrease rapidly. To date, it makes up approximately 25% of the publications in the full-text corpus. **(c)** Development in the number of pages per article in the period 1823–2016. The range of pages varies from 1–1,572 pages. Until year 1900 the number of one-page articles were increasing, at one point making up 75% of all articles. At the end of the 19th century, the number of one-page articles started to decrease, and by the start of the 21th century they made up less than 20%. Conversely, the number of articles with 11+ pages has been increasing, and by the start of the 21th century made up more than 20% of all articles.

https://doi.org/10.1371/journal.pcbi.1005962.g001

It has previously been speculated if text mining of full-text articles may be more difficult and lead to an increased rate of false positives [3]. To investigate this we altered the weights of the scoring system (See Methods, Eqs 1 and 2). The scoring scheme used here has weights for within sentence, within paragraph and within document co-occurrences (see Methods). When setting the document weight to zero versus using the previously calibrated value found in an earlier study we found that having a non-zero small value does indeed improve extraction of known facts in all cases (S5 Fig) [33]. We observed that the increase in AUC is less than when using a lower document weight (S2 Table). In one case, protein–protein associations, the MEDLINE abstract corpus outperforms the full-text articles. Abstracts are generally unaffected by the document weight, mainly because abstracts are almost always one paragraph. Overall, the difference in performance gain is largest for full-texts and lowest for abstracts and MEDLINE. Hence, all the full-text information is indeed valuable and necessary.

For practical applications, it is often necessary to have a low False Positive Rate (FPR). Accordingly, we evaluated the True Positive Rate (TPR) of the different corpora at the 10% FPR (TPR@10%FPR) (Fig 3). We found that full-texts have the highest TPR@10%FPR for disease-gene associations (S2 Table). When considering protein–protein associations and protein-compartment associations, full-texts perform equivalently to Core Abstracts and Core Full-texts. The result was similar to when we evaluated the AUC across the full range, removing the document weight has the biggest impact on the full-texts (S5 Fig and S6 Fig), while abstracts remain unaffected.

Discussion

We have investigated a unique corpus consisting of 15 million full-text articles and compared the results to the most commonly used corpus for biomedical text mining, MEDLINE. We found that the full-text corpus outperforms the MEDLINE abstracts in all benchmarked cases, with the exception of TPR@10%FDR for protein–compartment associations. To our knowledge, this is the largest comparative study to date of abstracts and full-text articles. We envision that the results presented here can be used in future applications for discovering novel associations from mining of full-text articles, and as a motivation to always include full-text articles when available and to improve the techniques used for this purpose.

The corpus consisted of 15,032,496 full-text documents, mainly in PDF format. 1,504,674 documents had to be discarded for technical reasons, primarily because they were not in English. Further, a large number of documents were also found to be duplicates or subsets of each other. On manual inspection we found that these were often conference proceedings, collections of articles etc., which were not easily separable without manual curation. We also managed to identify the list of references in the majority of the articles thereby reducing some repetition of knowledge that could otherwise lead to an increase in the false positive rate.

We have encountered and described a number of problems when working with full-text articles converted from PDF to TXT from a large corpus. However, the majority of the

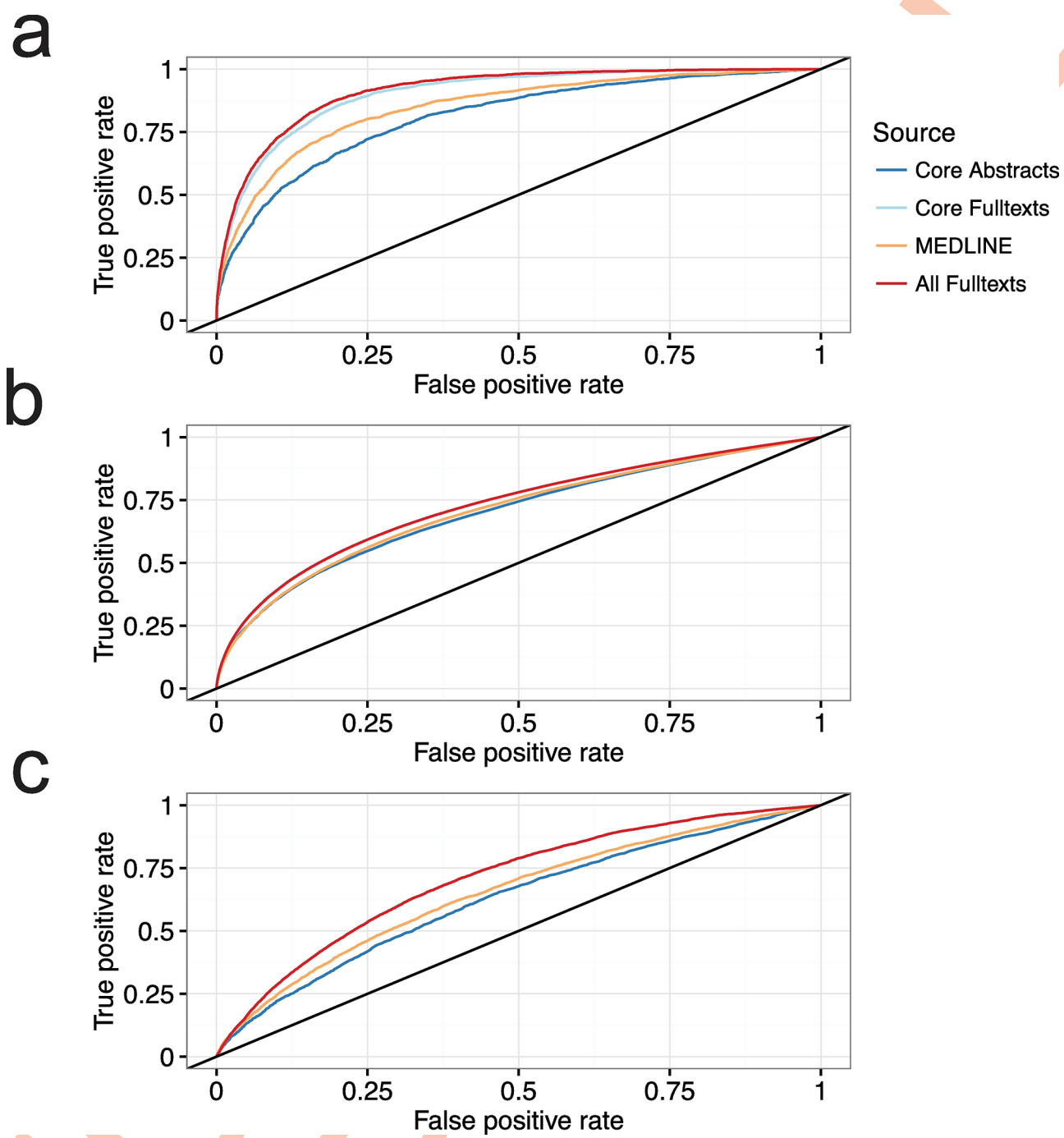

Fig 2. Benchmarking the four different corpora. In all cases the AUC is far greater than 0.5, indicating that the results obtained are better than random. The biggest gain in AUC is seen for disease-gene associations (**a**), followed by protein-compartment associations (**c**) and protein-protein associations (**b**).

https://doi.org/10.1371/journal.pcbi.1005962.g002

problems did not stem from the PDF to TXT conversion, which could potentially be solved using a layout aware conversion tool. Examples include PDFX [25], SectLabel [26] and LA-PDFText [27] of which the first is not practical for very large corpora as it only exists as an online tool. Nonetheless, to make use of the large volume of existing articles it is necessary to solve these problems. Having all the articles in a structured XML format, such as the one

Table 1. Area under the curve (AUC) for the four different corpora, with and without document weight for scoring co-occurrences.

	Disease–gene		Protein–protein		Protein–compartment	
	With document weight	Without document weight	With document weight	Without document weight	With document weight	Without document weight
All full-texts	0.91	0.86	0.73	0.70	0.71	0.67
Core full-texts	0.90	0.86	0.73	0.70	0.71	0.67
Core abstracts	0.81	0.82	0.70	0.70	0.62	0.64
MEDLINE	0.85	0.85	0.70	0.71	0.65	0.66

https://doi.org/10.1371/journal.pcbi.1005962.t001

provided by PubMed Central, would with no doubt produce a higher quality corpus. This may in turn further increase the benchmark results for full-text articles. Nevertheless, the reality is that many articles are not served that way. Consequently, the performance gain we report here should be viewed as a lower limit as we have sacrificed quality in favor of a larger volume of articles. The solutions we have outlined here will serve as a guideline and baseline for future studies.

The increasing article length may have different underlying causes, but one of the main contributors is most likely increased funding to science worldwide [45,46]. Experiments and protocols are consequently getting increasingly complex and interdisciplinary–aspects that also contribute to driving meaningful publication lengths upward. The increased complexity has also been found to affect the language of the articles, as it is becoming more specialized [47]. Moreover, we observed a steep increase in the average number of mentions of genes and compartments. This finding can most likely be attributed to recent developments in molecular biology, such as the sequencing of the human genome, Genome Wide Association Studies (GWAS), and other high-throughput technologies in 'omics [48,49]. It was outside the scope of this paper to go further into socio-economic impact. We have limited this to presenting the trends from what could be computed from the meta-data.

Previous papers are–in terms of benchmarking–only making qualitative statements about the value of full-text articles as compared to text in abstracts. In one paper a single statement is made on the potential for extracting information, but no quantitative evidence is presented [50]. In a paper targeting pharmacogenomics it is similarly stated that that there are associations that only are found in the full-text, but no quantitative estimates are presented [20]. In a paper analyzing around 20,000 full-text papers a search for physical protein interactions was made, concluding that these contain considerable higher levels of interaction [19]. Again, no quantitative benchmarks were made comparing different sources. In this paper, we have made a detailed comparison of four different corpora that provides a strong basis for estimating the added value of using full-text articles in text mining workflows.

We have used quite difficult, but still well established benchmarks, to illustrate the differences in performance when comparing text mining of abstracts to full-text articles. Within biology, and specifically in the area of systems biology, macromolecular interactions and the relationships between genes, tissues and diseases are key data that drive modeling and the analysis of causal biochemical mechanisms. Knowledge of interactions between proteins is extremely useful when revealing the components, which contribute to mechanisms in both health and disease. As many biological species from evolution share protein orthologs, their mutual interactions can often be transferred, for example from an experiment in another organism to the corresponding pair of human proteins where the experiment has not yet been performed. Such correspondences can typically be revealed by text mining as researchers in one area often will not follow the literature in the other and *vice versa*.

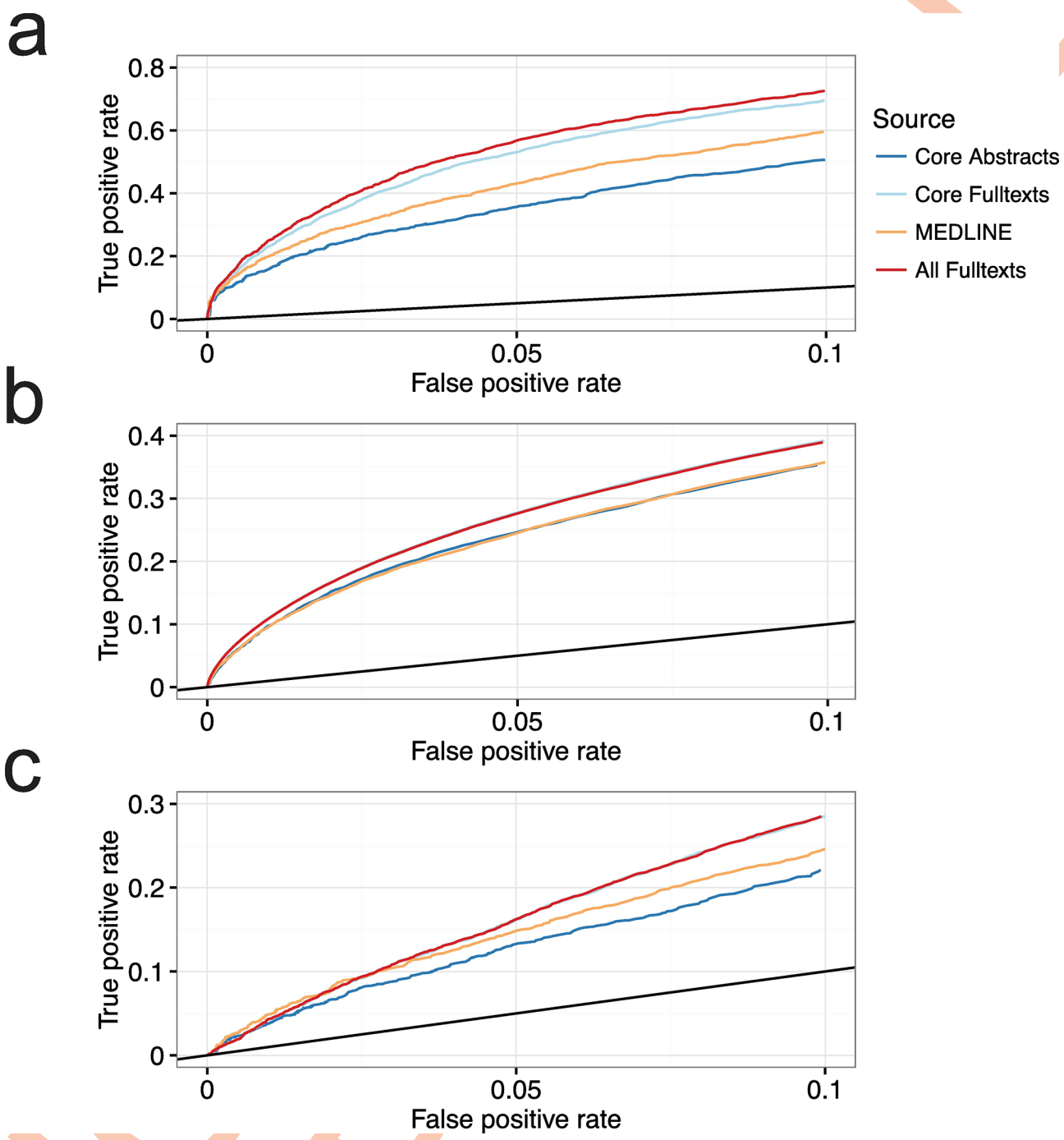

Fig 3. Benchmarking the four different corpora at low false positive rates. At a false positive rate of 10%, relevant to practical applications, the full-text corpus still outperforms the collection of MEDLINE abstracts for the extraction of **(a)** disease-gene associations. Conversely, the performance is the same for **(b)** protein-protein associations and **(c)** protein-compartment associations.

https://doi.org/10.1371/journal.pcbi.1005962.g003

The results presented here are purely associational. Through rigorous benchmarking and comparison of a variety of biologically relevant associations, we have demonstrated that a substantial amount of relevant information is only found in the full body of text. Additionally, by modifying the document weight we found that it was important to take into account the whole

document and not just individual paragraphs. The improvement in AUC that we present here were not overwhelming. One reason could be that associations have a higher probability of being curated if they are mentioned in the abstract. Moreover, most tools are geared towards abstracts. Thus, what we present is a lower limit on the performance gain. Consequently, as more full-text articles become available and text-mining methods improve, the quantitative benchmarks will improve. However, because the literature is highly redundant diminishing returns in terms of performance gain should be expected when adding evermore text. Event-based text mining will be the next step for a deeper interpretation and extending the applicability of the results [5]. With more development it may also be possible to extract quantitative values, as has been demonstrated for pharmacokinetics [51]. Other work is also going into describing the similarity between terms, and how full-text articles can augment this [52]. However, this is beyond the scope of this article.

The Named Entity Recognition (NER) system used depends heavily on the dictionaries and stop word lists. A NER system is also very sensitive to ambiguous words. To combat this we have used dictionaries from well-known and peer-reviewed databases, and we have included other dictionaries to avoid ambiguous terms. Other approaches to text mining have previously been extensively reviewed [10,14,51].

The full-text corpus presented here consists of articles from Springer, Elsevier and PubMed. However, we still believe that the results presented here are valid and can be generalized across publishers, to even bigger corpora. Preprocessing of corpora is an ongoing research project, and it can be difficult to weed out the rubbish when dealing with millions of documents. We have tried to use a process where we evaluate the quality of a subset of randomly selected articles repeatedly and manually, until it no longer improves.

Supporting information

S1 Fig. Category overview across all journals and years. The bar chart indicates the frequency, whilst the line is the cumulative sum. The first six categories contribute 74.8%.
(EPS)

S2 Fig. Number of publications per year on the log scale.
(EPS)

S3 Fig. Temporal trend for the categories embedded in the "Other" category from Fig 1B. Note that the "Other" category has grown as a whole, but that the growth is not tied to one category.
(EPS)

S4 Fig. Average number of mentions of compartments, diseases, and genes per year in the period 1823–2016.
(EPS)

S5 Fig. Benchmarking of the four different corpora when not using a document weight. (a) protein-disease associations, **(b)** protein-protein associations, **(c)** protein-compartment associations. **(a-c)**, Compared to including a document weight, the increase in performance has been reduced. In one case, protein-protein associations, the MEDLINE corpus outperforms the full-text articles. The corresponding figure with document weights is included in Fig 2.
(EPS)

S6 Fig. Benchmarking the four different corpora at a low false positive rate when not using a document weight. The increase in performance is reduced. In one case, for protein-protein associations, the MEDLINE corpus outperforms the full-text articles. The corresponding

figure with document weights is included in Fig 3.
(EPS)

S1 Table. The top 15 journals in the four different corpora.
(DOCX)

S2 Table. True positive rate at 10% false positive rate (TPR@10%FPR) for the four different corpora, with and without document weight for scoring co-occurrences.
(DOCX)

Author Contributions

Conceptualization: Søren Brunak.

Data curation: David Westergaard.

Formal analysis: David Westergaard, Lars Juhl Jensen, Søren Brunak.

Funding acquisition: Søren Brunak.

Investigation: Søren Brunak.

Methodology: David Westergaard, Lars Juhl Jensen, Søren Brunak.

Project administration: Søren Brunak.

Resources: Hans-Henrik Stærfeldt, Christian Tønsberg.

Software: David Westergaard, Lars Juhl Jensen.

Supervision: Lars Juhl Jensen, Søren Brunak.

Visualization: David Westergaard.

Writing – original draft: David Westergaard, Lars Juhl Jensen, Søren Brunak.

Writing – review & editing: David Westergaard, Lars Juhl Jensen, Søren Brunak.

References

1. Azevedo A. Integration of Data Mining in Business Intelligence Systems. 1st Editio. Azevedo A, Santos MF, editors. Integration of Data Mining in Business Intelligence Systems. IGI Publishing Hershey, PA, USA; 2014. 314 p.
2. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. Vol. 6, Genome biology. 2005. 6(7):224. https://doi.org/10.1186/gb-2005-6-7-224 PMID: 15998455
3. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. Vol. 74, Methods. 2015. 74:97–106. https://doi.org/10.1016/j.ymeth.2015.01.015 PMID: 25641519
4. Luo Y, Riedlinger G, Szolovits P. Text Mining in Cancer Gene and Pathway Prioritization. Vol. 13, Cancer Informatics. 2014. 13(Suppl 1):69–79. https://doi.org/10.4137/CIN.S13874 PMID: 25392685
5. Ananiadou S, Thompson P, Nawaz R, McNaught J, Kell DB. Event-based text mining for biology and functional genomics. Vol. 14, Briefings in functional genomics. 2015. 14(3):213–30. https://doi.org/10.1093/bfgp/elu015 PMID: 24907365
6. Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A. Text mining for metabolic pathways, signaling cascades, and protein networks. Vol. 283/pe21, Sci. STKE. 2005. 283/pe21:e21.
7. Liu F, Chen J, Jagannatha A, Yu H. Learning for Biomedical Information Extraction: Methodological Review of Recent Advances. arXiv:1606.07993 [cs]. 2016.
8. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Vol. 9 Suppl 2, Genome biology. 2008. 9 Suppl 2(Suppl 2):S8.
9. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. Vol. 17, Briefings in Bioinformatics. 2016. 17 (1):33–42. https://doi.org/10.1093/bib/bbv087 PMID: 26420781

10. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Vol. 13, Nature Reviews Genetics. 2012. 13(12):829–39. https://doi.org/10.1038/nrg3337 PMID: 23150036

11. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Vol. 13, Nature Reviews Genetics. 2012. 13(6):395–405. https://doi.org/10.1038/nrg3208 PMID: 22549152

12. Rodriguez-Esteban R, Bundschus M. Text mining patents for biomedical knowledge. Vol. 21, Drug Discovery Today. 2016. 21(6):997–1002. https://doi.org/10.1016/j.drudis.2016.05.002 PMID: 27179985

13. Simmons M, Singhal A, Lu Z. Text mining for precision medicine: Bringing structure to ehrs and biomedical literature to understand genes and health. In: Vol. 939, Advances in Experimental Medicine and Biology. Springer Singapore; 2016. p. 139–66. https://doi.org/10.1007/978-981-10-1503-8_7 PMID: 27807747

14. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Vol. 7, Nature reviews. Genetics. 2006. 7(2):119–29. https://doi.org/10.1038/nrg1768 PMID: 16418747

15. Winnenburg R, Wächter T, Plake C, Doms A, Schroeder M. Facts from text: Can text mining help to scale-up high-quality manual curation of gene products with ontologies? Vol. 9, Briefings in Bioinformatics. 2008. 9(6):466–78. https://doi.org/10.1093/bib/bbn043 PMID: 19060303

16. Wei C-H, Kao H-Y, Lu Z. Text mining tools for assisting literature curation. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics—BCB '14 [Internet]. New York, New York, USA: ACM Press; 2014. p. 590–1.

17. Cohen KB, Johnson HL, Verspoor K, Roeder C, Hunter LE, Verspoor K, et al. The structural and content aspects of abstracts versus bodies of full text journal articles are different. Vol. 11, BMC Bioinformatics. 2010. 11(1):492.

18. Papanikolaou N, Pavlopoulos GA, Theodosiou T, Iliopoulos I. Protein-protein interaction predictions using text mining methods. Vol. 74, Methods. 2015. 74:47–53. https://doi.org/10.1016/j.ymeth.2014.10.026 PMID: 25448298

19. Samuel J, Yuan X, Yuan X, Walton B. Mining online full-text literature for novel protein interaction discovery. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2010 [Internet]. IEEE; 2010. p. 277–82.

20. Garten Y, Altman R. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. Vol. 10, BMC bioinformatics. 2009. 10(Suppl 2):S6.

21. Müller HM, Kenny EE, Sternberg PW. Textpresso: An ontology-based information retrieval and extraction system for biological literature. Vol. 2, PLoS Biology. 2004. 2(11):e309. https://doi.org/10.1371/journal.pbio.0020309 PMID: 15383839

22. Martin EPG, Bremer EG, Guerin M-C, DeSesa C, Jouve O. Analysis of protein/protein interactions through biomedical literature: Text mining of abstracts vs. text mining of full text articles. In: Vol. 3303, Knowledge Exploration in Life Science Informatics. Springer, Berlin, Heidelberg; 2004. p. 96–108.

23. Corney DPA, Buxton BF, Langdon WB, Jones DT. BioRAT: Extracting biological information from full-length papers. Vol. 20, Bioinformatics. 2004. 20(17):3206–13. https://doi.org/10.1093/bioinformatics/bth386 PMID: 15231534

24. Blake C. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. Vol. 43, Journal of Biomedical Informatics. 2010. 43(2):173–89. https://doi.org/10.1016/j.jbi.2009.11.001 PMID: 19900574

25. Constantin A, Pettifer S, Voronkov A. Pdfx. Proceedings of the 2013 ACM symposium on Document engineering—DocEng '13. 2013.:177.

26. Luong M-T, Nguyen TD, Kan M-Y. Logical Structure Recovery in Scholarly Articles with Rich Document Features. Vol. 1, International Journal of Digital Library Systems. 2012. 1(4):1–23.

27. Ramakrishnan C, Patnia A, Hovy E, Burns GAPC. Layout-aware text extraction from full-text PDF of scientific articles. Vol. 7, Source Code for Biology and Medicine. 2012. 7(1):7. https://doi.org/10.1186/1751-0473-7-7 PMID: 22640904

28. Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, et al. Text mining the history of medicine. Rocha LM, editor. Vol. 11, PLoS ONE. 2016. 11(1):e0144717. https://doi.org/10.1371/journal.pone.0144717 PMID: 26734936

29. Lopresti D. Optical character recognition errors and their effects on natural language processing. Vol. 12, International Journal on Document Analysis and Recognition. 2009. 12(3):141–51.

30. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Vol. 43, Nucleic Acids Research. 2015. 43(D1):D447–52.

31. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease-gene associations. Vol. 74, Methods. 2015. 74:83–9. https://doi.org/10.1016/j.ymeth.2014.11.020 PMID: 25484339

32. Santos A, Tsafou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, Jensen LJ. Comprehensive comparison of large-scale tissue expression datasets. Vol. 3, PeerJ. 2015. 3:e1054. https://doi.org/10.7717/peerj.1054 PMID: 26157623

33. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, et al. COMPARTMENTS: Unification and visualization of protein subcellular localization evidence. Vol. 2014, Database. 2014. 2014(0):bau012–bau012. https://doi.org/10.1093/database/bau012 PMID: 24573882

34. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al. Disease ontology: A backbone for disease semantic integration. Vol. 40, Nucleic Acids Research. 2012. 40(D1):D940–6.

35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for The Unification of Biology. Vol. 25, Nature Genetics. 2000. 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651

36. Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. Vol. 44, Nucleic Acids Research. 2016. 44 (D1):D380–4. https://doi.org/10.1093/nar/gkv1277 PMID: 26590256

37. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The BRENDA Tissue Ontology (BTO): The first all-integrating ontology of all organisms for enzyme sources. Vol. 39, Nucleic Acids Research. 2011. 39(SUPPL. 1):D507–13.

38. Smith CL, Eppig JT. The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. Vol. 1, Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2009. 1(3):390–9. https://doi.org/10.1002/wsbm.44 PMID: 20052305

39. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Vol. 41, Nucleic Acids Research. 2013. 41(D1):D808–15.

40. Mørk S, Pletscher-Frankild S, Caro AP, Gorodkin J, Jensen LJ. Protein-driven inference of miRNA-disease associations. Vol. 30, Bioinformatics. 2014. 30(3):392–7. https://doi.org/10.1093/bioinformatics/btt677 PMID: 24273243

41. Kanehisa M, Goto S. Kyoto Encyclopedia of Genes and Genomes. Vol. 28, Nucleic Acids Research. 2000. 28(1):27–30. PMID: 10592173

42. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Vol. 45, Nucleic Acids Research. 2017. 45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662

43. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Vol. 44, Nucleic Acids Research. 2016. 44(D1):D457–62. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

44. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, et al. UniProt: A hub for protein information. Vol. 43, Nucleic Acids Research. 2015. 43(D1):D204–12.

45. Adams J. Collaborations: The rise of research networks. Vol. 490, Nature. 2012. 490(7420):335–6. https://doi.org/10.1038/490335a PMID: 23075965

46. Eckhouse S, Lewison G, Sullivan R. Trends in the global funding and activity of cancer research. Vol. 2, Molecular Oncology. 2008. 2(1):20–32. https://doi.org/10.1016/j.molonc.2008.03.007 PMID: 19383326

47. Plaven-Sigray P, Matheson GJ, Schiffler BC, Thompson WH. The Readability Of Scientific Texts Is Decreasing Over Time. bioRxiv. 2017.:119370.

48. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Vol. 409, Nature (London). 2001. 409(6822):860–921.

49. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. Lewitter F, Kann M, editors. Vol. 8, PLoS Computational Biology. 2012. 8(12):e1002822. https://doi.org/10.1371/journal.pcbi.1002822 PMID: 23300413

50. Mallory EK, Zhang C, Ré C, Altman RB. Large-scale extraction of gene interactions from full-text literature using DeepDive. Vol. 32, Bioinformatics. 2015. 32(1):106–13. https://doi.org/10.1093/bioinformatics/btv476 PMID: 26338771

51. Fluck J, Hofmann-Apitius M. Text mining for systems biology. Vol. 19, Drug Discovery Today. 2014. 19 (2):140–4. https://doi.org/10.1016/j.drudis.2013.09.012 PMID: 24070668

52. Zhu Y, Yan E, Wang F. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. Vol. 17, BMC Medical Informatics and Decision Making. 2017. 17(1):95. https://doi.org/10.1186/s12911-017-0498-1 PMID: 28673289