Published in final edited form as:

Expert Rev Proteomics. 2019 ; 16(11-12): 871–873. doi:10.1080/14789450.2019.1703678.

Working the literature harder: what can text mining and bibliometric analysis reveal?

Yu Han[1,2], **Sara Wennersten**[1,2], **Maggie P. Y. Lam**[1,2,3]

[1]Department of Medicine, Division of Cardiology, University of Colorado Denver, United States

[2]Consortium for Fibrosis Research and Translation, University of Colorado Denver, United States

[3]Department of Biochemistry and Molecular Genetics, University of Colorado Denver, United States

Keywords

bibliometrics; human disease; molecular interaction; protein function; text-mining

1. Introduction

Text-mining and bibliometrics provide tools to extract and organize large volumes of information from the scientific literature across fields. Text-mining and bibliometrics provide tools to extract and organize large volumes of information from the scientific literature. Among applications particularly germane to proteomics researchers are the prediction of protein-protein interaction and protein function. Systematic analysis of literature data can further reveal hidden relationships among biological concepts from genes and proteins to drugs and diseases, whereas bibliometrics can also identify trends in research topics, collaboration networks, and resource allocation. With increasingly sophisticated methods and the availability of user-friendly web tools, large-scale literature analyses can add considerable value to research workflows across disciplines.

2. Main text

The scientific literature is growing rapidly. By the time an average reader finishes reading this editorial (~5 min), 13 new articles will have been added to PubMed (>1.3 million articles added in 2018). The need for systematic approaches to combat this information overload has fueled adoption of text-mining and bibliometrics methods. Generally speaking, text-mining refers to the automated extraction of structured information from a free-text

Correspondence: Maggie P. Y. Lam, Department of Medicine, Division of Cardiology, Anschutz Medical Campus, University of Colorado Denver, Aurora, Colorado 80045, United States, maggie.lam@cuanschutz.edu, Telephone: +1-303-724-3520.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

document, whereas bibliometrics concerns the statistical analysis of trends and patterns across related publications. In practice, the approaches and objectives of the two often overlap; both share the overarching goals of summarizing literature information at scale and identifying the relationships between concepts.

A frequent use of text-mining is to recognize the mention of gene and protein names from the text of published research articles, including those written in non-standard common names or synonyms—e.g., Akt, NCAM-180, cTnT, etc. By automating this “named entity recognition” task over a set of articles, a large database of protein-publication relationship can be built where each entry specifies a gene/protein that is mentioned in a specific publication — e.g., the gene RBM5, mentioned in the article PMID 28061901. Couple that to a PubMed query that returns a list of publications on a search topic, and one can then retrieve all proteins that are associated with the topic in the entire literature. This “guilt-by-association” strategy of inferring functional association from co-occurrence can be generalized to other concepts including phenotypes, drugs, and metabolites. Both European PMC and the NCBI PubTator offer the text-mining results of associated concepts from PubMed articles [1], and a growing number of user-friendly web tools now exist to help researchers analyze concept associations in the literature without requiring programming efforts (Table 1).

2.1 Prediction of molecular interactions from text-mining

What can these analyses reveal? One fruitful area is in the prediction of protein-protein interactions (PPI) which can help formulate hypotheses and guide experiments. As mentioned, text-mining can automatically recognize protein names in an abstract or full-text article. Hence if two proteins co-occur in the same publication, one might predict that they are associated, such as through direct interaction. Earlier work showed that co-occurrence reliably predicts PPI in gold-standard databases with good specificity, although understandably sensitivity is lower for PPI from large-scale experiments whose results are not detailed in articles [2]. The widely-used protein interaction database STRING v11 includes in its protein association score a text-mining component that looks for co-mentions of two proteins within a paragraph or an article [3]. Comparing the current STRING database (v11; released 2019-01-19) to a previous release (v10, 2016-04-16), over 2,000 PPI pairs that were once primarily supported by text-mining evidence in 2016 have now become corroborated by experimental evidence score (e.g., CCL2–CXCL13), demonstrating text-mining can reliably anticipate experimentally valid PPI pairs.

2.2 Inference of protein functions in health and disease

Literature analysis can also help annotate protein function by summarizing whether a protein is closely associated with certain diseases or pathways. To be able to quantitatively compare the importance of two proteins in a disease (e.g., is PDX1 more associated with diabetes than INSM2?), it is necessary to not only count the number of co-occurrences but also account for the specificity of association. In other words, is a protein statistically more likely to be associated with papers focusing on the disease of interest over other diseases? To achieve this, different bibliometrics algorithms have been applied to quantify the semantic similarity between two concepts in text-mined results, including normalized compression distance, term-frequency inverse-document-frequency, and other metrics [4–7]. Finding the list of most frequently mentioned proteins in publications related to a disease can not only help researchers formulate hypotheses in targeted disease studies, but also create gene lists for statistical overrepresentation or GSEA-type analyses commonly used to analyze large expression datasets [5,8].

Technically speaking, these associations only reveal which proteins are the most “popular” in a disease, but do not directly assess the strength of evidence in each study and whether it relates to bona fide biological significance. However, the logic behind literature analysis often assumes the “wisdom of the crowd” — with researchers acting as rational agents that invest their time and resources judiciously, and over time the research community should collectively expend most efforts toward truly significant proteins. Benchmarking the popularity lists against orthogonally curated annotations (e.g., GO or GWAS targets) suggests that they do often accurately predict functional significance. Recent developments further improve upon the quality of results by adjusting for the year and impact factor of associated publications [5,7] or by integrating text-mining and gene co-expression data to predict additional proteins that might be associated with a query term [4].

2.3 Revealing unknown relationships across studies

Perhaps most importantly, literature analysis could unveil hidden relationships between concepts that are not explicitly mentioned in any single publication but only coalesce when analyzing the total body of publications. This logic underlied the “Swanson ABC” method of literature discovery, first expounded in 1986 when Don Swanson noticed in multiple articles that patients with Raynaud’s syndrome had high blood viscosity, whereas other unrelated articles suggested that fish oil use reduces blood viscosity. Putting two and two together, Swanson hypothesized that fish oil may be used as a treatment for Raynaud’s syndrome, which was later validated. Generally, the model proposes that if two concepts A and B (e.g., a disease A and a protein B) are associated in the literature and concept B is in turn associated with concept C (e.g., the protein B with a compound C), then one might hypothesize an indirect association path between A and C even if this linkage is not explicitly mentioned in the literature.

This discovery model can be accelerated by text-mining and may become increasingly valuable for drug repurposing by joining together separate disease-protein (A–B) and protein-drug (B–C) relationships [9]. It can also uncover unexpected commonalities in disease mechanisms. As mentioned, literature analysis can associate a disease term (e.g., “hypertension”) with a list of proteins [5] or phenotype terms [10]. It is therefore possible to quantify how closely related two diseases are, based on how many associated proteins or phenotypes they share. Upon analyzing a collection of disease terms (e.g., Disease Ontology), a network of relationships between disorders, or “diseasome”, can be constructed to identify unexpected similarities across diseases. For example, network analysis of phenotype associations led to the hypothesis that some forms of spinal muscular atrophy may be more closely related to lysosomal storage disorders than previously anticipated [10].

2.4 Collaboration networks and resource allocations

Aside from guiding individual research questions, text-mining and bibliometrics can uncover global trends in research activities and inform resource allocation, in so-called meta-research or science of science studies. Extensive bibliometric analyses of the biomedical literature have revealed among other observations the rising prominence of team science, a trend that is however accompanied by inequitable credit allocation [11]. Collaborative networks and research "hot topic" nodes have also been analyzed among proteomics researchers in the American Society for Mass Spectrometry (ASMS) [12]. Analyzing NIH-funded studies, one bibliometrics study showed how both basic and applied research provides immense value to commercial patents [13], whereas another investigation ranked genes by total funding received and shed light on underfunded disease targets [4]. Finally, the Human Proteome Organization (HUPO) has adopted bibliometrics to identify highly published proteins across research fields, as a means to prioritize the development of proteomics assays for promising targets that are more likely to be adopted by domain researchers.

3. Concluding remarks

We highlighted here some emerging applications of large-scale literature analyses. It should come as no surprise that the methodologies of text-mining and bibliometrics are themselves active areas of research. Co-occurrence metric can be refined by rules, e.g., predicting PPI only if two proteins are co-mentioned in a paragraph that includes additional recognizable words such as "interacts" and "binds". More sophisticated methods in natural language processing and machine learning can extract richer relationships from subject-verb-object triplets in phrases and provide context to different types of associations (positive vs. negative regulation). Secondly, the increasing availability open-access full-text articles will substantially improve accuracy and timeliness for the mining of information not present in article abstracts [14], which will surely benefit all above-mentioned applications. With increasing sophistication of methods and availability of user-friendly tools (Table 1), text-mining and bibliometrics will be increasingly valuable for helping researchers formulate hypothesis and discover biology.

Funding

This paper was funded by the U.S. Department of Health and Human Services, National Institutes of Health, National Heart, Lung, and Blood Institute, grants: F32 HL149191, R00 HL127302; and University of Colorado Consortium for Fibrosis Research and Translation Pilot Grant.

References

Papers of special note have been highlighted as:

* of interest

** of considerable interest

1. Wei CH, Allot A, Leaman R, et al. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1): W587–W593. [PubMed: 31114887] * This paper describes PubTator central, an automated annotation tool for bioentities via text-mining in PubMed Abstracts and PMC full-text literature repositories. The tool allows users to construct,

visualize, and download annotations for genes, diseases, chemicals, mutations, species, and cell lines.

2. He M, Wang Y, Li W. PPI finder: a mining tool for human protein-protein interactions. PLoS ONE. 2009;4(2): e4554. [PubMed: 19234603]
3. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1): D607–D613. [PubMed: 30476243]
4. Lachmann A, Schilder BM, Wojciechowicz ML, et al. Geneshot: search engine for ranking genes from arbitrary text queries. Nucleic Acids Res. 2019;47(W1): W571–W577. [PubMed: 31114885]
5. Lau E, Venkatraman V, Thomas CT, et al. Identifying High-Priority Proteins Across the Human Diseasome Using Semantic Similarity. J Proteome Res. 2018;17(12):4267–4278. [PubMed: 30256117]
6. Lee S, Kim D, Lee K, et al. BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. PLoS ONE. 2016;11(1): e0164680. [PubMed: 27760149]
7. Yu KH, Lee TLM, Wang CS, et al. Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining. J Proteome Res. 2018;17(4):1383–1396. [PubMed: 29505266]
8. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1): D833–D839. [PubMed: 27924018]
9. Yang HT, Ju JH, Wong YT, et al. Literature-based discovery of new candidates for drug repurposing. Brief Bioinform. 2017;18(3):488–497. [PubMed: 27113728]
10. Hoehndorf R, Schofield PN, Gkoutos GV. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. Sci Rep. 2015; 5:10888. [PubMed: 26051359] * This paper shows the use of a text-mining approach to find similar phenotypes among over 6,000 complex diseases. The resulting network is useful for exploration of novel disease-signature genes and disease-disease similarities.
11. Fortunato S, Bergstrom CT, Börner K, et al. Science of science. Science. 2018;359. [PubMed: 29700239]
12. Palmblad M, van Eck NJ. Bibliometric analyses reveal patterns of collaboration between ASMS members. J Am Soc Mass Spectrom. 2018;29(3):447–54. [PubMed: 29305796]
13. Li D, Azoulay P, Sampat BN. The applied value of public investments in biomedical research. Science. 2017;356(6333):78–81. [PubMed: 28360137]
14. Westergaard D, Stærfeldt HH, Tønsberg C, et al. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018;14(2): e1005962. [PubMed: 29447159] * This paper measures the growth of the scientific literature and shows that full-text mining outperforms abstract mining in the accuracy of extracted protein-protein interaction, as evaluated by benchmark datasets.
15. Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res. 2015;43(W1): W535–W542. [PubMed: 25925572]

Table 1.

Text-mining and Bibliometrics Web Tools

Web Tool	Main Function(s)	Citation	Web Address
BEST	Searches for relevant biomedical entity mentions in PubMed articles including genes and compounds	[6]	http://best.korea.ac.kr/
DisGeNET	Retrieves gene-disease and disease-disease relationship from disease search terms	[8]	http://disgenet.org
GeneShot	Finds associated genes to a query topic, predicts additional associated genes from mining other data types	[4]	https://amp.pharm.mssm.edu/geneshot/
PolySearch	Given a disease, genes, pathways query, finds other associated entities and concepts in text-mined corpora	[15]	http://polysearch.ca
PubPular	Finds and ranks proteins associated with any topic search terms, analyzes gene lists with precompiled terms	[5]	http://pubpular.net
PubTator Central	Highlights text-mined biological entities and concepts in the results of PubMed queries	[1]	https://www.ncbi.nlm.nih.gov/research/pubtator/
PURPOSE	Finds and ranks proteins associated with any topic search terms including disease and research focus areas	[7]	http://rebrand.ly/proteinpurpose