Elsevier R&D Solutions
FOR PHARMA & LIFE SCIENCES

WHITE PAPER

Harnessing the Power of Content

Extracting value from scientific literature: the power of mining full-text articles for pathway analysis

EXECUTIVE SUMMARY

Biological researchers face an overwhelming and constantly growing body of scientific literature that they must search through in order to keep up with developments in their field. Staying up-to-date with new findings is critical, yet seemingly impossible. With more than 1 million new citations in PubMed each year, how can anyone possibly read through all of the relevant full-text articles to uncover important data? For years manual review and curation of scientific publications has been the gold standard, with computer-based solutions ranking far behind in terms of accuracy and completeness. Fortunately times have changed; researchers now have ready access to significant computational power, and the capabilities of text mining software have improved dramatically. Elsevier's proprietary text-mining technology has proven that well-designed software applications can effectively "read" full-text articles, extracting far more relevant biological information than is found in article abstracts alone, with accuracies comparable to human researchers, and with much higher throughput. This technology helps ensure researchers are not missing out on valuable insights from the literature, giving them a better understanding of the biology behind specific pathologies and drug responses, supporting the target discovery process, and offering the competitive advantage that they need to succeed in their work.

Overall, the current version of Elsevier's text-mining technology has a 98% accuracy rate for entity detection and an 88% accuracy rate for relationship extraction-figures that are essentially identical to high-quality annotation by human experts.

Understanding Biology in the Era of Big Data

Biological research today can be summarized in one word – data. Massive quantities of information are generated daily from large and small laboratories, individuals, universities, and commercial organizations. Their findings, along with those of other individuals or groups, are published in multiple papers, across a wide range of scientific journals, over a period of years. On average, each year more than a million new records are added to PubMed – a service of the National Library of Medicine that indexes more than 23 million references comprising the MEDLINE bibliographic database at present. New findings are published, and yet it seems that each new result engenders more questions. The information needed to understand experimental results, describe fully an individual cellular pathway, or identify the interactions of multiple regulatory networks is scattered throughout dozens if not hundreds or thousands of articles and publications. The volume and complexity of the data have outstripped researchers' ability to keep up with the progress of science, even in their chosen fields. Researchers are left with a question that's impossible to answer – "What have I missed?"

What if you had a tool to dramatically reduce your odds of missing critically important research results? One way to speed up this process is to create a database of scientific knowledge (a "knowledgebase") from published literature that can then be used to quickly find and summarize information related to a particular disease or phenotype, as an alternative to reading hundreds of scientific articles. Biological knowledgebases can also be used to build detailed models of complex diseases to help uncover the underlying cause of a disease at the level of gene and protein networks, thereby helping to guide early target discovery, assessment, and prioritization.

Elsevier's proprietary text-mining technology can help researchers find their way in this ocean of data. This technology extracts facts from abstracts and full-text scientific publications from Elsevier and other trusted publishers, as well as from PubMed abstracts.

Although this technology is applicable to information all along the drug discovery and development pipeline, the first application has been to create the knowledgebase underlying Pathway Studio, an integrated data mining and visualization solution that allows investigators to search for facts on proteins, genes, diseases, drugs, and other entities; analyze experimental results and interpret data in a biological context; and build and visualize disease models representing the underlying biology of a particular condition or response. Pathway Studio provides direct access to underlying evidence extracted from the literature so that scientists can decide for themselves which results are relevant to their research.

Having access to the information derived from all those publications gives researchers a measurable advantage in rendering a complete picture of the genes and proteins involved in the biology of a disease or response to a drug.

The Challenge of Mining the Scientific Literature

Researchers have three main choices for building knowledgebases derived from the burgeoning literature: reading the articles (or abstracts) themselves, manual curation by experts, or using a specialized automated text-mining tool.

Researchers are used to reading dozens if not hundreds of scientific articles to understand their area of research. But reading a full paper takes time – even very experienced researchers can only read and fully comprehend a few articles an hour. One option to save time is to simply read the abstract of the paper. They're short, most are available on PubMed, and many can be read quickly. However, authors cannot include everything important in an abstract.

Overall, the process of deciding what goes into the abstract is often somewhat subjective. For example, when other researchers cite a particular paper, 20% of keywords they mention are not present in the abstract, which means that it didn't contain all the important information[1]. Multiple studies comparing the full text and abstract from the same paper concluded that less than half of the key facts from the body of a paper are present in its abstract[2–5]. Authors may choose to exclude from the abstract some technical or secondary information[5], or information that is less positive or less supportive of the main idea of the publication: a study examining randomized controlled trials published in high-profile journals showed that nearly half of papers that mentioned harm in the body of the paper did not report it in abstract[6]. Also, authors sometimes underestimate a finding's ultimate value; what may appear to be a relatively minor finding in one paper may lead to extensive research years later, resulting in hundreds of follow-up articles.

Based on this information, every researcher would agree that it is not enough to read just the abstract; one must read the full paper. But researchers' time is limited – how can they identify all the articles relevant to their research, much less gather all the critical information present in those articles?

Manual Curation is a Problematic Option

The alternative approach to personally reviewing the body of scientific literature is to rely on PhD researchers trained as curators. Traditionally, manual curation of papers has been viewed as the gold standard and assumed to be more accurate than automated systems. Although this was certainly true in the past, current text analysis technology is highly accurate, rivaling that of even experienced curators.

Even the best human curators are not perfectly accurate, nor are they completely consistent. Numerous scientific publications highlight errors and inconsistencies of human curators. For example, in one study using manual annotation of PubMed articles with Gene Ontology Annotation terms, only 39% of the terms assigned by three different curators were identical[7]. In another study[8], the average precision of annotating medical events in clinical narratives by three experts was reported to be 88%. Drews et al. reported 73% inter-annotator agreement of extracting clinical data from medical records[9]. And low inter-annotator agreement (precision among pairs of annotators ranged from 31% to 77%, and recall ranged from 49% to 71%) was also reported for phenotype annotation for the BioCreative workshop[10].

Text-mining Technology Provides a Better Solution

The ability to "machine-read" and "understand" published articles has been in existence for approximately 35 years, but only recently has the sophistication of those machine-reading applications aligned with the wide availability of sufficient computational power to make fully automated text extraction a viable alternative to reading and manually curating scientific literature. Elsevier has a proprietary natural language processing (NLP)-based technology for extracting structured information (e.g., gene and protein names, functions, interactions) from unstructured data (the text content of scientific articles). When applied to Elsevier's industry-largest collection of scientific literature, the value for researchers is obvious.

Automated text processing is a complex process that involves two critical phases: identifying entities of interest, and then identifying relationships between those entities. The process begins by putting structure around the collection of key biological concepts. Domain experts (PhD researchers) create highly-curated concept models called ontologies that consist of a set of named entities (e.g., proteins, small molecules, cellular processes, diseases). The text mining software is guided by these ontologies and a set of advanced matching algorithms to detect those concepts in the input text.

It then analyzes the sentence structure and "decomposes" each sentence into a set of Subject-Verb-Object triplets (e.g., insulin regulates glucose uptake). Using strict linguistic rules, the software analyzes each triplet to determine how these concepts are related to each other and extracts the relationship between the Subject and the Object, guided by a list of specific, categorized relationships between them (e.g., binding, protein modification, expression regulation, molecular transport).

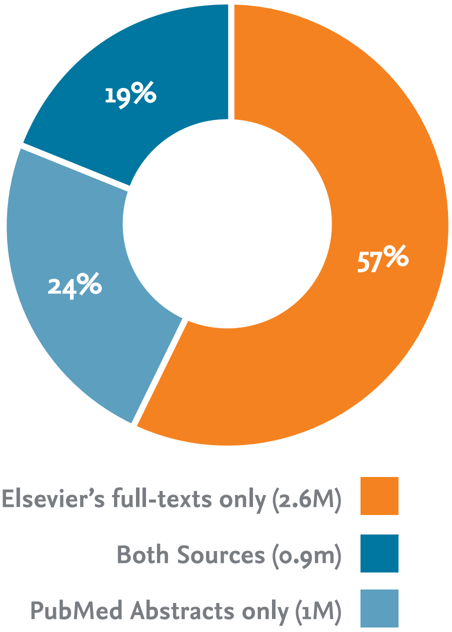
Overall, the current version of Elsevier's text-mining technology has a 98% accuracy rate for entity detection and an 88% accuracy rate for relationship extraction – figures that are essentially identical to high-quality annotation by human experts. Automated systems have other advantages over manual curation. They apply the same rules consistently to all articles over time, leading to greater reproducibility from article to article. And they can be easily updated with new terms and concepts that appear in the literature simply by changing the ontologies and linguistic rules, responding much more quickly than human annotators can.

A Wide Net is needed to Capture Relevant Results

The ability to get published, get a grant, or bring a new drug to market can all be derailed by coming in second in the race to discovery, in academia as well as in commercial laboratories. A delay of a single day bringing a new drug to market can cost $1 million or more in lost sales. How can you be sure you've cast a wide enough net to find relevant results as quickly as possible after they're published?

The solution is to look at as many articles, from as many journals as possible across your field of research. But with more than 1 million new research papers published each year, that's a massive amount of reading. A trained document curator can read and annotate at most 20-25 papers a day, but that translates into about 150 person-years to annotate 1 million papers. The ability to easily handle these massive and growing collections of data is what really tips the balance in favor of automated text mining applications. The current version of Elsevier's text-mining technology can process more than 23 million abstracts overnight, or between 80,000 and 100,000 full-text articles each hour on a regular PC. Users of the web version of Pathway Studio get weekly updates to the content in their knowledgebase, ensuring they are always up-to-date on the latest research findings.

Distribution of co-occurrences published in 2003-2012 by publication source.



More Facts are found in Full-Text Articles

To get a true measure of the advantage of full-text articles versus abstracts, we compared two corpora: 23 million PubMed abstracts and 2.5 million full-text articles from Elsevier's biological publications, and examined relationships that are especially important to researchers – facts linking proteins to diseases or clinical parameters, to biological functions, to small molecules, or to other proteins.

To mimic the human ability to infer connections from text, we first focused on instances where two entity terms occur in the same sentence. Linguistic rules were not applied, and our text-mining tool was used only to recognize pairs of concepts co-occurring within one sentence. Each pair was then annotated with four pieces of information: the year it was first published in either a full-text article or in an abstract, and the number of times it was published in full-text or abstract. The results (shown below) indicate that of co-occurrences first published between 2003 and 2012, 57% (2.6 million) were found only in the full-text articles; they did not appear in an abstract.

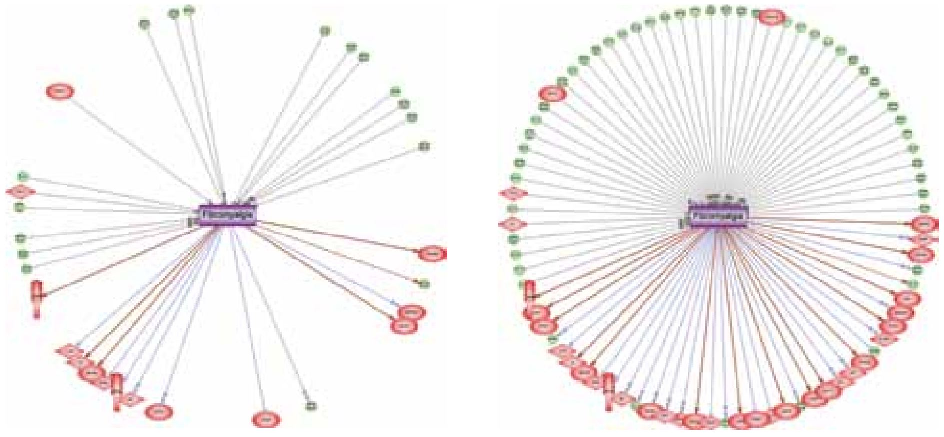
Type of co-occurrence	Estimated number of accurate facts in full-text articles
Protein – Disease or Clinical Parameter	112,000
Protein – Biological Process	156,000
Protein – Small Molecule	66,000
Protein – Protein	204,000

Sentence-level co-occurrence, by itself, does not necessarily mean that two concepts are related. Therefore, to estimate the number of accurate facts that are unique to full-text articles, we used PhD researchers to manually review approximately 1,200 random co-occurrences, in order to independently assess the proportion of true facts across different fact types and reference counts. We then extrapolated those rates of verified accurate co-occurrences identification versus false positives to the full corpus of PubMed abstracts. The estimated number of accurate facts found in the full-text articles but not present in PubMed abstracts is shown in the table (left).

The fact that full-text papers contain more connections between biological entities is reflected by the number of relations automatically extracted from full-texts compared to abstracts by Elsevier's text-mining tool. On average, this software extracts approximately 12 times more relationships from the full text of an article than from the abstract alone. Another important factor is that authors often restate the same facts and relationships multiple times in slightly different ways throughout the article, thereby increasing the likelihood that biological facts will be extracted and accurately interpreted by a text-mining engine if the full text is examined[3,11].

Another example of the advantage of using full-text articles rather than just abstracts is graphically demonstrated in the following example.

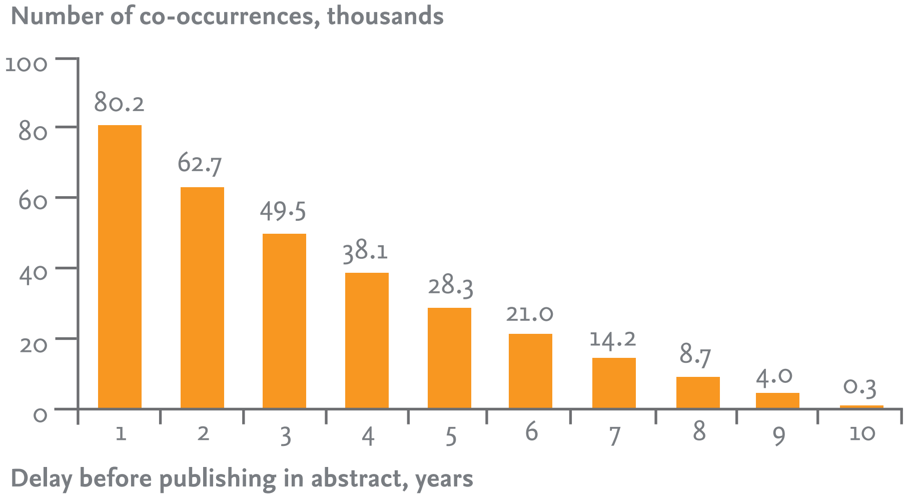
Using only abstracts, we found a total of 31 relationships in the literature for Fibromyalgia, but when full-text articles were added, an additional 53 relationships not mentioned in the abstracts were found and could be mapped to the disease. When a researcher is trying to understand the underlying biology of a disease or response to a drug, having a complete picture of the genes and proteins involved in that response and of the biological interactions between them is critical.



A. Abstracts only (31 relationships)

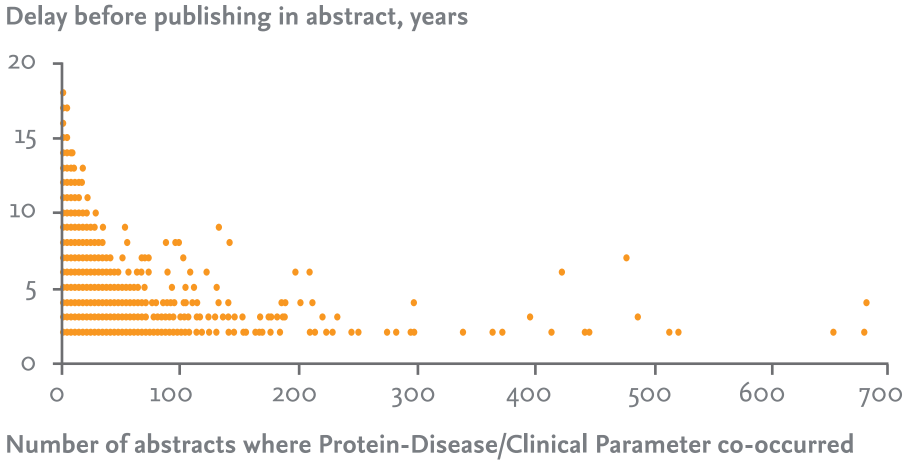
B. Abstracts plus full-text articles (84 relationships)

Figure legend: Facts about drugs and proteins affecting fibromyalgia progression were extracted from abstracts only (A) or from both abstracts and full-text (B). Each line represents at least one relationship between the disease in the center and a drug or protein on the outside.



Co-occurrences Often Mentioned in Full-text First

Almost every new discovery appears first in a single publication – although that research is often subsequently repeated and extended in multiple follow-on publications. The ability to find information rapidly after initial publication can convey a significant advantage in the competitive research arena. To attempt to quantify the advantage of reading full-text articles versus simply reading abstracts in terms of timely access, we examined the gap in time between the first mention of a co-occurrence in the full text of an article and the first time that same co-occurrence was first mentioned in an article abstract. Between 2003-2012, more than one third of co-occurrences appeared in the body of an article prior to being published in abstracts. Although in many cases the delay was relatively small (1-2 years), there are over 75,000 co-occurrences that were published in abstracts with a delay of 5 or more years (see the image left).

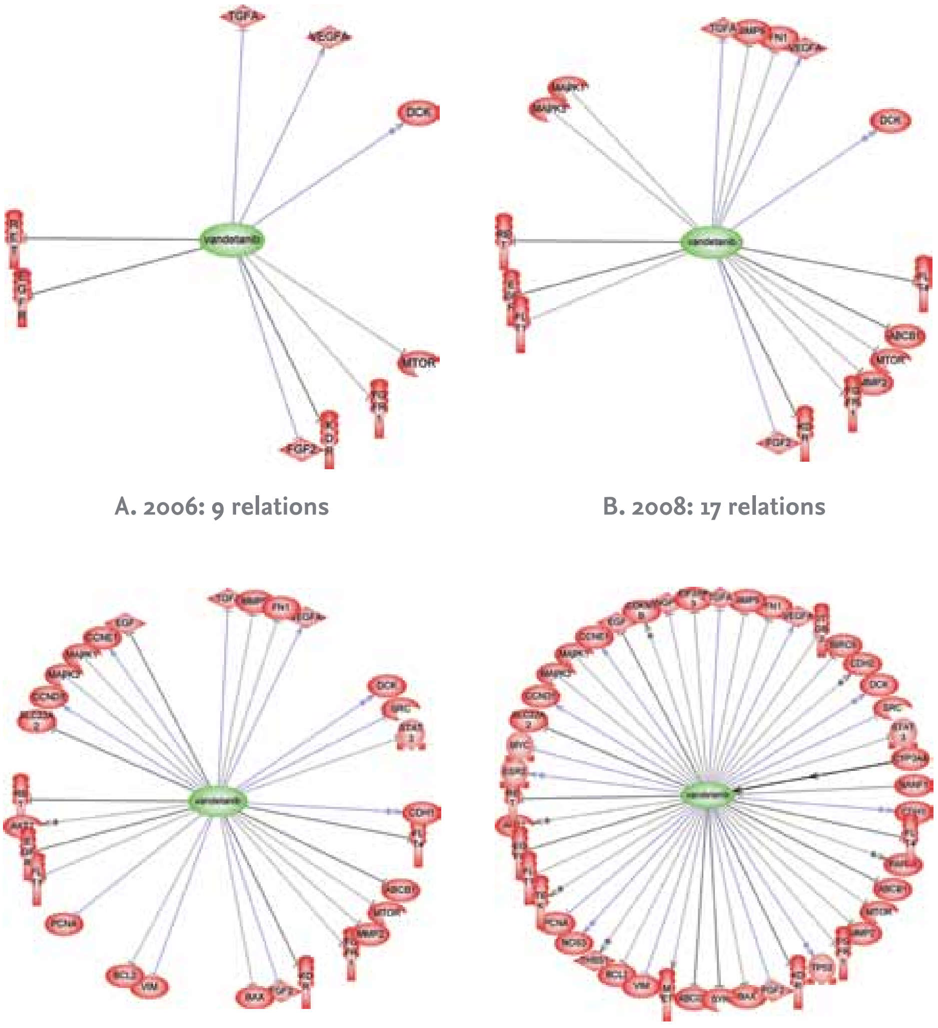


Example: Relationship between SIRT1 and senescence was mentioned in over 1,500 abstracts (data not shown) 4 years after first appearing in a full-text article

The next image illustrates the value of co-occurrences that appeared in the abstracts with a delay. If a finding was recognized as valuable, it will be mentioned in many abstracts. For example, association of SIRT1 and senescence was suggested in the body of a full-text publication 4 years earlier that it first appeared in an abstract. Since then it has been mentioned in over 1,500 abstracts, which demonstrates the great value of the finding.

A graphical example of how much new information accumulates over time is shown in the figure below. For Caprelsa (vandetanib) – a drug for late-stage (metastatic) medullary thyroid cancer in patients who are ineligible for surgery – we examined how many relationships were found in the full-text scientific literature over time. In this example, each 2-year interval resulted in approximately twice the number of identified relationships. Clearly, researchers gain demonstrable benefits from regular literature updates.

The content knowledgebase of the Pathway Studio Enterprise is updated quarterly, whereas the web version, in which all the information is managed in a secure cloud environment, allows for weekly updates to ensure researchers always have the most complete picture of their areas of interest.



A. 2006: 9 relations

B. 2008: 17 relations

C. 2010: 29 relations

D. 2012: 47 relations

Figure legend: Facts for proteins affected by drug Caprelsa were extracted from abstracts and full-text published up to year 2006 (A), 2008 (B), 2010 (C), and 2012 (D).

Conclusion

Biological research in the information age is highly competitive – whether you're in an academic, government, or a commercial setting. Staying current with the facts most relevant to your research can be very difficult, but understanding the complex biology behind a specific pathology or drug reaction is often critical to your research goals – so you need every advantage you can get. Pathway Studio can provide that crucial advantage. The combination of content from top-quality journals, from Elsevier and other respected publishers, coupled with our proprietary automated text-mining technology that can process and extract critical information from literally millions of full-text scientific articles and tens of millions of PubMed abstracts in a matter of hours, provides a compelling competitive advantage for your work. Add a collection of expertly-curated reference pathways along with sophisticated and flexible analytical and visualization tools that help you build and share complex models of pathways, processes, and diseases, and you can see why Elsevier's Pathway Studio is the obvious choice for researchers driven to succeed.

1. Divoli, A., Nakov, P. & Hearst, M. A. Do peers see more in a paper than its authors? Adv. Bioinforma. 2012, 750214 (2012).

2. Corney, D. P. A., Buxton, B. F., Langdon, W. B. & Jones, D. T. BioRAT: extracting biological information from full-length papers. Bioinforma. Oxf. Engl. 20, 3206–3213 (2004).

3. McIntosh, T. & Curran, J. R. Challenges for automatically extracting molecular interactions from full-text articles. BMC Bioinformatics 10, 311 (2009).

4. Schuemie, M. J. et al. Distribution of information in biomedical abstracts and full-text publications. Bioinforma. Oxf. Engl. 20, 2597–2604 (2004).

5. Shah, P. K., Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics 4, 20 (2003).

6. Bernal-Delgado, E. & Fisher, E. S. Abstracts in high profile journals often fail to report harm. BMC Med. Res. Methodol. 8, 14 (2008).

7. Camon, E. B. et al. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 6 Suppl 1, S17 (2005).

8. Raghavan, P., Fosler-Lussier, E. & Lai, A. M. Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise. AMIA Annu. Symp. Proc. AMIA Symp. AMIA Symp. 2012, 1366–1374 (2012).

9. Warner, J. L., Anick, P. & Drews, R. E. Physician inter-annotator agreement in the Quality Oncology Practice Initiative manual abstraction task. J. Oncol. Pract. Am. Soc. Clin. Oncol. 9, e96–102 (2013).

10. Arighi, C. N. et al. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database J. Biol. Databases Curation 2013, bas056 (2013).

11. Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinforma. Oxf. Engl. 17 Suppl 1, S74–82 (2001).