Text Mining Systems Biology: Turning the microscope back on the observer

July 3, 2018

Abstract

In this review, we describe the relationship between systems biology and text mining. On the one hand, text mining functions as a practical tool for systems biology research, which integrates diverse sub-fields and and takes an overall, systems-level view of biological phenomena. In this vein, various analyses have been done to recognize biological entities, construct networks of them (e.g. protein-protein interaction maps) and even find disease-associated genes directly from texts. On the other hand, text mining can also be applied to study systems biology itself, giving a "distant-reading perspective" on the evolution of the field. For example, by examining changes in the frequencies of terms in systems biology publications, we can analyze trends in research focus and in the popularity of systems approaches in various subdomains (e.g., in cancer research). Given these two uses of text mining for systems biology, we close with suggestions for adapting current publication formats for facilitating text mining and enabling its broader use.

Introduction

FIX

Systems biology has an interesting relationship with text mining. On the one hand, text mining functions as a practical tool for the field and has been employed in a number of contexts (eg protein protein interactions). As systems biology tends to have a large-scale perspective, the field embraces the "distant-reading" approach provided by text mining. On the other hand, we can turn the tool itself onto its observer and use text mining to study system biology as a field. In particular, compared to other domains in biological sciences, systems biology is a relatively new field, uniting a number of disparate sub fields in biology and systems theory, such as large scale analysis of networks, pathways and proteomes [20, 19, 31] . Insofar as this is the case, the umbrella term systems biology is quite important to the sub-fields being united under it. Thus, It is meaningful to ask to what degree it is functioning in a cohesive sense. Furthermore, it is interesting to ask whether it is a term rising in popularity or being supplanted by other new fields in todayâs biological science such as genomics and quantitative biology.

Here we survey both these aspects of systems biology and text mining - first as a tool for the field and then as way of understand the evolution of the field itself.

Text Mining as a tool in Systems Biology

MAKE

The current usage of the term text mining only dates to 1999 (though there are uses of text mining approaches before the with early applications dating back to 1980s).[16, 15]Since 2000, text mining has risen to become a popular jargon word. As a method for systematic information retrieval from large amount of publications, text mining is naturally a handy tool for systems biologists. In fact, biomedical text was one of the earliest applications of text data mining.[2, 33, 8]. In regard to systems biology, this approach has been applied in a few interesting scenarios in "omics" studies, such as mining the frequently studied genes in human genome[7] and analysis of the human phenome[29]. Although these studies demonstrated the potential of text mining and showed some interesting observations, they did not directly address traditional biological questions, such as identifying drug candidates or disease-related genes, building protein-protein interaction (PPI) network or help testing hypothesis. Here we try to systematically compare traditional literature screening and automated text mining at each level of the literature search, summarize what text mining has achieved and envision what text mining can do to help address biological questions in the future .

Figure 2 shows the comparison between traditional reading and text mining in the full cycle of scientific discoveries. If we look back for only a few decades, research articles were still in paper copies or camera-ready form, which required manual curation to tag the subjects and topics. As a result, it is impossible for researchers to perform automated literature screening for researchers. In the past two decades, electronic collections of research articles like PubMed(1996) and research article search engines like Google Scholar(2004) have made the searching for relevant publications much easier. Moreover, the search result is much less dependent on the keywords that the authors provided or quality of curation[4]. Rather, the result is automated by semantic matching in the scope defined by the user. Databases and servers are able to integrate large amount of biomedical publication.[30, 1, 9]

TOOLS NOT CORPORA

The next step for literature searches is reading and screening publications with a focus on specific topic or subject to form a picture and assumptions. With manual reading, this depends on the prior knowledge of the reader, introduces bias when the reader decides which article is relevant and requires human time and effort. It can now be automated at many levels by retrieving information from word frequency to topic modeling in either static or dynamic fashion. Popular tools such as Term Frequency-Inverse Document Frequency, Latent Dirichlet Allocation (LDA)[5], and dynamic LDA have been used to model large corpora such as Twitter, news, as well as scientific publications. [24, 13]. Although the application case for modeling research articles examined data from a rather broad scope and therefore requires less biological domain knowledge. Scientific symbols and vocabulary have made the task of customizing the tools for a specific research area more challenging.[22] To address this problem, software tools for named entity recognition (NER) in biological context have been developed.[28, 11, 23, 34, 35] Most of these tools are based on machine-learning methods and either learn the feature of biological terms of identifies terms from the context.

The most challenging step is to answer biological question from the processed text. A classic case is constructing Protein-Protein Interaction (PPI) network, other molecular interactions network from text mining. Several successful tools have been developed to aid in this process, many through BioCreative challenges.[26, 32, 3, 21, 14] Recent works also tried to find genes associated with diseases like breast cancer and find drug targets and novel application of existing drugs through text mining. [18, 10] These types of automated text searches bear some common traits: 1) the task requires large amount of lab result reports. 2) the results reported are easy to interpret from straightforward semantic relations such as word distance, concurrence, and connecting verbs without human judgment. As a result, those tasks requires less specific biological knowledge but more raw data in the form of larger textual corpora. This feature makes them good target for machine reading to reduce human effort. Although this is already a great efficiency improvement from manual screening from an efficiency perspective, we can imagine that more complicated biological relations can be automatically mined from text with customized tools. To facilitate this advancements, the language used to report biological experiment events need to be more standardized for the machines to read. For example, set default choices of verbs for specific types of interactions, simpler sentence structure, even standard forms for graphs and plots would make machine screening much easier.

LATER

The ultimate goal for automated text mining is to draw conclusion from the texts, generate new ideas, and even propose hypothesis from existing literature. This is yet to happen in the the biological text mining world, but can be made possible when combined with other machine learning tools. For instance, network analysis tools have enabled imputation of missing nodes[17] and evaluation of distance and centrality of [27] for social or biological network analysis. When combined with text mining results of molecule interaction network, automated pipelines that infer interactions between molecules that haven't been discovered or hard to crystallize experimentally. Furthermore, drug search for cancers can also be expedited with text mining on both the pathway and molecules related to the condition, and the drug molecules that interacts with those molecules.

Text mining reveals trends within systems biology

LOGIC

Overall, the number of research publications has been growing exponentially. More than one million manuscripts were published in year of 2017 alone. [1] As a result, it is challenging for researchers to conduct extensive literature searches on any topic or subject. Among all the areas in biological sciences, systems biology may be the most challenging to conduct thorough literature screening. Different from other fields in biology that focus on specific subjects, systems biology integrates knowledge from various fields by studying their interactions and relations. Therefore literature screening can give a broad picture of what system biology encompasses. Figure 1a shows a wordcloud from vocabulary in abstract of articles with "systems biology" in the title. The subjects studied include cells, genes, and proteins and popular methods include both experiments and computations, such as pathway analysis, network analyses. Figure 1b shows the number of annual publications in some of the subjects normalized by total amount of articles annual since the 1990s. Although the amount of publications are growing exponentially overall, the fields within systems biology family are not growing in all the same speed. The number of genes and cells related articles are growing with higher speed than more molecules and pathways. If the current trend continues, the exponential change amplify the difference over time.

Furthermore, probing the dynamics of relative word frequencies can reveal the trend change of focus and approaches within systems biology. Figure 2 shows normalized words frequency from abstracts of schizophrenia and autism publications from 1990s to 2016. The graphs show, in a sense, that systems biology is often a subtext on particular disease studies. For instance, one can study the systems biology related to cancer, the systems biology related to mental disease, etc. Here in these graphs, we see how one subtext is waxing and waning relative to another. We can see, for instance, the forever climbing usage of genes and mutations and a decreasing emphasis on whole-organism analysis. More generally, this approach allows us to visualize how the common view and research focus of the diseases evolved over time. The research nowadays focus more on the genetic and molecular level previous research that focused on the behavioral level. Thus, researchers can leverage text mining to appropriately make hypotheses on complex issues like diseases. Meanwhile, we have to be aware that the trend in research topics are not only a result of advances in our scientific understanding, but also affected by the available techniques and contemporary beliefs.

LOGIC

Conclusions

REW INV.

Here in this review, we have attempted to study the interrelationship between systems biology and text mining. On one hand, using text mining to look at the waxing and waning itself, of the term of systems biology, and on the other hand, to look at the way text mining functions as an integral tool within the large discipline of systems biology. Future research will harness the power of text mining to save researchers time from tedious literature screening. However, there are a few challenges specific to literature mining of biological publications. One is the biological and chemical terms that not only increase the vocabulary space of the models but also confuse the algorithms by introducing unusual synonyms, stop words and add complexity to the experiment description even when most other words are the same. Moreover, the way that biological finding are reported can vary a lot, considering the scientific training background and native languages of the researchers. Left alone the intrinsic logic of the research articles.

To make the text mining of biological literature easier to automate, there are a few approaches we can take. One is to accompany each new scientific report with a standard result summary in a easy to mine format, such as Structured Digital Abstract.[6, 25, 12] This has already been implemented in a few domains that structured reporting of the results is straightforward, such as PDB for protein structures and KEGG for pathways and genes. A step forward is easy to realize that also collects protocols, X-ray or cryoEM images for experimental reports and even choices of algorithms, hardware, and platform for computational report. Another is to scrutinize the new publications upon publishing using a standard topic modeling robot to systematically classify, extract keyword, and curate the results for others' reference. This also puts a requirement on the writing style of the researchers. Some may argue that this would hurt the diversity of and freedom of scientific writing but is aligned with the purpose of research publications, especially given that most contributors are writing in their second language. To help the automated processing of the articles, a clear structure requirement for the articles by each journal could be imposed.

MORE

Figure Legends

Figure 1 Fig 1a: Comparison of steps taken in a traditional literature search and text mining. Fig 1b: Text mining pipeline and tools.

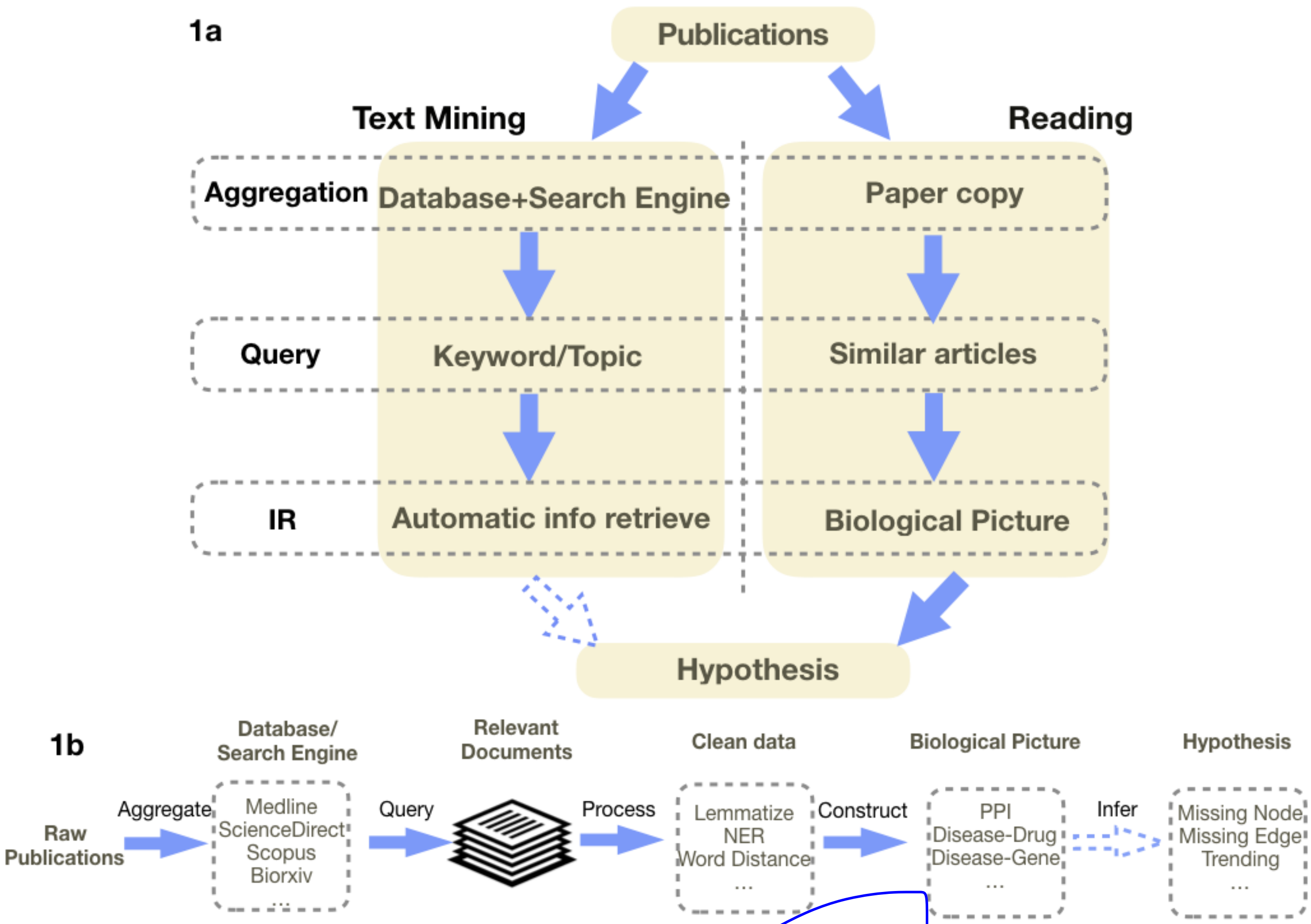


Figure 2 Fig 2a, 2b: Number of publications and normalized word frequency from abstract of publications regarding autism(a) and schizophrenia(b). Here blue and light blue dots represent gene related terms, and red dot represent whole-organism related terms.

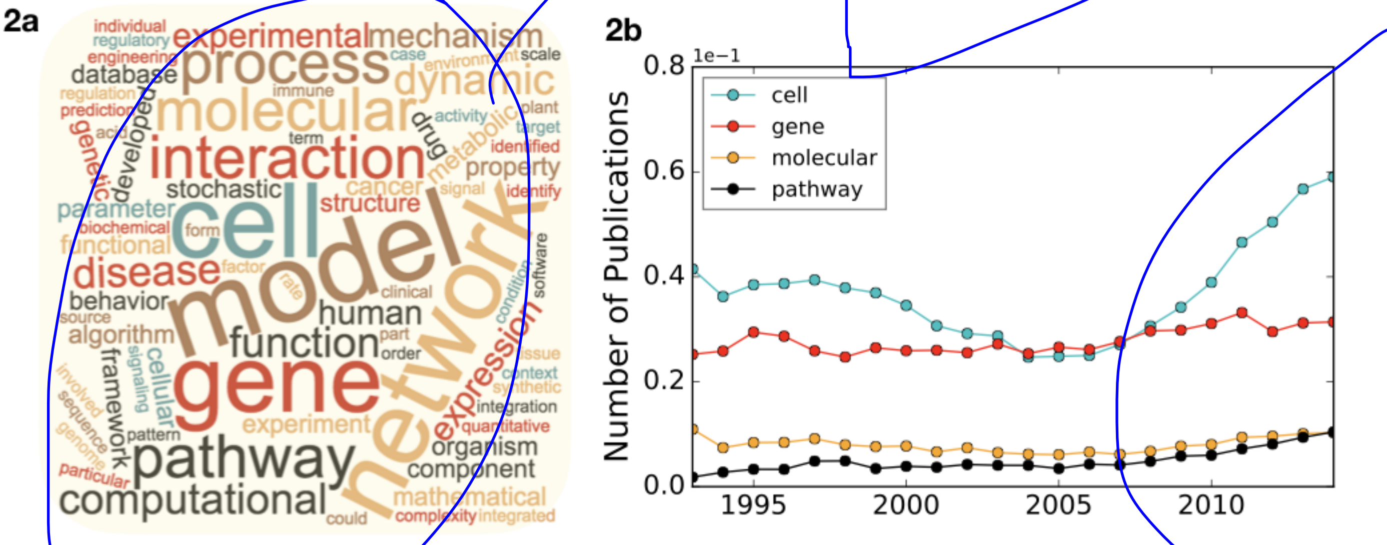
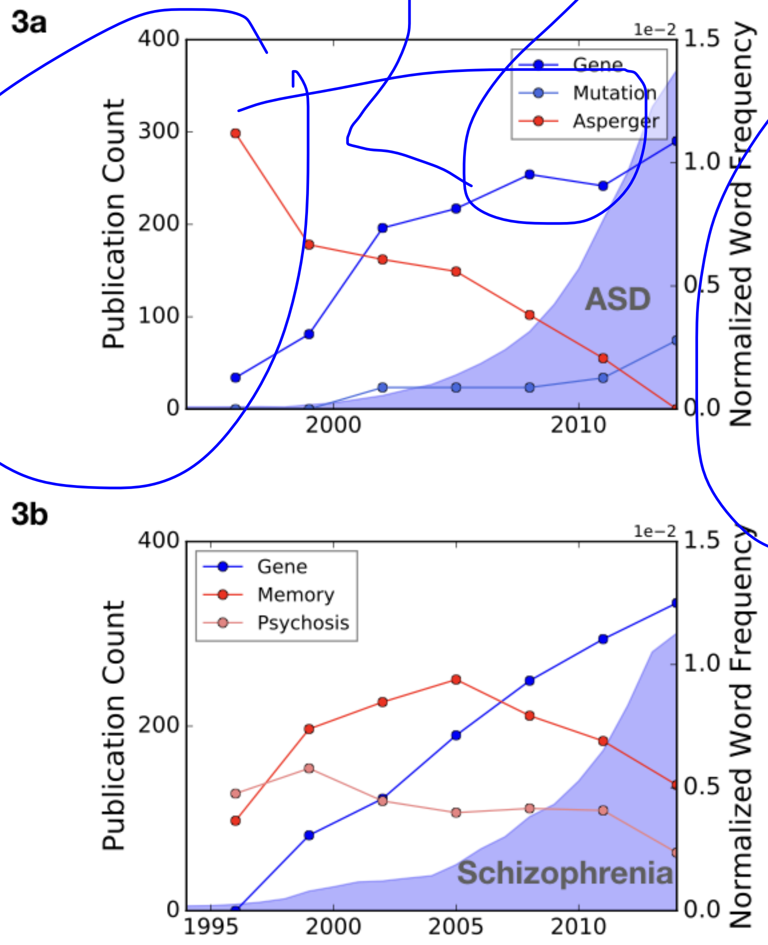


Figure 3 Fig 3a: Wordcloud from abstracts of publications with title containing "systems biology". Fig 3b: Number of publications annually in some subdomains of systems biology, normalized by total number of publications annually.



References

[1] A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan, and Z. Lu. Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc. *Nucleic Acids Research*, page gky355, 2018.

[2] S. Ananiadou and J. McNaught. *Text mining for biology and biomedicine.* Citeseer, 2006.

[3] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381 – 390, 2010.

[4] W. A. Baumgartner Jr, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, and L. Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–i48, 2007.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] K.-H. Cheung, M. Samwald, R. K. Auerbach, and M. B. Gerstein. Structured digital tables on the semantic web: toward a structured digital literature. *Molecular Systems Biology*, 6(1), 2010.

[7] E. Dolgin. The greatest hits of the human genome: A tour through the most studied genes in biology reveals some surprises. 551:427–431, 11 2017.

[8] R. A. Erhardt, R. Schneider, and C. Blaschke. Status of text-mining techniques applied to biomedical text. *Drug discovery today*, 11(7-8):315–325, 2006.

[9] W. W. Fleuren and W. Alkema. Application of text mining in the biomedical domain. *Methods*, 74:97 – 106, 2015. Text mining of biomedical literature.

[10] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, and W. Alkema. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLOS Computational Biology*, 6(9):1–11, 09 2010.

[11] M. Gerner, F. Sarafraz, C. M. Bergman, and G. Nenadic. Biocontext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. In *Bioinformatics*, 2012.

[12] M. Gerstein, M. Seringhaus, and S. Fields. Structured digital abstract makes text mining easy. *Nature*, 447:142 EP –, 05 2007.

[13] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[14] M. He, Y. Wang, and W. Li. Ppi finder: A mining tool for human protein-protein interactions. *PLOS ONE*, 4(2):1–6, 02 2009.

[15] M. A. Hearst. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 3–10. Association for Computational Linguistics, 1999.

[16] J. R. Hobbs, D. E. Walker, and R. A. Amsler. Natural language access to structured text. In *Proceedings of the 9th Conference on Computational Linguistics - Volume 1*, COLING '82, pages 127–132, Czechoslovakia, 1982. Academia Praha.

[17] M. Huisman. Imputation of missing network data: some simple procedures. *Journal of Social Structure*, 10(1):1–29, 2009.

[18] K. Kawashima, W. Bai, and C. Quan. Text mining and pattern clustering for relation extraction of breast cancer and related genes. In *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 59–63, June 2017.

[19] H. Kitano. Computational systems biology. *Nature*, 420(6912):206, 2002.

[20] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[21] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology*, 9(2):S1, Sep 2008.

[22] M. Krallinger, A. Valencia, and L. Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology*, 9(2):S8, Sep 2008.

[23] R. Leaman and G. Gonzalez. Banner: An executable survey of advances in biomedical named entity recognition. 13:652–63, 02 2008.

[24] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, Sep 2016.

[25] S. Michael and G. Mark. Manually structured digital abstracts: A scaffold for automatic text mining. *FEBS Letters*, 582(8):1170–1170.

[26] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, and I. Iliopoulos. ProteinÃÂÂprotein interaction predictions using text mining methods. *Methods*, 74:47 – 53, 2015. Text mining of biomedical literature.

[27] Z. Shi and B. Zhang. Fast network centrality analysis using gpus. *BMC bioinformatics*, 12(1):149, 2011.

[28] R. T.-H. Tsai, C.-L. Sung, H. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu. Nerbio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. 7 Suppl 5:S11, 02 2006.

[29] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen. A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14:535–542, 2006.

[30] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak. Text mining of 15 million full-text scientific articles. *bioRxiv*, 2017.

[31] D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116, 2007.

[32] C. Wu, J.-M. Schwartz, G. Brabant, S.-L. Peng, and G. Nenadic. Constructing a molecular interaction network for thyroid cancer via large-scale text mining of gene and pathway events. *BMC Systems Biology*, 9(6):S5, Dec 2015.

[33] D. Zhang, S. Simoff, and J. Debenham. Text mining techniques. *E-Service Intelligence: Methodologies, Technologies and Applications*, 37:191, 2007.

[34] J. Zhang, D. Shen, G. Zhou, J. Su, and C.-L. Tan. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics*, 37(6):411 – 422, 2004. Named Entity Recognition in Biomedicine.

[35] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, 2004.