


Computational Analyses Reveal Fundamental Properties of the Hemophilia Literature in the Last 6 Decades

Tiago JS Lopes¹, Ricardo Rios² and Tatiane Nogueira²

¹Department of Regenerative Medicine, National Center for Child Health and Development Research Institute, Tokyo, Japan. ²Department of Computer Science, Federal University of Bahia, Salvador, Brazil.

Bioinformatics and Biology Insights
Volume 16: 1–7
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322221125604


ABSTRACT: Hemophilia is an inherited blood coagulation disorder caused by mutations on the coagulation factors VIII or IX genes. Although it is a relatively rare disease, the research community is actively working on this topic, producing almost 6000 manuscripts in the last 5 years. Given that the scientific literature is increasing so rapidly, even the most avid reader will find it difficult to follow it closely. In this study, we used sophisticated computational techniques to map the hemophilia literature of the last 60 years. We created a network structure to represent authorship collaborations, where the nodes are the researchers and 2 nodes are connected if they co-authored a manuscript. We accurately identified author clusters, namely, researchers who have collaborated systematically for several years, and used text mining techniques to automatically synthesize their research specialties. Overall, this study serves as a historical appreciation of the effort of thousands of hemophilia researchers and demonstrates that a computational framework is able to automatically identify collaboration networks and their research specialties. Importantly, we made all datasets and source code available for the community, and we anticipate that the methods introduced here will pave the way for the development of systems that generate compelling hypothesis based on patterns that are imperceptible to human researchers.

KEYWORDS: Coauthor network, text mining, knowledge discovery, hemophilia

RECEIVED: June 9, 2022. **ACCEPTED:** August 24, 2022.

TYPE: Original Research Article

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Council for Science, Technology, and Innovation (CSTI), Cross-ministerial Strategic Innovation Promotion Program (SIP), “Innovative AI Hospital System,” by the National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN), grant number SIPAIH20D01, JSPS KAKENHI (JP22K06119), and the National Center for Child Health and Development internal grant (2022B-2). This work was also supported by CAPES (Coordination for the Improvement of Higher Education Personnel, a Brazilian federal

government agency), FAPESP (São Paulo Research Foundation) under grant number 2013/07375-0, and by the Terumo Life Science Foundation, Japan.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: TJSL received consulting fees from Pola Chemical Industries, Japan, for projects unrelated to the present study, and speaker honoraria from Sanofi Japan. The other authors do not have any conflict of interests to declare.

CORRESPONDING AUTHOR: Tiago JS Lopes, Department of Regenerative Medicine, National Center for Child Health and Development Research Institute, 2-10-1 Okura, Setagaya-ku, Tokyo 157-8535, Japan. Email: tiago-jose@ncchd.go.jp

Introduction

Hemophilia is an X-linked inherited blood coagulation disorder, affecting approximately 1 in 5000 to 25 000 live male births.¹ It is caused by the presence of a defective copy of the coagulation factor VIII (hemophilia A) or the coagulation factor IX (hemophilia B). In turn, these defective genes synthesize a partially functional or nonfunctional protein, resulting in a missing component in the precisely orchestrated coagulation cascade.

Although it is a relatively rare disorder, research in hemophilia is intense. Several groups are working to uncover the fundamental aspects of coagulation factor biology,^{2,3} improve patient care,⁴ develop physiotherapy programs,⁵ improve therapeutics,⁶ and advance gene therapies.⁷ In all, hemophilia research encompasses all areas of biomedical research.

As in other fields, the main channels used by hemophilia researchers to communicate their findings are the peer-reviewed scientific journals in English. Due to the transition from printed to the electronic form, the number of scientific journals and articles increased dramatically in the last 2 to 3 decades. Thus, even considering only hemophilia research, it is already difficult for professionals to stay up-to-date with all the latest findings. Similar to other hematological disorders, this trend points to a near future where it will be unfeasible for humans to read all published studies.

In other areas, researchers started addressing this issue by creating automatic text summaries,⁸ classifying studies according to

its contents,⁹ recommending articles based on reading records,¹⁰ and notably, making automatic discoveries by connecting dispersed factual information.¹¹ For hemophilia research, in particular, these applications are still lacking.

In this study, we gave the first step in this direction by developing a computational framework that maps the knowledge accumulated in hemophilia research in the last 60 years. First, we created a network where the nodes represent the hemophilia researchers, and 2 nodes are connected if they co-authored a manuscript. In previous studies, coauthorship networks proved itself useful to reveal meaningful patterns of scientific collaboration,¹² as well as serve as a historical record for future generations.¹³

We used this hemophilia coauthorship network to automatically find groups of authors (ie, clusters), who have collaborated systematically for many years. We used this information as input for text mining algorithms and found that even with minimum processing, it is already possible to automatically identify the topics representing the essence of the work performed by each group.

Thus, the contribution of this study in the short term is that it helps researchers to visualize and identify potential competitor and collaborator groups, and in the long term, the computational methodology introduced here paves the way for the development of automatic knowledge curation and discovery systems that are tailor-made for hemophilia research.



Materials and Methods

The hemophilia literature corpus

We searched the PubMed database on February 4, 2022, using the following terms (“hemophilia B” [Title/Abstract] OR “hemophilia B” [Title/Abstract] OR “haemophilia A” [Title/Abstract] OR “hemophilia A” [Title/Abstract] OR “FVIII” [Title/Abstract] OR “factor VIII” [Title/Abstract] OR “factor IX” [Title/Abstract] OR “factor VIIIA” [Title/Abstract] OR “factor IXa” [Title/Abstract]) NOT (“von willebrand disease”). We considered only articles that had an English abstract available. In total, this returned ~20 600 abstracts.

We downloaded all abstracts in the Medline format and processed them by in-house scripts to extract the abstract text and the authors.

Extracting author names to build a network

We extracted the author names from each abstract record using Python scripts and the Biopython package.¹⁴ We considered only articles with more than one author. Next, we built an undirected graph where the nodes represent the authors and created an edge between 2 nodes if they co-authored a manuscript. The weight on each edge is the number of manuscripts co-authored by the 2 authors. Moreover, we considered symmetrical edges, meaning that A-B is the same as B-A. We pruned the complete network by leaving only authors with 2 or more publications related to hemophilia (Supplementary Table 1 has the complete network).

Network processing and visualization

To calculate the centrality measures of the coauthorship network, we used the R statistical package (www.r-project.org) and the iGraph package.¹⁵ We used its functions to calculate the degree, betweenness, closeness, Burt’s constraint, authority score, PageRank-like, and Kcore, with their standard parameters.

We visualized the network and prepared the manuscript figures using Cytoscape¹⁶ version 3.8.

SPICi for finding clusters and text processing

To find clusters of authors in the coauthorship network, we used SPICi,¹⁷ with the parameters [-s 3 -d 0.1 -g 0.4]. All clusters are available in Supplementary Table 4.

For each cluster, we selected the manuscripts authored by at least 3 of the authors who are members of the given cluster. We considered only clusters that had at least 3 representative studies. Next, we concatenated the abstracts from the selected manuscripts and processed them to combine plurals (eg, inhibitors and inhibitor) and removed words and synonyms that are common in hemophilia (eg, “hemophilia,” “hemophilia,” “FVIII,” “factor,” “FIX”).

Finally, we used an online server to process and depict the contents of the corpus containing the abstracts from each author-cluster (<https://www.wordclouds.com/>).

Prediction of the number of manuscripts to be published in the future

The prediction of the number of papers published annually in this area was performed using an ARIMA model, available in the statsmodels package,¹⁸ version 0.13.1, which describes the time series behavior by combining 3 different methods. We used the R statistical package version 3.4 (www.r-project.org) and Python version 3.6.9 (<https://www.python.org/>).

Code and data availability

The source code and the datasets used in the study are available at <https://github.com/madlopes/Hem-AuthNet>.

Results

Properties of the hemophilia authorship network

The representation of information as a network is a convenient way to depict a relationship between entities. To build a coauthorship network of hemophilia studies, we queried the PubMed database using a carefully built search term with several synonyms and aimed to include abstracts genuinely related to hemophilia while excluding articles that only occasionally mentioned terms from this field (see Methods). We downloaded a set of more than 20 000 textual abstracts in English, covering the period of 1960–2022.

Next, we created an undirected graph, where the *nodes* are the manuscript authors, and 2 nodes are connected by an *edge* if they co-authored at least 1 manuscript. In this network, the weights of the edges are the number of studies that the 2 researchers co-authored.

This approach yielded a network with more than 54 000 nodes and 305 000 edges. Upon closer inspection, we noticed that this network was too large to be processed using current algorithms, and several authors had only a single publication in the field; therefore, we pruned the network by including only authors with 2 or more publications related to hemophilia. In the end, our coauthorship network had 14 767 nodes and 117 257, and retained only the authors who made a continuous contribution to the field; we termed this network *Hem-AuthNet* (Supplementary Table 1).

In general, the Hem-AuthNet is a very dense and compact network, as evidenced by its more than 14 000 nodes connected and forming a very large central component (Figure 1A). Moreover, given its diameter, we found that the number of intermediates between any 2 researchers consistently working in this field is at most 15. Although Hem-AuthNet does not take into account the *time* component (ie, some studies

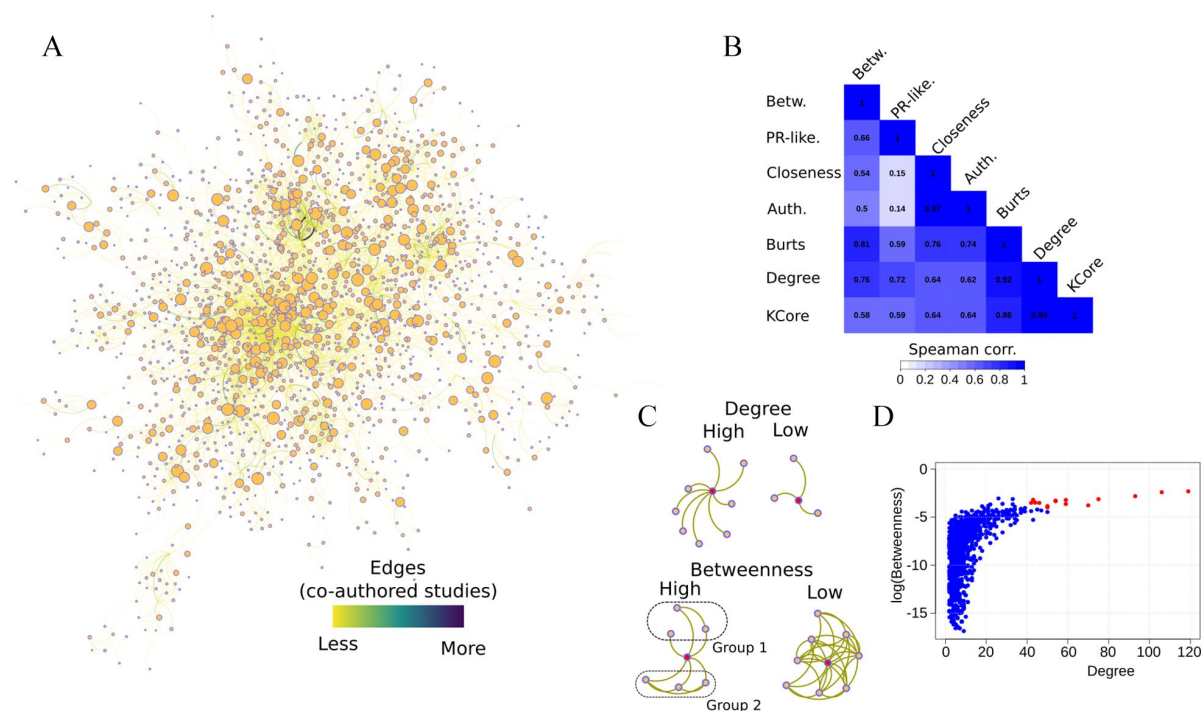


Figure 1. (A) In the hemophilia coauthorship network, each node represents an author and 2 authors are connected by an edge if they are listed as co-authors in a manuscript. The node diameter is proportional to the number of studies that each author published in 1960–2021. The color shade of the edges indicates the number of manuscripts co-authored by 2 researchers. (B) The Spearman correlation between the centrality measures derived from the pruned Hem-AuthNet. Burts' constraint is represented as $1/\text{value}$ to yield a positive correlation with the other measures. (C) In this network, high-degree nodes represent the authors who co-authored studies with several other researchers, whereas low-degree nodes represent those who collaborated with only a few others. High-betweenness nodes are those who served as a “bridge” between groups that would have no connection otherwise; on the contrary, low-betweenness nodes are the members of groups where most members are connected directly to each other. (D) While the vast majority of researchers in the Hem-AuthNet co-authored manuscripts with less than ~20 other researchers, a few of them served as the “hubs” of the network (ie, high-degree and high-betweenness). PR-like, PageRank-like.

were published decades apart), the presence of a large central component and the possibility of reaching all nodes with a small number of steps indicate that hemophilia is a highly collaborative field, likely due to the rarity of this disease and the small number of groups actively working on it.

Next, we investigated the connectivity properties of all authors in this network, namely, what kind of position they occupy within the hemophilia research landscape. For this purpose, we calculated several centrality measures of the Hem-AuthNet nodes; however, given the strong correlation that these measures displayed to each other, we found that only 2 measures sufficed (Figure 1B). Thus, for this analysis, we used the degree (how many connections a node has) and the betweenness (to what extent a node serves as a bridge to groups that otherwise would not be connected) (Figure 1C). We found that while most nodes make only a few connections, a few nodes have several dozen connections, for instance, among the most connected authors, ~100 co-authored manuscripts with more than 150 researchers. Moreover, the broad betweenness distribution displayed in Figure 1D indicates that while some authors co-authored studies only with their immediate contacts, other authors served as “bridges” between different groups and most likely participated in large multidisciplinary studies. Interestingly, the distribution of these

centrality measures is analogous to the properties exhibited by networks of a completely different nature, like the population size of cities¹⁹ and the magnitude of earthquakes.²⁰

Finally, we wondered what are the most central nodes in the whole Hem-AuthNet. To answer this question, we consider both the degree and the betweenness measures in conjunction (top 1% in both) and found that at least 108 authors filled this criteria (Supplementary Table 3); with their publication records combined, these authors have published more than 1000 manuscripts, have collaborated with thousands of researchers, and have a career spanning several decades (Supplementary Tables 2 and 3).

Taken together, these results indicate that Hem-AuthNet automatically identifies emerging authorship patterns in the hemophilia scientific literature. This approach offers a method to quickly identify the most prolific authors, their position within their collaboration network, and encode these patterns digitally, in a format that can be used for further *in silico* analyses.

Characterizing clusters of collaborators and their work

After studying the network characteristics of individual authors, we wondered about the properties that can be derived from groups of authors. For this purpose, we used a network

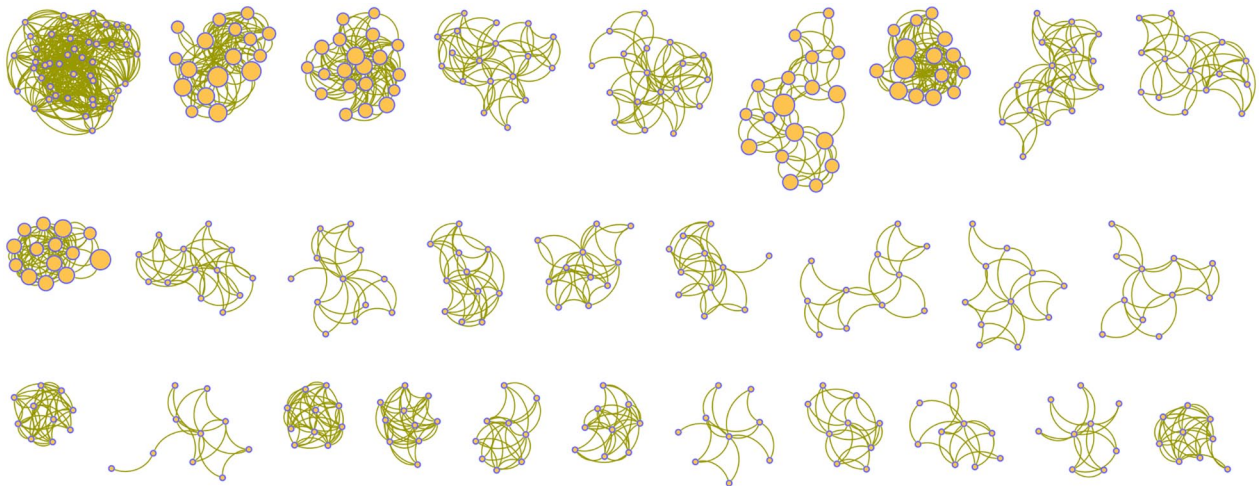


Figure 2. Using the Hem-AuthNet as input to a cluster-finding algorithm,¹⁷ we identified groups of researchers who co-authored several studies together over the years. Depicted are some of the representative clusters we found, and the node sizes are proportional to the number of studies published in the last 6 decades. While some clusters are tightly connected, indicating that most of their members appeared together as study authors, other clusters are less connected, suggesting more intermittent collaborations (the full network and the clusters are available in the Supplementary Material).

processing algorithm to identify *clusters* in the Hem-AuthNet—namely, groups of authors who have collaborated and published numerous hemophilia studies together.

In total, we found 25 clusters with sizes ranging from 3 to 15 authors (Figure 2; Supplementary Table 4). Although hemophilia researchers sporadically participate in studies involving several groups, our cluster detection algorithm identified the main network of each researcher—namely, the group of collaborators with whom they produced most of their studies. Interestingly, the clusters we found were marked by the presence of 1 or 2 senior authors and by several junior members. As Figure 2 depicts, the senior authors are easily distinguished by the node sizes, reflecting their number of publications. Moreover, it is clear that while most senior authors have close, persistent collaborations with only a few other researchers, most of the collaborations are only transient (Figure 1; Supplementary Table 1), probably due to the structure of most modern academic institutions.

Next, we used text mining algorithms to automatically analyze and determine the research specialties in the body of scientific work produced by the members of each cluster. For this purpose, we processed the ~20 000 manuscript abstracts related to hemophilia and selected those that had at least 3 authors from each cluster. In these texts, we found that its terms and sentences could readily identify the research interests from each group. As shown in Figure 3, these algorithms accurately found the groups working on the development of emicizumab,²¹ the bispecific antibody for hemophilia A prophylaxis (cluster 2), patient care (cluster 3), gene therapy (cluster 5), and inhibitor development (cluster 9), demonstrating that even with minimal processing and using only a handful of abstracts per group, the research topics in hemophilia are so specific that they are surprisingly suitable for algorithmic analysis.

Evidently, the information captured and represented by the Hem-AuthNet platform is a “snapshot” of the hemophilia literature, and this field is undergoing a permanent increase in the number and variety of topics (until 2025, we predict it will reach more than 1000 studies per year, or 1 study every ~8 hours; Supplementary Figure 1). In addition, we understand that the Hem-AuthNet layout and connectivity changes based on its input parameters, therefore, we took special care to make all datasets and source code available in a simple and intuitive format to enable the community to reproduce and extend our findings (see Data availability).

In summary, these results demonstrate the feasibility of representing the hemophilia research landscape as a network and show that this structure contains all information required for algorithms to reveal informative patterns and trends. Given the accelerating pace at which the hemophilia literature is growing, it is encouraging to verify that text mining techniques can promptly identify the research topics of each group based solely on the abstracts of their work.

Discussion

In this study, we created a comprehensive map spanning 6 decades of research in hemophilia (we named it the Hem-AuthNet). In this framework, we represented thousands of researchers, their collaborations, and the contents of their work. From a historical perspective, this work is an appreciation of thousands of careers dedicated to understanding the details of this bleeding disorder; from a practical point of view, the computational methods presented here enable researchers to make sense of the current vast hemophilia research landscape and to narrow down the scientific material that best aligns with their interests.

Even for a field with a scientific body of modest size (~20 000 articles), the complexity and number of authors

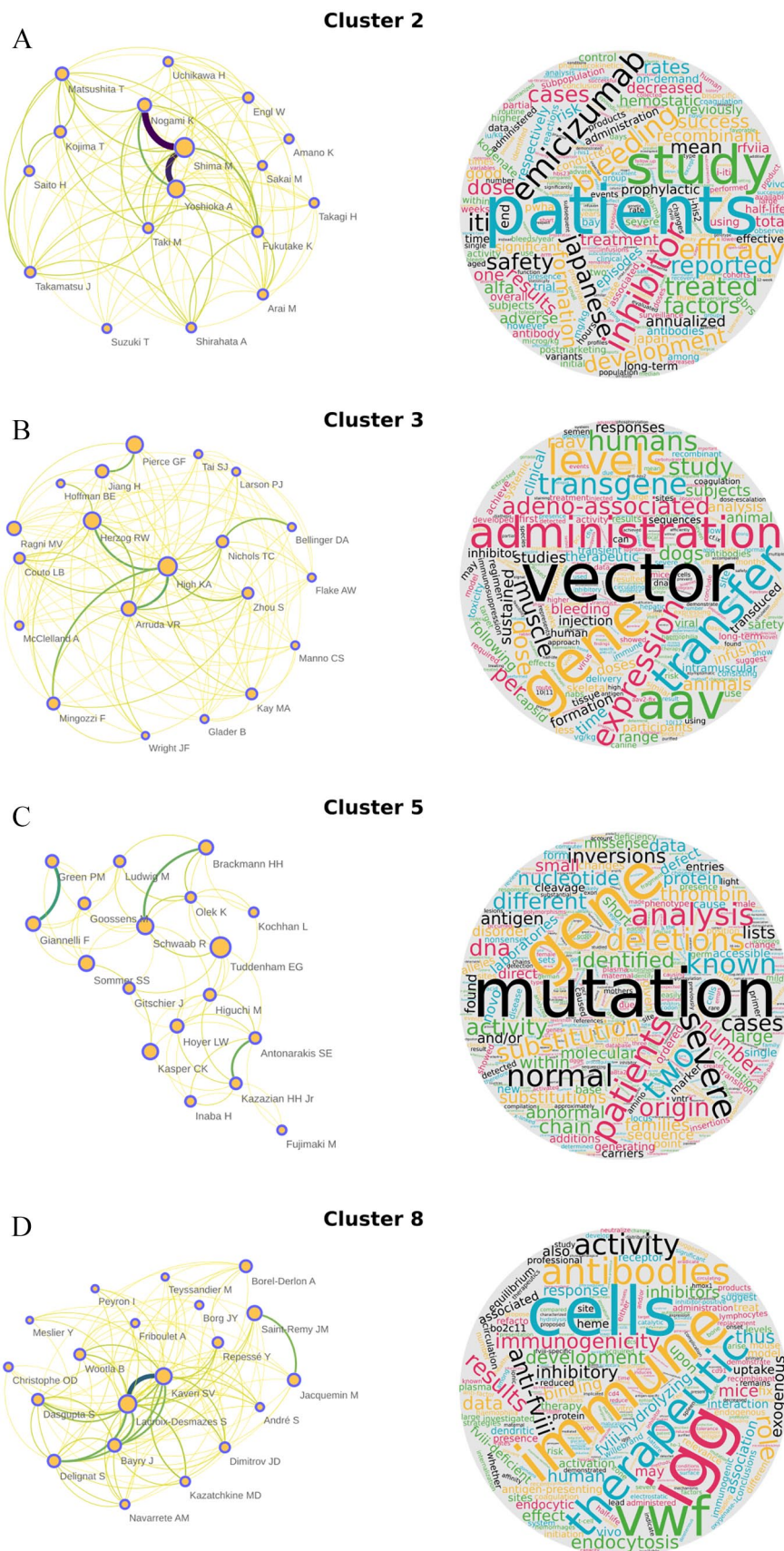


Figure 3. After automatically finding the most representative studies from each author-cluster (ie, the manuscripts with several authors from a given cluster), text mining techniques automatically captured the topics representing the essence of the research theme of each group. (A) Clinical studies and the novel therapeutics. (B) Topics related to gene therapy development and its underlying technologies. (C) Reports of mutation data and basic biology studies. (D) Studies related to the development of inhibitory antibodies and the immune system mechanisms.

composing the Hem-AuthNet largely surpass the human capacity to derive meaningful patterns from this structure. The representation of scientific collaborations as a network is a convenient way to create a structure that can be explored by algorithms. Our network analysis methods found that as in other research fields (eg, physics²²), the hemophilia research network also has “hubs”—namely, the authors who collaborated with hundreds of researchers and published dozens of articles (Figure 1). Interestingly, some of these hub researchers also served as “connectors” between different research groups (Supplementary Table 3); given that academic groups are often highly specialized in a few techniques, these researchers probably played a pivotal role in facilitating the development of studies that would otherwise not be conducted. The importance of persons interfacing and connecting different groups is a recurrent topic in social science studies,²³ and the Hem-AuthNet framework was able to detect and quantify this phenomenon in hemophilia research as well.

Interestingly, using the coauthorship network as input, we used graph analysis algorithms to find parts of the network that were strongly connected (ie, clusters). As in other networks derived from a variety of human activities,²⁴ we observed that in the more than 20 clusters, some collaborations were persistent and spanned several years, and others were only transient and sporadic (Figure 2). This is likely an emergent property of scientific collaboration networks, given that there is a small number of senior researchers, and a large number of junior members who undergo scientific training for only a few years.

Although it is important to visualize the connections made by the researchers working in the hemophilia field, it is essential to develop algorithms that make sense of their work. We found that using text mining techniques, we could identify the research topics representing the essence of each cluster—ranging from clinical care to molecular biology and drug development (Figure 3). This feature is particularly important because we predict that the literature related to hemophilia will increase dramatically (Supplementary Figure 1). Thus, it is important to create algorithms to automatically discover relevant content before researchers miss key studies due to the notorious information overload that already affects other fields.²⁵

In this sense, the research presented here opens interesting avenues for research. Perhaps the most exciting is the automatic discovery of patterns and connections between factual data that are not apparent to humans. These powerful techniques are already used to uncover the role of mutant genes in disease pathways,^{26,27} and to help synthesize novel materials that display notable physical properties.¹¹ If applied to hemophilia research, we anticipate that these methods will foster even better clinical care, physiotherapy programs and help in the resolution of issues threatening hemophilia patients (eg, the development of inhibitory antibodies and events of intracranial hemorrhage).

Conclusions

In summary, the framework presented here accurately represents the work produced by a large collaboration network established by thousands of hemophilia researchers in the last 6 decades. We expect that this system will facilitate knowledge discovery and will accelerate the development of superior treatments for people living with hemophilia.

Author Contributions

TJSL conceptualized the study designed the analysis. TL, TN, and RR performed the analyses, interpreted the results and wrote the manuscript.

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Lee CA, Berntorp E, Hoots K. *Textbook of Hemophilia*. 3rd ed. Chichester, England: John Wiley & Sons, Ltd; 2014.
- Fay PJ. Activation of factor VIII and mechanisms of cofactor action. *Blood Rev*. 2004;18:1-15.
- Lenting PJ, van Mourik JA, Mertens K. The life cycle of coagulation factor VIII in view of its structure and function. *Blood*. 1998;92:3983-3996.
- Berntorp E, Hermans C, Solms A, Poulsen L, Mancuso ME. Optimising prophylaxis in haemophilia A: the ups and downs of treatment. *Blood Rev*. 2021;50:100852.
- McLaughlin P, Hurley M, Chowdhary P, Khair K, Stephensen D. Physiotherapy interventions for pain management in haemophilia: a systematic review. *Haemophilia*. 2020;26:667-684.
- Peters R, Harris T. Advances and innovations in haemophilia treatment. *Nat Rev Drug Discov*. 2018;17:493-508.
- Leebeek FWG, Miesbach W. Gene therapy for hemophilia: a review on clinical benefit, limitations, and remaining issues. *Blood*. 2021;138:923-931.
- Wang M, Wang M, Yu F, Yang Y, Walker J, Mostafa J. A systematic review of automatic text summarization for biomedical literature and EHRs. *J Am Med Inform Assoc*. 2021;28:2287-2297.
- Simon C, Davidsen K, Hansen C, Seymour E, Barnkob MB, Olsen LR. BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics*. 2019;19:57.
- Yoneya T, Mamitsuka H. PURE: A PubMed article recommendation system based on content-based filtering. *Genome Inform*. 2007;18:267-276.
- Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*. 2019;571:95-98.
- Newman ME. Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci USA*. 2004;101:5200-5205.
- Newman ME. Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim E, Frauenfelder H, Toroczkai Z, eds. *Complex Networks*. Berlin, Germany: Springer; 2004:337-370.
- Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25:1422-1423.
- Csardi G, Nepusz T. The igraph software package for complex network research. *Int J Complex Syst*. 2006;1695:1-9.
- Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498-2504.
- Jiang P, Singh M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*. 2010;26:1105-1111.
- Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. Paper presented at: 9th Python in Science Conference; June 28-July 3, 2010:61; Austin, TX. <https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf>.
- Mori T, Smith TE, Hsu WT. Common power laws for cities and spatial fractal structures. *Proc Natl Acad Sci USA*. 2020;117:6469-6475.
- Meng F, Wong LNY, Zhou H. Power law relations in earthquakes from microscopic to macroscopic scales. *Sci Rep*. 2019;9:10705.
- Kitazawa T, Igawa T, Sampei Z, et al. A bispecific antibody to factors IXa and X restores factor VIII hemostatic activity in a hemophilia a model. *Nat Med*. 2012;18:1570-1574.

22. Newman ME. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys Rev E*. 2001;64:016132.
23. Burt RS. *Structural Holes: The Social Structure of Competition*. Cambridge, MA: Harvard University Press; 2009.
24. Palla G, Barabasi AL, Vicsek T. Quantifying social group evolution. *Nature*. 2007;446:664-667.
25. Landhuis E. Scientific literature: information overload. *Nature*. 2016;535:457-458.
26. Singhal A, Simmons M, Lu Z. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS Comput Biol*. 2016;12:e1005017.
27. Shen L, Shi Q, Wang W. Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis*. 2018;7:25.