Database, 2022, 1–3
DOI: https://doi.org/10.1093/database/baac039
Perspective/Opinion

New reasons for biologists to write with a formal language

Raul Rodriguez-Esteban*

Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Grenzacherstrasse 124, Basel 4070, Switzerland

*Corresponding author: Tel: +41-61-68-77533; Fax: +41-61-691 9391; Email: raul.rodriguez.esteban@gmail.com

Citation details: Rodriguez-Esteban, R. New reasons for biologists to write with a formal language. *Database* (2022) Vol. 2022: article ID baac039; DOI: https://doi.org/10.1093/database/baac039

Abstract

Current biological writing is afflicted by the use of ambiguous names, convoluted sentences, vague statements and narrative-fitted storylines. This represents a challenge for biological research in general and in particular for fields such as biological database curation and text mining, which have been tasked to cope with exponentially growing content. Improving the quality of biological writing by encouraging unambiguity and precision would foster expository discipline and machine reasoning. More specifically, the routine inclusion of formal languages in biological writing would improve our ability to describe, compile and model biology.

> […] language is just as indispensable a tool for the pursuit of biology as microscopes, kymographs and other instruments (46)

The primary way to describe biology is still document-centric natural language (1). Language is, therefore, fundamental to the development of biological research. Improving its quality to encourage unambiguity and precision fosters expository discipline and machine reasoning (2, 3) while striving towards Boole's ideal of a language '[…] freed from idioms and divested of superfluity, […] in a manner the most simple and literal […]' (4). This includes the use of standard nomenclatures, identifiers and reference databases, clear factual statements and computational or symbolic languages, i.e. formal languages.

The history of attempts to improve biological writings in such a way starts, at least, with the creation of standard nomenclatures for species dating back to Linnaeus in the 18th century (5) and Woodger's 1929 critique of the language used in biology (6, 7). Such attempts have had limited impact. Biological documents still contain many ambiguous names, convoluted sentences, vague statements and narrative-fitted storylines (8–11).

The need for better writing is, nonetheless, increasing, because the number of biology-related documents, such as scientific articles, patents and grants, keeps growing exponentially (12), as noticed even by the public during the COVID-19 pandemic (13, 14). Currently proposed solutions to cope with this growth appear to be insufficient. First, there is a scarcity of accessible and structured biological data derived from these documents. Biological databases, which are primary repositories for such data, are not growing to match current needs (15, 16), and the sustainability of their business model has been questioned (17–22).

Second, text mining is not, at the moment, a sufficient solution for the extraction of structured data from text. Arguably, taking off in the late 1990s with the release of PubMed (23), text mining went through a period of stagnation in performance benchmarks until recent advances in natural language processing (NLP). While new NLP algorithms have been able to master general linguistic tasks with greater ability than non-specialist college-educated humans, they are not yet able to extract complex biological relations (24–27) with acceptable performance, according to past community challenges (28), and with the exception of certain niche relation types. Crucially, complex relations, and associated contextual information, play a large role in the description of biological processes (29, 30).

Moreover, NLP algorithms have also lagged in tasks for which a certain level of factual knowledge is necessary, such as open-domain question-answering (31, 32). Knowledge graphs, which compile and organize knowledge of the world (33, 34), have been used to enrich NLP algorithms, powering them to state-of-the-art performance in both linguistic and factual applications (35–39).

Knowledge graphs can be partially created automatically but, in order to increase and maintain their quality, they need manually curated data (40), which can also be introduced through semi-automatic curation workflows based on artificial intelligence (AI) algorithms (41, 42). The increased use of knowledge graphs, including by companies such as Google and Meta, shows that improvements in NLP have not led to a decrease in, and one could say it has fed, the need for duly compiled, manually extracted knowledge. Thus, and perhaps counterintuitively, a golden era for NLP, and for AI in general, has been paralleled by growth in the use of knowledge graphs.

Improving biological writing has been recognized as one way to address the bottlenecks in the extraction of biological

Received 25 January 2022; Revised 18 March 2022; Accepted 17 May 2022

data from text. Recently, tips for scientific authors to make their articles more 'text-mining ready' have been proposed (43), and there has been yet another call for the use of standard names in biological texts, in this case for gene products (44). There is, however, a need to improve biological writing that is beyond the still-insufficient adoption of standard terminologies and text-mining-ready writing tips. Specifically, with the adoption of formal languages, such as Biological Expression Language (BEL) (2, 45), as part of regular writing practice.

Content written in a formal language, such as that related to protein interactions, phylogenies, drug–disease interactions or post-translational modifications, could be embedded inline within documents or in tables, metadata, equations or supplementary files or directly submitted to databases. Chemistry provides examples of how this could look in practice (46). A realization of the limitations of natural language and alchemical symbols (47) led to multiple successful initiatives in the 19th and early 20th centuries on the subject of standard nomenclatures, formulae and equations. Because of this, the text mining of chemical names is easier than the text mining of, for instance, gene names (48). Incidentally, these efforts were originally inspired by Linnaeus's work in biology (49).

Within biology, the field of systems biology has also had a strong interest in the use of formal languages (50), such as SBGN (51) or BioPAX (52). The latter particularly describes signalling pathways and, unlike BEL, represents direct biological mechanisms with a higher degree of granularity and complexity. Signalling pathways represented in a formal language offer a stark contrast with the unsystematic way in which they are described in biological writings (53).

Imagine, for instance, if the phrase 'TNF activates SYK' (54) were written as 'TNF activates SYK (p(HGNC:TNF)) -> act(p(HGNC:SYK))', using, in this case, the BEL language inside parentheses. This type of content could easily be extractable and would provide a source of readily available knowledge that would help improve the yield of database curation and the performance of AI/NLP algorithms. The ultimate goal would not be to improve AI/NLP algorithms or curation for their own sake but to improve our ability to describe, compile and model biology. For authors, this could also increase the visibility and impact of their work (55).

Formal languages should not be seen as computational biology any more than chemistry formulae are computational chemistry. Biology students can get acquainted (56) with ways to apply standard nomenclatures, write clearer factual statements and integrate formal languages in their writing. In the end, better biological writing would help both biologists and algorithms.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The author is an employee of F. Hoffmann-La Roche Ltd.

Conflict of interest

None declared.

References

1. Auer,S., Kovtun,V., Prinz,M. *et al.* (2018) Towards a knowledge graph for science. In: *WIMS '18: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, Association for Computing Machinery, New York, NY, USA, Novi Sad, Serbia, pp. 1–6.
2. Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198.
3. Boniolo,G., D'Agostino,M. and Di Fiore,P.P. (2010) Zsyntax: a formal language for molecular biology with projected applications in text mining and biological prediction. *PLoS One*, **5**, e9511.
4. Boole,G. (1854) *An Investigation of the Laws of Thought*. Walton & Maberly, London and Cambridge, UK.
5. Winston,J.E. (2018) Twenty-first century biological nomenclature—the enduring power of names. *Integr. Comp. Biol.*, **58**, 1122–1131.
6. Woodger,J.H. (1929) *Biological Principles: A Critical Study*. Routledge & Kegan Paul Ltd, London.
7. Nicholson,D.J. and Gawne,R. (2014) Rethinking Woodger's legacy in the philosophy of biology. *J. Hist. Biol.*, **47**, 243–292.
8. Hirschman,L., Morgan,A.A. and Yeh,A.S. (2002) Rutabaga by any other name: extracting biological names. *J. Biomed. Inform.*, **35**, 247–259.
9. Chen,L., Liu,H. and Friedman,C. (2005) Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, **21**, 248–256.
10. Rodriguez-Esteban,R. and Jiang,X. (2017) Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med. Genomics*, **10**, 59.
11. Jonnalagadda,S., Tari,L., Hakenberg,J. *et al.* (2009) Towards effective sentence simplification for automatic processing of biomedical text. In: *Proceedings of Human Language Technologies*. Association for Computational Linguistics, New York, NY, USA, Boulder, Colorado, pp. 177–180.
12. Bornmann,L. and Mutz,R. (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.*, **66**, 2215–2222.
13. Else,H. (2020) How a torrent of COVID science changed research publishing—in seven charts. *Nature*, **588**, 553.
14. Brainard,J. (2020) Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? *Science*.
15. Baumgartner,W.A. Jr, Cohen,K.B., Fox,L.M. *et al.* (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, **23**, i41–i48.
16. Rodriguez-Esteban,R. (2015) Biocuration with insufficient resources and fixed timelines. *Database (Oxford)*, **2015**, bav116.
17. Chandras,C., Weaver,T., Zouberakis,M. *et al.* (2009) Models for financial sustainability of biological databases and resources. *Database (Oxford)*, **2009**, bap017.
18. Reiser,L., Berardini,T.Z., Li,D. *et al.* (2016) Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database (Oxford)*, **2016**, baw018.
19. Karp,P.D. (2016) How much does curation cost? *Database (Oxford)*, **2016**, baw110.
20. Karp,P.D. (2016) Can we replace curation with information extraction software? *Database (Oxford)*, **2016**, baw150.
21. Poux,S., Arighi,C.N., Magrane,M. *et al.* (2017) On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*, **33**, 3454–3460.
22. Bourne,P.E., Lorsch,J.R. and Green,E.D. (2015) Perspective: sustaining the big-data ecosystem. *Nature*, **527**, S16–7.
23. Rodriguez-Esteban,R. (2016) Understanding human disease knowledge through text mining: what is text mining? In: Loging W (ed). *Bioinformatics and Computational Biology in Drug Discovery and Development*. Cambridge University Press, Cambridge, UK, pp. 47–62.

24. Zhu,L. and Zheng,H. (2020) Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks. *BMC Bioinform.*, **21**, 47.
25. Percha,B. and Altman,R.B. (2018) A global network of biomedical relationships derived from text. *Bioinformatics*, **34**, 2614–2624.
26. Mehryary,F., Kaewphan,S., Hakala,K. *et al.* (2016) Filtering large-scale event collections using a combination of supervised and unsupervised learning for event trigger classification. *J. Biomed. Semant.*, **7**, 27.
27. Gyori,B.M., Bachman,J.A., Subramanian,K. *et al.* (2017) From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.*, **13**, 954.
28. Huang,C.C. and Lu,Z. (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.*, **17**, 132–144.
29. Kim,J.D., Ohta,T. and Tsujii,J. (2008) Corpus annotation for mining biomedical events from literature. *BMC Bioinform.*, **9**, 10.
30. Thompson,P., Nawaz,R., McNaught,J. *et al.* (2011) Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinform.*, **12**, 393.
31. Cao,B., Lin,H., Han,X. *et al.* (2021) Knowledgeable or educated guess? Revisiting language models as knowledge bases. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, New York, NY, USA.
32. Wang,C., Liu,P. and Zhang,Y. (2021) Can generative pre-trained language models serve as knowledge bases for closed-book QA? In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, New York, NY, USA.
33. Hogan,A., Blomqvist,E., Cochez,M. *et al.* (2021) Knowledge graphs. *ACM Comput. Surv.*, **54**, 1–37.
34. Hoyt,C.T., Domingo-Fernández,D., Aldisi,R. *et al.* (2019) Re-curation and rational enrichment of knowledge graphs in Biological Expression Language. *Database*, **2019**, baz068.
35. Sun,Y., Wang,S., Feng,S. *et al.* (2021) ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv*: 2107.02137.
36. Fei,H., Ren,Y., Zhang,Y. *et al.* (2021) Enriching contextualized language model from knowledge graph for biomedical information extraction. *Brief. Bioinform.*, **22**, bbaa110.
37. Yuan,Z., Liu,Y., Tan,C. *et al.* (2021) Improving biomedical pretrained language models with knowledge. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*, Association for Computational Linguistics, New York, NY, USA, pp. 180–190.
38. Zhao,W., Zhao,Y., Jiang,X. *et al.* (2020) A novel method for multiple biomedical events extraction with reinforcement learning and knowledge bases. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 402–407.
39. Balabin,H., Hoyt,C.T., Birkenbihl,C. *et al.* (2022) STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs. *Bioinformatics*, **38**: 1648–56.
40. Weikum,G., Dong,L., Razniewski,S. *et al.* (2021) Machine knowledge: creation and curation of comprehensive knowledge bases. *Found. Trends® Databases*, **10**. *arXiv*, 11564.
41. Yun,W., Zhang,X., Li,Z. *et al.* (2021) Knowledge modeling: a survey of processes and techniques. *Int. J. Intell. Syst.*, **36**, 1686–1720.
42. Ge,T., Wang,Y., de Melo,G. *et al.* (2016) Visualizing and curating knowledge graphs over time and space. In: *Proceedings of ACL-2016 System Demonstrations. Association for Computational Linguistics*. Berlin, Germany, pp. 25–30.
43. Leaman,R., Wei,C.H., Allot,A. *et al.* (2020) Ten tips for a text-mining-ready article: how to improve automated discoverability and interpretability. *PLoS Biol.*, **18**, e3000716.
44. Fujiyoshi,K., Bruford,E.A., Mroz,P. *et al.* (2021) Opinion: standardizing gene product nomenclature-a call to action. *Proc. Natl. Acad. Sci. U. S. A.*, **118**, e2025207118.
45. Biological Expression Language. https://biological-expression-language.github.io/.
46. Woodger,J.H. (1952) *Biology and Language*. The University Press, Cambridge.
47. Fabbrizzi,L. (2008) Communicating about matter with symbols: evolving from alchemy to chemistry. *J. Chem. Educ.*, **85**, 1501.
48. Krallinger,M., Leitner,F., Rabal,O. *et al.* (2015) CHEMDNER: the drugs and chemical names extraction challenge. *J. Cheminform.*, **7**, S1.
49. Crossland,M.P. (1962) *Historical Studies in the Language of Chemistry*. Heinemann, London.
50. Strömbäck,L. and Lambrix,P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, **21**, 4401–4407.
51. Le Novère,N., Hucka,M., Mi,H. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.*, **27**, 735–741.
52. Demir,E., Cary,M.P., Paley,S. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 935–942.
53. Wu,C., Schwartz,J. and Nenadic,G. (2013) PathNER: a tool for systematic identification of biological pathway mentions in the literature. *BMC Syst. Biol.*, **7**, S2.
54. Takada,Y. and Aggarwal,B.B. (2004) TNF activates Syk protein tyrosine kinase leading to TNF-induced MAPK activation, NF-kappaB activation, and apoptosis. *J. Immunol.*, **173**, 1066–1077.
55. Cokol,M., Rodriguez-Esteban,R. and Rzhetsky,A. (2007) A recipe for high impact. *Genome Biol.*, **8**, 406.
56. Lichtenwalter,M.C. (1951) *How to Succeed in the Study of Biology*. University of California Press, Oakland, CA, USA, pp. 180–182.