

Linguistic Analysis of the bioRxiv Preprint Landscape

This manuscript ([permalink](#)) was automatically generated from [greenelab/annoxiver manuscript@59f2035](#) on June 16, 2020.

Authors

- **David N. Nicholson**

 [0000-0003-0002-5761](#) ·  [danich1](#)

Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552; T32 HG00046

- **Jane Roe**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [janeroe](#)

Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

1. What is a preprint
2. Why are preprints important?
3. Mention how preprints are being integrated into scientist's everyday workflow
4. Talk about biorxiv and discuss how it is one of the repositories that maintains preprints along with citation of others such as arxiv/medrxiv etc.
5. Discuss works that analyze biorxiv from an audience perspective (quantifying tweets etc.)
6. Mention the gap which consists of analysing the language of biorxiv preprints (first to do this)
7. ^ Why is this important? What will this allow for future research projects?
8. Provide list of contributions within this manuscript

Methods

Datasets

BioRxiv

BioRxiv [1] is a repository of biological and biomedical research preprints. We downloaded an xml snapshot of this repository on February 3, 2020 from bioRxiv's Amazon S3 resource [???] that contained the full text and image content of 98,023 preprints. Preprints on bioRxiv are versioned, and in our snapshot 26,905 of 98,023 contained more than one version. When preprints had multiple versions, we used only the latest one. Preprints in this snapshot were grouped into one of twenty-nine different categories. Each preprint was also classified as a new result, confirmatory finding, or contradictory finding. Some preprints in this snapshot have been withdrawn from bioRxiv. When a preprint is withdrawn, its content is replaced with the reason for withdrawal. Because we used the latest version, withdrawn preprints in our analysis contained only statements indicating their removal.

PubMed Central

PubMed Central (PMC) [2] is a repository that contains free-to-read articles. PMC contains two types of contributions: closed access articles from research funded by the United States National Institutes of Health (NIH) appearing after an embargo period and articles published under Gold Open Access [3] publishing schemes. Paper availability within PMC is largely dependent on the journal's participation level [4]. Individual journals have can fully participate in submitting articles to PMC, selectively participate sending only a few few of papers to PMC, only submit papers according to NIH's public access policy [5], or not participate at all. As of September 2019, PMC had 5,725,819 articles available [6]. Out of these 5 million articles, about 3 million were open access and available for text processing systems [7,8]. We downloaded a snapshot of this open access subset on January 31, 2020. This snapshot contains papers such as literature reviews, book reviews, editorials, case reports, research articles and more; however, we used only the research articles.

Comparing Corpora

We used gensim [9] (version 3.8.1) to preprocess the bioRxiv and PubMed Central corpora. We removed the 337 gensim-provided stopwords. Throughout our analysis we encountered non-word symbols (e.g., \pm), so we refer to words and symbols as tokens to avoid confusion.

Following the cleaning process, we calculated the frequency of every token shared between both corpora. Because many tokens were unique to one set or the other and observed at low frequency, we used the union of the top 100 most frequent tokens from each corpus to compare them. We generated a contingency table and calculated the odds ratio for each token. Furthermore, we also calculated the 95% confidence interval for each odds ratio [???].

Visualizing the Preprint Landscape

Generate Document Representation

We used gensim [9] (version 3.8.1) to train a word2vec continuous bag of words (CBOW) [10] model over the bioRxiv corpus. Our neural network architecture had 300 hidden nodes, and we trained this model for 20 epochs. We set a fixed random seed and otherwise used gensim's default settings. Following training, we generated a document vector for every article within bioRxiv and PubMed Central. This document vector is calculated by taking the average of every token present within a given article, ignoring those that were absent from the word2vec model.

Dimensionality Reduction of Document Embeddings

We used principal component analysis (PCA) [11] to project bioRxiv document vectors into a low dimensional space. We trained this model using scikit-learn's [12] implementation of a randomized solver [13] with a random seed of 100, output of 50 principal components, and default settings for all other parameters. For each principal component we calculated its cosine similarity with all tokens in our word2vec model's vocabulary. We report the top 100 positive and negative scoring tokens in the form of word clouds, where the size of each word corresponds to the magnitude of similarity and color represents positive (blue) or negative (orange) association.

Journal Recommendation

We aimed to predict the journal a paper would eventually be published in based on its embedding representation. We illustrate our recommendations as a short list along with a network visualization available at [website link here once available](#). Since we sought to examine if embeddings were related to publication venue, we used a simple k-nearest neighbors approach with Euclidean distance to recommend journals.

First, we filtered all journals that had fewer than 100 papers in the PMC dataset. A subset of our PMC corpus was directly linked to papers in bioRxiv as they had been published as open access articles. We held out this subset and treated it as our gold standard test set. We used the remainder of the PMC corpus for training and initial evaluation via cross validation. We considered a list of ten journal suggestions to be an appropriate number and we considered a prediction to be a true positive if the correct journal appeared within the ten closest neighbors of the query article.

Certain journals publish articles in a focused topic area, while others publish articles that cover many topics. Likewise, some journals have a publication rate of at most hundreds of papers per year while others publish at a rate of at least ten-thousand papers per year. Accounting for these characteristics, we designed two approaches for recommending journals.

The first approach is based on individual paper proximity, which enabled us to provide an example of the specific article or articles that led to the prediction. Conversely, predictions using this technique

could be biased due to the overrepresentation of general topic journals. We call this approach the paper-based classifier. This classifier takes a query article that has been projected onto the embedding space trained on bioRxiv preprints as input and reports the journals of the top ten closest papers. The number of journals returned via this method could be less than ten as multiple papers in close proximity to query article may belong to the same journal.

The second approach is based on close proximity to a journal's centroid. This technique provides recommendations that are more focused on domain-specific publication venues. We call this approach the journal-based classifier. This classifier was trained by computing journal centroids via taking the average embedding of all papers published in each journal. Following the centroid calculation, this classifier takes a query article projected onto the same embedding space as above for input and reports the top ten nearest journals centroids. Both the paper-based classifier and the journal-based classifier were optimized via 10-fold cross validation. Lastly, we evaluated performance of both classifiers on our gold standard test set of published preprints.

Results

Comparing bioRxiv to PubMed Central

Global View

1. Create a treemap visualization of top X terms that are different between bioRxiv and PubMed Central (based on odds ratio)

Published Preprint Differences

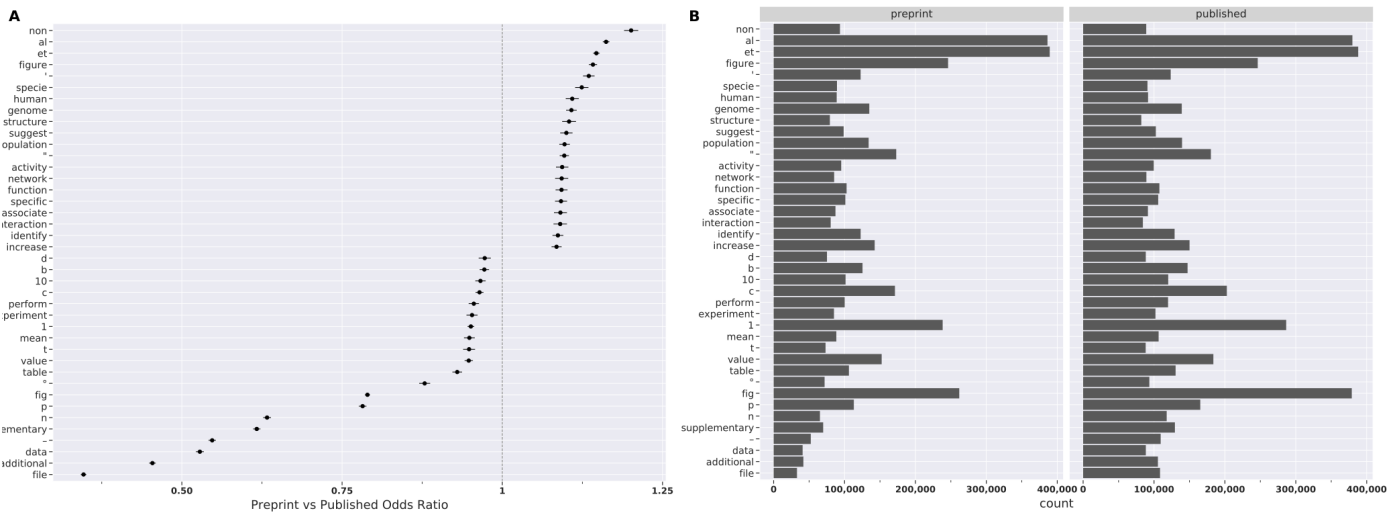


Figure 1: Top scoring tokens for preprints are focused on figure citations whereas their published versions are more focused on data availability. The plot on the left (A) is a point range plot of the odds ratio with respect to preprints. Values greater than one indicate a high association with preprints while values less than one indicate a high association with published articles. The dotted line provides a breaking point between both categories. The plot on the right (B) is a barchart of token frequency appearing in preprints and published versions of preprints respectively.

A preprint's linguistic style can change once a preprint has undergone the revision process prior to being published. We quantified this linguistic difference by calculating the odds ratio of tokens appearing in the union in bioRxiv preprints and their published counterparts within PMC. Tokens with an odds ratio greater than one are mainly centered on paper/figure references and research specific terms (Figure {[[???]]: biorxiv_pmc_pre_published_comp}). Tokens with an odds ratio of less than one are focused on data availability, and research measurements such as number of cases and controls or significance testing (Figure {[[???]]: biorxiv_pmc_pre_published_comp}).

The bioRxiv Preprint Landscape

1. Provide the tSNE figure of the bioRxiv
2. Discuss the results of the tSNE figure and highlight that there are category clusters within the figure

Topic Analysis of Principal Components

1. Provide an example of the word cloud for principal components
2. Show plot of the principal components and the scatterplot
3. Mention that the word clouds can be found at xyz

Journal Recommendations/Audience Associations

1. Title will change once analysis is finished
2. Provide key figure for this section and take-home message

Discussion

Conclusion

Acknowledgements

References

1. **bioRxiv: the preprint server for biology**

Richard Sever, Ted Roeder, Samantha Hindle, Linda Sussman, Kevin-John Black, Janet Argentine, Wayne Manos, John R. Inglis

bioRxiv (2019-11-06) <https://doi.org/ggc46z>

DOI: [10.1101/833400](https://doi.org/10.1101/833400)

2. **PubMed Central: The GenBank of the published literature**

R. J. Roberts

Proceedings of the National Academy of Sciences (2001-01-16) <https://doi.org/bbn9k8>

DOI: [10.1073/pnas.98.2.381](https://doi.org/10.1073/pnas.98.2.381) · PMID: [11209037](https://pubmed.ncbi.nlm.nih.gov/11209037/) · PMCID: [PMC33354](https://pubmed.ncbi.nlm.nih.gov/pmc/PMC33354/)

3. **Gold open access: the best of both worlds**

M. A. G. van der Heyden, T. A. B. van Veen

Netherlands Heart Journal (2017-12-01) <https://doi.org/ggzfr9>

DOI: [10.1007/s12471-017-1064-2](https://doi.org/10.1007/s12471-017-1064-2) · PMID: [29196877](https://pubmed.ncbi.nlm.nih.gov/29196877/) · PMCID: [PMC5758455](https://pubmed.ncbi.nlm.nih.gov/pmc/PMC5758455/)

4. **How Papers Get Into PMC** <https://www.ncbi.nlm.nih.gov/pmc/about/submission-methods/>

5. **8.2.2 NIH Public Access**

Policy https://grants.nih.gov/grants/policy/nihgps/html5/section_8/8.2.2_nih_public_access_policy.htm

6. **PMC Overview** <https://www.ncbi.nlm.nih.gov/pmc/about/intro/>

7. **PMC text mining subset in BioC: about three million full-text articles and growing**

Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, Zhiyong Lu

Bioinformatics (2019-09-15) <https://doi.org/ggzfsb>

DOI: [10.1093/bioinformatics/btz070](https://doi.org/10.1093/bioinformatics/btz070) · PMID: [30715220](https://pubmed.ncbi.nlm.nih.gov/30715220/) · PMCID: [PMC6748740](https://pubmed.ncbi.nlm.nih.gov/pmc/PMC6748740/)

8. **PubTator central: automated concept annotation for biomedical full text articles**

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, Zhiyong Lu

Nucleic Acids Research (2019-07-02) <https://doi.org/ggzfsc>

DOI: [10.1093/nar/gkz389](https://doi.org/10.1093/nar/gkz389) · PMID: [31114887](https://pubmed.ncbi.nlm.nih.gov/31114887/) · PMCID: [PMC6602571](https://pubmed.ncbi.nlm.nih.gov/pmc/PMC6602571/)

9. **Software Framework for Topic Modelling with Large Corpora**

Radim Řehůřek, Petr Sojka

Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (2010-05-22)

10. **Efficient Estimation of Word Representations in Vector Space**

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

arXiv (2013-09-10) <https://arxiv.org/abs/1301.3781>

11. **Probabilistic Principal Component Analysis**

Michael E. Tipping, Christopher M. Bishop

Journal of the Royal Statistical Society: Series B (Statistical Methodology) (1999-08)

<https://doi.org/b3hjw7>

DOI: [10.1111/1467-9868.00196](https://doi.org/10.1111/1467-9868.00196)

12. **Scikit-learn: Machine learning in Python**

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P.

Prettenhofer, R. Weiss, V. Dubourg, ... E. Duchesnay
Journal of Machine Learning Research (2011)

13. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions

Nathan Halko, Per-Gunnar Martinsson, Joel A. Tropp
arXiv (2014-04-29) <https://arxiv.org/abs/0909.4061>