Assignment 1: PubMed Query to Track Histone Biology in the 1980s

Laxmi Vijayan

May 28, 2023

1 Introduction

Each diploid cell in the human body contains approximately six billion base pairs of DNA, roughly six feet of genetic material.[8] Some amount of packing is required to fit this genetic material into the cell nucleus, which, in humans is approximately 10 µm in diameter. [9] One large reduction in the size of DNA is accomplished when it is wound around proteins called histones, like thread around a spool, forming chromatin. In chromatin form, six feet of DNA are reduced to approximately a foot in length. [8]

The aforementioned histones are a small family of alkaline proteins that help stabilize, structure, and organize DNA in eukaryotic cells and some prokaryotic cells. Today, we know an octamer consisting of positively-charged histone variants (i.e., two of each: H2A, H2B, H3, and H4) bind to approximately 146 base pairs of negatively-charged DNA to form a nucleosome. [5] With the addition of H1 and another 20 base pairs of DNA, a chromatosome is formed. [2] Though histones were discovered in 1884, this knowledge, as well as our knowledge about the role of histones in gene regulation, was largely accumulated over thirty years starting in the early 1970s. [2] The discovery of the structure of chromatin and the existence of nucleosomes in the mid-1970s and several concurrent advances in technology, such as improvements in protein analysis and microscopy, during this period catalyzed the golden age of histone biology. [5]

In this report, we use bibliographic data such as DOI, references, and citations to create a citation network to visualize this period in histone biology.

2 Methods

The bibliographic data was collected from PubMed, a database supporting search and retrieval of biomedical and life sciences literature. While databases such as Scopus and EuropePMC are alternative venues for data collection, they were not considered due to stringent limitations on the number of records that could be downloaded at one time (SCOPUS) and sparse results (EuropePMC).

Using PubMed's GUI, these Boolean queries were constructed and used:

```
Query 1:
((((‘journal article’[Publication Type]) AND (‘english’[Language])) AND
    ((‘1980/01/01’[Date – Publication] : ‘1990/12/31’[Date – Publication])))) AND
     ((histone) OR (methylase) OR (deacetylase) OR (epigenomics) OR (histones) OR
     (methylases) OR (deacetylases) OR (epigenomic) OR (epi genomic) OR (epi
    genomics))

Query 2:
((((‘journal article’[Publication Type]) AND (‘english’[Language])))) AND ((
    histone) OR (methylase) OR (deacetylase) OR (epigenomics) OR (histones) OR (
    methylases) OR (deacetylases) OR (epigenomic) OR (epi genomic) OR (epi
    genomics))
```

The first query was used directly with PubMed's GUI and returned 10,238 results. PubMed enforces a limit on the number of results that can be exported as .CSV files, so results were exported in three batches and collated using the following command:

```
cat queryp1.csv <(tail -n +2 queryp2.csv) <(tail -n +2 queryp3.csv) > query3.csv
```

Another drawback to using the PubMed GUI was the inability to customize which metadata to include in the export file. As such, a function was written using the Entrez module from the Bio package in Python to query PubMed programmatically. The second query was passed to the function, along with the start and end dates, so multiple queries spanning smaller time intervals could be conducted and exported as one .CSV file. The function used the ESearch and ESummary to retrieve IDs and then recover detailed summary information for each result in the query. The fields that were retrieved were 'PMID,' 'Publication Year,' 'Title,' 'Authors,' 'DOI,' and 'Publication.'

Both, the manual search and retrieval and the programmatic retrieval, resulted in 10,238 results. Once the results were visually verified, the PMID and DOI were exported to a PostgreSQL table. To create a network the DOIs were modeled as nodes and joined to the Open Citations dataset. The Open Citations dataset is a PostgreSQL table that models a network with 75,025,194 nodes and 1,363,605,603 edges. The resultant table of matched DOIs contained 399,762 non-distinct DOIs. From this table, distinct DOIs were extracted to identify the nodes in the network. For each node, the number of references (in-degree), number of citations (out-degree), and total degree were calculated. The publication year for each node was also captured from the Open Citations dataset for each distinct DOI. This table was exported as a .CSV file, and descriptive statistics were calculated using Python.

3 Results

Of the 10,238 PubMed query results, only 8,310 results had DOI. When these results were checked for case differences, we found that there were only 8,119 distinct DOIs. The final dataset contained 209,005 nodes and 799,524 total edges. Distinct DOIs from the matched table were considered nodes and the sum of the number of references (in-degree edges) and the number of citations received (out-degree edges) were included for the total degree. Eighty-nine percent of DOIs from the original query were represented in the final dataset. The DOIs that were not represented can be assumed to have no references to or citations from the DOI represented in the Open Citations dataset.

3.1 Citation Distribution

Of the 209,005 DOIs, 154,267 had references, and 64,278 DOIs were cited by another DOI. The average number of references was 2.59 and the average number of citations was 6.17. For DOIs in the top 10 percentile for number of references, the average median of references was 7. Similarly, for DOIs in the top 1 percentile for number of citations, the average median of citations was 23. This heavily right-skewed data is shown in Figure 1. The paper with the greatest number of references was "Chromatin Structure and Gene Activity: The Role of Nonhistone Chromosomal Protein" by Iain Cartwright et al. (1982) which was a review paper published in Critical Reviews for Biochemistry.[3] It cited a total of 752 papers according to the PubMed metadata, of which 657 were captured by our network. The paper with the greatest number of citations was "Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids, and other polar compounds" by Andrew J. Alpert (1990) in the Journal of Chromatography.[1]

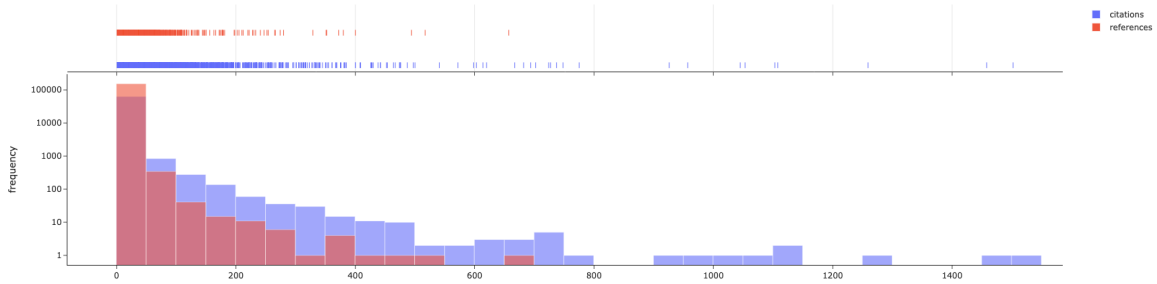
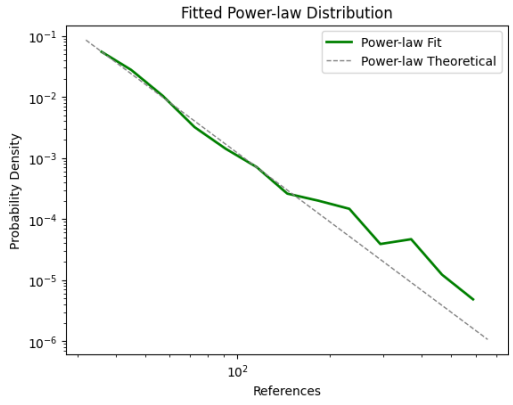
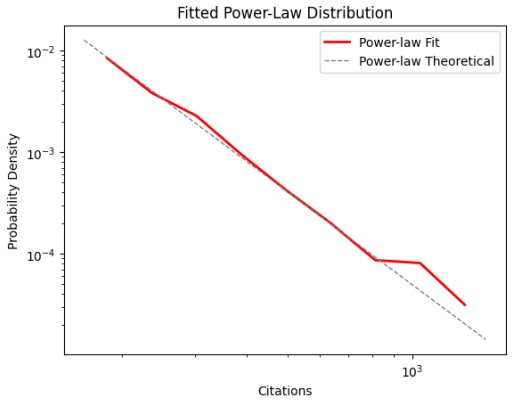


Figure 1: References and citation frequency for 209,005 distinct DOI

Given the heavy tail, the power-law distribution was fit to the data for number of references and number of citations. As seen in Figure 2, the data for both mostly match the theoretical fit, though there are deviations near the tail.



(a) The data for number of references mostly matches the theoretical power-law



(b) Similarly, the data for number of citations also closely matches the theoretical power-law

Figure 2: Number of references and citations fit to a power-law distribution

Consequently, the goodness of fit was compared to the lognormal distribution which is another heavy-tailed distribution. The log-likelihood ratio for references was -0.034 and the p-value was 0.70, which suggests a lognormal distribution. Similarly, the log-likelihood ratio for citations was -0.800 and the p-value was 0.42, which again suggests a lognormal distribution.

3.2 Citations Over Time

This network includes papers published between 1870 and 2022. The publication year was collected from the Open Citation database, and when one paper purportedly cited in 1870 was cross-checked by looking up the DOI on PubMed, it was discovered that the dates did not match. This paper, according to its PubMed metadata, was published in 1984.[7] Even accounting for minor discrepancies, the number of citations and references were examined over each year across the dataset (Figure 3).

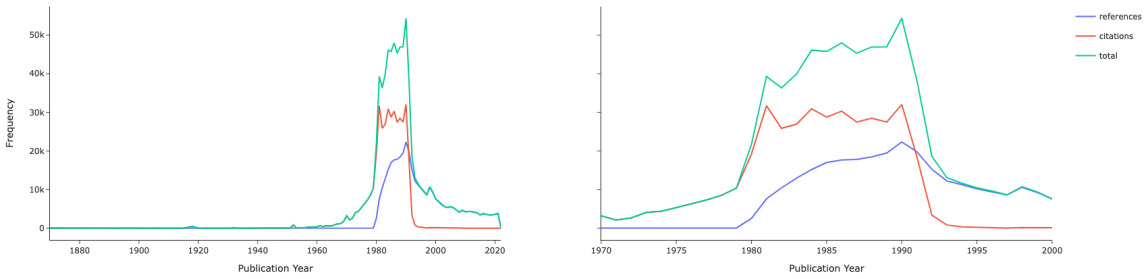


Figure 3: The number of references, citations, and the sum of both in each year across the dataset. The figure on the right

Most of the papers were published in the 1980 - 1990 time period. This was expected, given that our original query was for papers published in this period. The number of references in papers published during the decade 1980 to 1990 rapidly grows. This can be attributed to accumulating interest and growth in this scientific area. It can also be attributed to the change in reference patterns over time. The average number of references per paper published in 1961 was about 15.[6] The average number of references per paper published in 2019 was 45.[4] The number of citations per year grows as we enter the decade, but is more steady through the decade. This can be interpreted as a steady amount of activity in the scientific area.

4 Conclusion

A network of scientific papers related to histone biology focusing on the time period of 1980 to 1990 was created using a PubMed query and a large bibliometric citation network, Open Citations. This network had 209,005 nodes and 799,524 edges. Descriptive analysis showed that a majority of the papers in the network cite few papers and are cited by few papers. Using data visualization, we were able to demonstrate how the field of histone biology grew over the course of a decade and the resulting patterns through bibliometric data.

References

[1] Andrew J. Alpert. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of Chromatography A*, 499:177–196, 1990.

[2] Anthony T. Annunziato. Dna packaging: Nucleosomes and chromatin. *Nature Education*, 1(1):26, 2008.

[3] Iain L. Cartwright, Susan M. Abmayr, Gerhard Fleischmann, Ky Lowenhaupt, Sarah C. R. Elgin, Michael A. Keene, and Gary C. Howard. Chromatin structure and gene activity: The role of nonhistone chromosomal protein. *Critical Reviews in Biochemistry*, 13(1):1–86, 1982.

[4] Can Dai, Quan Chen, Tao Wan, Fan Liu, Yanbing Gong, and Qingfeng Wang. Literary runaway: Increasingly more references cited per academic research article from 1980 to 2019. *PLOS One*, 16(8), 2021.

[5] Donald E. Olins and Ada L. Olins. Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology*, 4:809–814, October 2003.

[6] Derek J. De Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[7] Alexander Rich, Alfred Nordheim, and Andrew H.-J. Wang. The chemistry and biology of left-handed z-dna. *Annual Review of Biochemistry*, 53(1):791–846, 1984.

[8] Brittany Simpson, Connor Tupper, and Nora M. Al Aboud. Genetics, dna packaging.

[9] Hui Bin Sun, Jin Shen, and Hiroki Yokota. Size-dependent positioning of human chromosomes in interphase nuclei. *Biophysical Journal*, 79(1):184–190, 2000.