From Big Data to Small Data in Bioinformatics

Abstract

Larger biological datasets better approximate the normal distribution and hence big data allows researchers to study the entire distribution. However due to the computational power required, numerous researchers are unable to conduct studies involving big data. Among the numerous solutions to tackle this solution, this study focuses on developing a pipeline to reduce the dataset to make it computationally feasible, without sacrificing data of significance. This was done by clustering the dataset and then improving the genetic diversity per cluster. The pipeline was developed using Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), which was responsible for the 2019 pandemic and has over 9 million sequences available on NCBI web server.

Introduction

According to the central limit theorem, larger and larger datasets approximate the normal distribution better, and biological data such as nucleotide sequences are no different [1]. Majority of the sequences correspond to similar properties while a minority of the sequences correspond to dissimilar properties found on the extreme end of the distribution. However, given that the population of different species can vary between a few hundred to several trillions, it becomes difficult to study the entire distribution [1]. Hence the focus of big data in bioinformatics is to "capture" the extreme ends of the distribution by sequencing more and more samples. And with over 3.7 billion nucleotide sequences hosted by GenBank in 2024, big data has facilitated the development of numerous fields - such as evolutionary biology, molecular biology, metagenomics, medicine, and forensic investigations - it has several caveats which limits its usage by the average researcher [2,3,4].

As big data becomes bigger, so does the computational power and storage requirements; which the average researcher may not have access to, and hence restricts their contributions to their fields. Hence a big focus of research is on making research with big data accessible; whether it by developing better and cheaper hardware, or developing softwares with better time complexities [3,4]. However, a point of focus which often goes unseen is the data itself: that is to reduce the dataset in such a manner that data of significance (i.e. corresponding to extreme ends of normal distribution) is not lost.

While there are some methods available for this purpose, the most optimal is perhaps to improve the nucleotide diversity of the sample. In addition to making the analysis computationally feasible, improving the diversity can change the corresponding distribution from normal to uniform. In normal distributions, there is chance of data of significance resembling and being regarded as random noise, but this chance is signficantly reduced in a uniform distribution.

Objective

The objective of this study is to develop a simplistic pipeline which reduces a large dataset into a relatively small dataset without losing data of significance. This is done by clustering the dataset and then improving the genetic diversity of the each cluster.

For the purpose of developing the pipeline, this study will use Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), which was responsible for the 2019 pandemic and has over 9 million sequences available on NCBI web server.

Method

Databases such as NCBI are often plagued by poorly read sequences which can negatively impact results, but they can be removed using a filtration criteria available on the web server. However, the web server can not download large datasets. And while the command line tool "datasets" by NCBI is developed for this purpose, it lacks the filtration capabilities of the web server. Hence after filtration, the results table should be downloaded from which the IDs can be extracted and fed to datasets. Depending on the internet stability, it may be desirable to download the sequences in batches. After downloading the dataset, duplicates should be removed using tools such as SeqKit.

Depending on the study, it may be desirable to extracting open reading frames (ORFs) from the dataset as it can reduce the nucleotides per sequence and reduce the overall dataset significantly since distinct sequences need not have distinct ORFs. This can be done using command line tools such as EMBOSS and orfipy.

Clustering is the crux of the pipeline but traditional methods such as k and hierarchical clustering require the dataset to be aligned which can be time-consuming and computationally expensive. Hence, alignment-free softwares should be used to cluster the sequences, such as MeShClust and MMSeqs2. It is recommended to review literature, as there may already exist some classifications for the dataset, such as Pangolin lineage for SARS-CoV-2, which can be used for clustering. Moreover, if the clusters are too large, then it is recommended to further sub-cluster them for ensuring maximum reduction of the dataset.

Unfortunately, there is currently no reliable method for improving genetic diversity without an alignment, and hence the clusters must be aligned before proceeding. While computationally expensive, it can be made feasible using loose parameters and/or tools suited for large databases, such as MAFFT, HAlign, and clustalw.

To improve the genetic diversity of a cluster, construct an identity matrix and then convert it into a distance matrix. Determine the set of sequences S_t, whose distance is greater than some proportion of the maximum distance, i.e.

$$S_t = \{m_i, m_j : d(m_i, m_j) \geq t \cdot \max(\{d \in M\})\} \text{ where } t \in \{0.9, 0.95, 0.999, 1\}$$

Determine Nei's nucleotide diversity of all sets S_t. The set with maximum diversity comprises the genetically improved cluster.

Finally, we merge the unaligned sequences corresponding to the sequences in the improved cluster to form the reduced dataset. This is important as the alignment was rough and not precise. The reduced dataset should now be significantly less than the original dataset while maintaining the entire distribution.

General Pipeline

1. Using an inclusion/exclusion criteria, filter the sequences on NCBI web server and download the results table.
2. Extract the IDs from the results table and turn into multiple files.
3. Download sequences by feedings IDs to datasets via loops and then remove duplicates using SeqKit.
4. Use EMBOSS to extract ORFs and remove duplicates using SeqKit.
5. Cluster and sub-cluster the sequences.
6. Roughly align clusters using MAFFT.
7. Improve genetic diversity per cluster using provided method.

Pipeline Requirements

Hardware Details:

- Model: Dell Inc. Precision 7730
- CPU: Intel Core i7-8750H × 12
- RAM: 32 GB
- VRAM: 8 GB
- OS: Fedora Linux 40 (Workstation) with Linux 6.11.4-201.fc40.x86_64

Softwares Details:

- R 4.4.1
  - tidyverse 2.0.0
  - stringr 1.5.1
  - kmer 1.1.2
  - ape 5.8.1
  - pegas 1.3
  - dendextend 1.19.0
- GNU bash 5.2.26
- datasets 16.22.1
- SeqKit 2.8.2
- EMBOSS 6.6.0.0
- MAFFT 7.526

Results and Discussion

The data was collected on 28th January 2025 from NCBI using the following criteria:

- Taxon ID: 2697049
- Ambiguous Characters < 30
- Nucleotide Completeness: Complete
- Host: 9696 (human)
- Collection date: 30th December 2003 - current date

The results table was downloaded and the files with 10k IDs each were fed into datasets to download the sequences in batches. The S gene (of length ~3kb) was extracted using EMBOSS getORF. The remaining sequences were clustered first by pangolin lineage and then by hierarchical clustering, forming 1546 clusters which were roughly aligned using MAFFT. The nucleotide diversity was improved per cluster and finally the unaligned sequences were merged. The number of sequences was reduced significantly as shown below.

Original Dataset	Downloaded from NCBI	Unique ORFs	Reduced Dataset	Reduction (%)
9,064,523	1,365,419	181,940	6,326	143,190

While calculating the nucleotide diversity of the original dataset is difficult as it requires it to be aligned, the nucleotide diversity per cluster is compared in the table below.

Dataset	Average nucleotide diversity per cluster	Minimum nucleotide diversity per cluster	Maximum nucleotide diversity per cluster
Original	0.0012775	0.0000000	0.4823382
Reduced	0.0082952	0.0000000	0.6081363
Change (%)	549.3	0.0	26.0

Conclusion

While big data has facilitated the development of numerous fields, it has also restricted numerous researchers from contributing to research due to the computational power required. Among the numerous methods to tackle this issue, one is to reduce the dataset without losing data of significance. This study provides a simplistic pipeline for this purpose which is based on improving the nucleotide diversity of clustered sequences. Using SARS-CoV-2, the number of sequences was reduced by 143190% while the average nucleotide diversity per cluster was improved by 26%. Hence this pipeline allows researchers with limited resources to conduct research involving big data.

References

1- Górski, A. Z., & Piwowar, M. (2021). Nucleotide spacing distribution analysis for human genome. Mammalian Genome, 32(2), 123–128. https://doi.org/10.1007/s00335-021-09865-5

2 - Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., & Karsch-Mizrachi, I. (2023). GenBank 2024 Update. Nucleic Acids Research, 52(D1), D134–D137. https://doi.org/10.1093/nar/gkad903

3 - Gupta, A., Kumar, S., & Kumar, A. (2023). Big Data in Bioinformatics and Computational Biology: Basic Insights. Methods in Molecular Biology, 153–166. https://doi.org/10.1007/978-1-0716-3461-5_9

4 - Nazipova, N. (2017). Big data in bioinformatics. Mathematical Biology and Bioinformatics, 12(1), 102–119. https://doi.org/10.17537/2017.12.102