

RESEARCH

Introduction to Focus Areas in Bioinformatics Project Week 5

Sina Glöckner^{*}, Christina Kirschbaum, Swetha Rose Maliyakal Sebastian and Gokul Thothathri

^{*}Correspondence:

sina.gloeckner@fu-berlin.de

Institute for Informatics, Freie

Universität Berlin, Takustr. 9,

Berlin, DE

Full list of author information is
available at the end of the article

Abstract

Goal of the project: The goal was to extract the accession numbers of the plastid genomes fulfilling the given requirements from NCBI and visualize their release dates.

Main results of the project: The total number of entries for our query on GenBank rose much faster than the linked entries in the RefSeq database.

Personal key learnings:

Sina & Christina: To use Entrez Direct in the bash.

Swetha: concepts of data mining steps in NCBI

Gokul: understood methods of query building, how for loop is used

Estimation of the time:

Sina: 14 hours

Christina: 15 hours

Swetha: 5 hours

Gokul: 6 hours

Project evaluation: 4

Number of words: 1137

1 Scientific Background

As a large bioinformatics facility, NCBI has invested in developing robust process flows to generate annotation and perform quality assurance tests for eukaryotic and prokaryotic genomes, transcripts, and proteins [1].

Direct submissions from researchers, national and international collaborations or agreements with data providers and research consortia, and internal curation activities are all sources of data for NCBI. NCBI, for example, operates the GenBank database and is a collaborator in the International Nucleotide Sequence Database Collaboration (INSDC) alongside the EMBL-EBI European Nucleotide Archive (ENA) [2] and the DNA Data Bank of Japan (DDBJ)[3].

Entrez is an integrated database retrieval system that gives users access to 34 databases with a total of 3 billion records. On the Entrez global search page, there are links to the web portals for each of these databases [4]. The E-utilities are an Application Programming Interface for Entrez functions, with extensive documentation [5].

2 Goal

For this project, the NCBI database was searched for nucleotide sequences that recorded complete and validated plastid genomes, had an length between 50,000

and 250,000 base pairs, and were first released in 2020. The entries were saved in GenBank's nucleotide section or NCBI RefSeq, while NCBI SRA contained the raw reads of the sequences. The release dates of this sequences were extracted and visualized.

3 Data

We used three databases, the most important being GenBank. The data gathered from GenBank was further verified by comparing it to RefSeq and the sequence read archive (SRA). All three databases are part of the database resources of the National Center for Biotechnology Information [6].

3.1 GenBank

GenBank is a public database containing nucleotide sequences for about 260,000 species. Laboratories worldwide can submit their own sequencing results [7]. This project focused on complete and verified records of plastid genomes with sequence lengths ranging from 50,000 to 250,000 base pairs, which were released within 2020.

3.2 NCBI RefSeq

Another part of the NCBI databank is the RefSeq database. This database is a collection of curated, stable, and non-redundant reference sequences [8]. Since RefSeq is a collection of genomic, transcript and protein sequence records from other public archives, the records gathered from GenBank are part of RefSeq. Therefore, a comparison was conducted.

3.3 The Sequence Read Archive (SRA)

The Sequence Read Archive (SRA) is an international public archival resource for next-generation sequence data established under the guidance of the International Nucleotide Sequence Database Collaboration (INSDC) [9]. The SRA is a repository of raw sequence data with the aim to balance the cost of long-term archival with the requirement to store sufficient information to support re-use of the submitted data [10].

4 Methods

The data was extracted from the GenBank nucleotide database using a shell script in the unix bash [11]. The results were visualized using R [12].

4.1 Sequence Data Mining

The preliminary step for data mining is to retrieve the data from the NCBI repository by building the a query. This query consists of several constraints. That way the number of records is reduced and more feasible to work with.

In this case, the query filters for complete plastid genomes. Since some researchers use chloroplasts synonymous, both are included in the query. To strengthen these constraints, unverified or partial sequence records are excluded. Further, we filter by date. Only sequence records from 2020 are imported. The last necessary restriction is the sequence length. Sequences are longer than 50,000 and shorter than 250,000 base pairs.

In the beginning we counted the number of records, this query applies to. For this the `esearch` utility [13] and `xtract` was used. Next, the accession numbers of all records were written into a file. This file was utilized to find the corresponding records in the RefSeq database and in the NCBI SRA database. Lastly, the dates of each record were accessed and saved.

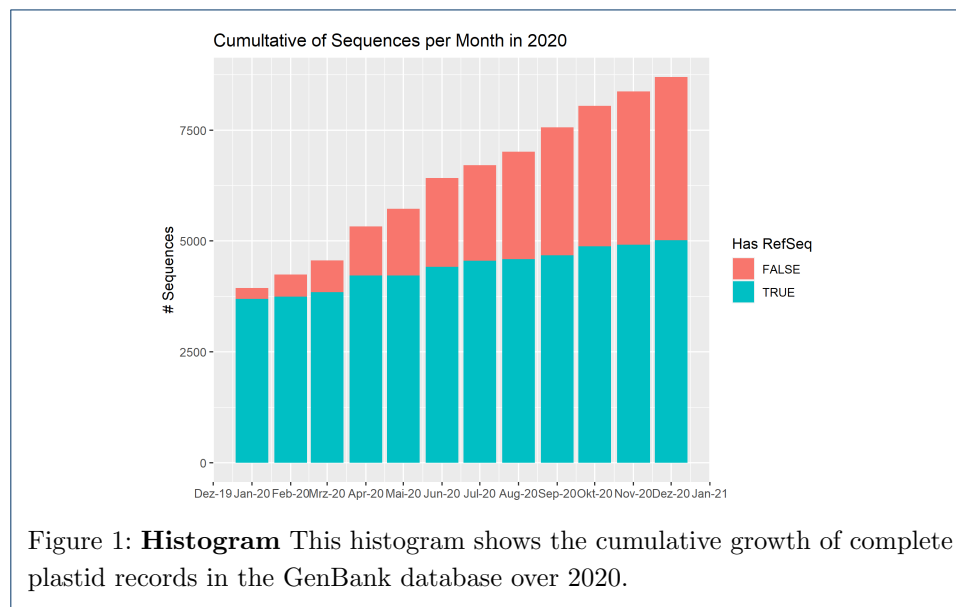
4.2 Data Visualization

This data was then visualized using R [12]. First, a file containing accession number and date is read. Every date is transformed into the year-month-form with the help of `lubridate` [14]. Lastly, the number of new sequences in every month as well as the cumulative growth of the absolute number of sequences is visualized in a histogram. For this, the tidyverse package `ggplot` [15] was utilized.

5 Results

With the help of `esearch`, 5016 records fitting to the criteria could be found. All of them were unique. These records could be linked to 1331 RefSeq records. It is noticeable, that only sequences starting with "NC_" are part of RefSeq.

The identification of SRA records proved to be more difficult. Several different approaches were used, but no connection was found. The team tested accessing the records by starting a similar query on SRA for raw reads instead of whole genomes. The SAMN id and the taxon number of each found record were then used to find a connection to the nucleotide database. No conclusive results were found. Another approach was to use the known sequence records and connect them to their SRA records. However, no overlap was found.



The data was then visualized in a histogram for each month (Figure 1). The number of sequence records rises at a constant rate until in December a cumulative sum of 5016 records are included, while the number of RefSeq records barely rose after January.

6 Discussion

Many problems occurred during our attempts to include the SRA in our data mining. A multitude of different approaches failed during computation, and the two successful runs provided insufficient results. The struggle with this process showed a difficulty of data mining.

That is, what this weeks project mainly focused on, the data mining part of a data scientists work. It is important to know how to retrieve the required data from different databases, in this case GenBank, NCBI RefSeq, and NCBI SRA. In a minor task data visualization was included, which is also often demanded in this field of work.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

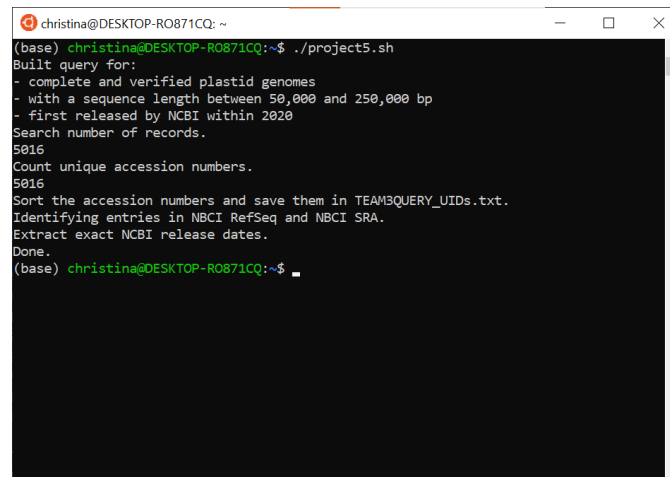
Sina Glöckner and Christina Kirschbaum implemented the shell script and visualized the results.

Swetha Rose Maliyakal Sebastian and Gokul Thothathri informed about the scientific background and wrote the report.

References

- Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A.: Genenames.org: the HGNC resources in 2015. *Nucleic Acids Research* **43**(D1), 1079–1085 (2014)
- Harrison, P., Ahamed, A., Aslam, R., Alako, B., Burgin, J., Buso, N., Courtot, M., Fan, J., Gupta, D., Haseeb, M., Holt, S., Ibrahim, T., Ivanov, E., Jayathilaka, S., Balavenkataraman Kadhirvelu, V., Kumar, M., Lopez, R., Kay, S., Leinonen, R., Cochrane, G.: The european nucleotide archive in 2020. *Nucleic Acids Research* **49** (2020). doi:[10.1093/nar/gkaa1028](https://doi.org/10.1093/nar/gkaa1028)
- Ogasawara, O., Kodama, Y., Mashima, J., Kosuge, T., Fujisawa, T.: Ddbj database updates and computational infrastructure enhancement. *Nucleic acids research* **48** (2019). doi:[10.1093/nar/gkz982](https://doi.org/10.1093/nar/gkz982)
- Baxeavanis, A.: Searching the ncbi databases using entrez. *Current protocols in bioinformatics* / editorial board, Andreas D. Baxeavanis ... [et al.] **Chapter 1**, 1–3 (2006). doi:[10.1002/0471250953.bi0103s13](https://doi.org/10.1002/0471250953.bi0103s13)
- Entrez Direct: Programming Utilities Help. <https://www.ncbi.nlm.nih.gov/books/NBK25501/>
- Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **41**(Database issue), 8–20 (2013). doi:[10.1093/nar/gks1189](https://doi.org/10.1093/nar/gks1189). Accessed 2021-11-27
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: GenBank. *Nucleic Acids Res* **39**(Database issue), 32–37 (2011). doi:[10.1093/nar/gkq1079](https://doi.org/10.1093/nar/gkq1079)
- O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D.: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**(D1), 733–745 (2016). doi:[10.1093/nar/gkv1189](https://doi.org/10.1093/nar/gkv1189)
- Cochrane, G., Karsch-Mizrachi, I., Nakamura, o.b.o.t.I.N.S.D.C. Yasukazu: The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research* **39**(suppl_1), 15 — —18(2010)
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079 (2009)
- GNU, P.: Free Software Foundation. Bash (3.2. 48)[Unix shell program] (2007)
- R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Kans, J.: Entrez Direct: E-utilities on the Unix Command Line. National Center for Biotechnology Information (US), ??? (2021). Publication Title: Entrez Programming Utilities Help [Internet]. <https://www.ncbi.nlm.nih.gov/books/NBK179288/> Accessed 2021-11-25
- Grolemund, G., Wickham, H.: Dates and times made easy with lubridate. *Journal of Statistical Software* **40**(3), 1–25 (2011)
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H.: Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686 (2019). doi:[10.21105/joss.01686](https://doi.org/10.21105/joss.01686)

Figures

A screenshot of a terminal window titled 'christina@DESKTOP-RO871CQ: ~'. The terminal shows the execution of a script './project5.sh'. The script's output includes a list of search criteria, the number of records found (5016), and a series of processing steps like counting unique accession numbers and sorting them. The terminal ends with a prompt for the user to press a key.

```
christina@DESKTOP-RO871CQ: ~  
(base) christina@DESKTOP-RO871CQ:~$ ./project5.sh  
Built query for:  
- complete and verified plastid genomes  
- with a sequence length between 50,000 and 250,000 bp  
- first released by NCBI within 2020  
Search number of records.  
5016  
Count unique accession numbers.  
5016  
Sort the accession numbers and save them in TEAM3QUERY_UIDs.txt.  
Identifying entries in NCBI RefSeq and NCBI SRA.  
Extract exact NCBI release dates.  
Done.  
(base) christina@DESKTOP-RO871CQ:~$ █
```

Figure 2: **Screenshot Terminal.** The screenshot shows the run of the script.