Published on Jul 19, 2023 by Arcadia Science **DOI**: 10.57844/arcadia-w8xt-pc81

Clustering the NCBI nr database to reduce database size and enable faster BLAST searches

The increasingly large number of sequences available in public databases makes searches slower and slower. We clustered the NCBI non-redundant protein database and calculated taxonomic info for each cluster. This collapses similar sequences and reduces the database by over half.

Contributors (A-Z)

Feridun Mert Celebi, Jonathan A. Eisen, Megan L. Hochstrasser, Taylor Reiter

Version 2 · Mar 31, 2025

Purpose

NCBI recently created a clustered nr protein database that reduces search times and returns richer taxonomic diversity for matched sequences. This database, however, is currently only searchable through NCBI's online BLAST interface as it won't be available for download until fall 2023 **[1]**. To capture these benefits for local searches, we made an equivalent database. We performed the clustering as indicated by NCBI and created an accompanying taxonomy file that annotates the taxonomic identifier for each cluster member and the lowest common ancestor for each protein cluster. We

are sharing the database to help others doing BLAST protein searches outside of the online interface and hope the added taxonomy files are in a useful format to support a variety of use cases.

- This pub is part of the **platform effort**, "Software: Useful computing at Arcadia." Visit the platform narrative for more background and context.
- All associated code is available in this GitHub repository.
- You can download our clustered nr database and accompanying taxonomy files from the OSF.

The motivation

As sequencing costs continue to decrease, the size of sequencing data repositories like GenBank continue to grow. GenBank contained over 2.5 billion nucleotide sequences from 504,000 species as of December 2021 **[2]**. As sequencing databases get larger, searching these databases becomes increasingly expensive.

One way to reduce search times and maintain diversity in search results is to cluster the database using sequence similarity. This strategy is especially useful for protein sequences, which compared to other units of sequencing (chromosomes, genomes, etc.), are relatively short and maintain similarity across large evolutionary time scales. When applied to protein sequences, clustering collapses similar sequences to a representative sequence while maintaining the diversity of sequences in the original database.

In May 2022, the National Center for Biotechnology Information (NCBI) introduced an experimental database, ClusteredNR, that is available via the NCBI BLAST web server for BLASTp and BLASTx searches **[3][4]**. The clustered database is faster to search and more taxonomically and functionally balanced given that GenBank itself has many overrepresented organisms and classes of proteins. NCBI generated ClusteredNR by clustering the standard nr database at 90% length and 90% identity using MMseqs2 **[5]**.

The problem

NCBI's ClusteredNR database is not yet downloadable, meaning it cannot yet be used for local searches. For many of the pipelines in development at Arcadia, we sought a faster alternative to BLAST nr searches. We also hoped to capture taxonomically diverse results even when filtering to top hits. In trials on the NCBI web server, we were pleased with search times and results using the experimental ClusteredNR database and thus sought to recreate this resource for local usage.

In our communication with them, the NCBI National Library of Medicine's (NLM) support team (nlm-support@nlm.nih.gov) mentioned that they are preparing the ClusteredNR for users to download. The main challenge they are facing is finding a way to let users view the contents of these clusters directly through the command line interface. In response to this, we thought carefully about and shared our needs regarding cluster members and taxonomic labels. We hope that this feedback will be useful to the NCBI team.

Our solution

Following NCBI's lead with ClusteredNR, we created a clustered nr database using the same parameters (90% length and 90% identity) and algorithm (MMseqs2 `mmseqs easy-linclust`) **[5]**. This produces a FASTQ file of representative sequences that one can use as a BLAST database for algorithms like NCBI's BLAST **[6]** or DIAMOND ***[7]***. We also calculated and recorded the lowest common ancestor — the deepest taxonomic lineage shared by all cluster members — for each cluster to supply taxonomic lineages for each sequence in the database. We created an SQLite database from this file to decrease search times for each protein identifier, as the indexed organization and optimized query processing of SQLite enable faster data retrieval than linear searches in flat files.

> The **Snakemake pipeline [8]** we used to build this database is available on GitHub (DOI: 10.5281/zenodo.7998838). You can download the **database files** from OSF (DOI: 10.17605/osf.io/tejwd).

The resource

Clustering the nr database

We started by downloading the FASTA file for the NCBI nr database (downloaded on March 13, 2023; click here to download). We then used `mmseqs easy-linclust` with parameters `--min-seq-id 0.9 -c 0.9 --similarity-type 2 --cov-mode 1` to cluster the FASTA file **[5]**. We selected these parameters to match those used by NCBI **[3]**. This process produced a FASTA file with representative cluster sequences, as well as a TSV file recording cluster membership for each protein identifier. The input FASTA file contained 533,074,657 sequences, and the clustered file contained 239,387,198 sequences, reducing the FASTA file size from 144 GB to 59 GB. This constitutes a 55% reduction in the number of sequences in the database and a 58% reduction in the file size.

Calculating the lowest common ancestor for each protein cluster

To calculate the lowest common ancestor for each protein cluster, we started by downloading the NCBI taxdump (click here to download) and prot.accession2taxid files (click here to download). For each prot.accession2taxid file, we used csvtk `join` with the flag `--left-join` to join the protein accession recorded in the cluster membership TSV file output by MMseqs2 to its NCBI Taxonomy identifier. While NCBI provides a single prot.accession2taxid file, we used the segmented version to reduce the RAM required to run the csvtk `join` commands. We combined the joined files together and used csvtk `fold` **[9]** to collapse the taxonomy identifiers for each representative sequence into a single line, each separated by a semi-colon. We then used taxonkit `lca` with flag `--buffer-size 1G` to calculate the taxonomic identifier for the lowest common ancestor for each representative sequence and reformatted the taxonomic identifier into a named lineage using taxonkit `reformat` **[10]**. We chose to report superkingdom, kingdom, phylum, class, order, family, genus, species, and strain. Lastly, we built an SQLite database from this TSV file to facilitate fast retrieval of taxonomic lineages from protein identifiers.

Hosting the database

We provide our clustered nr database and accompanying lineage files for download in an OSF repository **[11]**.

We emphasize that this is a temporary solution. We anticipate that the NCBI will release their ClusteredNR database for download in fall 2023, and at that time, plan to consider this database deprecated. The NCBI is better positioned to routinely integrate new sequences into their databases and to adapt their distributed files to a wider variety of use cases.

Key takeaways

- Inspired by the NCBI's experimental ClusteredNR database that is available for BLAST searches on the BLAST web interface, we built a clustered nr database that we can use locally.
- Clustering the database reduced the number of sequences in the database by 55% and the database file size by 58%.
- We provide accompanying taxonomy files in TSV and SQLite format. They report the taxonomic identifier and named lineage of the lowest common ancestor of each cluster.
- The database is available for download in this OSF repository.

Next steps

We anticipate that NCBI will release their ClusteredNR database for download soon. When this happens, we'll evaluate the two databases and plan to mark our database as superseded. We will then update our pipelines to use the NCBI version.

NCBI has requested that questions or feedback related to their ClusteredNR database be directed to blast-help@ncbi.nlm.nih.gov. We hope that if you find the database presented in this pub useful, or if it is missing a key element for your use case, that you communicate that both with us and to the NCBI.

Acknowledgements

We would like to thank Scott D. McGinnis for his responses regarding NCBI's ClusteredNR database, and Milot Mirdita and Wei Shen for their fast and helpful responses to GitHub issues.

References

1. NCBI Staff. (2023). Comment on "New ClusteredNR database: faster searches and more informative BLAST results." https://ncbiinsights.ncbi.nlm.nih.gov/2022/05/02/clusterednr_1/

2. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I. (2021). GenBank. https://doi.org/10.1093/nar/gkab1135

3. (2023). New ClusteredNR database: faster searches and more informative BLAST results. https://ncbiinsights.ncbi.nlm.nih.gov/2022/05/02/clusterednr_1/

4. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. (2008). NCBI BLAST: a better web interface. https://doi.org/10.1093/nar/gkn201

5. Steinegger M, Söding J. (2018). Clustering huge protein sequence sets in linear time. https://doi.org/10.1038/s41467-018-04964-5

6. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I. (2013). BLAST: a more efficient report with usability improvements. https://doi.org/10.1093/nar/gkt282

7. Buchfink B, Reuter K, Drost H-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. https://doi.org/10.1038/s41592-021-01101-x

8. Köster J, Rahmann S. (2012). Snakemake—a scalable bioinformatics workflow engine. https://doi.org/10.1093/bioinformatics/bts480

9 Shen W. csvtk - a cross-platform, efficient and practical CSV/TSV toolkit. https://github.com/shenwei356/csvtk

10 Shen W, Ren H. (2021). TaxonKit: A practical and efficient NCBI taxonomy toolkit. https://doi.org/10.1016/j.jgg.2021.03.006

11 Foster ED, Deardorff A. (2017). Open Science Framework (OSF). https://doi.org/10.5195/jmla.2017.88