TEAM - 2

NON-REDUNDANT DATASET CONSTRUCTION FOR RSV GENOME SEQUENCES

1. Novelty:

Human respiratory Syncytial virus is a major cause of lower respiratory tract infections in infants and elderly people. High mutation rates complicate the development of vaccine and antiviral development. We used full genome sequences instead of partial Sequences and applied k-mer based cosine similarity clustering instead of using only blast. We generated a large non-redundant dataset whose size is greater than 500.Training a LSTM model to predict the future predictions.Perform protein stability analysis to check whether the mutation is stabilizing the virus or destabilizing the virus focusing on a specific gene(F-protein).

2. Source:

The RSV genome sequences used in this study were obtained from the National Center for Biotechnology Information (NCBI) public database. All sequences were downloaded in FASTA format. The initial sequences retrieved are greater than 5000 and the genome length is approximately 15000.

3. Preprocessing:

The dataset we took, consisted full length genome sequences of Respiratory Syncytial Virus (RSV). We did Preprocessing to ensure biological validity and data quality before redundancy reduction.
Steps performed:

- Length Filtering: We retained sequences between 14000 and 16000 base pairs. This helped us exclude incomplete or unusually long genome sequences that could introduce noise into the analysis.
- Ambiguity Filtering: We removed sequences with more than 10% ambiguous bases (N).
- Exact Duplicate Removal: Removed identical sequences to eliminate computational redundancy.

After completing all these preprocessing steps, we obtained 5,465 high-quality RSV genome sequences, which were then used for the clustering stage.

4. Non-Redundant Dataset Construction:

To reduce sampling bias and prevent overfitting during model training, we constructed a non-redundant dataset using an alignment-free clustering approach based on k-mer frequency patterns and cosine similarity.
Methodology:

- We converted each genome into a numerical feature vector using 6-mer frequency counts. These 6-mers represent all possible nucleotide combinations of length six and capture mutation patterns across the genome.
- We then calculated cosine similarity between the k-mer vectors to measure how similar two sequences were. If the similarity score was 0.995 or higher, the sequences were

grouped into the same cluster using a greedy clustering strategy. From each cluster, we retained one representative sequence.

We chose this alignment-free method instead of traditional alignment-based tools because it captures overall mutation distribution patterns, scales efficiently to large genomic datasets, and generates numerical feature representations that are well-suited for deep learning models.

5. Justification for k-mer based Genome representation:

The dataset that is being used for this project has more than 5000 sequences of length 15000 performing a pairwise alignment based similarity such as BLAST on such large datasets becomes computationally huge as the comparisons increases quadreatically with the number of sequnces. To overcome the limitation of alignment free k-mer frequency representation we tried using k-mer representation where each genome sequence was decomposed into overlapping substrings of length k=6.Smaller k values lack sufficient specificity while the large k values generate extreme sparse high dimensional vectors.

Advantages of k-mer representation:

- Alignment-free and computationally efficient
- Scalable for thousands of viral genomes
- Captures mutation-driven compositional changes
- Suitable for downstream deep learning models

6. Justification for cosine similarity:

After converting each genome into normalized k-mer frequency vector,pairwise similarity between sequences was computed using cosine similarity.

- **Length Normalization:** The RSV genomes are approximately 15000b,minor length variations exist. To overcome those minor length variations Cosine similarity is introduces which normalizes vector magnitude, ensuring fair comparison.
- **Pattern-Based Similarity:** As we are interested in finding out the similarity of mutation patterns than the absolute k-mer counts. This Cosine similarity focuses on proportional similarity.
- **Computational Efficiency:** Cosine Similarity can be easily computed using matrix operations which makes it more efficient for sequences which have length over 5000.
- **Widely Used in Genomic Alignment-Free Studies:** Cosine similarity is more commonly used in alignment-free sequence comparison and text-vector similarity analysis, making it mathematically robust and interpretable.

7. Importance of the Evaluation Metrics Used:

We validated the clustering results using scientific methods:

Redundancy Reduction Rate:

- Public databases like National Center for Biotechnology Information (NCBI) often contain many similar sequences.
- This happens beacause of the same strain may be submitted multiple times by different research groups
- Sometime the sequences differ only in small metadata details.
- In other cases, the samples come from the same outbreak and are almost identical.

- A high redundancy Reduction Rate(RRR) shows that the clustering method successfully removed these repeated genomes At the same time it keeps a set of sequences that still represents the overall diversity.
- In simple terms, RRR indicates how effectively the method eliminates unnecessary duplicates without losing important biological variation.

Intra-cluster Similarity:

- Intra-cluster similarity measures how similar the sequences are within the same cluster.
- A high intra-cluster similarity means:
- The sequences placed in the same cluster are truly very similar to each other.
- Each cluster is consistent and uniform in its composition.
- The similarity threshold used for clustering was chosen appropriately.
- This shows that the clustering method is reliable and does not combine biologically different genomes into the same group.

Inter-Cluster Similarity:

- Inter-cluster similarity measures how similar sequences are between different clusters.
- A lower inter-cluster similarity indicates that:
- The clusters are clearly separated from each other.
- Different genomic variants are kept in separate groups.
- Biologically unrelated genomes are not combined into the same cluster.
- This shows that the clusters are properly separated and that the chosen similarity threshold is strong and reliable.

Shannon Diversity Index:

- Shannon Diversity Index measures how evenly sequences are distributed across different clusters
- Respiratory Syntactical Virus(RSV) in an RNA virus with high mutation rate and significant genetic variation.
- For mutation prediction we used depp learning models like LSTM,it is importatnt that:
  - ➢ The dataset includes diverse strains.
  - ➢ No single cluster dominates the data.
  - ➢ Evolutionary variation is preserved.
- A higher shannon Diversity Index indicates:
  - ➢ More balanced representation of clusters.
  - ➢ Maintained genetic diversity.
  - ➢ Reduced sampling bias.
- This helps the mutation prediction model generalize better and prevents it from overfitting to a single type.

8. Validation of dataset Specificity to RSV:

- To test whether the curated dataset is specific to Human respiratory syncytial virus, additional biological validation was performed.
- All sequences were retrieved using organism-specific queries from NCBI database, ensuring correct taxonomic annotation.

- The conserved RSV genomic organization (NS1–NS2–N–P–M–SH–G–F–M2–L) was verified across representative sequences, confirming characteristic RSV gene structure.
- Similarity based clustering helps us prove that all the sequences that are within the RSV clusters without any unrelated outliers, supporting the homogeneity of the dataset.
- This analysis tells us that this sequence is correct for the specific disease that we have considered for our project.

9. Results:

The initial dataset contained 5,466 RSV genome sequences.

- After preprocessing, 5,465 high-quality sequences remained.
- Using the alignment-free clustering approach, these sequences were grouped into 377 clusters, with one representative genome selected from each cluster. This resulted in a redundancy reduction rate (RRR) of 0.931
- The average intra-cluster similarity was 0.99, showing that sequences within each cluster were nearly identical. In contrast, the average inter-cluster similarity was 0.9201, indicating meaningful differences between clusters.
- We also calculated the Shannon Diversity Index, which was 3.7054. This suggests that even after redundancy reduction, the overall genetic diversity of the dataset was preserved.

Overall, we were able to effectively clean, refine, and structure the dataset in a way that reduced redundancy while maintaining biological diversity, making it suitable for robust model training.