Dataset Collection and Processing

Collecting the Mutation Data:

To begin the project, I needed real mutation information from cancer patients rather than artificial sequences. For this purpose I used the cancer genomics resource cBioPortal. From there I selected the lung adenocarcinoma study belonging to the TCGA PanCancer Atlas collection.

Inside the study page there is a downloads section that provides mutation records in MAF format. After downloading it, I obtained a text file named data_mutations.txt.
This file does not contain DNA sequences. Instead, each row describes a mutation observed in a patient sample. It specifies the affected gene and the nucleotide change in coding DNA notation.

From this dataset I filtered only the genes relevant to lung adenocarcinoma such as TP53, KRAS, EGFR, BRAF, ALK and others. I also kept only SNP mutations because they represent single base substitutions and can be reconstructed reliably.

So at this point I had real patient mutation records but still no mutated gene sequences.

Getting the Normal Gene Sequences:

Since the mutation file only tells what changed and not the full sequence, I needed the normal gene sequence as a starting point.

For each gene I searched in the NCBI nucleotide database using its RefSeq ID and downloaded only the CDS (coding sequence) in FASTA format.
This sequence represents the healthy version of the gene and acts as the reference sequence.

Now I had two types of data:

- Mutation coordinates from TCGA
- Normal gene sequence from NCBI

Reconstructing the Mutated Gene:

The mutation dataset gives the position and the nucleotide change.
So I recreated the tumor gene by applying that change on the normal gene.

For every mutation entry:

1. I located the mutation position in the reference CDS
2. I replaced the original nucleotide with the mutated nucleotide
3. I saved the resulting sequence as a mutated gene

This process was repeated for all mutation entries.
Therefore the sequences were not generated randomly — they were reconstructed from real mutation records.

Identifying Mutation Positions:

To record the mutation information in a structured way, each mutated sequence was compared with the normal sequence nucleotide by nucleotide.

The position where the difference first appears was stored along with:

• original base
• mutated base

This produced a clean mutation dataset for analysis.

Redundant Sequence Removal using Cosine Similarity

After reconstructing mutated gene sequences for each patient, many sequences were found to be highly similar because different patients can share identical or near-identical mutation patterns. Such sequences do not contribute additional biological information and may bias downstream analysis.

To address this, each sequence was converted into a numerical vector based on nucleotide composition (frequency of A, T, G and C). Cosine similarity was then calculated between every pair of mutated sequences.

If the similarity between two sequences was greater than 0.80, they were considered redundant and only one representative sequence was retained. Sequences with similarity below this threshold were preserved as unique mutation profiles.

This step ensured that the final dataset contained biologically distinct gene variants rather than repeated mutation patterns, thereby improving the reliability of alignment scoring and comparative analysis.

Pairwise Alignment:

After reconstructing the variants, each mutated gene was aligned with its reference gene.

I used global alignment because the entire gene length needs to be compared.
Local alignment would only compare a matching region, but here the sequences differ only at specific positions while the rest of the gene remains the same.

So comparing the full sequence provides a proper similarity measure.

Alignment Score Calculation:

The alignment score was calculated using a simple similarity count:

match = +1
mismatch = 0
no gap penalty

Since most mutations are single nucleotide substitutions, the score roughly equals the gene length minus the number of changed bases.

This score indicates how similar the mutated gene remains to the normal gene.

Final Result:

Mutation data from the MAF file was mapped onto reference coding sequences of selected *Homo sapiens* cancer genes.
For each patient, all reported nucleotide substitutions were incorporated into a single reconstructed sequence. This produced realistic multi-mutation gene sequences instead of treating every mutation independently.

Sequences that did not contain valid nucleotide substitutions were ignored during reconstruction.

After reconstruction and redundancy filtering:

- Each sequence corresponds to a unique patient mutation pattern
- Duplicate mutation profiles were removed
- The dataset preserves mutation diversity instead of mutation repetition
- Sequences remain biologically close to the reference while still representing distinct variants

Novelty of our project:

Instead of treating each mutation as a separate sequence, this project reconstructs one complete mutated gene sequence per patient by combining all mutations occurring in that patient. This better reflects real biological conditions where multiple mutations exist together in the same gene. After reconstruction, cosine similarity is used to remove repeated mutation patterns and keep only distinct mutation profiles. Finally, pairwise alignment is applied to verify that the sequences are valid variations of the reference gene. Thus, the project produces a realistic, non-redundant patient-level mutation dataset rather than a simple list of mutations.