
Name: Arunannamalai Sujatha Bharath Raj

Andrew ID: asujatha

Course: Practical Computing for Biologists (03-701)

TP53 Mutation Hotspot Analyzer

Introduction:

The tumor-suppressor gene *TP53* encodes the p53 protein, a key regulator of cell-cycle arrest, apoptosis, and DNA repair. Mutations in *TP53* are among the most frequent alterations in human cancers, compromising the protein's DNA-binding ability and promoting tumor progression. Identifying recurrent "hotspot" mutations helps reveal functional regions of the gene that drive oncogenesis. This project developed a Python-based *TP53 Mutation Hotspot Analyzer* to retrieve, parse, and visualize clinically reported mutations from the NCBI ClinVar database, providing a reproducible computational framework for genomic variant analysis.

Methods and Data Usage:

Data source and format:

Variant data were obtained from the **NCBI ClinVar** public database (<https://www.ncbi.nlm.nih.gov/clinvar>). The dataset used was **variant_summary.txt.gz**, a tab-delimited text file containing annotated variants (fields: GeneSymbol, Assembly, Protein_Change, ClinicalSignificance, and PhenotypeList). The dataset was downloaded directly from: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz.

The *TP53* reference sequence (RefSeq ID: **NM_000546.6**) was also retrieved from NCBI's nucleotide database in **FASTA format** using Biopython's Entrez API.

Analysis workflow:

1. **Sequence retrieval:** The *TP53* coding DNA sequence (NM_000546.6) was fetched via Biopython and translated into the protein sequence to serve as a reference for mutation mapping.

2. **Variant extraction:** Using `pandas.read_csv()`, the ClinVar summary file was filtered for *TP53* variants corresponding to the **GRCh38** genome assembly.
3. **Data cleaning:** Phenotype entries were standardized by removing database identifiers (MedGen, OMIM, MONDO, etc.) and excluding non-informative rows labeled “not provided” or “NA.”
4. **Variant parsing:** HGVS notations beginning with p. were parsed using regular expressions to extract amino-acid positions and classify variants as *missense*, *nonsense*, *frameshift*, or *other*.
5. **Visualization and aggregation:** Using matplotlib, variant counts were grouped by amino-acid position and visualized as a **lollipop plot** to highlight frequently mutated codons. Known cancer hotspots (codons 175, 248, and 273) were annotated for reference.
6. **Statistical summary:** The frequency of each mutation class and the top five associated cancer phenotypes were tabulated to quantify variant distributions across clinical conditions.

Results:

After filtering, approximately **3,000 unique TP53 variants** were analyzed. **Missense mutations** comprised the majority ($\approx 65\%$), followed by **frameshift** and **nonsense** mutations. The mutation frequency plot revealed pronounced peaks at **codons 175, 248, and 273**, aligning with canonical *TP53* hotspots in the DNA-binding domain known to disrupt p53’s tumor-suppressive function. Phenotype analysis showed strong associations with **Li-Fraumeni syndrome**, **Hereditary cancer-predisposing syndrome**, and **Breast and Ovarian cancers**. These results replicate published mutation patterns, validating the computational pipeline.

Conclusion:

In conclusion, the *TP53 Mutation Hotspot Analyzer* effectively integrates database retrieval, variant parsing, and visualization to identify biologically meaningful hotspots. The workflow can be adapted to analyze other cancer driver genes and supports data-driven exploration in precision oncology.