GPCR Database

An algorithm to automatically clean and identify protein paralogs in human GPCRs from data extracted from UniProt

**Script 1: Generate Protein Data**

This script processes a FASTA file containing protein sequences and tags, creating a new formatted file where sequences have no whitespace. It then generates three output files (protein\_table2.csv, protein\_pos\_table2.csv, variant\_table2.csv) with data for database upload, including information about proteins, their positions, and associated variants. The script extracts relevant information from the input files and formats it into the specified CSV files.

**Script 2: Generate Variant Data**

This script takes a protein position file and a variant file as input, generating a new CSV file (variant\_test\_out.csv) with variant data. It matches protein positions and variants based on UniProt IDs and positions, extracting relevant information about each variant and writing it to the output file.

**Script 3: Identify Variants**

This script analyzes damaging variants in conserved regions of proteins. It takes as input a file of UniProt IDs, a conservation file, and a variant file. The script calculates the total number of damaging variants, identifies conserved positions using conservation scores, and counts damaging variants located in conserved regions. The output includes statistics on the total number of damaging variants and the percentage located in conserved regions for each UniProt ID.