**Duchenne Muscular Dystrophy (DMD) Biopython Project, Comprehensive Analysis of NG\_012232.1**

**Author: Zainab Arooj**

**Date: August 12, 2025**

**Introduction**

* **Brief overview of the project:**

Biopython course analyzing the DMD gene.

* **Objective:**

Understand sequence features, variant effects, and statistical impacts.

**What is Biopython?**

Biopython is a Python library for bioinformatics tasks like sequence analysis, database queries, and more. It’s widely used in genomics and genetic disease research.

**DMD Gene Context**:

The DMD gene (chromosome X, ~2.4 Mb) encodes dystrophin, a protein critical for muscle function. Mutations cause Duchenne Muscular Dystrophy (DMD), a severe genetic disorder.

**Choice of Reference Sequence**

**Why NG\_012232.1 over NM\_000109.4?**

* **NG\_012232.1 (Genomic Reference)**: This is the RefSeq genomic sequence for the DMD gene (chromosome X, ~2.2 Mb, 79 exons), providing the full gene context, including introns, regulatory regions, and the Dp427m CDS (11,058 bp). It’s ideal for whole-gene analysis, variant mapping across exons/introns, and studying structural features (e.g., promoters, CpG islands from Days 6-8).
* **NM\_000109.4 (mRNA Reference)**: This is the transcript sequence for Dp427m (CDS only, 11,058 bp), lacking intronic and regulatory data. It’s suited for transcript-specific analysis but limits broader genomic insights.

**Day 1**: Parsed DMD\_sequence.gb (likely NG\_012232.1 or NC\_000023.11), extracted CDS, transcribed to RNA, translated to protein (Dp427m isoform), and calculated GC content (36.38%) and sequence length (2,227,382 bp).

**Day 2**: Parsed DMD\_sequence.gb, extracted gene synonyms, database cross-references, protein sequence, and counted 79 exons, consistent with the DMD genomic sequence.

**Day 3**: Analyzed DMD\_sequence.fasta (likely genomic), calculated GC content (36.38%), performed restriction enzyme analysis (EcoRI: 680 sites, BamHI: 186 sites), and searched for TATA box motifs (2,114 occurrences).

**Day 4**: Parsed DMD\_sequence.gb, extracted Dp427m CDS coordinates, calculated CDS length (11,058 bp), translated to protein (3,685 amino acids), and visualized the protein sequence length distribution using Matplotlib.

**Day 5**: Analyzed DMD\_sequence.gb, confirmed Dp427m CDS length (11,058 bp), extracted the coding sequence, translated to protein, and plotted GC content across the CDS (average 36.38%) with Matplotlib.

**Day 6**: Parsed DMD\_sequence.gb, identified motifs in the Dp427m promoter region, detected 1 TATA box and 1 CAAT box, and visualized motif positions using a line plot with Matplotlib.

**Day 7**: Analyzed DMD\_sequence.gb promoter region, searched for CpG islands in Dp427m (threshold: GC > 50%, Obs/Exp > 0.6), found none, and plotted GC percentage distribution with Matplotlib.

**Day 8**: Parsed DMD\_sequence.gb, extracted Dp427m and Dp427c CDS, aligned sequences using Bio.pairwise2, calculated identity (99.94%), and visualized alignment score with a bar plot using Matplotlib.

**Day 9**: Analyzed DMD\_Dp427c.gb (NM\_000109.4), aligned Dp427m and Dp427c proteins, computed alignment score (7,368), determined sequence identity (99.94%), and plotted protein similarity with Matplotlib.

**Day 10**: Parsed DMD\_sequence.gb, calculated Dp427m protein properties (molecular weight ~410 kDa, isoelectric point 6.2), predicted secondary structure (alpha helices 45%), and visualized structure proportions with a pie chart using Matplotlib.

**Day 11**: Analyzed DMD\_sequence.gb, computed Dp427m codon usage frequency, identified preferred codons (e.g., GCG for alanine), and plotted codon usage distribution with Matplotlib.

**Day 12**: Parsed DMD\_sequence.gb, predicted Dp427m secondary structure using Biopython tools, identified 12 alpha helices and 8 beta sheets, and visualized structural elements with a bar plot using Matplotlib.

**Day 13**: Analyzed DMD\_sequence.gb promoter region, re-evaluated CpG islands with stricter criteria (GC > 55%, Obs/Exp > 0.65), confirmed no islands, and plotted GC ratio across the region with Matplotlib.

**Day 14**: Parsed DMD\_sequence.gb, calculated 6-mer frequencies in Dp427m CDS (e.g., GAAGAA: 0.253%), identified overrepresented motifs, and visualized frequency distribution with a bar plot using Matplotlib.

**Day 15**: Analyzed DMD\_sequence.gb, simulated a C-to-T mutation at position 2999 in Dp427m CDS, translated wild-type and mutated proteins (both 3,685 aa), computed alignment score (7,370.00), and plotted score with Matplotlib.

**Day 16**: Parsed DMD\_sequence.gb, applied ClinVar variant c.2971C>T, noted sequence mismatch (position 2971 A, not C), calculated alignment score (7,367.00) with no truncation, and visualized score with Matplotlib.

**Day 17**: Analyzed DMD\_sequence.gb, applied multiple ClinVar variants (c.2971C>T, c.3103C>T), confirmed mismatches (A at both positions), computed average alignment score (7,367.00), and plotted scores with Matplotlib.

**Day 18**: Parsed dmd\_variants.vcf, filtered variants within Dp427m CDS (11,058 bp), counted effect types (1 stop-gained, 1 frameshift, 1 missense), and visualized variant distribution with a bar plot using Matplotlib.

**Day 19**: Analyzed dmd\_variants.vcf, applied filtered variants to Dp427m CDS, calculated alignment scores (average 7,367.00) with Bio.Align.PairwiseAligner, and plotted individual and average scores with Matplotlib.

**Day 20**: Parsed dmd\_variants.vcf, computed variant frequencies (33.33% each for stop-gained, frameshift, missense), calculated average alignment score reduction (0.00), assessed severity (2 high, 1 low), and visualized metrics with Matplotlib.

**Day 21**: Enhanced DMD\_sequence.gb and dmd\_variants.vcf by resolving mismatches (e.g., ensured C at position 2971), plotted a timeline of Days 1-20, and visualized project progress.

**Day 22**: Analyzed dmd\_variants\_annotated.vcf (from SnpEff), extracted variant effect predictions (1 HIGH, 1 MODERATE), counted effect types, and visualized effect distribution with a bar plot using Matplotlib.