PRNP Severity Modeling Using Engineered Features and Protein Language Models

Team Members:

C. Shashank – BL.SC.U4AIE24011

R. Abhitej Reddy – BL.SC.U4AIE24040

Saanvi Verma – BL.SC.U4AIE24044

1. Introduction

Prion diseases are fatal neurodegenerative disorders caused by the misfolding of the prion protein (PrP). The cellular prion protein (PrP^{D}) undergoes conformational change into a misfolded β-sheet rich structure (PrP^{sc}), leading to aggregation and neurotoxicity.

One of the most studied genetic determinants in prion disease is the **PRNP gene**, especially the polymorphism at codon 129 (Methionine/Valine). Variants in PRNP influence disease susceptibility, progression, and phenotype.

The aim of this project was to:

1. Construct a clean, non-redundant human PRNP protein dataset
2. Engineer biologically meaningful sequential and structural features
3. Extract deep contextual embeddings using a pretrained protein BERT model
4. Prepare a tabular dataset suitable for severity modeling

---

2. Data Collection

2.1 Retrieval of PRNP Sequences

We queried the NCBI Nucleotide database using:

```
PRNP[Gene] AND "Homo sapiens"[Organism] AND CDS
```

We downloaded all coding sequences (CDS) in FASTA format.

Initial Dataset Statistics:

- ~11,235 CDS sequences
- Length range: 24 bp – 14,406 bp
- Average length: ~1524 bp

This clearly indicated:

- Many partial fragments
- Some genomic regions
- Redundant submissions

---

3. Data Cleaning and Filtering

3.1 Length Filtering

The expected coding length of human PRNP is ~762 base pairs.

We filtered sequences with:

```
700 ≤ length ≤ 900 bp
```

After filtering:

- 1,314 sequences remained

3.2 Removing Redundancy

We removed duplicate DNA sequences using set-based filtering.

After deduplication:

- 516 unique CDS sequences

3.3 Translation to Protein

All 516 CDS were translated into protein sequences.

Expected PRNP protein length: ~253 amino acids.

We filtered translated sequences to retain only proteins between:

```
240 – 260 amino acids
```

This removed:

- Partial translations
- Frameshift errors
- Truncated proteins

Final Clean Protein Dataset:

- 193 unique human PRNP protein variants

---

4. Dataset Conversion to Tabular Format

The FASTA file was converted into a structured tabular format.

Each row represents one protein sequence.

Initial columns:

- sequence_id
- sequence
- length

Dataset shape after conversion:

```
(193, 3)
```

---

5. Feature Engineering

To model severity, we extracted biologically relevant features reflecting protein stability, aggregation tendency, and physicochemical properties.

5.1 Physicochemical Properties

Using BioPython ProtParam:

- Molecular Weight
- Isoelectric Point (pI)
- Aromaticity
- Instability Index
- GRAVY (Grand Average of Hydropathy)
- Aliphatic Index
- Net Charge

These properties relate to folding stability and hydrophobic exposure.

5.2 Aggregation-Related Features

Prion diseases involve protein aggregation. We engineered features specifically targeting aggregation tendencies:

- Q% (Glutamine fraction)
- N% (Asparagine fraction)
- Longest Q/N stretch
- Hydrophobic stretch count
- Aromatic residue density
- Glycine density
- Cysteine count
- Proline count
- Octapeptide repeat motif count

The octapeptide repeat region is known to influence prion pathogenicity.

5.3 Amino Acid Composition

We added 20 features representing normalized amino acid frequency:

```
AA_A, AA_C, AA_D, ..., AA_Y
```

This captures global compositional bias across variants.

5.4 Structural Proxies

Since experimental structures were not used, we included structural tendency proxies:

- Polar %
- Nonpolar %
- Tiny residue %
- Bulky residue %
- Disorder-promoting residue fraction
- Order-promoting residue fraction

These approximate flexibility and folding tendencies.

5.5 Codon 129 Polymorphism

We extracted the amino acid at position 129 (index 128).

Encoded features:

- residue_129
- is_M129
- is_V129

Codon 129 is strongly associated with disease phenotype and susceptibility.

5.6 Feature Summary

Total engineered features:

~48 features

Dataset shape after feature engineering:

```
(193, ~48+ columns)
```

---

6. BERT-Based Protein Embeddings

Handcrafted features capture global properties but do not model contextual residue interactions.

To address this, we used a pretrained protein transformer model:

Model Used:

facebook/esm2_t33_650M_UR50D

Embedding size: 1280 dimensions

6.1 Embedding Extraction Process

For each sequence:

1. Tokenized the protein sequence
2. Passed it through the model
3. Extracted last hidden state
4. Applied mean pooling across residues

Output:

```
Embedding shape: (193, 1280)
```

Each protein is now represented as a 1280-dimensional contextual vector.

---

7. Final Dataset Construction

We concatenated:

- 48 engineered features
- 1280 BERT embedding features

Final dataset size:

```
(193, ~1328 total features)
```

Saved as:

```
prnp_full_dataset_with_bert.csv
```

---

8. Interpretation of Outputs

Dataset Reduction

Stage	Sequences
Initial CDS	~11,235
Length filtered	1,314
Unique DNA	516
Clean proteins	193

This progressive filtering ensured:

- Removal of fragments
- Removal of redundancy
- Retention of biologically valid variants

Feature Matrix

The final dataset contains:

- Physicochemical properties
- Aggregation indicators

- Structural proxies
- Codon 129 encoding
- Deep contextual embeddings

This creates a multi-layer representation of each PRNP variant.

---

9. Most Relevant Features for Severity Modeling

Based on biological reasoning, the strongest features are:

Primary Determinants

- Codon 129 polymorphism
- Q/N content
- Longest QN stretch
- Hydrophobic stretch count
- Instability index
- GRAVY

Structural Determinants

- Disorder-promoting fraction
- Aromatic density
- Aliphatic index

Deep Contextual Representation

- 1280-dimensional BERT embedding

The BERT embeddings likely encode structural rearrangement potential and mutation context effects that are not captured by simple physicochemical descriptors.

---

10. Current Status and Next Steps

Completed:

- Data collection and filtering
- Translation and cleaning
- Feature engineering
- Codon 129 encoding
- BERT embedding extraction
- Final tabular dataset creation

Next steps (for full severity modeling):

- Incorporate clinical severity labels
- Train supervised models (Random Forest / MLP)
- Perform feature importance analysis
- Evaluate model performance

---

11. Conclusion

We constructed a high-quality human PRNP protein dataset containing 193 unique variants. Through systematic filtering and translation, we ensured biological validity.

We engineered physicochemical and aggregation-relevant features and encoded codon 129 polymorphism. To enhance representational power, we extracted 1280-dimensional contextual embeddings using a pretrained protein transformer.

The final dataset integrates handcrafted biological insight with deep sequence representation, making it suitable for downstream severity prediction once labels are incorporated.