Module 10: BRCA1 Variant Effect Prediction Using Nucleotide Transformer

AI and Biology Perspectives on Variant Effects

Biology: Functional Impact

BRCA1 Variant Study

- **Subject:** Human BRCA1 gene
- **Scope:** Single Nucleotide Variants (SNVs) and their effect on protein function
- **Key Insight:** Mutation **surrounding** sequence is critical for understanding variant effects

AI: Variant analysis

Nucleotide-Transformer (BERT-style)

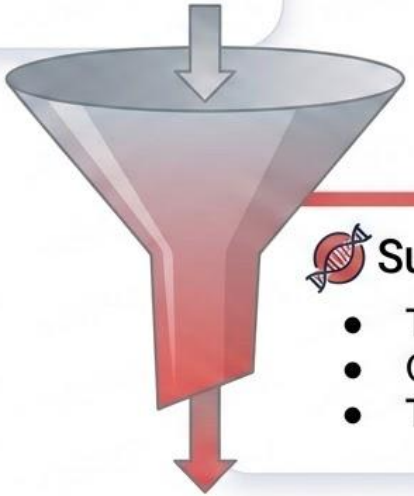
language model trained on genomic sequences, to analyze BRCA1 variants using two complementary approaches:

1. **Zero-shot method:** Direct use of model representations
2. **Supervised method:** Training on specific functional labels

Different approaches of variant analysis

Pretrained Model (General)

- Trained on vast genomic sequences
- Understands DNA “grammar”
- Wide range of applicability

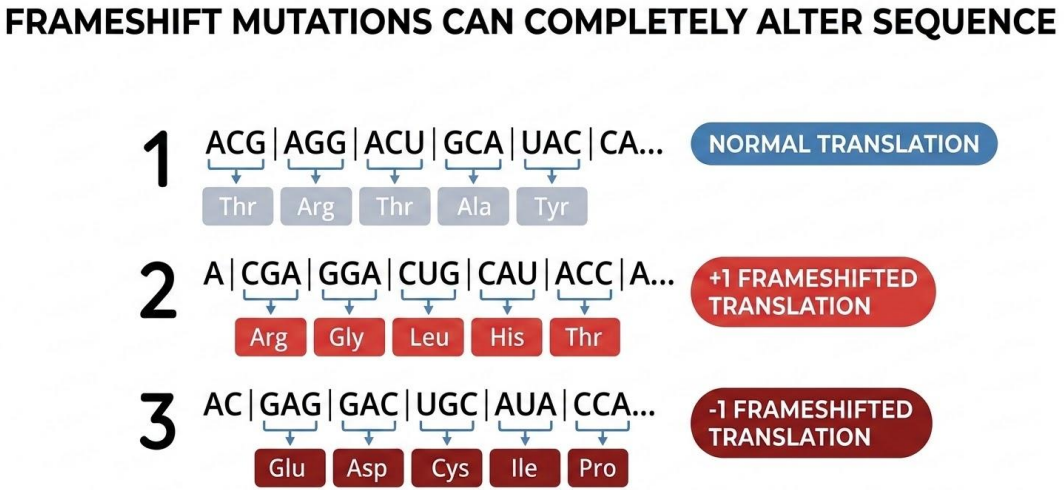


Supervised Model (Specific)

- Trained on specific variant labels
- Optimized for variant effect prediction
- Targeted to BRCA1 function

Intro: Human BRCA1 Gene

- Encodes a protein crucial for DNA repair
- Clinical relevance: variants can cause **loss of function (LOF)** which is associated with significantly increased risk of breast and ovarian cancers



Aim

The Gap:

Understanding genomic context is **essential** to define a variant and its impact on protein function

The AI Solution:

Nucleotide transformer models assist in distinguishing between **loss-of-function** and **benign variants**

Learning Objectives

By the end of this lab, you will understand:

01. How transformer models process genetic sequences
02. Different methods for extracting meaningful sequence representations
03. The difference between zero-shot and supervised approaches in genomics
04. How to evaluate variant effect prediction models

Inputs

Datasets

- **SNVs:** experimentally validated BRCA1 variants from Findlay et al. (2018)
    - Contains function scores for 3,893 SNVs
- **Reference:** DNA sequence (hg19)

Models

Pretrained Model(general)

Supervised Model(specific): (train on labels)

1. logistic regression
2. random forest

Methods: The Nucleotide Transformer

Pre-trained Model

A specialized model designed for genomics:

- Trained on vast amounts of **genomic sequence data**.
- Understands the **"grammar"** of DNA sequences, unlike language models.
- Includes a **tokenizer** (DNA to inputs) and the model itself.
    - Example: Nucleotide-Transformer (BERT-style)

Example: BERT Architecture

Bidirectional Encoder Representations:

- Reads all parts of a sequence **at once** [both left and right directions]
- Captures **deep contextual meaning** within sequences.
- Ideal for **nuanced understanding** of genomic context.

Workflow

01 Setting Up the Nucleotide Transformer Model

02 Loading the BRCA1 Variant Dataset

03 Extracting Genomic Sequences Around Variants

04 Finding the Variant-Affected Tokens

05 Embedding Extraction Methods

06 Supervised Feature Extraction and Evaluation

07 Visualization and Model Evaluation

Key Takeaways

Zero-shot Utility

Embedding deltas (ref vs var) provide a fast, interpretable proxy for variant impact without retraining.

Supervised Performance

Training on concatenated embeddings reliably boosts AUROC (e.g., LR ~0.766 → RF ~0.802).

Best Embedding Strategy

Concatenated (mean-pooled + variant-token) captures global context and local effect for strongest signals.

Tokenization Caveat

k-mer tokenization can split SNV signal—comparing tokenized ref/var is essential for signal integrity.

Context vs Compute

Large windows (4,096 bp) improve context; cache embeddings to avoid repeated transformer calls.

Practical Pipeline

Extract sequences → find variant token → compute embeddings → store combined scaled deltas.

Limitations: Overlap between LOF and FUNC/INT remains; pretraining biases and label noise limit clinical certainty