Transformer-Based DNA Sequence Similarity Search

GROUP
Bhavinkumar V Kuwar [MT24212]
Spoorthy BB [MT24228]

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Problem Statement

Traditional DNA sequence comparison methods like BLAST are slow and often fail to detect functional similarities in divergent sequences. They rely on alignment, which is computationally expensive and misses relationships when sequence similarity is low.

Project Overview

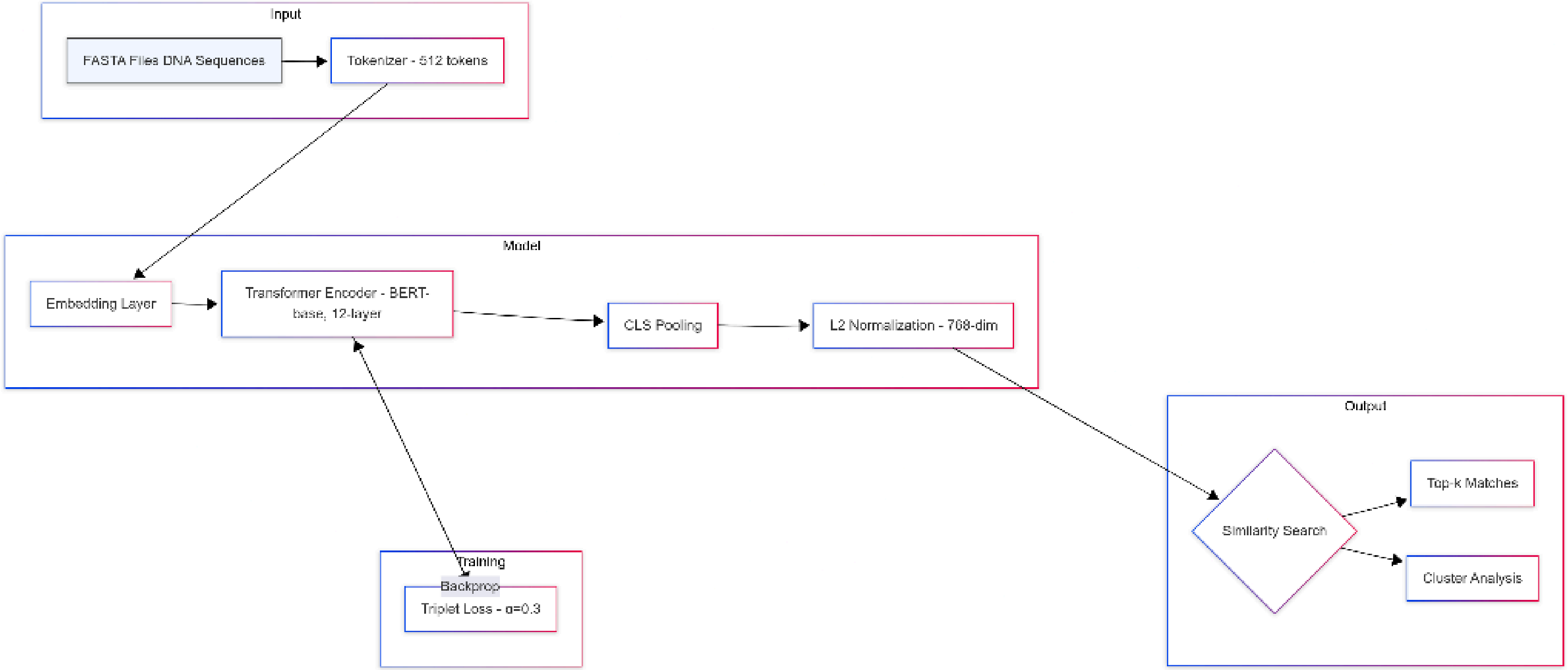
- Problem: Traditional BLAST alignment is slow for large databases

- Solution: Neural embeddings enable O(1) similarity searches.

Key Innovation:

- Transformer architecture (BERT-base)
- Triplet loss optimization (α = 0.3)
- L2-normalized embeddings

System Architecture



Input

FASTA Files DNA Sequences

Tokenizer - 512 tokens

Model

Embedding Layer

Transformer Encoder - BERT-base, 12-layer

CLS Pooling

L2 Normalization - 768-dim

Training

Backprop

Triplet Loss - α=0.3

Output

Similarity Search

Top-k Matches

Cluster Analysis

Key Algorithms

- Triplet Loss:
    - $L = max(||f(a) - f(p)||2 - ||f(a) - f(n)||2 + \alpha, 0$

- Embedding Similarity:
    - $sim(x, y) = x \cdot y / ||x|| \cdot ||y||$

- Recall@k:
    - *# correct in top k / total queries*

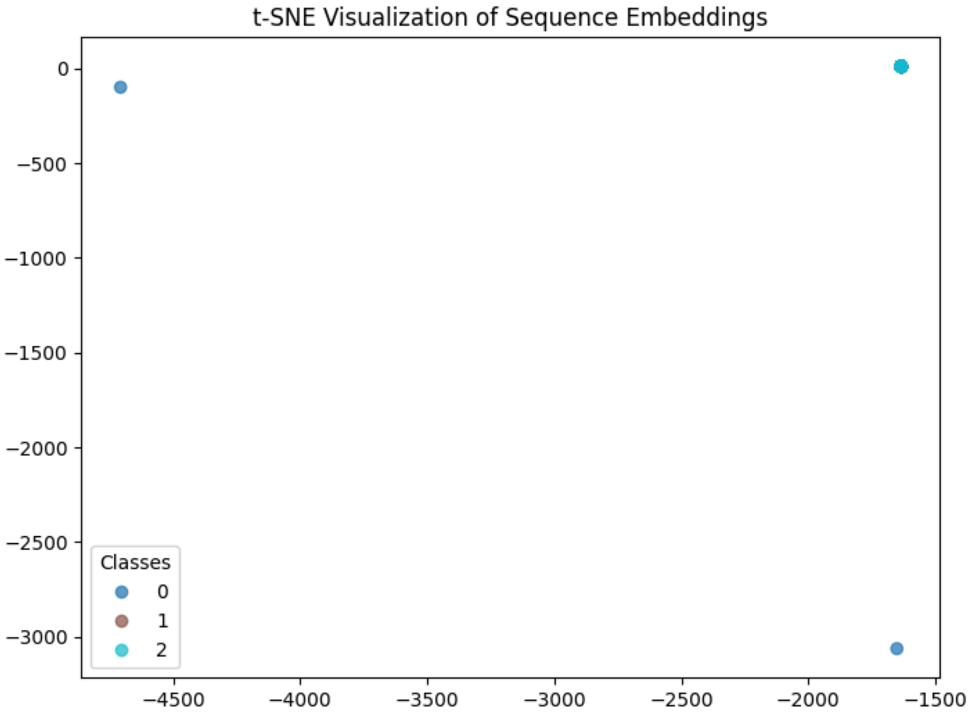
Implementation Details

Component	Choice
Base Model	BERT-base-uncased
Embedding Dim	768
Batch Size	4
Learning Rate	2e-5
Epochs	5
Margin (α)	0.3

Results: Quantitative

Metric	**Original Paper**	**Ours**
Recall@10	53-66%	84%
Training Classes	1,000	30
Embedding Dim	256	768

Results: Qualitative



Critical Analysis

Advantages:

- Higher recall than CNN-BiLSTM
- Better long-range dependency capture
- GPU-accelerated searches

Limitations:

- Small synthetic dataset
- Memory intensive (BERT-base)
- Needs biological validation

Thank You