Llama-DNA: Functionally Annotating Unknown Genomic Sequences

Syed Ashal Ali[1]

Department of Computer Science, Stanford University

Stanford
Computer Science

Problem

- **Challenge:**
  Traditional reference-based models struggle to annotate novel, highly divergent, or repetitive genomic sequences.
- **Why It Matters:**
  Incomplete reference genomes limit biomedical discovery and obscure insights into genetic diversity and disease mechanisms.
- **Aim:**
  To develop a scalable, cross-modal framework that leverages genomic embeddings alongside natural language generation for accurate functional annotation.

Background

- Genomic sequence annotation is essential yet hindered by the reliance on incomplete reference genomes.
- Prior models (e.g., DNABERT-2 *[1]*) have advanced sequence modeling but fall short when handling novel or rare sequences. SPLASH *[2]* provides statistical information about a sequence without a reference.

Data

Sequence	Dataset	SPLASH_Effect	SPLASH_pval	SPLASH_entropy	Annotation
GTCA..	ClinVar	0.6075	0.1207	0.595	This sequence is...

Figure 1: Sample observation from the curated dataset. *Genomic sequences from 6 human chromosomes were extracted from FASTA files and mapped to Ensembl Gene IDs. Gene annotations from RefSeq/Ensembl were integrated with mutation data from the UCSC Genome Browser using chromosomal coordinates. Finally, SPLASH was applied to enrich the dataset with statistical scores (effect size, p-values, and entropy).*

Methods

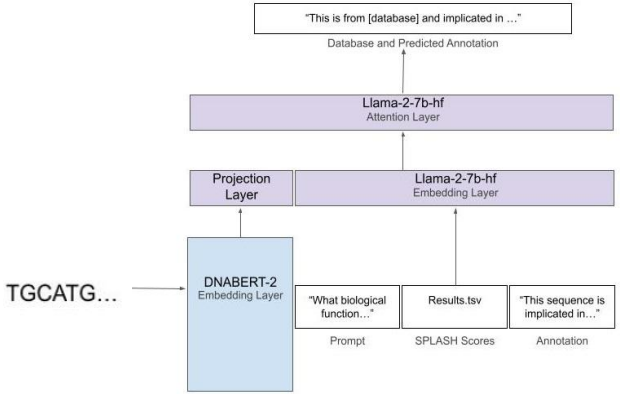

Figure 2: Overview of the Llama-DNA architecture. *Sequence first embedded by DNABERT-2, then projected into textual embedding space of Llama-2-7 using a projection layer. Simultaneously, textual annotations, prompts, and SPLASH-derived statistical scores are embedded directly using Llama-2-7b-hf's embedding layer. The concatenated embeddings are processed through LoRA-based attention layers in attention layers of Llama-2-7b-hf to generate the final predicted annotation and database classification.*

Experiments and Results (I)

Metric	Baseline	Llama-DNA
Accuracy (Dataset Classification)	0.8065	**0.8719**
F1 (Dataset Classification)	0.4464	**0.6422**
Avg. BLEU (Annotation)	0.5685	**0.6118**
Avg. ROUGE-L (Annotation)	0.7172	**0.7591**
Avg. METEOR (Annotation)	0.7833	**0.8413**

Figure 3: Quantitative Results table. *This table reports the quantitative scores I achieved after running my experiment with Llama-DNA and testing it on the test set. It compares the scores that Llama-DNA achieved on the test set with the baseline values.*

Experiments and Results (II)

Ground Truth
This sequence is part of **chr1** chromosome part of **TUFT1** gene. This gene is described as **tuftelin 1** and a **protein coding** transcript type. It is of **skin fragility** and **woolly hair syndrome** phenotype. It is part of the **ClinVar** dataset.
Predicted Annotation
This sequence is found in **chr1** chromosome part of **TUFT1** gene. This gene is described as **tuftelin 1** and a **protein coding** transcript type. It is associated with **skin fragility** and **woolly hair syndrome**. It is part of the **ClinVar** dataset.

Figure 4: Qualitative Results table. *This table compares the ground truth annotation and the annotation predicted by Llama-DNA for a particular DNA sequence. All the functional parts of the annotation are correct but there are deviations from the ground truth in terms of exact n-gram.*

Conclusion

- Llama-DNA demonstrates that cross-modal integration can significantly enhance genomic sequence annotation, offering a robust, interpretable, and scalable solution.
- Future directions could involve scaling dataset to include greater genomes and gaining access to more annotations.

References:
[1] Zhihan Zhou et al. "DNABERT-2: Efficient foundation model and benchmark for multi-species genomes." *ICLR*, 2024.
[2] Chaung et al. "Splash: A statistical, reference-free genomic algorithm unifies biological discovery." *Cell*, 2023.