Enhancing Biological Reasoning via Ensemble Embeddings in BioReason Model

Progress Report

Prabashwara D.G.H. - 210483T

August 24, 2025

Literature Review

Recent DNA foundation models (FMs) like Evo2 and Nucleotide Transformer (NT) [1], [2] have advanced genomics by learning dense representations from nucleotide sequences that enable tasks like variant effect prediction and regulatory element identification. Although these models are powerful they lack interpretability which limits its mechanistic insights. On the other hand the LLMs lack the deep understanding about the nucleotide sequences but are good with reasoning and intepreting.

The BioReason[3] was introduced to address this issue. It utilised the power of both DNA foundation models and LLMs for the first time to make enhanced a biological reasoning model. For the DNA foundational modules they have utilised the NT-500M model variant and the Evo2-1B variant model. For the LLM they have incorporated the Qwen3-1.7B and Qwen3-4B variants.

The BioReason was trained on three different datasets. The first dataset is a biological reasoning dataset that was derived from the KEGG(Kyoto Encyclopedia of Genes and Genomes) database[4]. The second dataset focuses on DNA variant effect prediction of coding sequences and third one focuses on variant effect prediction of coding non-SNVs(non single nucleotide variant). All the three datasets are question and answer datasets. Each dataset has questions, answers(diseases), reasonings, reference DNA sequences and variant DNA sequences so the model can be finetuned easily.

The authors have frozen the DNA foundation model and added a linear projection layer to project the DNA embedding into LLM embedding dimensions. Then they have added the DNA sequences using a specialized token with the question in a chat template, and then tokenized it. Then they have taken the embeddings for those tokens and then stacked those embeddings along with the projected DNA embeddings. Then it was sent to the LLM and the output was taken. In the output, the model provides the reasoning and the

answer(the disease). This answer was taken to evaluate the accuracy and the F1 score of the model. To train the model they have used supervised fine tuning with the Low-Rank Adaptation(LoRA)[5] along with the reinforcement learning with the Group Relative Policy Optimization (GRPO)[6].

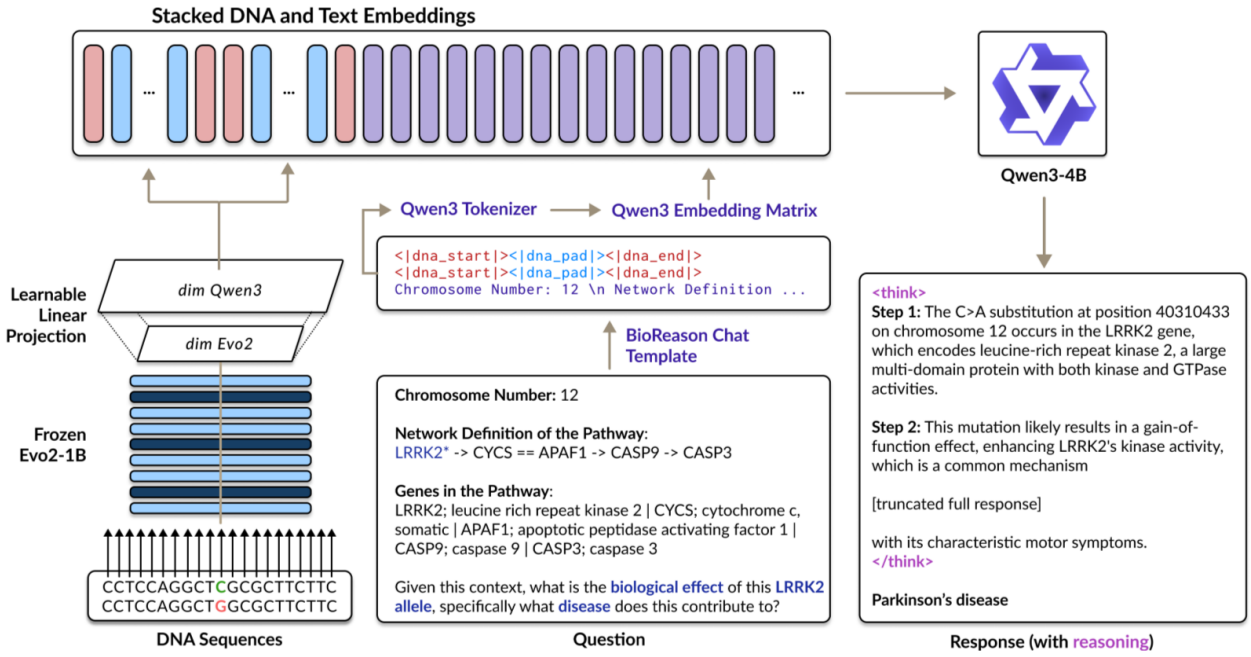


Figure 1: The BioReason Architecture from [3].

Ensembling models have given improved performances in genomic models [7], [8]. To further enhance the model, we can consider ensembling the DNA foundation models instead of relying on a single NT or Evo2 variant. Embeddings from the two DNA models can be projected and then aggregated before integration into the LLM. This approach will allow the model to capture complementary representations of the nucleotide sequences, improving the robustness, generalization, and accuracy of the model.

Methodology Outline

This research focuses mainly on assessing the effect of ensembling DNA foundational models on the BioReason model. Considering the resources and the time available, only the biological reasoning dataset derived from the KEGG database will be used to train the model. The rest of the architecture and training procedure will be the same as in the original BioReason paper. The methodology outline for this research will be as follows.

1. **Dataset**

    - Use the biological reasoning dataset derived from the KEGG database.

- Contains questions, answers (diseases), reasoning, reference DNA sequences and variant DNA sequences.
- The dataset is split into training/validation(1159 data points) and test(290 data points) sets for model development and evaluation.

2. Model Architecture

- For the DNA Foundation Model: NT-500M and Evo2-1B variants will be utilised.
- A Projection Layer will map both DNA embeddings into the LLM embedding dimension.
- Embedding Integration will be done by combining projected DNA embeddings with tokenized question embeddings.
- For the LLM, Qwen3-4B will be used which is the state of the art LLM for the selected dataset according to the BioReason paper.

3. Ensemble Embedding Strategy

- Multiple ensemble methods for DNA embeddings will be explored
  - Averaging embeddings before projection.
  - Weighted Averaging for each model embedding.
  - Concatenating embeddings and projecting into LLM space.
- Compare ensemble approaches against single DNA FM baselines.

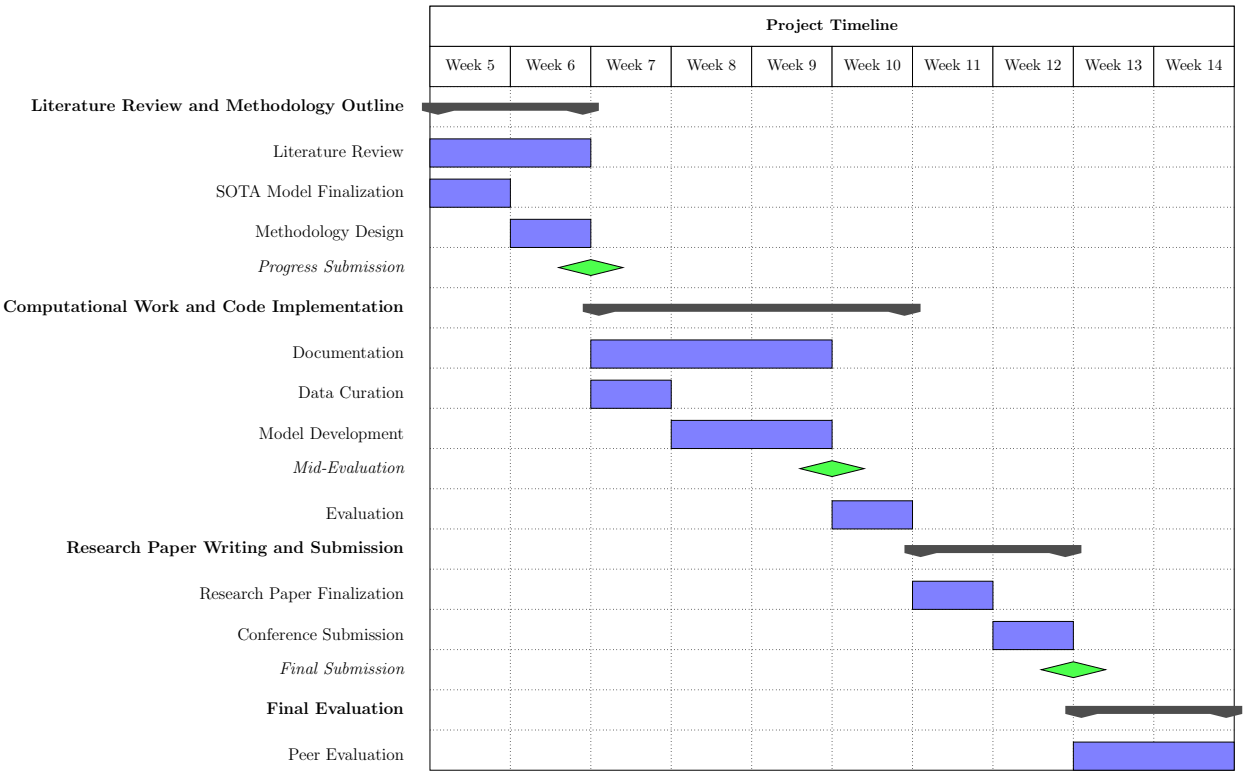
4. Training Procedure

- Supervised Fine-Tuning (SFT) will be carried out with Low-Rank Adaptation (LoRA) using the QA pairs from the dataset.
- Reinforcement Learning will be carried out to improve reasoning with Group Relative Policy Optimization (GRPO) as in the BioReason paper.

5. Evaluation Metrics

- Accuracy of predicted diseases.
- F1 Score for prediction performance.

Project Timeline



References

[1] G. Brixi *et al.*, "Genome modeling and design across all domains of life with Evo 2," Feb. 21, 2025, *Genomics*. doi: 10.1101/2025.02.18.638918.

[2] H. Dalla-Torre *et al.*, "Nucleotide Transformer: building and evaluating robust foundation models for human genomics," *Nat. Methods*, vol. 22, no. 2, pp. 287–297, Feb. 2025, doi: 10.1038/s41592-024-02523-z.

[3] A. Fallahpour *et al.*, "BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model," May 29, 2025, arXiv:2505.23579. doi: 10.48550/arXiv.2505.23579.

[4] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes."

[5] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 16, 2021, arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[6] Z. Shao *et al.*, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," Apr. 27, 2024, arXiv:2402.03300. doi: 10.48550/arXiv.2402.03300.

[7] L.-L. Gu, R.-Q. Yang, Z.-Y. Wang, D. Jiang, and M. Fang, "Ensemble learning for integrative prediction of genetic values with genomic variants," *BMC Bioinformatics*, vol. 25, no. 1, p. 120, Mar. 2024, doi: 10.1186/s12859-024-05720-x.

[8] S. Tomura, M. J. Wilkinson, M. Cooper, and O. Powell, "Improved Genomic Prediction Performance with Ensembles of Diverse Models."