Background

Genomes encode diverse molecular functions through a compact 4-nucleotide vocabulary, yet computational modeling of DNA remains challenging compared to protein language models, which have revolutionized tasks like structure prediction and design.

While prior genomic models focused on extending context length (e.g., 100k+ nucleotides) to capture long-range interactions, they achieved limited functional accuracy and will hinder their ability to generalize across functional genomics, variant effect prediction, and synthetic biology tasks.

To enhance the sufficient representational capacity, he paper Accurate and General DNA Representations Emerge from Genome Foundation Models at Scale propose an AIDO.DNA model for short sequences (≤4k nucleotides), but it can handle a limited length of sequences, since it is a transformer-like architecture and the performance will descend fast as sequence length exceed the upper bound.

Aims

We wanna apply an positional encoding extension method to enlarge the sequence length (≤4k nucleotides), AIDO.DNA can handle without loss of representational capability. We will modify base on AIDO.DNA model and we expect that the sequence length will increase 8-16 times compared with AIDO.DNA.

methodology

We intend to test PI, GeNE in Context Length Extension via Generalized Extrapolation Scale, YaRN in YaRN: Efficient Context Window Extension of Large Language Models, NTK-aware, CLEX, several positional embedding method to enlarge the context length to $4\times, 8\times, 16\times$ compared with the original length. And we will test its performance on downstream tasks compared with AIDO.DNA.

These positional embedding methods don't need to modify the basic architecture and enable to enhance the capability to handle long context via modification of PE layer and a simple fine-tuning or no fine-tuning. Hence, it is a low overhead experiment.

Maybe we can aggregate the advantage of the PE method above, and propose a novel PE method.

Remark: There is no papers applying the method (only need a simple fine-tuning based on pre-trained model, do not need to modify the architecture and pre-train it from scratch) to extend the length of input DNA sequence without loss of performance.