Title: A Cross-Species Transformer Model to Infer Gaps in Incomplete Genome Assemblies

Who: Luke Nguyen (lnguye61), Ashley Xu (axu71), Romer Miranda (remirand), Khoi Le (kcle)

Introduction: We are trying to solve the problem of incomplete genome assemblies. When an organism's genome is sequenced, it must be done in chunks, as current DNA sequencing technologies only work for certain lengths of reads. These reads are then aligned together into a reference genome for that organism. However, there are certain instances where a complete reference genome can not be assembled due to challenges in alignment, such as repetitive regions. As a result, we propose the implementation of a transformer model trained to learn long-range dependencies in DNA sequences, in order to fill these gaps in the genome. The problem we are trying to solve is that of generation: we want our model to learn the underlying distribution of DNA sequence data and be able to generate new sequences.

In larger-scale genomic projects, such as metagenomic studies, sequences can be unlabeled, meaning the species they came from is unknown. Different species have different genomic features, and knowing these unique features can vastly improve the accuracy of a genome assembler. A generalized model that aims to fill in gaps in incomplete assemblies will likely have less accurate results than one that is fine-tuned for that specific species. Thus, we propose adding a classifier prior to the main transformer that generates the missing gaps. The result of this classifier can then be used to fine-tune our main transformer to the specific species the classifier predicted the input sequence to be from.

Related Work: We take inspiration from DNABERT and GapPredict to fill in gaps in incomplete genome assemblies. The DNABERT paper (Davuluri et. al, 2021) describes a model adapted from BERT, a transformer-based language model, and designed to have a general understanding of DNA, learn contextual information, and be extended to various sequence-related applications. The authors pre-trained the model on human genome data and fine-tuned DNABERT with task-specific data for prediction of promoters, transcription factor binding sites, and splice sites, and they also fine-tuned DNABERT on Mouse ChIP-seq datasets to test its generalizability to other organisms. Overall, DNABERT outperformed many existing tools used for prediction of genome-wide regulatory elements.

The GapPredict paper (Birol et. al, 2021) describes an LSTM-based model that attempts to solve the same problem of filling in incomplete genome assemblies. The model is composed of 3 layers, an embedding layer to encode individual bases in sequences, an LSTM layer, and finally a dense layer that outputs probabilities for each nucleotide using softmax. The paper also uses other gap-filling tools, namely Sealer and GAPPadder, 2 non deep learning tools, to benchmark GapPredict's performance. GapPredict predicted 78.9% of gaps that Sealer also predicted, and 65.6% of gaps that Sealer could not predict, with an accuracy of at least 90%. However, when Sealer and GAPPadder were used together, it was found that sequences predicted by GapPredict were 1.6 to 1.8 times less accurate.

Data: The primary datasets we intend to implement are the GRCh38 human assembly and the more modern T2T-CHM13 gapless telomere-to-telomere human genome. There is a lot of data stored in these assemblies, so we may use subsections of each, such as focusing on only protein-coding sequences. We intend to break down our data into smaller sequences using sliding window techniques before tokenizing segments of sequences by embedding on tested k-mer lengths to find a performance/efficiency compromise that works with our machines. If time permits, we would add additional datasets, most likely from other organisms, in which we can train our model to make a prediction about the organism and then based on what organism it thinks the input sequences are, it would then feed the sequences into a tailored model that would then do an organism specific sequence imputation.

Links to Data:

Human

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/

Mouse

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001635.27/

Sea urchin

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002235.5/

Methodology: The architecture of our model will be a transformer. Transformers perform well with sequential data and offer an improvement over RNNs especially when considering long-range dependencies. In particular, the attention mechanism is advantageous for focusing on certain patterns in the genome that are relevant for gap filling. For our base goal, we will first train the model on masked human genome sequences, providing the DNA sequences surrounding the gap and having the model infer the masked sequence. The data will be tokenized as k-mers. This model will serve as a base, somewhat like a foundational model, that can then be finetuned for transformer submodels that are species-specific. For our target goal, we will implement a transformer-based classifier to predict species based on input genome sequences. The classifier's output can then be used to determine which species-specific submodel should be used for sequence imputation.

Metrics: We plan to evaluate accuracy by comparing the model's inferred sequence with the true sequence (before masking). We will first train and test our transformer model on human genome sequences (base goal), and after that, we will train and test our submodels on mouse and sea urchin genome sequences (target goal). We will try different metrics such as sequence alignment score between actual and imputed sequences.

Base goal: Sequence imputation on only human genome sequences, we would mask out parts of the sequences and train a transformer to predict missing nucleotides/codons. We would first test out the feasibility of this, as it may not be computationally possible with our given resources. If it is too computationally intensive, we may want to work on a pretrained model and work on the training process by feeding it more curated data (maybe only protein coding regions, sections that are rich with exons and introns, etc)

Target goal: Species-aware sequence imputation. Adding a classifier to determine species based on genome sequence and passing the classification output to a transformer that is fine tuned to that specific output species. If it is not feasible to first pre-train on human data and then fine-tune for other species, then we may try to train each species submodel separately using more curated (smaller, species-specific) data.

Stretch goal: If the model is working appropriately, it would be really interesting to train it on longer sequences with known exon-intron gaps and have it make predictions. From these predictions, we could then keep a history of the cross-attention patterns used to make predictions. It would be extremely rewarding if we could see protein-coding sequences being predicted with self-attention to other exonic parts of the sequence as the context (basically validating based on the known biological trend that protein-coding regions matter more for protein-coding region prediction). What would the intronic predictions call self-attention to?

Ethics: The stakeholders of this research are scientists who may use the inferred sequences downstream in their research. If there are errors in our inferred sequences, then these errors may be propagated in others' research. In addition, our training data is a reference genome, which comes from an individual person. That may make the model biased toward generating sequences that have a distribution close to the individual's ethnic origin or other characteristics.

Division of labor:

We are meeting biweekly (twice a week), where we will describe roles in more detail as the project evolves. Described here is a high-level overview of everyone's primary focus

Luke: Visualization, properly showing model training and accuracy gains (if prevalent), showing examples of successful and unsuccessful sequence imputation to put in presentations, creating different graphs and charts that show a narrative of how we designed our model, the lessons, and inferences we learned from earlier implementations, and how we then adjusted the model.

Romer: Transformer-based model(s) hyperparameter fine-tuning

Ashley: Data curation (picking specific regions that hold higher biological significance or with high density of exonic and intronic regions), training and test set creation

Khoi: Data tokenization, creating different ways of tokenizing DNA sequences that reflect biology.

Github Link: https://github.com/LukeLose/ACTGap.git