COMP 565 Fall 2023 Final project
Leveraging genome foundation models to predict personal gene expression from DNA sequence

1 General guideline for the final project report

You may work on your own or with another classmate. For the latter, the two students will be assigned the same grade for the project. Therefore, it will be your responsibility to make sure the work is divided equally between you and your teammate. Regardless your choice of the project, write your report in maximum 10 pages excluding figures and references. Use 0.75 inch margin all around, 12 font size, and single-space (which are the same as this instruction).

1. Introduction
    - Opportunities and challenges
    - Related methods/studies
    - One paragraph summarizing the approach and the results
2. Materials and Methods
    - Data processing
    - Model description
    - Evaluation metric
3. Results
    - Benchmark table comparing GFM-based with baselines
    - Scatter plot of observed and predicted gene expression
    - Qualitative analysis (see details below)
4. Discussion
    - Summarize what have been accomplished in this study
    - Limitations of the presented method
    - Future work

2 Seek help if you need them

You may consult with your instructor and TA as well as Adrien Osakwe (QLS student in Li lab) and Shadi Zabad (PhD CS student in Li lab) either on Ed or by appointment.

3 Overview

The advent of Genome Foundation Models provides promising opportunities of predicting molecular phenotypes using individual genomic sequences as unstructured data. This may address the gap observed in [1], where Enformer (i.e., transformer + CNN) performed poorly on the task possibly because of the lack of pre-training on the human genome and the inability to operate at the single nucleotide resolution.

In this project, you will explore GFMs on predicting personalized gene expression using the DNA sequences from the GTEx subjects measured by whole genome sequencing (WGS). This is a classic DNA-to-gene-expression problem. While traditional methods use tabular genotype matrix to predict gene expression in a transcriptome-wide association studies (TWAS) (e.g., PrediXcan [2]), we will use unstructured sequences as input to predict gene expression. To do this, we will make use of the Genotype-Tissue Expression (GTEx) version 8 data [3], which contains genotype and gene expression measured in the same subjects over diverse tissues (`https://gtexportal.org/`). For the purpose of the course project, we will focus on predicting gene expression in the Whole Blood, which is the tissue that contain most samples among the 54 tissues in total (`https://www.gtexportal.org/home/tissueSummaryPage`).

4 Data processing

The gene expression and genotype data are available at `https://drive.google.com/drive/folders/17dOyxbSax-i18pvdnYL-hqmXUzZWkcjs?usp=sharing`.

4.1 Gene expression

The gene expression file Whole_Blood.v8.normalized_expression.bed.gz (also available publicly at `https://www.gtexportal.org/home/downloads/adult-gtex#qtl`) contains fully processed and normalized gene expression in BED format (which were used for eQTL mapping in the GTEx analysis). The bed file looks like this:

#chr	start	end	gene_id	GTEX-111YS	...	GTEX-ZXG5
chr1	29552	29553	ENSG00000227232.5	-1.225	...	0.765
chr1	129222	129223	ENSG00000238009.6	-0.673	...	-0.321
chr1	131024	131025	ENSG00000233750.3	0.301	...	0.1518

The columns are chromosome, start, end, Ensembl gene ID, and GTEx sample ID (670 Whole Blood samples in total), respectively. Note that the start and end position is just 1-based nucleotide position of the TSS of the gene, which is not the real end of the gene (otherwise every gene would have been only one nucleotide long). The gene annotations are in references_v8_gencode.v26.GRCh38.genes.gtf. For example, gene ENSG00000233750.3 (gene name: *CICP27*, which is a pseudogene) has gene expression value 0.301 for the first subject GTEX-111YS and 0.1518 for the last subject GTEX-ZXG5. You can find the genotype for the same subjects from the genotype file described next.

4.2 Genotype

Note the genotype access is protected and only granted to the instructor Prof. Li for research purpose. **Do not share them with anyone outside of the class without your instructor's permission.** The file phASER_WASP_GTEx_v8_merged.vcf.gz contains the variant calling file (VCF). This file is a large gzip-compressed file. Do not uncompress or open with text editor but rather view it under command line with `zless phASER_WASP_GTEx_v8_merged.vcf.gz` or `gunzip -c phASER_WASP_GTEx_v8_merged.vcf.gz | less`. All the lines that start with the # sign are the descriptor line. The last line that starts with the # sign is the table header:

```
#CHROM   POS     ID                  REF  ALT  QUAL  FILTER  INFO   FORMAT  GTEX-1117F  ...
chr1     13526   chr1_13526_C_T_b38  C    T    .     PASS    AF=... GT:...  0|0...      ...
```

The columns are described in Wikipedia (`https://en.wikipedia.org/wiki/Variant_Call_Format`). GTEX-1117F ... GTEX-ZZPU are the Subject IDs.

The actual data line starts from the line without the # sign. The FORMAT column always has 'GT:PG:PB:PI:PW:PC:PM', which are the names of the fields coded as binary in the individual genotype columns that follow it (i.e., subject columns GTEX-1117F, ... GTEX-ZZPU). Here GT stands for genotype, which is the data we care about. You may ignore the rest of the fields (i.e., PG:PB:PI:PW:PC:PM). The GT field for each subject is coded as $0|0$, $1|0$, $0|1$, or $1|1$, which indicates the four combinations of allelic specific mutation.

Your first task is to write a script that can create personal DNA sequences by changing the nucleotide in the reference genome (hg38) (i.e., the letter under the REF column above) to the alternative allele (i.e., the letter under the ALT column) whenever necessary based on the individual genotype. To simplify your task, you may consider mutations on any of the two alleles as a mutation (i.e., $0|1$, $1|0$, or $1|1$ as a positive mutation and only $0|0$ as no mutation).

4.3 Create personal DNA sequences

The reference genomic sequences in FASTA format can be downloaded in here: `https://hgdownload.cse.ucsc.edu/goldenpath/hg38/chromosomes/`. However, you can extract genomic sequences more efficiently from the human reference genome (version **hg38** or **GRCh38** – make sure you use the right version of the reference) with `genome retrieval` function in R `https://cran.r-project.org/web/packages/biomartr/vignettes/Sequence_Retrieval.html`. There is no well known equivalent package in Python to my best knowledge. Some options are GenomePy at `https://vanheeringen-lab.github.io/genomepy/content/api_core.html` and `gget` at `https://pachterlab.github.io/gget/en/seq.html`.

We do not create the entire personal genome sequence of 3.2 billion nucleotide of every subject but rather only the *in-cis* regulatory regions of the target genes that are loosely defined as 1000 nucleotides upstream and 1000 nucleotides downstream from the transcription start site (TSS) of a target gene (i.e., 2001 length of sequence per target gene). You can do this very efficiently with `bedtools` in command-line (`https://bedtools.readthedocs.io/en/latest/content/tools/window.html`):

```
bedtools window -u -a phASER_WASP_GTEx_v8_merged.vcf.gz -b
 ↪  Whole_Blood.v8.normalized_expression.bed.gz -header | gzip >
 ↪  phASER_WASP_GTEx_v8_merged_filtered.vcf.gz
```

The output file from the above command `phASER_WASP_GTEx_v8_merged_filtered.vcf.gz` was saved for you in the same Google Drive (149M in size in contrast to 9.6G for the original genotype file). You may work with the filtered file for the project. To extend the sequence length, you will need to set the window size with the optional flag `-w`. For example, if we want 5000 bp +/- around TSS (i.e., 10kb + 1 sequence length), we do `bedtools window -w 5000 -u -a genotype.vcf.gz -b tss.bed.gz`.

5 Training GFMs on predicting personal gene expression from DNA sequences

5.1 GFMs

Experiment the pre-trained GFMs on the regression task of predicting personal gene expression using individual DNA sequences.

- HyenaDNA (`https://github.com/HazyResearch/hyena-dna`)
- DNABERT (`https://github.com/jerryji1993/DNABERT`)
- DNABERT2 (`https://github.com/Zhihan1996/DNABERT_2`)
- Nucleotide Transformers (`https://github.com/instadeepai/nucleotide-transformer`)

Depending on your capacity, you don't need to experiment on all GFMs. To work with the existing GFM, you will need to convert the sequences you obtained above into the GFM-specific formats. For example, while HyenaDNA takes single-nucleotide resolution sequences, DNABERT and NT may need k-mer representation.

5.2 Target genes

Here we can work with the genes with strong genetic signals. For instance, 2589 genes have an in-cis eQTL SNP multi-testing corrected q-value $<$ 0.01 based on Whole_Blood.v8.egenes.txt.gz file. For the scope of this project, you don't need to train GFM model to predict the expression of all of these genes. You may either train a model to predict expression of a single gene using 80% of the 670 samples as the training examples (20% for testing) or you may train a single-output model to predict expression of G genes. In the latter case, your training and testing examples are G times bigger than the single-gene model.

5.3 Training GFMs

To leverage the pre-trained GFM, you may experiment **probing** and **fine-tuning**. For the former, you don't need to exhaustively probe all embedding layers. Try a few embedding outputs from the last, second last, etc. Here you may save the embedding first and then train an external regression model such as linear regression, XGBoost, MLP, Random Forest, etc. Be sure to document your choices. For fine-tuning, for a large pre-trained GFM, you may also selectively fine-tune some layers or experiment with learning the rescaling weights as done in the NT.

If you need to access GPU, your instructor can support your application for a Compute Canada account. Apply using his CCRI: yfh-205-01 for the supervisor info. If you are not familiar with Compute Canada, learn how to submit batch or interactive jobs (`https://docs.alliancecan.ca/wiki/Running_jobs`). When you do submit use `--account=ctb-liyue` to enjoy the priority given to us. For example, to submit an interactive job for one hour using one GPU and 125G CPU RAM, you may do:

```
salloc --time=1:0:0 --ntasks=1 --account=ctb-liyue --gres=gpu:1 --mem=125G
```

5.4 Benchmark

Compare the performance of the GFM+regressor with the standard CNN or the SOTA Enformer (optional for the purpose of this project) in terms of R-squared, Pearson, Spearman correlation, and MSE. Extend the sequences to longer than 1k nucleotide for one or two target genes to see whether your prediction improves. FYI: the *in-cis* eQTL analysis is based on 1 Mb +/- around the TSS, which is 2 Mb long sequence and perhaps only HyenaDNA can handle.

6 Model explainability (optional)

Perform in-silico mutagenesis on the individual predictions from the trained model to identify putative causal variants and then check how well it matches with the eQTL fine-mapped SNPs (`https://gtexportal.org/home/downloads/adult-gtex#qtl`).

Examine the attention heatmap to see whether the model can pay attention to known TF binding sites `https://resources.aertslab.org/cistarget/regions/`.

You may compute Kernel SHAP values or DeepLIFT/DeepSHAP values at each nucleotide of the test sequences but consider this as optional if you run out of time.

7 Explore

The above are only a few suggestions. You can explore further directions should your bandwidth allow. For example, you may explore in-context learning with a few shot examples and binary coding of the gene expression as the prepended sequence in the prompt (Lecture 12).

References

[1] Alexander Sasse, Bernard Ng, Anna E. Spiro, Shinya Tasaki, David A. Bennett, Christopher Gaiteri, Philip L. De Jager, Maria Chikina, and Sara Mostafavi. Benchmarking of deep neural networks for predicting personal gene expression from dna sequence highlights shortcomings. *bioRxiv*, 2023.

[2] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091–1098, 2015.

[3] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.