Implementing Hubble.2d6

A tool for CYP2D6 phentoype prediction

August 11, 2021

Author
Coleton Annett
University of Victoria
cannett@uvic.ca

Data Availability
All data used can be found on the original paper's zenodo, or this project's repository.

Introduction

CYP2D6 is a highly polymorphic gene whose produced protein metabolises up to 25% of all clinically prescribed drugs today. Predicting an individual's activity to metabolise these drugs allows for more personalised treatment and lower risk of drug toxicity.

Hubble.2D6 is a tool that is presented to help determine CYP2D6 phenotype (metabolic activity) based on haplotype function predictions. Efforts were made to implement Hubble.2D6 from scratch based on methodology, etc. from the original paper. In particular, the method of transfer learning – commonly used in image recognition – is utilised by Hubble.2D6 do aid development of complex models even within a domain suffering from data scarsity.

Objective

Implement Hubble.2D6, following as closely as possible with the original paper. This includes use of transfer learning and functional annotation of one-hot sequence data.

Methodology

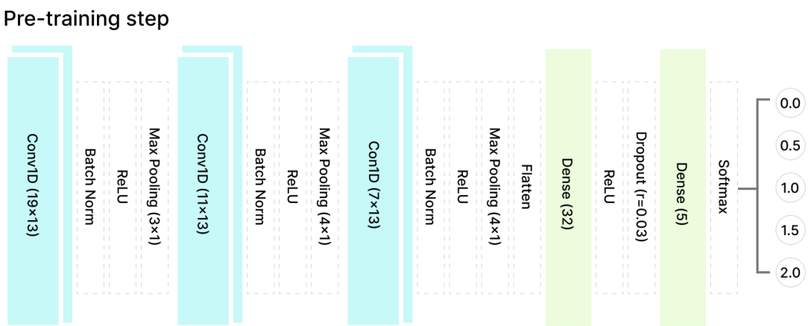
2 training steps were implemented, with the final ensemble model containing 7 individual networks whose weights are fine-tuned from the first step's model.

The output of the final model is decoded into a haplotype functional class (*no function, decreased function, or normal function*)

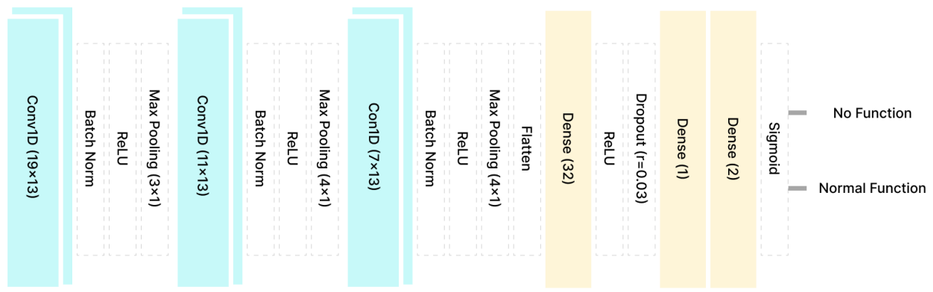
Network Architectures

Implementation involves building 2 separate models, one trained on activity score classification, and the other on ordinal regression of probability scores. Learned weights from one are then transferred to the other for fine-tuning.

Pre-training step



Final training step



Pre-processing

Custom scripts were used to convert data from Variant Call Format (VCF) into one-hot encode sequence data with functional annotations, resulting in a *14868 x 13* matrix per sample.

Post-processing

Output of the final ensemble model are two probability scores. Cutoff values are used to convert these two scores into haplotype function classes based on their values. Final prediction classes are "No Function", "Decreased Function", and "Normal Function".

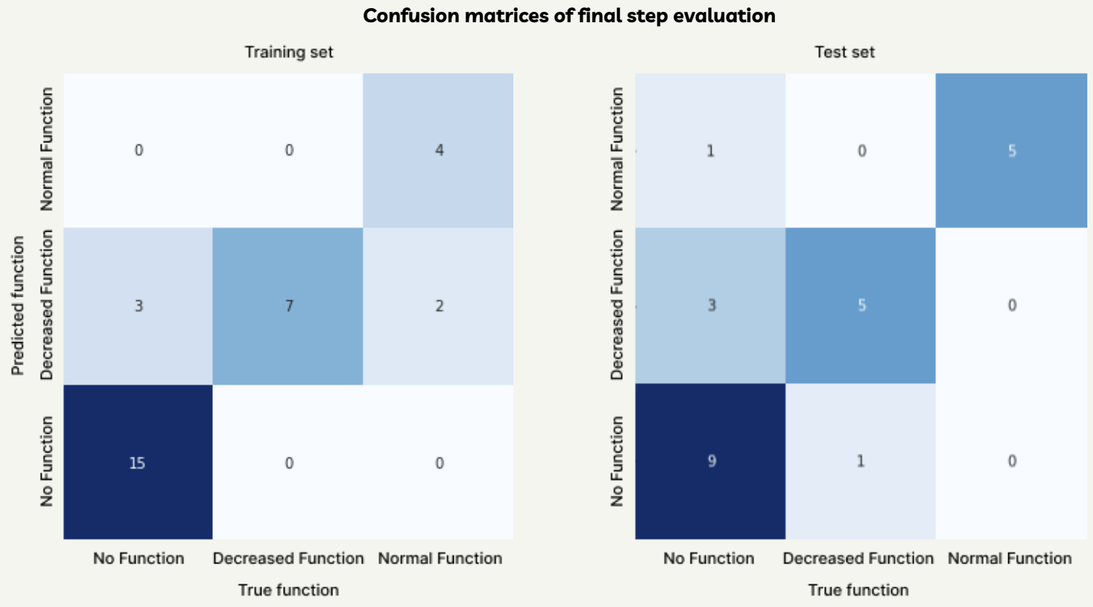
Complications

Data for the 2nd pre-training step was not available and therefore was omitted from the implementation.

Training and results

The **pre-training step** consisted of 50,000 simulated diplotypes as the training set, and 10,000 within the test set. The final model achieved *83% accuracy* on the testing set.

The **final training step** inherited the weights of the convolution layers from the previous model and were fine tuned. Training set comprised of 31 star alleles and their respective suballeles, with 24 star alleles being held for evaluation. The final ensemble model resulted in a 80% *accuracy score* on the test set.



Evaluation

To compare with the official Hubble.2d6 tool, both it and my implementation were run on the entire PharmVar star allele set, as well as only the uncurated star alleles.

In both cases, my implementation attains an *agreement rate of approximately 60%* with Hubble.2d6's predictions.

Conclusion

Due to data availability and other concerns, a true implementation of Hubble.2D6 could not be achieved. However, the attempted implementation was able to generate predictions that agree with the official tool at a rate of 60%, and predict to 80% accuracy the haplotype functions contained in the evaluation set.

Related literature

Gregory McInnes, Rachel Dalton, Katrin Sangkuhl, Michelle Whirl-Carrillo, Seung-been Lee, Russ B. Altman, andErica L. Woodahl. Hubble2d6: *A deep learning approach for predicting drug metabolic activity*. bioRxiv, 2019.doi:10.1101/684357.
URL: https://www.biorxiv.org/content/early/2019/06/27/684357.