A Deep Learning Approach for Predicting Function of Non-coding Genomic Variants

Fred Lu

Advised by Zihuai He, PhD, Dept. of Neurology

BACKGROUND

A large variety of single-nucleotide polymorphisms in the genome are associated with specific diseases. Most such genomic variants occur in non-coding DNA sequences, so they are not directly involved in protein variation. This makes it challenging to understand their function.

Goal: Build neural networks to predict functional variants using epigenetic markers as predictors.

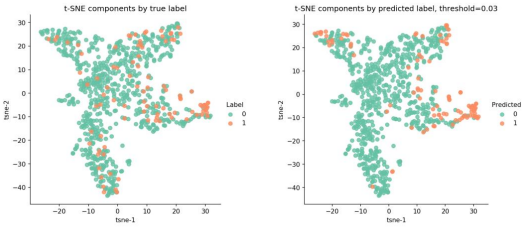
DATA & FEATURES

MPRA dataset for GM12878 (lymphoblastoid) cell line:

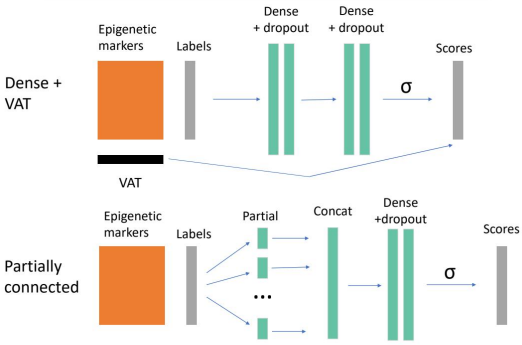
- 693 experimentally confirmed functional variants
- >22,000 negative variants

Cell/tissue-specific epigenetic features from ENCODE:

- 1016 features for each variant site
- Scores for each of 8 different markers in 127 different cells/tissues



MODELS



- **Dense:** Fully connected net with dropout
- **PC Net:** Partially connected net, use sparseness to leverage inter-feature relationships
- **Dense+VAT**: add perturbation regularization to Dense

Benchmarks:

- *GenoNet* (He et al.): Published elastic net predictions
- *Logistic* Regression: L2 reg. with 3-fold CV

SETUP

Data first split into **train (85%) / test (15%)**. Models trained with **iterated train (80%) / dev (20%)** splits within train set.

The following metrics are used:

- Average precision-recall **(AUPR)**
- Area under ROC curve **(AUROC)**

RESULTS

	Avg. validation		Test set	
Model	AUPR	AUROC	AUPR	AUROC
Logistic	0.259	0.764	0.228	0.738
GenoNet	0.251	0.740	0.222	0.728
Dense	0.266	0.761	**0.232**	0.747
PC Net	**0.275**	**0.769**	0.228	**0.750**
Dense+VAT	0.265	0.753	0.226	**0.750**

- Our model outperforms the benchmarks
- Models are relatively stable with architecture modifications

DISCUSSION

- Scores vary across chromosomes, but do not depend on number of training examples.
- PC and VAT may have tendency to overfit.
- Incorporate semi-supervised learning in the future