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

Fred Lu
Advised by Zhihua 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.


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 + VAT:** 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

RESULTS

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg. validation AUPR</th>
<th>AUROC</th>
<th>Avg. validation AUROC</th>
<th>Test set AUPR</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>0.259</td>
<td>0.764</td>
<td>0.228</td>
<td>0.738</td>
<td></td>
</tr>
<tr>
<td>GenoNet</td>
<td>0.251</td>
<td>0.740</td>
<td>0.222</td>
<td>0.728</td>
<td></td>
</tr>
<tr>
<td>Dense</td>
<td>0.266</td>
<td>0.761</td>
<td><strong>0.232</strong></td>
<td>0.747</td>
<td></td>
</tr>
<tr>
<td>PC Net</td>
<td><strong>0.275</strong></td>
<td><strong>0.769</strong></td>
<td>0.228</td>
<td><strong>0.750</strong></td>
<td></td>
</tr>
<tr>
<td>Dense + VAT</td>
<td>0.265</td>
<td>0.753</td>
<td>0.226</td>
<td><strong>0.750</strong></td>
<td></td>
</tr>
</tbody>
</table>

- 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

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)