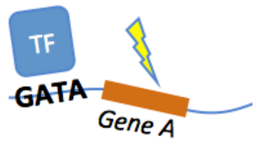
Transcription Factor Binding Prediction Using Constrained Convolutional Neural Networks

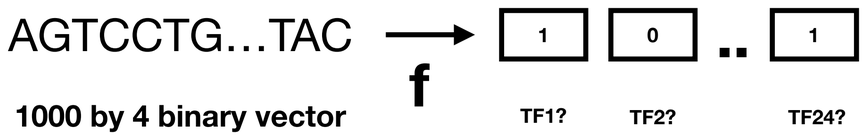
Yakir Reshef, Fulton Wang

Background



- Different transcription factors (TFs) bind at different locations in the genome in a sequence-dependent manner
- Presence of TFs at a location can determine whether nearby genes are activated and ultimately the presence of associated diseases
- Learning relationship between sequence and TF binding leads to understanding effect of genetic variants on disease

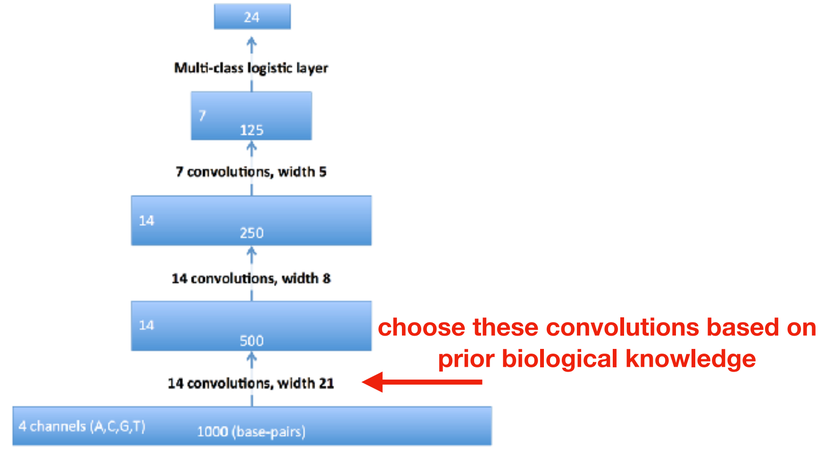
Problem Statement



- Multitask classification of length 1000 DNA sequences.
- Each of 24 tasks is to predict presence of 1 of 24 TFs “in” the sequence
- Dataset: ~400k sequences, each labelled with bound TF

Can we learn a model f that is biologically interpretable and thus generalizable?

Approach

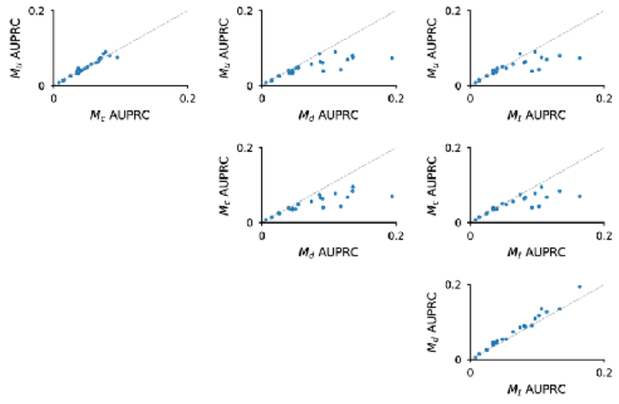


- Convolutional neural network whose first layer filters are **chosen a priori**
- Filters chosen to be short sequences, i.e. “motifs” experimentally shown to bind to TFs.

Models Compared

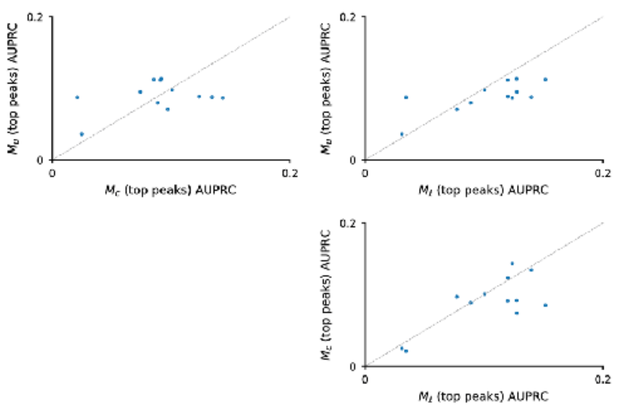
- **M_C**: CNN, with chosen filters
  - Biologically plausible filters may generalize outside training data better
  - Interpretable features
- **M_U**: equivalent model, with learned filters
  - Filters may not correspond to biological mechanisms
- **M_d**: equivalent model, with random filters
  - Unlikely to capture useful local information
- **M_I**: vanilla logistic regression
  - Global model, i.e. entire sequence is 1 “motif”

Results



- $M_D > M_I > (M_U \sim M_C)$
- No sacrifice when replacing non-interpretable first layer filters with interpretable filters.
- Suggests importance of global determinants for TF binding
- Different tasks may have different suitable representations - task specific models may do better

Results on cleaner data



- Same comparison on subset of data with more certainty on labels
- Same trend: $M_I > (M_U \sim M_C)$, though perhaps $M_U \neq M_C$