Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

Christina Ji

April 6, 2017

DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences

Daniel Quang and Xiaohui Xie

http://mit6874.github.io

Overview

- Key Claim
  - Novel hybrid convolutional and bi-directional long short-term memory recurrent neural network framework for predicting non-coding function *de novo* from sequence
- Importance
  - Non-coding DNA has many disease-related variants
- Issues
  - Overstated claims and lack of explanation for selected hyperparameters

Assumption

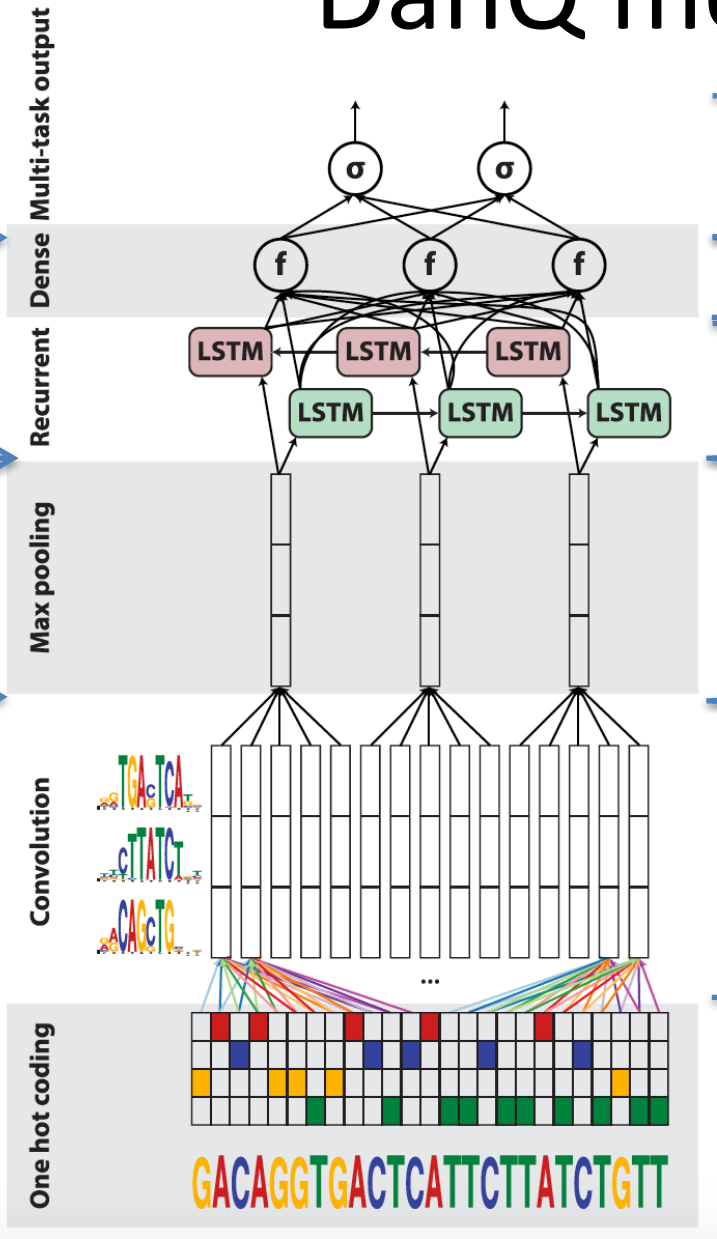
- Long-term dependencies exist between motifs and can be exploited by RNNs when learning regulatory grammars

Data Used

- Same features and data as DeepSEA
  - Segment human genome into non-overlapping 200-bp bins
  - Samples: 1000 bp centered on a bin overlapping at least one TF binding ChIP-seq peak
  - Targets: 919 ChIP-seq and DNase-seq peak sets from ENCODE and Roadmap Epigenomics

DanQ model

60 epochs
~6h each to train



ReLU + Dropout probability 0.5

Dropout probability 0.2

ReLU

919 output features

925 neurons

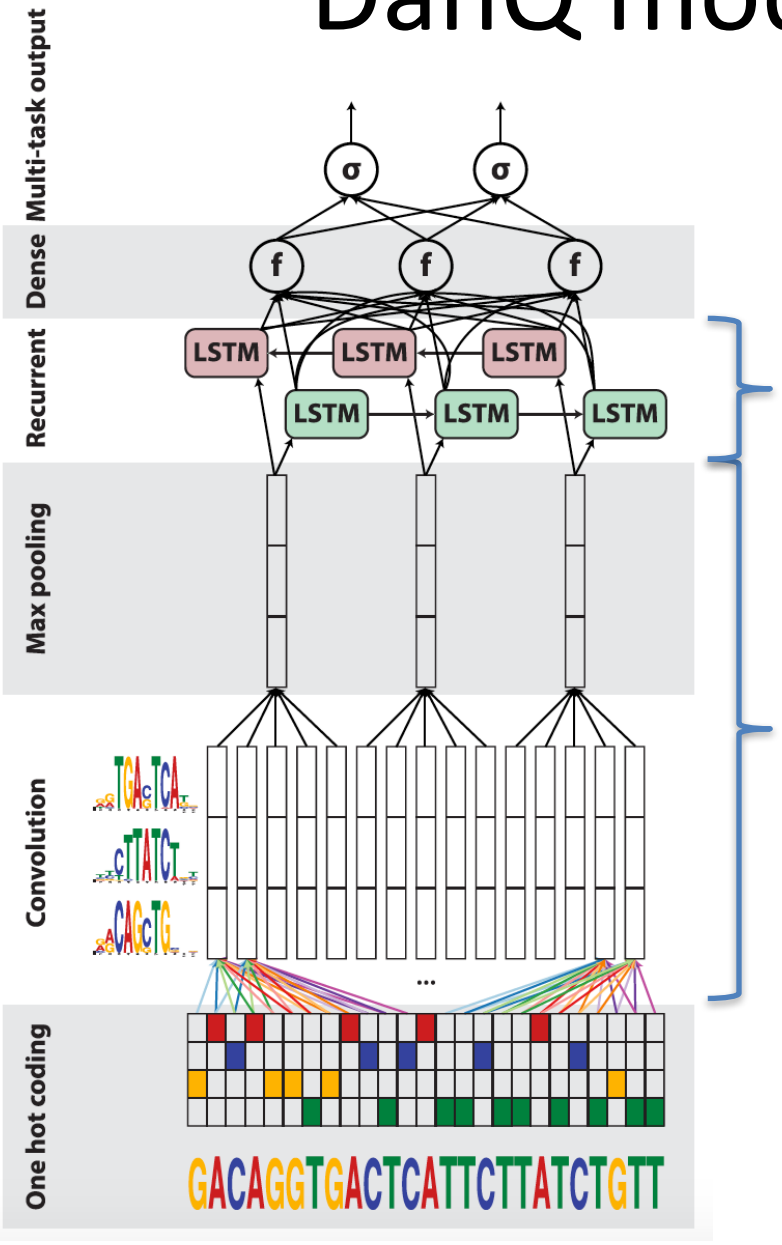
320 forward neurons
320 backward neurons

Window size 13
Stride 13

320 kernels
Window size 26
Stride 1

Random weight initialization from *unif*(-.05, .05)
Bias initialization to 0

DanQ model

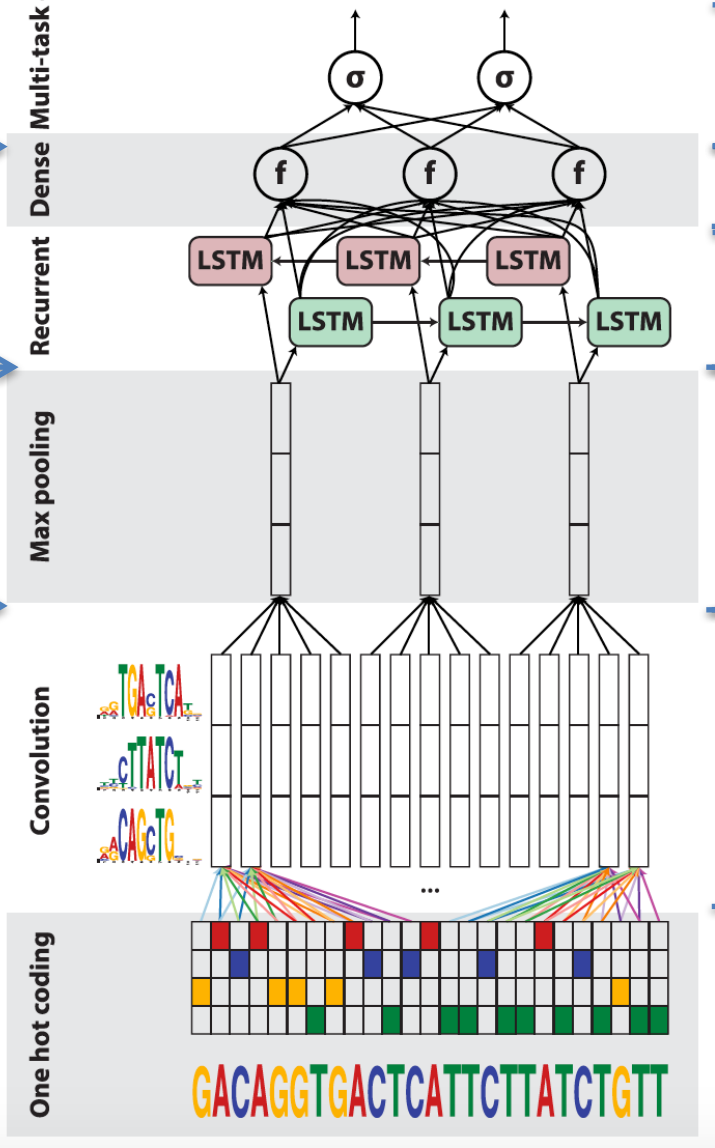


Learn regulatory grammar governed by physical constraints that dictate the *in vivo* spatial arrangements and frequencies of combinations of motifs, a feature associated with tissue-specific functional elements such as enhancers

Capture regulatory motifs

DanQ-JASPAR model

30 epochs
~12h each to train



ReLU + Dropout probability 0.5

Dropout probability 0.2

ReLU

919 output features

925 neurons

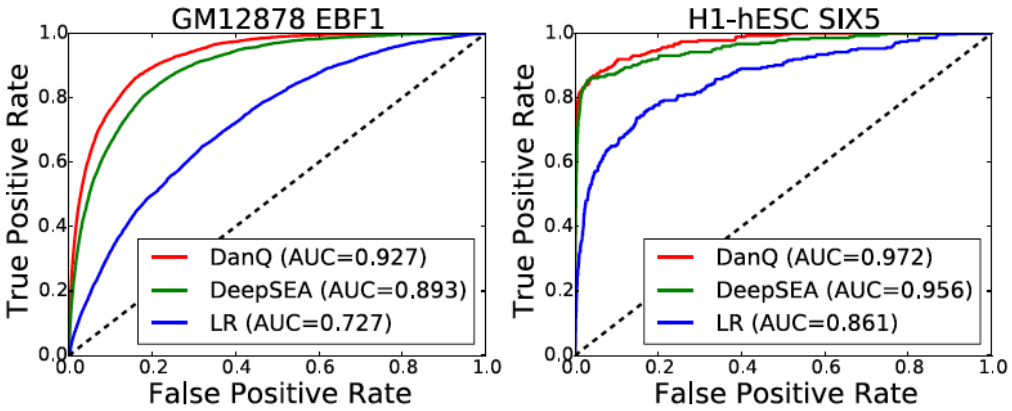
512 forward neurons
512 backward neurons

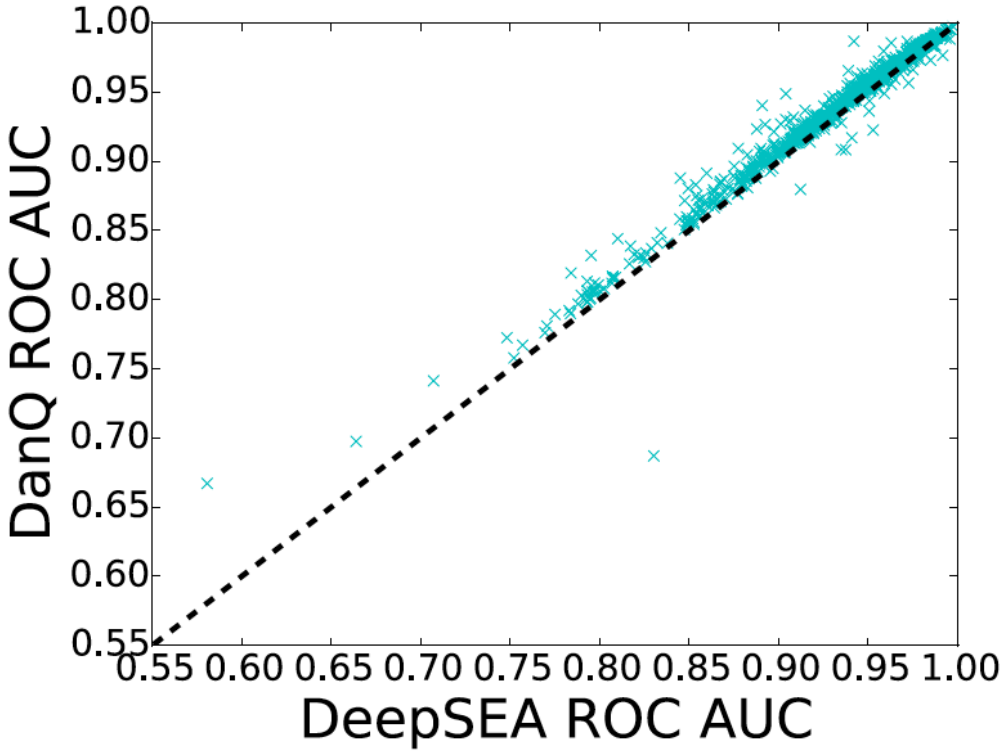
Window size 15
Stride 15

1024 kernels
Window size 30
Stride 1

Half of weights initialized with motifs from JASPAR
Corresponding biases from *unif*(-1.0, 0.0)

ROC AUC Comparisons

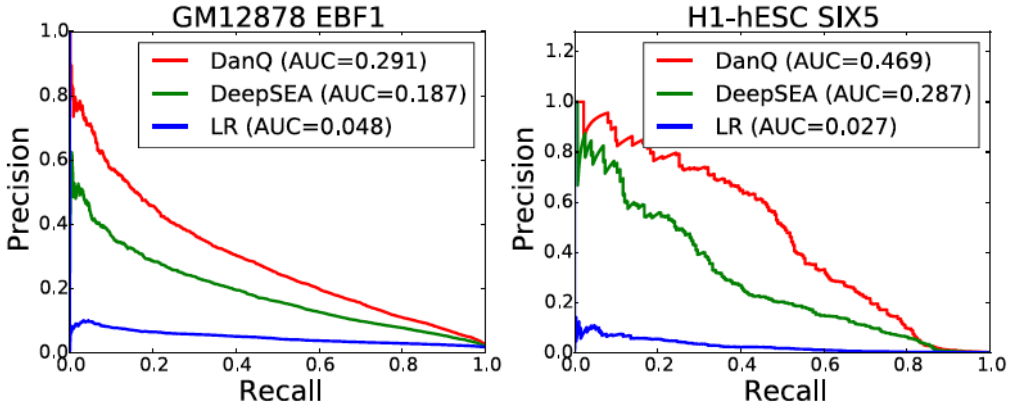


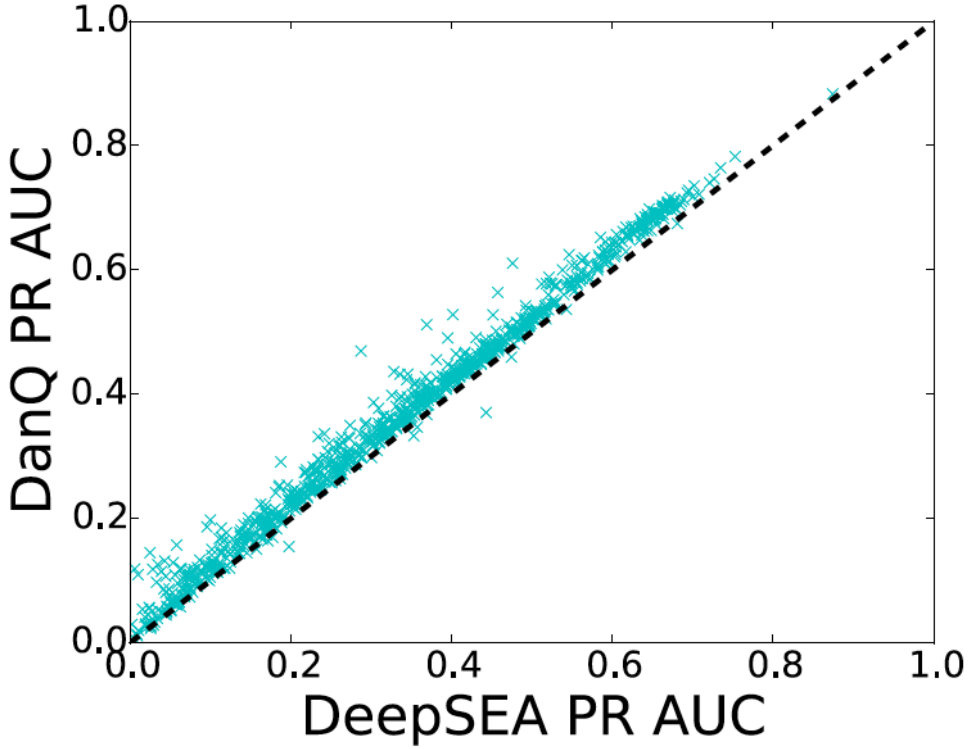


DanQ outperforms DeepSEA for 94.1% of targets, although absolute improvement only around 1-4% for most targets

Given sparsity of positive binary targets (~2%), ROC AUC inflated by class imbalance

PR AUC Comparisons

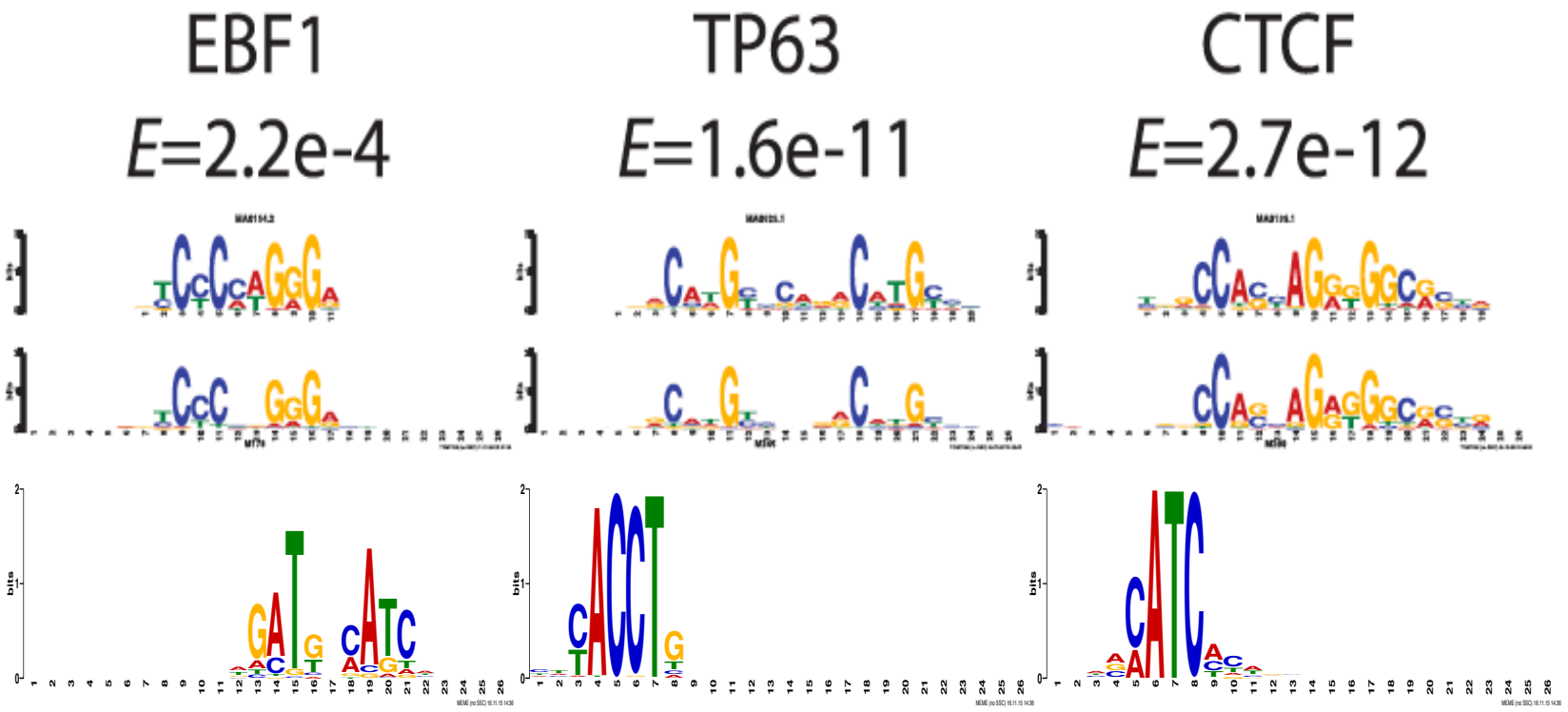




DanQ outperforms DeepSEA for 97.6% of targets, with absolute improvement over 10% and relative improvement over 50% for some samples

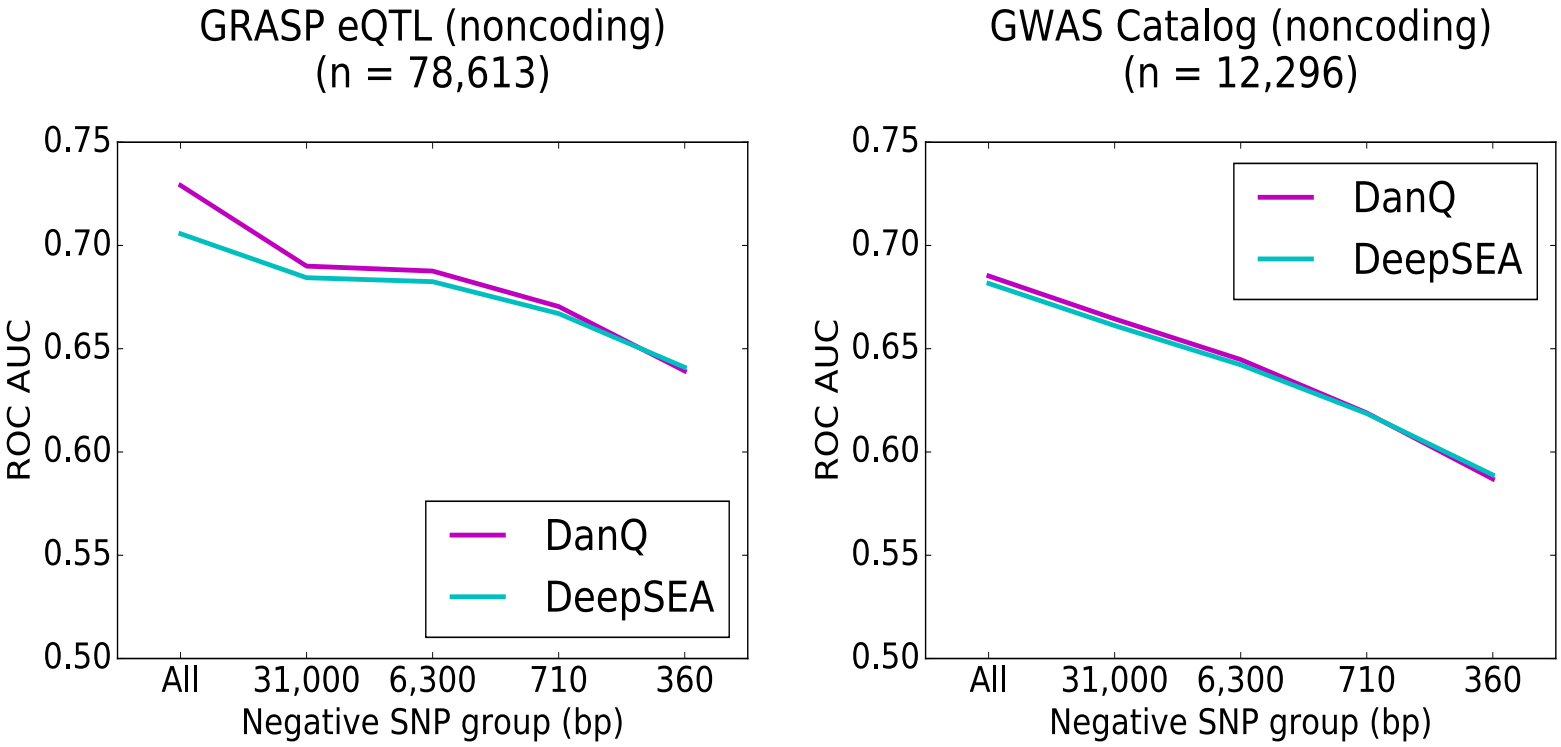
Adding recurrent connections significantly increases modeling power

Kernel Visualization



166 of 320 learned motifs align with known motifs

Functional SNP Prioritization



DanQ outperforms DeepSEA for most of testing sets with 0.5 - 2% difference

X-axis: Distance from negative SNP group to nearest positive SNP in 1000 Genomes Projects

Key Claims

- Hybrid CNN-RNN architecture is better than only a CNN for predicting function of non-coding variants
- DanQ is an effective motif discoverer

Claim 1 Evaluation

- Claim: Hybrid CNN-RNN architecture is better than only a CNN for predicting function of non-coding variants
- Evidence 1: 97.6% of DanQ PR AUC scores surpassed DeepSEA scores
  - Analysis: Holds
- Evidence 2: Over 50% relative improvement compared to DeepSEA for some samples
  - Analysis: Figure shows two best samples, most around 2-7% relative improvement

Claim 2 Evaluation

- Claim: DanQ is an effective motif discoverer
- Evidence 1: DanQ model learned diverse motifs and 166 of 320 matched known motifs
  - Analysis: Effect of other 154 motifs?
- Evidence 2: DanQ-JASPAR showed 2-3% absolute improvement in PR AUC scores over DanQ
  - Analysis: Additional motifs learned?

Hyperparameter selection

Hyperparameter	DanQ	DanQ-JASPAR	DeepSEA
Convolutional kernel size	26	30	8
Max pooling window size and stride	13	15	4
Fully connected layer number of neurons	925	925	925
Weight regularization	None	None	5e-7 for L2, 1e-8 for L1

- Did not mention which hyperparameters were tested
- Would a smaller size result in faster training without loss of motif detection?
- Visualize connections from fully connected layer to output?

Future Improvements

- Since causal variants are likely in linkage disequilibrium with SNPs labeled as positive variants, study the link between phenotypes and haplotypes instead
- Fully recurrent to process sequences of arbitrary length
- Incorporate new ChIP-seq and DNase-seq datasets from more cell types
- Distributed computing-based hyperparameter tuning

Impact

- Improvement in chromatin effect prediction from DNA sequences
- All source code and trained models available on Github for future use
- Further research: Can it translate to an improvement in functional variant prediction?

Summary

- Key Claim
  - First hybrid CNN and RNN to predict the function of non-coding variants from DNA sequences
- Importance
  - 98% of human genome is non-coding, containing 93% of disease-associated variants, yet majority of function is unknown
- Issues
  - Improvements may not be as significant as stated
  - Lack of transparency in hyperparameter tuning
  - Author's ego, eh? **Dan**iel **Q**uang

FIN - Thank You