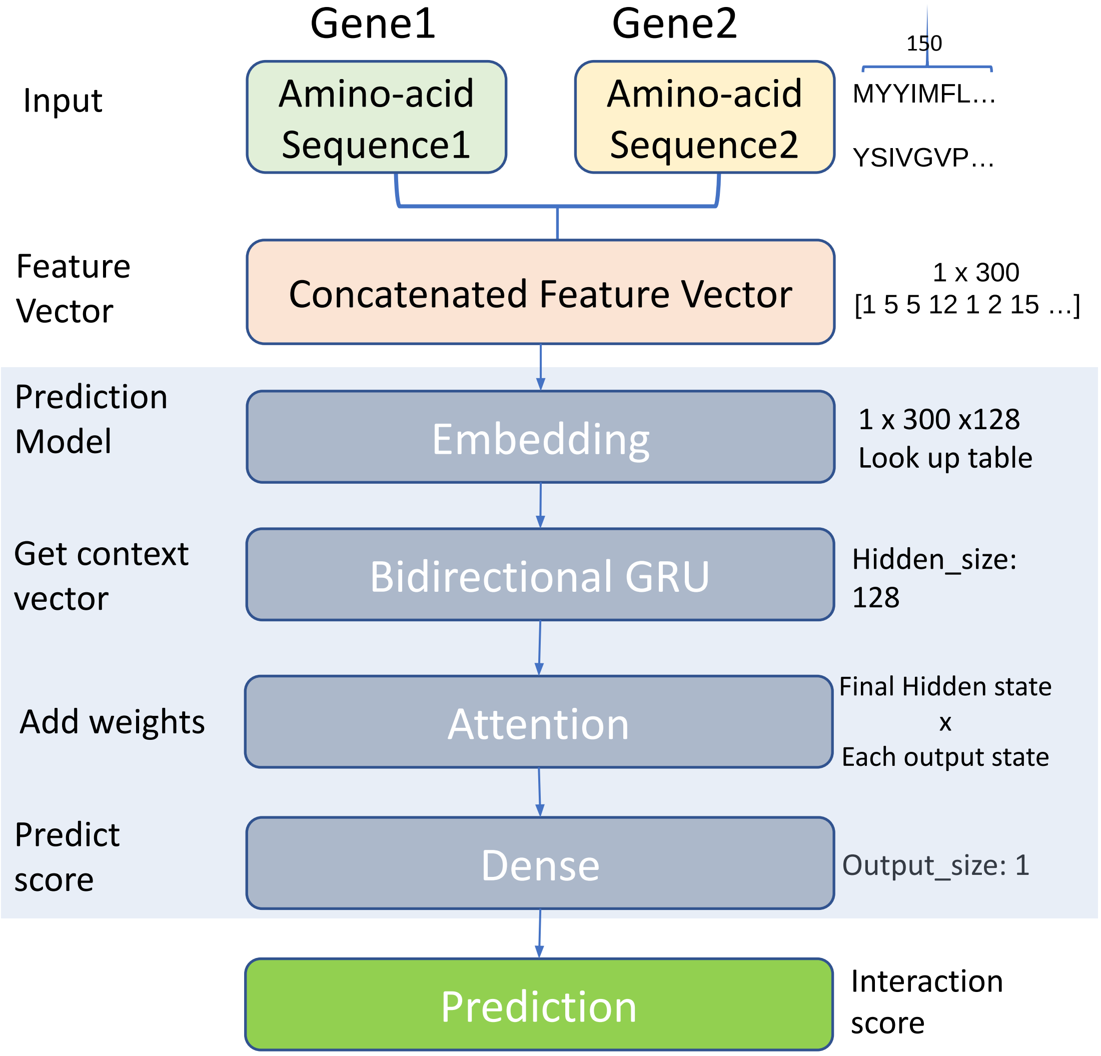
Predicting Phenotype from Genomic Sequence with Deep Neural Networks

Motivation

- Available DNA sequence data and no complex/hand-picked features
- Avoid influences of errors in existing term annotations (Yu M. et al. (2016))
- Gated RNNs (LSTMs and GRUs) are successfully used in predicting protein functions, enhancer-promoter interactions (SPEID) and quantifying the function of DNA sequences(DanQ)

Network Architecture – Attention based GRU



Gene1 Gene2

Input: Amino-acid Sequence1 | Amino-acid Sequence2

150

MYYIMFL…

YSIVGVP…

Feature Vector: Concatenated Feature Vector

1 x 300
[1 5 5 12 1 2 15 …]

Prediction Model: Embedding

1 x 300 x128
Look up table

Get context vector: Bidirectional GRU

Hidden_size: 128

Add weights: Attention

Final Hidden state
x
Each output state

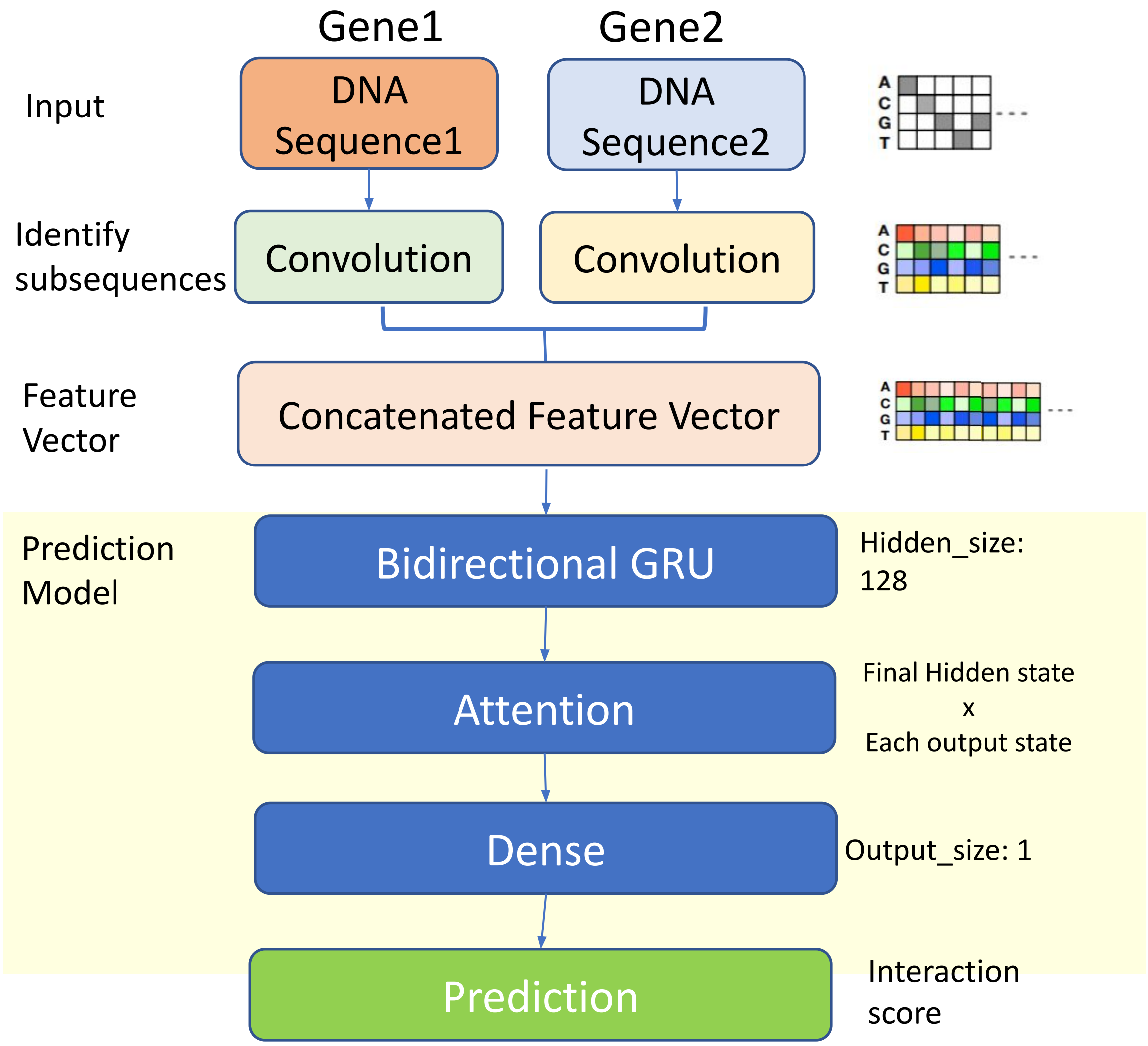
Predict score: Dense

Output_size: 1

Prediction

Interaction score

Model 1



Gene1 Gene2

Input: DNA Sequence1 | DNA Sequence2

A C G T ---

Identify subsequences: Convolution | Convolution

A C G T ---

Feature Vector: Concatenated Feature Vector

A C G T ---

Prediction Model: Bidirectional GRU

Hidden_size: 128

Attention

Final Hidden state
x
Each output state

Dense

Output_size: 1

Prediction

Interaction score

Model 2

Experiments & Results

Interaction Data: Collins, et al. (Nature, 2007)

Genomic Sequences: Genome Browser database

Number of Unique Genes: 664

Table.1 Pearson correlation coefficients

Method	GO (Yu et.al)	Model 1	Model2
Pearson Correlation	0.479	0.372	0.344

Table.2 Classification results

	Precision			Recall			F1-score		
	Yu et.al	M1	M2	Yu et.al	MI	M2	Yu et.al	MI	M2
Negative	0.56	0.56	0.54	0.29	0.29	0.10	0.38	0.38	0.17
Non-interaction	0.93	0.72	0.92	0.98	0.94	0.95	0.96	0.82	0.95
Positive	0.58	0.80	0.00	0.07	0.00	0.00	0.13	0.01	0.00

M1: 20 epochs with 1/5 non-interaction data
M2: 15 epochs with all non-interaction data

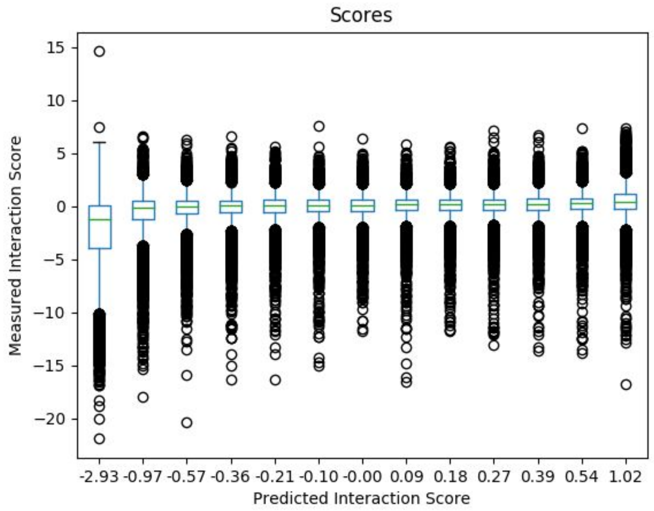


Fig.1 Measure vs. predicted (Yu et.al)

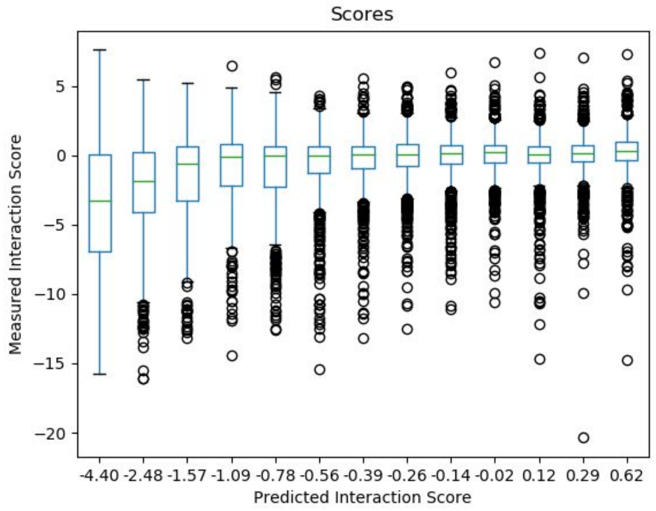


Fig.2 Measure vs. predicted (Model 1)

Results

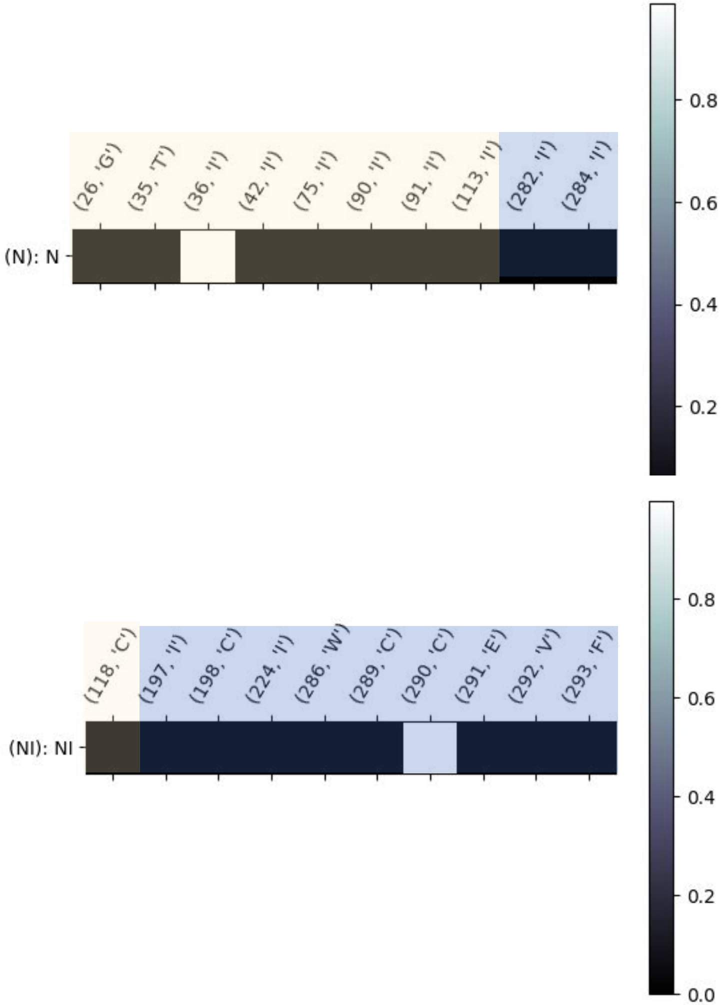


Fig.3 Attention analysis

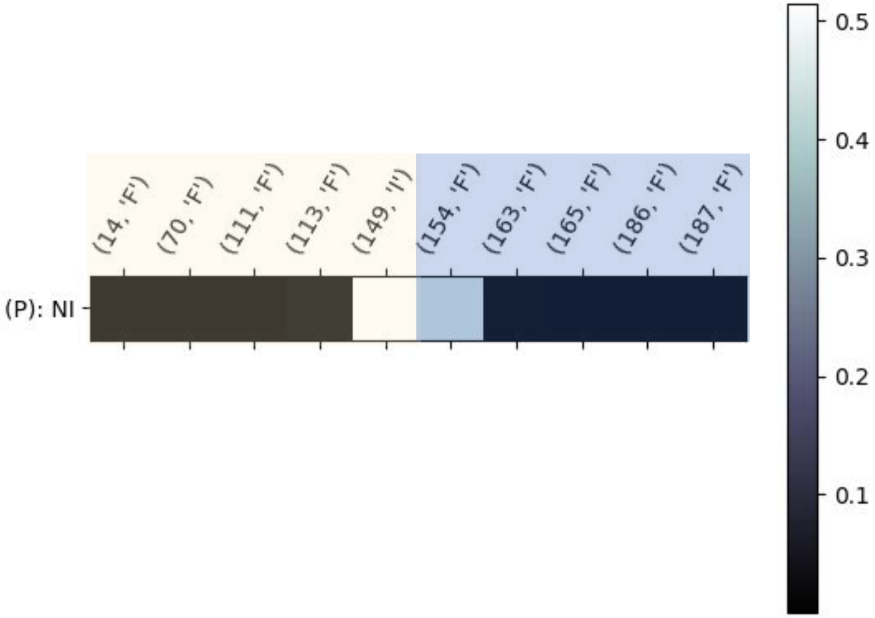


Fig.3 Prediction Failure case

Future Directions

- Improving results :
  - Data: Costanzo, et al. (Science, 2010)
  - Model: LSTM, fine-tuning parameters, convolution after embedding
  - Validation: Cross Validation
- Use Prot2Vec embeddings/pretrained embeddings
- Pairwise interaction on the specific base/amino-acid/protein
- Biological interpretation
- Inter-species genetic interactions prediction