Predicting Phenotype from Genomic Sequence with Deep Neural Networks

Wenyan Li, December 8, 2019

Motivation

A major problem in genetics is to understand how different DNA sequence variations can yield to similar phenotypes." (Waddington, 1942). In yeast genetic interactions, the deletion of two or more genes results in an unexpectedly slow or fast cellular growth phenotype. Detection of negative (i.e. synthetic lethal) interactions between mutated oncogenes or tumour-suppressor genes and target proteins may provide inspiration for cancer treatments development. In Yu M. et al. (2016), phenotype is translated from genotype based on gene ontology, and predicted interaction scores may be influenced by errors in gene annotations or relationship between terms. As deep learning being effective in identifying complex patterns from feature-rich datasets, especially as recurrent neural networks(RNNs) such as long short term memory(LSTM) and gated recurrent unit(GRU) are capable of dealing with long-distance sequential data, predicting genetic interactions directly from DNA or amino-acid sequences using deep learning techniques would help us gain insights into underlying complex phenotypes, including cancer and other common diseases.

Objective and Research Methods

In order to translate phenotype directly from genomic sequences, the problem is quantified as predicting genetic interaction scores from pairwise amino-acid/DNA sequences. Following research questions need to be addressed: (1) data representation, (2) neural network structure of the prediction model, (3) result validation.

Data Representation

Amino-acid sequences For a pair of genes, amino-acid sequences with fixed lengths are extracted for each gene and are then converted into indexes and concatenated as feature vectors. Character embedding, as used in [5], of the feature vector is used for its dense representation. The embedded pairwise amino-acid sequence representing the gene pair is then fed into the interaction prediction model as the input.

DNA sequences Similar to the procedure in [1], DNA sequences with fixed lengths are extracted for each gene. In order to obtain low dimensional and dense representation vectors, the DNA sequences are converted with one-hot encoding and then fed into a convolution layer, a rectified linear unit (ReLU) and an activation layer which together extract subsequence features from the input, followed by a pooling layer, which reduces dimensionality.

Neural Network Structure

Encoder As gated RNNs such as LSTMs and GRUs being capable of learning long-term dependencies, and comparing to LSTM, GRUs are computationally easier and have on par performance. Thus, the dense representation vector $V = [V_1, \ldots, V_T]$ is fed into a GRU and encoded into a "context vector" O_T.

Attention When input sequence is long, the summaries vector O_T is likely to contain noisy information from many irrelevant features, we thus apply attention mechanism to obtain a

weighted context vector. The attention score A = [a_1, ..., a_T] of the sequence is calculated by using the "dot production method", which computes the cosine similarity between the sequence vector O_T and the representation vector V according to the algorithm proposed in [3].

Dense The output layer is a linear layer which outputs the predicted interaction score with the input weighted context vector.

Loss function As we want to predict the exact interaction score between each gene pair, mean squared error (MSE) between the predicted interaction scores and the targets are calculated and minimized during training. Classification of the interactions can then be done with the predicted scores.

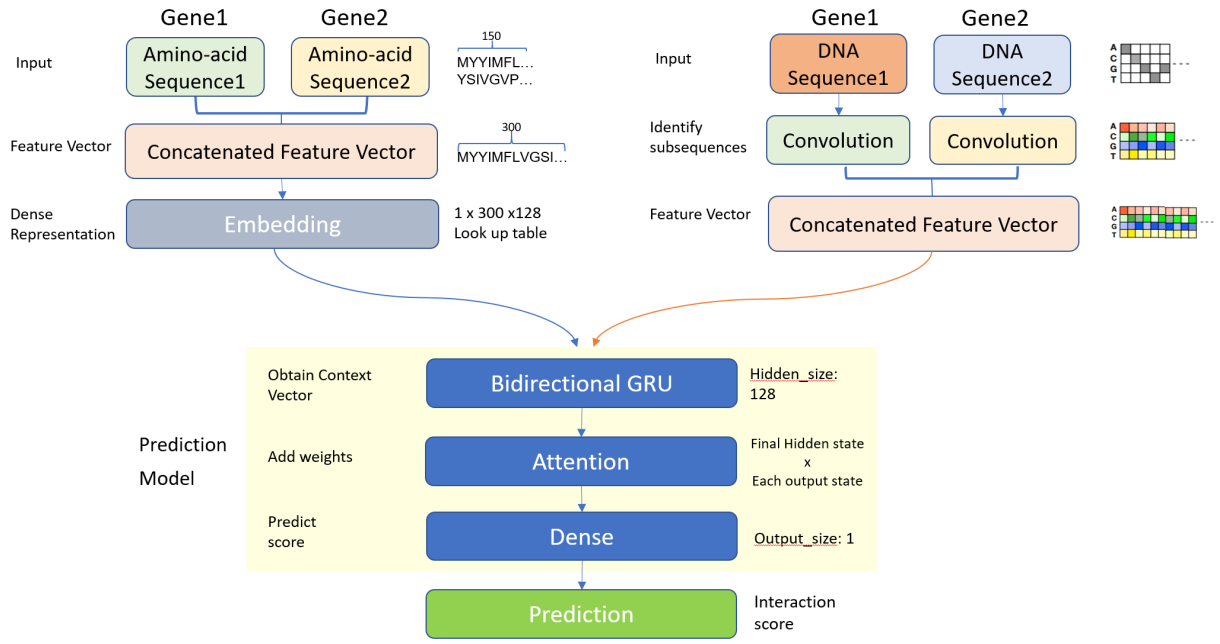

Figure.1 Diagram of the attention-based model to predict genetic interactions based only on genomic sequences

Result validation

The results of the prediction are verified on the hold-out set of genetic interactions data from Collins, et al. (Nature, 2007). Pearson correlation coefficient and precision and recall, as well as the F1 score of the classification result are further calculated and analyzed for evaluation.

Training Procedure

Data preparation

Genomic sequences of S. cerevisiae (baker's yeast) genes, downloaded from Genome Browser database, are used as raw data inputs and are further processed to obtain representation vectors. Genetic interactions data from Collins, et al. (Nature, 2007) which contains 150309 pairwise gene interaction scores of 664 unique genes are used as our targets. As the the negative interaction (i.e. synthetic lethal) class greatly outnumbers the other two interaction classes. In the experiment with amino-acid sequences, only one fifths of the negative interaction samples are randomly selected into the dataset as suggested in [4]. However, in the experiment with DNA sequences, in order to have a direct compare with the method proposed in Yu M. et al. (2016),

the full dataset is used. Our datasets are then splitted into training(64%), evaluation (16%) and testing(20%) sets.

Parameter selection

Referring to the model parameters and sequence lengths used in [1] and [4], amino-acid sequence of length 150 is used for representing each gene, thus the input of the encoder is a vector of length 300. The embedding size is set as 128. The length for DNA sequences is selected to be 1000. For the convolutional layer and the max-pool layer, we set kernel size as 200, filter length as 40, pool length as 20, stride as 10. For the prediction model, we use an embedding size of 128, a drop out rate of 0.4, and a 2-layer bidirectional GRU. Mini-batching is used with a batch size of 64 and we choose Adam for optimizing with a default learning rate of 0.001. When training on DNA sequences, we also set weight decay to 10^{-5} as suggested in [1]. Due to limitations on time and computation resources, we train our prediction model for 20 epochs on amino-acid sequences and 15 epochs on DNA sequences. Note that the parameters are not fine-tuned and the number of samples and epochs are not well controlled and balanced when using different sequences, which may lead to bias in the results and hence should be further experimented and studied.

Results

Consistent with Yu M. et al. (2016), pearson correlation coefficient between the predicted and true genetic interaction scores is used to evaluate the model. The result and comparison between different methods/sequences are displayed in Table.1. Measured genetic interaction scores versus predicted scores using amino-acid and DNA sequences are shown in Fig.2 and Fig.3. Compared to model proposed in Yu M. et al. (2016), which uses random forests and feature vectors constructed using Gene Ontology(GO), our proposed method achieves a reasonable correlation score with direct use of genomic sequences within limited training epochs, and reaches on par performance on classifying synthetic lethal and gene-pairs with non-interaction. It can also be observed that the proposed method performs relatively poorly on predicting positive interactions, which may be caused by too few positive training samples, making it very hard for the neural network to learn within limited time.

Table. 1 Pearson correlation and classification results of different models and input sequences

	Yu et.al			**Our Model(Attention-based GRU)**					
Input	Feature vector based on Gene Ontology			Amino-acid Sequence			DNA Sequence		
Pearson Correlation	0.479			0.372			0.344		
Classification	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1score
Negative	0.56	0.29	0.38	0.56	0.29	0.38	0.54	0.10	0.17
Non-interaction	0.93	0.98	0.96	0.72	0.94	0.82	0.92	0.95	0.95
Positive	0.58	0.07	0.13	0.80	0.00	0.01	0.00	0.00	0.00

With the use of attention mechanism proposed in [3], the underlying relationship between phenotype and bases/proteins can be extracted and analyzed from the attention weights assigned to each element in the sequence. As shown in Fig.4, top ten positions and amino acids of the gene pairs are extracted and it can be noticed that for some gene pairs the model focuses on the

specific amino-acid in both genes during the prediction process. Attention used for DNA bases can also be obtained though could be more complex, as we can only directly get the attention weights added to the subsequences after max pooling. But it's still possible to trace back to the position of the base in the original DNA sequence as the stride and filter length are fixed.

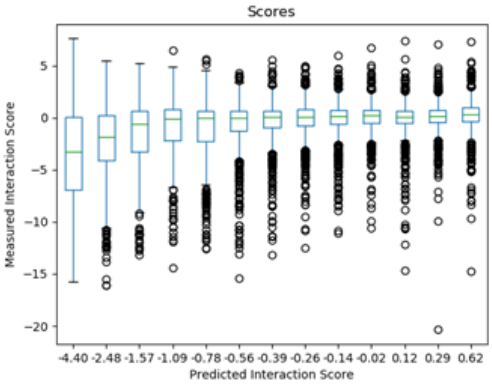

Figure.2 Measured genetic interaction scores versus predicted scores using amino-acid sequences

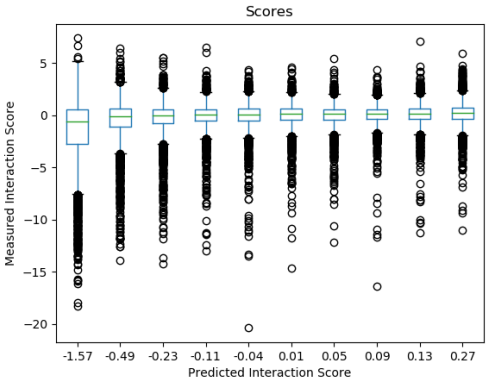

Figure.3 Measured genetic interaction scores versus predicted scores using DNA sequences

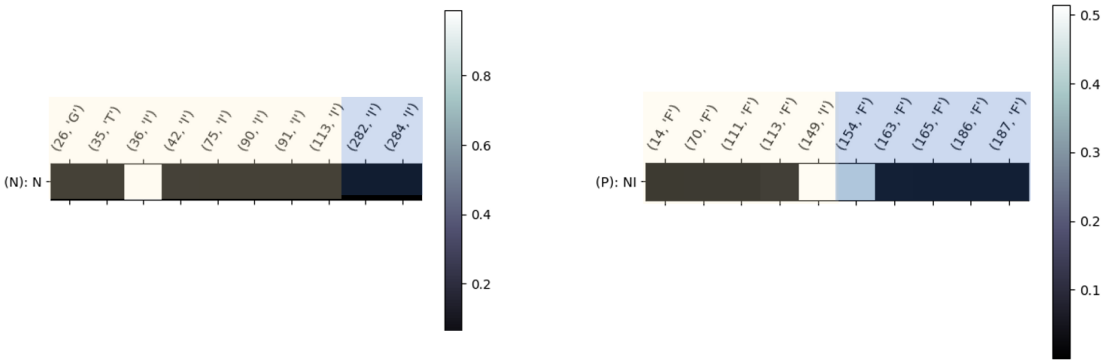

Figure. 4 Attention display for two gene pairs

Limitations and Future Directions

Considering the parameter selection and data processing procedures, there are still limitations that should be overcomed. Amino-acid and DNA sequences for S. cerevisiae (baker's yeast) genes vary greatly in lengths before processing. Hence, the sequences are cut off or padded with zeros to a fixed length as processed in [4]. The noises and information loss introduced by this procedure are ignored in our experiment, however better solutions for this problem should be further studied. We also assume concatenation order of the gene sequences are symmetrical for a gene pair, which may not be true and should be experimented and validated in future work. Meanwhile, as the current dataset used for training the prediction model is curated and highly unbalanced, i.e. with much more non-interacting gene pairs than the other two kinds. Training data need to be augmented properly to ease the problem. The proposed method should be examined and validated on a larger dataset such as Costanzo, et al. (Science, 2010) and parameters of the prediction model remain to be fine-tuned. There are also minor modifications

which may help with the result, such as changing GRU into LSTM, and using convolution after embedding rather than replace it. As mentioned in the previous section, the application of attention mechanism would allow us to reveal the corresponding biological interpretation on genetic interactions. Thus, detailed analysis is expected to follow up and pair-wise interaction on a base/amino-acid level should be considered.

Moreover, as we trained our model in an end-to-end manner, character-level embeddings for the amino-acid sequences and the dense representation for the DNA sequences are trained on-the-fly. The use of pretrained embeddings such as ProtVec and GeneVec proposed in [5] may lead to future improvements, and the 3-gram modelling for studying the sequences is also inspiring.

Due to time constraints, we only validate our results on the hold out data. Cross-validation as was done in Yu M. et al. (2016) can also be helpful.

Additionally, with the efforts that synthetic lethality is able to be predicted using interspecies networks and connectivity homology as declared in [7] and zero-shot translation is achieved in [6], inter-species genetic interactions prediction may be achieved by modifying the current model and combine it with the single translation model in [6] into a regression or classification model.

References

[1] Singh, Shashank, et al. “Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks.” Feb. 2016, doi:10.1101/085241.

[2] Daniel Quang, Xiaohui Xie; DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences, Nucleic Acids Research, Volume 44, Issue 11, 20 June 2016, Pages e107, https://doi.org/10.1093/nar/gkw226

[3] Shen, S., & Lee, H. (2016). Neural Attention Models for Sequence Classification: Analysis and Application to Key Term Extraction and Dialogue Act Detection. Interspeech 2016. doi:10.21437/interspeech.2016-1359

[4] Xueliang Liu. Deep recurrent neural network for protein function prediction from sequence. arXiv preprint arXiv:1701.08318, 2017.

[5] Asgari E, Mofrad MRK. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. Kobeissy FH, ed. PLoS ONE. 2015.

[6] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s multilingual neural machine translation system: Enabling zero-shot translation. CoRR, abs/1611.04558, 2016. URL http://arxiv.org/abs/1611.04558.

[7] Jacunski A, Dixon SJ, Tatonetti NP (2015) Connectivity Homology Enables Inter-Species Network Models of Synthetic Lethality. PLoS Comput Biol 11(10): e1004506.

Appendix

Training loss plots for using amino-acid and DNA sequences.

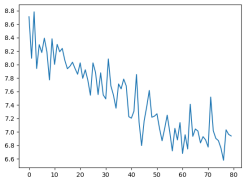

Figure. A1 Training loss for amino-acid sequence

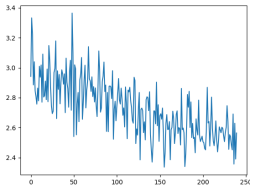

Figure. A2 Training loss for DNA sequence