Neural Network Approaches to DNA Sequence Denoising

Christine Tataru, Abhishek Roushan, Clara McCreery
{ctataru5,aroushan,mccreery}@stanford.edu

Deep Learning

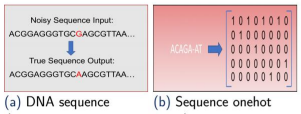
STANFORD

Background

- Noise introduced to DNA sequences from technological errors is a significant problem in the field of microbiome research.
- The gut microbiome is extremely diverse; difficult to distinguish nucleotide variations due to sequencing error from bacterial evolution
- Current techniques to address this problem (eg. alignment graph consensus) scale poorly

Introduction

- Few nucleotide differences between 16S sequences->1000 years of evolutionary change
- Deep learning approach to resolve sequence noise without reference sequence.
- Investigate neural network models for converting noisy sequences to their denoised counterparts



(a) DNA sequence denoising (b) Sequence onehot encoding

Figure: Sequence Denoising and Encoding Representation

Data Acquisition

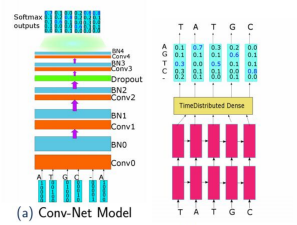
- Synthetic data generated by experts in the field
- 12% of the 500,000 sequences in data set are noisy
- Common sequence length is 250 nucleotides
- A "-" was used to align sequences despite insertions and deletions

Model Input Data

- Train set: 112,089 sequences (50% noisy) (more noise in Train to speed up training)
- Val set: 12,269 sequences (12% noisy)
- Test set: 12,107 sequences (12% noisy)

Architecture Overview

- Investigated convolutional & recurrent approach
- Trained with fair share of noisy+exact sequences on shallow architectures and iterate quickly.
- Consistent train/val/test data for fair comparison of CNN and LSTM architectures

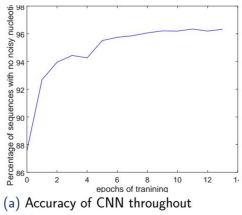


(a) Conv-Net Model (b) RNN model

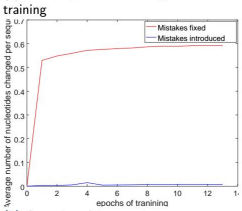
Figure: Architectures for training data

Convolutional Approach

- 5 layers (each with BN, ReLU, kernel size = 7)
- Dropout(0.3) after layer 3
- Plateaued after 12 epochs



(a) Accuracy of CNN throughout training



(b) Correct and incorrect nucleotide changes

Reccurent Approach

- Many-to-many VS encoder-decoder architecture
- Winner: 2-layered Bidirectional LSTM (latent dim=100, Inter-layer Dropout [p=0.5])
- 'Adam' optimizer, categorical cross-entropy loss

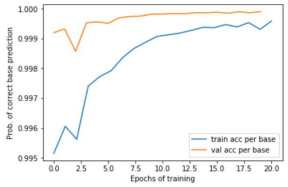


Figure: Prob. of correct base prediction throughout training

Results

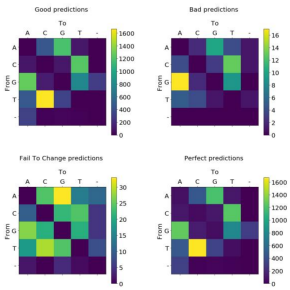


Figure: LSTM model predicts mostly G-A and C-T subs, and fails uniformly for most subs.

- The convolutional network converted a data set in which 12.5% of sequences were noisy to an output in which 3.7% of sequences had some noise. The recurrent network produced output where .095% of sequences had some noise.

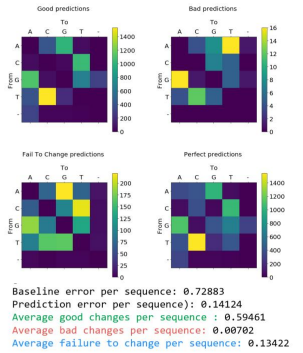


Figure: Conv model 1. predicts mostly G-A and C-T subs 2. Fails less uniformly but with similar frequency to recurrent model 3. Outputs more error per sequence than recurrent model.

Conclusion

- Both networks reduce the number of incorrect nucleotides on our data set; Recurrent architecture seems more successful.
- Each tended to incorrectly change a small number of noiseless sequences incorrectly
- The correct & incorrect substitutions by the model-> G-A or C-T (consistent with baseline). Occasional A-C and G-T; rare A-T or C-G.

Future Work

- Generalization of the network architecture and learned weights to other data sets
- Hyperparameter tuning to get depth of layer, activations etc.

References

- Convolutional Sequence Modeling
- Sequence to sequence learning keras