Unsupervised Representation Learning of DNA Sequences

Vishal Agarwal | N. Jayanth Kumar Reddy | Dr. Ashish Anand
Indian Institute of Technology Guwahati, India
vishalagarwal.jss@gmail.com | jayanth.reddy@iitg.ac.in | anand.ashish@iitg.ac.in

Motivation

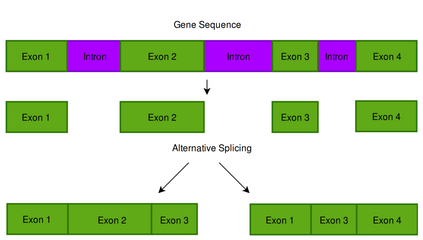
- Gene sequences are very long and variable length sequences extending upto millions in length.
- Therefore generally a small fixed length window is used around splice junctions which captures local neighbourhood splicing signals.
- Fails to capture global splicing signals and long range dependencies.

Objective

- Learn fixed-length latent representation of gene sequences in an unsupervised manner.
- Evaluate learned representations on a supervised task of splice site prediction in two different settings.
- Model attribution using a known axiomatic approach, *Integrated Gradients*, to identify important regions and motifs which influence splicing.

Introduction

- Splicing : Removal of introns and combining exons from exon-intron interface in genes to form proteins.
- Alternative Splicing : Regulated process where introns and exons are alternatively joined to generate more than one mRNA.
- Responsible for protein diversity in the body.
- Mis-splicing may lead to genetic disorders.



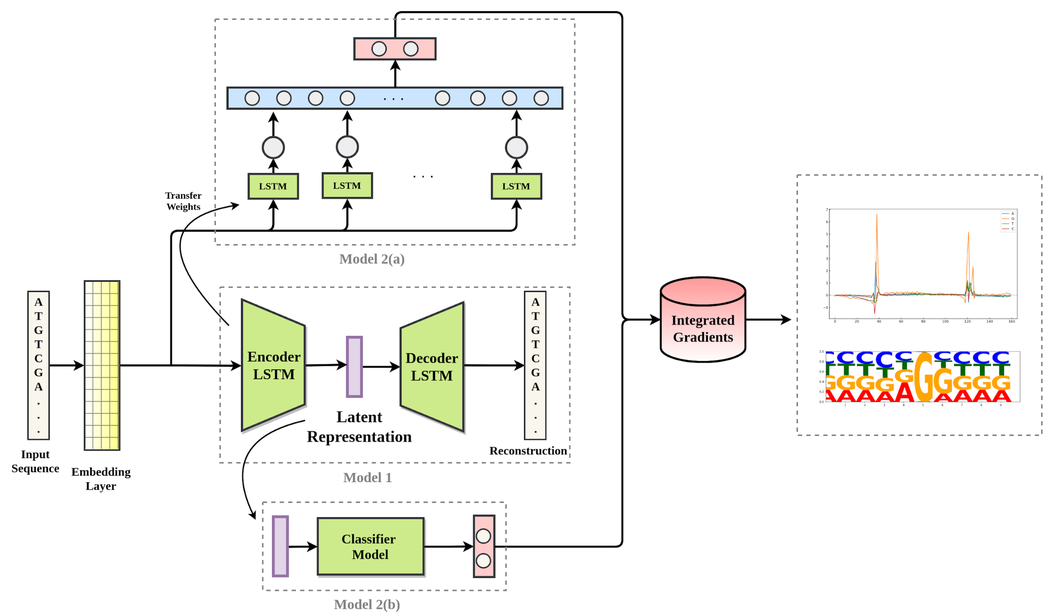
Proposed Approach

- Using sequence-to-sequence autoencoder model to learn latent embeddings.
- The model consists of an encoder BiLSTM and a decoder LSTM.
- The encoder maps input sequence to a latent representation which is then used by the decoder to reconstruct back the actual input sequence.
- Quantitative and Qualitative analysis to evaluate learned representations.
- Quantitative analysis on a supervised task of splice site prediction in two settings.
- Qualitative analysis by axiomatic approach known as Integrated Gradients.
- We use GENCODE annotations based on human genome data GRCh38 to prepare datasets for all experiments.

Experiment Setting

- For false data generation, the consensus dimer GT and AG is searched randomly such that both donor and acceptor are in the same chromosome and its distance is in the range of true data length range.
- A LSTM model initialized with learned encoder weights instead of random initialization for supervised task.
- Using learned latent embeddings as features for simple classifiers such as SVM, Feedforward network, vanilla RNN.
- Model attribution using Integrated Gradients to infer important regions in gene sequences.
- Using model attribution score to generate sequence logo and identify critical motifs.

Graphical Illustration of the entire work



Results

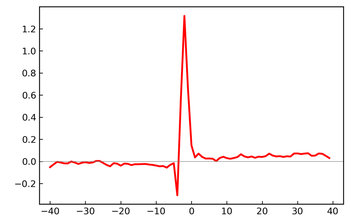
Experiment 1

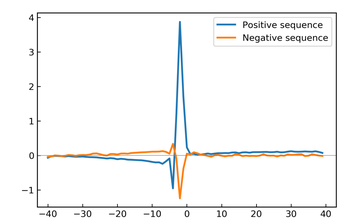
	Model	Accuracy
Random Weights	LSTM	95.43%
	BI-LSTM	96.04%
	BI-LSTM Attention	97.23%
Autoencoder initialized	LSTM	98.54%
	BI-LSTM	98.60%
	BI-LSTM Attention	99.07%

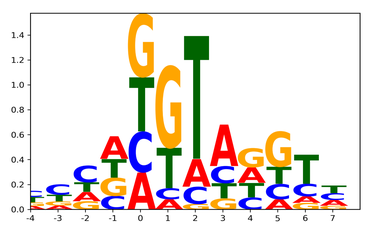
Experiment 2

Model	Accuracy
SVM	98.63%
Feedforward Network	98.88%
Vanilla RNN	98.93%

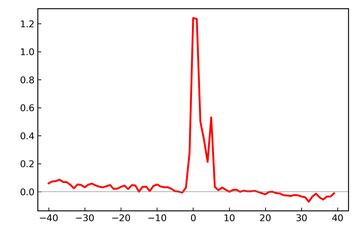
Donor

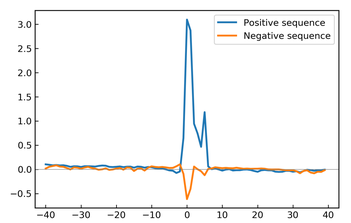


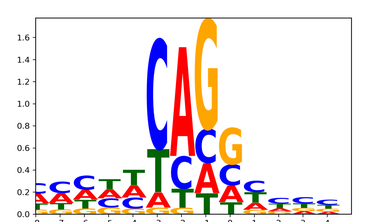




Acceptor







References

1. Ilya Sutskever, Oriol Vinyals and Quoc V Le. Sequence to Sequence learning with neural networks, NIPS 2014.
2. Nitish Srivastava, Elman Mansimov and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms, ICML 2015.
3. Mukund Sundararajan, Ankur Taly and Qiqi Yan. Axiomatic attribution for deep networks, ICML 2017.