Unsupervised Representation Learning of DNA Sequences

Vishal Agarwal[1*], **N. Jayanth Kumar Reddy**[1] and **Ashish Anand**[2]
[1]Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India
[2]Dept. of Computer Science and Engineering, Indian Institute of Technology Guwahati, India
{vishal.agarwal, jayanth.reddy, anand.ashish}@iitg.ac.in,

Abstract

Recently several deep learning models have been used for DNA sequence based classification tasks. Often such tasks require long and variable length DNA sequences in the input. In this work, we use a sequence-to-sequence autoencoder model to learn a latent representation of a fixed dimension for long and variable length DNA sequences in an unsupervised manner. We evaluate both quantitatively and qualitatively the learned latent representation for a supervised task of splice site classification. The quantitative evaluation is done under two different settings. Our experiments show that these representations can be used as features or priors in closely related tasks such as splice site classification. Further, in our qualitative analysis, we use a model attribution technique *Integrated Gradients* to infer significant sequence signatures influencing the classification accuracy. We show the identified splice signatures resemble well with the existing knowledge.

1 Introduction

Recently there is a surge in studies using deep learning models for DNA sequence based classification tasks. One of the primary reason for the adoption of such methods is representation learning or feature learning from raw data. In the case of DNA sequence based classification tasks, DNA sequence containing 4 nucleotides A, T, G, C constitute raw data. Most studies often choose fixed-length DNA sequences as input by choosing a context window. However, in many cases, important nucleotides may not lie within the same context window size in all input sequences. Hence, there is a requirement of models which can handle long as well as variable length DNA sequences as inputs. Such a model can then take into account of both short (local) and long (global) range dependencies.

In this work, we primarily focus on learning representation for long, variable length DNA sequences using sequence-to-sequence based autoencoder. We evaluate our model on splice site classification task. In genomics, splicing is an important phenomenon, leading to protein diversity in the body. We performed two quantitative and a qualitative evaluation of the learned latent representation of input DNA sequence. The quantitative evaluation is carried in two different settings. For qualitative analysis, we use Integrated Gradients, a model attribution technique proposed by [Sundararajan *et al.*, 2017]. This provides attribution of input feature to the predicted classification score and identify relevant region and motifs influencing splicing.

*Contact Author

2 Model Description

We use an autoencoder-like sequence-to-sequence model to learn fixed-length representations of sequences. The model consists of an encoder and a decoder LSTM inspired by [Sutskever *et al.*, 2014]. The encoder network uses a bidirectional LSTM to process the input sequence from both ends and map it to a fixed-length embedding, summarizing the input sequence which captures important motif information. The decoder network uses a unidirectional LSTM to reconstruct the input sequence using the latent embedding only. The motivation behind this is to capture relevant features that summarize the input sequence well-enough to be able to reconstruct it back.

3 Experiment and Results

The quantitative analysis of learned representations is done on a supervised task under two different settings. First, an LSTM is trained for splice site prediction in a DNA sequence. Instead of initializing the LSTM with random weights, it is initialized with the trained encoder weights to add apriori information. This provides a good starting point for the discriminative model to converge faster and improves classification accuracy. We compare this model with a baseline model of similar architecture but randomly initialized parameters. Table 1 shows the former model performs better. We also experimented with different architectures such as LSTM, bidirectional LSTM and bidirectional LSTM with Attention and compared the results. Table 1 shows the comparison of different types of models.

In the second evaluation setting, the latent embeddings are used as features on the same task of splice site identification. We use Support Vector Machine(SVM), 2-layer Artificial Neural Network(ANN) and a vanilla Recurrent Neural Network(RNN) model to conclude the effectiveness of latent

	Model	Accuracy
Random Weights	LSTM	95.43%
	Bi-LSTM	96.04%
	Bi-LSTM Attention	97.23%
Autoencoder Initialized	Bi-LSTM Attention	99.07%

Table 1: Classification accuracy for encoder initialized LSTM model

Model	Accuracy
SVM	98.63%
ANN	98.88%
Vanilla RNN	98.93%

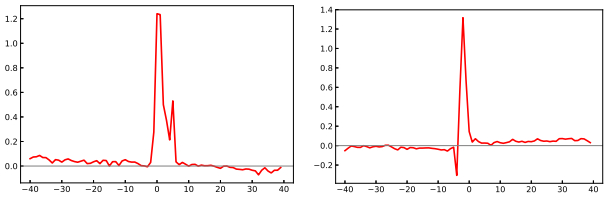
Table 2: Classification accuracy for simple classifier model using latent embedding

representations. If it did capture motif information, then we expect the classifier to perform well. The results for this setting are shown in table 2.

We use a popular visualization technique - Integrated Gradients proposed by [Sundararajan *et al.*, 2017]. The visualizations provide model attribution by identifying important regions in the sequence. These can be interpreted as motif signals which influence splicing.

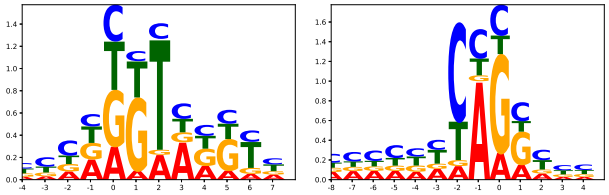
Integrated gradient requires a baseline against which it compares the prediction of the network and accordingly provides attribution to the feature which differs in the input and the baseline. In our case, we chose the baseline as zero embedding matrix. It calculates the attribution score by accumulating gradient of network prediction score with respect to the embedding at each point in the straight-line path from baseline to input, then multiplied by the difference in the baseline and input feature value. In our experiment, we used 200 steps for gradient calculation along the path.

For visualization, we take 40nt sequence window upstream and downstream around both at acceptor and donor sites. Figure 1(a) and 1(b) shows the attribution score averaged over all data for donor and acceptor respectively. It is evident that the nucleotides closer to the sites influenced the model most, which confirms with the existing knowledge. Also, we generate a sequence logo from the attribution score to identify important motifs or splicing signals. Figure 2(a) and 2(b) shows the sequence logo for both donor and acceptor around the most relevant region given by the attribution score. The



(a) Donor (b) Acceptor

Figure 1: Average attribution score per position



(a) Donor (b) Acceptor

Figure 2: Sequence logo to visualize important motifs attributed by the model

donor results show the presence of strong GT signal and the acceptor results shows the presence of strong AG signal. This validates the known consensus motif.

4 Conclusion

In this work, we presented an unsupervised representation learning approach to learn representations of DNA sequences in a latent space. We leveraged deep learning techniques to use a sequence-to-sequence autoencoder-like framework to learn representations in an unsupervised setting. We exploit this autoencoder model in two ways: first the learned weight parameters of this model can be used to initialize a classifier with similar architecture, and second, latent representation was used as input features for three different classifiers SVM, ANN and vanilla RNN.

The results indicate that the use of pre-trained weight parameters help in faster convergence with improved accuracy. Furthermore, our analysis shows that the learned latent embeddings are good features as three different classifiers gave similar performance using it as input feature.

Finally, attributional analysis shows that the model is able to pick significant regions, confirming with the existing knowledge, of input DNA sequence for the splice site classification task.

References

[Srivastava *et al.*, 2015] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 843–852. JMLR.org, 2015.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.