Cell type-agnostic representation of the human epigenome through a deep recurrent neural network

Kevin B. Dsouza, Adam Y. Li, Vijay K. Bhargava, Maxwell W. Libbrecht

1 Introduction

Sequencing-based assays such as ChIP-seq and ATAC-seq have recently been used to characterize the epigenome of hundreds of human cell types. These assays detailing a varied number of epigenomic functions like methylation status, local chromatin accessibility, histone modifications, factor binding and chromatin structure are hosted by consortia such as Roadmap Epigenomics [1] and ENCODE [2]. These data sets necessitate integrative methods that summarize them into a useful representation. A popular existing type of method is segmentation and genome annotation (SAGA) algorithms such as Segway [3] and ChromHMM [4], which produce an annotation of the epigenome of a given cell type.

Existing SAGA annotations are cell type-specific; that is, they annotate activity in a given cell type. This corresponds poorly to most definitions of genomic elements, which are relative to the genome sequence itself. For example, annotations of protein coding genes are cell type-agnostic: they contain an archetypal set of gene locations where the locations are fixed and only the activity varies across cell types. Moreover, connecting a genetic locus to a phenotype or disease requires a cell type-agnostic understanding of its function. Existing SAGA algorithms cannot be adapted for this task because they use simple discrete or linear models that cannot capture the complexity of the epigenome across all cell types.

We propose a method that produces a cell type-agnostic low-dimensional representation of the epigenome. This representation assigns a vector of features to each genomic position that represents that position's activity across all tissues. We do this using a deep long short-term memory (LSTM) [6] recurrent neural network autoencoder to reduce all existing epigenome data into a single low-dimensional representation. This LSTM uses an autoencoder architecture in which aims to produce a representation that can be used reconstruct the original data as accurately as possible.

One similar neural network representation learning method for epigenetics data exists. Like our method, Avocado produces a representation of the epigenome that assigns a low-dimensional vector to each genomic position. Avocado was initially developed for imputation. It uses distinct embeddings for each cell type, assay type and genomic position, and couples this with a feed-forward neural network that imputes unperformed assays. Our method has two advantages relative to Avocado. First, we use a sequential model that captures the spatial relationship of neighboring genomic positions. Second, our method can be applied genomic positions that were not used in training by inputting the relevant data into our encoder, while Avocado must use an expensive iterative optimization to do so.

We demonstrate the utility of this representation through several analyses. First, we show that this representation simultaneously captures cell type-specific activity across many cell types, including gene expression, replication timing and chromatin contacts. We do this by demonstrating that all of the above phenomena are accurately predictable using just the latent representation. Second, we show that this latent representation distinguishes functional and non-functional regions by showing that the representation accurately identifies conserved regions. Third, we demonstrate

how a sequential model leads to smoother and more interpretable representations than existing methods, which do not capture dependence among neighboring positions.

1.1 Methods

1.2 LSTM model

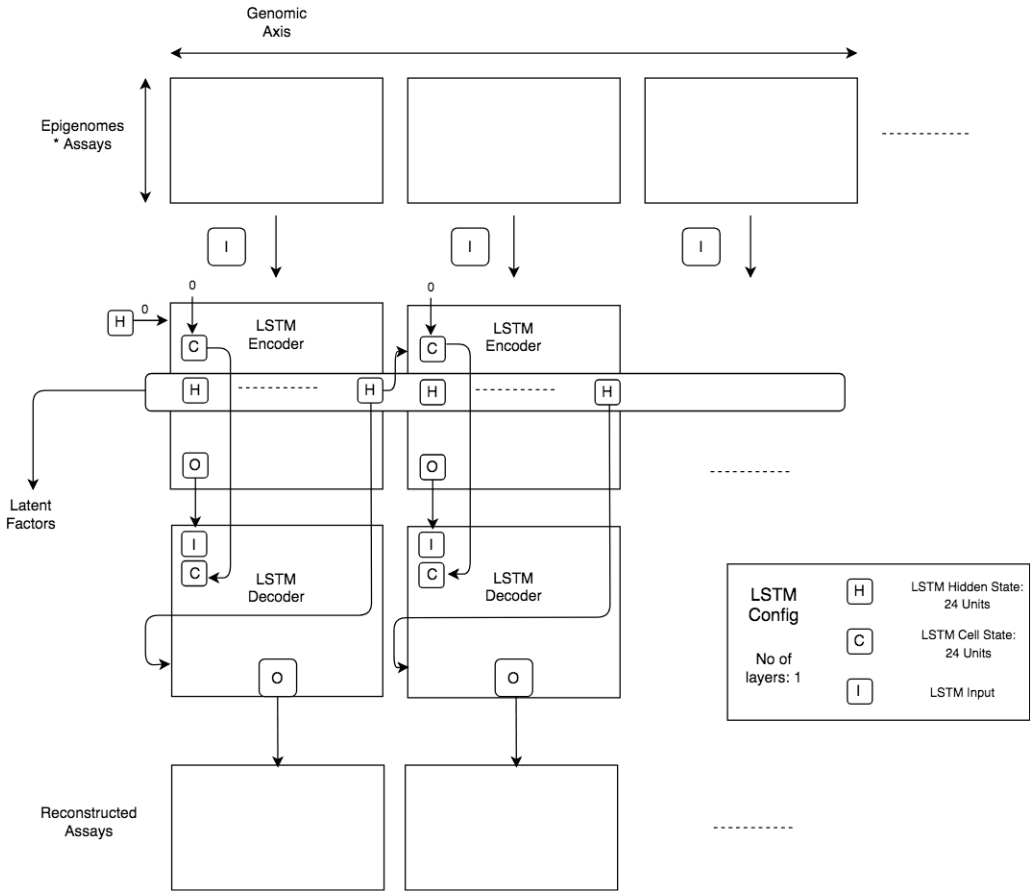

Figure (1) LSTM Autoencoder Framework. The Encoder and the Decoder are LSTM's with the given configuration. The assays that are serving as the input to the Encoder are arranged in a matrix of (Epigenomes*Assays) × Genome Length and fed in one Frame Length at a time (which is a design parameter).

The LSTM maintains a representation of long term dependencies and because of this ability it serves to be a good candidate for modelling sequential data. Using these internal representations of the LSTM, we can recreate the original sequential input in a encoder-decoder framework [7], which forms the backbone of advances in fields like language modelling [8], speech recognition [9], sequence-to-sequence prediction [10] and neural machine translation (NMT)[11]. The two stage framework consists of an encoder that creates an internal representation of the sequence that it sees, and a decoder that uses this representation to recreate the original input. The framework is as given in Figure 1.

1.2.1 Encoder

The first stage of the framework is an LSTM that acts as an encoder. The encoder reads the input sequence and creates a fixed length vector representation in the form of an embedding. This embedding can then be used for various tasks related to the input sequence like recreating the input sequence as in auto-encoding.

1.2.2 Decoder

The decoder uses the fixed length vector embedding as it's initial hidden state and tries to recreate the original sequence. Along with the initial hidden state seed, the decoder, at each step, receives as input the output of the encoder and its cell state. It then uses these states to output the original input at each step using mean squared error (MSE) as the loss function.

2 Results

2.1 Latent factors capture many types of genomic activity

We find that our LSTM model with cell-type agnostic features performs at par with Avocado's Deep Tensor Factorization model [5] (not shown) in classifying four important cell-type specific genomic phenomena, namely: gene expression, promoter-enhancer interactions, replication timing and frequently interacting regions (FIREs). We train a gradient boosted machine learning classifier as in [5] on the latent features obtained by training our recurrent model on the full set of ChIP-seq and DNase-seq assays in the Roadmap compendium and compare it with the Avocado features obtained from [12]. We achieve a good mean Average Precision (mAP) on the four downstream genomic tasks, at par with [5] (not shown). This is expected as the representation used to reconstruct the epigenomic data will be able to serve as a good basis for classification of genomic features, as many of these features manifest themselves in the epigenomic signals.

2.2 LSTM model captures evolutionary activity.

We find that our representation accurately identifies conserved regions in the genome. PhyloP scores (taken from [13]) measure evolutionary conservation at particular genomic sites. Negative scores indicate evolutionary acceleration, which is faster than neutral drift and positive scores indicate evolutionary conservation, which is slower than neutral drift. These p-scores help us understand the nature of selection at chosen genomic positions. The 2D histogram of the PhyloP scores with the feature values in Figure 2 demonstrates that certain chosen features are indicative of evolutionary conservation and acceleration. We also observe that the PhyloP scores show a positive correlation with our feature set (plot not shown).

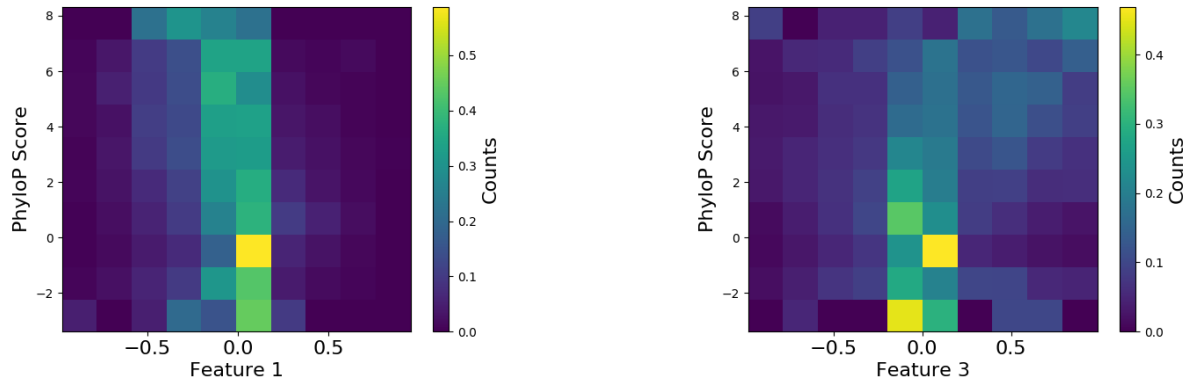

Figure (2) 2D histogram of PhyloP score and feature values. The x-axis bins the values of the chosen feature and the y-axis bins the values of the PhyloP score. The color gradient denotes the strength of points in each bin after columnwise normalization.

2.3 LSTM model provides smoother features

We found that the LSTM model produces latent factors that are smooth across the genome. Owing to the sequential nature of our model, we expect the features obtained to be smoother across the genome when compared to Avocado. To demonstrate this, we plot the Euclidean distance of features between pairs of positions for different distances and average it across the genome as shown in Figure 3 and it can be seen that the LSTM features are smoother than features from Avocado.

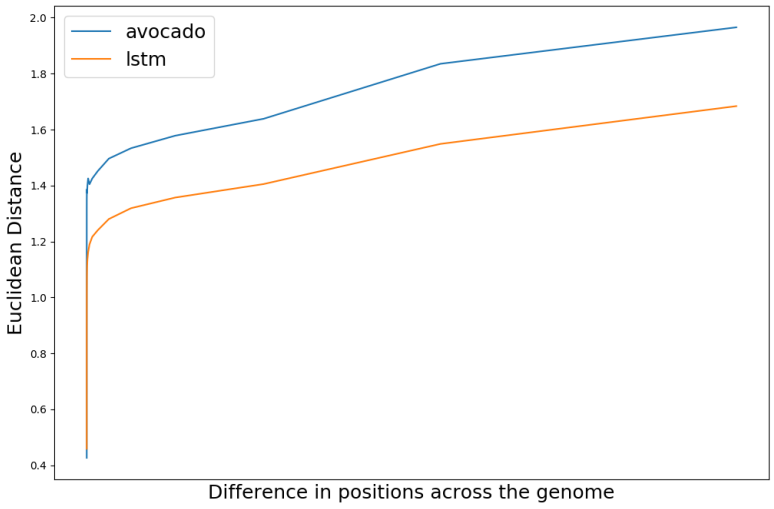

Figure (3) Euclidean distance of the latent representations of Avocado and LSTM. The x-axis represents the difference between the positions in the genome. The y-axis shows the Euclidean distance. The distance was calculated for pairs of positions for varied distances and averaged across the genome.

References

[1] Available Online: http://www.roadmapepigenomics.org/

[2] Available Online: https://www.encodeproject.org/

[3] Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods, 9(5), 473-476.

[4] Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature methods, 9(3), 215.

[5] Schreiber, J., Durham, T., Bilmes, J., & Noble, W. S. (2019). Multi-scale deep tensor factorization learns a latent representation of the human epigenome. BioRxiv, 364976.

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[7] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[8] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[9] Lu, L., Zhang, X., Cho, K., & Renals, S. (2015). A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In Sixteenth Annual Conference of the International Speech Communication Association.

[10] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.

[11] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[12] Available Online: https://noble.gs.washington.edu/proj/avocado/

[13] Available Online: http://compgen.bscb.cornell.edu/phast/