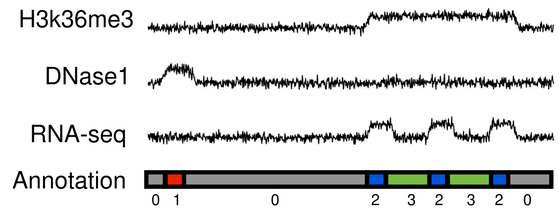
A cell type-agnostic representation of the human epigenome through a deep recurrent neural network model

SFU SCHOOL OF COMPUTING SCIENCE

Kevin D'Souza, Adam Li, Vijay Bhargava, Maxwell W Libbrecht

UBC Electrical and Computer Engineering

Semi-automated genome annotation



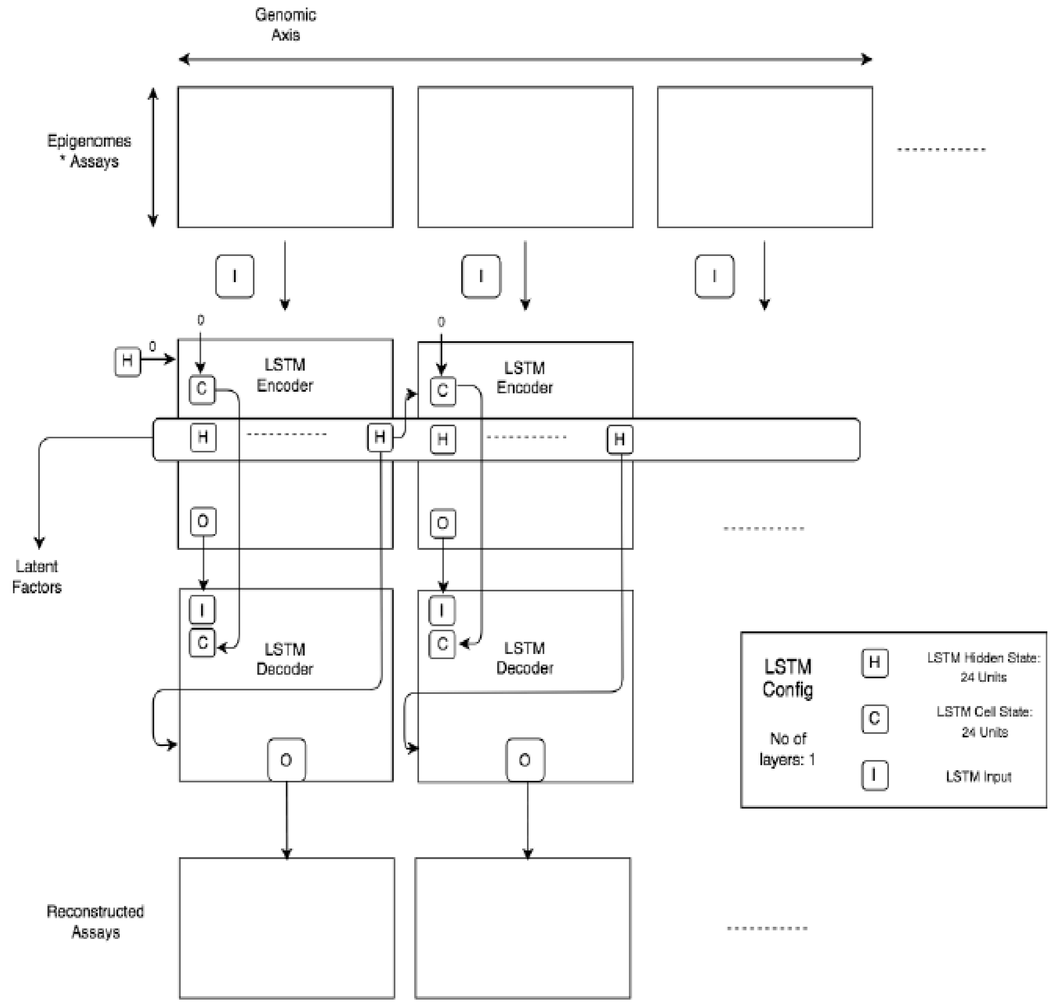
Human interpretation: 1 = “Enhancer”, 2 = “Exon”, …

- **Input**: Real-valued functional genomics data tracks defined over the genome.
- **Output**: Partition of the genome with integer labels assigned to each segment (*celll type specific*)
- **Model**: Hidden Markov model or dynamic Bayesian network.
- **Examples**: ChromHMM, Segway

Existing feed-forward Neural Model (Avocado)

- Model trained on a tensor of cell type, assay and genomic position and corresponding embeddings (factors) are learnt
- These embeddings are concatenated to impute the track value at each position in the tensor using a Neural Network
- Cell Type agnostic continuous annotations are produced by concatenating the resulting factors

Sequential LSTM Model



- Model captures spatial relationships of neighboring genomic positions by the virtue of the LSTM being able to maintain long term dependencies
- Backbone is an Autoencoder framework comprising of an Encoder and a Decoder
- The Encoder and the Decoder are LSTM’s with the given configuration
- The Assays serving as the input to the Encoder are arranged in a matrix format and fed in one frame length at a time
- The hidden states of the LSTM are used as annotations. These are continuous and cell type agnostic.

Ablations

- Different versions of the model are tried with increasing number of hidden nodes and its noted that a single layer with layer norm provides the best tradeoff with respect to the parameters and MAP

Results

Model classifies important genomic phenomena

Model captures evolutionary activity

Model provides smoother features