Evaluating Machine Learning Approaches for Structural Genomics

Jonathan C. Pickett*[ab], Arya Haji Taheri[ab], Ryan R. Cheng[a], Michele Di Pierro[a]
Vinícius G. Contessoto[a], and José Onuchic[a]

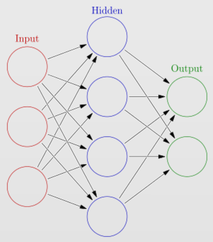
[a]*Center for Theoretical Biological Physics – CTBP – BRC, Rice University, Houston, TX*
[b]*University of Houston - UH - Houston, TX*
**j.pickett.2004@gmail.com*

Abstract:

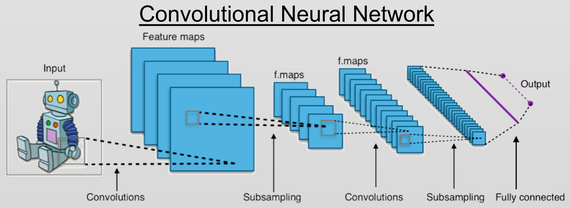
Modern molecular biology produces large amounts of data, which can be difficult to derive any useful information from. We are investigating correlations that exist between genetic annotations of human DNA and chromosome structural features. Chromatin Immuno-Precipitation Sequencing(ChIP-Seq) data tracks, made available through the ENCODE project, characterize the biochemical nature of chromosomal loci. Chromatin can be categorized into types that we call type A and type B which we further classify into chromatin sub-types(A1, A2, B1, B2, and B3). It has been previously shown that these chromatin structural types are strongly related to the overall genome architecture of cells.

Machine learning algorithms have proven to be especially adept at "learning" from correlations in very large data sets. We constructed a number of machine learning models and tested how accurately each performed when identifying chromatin sub-types. Our best approach so far is a recurrent neural network which produced a total error of less than 28% when classifying chromatin sub-types.

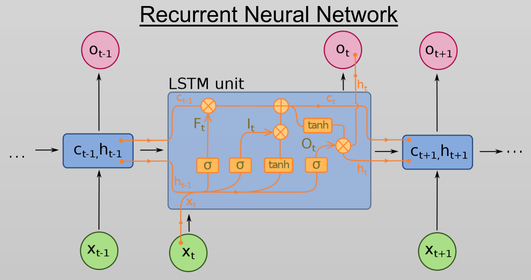
Deep Neural Network



Convolutional Neural Network



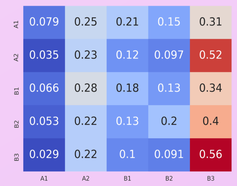
Recurrent Neural Network



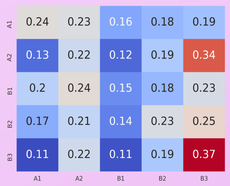
Source for images: wikipedia

Method and Results:

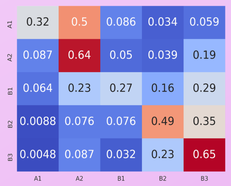
DNN
81.2% error



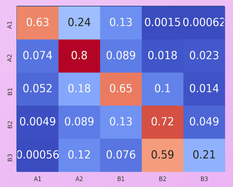
CNN
73.6% error



CNN to RNN
48.9% error



RNN
46.3% error



A1 = 4804, A2 = 5676, B1 = 3944, B2 = 2830, B3 = 8888

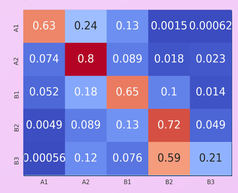
Model Comparison:

There are many types of machine learning model structures. Some of the most common are deep neural networks(DNN), convolutional neural networks(CNN) and recurrent neural networks(RNN). Each of these has a number of variants and limitless combinations that can be used to address unique problems.

Here, DNN, CNN, and RNN, as well as a model that combines aspects of all three, are compared through the use of confusion matrices. All of these models were trained with ChIP-Seq data from the even numbered chromosomes and tested with ChIP-Seq data from the odd numbered chromosomes.

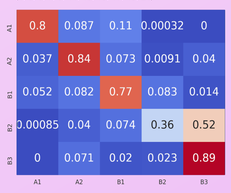
The RNN model outperforms any other model that was tested.

Trained on Even Chromosomes: 46% error
(2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)



A1=4804, A2=5676, B1=3944
B2=2830, B3=8888

Trained on Odd Chromosomes: 27% error
(1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21)



A1=3208, A2=5952, B1=3044
B2=5884, B3=8222

Trained on Random Chromosomes: 20% error
(9, 16, 15, 10, 2, 4, 21, 22, 1, 13, 18)

A1=2980, A2=6992, B1=3322
B2=4670, B3=11206

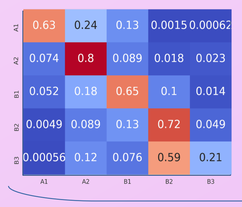
Trained on Random Chromosomes: 33% error
(16, 4, 21, 5, 6, 7, 13, 14, 8, 11, 17)

A1=4256, A2=4862, B1=3288
B2=4190, B3=6384

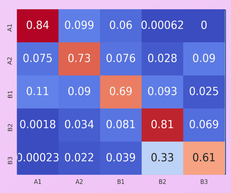
Model Consistency:

The ability to make accurate predictions on data that a model has not been previously exposed to is the ultimate goal of any machine learning model. We used different sets of ChIP-Seq data to train the RNN model, showing that it produces accurate results on testing data regardless of whether it is trained on the same data or different data. A curious observation from these results is the preferential prediction of B2 over B3 or vice-versa. The B3 sub-type is by far the most numerous, so when the model preferentially predicts B2 it is bound to have more errors. The models that preferentially predict B3 have far better accuracy.

RNN trained with all 272 protein data:
46% Error

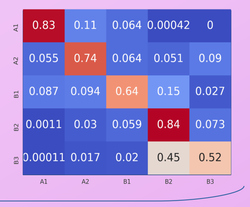


RNN trained with 20 max protein data:
29% Error



RNN trained with 20 max and 20 min (40)protein data:
30% Error

RNN trained with 20 min protein data:
33% Error



A1 = 4804, A2 = 5676, B1 = 3944, B2 = 2830, B3 = 8888

Protein Data Optimization:

In our RNN there are weights matrices which contain values which strongly express the importance of specific protein experiments. By evaluating which of these weights are given the highest values we can determine which protein experimental values are more important to training our model. This evaluation not only helps us to optimize our model's predictive capabilities, it also gives us a clue about the important role that some of these proteins may play in determining chromosome structure in vivo. Reducing the size of the data set also seems to improve the quality of our results, which might imply that training on larger data sets is not yet optimal. Further investigation is needed to determine the actual meaning of these findings.

Bibliography:

Di Pierro, M. et al. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. , 1714980114, (2017).

Acknowledgments:

NSF PHY-1427654
NSF ACI-1531814

Discussion:

The best model we have developed so far is a recurrent neural network made up of two bidirectional LSTM layers, the first having 272 memory units and the second having 350. The output layer is a Dense, time distributed layer with 5 nodes, correlating to the 5 chromatin sub-types, and a softmax activation function. Binary cross entropy is the loss function and Adam is used as the optimization function in this model. By reducing the protein data set, increasing the number of training epochs and decreasing the batch size, we were able to obtain a model that produces only 28% total error and correctly distinguishes between type A and type B with high accuracy.

Using machine learning to analyze large amounts of data in this manner has been shown to be very insightful in the fields of epigenetics and genomics. Applying this approach to similar problems will help us to learn more about chromosome structure and how it is related to gene expression.

Major Category Accuracy: