Recursive Convolutional Neural Networks for Epigenomics

Aikaterini Symeonidi*[1], Nicolaou Anguelos*[2], Frank Johannes[1], Vincent Christlein[2]

1. Technical University of Munich, Population Epigenetics and Epigenomics

2. Friedrich-Alexander University Erlangen-Nürnberg, Pattern Recognition Lab

*ksymeonidh@gmail.com, *anguelos.nicolaou@gmail.com

Technische Universität München

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Abstract

Deep learning methods have proved to be powerful classification tools in the fields of structural and functional genomics. In this poster, we introduce **Recursive** Convolutional Neural Networks (RCNN) for the analysis of epigenomic data. We focus on the task of predicting gene expression from the intensity of histone modifications. The proposed RCNN architecture can be applied to data of an arbitrary size, and has a single meta-parameter that quantifies the models capacity, thus making it flexible for experimenting. The proposed architecture outperforms state-of-the-art systems, while having several orders of magnitude fewer parameters.

The Problem

Chromatine consists of DNA knots coiled around proteins called histones. Histones can be modified through methilation and other processes, thus affecting the coiling and uncoilling of DNA locally. Histone modifications regulate gene expression and can be sequenced and mapped onto the genome as a 1D signal. Given the local intencity of H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3 we want to predict whether a gene is expressed.

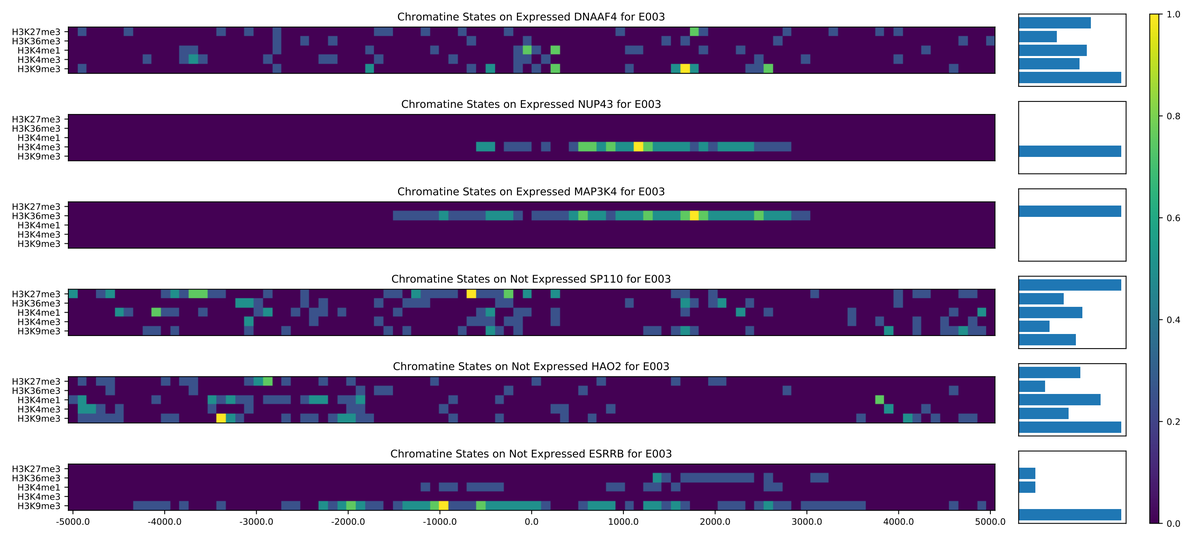

Figure 1:

We formulate the problem as a binary classification (expressed/not-expressed) of 5-channel signals sampled around each gene.

Method

- Ronald Coase:If you torture the data long enough, it will confess to anything.
- Reduce the model parameters by sharing all convolutional weights
- Multi-scale / Scale invariance
- Single capacity meta-parameter

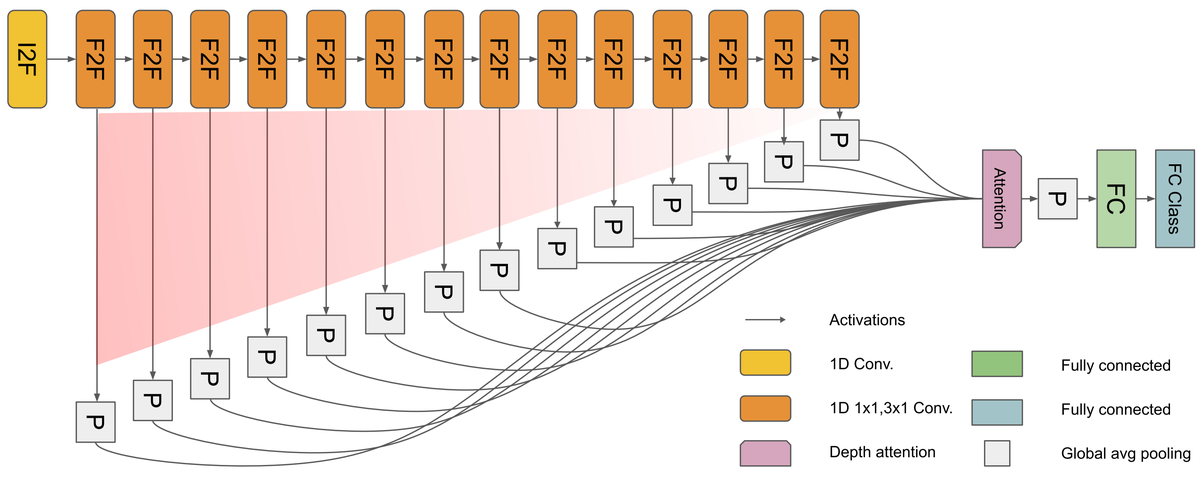

Figure 2: ReChrome architecture

Conclusions

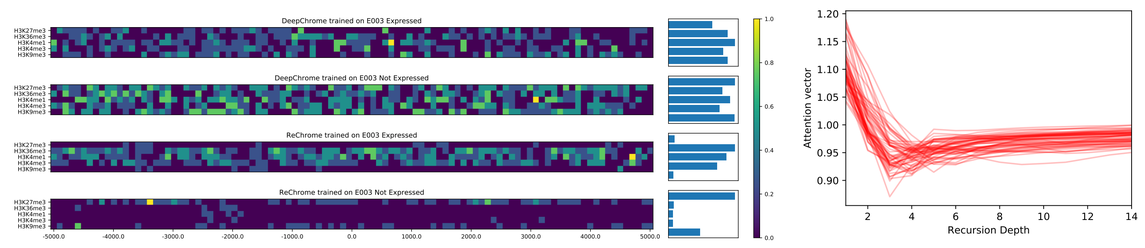

Figure 5: Dreamed samples of ReChrome and DeepChrome, and learned depth coefficients of ReChrome.

- ReChrome is extremely well regularised. Practically immune to over-fitting.
- Reduced Capacity makes the ReChrome more generic.
- Local structure of the signals is not informative.
- Up to this performance ($< 90\%$), the problem is easy. Very small capacity ReChrome performs almost as well.

Experiments

We worked on the epigenome of humans [1] we partitioned human genes into train, validation, and testing so that a distance of 100 Kbases is guarantied. We used 56 cell types to create 56 datasets.

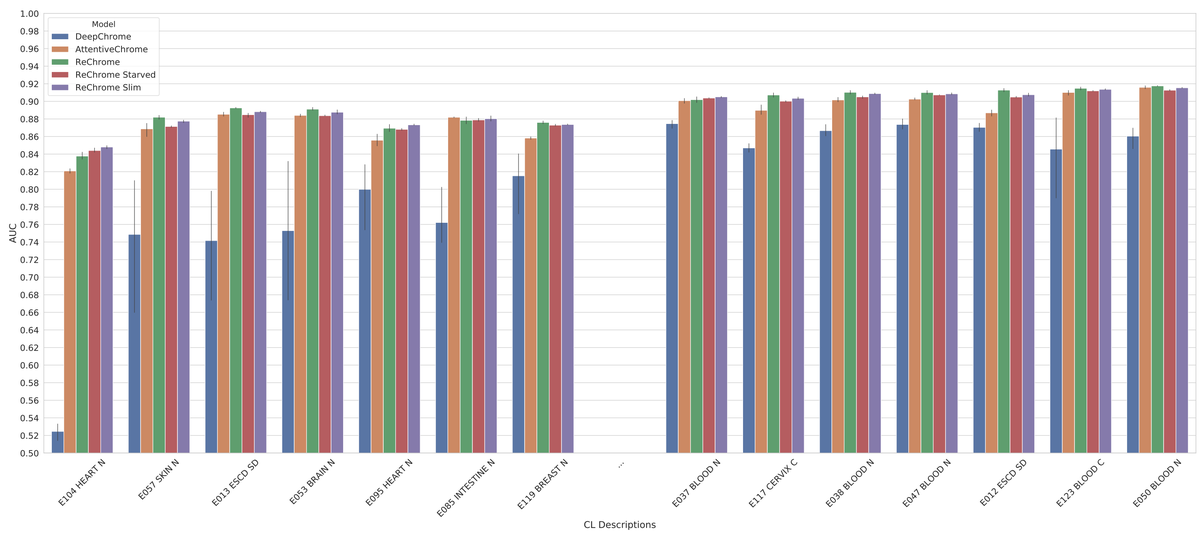

Figure 3: Model performance and consistency across datasets.

All experiments are performed on each dataset and averages across datasets are presented. In table I, we compare to the State-of-the-Art, a CNN classifier [2] and an RNN with attention [3]. We also perform a cross-dataset leave-one-out study: For each dataset, we train a model on the train-set and average the performance on the test-set of all others. All variants of ReChrome outperform both baselines.

TABLE I: Simple and cross-dataset performance

Model	Parameters	Val. set	Test set	Cross test set
DeepChrome	644 177	87.36	81.43	79.78
AttentiveChrome	55 681	86.36	86.80	NA
ReChrome	31 016	**87.54**	87.73	86.13
ReChrome Slim	3076	87.06	**87.75**	**86.56**
ReChrome Starved	416	86.55	86.45	86.45

An ablation study seen in table II, demonstrates the performance of ReChrome across several context sizes. Each sample, a gene's TSS, is sampled at several sampling rates from 1 base up-to 30,000 bases. Sparse sampling is realised by average pooling. The sample length context around the gene TSS also varies between ± 4950 and ± 15000 bases. It is apparent that ReChrome manages to work at various sampling rates and sample sizes with small variation in performance.

TABLE II: Sampling ablation study

Model	bin size	bin count	TSS context (bases)	AUC [%]
ReChrome	1	30 000	$\pm 15,000$	85.35
ReChrome	100	100	± 5000	87.54
ReChrome	150	66	± 4950	87.64
ReChrome	150	200	$\pm 15\,000$	87.63
ReChrome	300	34	$\pm 10\,200$	88.05
ReChrome	300	100	$\pm 15\,000$	88.07
ReChrome	400	26	± 5200	**88.14**
ReChrome	400	76	$\pm 15\,200$	88.09
ReChrome	30 000	1	$\pm 15\,000$	87.35

References:

[1] A. Kundaje and others Integrative analysis of 111 reference human epigenomes. Nature 2015.

[2] R. Singh, J. Lanchantin, G. Robins, Y. Qi. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics 2016.

[3] R. Singh, J. Lanchantin, A. Sekhon, Y. Qi. Attend and predict: Understanding gene regulation by selective attention on chromatin. NIPS 2017