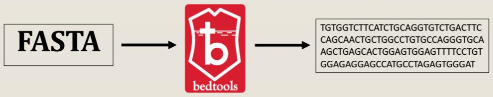
Don’t Drop the Base Pairs: Predicting Genetic Patterns associated with Leukemia

Brandon Benson, Alex McKeehan I Stanford University [mckeehan, bensonb]@stanford.edu

Model

Dataset includes full base pair sequence of one individual leukemia patient.

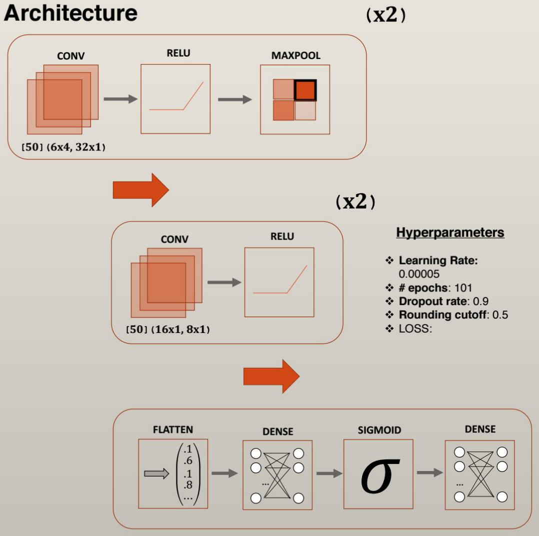
- **23** chromosomes sequenced at the base pair level.
- **Chromatin accessibility** of 18 cell types:



FASTA

TGTGGTCTTCATCTGCAGGTGTCTGACTTC
CAGCAACTGCTGGCCTGTGCCAGGGTGCA
AGCTGAGCACTGGAGTGGAGTTTTCCTGT
GGAGAGGAGCCATGCCTAGAGTGGGAT

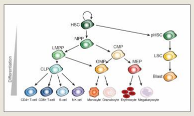
Architecture



Hyperparameters

- **Learning Rate:** 0.00005
- **# epochs**: 101
- **Dropout rate**: 0.9
- **Rounding cutoff**: 0.5
- LOSS:

Problem

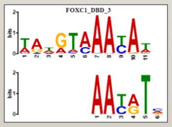


Leukemia results from imbalance in regulation of the typical protein pathways.
The 18 related cell types are shown above:

Gene therapy is possible if genetic sequences could be associated with certain diseases and traits (Hindorff et al). Specifically, we want to map genetic sequences to a chromatin accessibility binary, which can be used to identify anomalies in the protein pathways of these 18 related cell types.

Solution:

We achieve our test-metric goal of (0.70 auPRC) and accuracy of (0.90) using a CNN model. Now that we have predicted chromatin accessibility binaries, we can interrogate the model and interpret results.
We identify genetic motifs that strongly activate our first convolutional layer filters using only 6 first layer filters:



Increasing performance by using all 50 first layer filters, we find that we cannot accurately distinguish regulatory genes beyond the control group.

Confusion Matrix:

	Predict: 1	Predict: 0
Label: 1	0.46	0.54
Label: 0	0.4	0.6

TRUE POSITIVE
PPARD
MYB
ZBTB4
OTX1
DDIT3

FALSE POSITIVE
REST
CREB3L2
MAF
ZNF713
ZBED1

TRUE NEGATIVE
FEZF1
STAT4
BBX
POU3F4
DMRT3

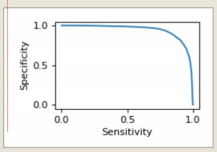
FALSE POSITIVE
ELK1
TAL1
SMAD1
SMAD4
YY1

Performance

Optimized **sigmoid cross entropy** loss function with a Area under precision-recall curve as a test metric.
Average confusion-matrices over **18** cell types:

	Predict: 0	Predict: 1
Label: 0	0.805	0.023
Label: 1	0.1	0.072

auPRC: 0.75
Imbalanced data makes this sensitivity and specificity curve a good description of the model:



Next Steps:

- ❖ Comparison of patterns among 23 chromosomes.
  - ➢ Only tested on a single chromosome.
- ❖ Training with GPU acceleration on full dataset.
  - ➢ Pending approval from Amazon.
- ❖ Clinical validation of genetic pathways in mice.
- ❖ Optimization of hyperparameters via telescope search.

References

[1] Shobhit Gupta, JA Stamatoyannopolous, Timothy Bailey and William Stafford Noble, "Quantifying similarity between motifs", Genome Biology, 8(2):R24, 2007.
[2] Kelly et al. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2015.
[3] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009.Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106: 9362–9367.
[4] Peyton Greenspan; Stanford Biology Department.
[5] CS 230; Coursera. Stanford University; Winter 2018. Andrew Ng.