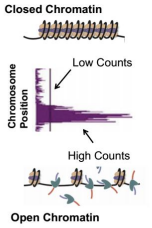
Denoising ATAC-seq with Convolutional Neural Networks

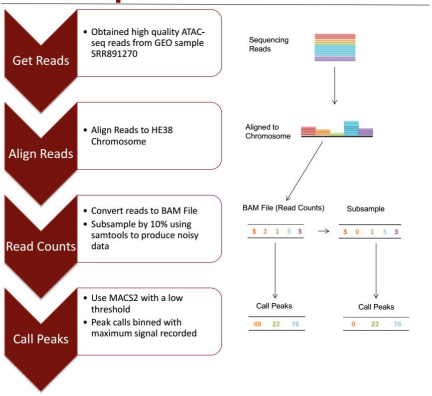
Nic Fishman and Sarah Gurev
{njwfish, sgurev}@stanford.edu

The Problem with ATAC-seq

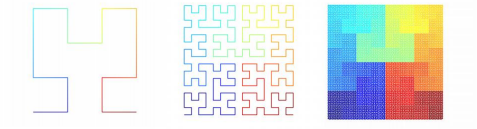
- Chromatin is a compressed version of DNA. A gene cannot be expressed when it is compressed or closed.
- ATAC-seq is a sequencing technique used to determine what parts of chromatin are open or closed.
- The quality of ATAC-seq data is heavily dependent on many environmental factors, and therefore prone to noisiness



Data Pipeline



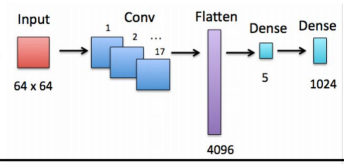
Hilbert Curve



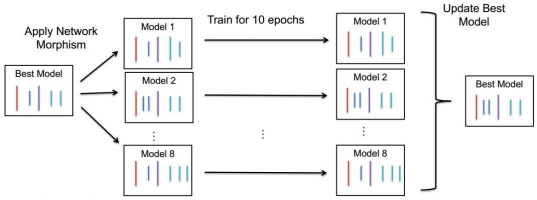
Model Design and Selection

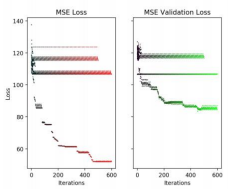
Architecture Search: Random Search

Initial Model:
- Random architecture search
- 1 Conv2D layer into 1 Dense layer
- Randomize:
  - Kernel size
  - Filters
  - Hidden units



Architecture Search: NASH

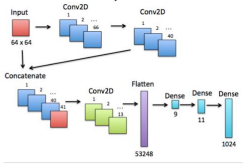




Iterations of Search	Train Error (9753 samples)	Test Error (972 samples)
0	150.156	128.45
20	137.878	122.953
40	92.135	90.574
60	89.778	88.805

Final Model:
Greedy Hill Search to find a new best model after randomly changing the parent

1. Apply Network Morphisms to develop new model:
2. Train child networks with cosine annealing
3. Move to most promising child of network

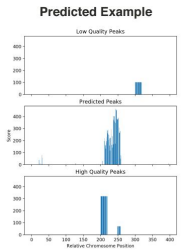


Results

We used an architecture search to develop a CNN architecture that is trained on high quality, low quality pairs of data.

Our model can reliably convert noisy data into high quality data.

An example is to the right, where the subsampled peak is shifted right, and is low score. The model shifts it left, and boosts the peak score.



Discussion

- Our model is able to map low quality to high quality ATAC-seq data using a CNN developed through an Architecture Search
- The results achieved are good, which is expected given the role of the Architecture Search to build a network with the best possible performance
- Our computationally simulated noise, done via subsampling, is a reasonable estimate of ATAC-seq noise caused by environmental factors like amount of input DNA and sequencing depth, so our model should do a good job of denoising true low-quality data.

Future Work

- Develop pipeline for cell-line specific training
- Confirm biological relevance of denoised data using properties of chromatin accessibility for evaluation
- Replicate work with other sources of noise

References

Elsken, T., Metzen, J. H., & Hutter, F. (2017). Simple and efficient architecture search for convolutional neural networks. arXiv preprint arXiv:1711.04528.

Koh, P. W., Pierson, E., & Kundaje, A. (2017). Denoising genome-wide histone ChIP-seq with convolutional neural networks. Bioinformatics, 33(14), I225-I233. doi:10.1093/bioinformatics/btx243

Loshchilov, I., & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. Proceedings of the International Conference on Learning Representations (ICLR).

Ou, J., Liu, H., Yu, J., Kelliher, M. A., Castilla, L. H., Lawson, N. D., & Zhu, L. J. (2018). ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC genomics, 19(1), 169.