Research Statement

Rick Farouni

Background

The mechanism governing gene regulation in eukaryotic organisms, such as humans, is a complex unsolved problem of great importance. The difficulty of the problem partly stems from our inability to read the genome. Although we can currently sequence an entire human genome in less than a day, we are still unable to ascribe semantic meaning to all of the various regions that make up the sequence itself. Admittedly, we have made great progress in annotating the 1.5% of the genome that contains the approximately 20,000 protein coding genes we presently know (GENCODE annotation, V25). Nonetheless, the rest of the non-coding genome is still far from fully understood.

In humans, gene expression is mainly controlled by the binding of roughly 2,000 transcription factor (TF) proteins to millions of regions on DNA known as transcription factor binding sites (TFBS). A single TF can regulate multiple genes by binding to multiple TFBS and any particular gene can be regulated my several TFs. A region of noncoding DNA where transcription factor binding sites cluster together is called a *cis-regulartory element*. Cis-regulatory elements can be sub-classified into promoters, enhancers, silencers, locus control regions, or insulators depending on their relative distances from the transciption start site (TSS) and the three-dimensional structure of chromatin, which is a DNA-protein complex consisting of DNA wraping around histone proteins forming nucleosomes packaged into chromosomes. When the genome is packed into closed chromatin (its default state), no transcriptional activity occurs since TFBSs are not accessible. For chromatin to open, repressive chromatin marks such as methylated histones and methylated cytosines (in CpG islands sequences) need to be modified by enzymes and replaced by active chromatin marks. These post-translational modifications are epigenetic since they can regulate gene expression without modifying the genetic code.

One class of non-coding DNA sequence that is of particular interest consists of cis-regulatory elements such as *enhancers* that exert a functional role in the regulation of gene expression. Mapping the regulatory elements specific to each cell type and condition is a first step to understanding how these regulatory regions are responsible for determining cell identity, orchestrating cellular differentiation, and how their dysregulation causes abnormal gene expression leading to diseases such as cancer.

Problem Statement

As of yet, we do not have a complete and reliable list of identified regulatory elements, especially for enhancers and their target promoters. Although there are experimental approaches such as STARR-seq, Hi-C, and ChIA-PET for identifying enhancers and enhancer-promoter associations, these assays are not context specific, their signal do not necessary imply functional relationships, and they are subject to high bias and noise. As a result of these limitations and

the dramatic increase in publicly available 'omics' datasets, many researchers (Yip *et al.*, 2013; Leung *et al.*, 2016) have called for the adoption of a different strategy for identifying genomic elements, one that leverages data-driven machine learning approaches to identify patterns in large datasets and to generate computational predictions which can then be experimentally validated in the lab.

Research Plan

My ongoing dissertation investigates the application of *deep generative latent models* to high-dimensional datasets. Deep generative models (Salakhutdinov, 2009; Kingma and Welling, 2013) are a class of unsupervised machine learning methods that bring together the strengths of two streams of machine learning: deep learning and Bayesian probabilistic modeling. Whereas deep learning allows us to build scalable models that can capture the most complex nonlinear dependencies in the data, Bayesian generative modeling provides a probabilistic framework in which we can model, not just the variables we observe, but also latent variables that represent hidden patterns which can explain the observed data. By combining a generative model with a deep neural network, we obtain a deep generative model that can capture a *hierarchy* of distributed dependencies between latent variables, learn an efficient lower dimensional representation of the data, and enable us to generate novel samples from the model's learned latent representational space, samples such as images of objects never seen before in training data. This latent representation equips generative models with a causal interpretive power that is absent from many black-box machine learning algorithms which are optimized for empirical prediction. Although predictive models have proven to be valuable in many scientific applications, it can be argued that for the purposes of basic scientific research, their value is contingent upon the supportive role they play in building causal explanatory models of natural phenomena. I would also argue that the generative approach to modeling keeps true to the principle of *verum factum*, namely, to understand an entity properly, one should be able to create it, or in the words of Richard Feynman "What I cannot create, I do not understand".

I realized the potential of applying deep generative models to problems in computational biology after I completed a summer internship at the lab of Professor Ewy Mathè in the Department of Biomedical Informatics. As part of internship program, I worked on developing an R package for the workflow analysis of data generated from genome-wide chromatin accessibility assays such as ATAC-seq and DNase-seq. Although the software implementation was more concerned with developing a downstream analysis workflow for mapping and comparing chromatin accessibility between different cell types (e.g. cancer vs normal cells), I also explored the characteristics of raw ATAC-seq data. I was intrigued to know that the paired-end reads generated by ATAC-seq (Buenrostro *et al.*, 2013) not only provide footprints of specific transcription factor occupancy, but also fingerprints of nucleosome packing and positioning with which we can infer the functional states of chromatin. Knowing that the most successful computational methods for learning chromatin states happen to be unsupervised generative models (e.g. *ChromHMM* (Ernst and Kellis, 2012) and *Segway* (Hoffman *et al.*, 2012)), I became interested in investigating whether unsupervised deep generative models can also be used to accomplish a similar feat by using the distribution of fragment sizes generated by ATAC-seq. This interest fits well with a more ambitious research question that I would like to pursue. Namely, how to determine the identity and interactions of cell-specific *active* regulatory regions

from high-throughput molecular features such as chromatin accessibility, histone modifications, and transcription factor binding sites using statistical and machine learning models trained on large datasets.

As a postdoctoral fellow, I would like to apply a principled statistical and computational approach to identifying enhancers and enhancer-promoter interactions that are specific to the cancer cell-type. Given that context-specific signals determine oncogenic enhancer activation and by extension their dysregulatory effect on gene expression, I think that a modeling framework that just relies on static feature such as DNA sequence won't be sufficient if we do not incorporate a range of other cell measurements, ranging from *sequence features* such as transcription factor binding sites to *epigenomic features* such as nuclesome positioning, histone modifications, and chromatin accessibility.

I see a great opportunity in applying machine learning models to data generated by high-throughput assays such as ATAC-seq (Buenrostro *et al.*, 2013) and CHIP-exo (Rhee and Pugh, 2008) with the goal of discovering enhancers and other regulatory elements that play a major role in cancer. Enhancers, by virtue of functioning as the signaling switchboard that brings about changes in transcriptional gene regulation in response to external cues, must be key in understanding how the healthy genotype can give rise to the cancer phenotype. To this goal I hope to devote my future scientific career.

References

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **10**(12), 1213–8.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. a., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, **9**(5), 473–6.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. *ICLR*, (Ml), 1–14.

Leung, M. K. K., Delong, A., Alipanahi, B., and Frey, B. J. (2016). Machine learning in genomic medicine: A review of computational problems and data sets. *Proceedings of the IEEE*, **104**(1), 176–197.

Rhee, H. S. and Pugh, B. F. (2008). ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. *Current Protocols in Molecular Biology*, **141**(4), 520–529.

Salakhutdinov, R. (2009). Learning Deep Generative Models. *Mit.Edu*, **2**, 1–84.

Yip, K. Y., Cheng, C., and Gerstein, M. (2013). Machine learning and genome annotation: a match meant to be? *Genome biology*, **14**(5), 205.