INFERRING GENOMIC DATA

Advanced Topics in Data Mining and Machine Learning

Dylan Marshall

April 18th, 2018

Abstract

Deep convolutional neural networks (CNN) are best known for their remarkable accuracy in classifying images according to their respective labels. Biologists, noting the utility of these recently developed methods, collectively wonder how their data may be amenable [1, 2]. Indeed, the nascent intersection of deep learning and genomics has already resulted in biological insights - such as the application of a deep CNN to DNA sequences which revealed disease associated nucleotides [3]. DNA, however, is only but one of many genomic data types. Other, orthogonal, genomic data types come from experiments which interrogate the molecular processes taking place on DNA, such as the location of certain proteins. Integrating these different datasets has proven difficult for the field and tools capable of doing so have yet to be developed.

In this talk I present FIDDLE, a multi-modal, deep CNN framework that learns the unified representation of multiple disparate genomic data types to infer another genomic data type [4]. As a case study, I demonstrate FIDDLE's ability to predict transcription-start-site sequencing (TSS-seq) genomic data in the model organism Saccharomyces cerevisiae (brewer's yeast). Through the modulation of input combinations, I show how FIDDLE can be used to draw conclusions about the non-linear relationships between genomic data types.

References

[1] review of deep learning in biology / medicine:

(2018) Ching, T., Himmelstein, D. S., … Gitter, A., Green, C. S. Opportunities And Obstacles For Deep Learning in Biology And Medicine. *bioRxiv 142760*; doi: https://doi.org/10.1101/142760

[2] *Nature* news article on promises and pitfalls of deep learning in biology:

(2018) Webb. S. Deep Learning for biology. *Nature* 554(7693):555-557; doi: https://doi.org/10.1038/d41586-018-02174-z

[3] Convolutional model uncovers disease-associated genome variants

(2015) Zhou, J., Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 931-4; doi: https://doi.org/101038/nmeth.3547

[4] preprint of project:

(2016) Eser, U., Churchman, L.S., FIDDLE: An integrative deep learning framework for functional genomic data inference. *bioRxiv 081380*; doi: https://doi.org/10.1101/081380

[5] video describing project:

(2016) Broad Institute of MIT & Harvard *Models, Inference & Algorithms* seminar. "FIDDLE: An integrative deep learning framework for functional genomic data inference." https://www.youtube.com/watch?v=pcLTUsOm5pc