Predict the Conditional Probability Distributions from Noisy Measurements with Neural Networks

Shangying Wang[1, 2] and Simone Bianco[1, 2, &]

Short Abstract — **Biological systems are intrinsically noisy. This imposes a big challenge in building traditional supervised machine learning models that can only predict determined phenotypic parameters or categories per specific genetic and/or environmental conditions as inputs. Furthermore, biological noise has been proven to play a crucial role in gene regulation mechanisms. The prediction of the average value of a given phenotype is not always sufficient to fully characterize a given biological system. In this work, we developed a deep learning algorithm that can predict the conditional probability distribution of a phenotype of interest with a small number of observations per input condition. The deep neural network can automatically generate the probability distributions based on as few as one noisy measurement for each input condition, with no prior knowledge or assumption of the probability distributions. This is extremely useful for exploring unknown biological systems with limited measurements for each input condition, which is linked not only to a better quantitative understanding of biological systems, but also to the design of new ones, as it is in the case of synthetic biology and cellular engineering.**

Keywords — **conditional probability distribution, noisy measurements, neural network, deep learning, biological systems.**

I. Introduction

Phenotypic variation is ubiquitous in biology and is often traceable to underlying genetic and environmental variation. However, identical genotypic and environmental conditions are not sufficient to guarantee a unique phenotype. This is mainly due to the inherent stochastic nature of biological processes. Random fluctuations may alter the levels of the biochemical components and drive the system to different phenotypes. Noise is a key component of a well-functioning biological system and plays an essential role in key cellular activities [1- 5]. However, the study of the nature and effect of noise on biological systems is still poorly studied and hence only partially understood. This implies that the use of machine learning tools in biology is somehow limited by biological fluctuations apparent in the data. Particularly, building machine learning tools to elucidate the relationship between underlying genetic and environmental conditions to phenotypic observations is a challenge. One critical deficiency of this type of machine learning predictors is the general inability to call a given observation as an outlier relative to the data that the model has been trained and tested on [6]. More specifically, there is a distribution of outputs (all possible phenotypic values) corresponding to each unique set of inputs (genetic and environmental condition).

A naive deep learning predictor will make a prediction based on the mean of all available observations for each unique set of inputs. This is problematic in the case of stochastic processes in the presence of limited observations, where the mean value cannot represent the entire dynamics of the system. Based on the central limit theorem in statistics, the sampling distribution of the mean for a variable will approximate a normal distribution given a sufficiently large sample size. The sampling distributions of the mean cluster more tightly around the population mean as the sample sizes increase. Conversely, the sampling distributions of the mean for smaller sample sizes are much broader. For small sample sizes, it's not unusual for sample means to be further away from the actual population mean. Furthermore, even with a very large sample size and a reliable machine learning predictor are trained, the sample mean cannot provide insightful information for specific inputs since not only the average, but the whole probability distribution is important in understanding the whole characteristic of the population.

The aim of this work is to introduce a machine learning method capable of inferring the conditional probability distribution *CPD(y|x)* of the output variable *y*, conditional on inputs *x* without knowing the prior knowledge or assumption of the distribution itself.

II. Conclusion

Our approach Overcomes the bottleneck of the need to collect many observations per input condition to eliminate the effect of the noise of phenotype observations or reconstruct the conditional probability distribution of a stochastic process, which is often expensive and time-consuming. It also reduces the barrier in implementing predictive machine learning model for "noisy" biological data.

References

[1] Raser, J. M. & O'shea, E. K. Noise in gene expression: origins, consequences, and control. Science 309, 2010{2013 (2005).

[2] Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. Nature 467, 167{173 (2010).

[3] Munsky, B., Neuert, G. & Van Oudenaarden, A. Using gene expression noise to understand gene regulation. Science 336, 183{187 (2012).

[4] Tsimring, L. S. Noise in biology. Reports on Progress in Physics 77, 026601 (2014).

[5] Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mrna abundance. Cell 165, 535{550 (2016).

[6] Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface 15, 20170387 (2018).

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant No. DBI-1548297.

[1]Department of Functional Genomics and Cellular Engineering, IBM Almaden Research Center, 650 Harry Rd, San Jose, CA 95120

[2]Center for Cellular Construction, San Francisco, CA, 94158

[&]Current address: Altos Labs, Redwood city, CA