Part I: Extracting the Hidden Principles of Biology with Machine Learning

March 23, 2022

By **Andrew Beam, Ph.D.**
Assistant Professor, Harvard University, and Founding Head of Machine Learning, Generate Biomedicines

Molly Gibson, Ph.D.
Co-Founder and Chief Strategy and Innovation Officer, Generate Biomedicines

Gevorg Grigoryan, Ph.D.
Co-founder and Chief Technology Officer, Generate Biomedicines

generatebiomedicines.com

History shows us that technological advances occur when a scientific field graduates from basic study of a natural system to knowing the core principles that govern that system's behavior. Take the transistor, for example: its invention was not possible until the field of physics had moved from understanding the electron itself to identifying the principles of electron mobility that enabled the electron to be controlled in a precise and predictable way. Knowledge of such principles is essential before biotech and pharmaceutical companies can intentionally engineer any natural system to create novel technologies.

The intentional engineering of biology has largely eluded us. Biological systems are governed by highly complex dynamics shaped by interactions between an astronomical number of components. This complexity has, so far, stymied attempts to identify a set of principles that would give us precise control over the cell or let us predictably and reproducibly design a protein for a specific purpose. As a result, almost all modern medicines have been derived from naturally occurring molecules and cells using slow trial-and-error processes, instead of being purpose-built from principles.

It may well be that evolution has not equipped our brains for the complex, probabilistic reasoning needed to find the principles in biology. But thanks to the digital age ushered in by the transistor, a new kind of intelligence is emerging that excels at this type of reasoning – machine learning, a type of artificial intelligence (AI) that uses computer algorithms to identify patterns in data with little or no human intervention. Machine learning is enabling a new paradigm for approaching science generally and biology specifically. Rather than positing the principles of a complex system and testing whether they are predictive (the "bottom-up" paradigm science has traditionally used), AI enables the inference of principles from a given set of observations (data). This is the goal of the emerging field known as generative biology.

Welcome to the future

Machine learning has revolutionized fields including speech recognition, medical imaging, computer vision, natural language processing, and computational biology over the past decade. The advent of large data sets coupled with advances in computing have made it practical to train large and performant neural networks. These neural networks, which are said to employ deep learning because of their many-layered nodes, are accomplishing tasks that were until recently thought to be too difficult for computers to tackle—for example, winning the ancient Chinese game of Go, a notoriously complicated game, and even surpassing the strategic thinking of Go champions.

These advances in machin`e learning are beginning to change the future of biology as well. For example, machine learning systems have now surpassed the ability of any traditional method to predict the three-dimensional structure of a protein, based solely on its amino acid sequence. This parallels the Go story: when a problem is correctly framed for machine learning and the right mix of data and computing power is available, the limitations imposed by traditional bottom-up approaches can be quickly overturned.

Besides structure prediction, machine learning is on the brink of discovering the principles that link a protein's amino acid sequence to its function. This capability results from an explosive increase over the last decade in the available number of protein sequences from which the required principles can be extracted. Once biologists have fully cracked that code for the protein universe, the application of machine learning to produce proteins with completely novel, purpose-built functions – proteins never before seen in nature – is inevitable. This will be a milestone for the field of generative biology and a transformative moment in the history of biology itself. It will also revolutionize the field of medicine by introducing the power to generate therapies across the entire spectrum of human diseases that were previously imaginable only in the realm of science fiction.

In our next article, we'll take a look at the inherent flaws in our current trial-and-error approach for drug discovery and why generative biology, aided by the power of machine learning, offers a new paradigm for transforming the future of human medicine.