Deep learning for genomics

Application of deep learning to genomic datasets is an exciting area that is rapidly developing and is primed to revolutionize genome analysis. We embrace the potential that deep learning holds for understanding genome biology, and we encourage further advances in this area, extending to all aspects of genomics research.

The human genome comprises more than 3 billion base pairs. Recent technological advances have increased the mechanistic understanding of genome biology to an incredible degree. However, the complexity and sheer amount of information contained in DNA and chromatin remain roadblocks to complete understanding of all functions and interactions of the genome. Connecting genotype to phenotype, predicting regulatory function, and classifying mutation types are all areas in which harnessing the vast genomic information from a large number of individuals can lead to new insights. However, working in this large data space is challenging when conventional methods are used. Therefore, new and innovative approaches are needed in genome science to enrich understanding of basic biology and connections to disease.

One exciting and promising approach now being applied in the genomics field is deep learning, a variation of machine learning that uses neural networks to automatically extract novel features from input data. Deep learning has been successfully implemented in areas such as image recognition or robotics (e.g., self-driving cars) and is most useful when large amounts of data are available. In this respect, using deep learning as a tool in the field of genomics is entirely apt. Although it is still in somewhat early stages, deep learning in genomics has the potential to inform fields such as cancer diagnosis and treatment, clinical genetics, crop improvement, epidemiology and public health, population genetics, evolutionary or phylogenetic analyses, and functional genomics.

In this issue, Zou et al. provide a primer on deep learning for genomics (https://doi.org/10.1038/s41588-018-0295-5) that is intended for a broad audience of biologists, bioinformaticians, and computer scientists. The authors include practical guidelines on how to perform deep learning on genomic datasets, and they have compiled a convenient list of resources and tools for researchers. It is our hope that this Perspective will aid the community in adopting deep learning techniques in their genomic analyses when appropriate. The authors have even generated an interactive tutorial demonstrating how to build a convolutional neural network for discovery of DNA-binding motifs. Because this is a relatively new and rapidly developing field, we recognize that this list is not exhaustive, but we consider it to be a good starting point for those who wish to learn more about applying deep learning methods to their datasets.

Functional genomic analysis is the field in which deep learning has made the most inroads to date. The availability of vast troves of data of various types (DNA, RNA, methylation, chromatin accessibility, histone modifications, chromosome interactions, and so forth) ensures that there are enough training datasets to build accurate prediction models relating to gene expression, genomic regulation, or variant interpretation. Other features such as identification of long noncoding RNAs or splice-site prediction can also be analyzed. As more data become available, better models will be able to be trained, thus resulting in even more precise and accurate predictions of genomic features and functions.

Although deep learning holds enormous promise for advancing new discoveries in genomics, it also should be implemented mindfully and with appropriate caution. Deep learning should be applied to biological datasets of sufficient size, usually on the order of thousands of samples. The 'black box' nature of deep neural networks is an intrinsic property and does not necessarily lend itself well to complete understanding or transparency. Subtle variations in the input data can have outsized effects and must be controlled for as well as possible. Importantly, deep learning methods should be compared with simpler machine learning models with fewer parameters to ensure that the additional model complexity afforded by deep learning has not led to overfitting of the data. Depending on the type and size of the datasets being analyzed and the questions being asked, deep learning can either offer benefits or introduce more uncertainty.

Even with these caveats, there is great potential for deep learning methods to make substantial contributions to the understanding of gene regulation, genome organization, and mutation effects. Beyond being applied to functional genomics, deep learning can also be applied to larger questions relating to health and disease or other areas in which genomic information is used, such as plant or population genomics. We are eager to embrace deep learning methods as an established tool for genomic analysis, and we look forward with great anticipation to the new insights that will emerge from these applications. ❐

Published online: 21 December 2018
https://doi.org/10.1038/s41588-018-0328-0