BIOINFORMATICS

Variants from the deep

Deep learning pushes the limits of accuracy in genomic variant calling.

Spotting the differences between a sequenced genome and a reference genome, amid sequencing errors, is a cardinal task in computational biology. Tools for calling genomic variants have evolved to be highly sophisticated, but "even state-of-the-art methods are producing thousands of errors per genome," says Mark DePristo. "There's an outstanding challenge, despite how well established the field is, to do better."

While at the Broad Institute, DePristo led the team that developed the Genome Analysis Toolkit (GATK), which has become a go-to sequence analysis engine. A few years ago, he sensed that customary approaches had almost exhausted their potential to improve. "We've thrown the traditional statistical modeling kitchen sink at this problem," he says. In 2015, DePristo moved to Google, Inc., and began work on DeepVariant, a variant caller based on deep learning.

Deep learning works by transforming data as they pass through processing units or 'neurons' that are arranged in connected layers, in order to accomplish a specific task. These transformations are encoded in the neural network during training—the process of optimizing the learning task by using data with a known ground truth. In effect, a network trained for variant classification learns which features of the data distinguish true variants from errors.

DeepVariant uses the open-source Inception TensorFlow framework, which was designed by Google for image classification. First, sequencing reads are aligned to a reference, and a sensitive caller finds all potential variants. At each variant site, the 'pileup' of aligned reads is then presented to the neural network as an image, on which six superimposed channels encode additional information such as base quality. The classifier outputs reference and variant allele genotype likelihoods.

A recent benchmarking study by cloud-based genome informatics company DNAnexus confirmed that DeepVariant substantially outperforms existing tools in calling single base variants and short insertions and deletions. When an earlier version of the software was found to lag on exome and PCR-amplified sequencing data, the researchers added these data types to the training suite, which greatly boosted performance. Whereas a traditional approach would need its model to be adapted or its parameters tweaked, "most of the problems that people encounter in variant calling become a data problem for DeepVariant," says DePristo. Retraining also worked well for long-read and other sequencing data sources, making DeepVariant something of a universal tool.

Application of deep learning to genomics has been challenging. The team had to figure out how to encode genomic data in TensorFlow, how to 'feed' the algorithm at massive scales, and how to engineer consistent results across the very heterogeneous hardware that is available—standard computer chips (CPUs), graphics processing units (GPUs) and tensor processing units (TPUs). They have been keen to share their findings through open-source code and tutorials, to encourage others to try deep learning.

DeepVariant's performance does come at a computational price, but the software is continually updated to take advantage of the rapidly evolving efficiencies in deep learning software and hardware, and an efficient implementation is being offered in the Google Cloud Platform.

In the future, DeepVariant may be expanded to call structural variants, a complex and diverse category of variation. DePristo notes that DeepVariant is at the start of its technical life span, and his team is betting on the trajectory of deep learning hardware and software improvements. "We're trying to ride that wave, and being inside of Google is a great place to see where that wave is going," he says.

Tal Nawy

Published online: 30 October 2018
https://doi.org/10.1038/s41592-018-0209-6

Research papers
Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4235 (2018).