GENETICS

Predicting the impact of genomic variation

An unsupervised approach with high predictive power provides a single measure of functional importance for each variant in the human genome.

Whole-genome sequencing has identified a great number of genetic variants, but predicting whether these variants will have benign or deleterious effects is currently one of the biggest challenges in genetics. Computational models are designed to help make this important distinction, but often they are based on one type of criterion, such as the degree of evolutionary conservation, or are limited to coding regions of the genome.

Recent *in silico* prediction tools, such as CADD (combined annotation-dependent depletion) and GWAVA (genome-wide annotation of variants), have enabled significant progress by integrating diverse annotations into one score that indicates the functional importance of any variant in the genome. These machine learning methods are supervised, meaning that they learn how to classify new variants by training on a set of variants that have already been labeled as benign or deleterious. Variant labels in the CADD training data, for example, are based on the extent of their evolutionary conservation as a proxy for their functionality. The predictive power of supervised approaches thus depends on the quality of the labeled data used in the training stage, which represents their major limitation, as some of the training data may be mislabeled.

Columbia University biostatistician Iuliana Ionita-Laza and her colleagues sought to improve on integrative approaches by reducing the bias introduced by labeling training data. Their Eigen algorithm is an unsupervised spectral approach that does not rely on labeled training data. As Ionita-Laza explains, this means that Eigen does not make any *a priori* assumptions about whether any type of annotation is more or less relevant to disease. Instead, the approach assigns weights for each functional annotation on the basis

of its predictive accuracy, which is derived from the correlation structure between the different annotations. The functional score for each variant is computed as an optimal weighted linear combination of individual annotations. This way, annotations with higher predictive accuracy are up-weighted, whereas those with lower accuracy are down-weighted.

As different types of annotations are relevant to coding and noncoding variants, the Eigen score is defined separately for these two categories.

The scientists tested Eigen in a wide range of scenarios and compared its performance to that of algorithms such as CADD and GWAVA. The comparative data showed that for coding variants, Eigen outperformed the CADD score for mutations in some Mendelian disease genes, such as those for Kabuki syndrome, cystic fibrosis and breast cancer. Eigen's predictive power was also superior in identifying *de novo* mutations associated with autism, schizophrenia, epileptic brain dysfunction and intellectual disability. Similarly, high Eigen scores for noncoding variants were more significantly enriched in results from genome-wide association studies, expression quantitative trait loci studies and noncoding cancer mutations from the COSMIC database, compared to CADD and GWAVA scores.

Ionita-Laza remarks that as high-quality data labels become more available, supervised learning will become preferable to unsupervised. In the meantime, the biostatistician suggests gradual adoption of a hybrid approach combining high-quality labeled data with unlabeled data to further increase the predictive power of variant functionality.

Vesna Todorovic

RESEARCH PAPERS

Ionita-Laza, I. *et al.* A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016). (4 January 2016).