COMPUTATIONAL GENOMICS

Neural network architecture search with AMBER

Deep learning applied to genomics can learn patterns in biological sequences, but designing such models requires expertise and effort. Recent work demonstrates the efficiency of a neural network architecture search algorithm in optimizing genomic models.

Yi Zhang, Yang Liu and X. Shirley Liu

Deep learning has been powerful in learning complex functions from data and has been applied in computer vision, natural language processing and biology. If we view the human genome as a book with three billion letters of nucleotides represented by A, C, G and T, genes and gene-controlling sequences are encoded in the book and variations in the genome can link to disease conditions. Neural network models that extract patterns from the sequences can help predict functional genomic elements and interpret genetic variations. However, the current deep learning models for genomics usually involve expert-designed neural network structures and require extensive tuning, making such models unapproachable for most other scientists. In a recent publication in *Nature Machine Intelligence*, Zhang and colleagues[1] present a framework called Automated Modelling for Biological Evidence-based Research (AMBER), which incorporates a deep learning architecture searching algorithm and demonstrates efficient and automatic model selection for genomic problems.

Recent work that applies deep learning to biological sequences has improved our understanding of the human genome. Examples include clinical impact inference for protein-coding mutations[2], pattern recognition among sequences bound by transcription factors (TF)[3], and epigenetic effect prediction for genetic variations[4]. Specifically, as epigenetic profiles of a genomic region reflect biological functions, deep learning models that predict epigenetic profiles from genomic sequences have been powerful in extracting patterns in functional biological sequences (Fig. 1a). Convolutional neural networks (CNNs) are well suited for this task due to their advantage in extracting spatial patterns in sequences. For instance, Zhou et al.[4] constructed a CNN-based model called DeepSEA to map genomic sequences to epigenetic profiles. The trained model can thus predict genetic variants with significant epigenetic effects. Built on the previous success with the expert-designed CNN in DeepSEA, Zhang et al. expanded the work to achieve an optimized model by incorporating a state-of-the-art neural network architecture searching algorithm called Efficient Neural Architecture Search (ENAS)[5]. Usage of ENAS avoided the combinatorial explosion in the CNN model search space and sufficiently explored various architectures to reach an optimized architecture.

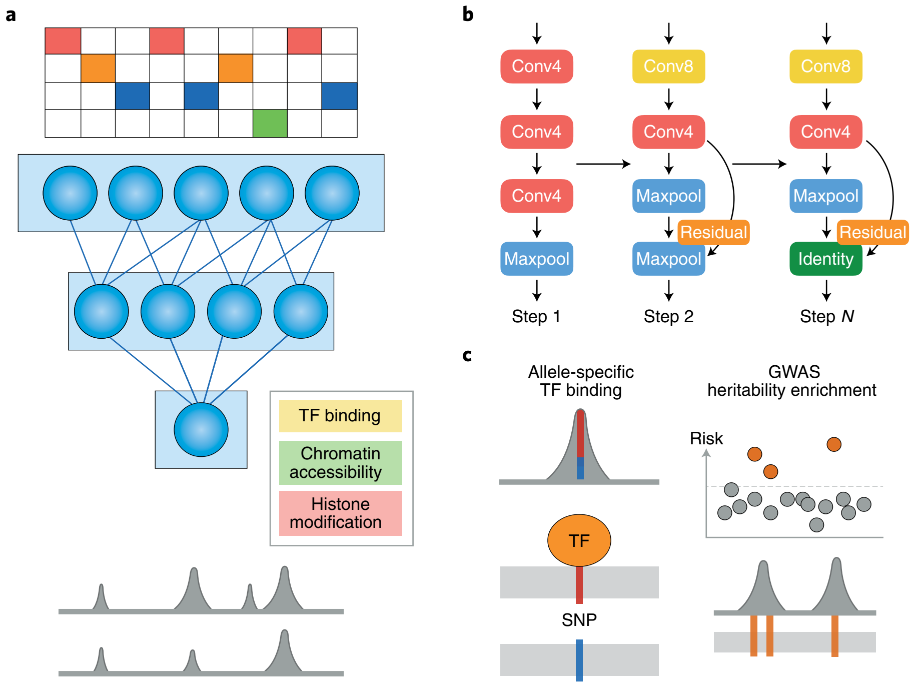

Fig. 1 | Neural architecture search for genomic problems. a, The convolutional neural network is able to model spatial patterns in genomic sequences and predict epigenetic profiles which are biological function indicators. **b**, Each candidate model is a convolutional neural network allowing connection among layers. The neural networks architecture is updated by reinforcement learning. **c**, The predicted functional annotations of the genetic variations can be evaluated by two independent analyses: allele-specific binding of transcription factor (TF), and heritability enrichment of single nucleotide polymorphism (SNPs) from genome-wide association studies (GWAS).

The idea of ENAS is to first define a set of basic layers used in the CNN, as well as ways of connecting the layers, followed by a recurrent neural network (RNN) controller to automatically decide what operation to choose at each layer and how to connect layers in each candidate model (Fig. 1b). Therefore, the parameters trained in the ENAS include both the candidate model weights and the parameters of the RNN controller. During the training process, the model is updated by reinforcement learning so that the performance of the current candidate architecture works as a reward. Finally, the search will converge at an optimized model. A previous Neural Architecture Search (NAS) algorithm[6], searches over each candidate model but has high computation cost. Another Hierarchical Neural Architecture Search

(HNAS)[7] method applies a hierarchical structure to constrain the search space, but depends on well-designed computing cells. Compared to the previous algorithms, the ENAS strategy significantly improves computing efficiency by allowing weights to be shared among all candidate models and by using reinforcement learning in RNN controller updates.

In Zhang et al., the optimal CNN model generated by ENAS took as input genomic sequences of length 1,000 base-pair and was trained to predict 919 epigenetic profiles in a multitasking regime. The model search space consists of 12 layers, each layer chosen from 7 commonly used operations including convolution, dilated, max-pooling and Rectified Linear Unit (ReLU). Based on multiple runs of ENAS, the authors showed that the optimal models (AMBER-Seq models) outperformed baseline sampled models (AMBER-Base) in the epigenetic profile prediction task. Interestingly, the authors observed that the optimal model prefers convolutional operation of kernel size 8 in the bottom and middle layers while selecting max-pooling operation in top layers closer to output. This reflects that the optimized model tends first to convolute the spatial sequence content and then pool and condense the feature to generate a prediction.

Given that the optimized model has good performance in predicting functional categories of genomic sequences, this model can predict the effects of genetic variations. The authors showed that the performance improvement by the AMBER-Seq optimized model could be confirmed by two analyses that independently measure biological significance of genetic variants (Fig. 1c). The first is allele-specific binding of TFs, quantified by allelic imbalance in genomic reads from chromatin immunoprecipitation with sequencing (ChIP-seq) technology. The other is heritability enrichment of genetic variants from genome-wide association studies (GWAS) that discover disease-associated risk variants. In predicting allele specificity of TFs, AMBER-Seq outperforms the expert-designed models, including DeepSEA, deltaSVM[8], and DeepBind[9]. Specifically, the authors presented a case where the AMBER-Seq models align better with biological data. Specifically, AMBER-Seq models predicted stronger binding of SPI1 at G allele at the single nucleotide polymorphism (SNP) rs11658786 compared to A allele, consistent with the DeepSEA model and the ChIP-seq data. Other models — including AMBER-Base, deltaSVM, DeepBind and Jaspar motif — give the opposite direction of allele-specificity. In GWAS heritability enrichment, the authors showed that the annotation generated by AMBER-Seq optimized model delivers more information of variant impact than the AMBER-Base sampled model.

Overall, Zhang et al. demonstrated the power of neural architecture searching to improve deep learning models in genomics. The proposed AMBER framework is an innovative step that enables the improvement of deep learning models built on genomic sequences. The framework is generalizable to other biological machine learning tasks that take biological sequences as input. Examples include imputing TF binding profile in a new cell type and predicting neoantigen capable of eliciting antitumor responses. Moving forward, algorithms that boost biological interpretability for deep learning models can lead to more scientific discoveries and insights in genomics. ❐

Yi Zhang, Yang Liu and X. Shirley Liu ✉

Department of Data Science, Dana-Farber Cancer Institute, Harvard T. H. Chan School of Public Health, Boston, MA, USA.

✉*e-mail: xsliu@ds.dfci.harvard.edu*

Published online: 7 May 2021
https://doi.org/10.1038/s42256-021-00350-x

References

1. Zhang, Z., Park, C. Y., Theesfeld, C. L. & Troyanskaya, O. G. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-021-00316-z (2021).
2. Sundaram, L. et al. *Nat. Genet.* **50**, 1161–1170 (2018).
3. Avsec, Ž. et al. *Nat. Genet.* **53**, 354–366 (2021).
4. Zhou, J. & Troyanskaya, O. G. *Nat. Methods* **12**, 931–934 (2015).
5. Pham, H., Guan, M., Zoph, B., Le, Q. & Dean, J. in *Proc. 35th Int. Conf. Machine Learning* 4095–4104 (PMLR, 2018).
6. Zoph, B. & Le, Q. V. in *Int. Conf. Learning Representations* (ICLR, 2017).
7. Liu, H., Simonyan, K., Vinyals, O., Fernando, C. & Kavukcuoglu, K. *Int. Conf. Learning Representations* (ICLR, 2018).
8. Lee, D. et al. *Nat. Genet.* **47**, 955–961 (2015).
9. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. *Nat. Biotechnol.* **33**, 831–838 (2015).

Competing interests

X.S.L. is a cofounder, board member, SAB, and consultant of GV20 Oncotherapy and its subsidiaries, SAB of 3DMedCare, consultant for Genentech, and stockholder of AMGN, JNJ, MRK, PFE, and receives sponsored research funding from Takeda and Sanofi.