

Machine-learning model makes predictions about network biology

A deep-learning model called Geneformer has been developed and pretrained using about 30 million single-cell gene-expression profiles to enable it to make predictions about gene-network biology in instances in which gene-expression data are limited. Geneformer can be tuned for many downstream applications to accelerate discovery of key gene-network regulators and candidate therapeutic targets.

This is a summary of:

Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* <https://doi.org/10.1038/s41586-023-06139-9> (2023).

Cite this as:

Nature <https://doi.org/10.1038/d41586-023-01504-0> (2023).

The problem

The mapping of gene networks – networks of interactions between different genes – is useful for understanding key biological and disease processes, and for identifying potential drug targets. However, gene-network mapping requires large amounts of gene-expression data so that models can learn the connections between genes. This poses a challenge in instances in which data are limited, such as when working with rare diseases or diseases that affect tissues that are hard to sample in the clinic.

Although data remain limited in such settings, advances in sequencing technologies have driven a rapid expansion in the amount of gene-expression data available from human tissues more broadly. A machine-learning technique called transfer learning has been pivotal to models achieving a greater understanding of natural languages^{1,2} and better computer vision recognition³. It works by tuning deep-learning models that have been pretrained on large, general data sets to carry out a wide array of tasks with limited task-specific data. In other words, transfer learning allows fundamental knowledge gained during pretraining to be transferred to tasks on which a model has not yet been trained. We sought to develop and pretrain a deep-learning model with a large, general gene-expression data set such that it could ‘understand’ gene-network dynamics and could offer predictions about gene interactions and cell state in a broad range of applications in cases impeded by a dearth of data.

The solution

Using publicly available data, we assembled a gene-expression data set (which we call Genecorpus-30M) that includes about 30 million gene-expression profiles of individual cells from a broad range of human tissues. We then used this data set to pretrain a deep-learning model that we developed, called Geneformer, to gain a fundamental understanding of gene-network dynamics (Fig. 1). Geneformer’s pretraining is self-supervised, meaning that it learns from unlabelled data. As an ‘attention-based’ model, Geneformer learns which genes to pay more of a machine-learning form of attention to. Through this self-supervised pretraining, Geneformer learnt to pay greater attention to genes with key roles in cells, such as those that encode transcription factors (proteins that regulate the expression of other genes) and central regulatory nodes in gene networks. Geneformer is also context-aware, enabling

it to make predictions specific to the context of each cell. This is crucial, given that genes’ functions vary in, for example, different cell types, developmental stages and disease states.

Pretraining allowed Geneformer to encode the hierarchy of gene networks, including which genes affect the expression of which other genes. We found that when Geneformer was modified for a diverse set of tasks relevant to the dynamics of gene networks or modifications of the DNA–protein complex chromatin, it consistently boosted predictive accuracy compared with standard alternative approaches. When fine-tuned using limited gene-expression data specific to heart-muscle disease, Geneformer identified candidate therapeutic targets. Targeting two of these candidates in a model of heart-muscle disease based on induced pluripotent stem (iPS) cells led to functional improvement in the contractile force generated by the heart muscle cells derived from the iPS cells. Notably, pretraining with larger and more-diverse data sets consistently improved Geneformer’s predictive power in downstream tasks. This suggests that as the amount of publicly available gene-expression data expands, even better predictions could be made in even more elusive settings for which task-specific data are even more scarce than they are for the examples we examined.

Future directions

By transferring its ‘understanding’ of gene-network dynamics, Geneformer can be applied to diverse questions to accelerate the discovery of key network regulators and candidate therapeutic targets when data are limited. Future models might be developed that can jointly process multiple data types, such as gene-expression and chromatin-accessibility profiles from the same individual cells, as more multimodal data become publicly available. Future models could also extract network hierarchies encoded in Geneformer to explicitly infer network connections in specific cell types and diseases. Geneformer’s context awareness could be especially useful for studying diseases in which multiple cell types are affected and progressive diseases in which therapeutic targets might differ depending on disease stage.

Christina V. Theodoris is at the Gladstone Institutes, San Francisco, California, USA, and **Patrick T. Ellinor** is at the Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

EXPERT OPINION

Geneformer is a deep-neural-network model, pretrained on about 30 million gene-expression profiles, that can be fine-tuned for analysing smaller data sets. Compared with models trained on smaller data sets, Geneformer boosts the accuracy of downstream tasks,

including prediction of dosage sensitivity and therapeutic targets.” (CC BY 4.0)

Natasa Przulj is at the Catalan Institution for Research and Advanced Studies, Barcelona Supercomputing Center, Barcelona, Spain, and at University College London, London, UK.

FIGURE

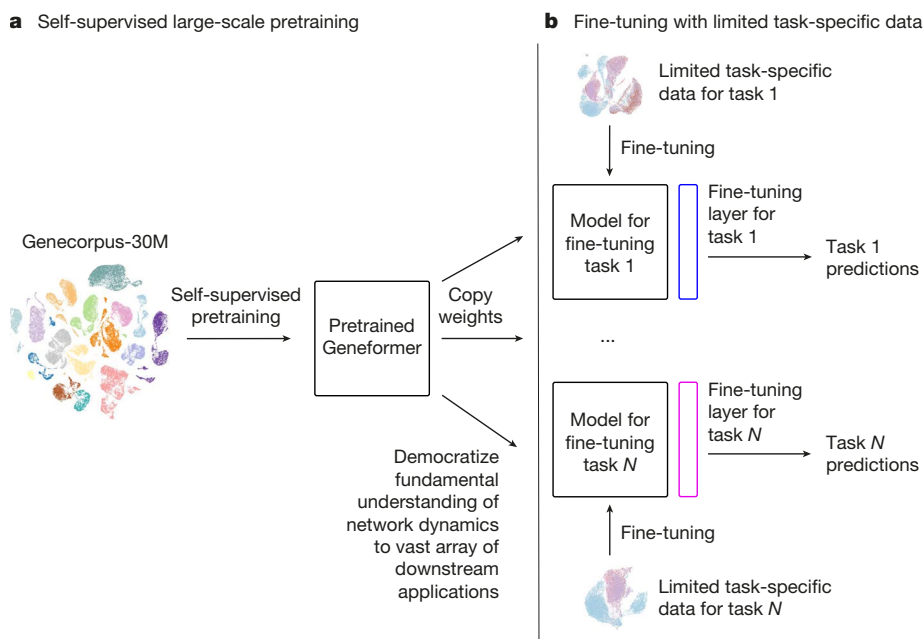


Figure 1 | A machine-learning model that learns gene-network dynamics. **a**, A machine-learning model called Geneformer was pretrained on a data set containing some 30 million single-cell gene-expression profiles (Genecorpus-30M), to learn about the fundamental interactions between different genes (gene networks). The pretraining was self-supervised, meaning Geneformer learnt from unlabelled data. Geneformer involves multiple network layers and learns to pay a machine-learning form of ‘attention’ to different genes by modifying connection strengths (or weights) in these layers. **b**, By copying the pattern of weights in its network, then adding a ‘fine-tuning’ layer, Geneformer could be modified to transfer its knowledge about gene networks to address a multitude of different tasks (1–N) for which there were limited data.

BEHIND THE PAPER

Mapping the gene-regulatory networks that drive human disease makes it possible to design network-correcting treatments that target core disease mechanisms rather than merely manage symptoms. We previously developed a framework for mapping disease-dependent gene networks to enable network-based screening that leveraged machine learning and modelling with induced pluripotent stem cells; this approach identified a promising network-correcting therapy for cardiac-valve disease^{4,5}. However, to computationally infer the network map, large amounts of gene-expression data were needed for the model to learn the

connections between genes; such data are often unavailable. Seeing the impacts of transfer learning in other fields^{1–3} inspired us to develop a transfer-learning model for gene-network biology that uses the large human-gene-expression data sets now available to improve predictions in settings with limited data. We were enthralled by the ability of this approach to allow predictions in so many applications relevant to gene-network dynamics, chromatin dynamics and disease modelling, including identifying candidate therapeutic targets for cardiomyopathy.

C.V.T.

REFERENCES

1. Vaswani, A. et al. in *Proc. 31st Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 6000–6010 (NIPS, 2017).
2. Devlin, J. et al. in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (eds Burstein, J. et al.) 4171–4186 (Assoc. Comput. Linguistics, 2019).
3. He, K. et al. in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
4. Theodoris, C. V. et al. *Cell* **160**, 1072–1086 (2015).
5. Theodoris, C. V. et al. *Science* **371**, eabd0724 (2021).

FROM THE EDITOR

Transfer learning is a powerful artificial-intelligence-based approach that lays a foundation of probabilistic associations on the basis of massive data sets. Geneformer is not only a tool for predicting changes in gene expression and cell state in the setting of *in silico* genetic manipulation (gene knockout or overexpression), but also a major step forward in our attempts to digitally probe biological systems.

George Caputa, Senior Editor, *Nature*