Article

Transfer learning enables predictions in network biology

https://doi.org/10.1038/s41586-023-06139-9

Received: 29 March 2022

Accepted: 27 April 2023

Published online: 31 May 2023

Check for updates

Christina V. Theodoris[1,2,3,4]✉, Ling Xiao[2,5], Anant Chopra[6], Mark D. Chaffin[2], Zeina R. Al Sayed[2], Matthew C. Hill[2,5], Helene Mantineo[2,5], Elizabeth M. Brydon[6], Zexian Zeng[1,7], X. Shirley Liu[1,7,8] & Patrick T. Ellinor[2,5]✉

Mapping gene networks requires large amounts of transcriptomic data to learn the connections between genes, which impedes discoveries in settings with limited data, including rare diseases and diseases affecting clinically inaccessible tissues. Recently, transfer learning has revolutionized fields such as natural language understanding[1,2] and computer vision[3] by leveraging deep learning models pretrained on large-scale general datasets that can then be fine-tuned towards a vast array of downstream tasks with limited task-specific data. Here, we developed a context-aware, attention-based deep learning model, Geneformer, pretrained on a large-scale corpus of about 30 million single-cell transcriptomes to enable context-specific predictions in settings with limited data in network biology. During pretraining, Geneformer gained a fundamental understanding of network dynamics, encoding network hierarchy in the attention weights of the model in a completely self-supervised manner. Fine-tuning towards a diverse panel of downstream tasks relevant to chromatin and network dynamics using limited task-specific data demonstrated that Geneformer consistently boosted predictive accuracy. Applied to disease modelling with limited patient data, Geneformer identified candidate therapeutic targets for cardiomyopathy. Overall, Geneformer represents a pretrained deep learning model from which fine-tuning towards a broad range of downstream applications can be pursued to accelerate discovery of key network regulators and candidate therapeutic targets.

Mapping the gene regulatory networks that drive disease progression enables screening for molecules that correct the network by normalizing core regulatory elements, rather than targeting peripheral downstream effectors that may not be disease modifying[4,5]. However, mapping the gene network architecture requires large amounts of transcriptomic data to learn the connections between genes, which impedes network-correcting drug discovery in settings with limited data, including rare diseases and diseases affecting clinically inaccessible tissues. Although data remain limited in these settings, recent advances in sequencing technologies have driven a rapid expansion in the amount of transcriptomic data available from human tissues more broadly. Furthermore, single-cell technologies have facilitated the observation of transcriptomic states without averaging the expression of genes across multiple cells, potentially providing more precise data for inference of network interactions, especially in diseases driven by dysregulation of multiple cell types.

Recently, the concept of transfer learning has revolutionized fields such as natural language understanding[1,2] and computer vision[3] by leveraging deep learning models pretrained on large-scale general datasets that can then be fine-tuned towards a vast array of downstream tasks with limited task-specific data that would be insufficient to yield meaningful predictions when used in isolation. Unlike modelling approaches that necessitate retraining a new model from scratch for each task[6,7], this approach democratizes the fundamental knowledge learned during the large-scale pretraining phase to a multitude of downstream applications distinct from the pretraining learning objective, transferring knowledge to new tasks (Fig. 1a and Extended Data Fig. 1a,b). The advent of the self-attention mechanism[1,2] has further transformed the deep learning field by generating context-aware models that are able to pay attention to large input spaces and learn which elements are most important to focus on in each context, boosting predictions in a wide realm of applications[2,8]. Gene regulatory network architectures are highly context-dependent, and attention-based models, known as transformers, may be exceptionally suited to context-specific modelling of network dynamics.

Here, we developed a context-aware, attention-based deep learning model, Geneformer, pretrained on large-scale transcriptomic data to enable predictions in settings with limited data. We assembled a large-scale pretraining corpus, Genecorpus-30M, comprising 29.9 million human single-cell transcriptomes from a broad range of tissues from publicly available data. We then pretrained Geneformer on this corpus using a self-supervised masked learning objective to gain a fundamental

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Cardiovascular Disease Initiative and Precision Cardiology Laboratory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [4]Harvard Medical School Genetics Training Program, Boston, USA. [5]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [6]Precision Cardiology Laboratory, Bayer US LLC, Cambridge, MA, USA. [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA. ✉e-mail: christina.theodoris@gladstone.ucsf.edu; ellinor@mgh.harvard.edu