Modeling and Predicting Single-Cell Multi-Gene Perturbation Responses

Gefei Wang, Tianyu Liu, Jia Zhao, Youshu Cheng, **Hongyu Zhao***

Department of Biostatistics, Yale University

ABSTRACT

Understanding cellular responses to genetic perturbations is essential for deciphering gene regulation and phenotype formation. While high-throughput single-cell RNA-sequencing has facilitated detailed profiling of heterogeneous transcriptional responses to perturbations at the single-cell level, there remains a pressing need for computational models that can decode the mechanisms driving these responses and accurately predict outcomes to prioritize target genes for experimental design. This presentation introduces a deep generative learning framework designed to model and predict single-cell transcriptional responses to genetic perturbations, including single-gene and combinatorial multi-gene perturbations. The method effectively integrates prior biological knowledge and disentangles basal cell states from perturbation-specific salient representations by leveraging gene embeddings derived from large language models. Through comprehensive evaluations on multiple single-cell CRISPR Perturb-seq datasets, the approach outperformed state-of-the-art methods in predicting perturbation outcomes, achieving higher prediction accuracy. Notably, it demonstrated robust generalization to unseen target genes and perturbations, and its predictions captured both average expression changes and the heterogeneity of single-cell responses. Furthermore, the predictions enable diverse downstream analyses, including identifying differentially expressed genes and exploring genetic interactions, demonstrating its utility and versatility. This is joint work with Gefei Wang, Tianyu Liu, Jia Zhao, and Youshu Cheng.

Keywords: Statistical genetics, foundation models, embedding, biostatistics, genomics

Transcriptomic Analysis and Image-Based Deep Learning Prognostic Model for Lung Adenocarcinoma

Yang-Ming Yeh[1], Yahsuan Chang[2], Liang-Yin Tao[3], **Hsuanyu Chen[1]**

[1]*Institute of Statistical Science, Academia Sinica, Taipei City, Taiwan,*

[2]*National Health Research Institutes - Zhunan Campus, Zhunan Township, Taiwan*

[3]*Department of Statistics, University of California, Davis, Davis, CA, United States*

ABSTRACT

Background

Therapies such as EGFR tyrosine kinase inhibitors (TKIs) and immune checkpoint inhibitors (ICIs) have improved outcomes for EGFR-mutant and wild-type lung adenocarcinoma patients, respectively. However, drug resistance and limited survival remain significant challenges. This study proposes a novel prognosis prediction model using whole-transcriptome data integrated with an image-based deep learning approach.

Methods

Four cohorts were analyzed: one training cohort (RNA-seq, n=391) and three independent validation cohorts (TCGA RNA-seq, n=394; GSE68465, n=443; GSE13213, n=117). RNA-seq data were converted into images using pixel-encoding strategies to preserve transcriptomic information. A convolutional neural network (CNN) model was trained on all gene expression data without feature selection. To mitigate overfitting, the training cohort was split into training (n=259), testing (n=62), and validation (n=70) subsets. The CNN was trained on the training subset, validated on the testing and validation subsets, and independently evaluated on the three external validation cohorts.

Results

RNA-seq data were successfully transformed into images, enabling effective CNN model training. The model stratified patients into high- and low-risk groups. In the training cohort, high-risk patients exhibited significantly shorter overall survival (OS; $P < 0.01$). Similar findings were observed in the external validation cohorts: TCGA ($P = 0.001$), GSE68465 ($P < 0.0001$), and GSE13213 ($P = 0.003$). Prediction accuracies were 0.71, 0.81, and 0.92 for the training, testing, and validation subsets, respectively, with an average accuracy of 0.70 across the external cohorts.

Conclusions

This study presents an innovative image-based deep learning approach for analyzing whole-transcriptome data without requiring differential gene selection. By capturing comprehensive transcriptomic information, this method offers potential for enhanced prognostic modeling and molecularly guided lung cancer treatments.

Back to Sessions List

Modeling the Impact of Personal Genome Variation on Molecular Phenotypes

Nilah Ioannidis

UC Berkeley; UCSC; CZ Biohub

ABSTRACT

Understanding inter-individual variation in molecular, cellular, and other clinically-relevant phenotypes is an important challenge in precision medicine. Sequence-based genomic deep learning models that predict gene expression and other molecular phenotypes directly from DNA sequence can be applied in silico to sequences containing any combination of rare or common genetic variants, with great potential to predict the genetic contribution to variation in such phenotypes. However, despite success in explaining variation in molecular phenotypes across the genome and across a variety of cell types, we and others recently found that current sequence-based genomic deep learning models have limited ability to explain variation in gene expression across different individuals based on their personal genome sequences. I will discuss our work to characterize the cross-individual performance of such models on gene expression and other molecular phenotypes, with resulting insights into their understanding of regulatory variation. I will also discuss our recent efforts to develop models with improved understanding of variation across individuals using several strategies, such as incorporating personal genome and transcriptome data during model training and using a hierarchical approach to first model more locally-regulated phenotypes such as chromatin accessibility.

Keywords: Genomics; Machine Learning; Deep Learning.

An AI System to Help Scientists Write Expert-Level Empirical Software

Eser Aygün[1,*], Anastasiya Belyaeva[2,*], Gheorghe Comanici[1,*], **Dr. Marc A. Coram[2,*]**, Hao Cui[2,*], Jake Garrison[3,*], Renee Johnston[2,*], Anton Kast[2,*], Cory Y. McLean[2,*], Peter Norgaard[2,*], Zahra Shamsi[2,*], David Smalling[1,*], James Thompson[2,*], Subhashini Venugopalan[2,*], Brian P. Williams[2,*], Chujun He[2,4,**], Sarah Martinson[2,5,**], Martyna Plomecka[2,6,**], Lai Wei[2], Yuchen Zhou[2], Qian-Ze Zhu[2,5,**], Matthew Abraham[2], Erica Brand[2], Anna Bulanova[1], Jeffrey A. Cardille[2,7], Chris Co[2], Scott Ellsworth[2], Grace Joseph[2], Malcolm Kane[2], Ryan Krueger[2,5,**], Johan Kartiwa[2], Dan Liebling[2], Jan-Matthis Lueckmann[2], Paul Raccuglia[2], Xuefei (Julie) Wang[2,8,**], Katherine Chou[2], James Manyika[2], Yossi Matias[2], John C. Platt[2], Lizzie Dorfman[2], Shibl Mourad[1,‡] and Michael P. Brenner[2,5,‡]

Google Research, Google Zurich

ABSTRACT

The cycle of scientific discovery is frequently bottlenecked by the slow, manual creation of software to support computational experiments. To address this, we present an AI system that creates expert-level scientific software whose goal is to maximize a quality metric. The system uses a Large Language Model (LLM) and Tree Search (TS) to systematically improve the quality metric and intelligently navigate the large space of possible solutions. The system achieves expert-level results when it explores and integrates complex research ideas from external sources. The effectiveness of tree search is demonstrated across a wide range of benchmarks. In bioinformatics, it discovered 40 novel methods for single-cell data analysis that outperformed the top human-developed methods on a public leaderboard. In epidemiology, it generated 14 models that outperformed the CDC ensemble and all other individual models for forecasting COVID-19 hospitalizations. Our method also produced state-of-the-art software for geospatial analysis, neural activity prediction in zebrafish, time series forecasting and numerical solution of integrals. By devising and implementing novel solutions to diverse tasks, the system represents a significant step towards accelerating scientific progress.

Keywords: empiricial software, LLM-based data analysis, forecasting, batch-normalization

[1]Google DeepMind, [2]Google Research, [3]Google Platforms and Devices, [4]Massachusetts Institute of Technology, [5]School of Engineering and Applied Sciences, Harvard University,

[6]Google Cloud, [7]Faculty of Agricultural and Environmental Sciences, McGill University, [8]California Institute of Technology

*Equal contribution in alphabetical order
** Carried out as part of a student researchership at Google Research
‡ To whom correspondence should be addressed: shibl@google.com, mbrenner@google.com

Back to Sessions List