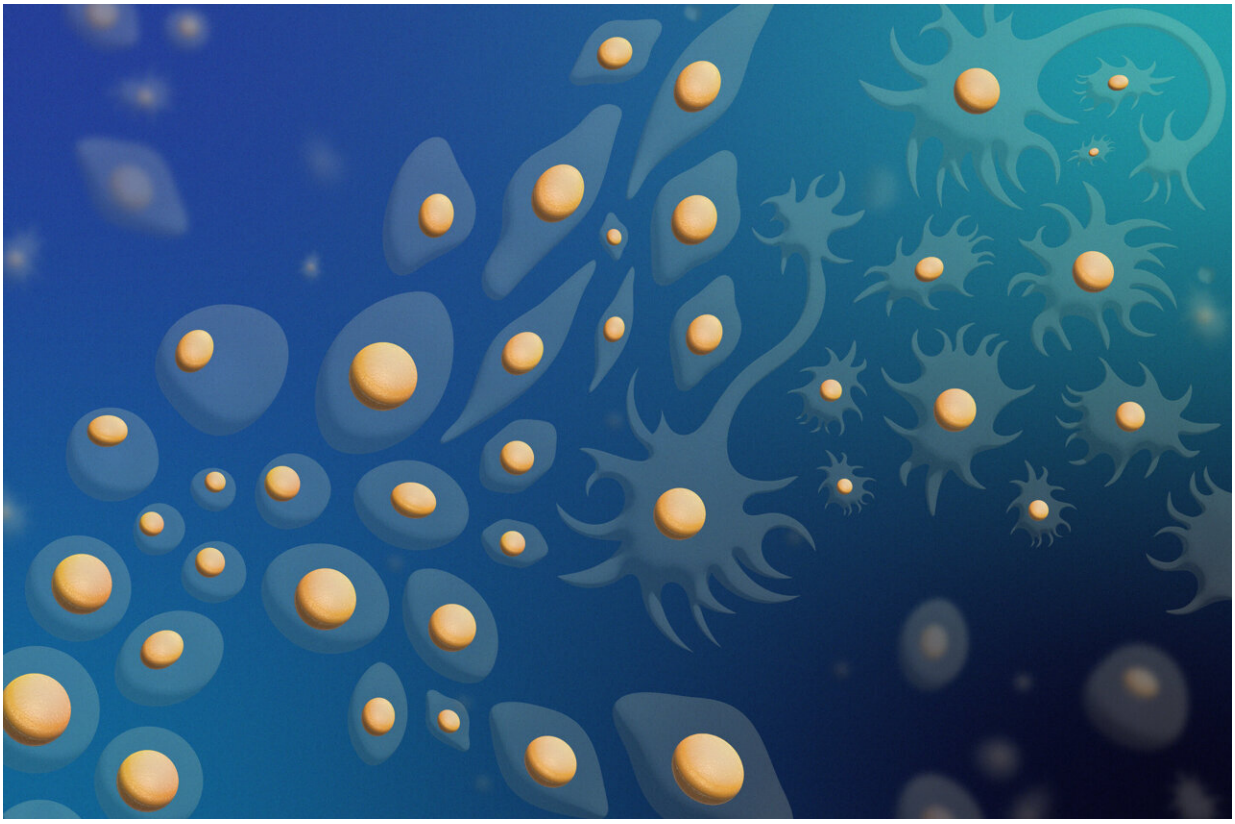
Transfer learning paves the way for new disease treatments

March 4 2024



Cells change shape and function when reprogrammed in response to the exogenous alteration of expression of a handful key genes identified by the computational approach. Credit: Ellie Mejía/Northwestern University

Technological advances in gene sequencing and computing have led to an explosion in the availability of bioinformatic data and processing

power, respectively, creating a ripe nexus for artificial intelligence (AI) to design strategies for controlling cell behavior.

In a new study, Northwestern University researchers have reaped fruit from this nexus by developing an AI-powered transfer learning approach that repurposes publicly available data to predict combinations of gene perturbations that can transform cell type or restore diseased cells to health.

The study, "Cell reprogramming design by transfer learning of functional transcriptional networks," is published this week in the *Proceedings of the National Academy of Sciences*.

Since the completion of the human genome project 20 years ago, scientists have known that human DNA comprises more than 20,000 genes. However, it has remained a mystery as to how these genes work together to orchestrate the hundreds of different cell types in our body.

Surprisingly, essentially by guided trial-and-error, researchers have demonstrated that it is possible to "reprogram" cell type by manipulating only a handful of genes. The human genome project also facilitated advances in sequencing technologies, making it cheaper not only to read the genetic code, but also to measure gene expression, which quantifies the precursors of the proteins that carry out cell functions.

This increase in affordability has led to the accumulation of a massive amount of publicly available bioinformatic data, raising the possibility of synthesizing these data to rationally design gene manipulations that can elicit desired cell behaviors.

The ability to control cell behavior, and thus transitions across cell types, can be applied to regrowing injured tissues or to transforming cancer cells back into normal cells.

Injured tissues resulting from strokes, arthritis and multiple sclerosis affect 2.9 million individuals each year in the United States, costing as much as $400 million per year. Meanwhile, cancers are responsible for around 10 million deaths annually worldwide with economic costs in the trillions of dollars.

Because the current standard of care does not regenerate tissues and/or has limited efficacy, there is a critical need to develop more effective treatments that are broadly applicable, which in turn requires identification of molecular interventions that can be inferred from high-throughput data.

In the new study, the researchers train their AI to learn how gene expression gives rise to cell behavior using publicly available gene expression data. The predictive model generated by this learning process is transferred to specific cell reprogramming applications. In each application, the approach finds the combination of gene manipulations that is most likely to induce the desired cell type transition.

Unprecedented exploration of the genome-wide dynamics

"Our work stands out from previous approaches to rationally design strategies to manipulate cell behavior," said Thomas Wytock, lead author of the paper and member of the Center for Network Dynamics at Northwestern University. "These approaches mostly fall into two categories: one in which genes are organized into networks according to their interactions or common properties; and another in which the expression of genes from healthy and diseased cells are compared to single out the genes that show the largest differences."

In the first category, there is a tradeoff between realism and scale. Some

network models comprise many genes but only say whether a relationship is present or absent. Other models are quantitative and experimentally validated but necessarily involve a small number of genes and relationships.

Northwestern's new work retains the strengths of both types of models: it is inclusive of all genes in the cell and quantitative in representing their expressions. This is achieved by reducing the expression of nearly 20,000 individual genes to no more than 10 linear combinations of such genes, which are weighted averages referred to as eigengenes.

"Eigengenes basically show how genes operate in concert, making it possible to simplify the dynamics of a large dynamical network to just a few moving parts," said Adilson Motter, the Charles E. and Emma H. Morrison Professor of Physics at the Weinberg College of Arts and Sciences, director of the Center for Network Dynamics at Northwestern University and the study's senior author. "Each eigengene can be thought of as a generalized pathway that is approximately independent of the others. So, eigengenes pick up the relevant correlations and independences in the gene regulatory network."

Approaches in the second category can find individual genes associated with a change in cell behavior but fail to specify how genes work together to enable this change. The new approach overcomes this challenge by recognizing that genes change their expressions in concert. The quantitative accounting of this property in terms of eigengenes makes it possible to additively combine their responses to different gene perturbations by suitably scaling them. The combined responses can then be input into the AI model to determine which perturbations elicit the desired cell behavior.

Averting combinatorial explosion

Equipped with this AI model, the researchers curated publicly available data to identify how gene expression changes when a single gene is perturbed by exogenously raising or lowering its expression. They then developed an algorithm to solve the inverse problem, which is to predict gene combinations that are most likely to induce a desired reprogramming transition, such as to cause diseased cells to behave as healthy cells.

The approach that results from integrating the data and algorithm circumvents combinatorial explosion that would result from testing all combinations in order to identify the effective ones. This is significant because experiments can test only a limited number of cases, and the algorithm provides a way to identify the most promising cases to be tested.

"The approach shines in its ability to examine myriad combinations computationally," said Wytock. "For example, the pairwise combinations of 200 perturbations yield 20,000 cases, triples yield over 1.3 million cases, and this number keeps growing exponentially. Because the algorithm employs optimization, the approach can compare predictions across a potentially infinite number of combinations through the magic of calculus."

Another challenge circumvented by the approach is that the gene perturbations can combine in a non-additive manner. For example, consider the impact of gene perturbations on cellular growth rate and imagine perturbations halve the growth rate when applied in isolation.

The effect of two such perturbations combines non-additively if they reduce growth to either significantly more or significantly less than half of a half (or one quarter). Even though there is a large body of research characterizing non-additive interactions between genes, the new approach is effective even without having to account for such deviations

from additivity.

"This is a case in which the whole is well approximated by the sum of the parts," Motter said.

"This property of the interventions needed to induce transitions between cell types is counterintuitive because the cell types themselves emerge from collective interactions among genes."

Because the approach addresses the main challenges to control cell behavior, it can be applied to many different biomedical conditions, including those that will benefit from future data.

A flexible model for forthcoming data

The fact that responses to gene perturbations combine additively facilitates generalization across cell types. For example, if a gene is disrupted in a skin cell, the resulting impact on expression would be largely the same in a liver cell.

Thus, the AI-powered approach can be thought of as a platform into which data pertaining to a specific disease in a specific patient may be inserted. The approach may be applied whenever curing the disease can be conceived as a reprogramming problem, as in the case of cancers, diabetes, and autoimmune diseases, which all result from cell dysfunction.

The versatility of the approach allows the gene expression in a single study to be rapidly contextualized across all available data in the National Center for Biotechnology Information's Sequencing Read Archive, which is the largest publicly available repository for gene expression data.

This archive has grown 100-fold from 10 terabytes to 1,000 terabytes between 2012 and 2022 and continues to grow exponentially as sequencing costs decrease. This work provides a critical tool for translating this wealth of data into specific predictions of how genes work together to control the behavior of normal and diseased cells.

More information: Thomas P. Wytock et al, Cell reprogramming design by transfer learning of functional transcriptional networks, *Proceedings of the National Academy of Sciences* (2024). DOI: 10.1073/pnas.2312942121

Provided by Northwestern University