Media contact:
Gina Kirchweger
gina@lji.org
848.357.7481

LJI scientists develop new method to match genes to their molecular "switches"

New research may help scientists hunt down solid tumors and better diagnose disease

LA JOLLA, CA—Scientists at La Jolla Institute for Immunology (LJI) have developed a new computational method for linking molecular marks on our DNA to gene activity. Their work may help researchers connect genes to the molecular "switches" that turn them on or off.

This research, published in *Genome Biology*, is an important step toward harnessing machine learning approaches to better understand links between gene expression and disease development.

"This research is about bringing a three-dimensional perspective to studying DNA modifications and their function in our genome," says LJI Associate Professor Ferhat Ay, Ph.D., who co-led the study with LJI Professor Anjana Rao, Ph.D.

Ay and Rao are working to pinpoint regions of the genome that contain molecular enhancers, or "switches," which fine tune the levels of gene expression and determine when and where genes will be on or off. This work requires researchers to develop computational tools that can harness complex genomic data and find which enhancers are connected to which genes.

For the new study, the LJI researchers employed machine learning tools called linear and graph neural networks to processgenomic data and make these connections. Neural networks are computational tools modeled after how neurons in the brain process information and identify patterns. Graph neural networks are able to integrate 3D information, such as the DNA physical interactions inside the cell.

Edahí González-Avalos, Ph.D., spearheaded the development of this graph neural network as a UC San Diego graduate student jointly mentored by Rao and Ay at LJI. "We can use this to prioritize DNA interactions within the genome," says González-Avalos, who now works at Guardant Health.

The neural network goes to work

The researchers trained new neural networks that learn how the presence of an important DNA modification called 5hmC, either near the gene or far away from it, is related to gene expression

activity. This attachment of a hydroxymethyl group to cytosine has been associated with enhancer activity.

In fact, 5hmC appears to have such an important influence on gene expression that scientists have termed 5hmC the "sixth letter" of the DNA alphabet alongside A, T, C, G, and an intermediate methylated form called 5mC (the fifth base). The conversion of 5mC to 5hmC on cytosine is associated with enhancer activity—the more 5hmC, the greater the level of enhancer activity.

In previous studies, researchers in the Rao Lab had discovered that the location of 5hmC in the genome changed depending on what cell types they were looking at—and what genes those cell types expressed. The actual DNA code would be the same, but 5hmC would be attached to the genome in different places in a liver cell versus a lung cell or a brain cell.

This 5hmC distribution controlled the expression of different gene sets in these different types of cells. The researchers had found that 5hmC attaches to regions of the genomes that work as enhancers—the same regions that help switch gene expression on and off—as well as to the genes themselves. These differences in active genes and enhancers are what distinguishes a liver cell from cells in the lung or neurons in the brain.

"The distribution of 5hmC differs from cell type to cell type," says Rao. "If you can tell where 5hmC is, you can infer what cell type is producing the DNA you are studying." For example, if a cell is a cancer cell, you can infer what type of cancer it is, even if it has metastasised (moved far away from) its original site in the body.

The new research method allows a simpler connection to be made between genes and enhancers than was possible with earlier methods.

"This paper was a proof-of-concept showing we could use these graph neural networks to predict interactions between genes and enhancers using 5hmC," says González-Avalos.

Ay says he was pleased to see how the neural network revealed connections between genes and 5hmC in far-away regions of the genome. These long-distance connections across the genome helped prioritize regions with the ability to enhance gene expression.

"What is exciting is that some of these distant enhancers are novel regulatory elements that have not been discovered before," says Ay.

Going forward, the researchers hope to take a closer look at 5hmC distribution to better understand enhancer and gene interactions in human cells. "This research was done with data from mouse cells," says Ay. "Next, we'd want to look at 5hmC and these interactions in immune cells and cancer cells from patients."

Hope for better cancer diagnostics

Just as in normal cells, 5hmC distribution differs between cancer cell types. This means the new LJI method may prove valuable for understanding the genetic mechanisms that drive cancer development.

Rao says the new method may also open the door to faster, more accurate cancer diagnoses. Currently, it is very hard for scientists to analyze blood samples for signs of solid tumors in the body. "Solid tumor cells aren't usually available in the blood. What's available is DNA, and it's usually DNA that's been partially degraded," says Rao.

As Rao explains, doctors could help more patients—and potentially detect cancers earlier—if they could look beyond the DNA itself and analyze 5hmC distribution instead.

More work needs to be done before scientists have the tools for this kind of cancer detection, but Ay says the new work shows the power of combining experimental data with new computational methods. "This suggests that by applying our new method we can identify new and unannotated distant enhancers," says Ay.

What to know about neural networks

Scientists use neural networks to understand extremely complicated datasets. "Neural" networks got their name because they are inspired by how the human brain processes information.

Imagine you are driving on the highway, and you see a sign for your exit. The neurons in your eyes communicate the road sign information to other neurons in your brain, which signal other neurons, and so on.

As the neurons process the information, they give each answer weight, or importance. Is the sign shaped like a road sign? Yes? Does the road sign include letters? Yes. What does it say? Let's see. Neurons route the information along the right path to quickly find an answer.

After what feels like a split second, your neurons tell you: YES. That's the sign for your exit.

You are aware of the input (the road sign) and your brain's conclusion. But your brain came up with its own way of processing the information.

Computational scientists build neural networks that come to conclusions in a similar way. Researchers can take complicated inputs, such as genomic data, and neural networks can process the information to spot hidden trends and make predictions.

In these artificial neural networks, information travels through artificial neurons, or nodes, that process the information and signal other artificial neurons. Every time the signal passes through a

neural connection, it carries weight, which alters how the neural network routes the signal for further processing.

Scientists don't have to intervene to help the network come to conclusions. Instead, the network adapts and learns as it goes—which is why scientists call this process machine learning.

###