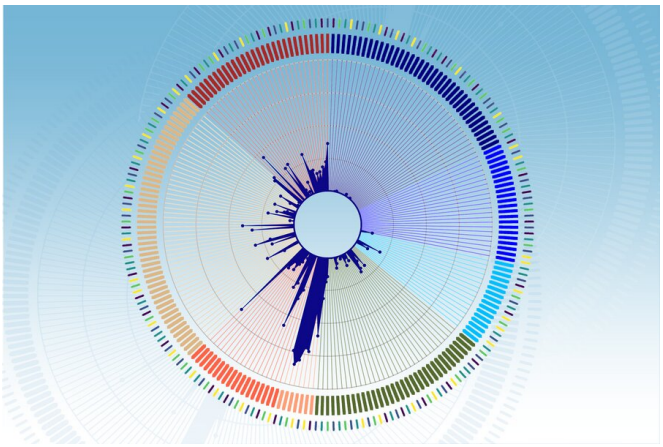
Using machine learning to identify undiagnosable cancers

2 September 2022, by Bendta Schroeder



“Machine learning tools like this one could empower oncologists to choose more effective treatments and give more guidance to their patients,” says Koch Institute clinical investigator Salil Garg. Credit: Massachusetts Institute of Technology

The first step in choosing the appropriate treatment for a cancer patient is to identify their specific type of cancer, including determining the primary site—the organ or part of the body where the cancer begins.

In rare cases, the origin of a cancer cannot be determined, even with extensive testing. Although these cancers of unknown primary tend to be aggressive, oncologists must treat them with non-targeted therapies, which frequently have harsh toxicities and result in low rates of survival.

A new deep-learning approach developed by researchers at the Koch Institute for Integrative Cancer Research at MIT and Massachusetts General Hospital (MGH) may help classify cancers of unknown primary by taking a closer look the gene expression programs related to early cell development and differentiation.

"Sometimes you can apply all the tools that pathologists have to offer, and you are still left without an answer," says Salil Garg, a Charles W. (1955) and Jennifer C. Johnson Clinical Investigator at the Koch Institute and a pathologist at MGH. "Machine learning tools like this one could empower oncologists to choose more effective treatments and give more guidance to their patients."

Garg is the senior author of a new study, published Aug. 30 in *Cancer Discovery*. The artificial intelligence tool is capable of identifying cancer types with a high degree of sensitivity and accuracy. Garg is the senior author of the study, and MIT postdoc Enrico Moiso is the lead author.

Machine learning in development

Parsing the differences in the gene expression among different kinds of tumors of unknown primary is an ideal problem for machine learning to solve. Cancer cells look and behave quite differently from normal cells, in part because of extensive alterations to how their genes are expressed. Thanks to advances in single cell profiling and efforts to catalog different cell expression patterns in cell atlases, there are copious—if, to human eyes, overwhelming—data that contain clues to how and from where different cancers originated.

However, building a machine learning model that leverages differences between healthy and normal cells, and among different kinds of cancer, into a diagnostic tool is a balancing act. If a model is too complex and accounts for too many features of cancer gene expression, the model may appear to learn the training data perfectly, but falter when it encounters new data. However, by simplifying the model by narrowing the number of features, the model may miss the kinds of information that would lead to accurate classifications of cancer types.

In order to strike a balance between reducing the number of features while still extracting the most relevant information, the team focused the model on signs of altered developmental pathways in cancer cells. As an embryo develops and undifferentiated cells specialize into various organs, a multitude of pathways directs how cells divide, grow, change shape, and migrate. As the tumor develops, cancer cells lose many of the specialized traits of a mature cell. At the same time, they begin to resemble embryonic cells in some ways, as they gain the ability to proliferate, transform, and metastasize to new tissues. Many of the gene expression programs that drive embryogenesis are known to be reactivated or dysregulated in cancer cells.

The researchers compared two large cell atlases, identifying correlations between tumor and embryonic cells: the Cancer Genome Atlas (TCGA), which contains gene expression data for 33 tumor types, and the Mouse Organogenesis Cell Atlas (MOCA), which profiles 56 separate trajectories of embryonic cells as they develop and differentiate.

"Single-cell resolution tools have dramatically changed how we study the biology of cancer, but how we make this revolution impactful for patients is another question," explains Moiso. "With the emergence of developmental cell atlases, especially ones that focus on early phases of organogenesis such as MOCA, we can expand our tools beyond histological and genomic information and open doors to new ways of profiling and identifying tumors and developing new treatments."

The resulting map of correlations between developmental gene expression patterns in tumor and embryonic cells was then transformed into a machine learning model. The researchers broke down the gene expression of tumor samples from the TCGA into individual components that correspond to a specific point of time in a developmental trajectory, and assigned each of these components a mathematical value. The researchers then built a machine-learning model, called the Developmental Multilayer Perceptron (D-MLP), that scores a tumor for its developmental components and then predicts its origin.

Classifying tumors

After training, the D-MLP was applied to 52 new samples of particularly challenging cancers of unknown primary that could not be diagnosed using available tools. These cases represented the most challenging seen at MGH over a four-year period beginning in 2017. Excitingly, the model classed the tumors to four categories, and yielded predictions and other information that could guide diagnosis and treatment of these patients.

For example, one sample came from a patient with a history of breast cancer who showed signs of an aggressive cancer in the fluid spaces around the abdomen. Oncologists initially could not find a tumor mass, and could not classify cancer cells using the tools they had at the time. However, the D-MLP strongly predicted ovarian cancer. Six months after the patient first presented, a mass was finally found in the ovary that proved to be the origin of tumor.

Moreover, the study's systematic comparisons between tumor and embryonic cells revealed promising, and sometimes surprising, insights into the gene expression profiles of specific tumor types. For instance, in early stages of embryonic development, a rudimentary gut tube forms, with the lungs and other nearby organs arising from the foregut, and much of the digestive tract forming from the mid- and hindgut. The study showed that lung-derived tumor cells showed strong similarities not just to the foregut as might be expected, but to the to mid- and hindgut-derived developmental trajectories. Findings like these suggest that differences in developmental programs could one day be exploited in the same way that genetic mutations are commonly used to design personalized or targeted cancer treatments.

While the study presents a powerful approach to classifying tumors, it has some limitations. In future work, researchers plan to increase the predictive power of their model by incorporating other types of data, notably information gleaned from radiology, microscopy, and other types of tumor imaging.

"Developmental gene expression represents only one small slice of all the factors that could be used

to diagnose and treat cancers," says Garg. "Integrating radiology, pathology, and gene expression information together is the true next step in personalized medicine for cancer patients."

More information: Enrico Moiso et al, Developmental Deconvolution for Classification of Cancer Origin, *Cancer Discovery* (2022). DOI: 10.1158/2159-8290.CD-21-1443

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology