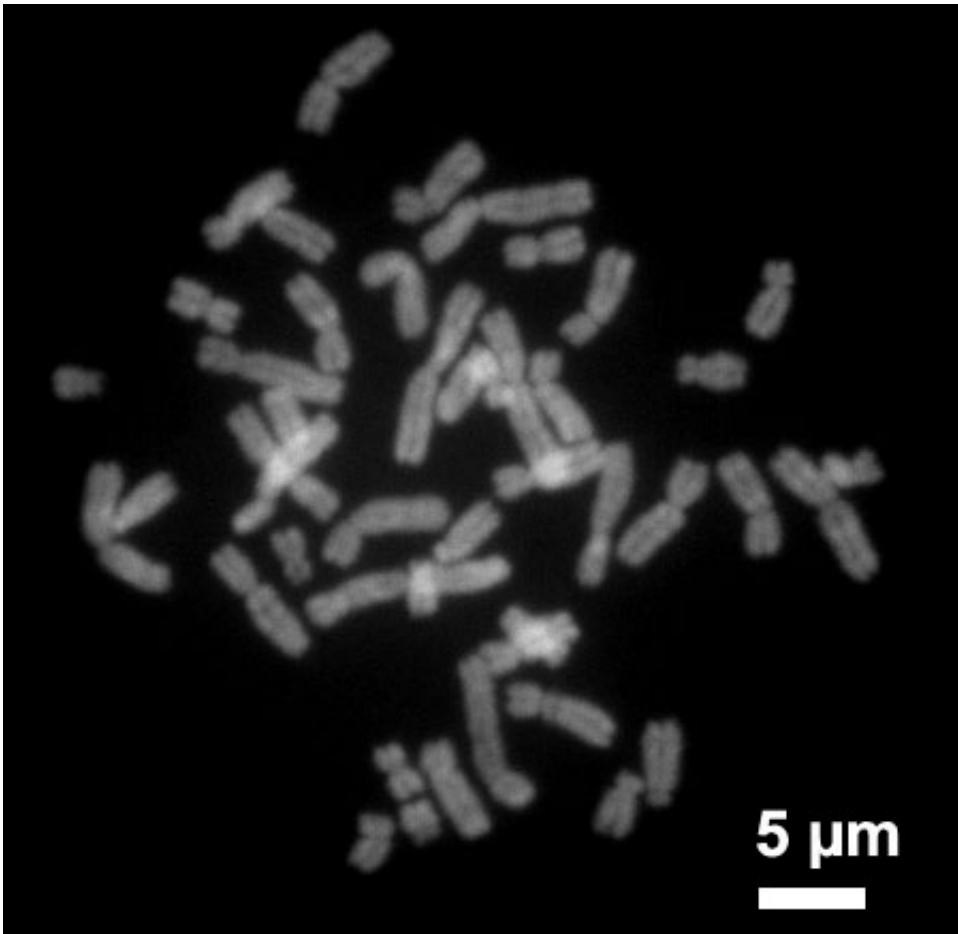# Neural network fills in data gaps for spatial analysis of chromosomes

November 7 2019

Human chromosomes during metaphase. Credit: Steffen Dietzel/Wikipedia

Computational methods used to fill in missing pixels in low-quality images or video also can help scientists provide missing information for

how DNA is organized in the cell, computational biologists at Carnegie Mellon University have shown.

Filling in this missing information will make it possible to more readily study the 3-D structure of chromosomes and, in particular, subcompartments that may play a crucial role in both disease formation and determining cell functions, said Jian Ma, associate professor in CMU's Computational Biology Department.

In a research paper published today by the journal *Nature Communications*, Ma and Kyle Xiong, a CMU Ph.D. student in the CMU-University of Pittsburgh Joint Ph.D. Program in Computational Biology, report that they successfully applied their machine learning method to nine cell lines. This enabled them, for the first time, to study differences in spatial organization related to subcompartments across those lines.

Previously, subcompartments could be revealed in only a single cell type of lymphoblastoid cells—a cell line known as GM12878—that has been exhaustively sequenced at great expense using Hi-C technology, which measures spatial interactivity among all regions of the genome.

"We now know a lot about the linear composition of DNA in chromosomes, but in the nuclei of human cells, DNA isn't linear," Xiong said. "Chromosomes in the cell nucleus are folded and packaged into 3-D shapes. That 3-D structure is critical to understanding the cellular functions in development and diseases." Subcompartments are of particular interest because they reflect spatial segregation of chromosome regions with high interactivity.

Scientists are eager to learn more about the juxtaposition of subcompartments and how it affects cell function, Ma said. But until now researchers could calculate the patterns of subcompartments only if they had an extremely high coverage Hi-C dataset—that is, the DNA had

been sequenced in great detail to capture more interactions. That level of detail is missing in the datasets for cell lines other than GM12878.

Working with Ma, Xiong used an artificial neural network called a denoising autoencoder to help fill in the gaps in less-than-complete Hi-C datasets. In computer vision applications, the autoencoder can supply missing pixels by learning what types of pixels typically are found together and making its best guess. Xiong adapted the autoencoder to high-throughput genomics, using the dataset for GM12878 to train it to recognize what sequences of DNA pairs from different chromosomes typically might be interacting with each other in 3-D space in the cell nucleus.

This computational method, which Ma and Xiong have dubbed SNIPER, proved successful in identifying subcompartments in eight cell lines whose interchromosomal interactions based on Hi-C data were only partially known. They also applied SNIPER to the GM12878 data as a control. But Xiong noted that it is not yet known how widely this tool can be used on all other cell types. He and Ma are continuing to enhance the method, however, so it can be used on a variety of cellular conditions and even in different organisms.

"We need to understand how subcompartment patterns are involved in the basic functions of cells, as well as how mutations can affect these 3-D structures," Ma said. "Thus far, in the few cell lines we've been able to study, we see that some subcompartments are consistent across cell types, while others vary. Much remains to be learned."