Using machine-learning to find mutations in similar genome sequences of cancer samples

July 20 2021, by Bob Yirka

Credit: CC0 Public Domain

A team of researchers working at the Francis Crick Institute has developed a way to find mutations in similar genome regions of cancer samples. In their paper published in the journal *Nature Biotechnology*, the group describes using a machine-learning algorithm to spot cancerous mutations in non-unique parts of the genome.

As part of human evolutionary history, sections of the genome have undergone rearrangement, and in some cases, duplication. Such duplications have been found to be problematic when attempting to find mutations. Current scanning methods toss out short sequences that are identified as ambiguous, which means that segments of the genome that are very similar to one another are not included in such reports—and that means that any mutations will be missed. In this new effort, the researchers have developed a means for finding mutations in non-unique parts of the genome.

The approach involved first developing a list of genome regions known to be similar to other regions and then using them to teach a machine-learning algorithm how to recognize them. Researchers then used the algorithm to spot mutations in different tissues—2,658 samples from the Pan-Cancer Analysis of Whole Genome dataset. The researchers uncovered mutations in 1,744 coding sequences along with thousands of other mutations in non-coding sequences. They also found that their algorithm had a false discovery rate of approximately 7% and a validation rate of more than 80%.

The researchers noted that those mutations that involved coding sequences have an impact on protein sequences, some of which have been linked to cancer types. They also found instances of mutations that led to protein changes, that have also been linked to specific kinds of cancers. As one example, they found a recurrent mutation in the KMT2C and PIK3CA genes. They also found mutations that have been linked to breast cancer. And they found mutations that are involved in regulatory regions, including some in the immunoglobulin family.

The researchers suggest their technique can be used by other teams as a means to overcome issues with overlooking mutations in near-duplicate genetic regions.

More information: Maxime Tarabichi et al, A pan-cancer landscape of somatic mutations in non-unique regions of the human genome, *Nature Biotechnology* (2021). DOI: 10.1038/s41587-021-00971-y