New algorithm rapidly finds anomalies in gene expression data

November 27 2019

Credit: CC0 Public Domain

Computational biologists at Carnegie Mellon University have devised an algorithm to rapidly sort through mountains of gene expression data to find unexpected phenomena that might merit further study. What's more, the algorithm then re-examines its own output, looking for mistakes it has made and then correcting them.

This work by Carl Kingsford, a professor in CMU's Computational Biology Department, and Cong Ma, a Ph.D. student in computational biology, is the first attempt at automating the search for these anomalies in gene expression inferred by RNA sequencing, or RNA-seq, the leading method for inferring the activity level of genes.

As they report today in the journal *Cell Systems*, the researchers already have detected 88 anomalies—unexpectedly high or low levels of expression of regions within genes—in two widely used RNA-seq libraries that are both common and not previously known.

"We don't yet know why we're seeing those 88 weird patterns," Kingsford said, noting that they could be a subject of further investigation.

Though an organism's genetic makeup is static, the activity level, or expression, of genes varies greatly over time. Gene expression analysis has thus become a major tool for biological research, as well as for diagnosing and monitoring cancers.

Anomalies can be important clues for researchers, but until now finding them has been a painstaking, manual process, sometimes called "sequence gazing." Finding one anomaly might require examining

200,000 transcript sequences—sequences of RNA that encode information from the gene's DNA, Kingsford said. Most researchers therefore zero in on regions of genes that they think are important, largely ignoring the vast majority of potential anomalies.

The algorithm developed by Ma and Kingsford automates the search for anomalies, enabling researchers to consider all of the transcript sequences, not just those regions where they expect to see anomalies. This technology could uncover many new phenomena, such as the 88 previously unknown common anomalies found in the multi-tissue RNA-seq libraries.

But Ma noted that identifying anomalies is often not clear cut. Some RNA-seq "reads," for instance, are common to multiple genes and transcripts and sometimes get mapped to the wrong one. If that occurs, a genetic region might appear more or less active than expected. So the algorithm re-examines any anomalies it detects and sees if they disappear when the RNA-seq reads are redistributed between the genes.

"By correcting anomalies when possible, we reduce the number of falsely predicted instances of differential expression," Ma said.

More information: *Cell Systems* (2019). www.cell.com/cell-systems/full … 2405-4712(19)30381-3

Provided by Carnegie Mellon University