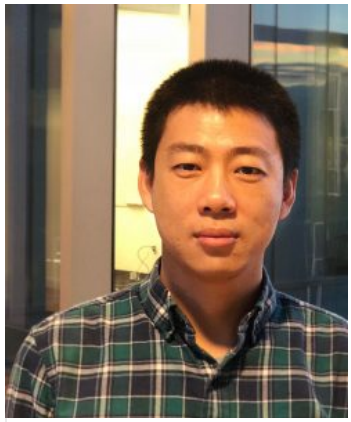
Computational Biology Department | Carnegie Mellon University School of Computer S

CMU Software Assembles RNA Transcripts More Accurately

Nov 13, 2017 | Faculty, Lane Fellows, Research

Method should help scientists understand regulation of gene expression.

By Byron Spice



Lane Fellow Mingfu Shao

PITTSBURGH—Computational biologists at Carnegie Mellon University have developed a more accurate computational method for reconstructing the full-length nucleotide sequences of the RNA products in cells, called transcripts, that transform information from a gene into proteins or other gene products.

Their software, called Scallop, will help scientists build a more complete library of RNA transcripts and thus help scientists better understand the regulation of gene expression.

A report on Scallop by Carl Kingsford, associate professor of computational biology, and Mingfu Shao, Lane Fellow in the School of Computer Science's Computational Biology Department, is being published online today by the journal Nature Biotechnology.

Scallop is a so-called transcript assembler, taking fragments of RNA sequences, called reads, that are produced by high-throughput RNA sequencing technologies (RNA-seq), and putting them back together, like pieces of a puzzle, to reconstruct complete RNA transcripts.

"There are many existing assemblers," Shao said, "but these existing methods are still not accurate enough."

When compared to two leading assemblers, StringTie and TransComb, Scallop is 34.5 percent and 36.3 percent more accurate for transcripts consisting of multiple exons —– subunits of a gene that encode part of the gene product.

Like other reference-based assemblers, Scallop begins by constructing a graph to organize reads that are mapped to the corresponding locations on the gene's DNA. Many alternative paths exist for connecting the reads together, however, so errors are easily made. Scallop improves its odds by using a novel algorithm to take full advantage of the information from reads that span several exons to guide it to the correct assembly paths.

Scallop proves particularly adept when assembling less abundant RNA transcripts, improving upon the accuracy of StringTie and TransComb by 67.5 percent and 52.3 percent.

The researchers have released Scallop as open software on the GitHub repository.

"We've had more than 100 downloads already and, based on the feedback we've received, people are really using it," Shao said. "We expect more users now that our paper is out."

The Gordon and Betty Moore Foundation, The Shurl and Kay Curci Foundation, the National Science Foundation and the National Institutes of Health supported this research.