Researchers develop tool to evaluate genome sequencing method

January 2 2013

Advances in bio-technologies and computer software have helped make genome sequencing much more common than in the past. But still in question are both the accuracy of different sequencing methods and the best ways to evaluate these efforts. Now, computer scientists have devised a tool to better measure the validity of genome sequencing.

The method, which is described in the journal *PLOS ONE*, allows for the evaluation of a wide range of genome sequencing procedures by tracking a small group of key statistical features in the basic structure of the assembled genome. Such sequence-assembly algorithm lays out the individual short reads (strings of DNA's four nucleic acid bases sampled from the target genome) to put together the complete genome sequence —much like a complex jig-saw puzzle. The method uses techniques from statistical inference and learning theory to select the most significant features. Surprisingly, the method concludes that many features thought by human experts to be the most important were actually highly misleading.

The work was conducted by researchers at New York University's Courant Institute of Mathematical Sciences, NYU School of Medicine, Sweden's KTH Royal Institute of Technology, and Cold Spring Harbor Laboratory.

Current evaluation methods of genome sequencing are typically imprecise. They rely on what amounts to "crowd sourcing," with scientists weighing in on the accuracy of a sequencing method. Other

evaluations use apples-to-oranges comparisons in making assessments, thus limiting their value.

In the *PLOS ONE* work, the researchers expanded upon an earlier system they created, Feature Response Curve (FRCurve), which offers a global picture of how genome-sequencing methods, or assemblers, are able to deal with different regions and different structures in a large complex genome. Specifically, it points out how an assembler might have traded off one kind of quality measure at the expense of another kind. For instance, it shows how aggressively a genome assembler might have tried to pull together a group of genes into a contiguous piece of the genome, while incorrectly rearranging their correct order and copy numbers.

However, FRCurve has a significant limitation—it can only gauge the accuracy of certain kinds of assemblers at one time, thereby excluding comparisons among the range of sequencing methods currently being employed. Many of these methods, where the original FRCurve failed, are becoming highly popular, as they are specifically designed to work with the most established next-generation sequencing technologies and are able to perform some error correction and data compression. However, by doing so, they also discard the original signature of key statistical features (e.g., position and orientation of the reads used to generate the candidate sequence) that FRCurve needs for evaluation.

The work reported in PLOS ONE unveils a new method, FRCbam, which has the capability to evaluate a much wider class of assemblers. It does so by reverse engineering the latent structures that were obscured by error-correction and data compression; and it performs this operation rapidly by using efficient and scalable mapping algorithms.

Instead of assumption-ridden simulation or expensive auxiliary methods, FRCbam validates its analysis by examining a large ensemble of assemblers working on a large ensemble of genomes, selected from

crowd-sourced competitions like GAGE and Assemblathons. This way, FRCbam can characterize the statistics that are expected and then validate any individual system with respect to it.

FRCbam and FRCurve are expected to be used routinely to rank and evaluate future genome projects. This method is currently employed to evaluate the sequence assembly of the Norway Spruce, one of the largest genomes sequenced so far—it is seven times longer than the human genome.

Provided by New York University