Researchers develop systematic approach for accurate DNA sequence reconstruction

December 5 2011

Researchers at the Genome Institute of Singapore (GIS) have, for the very first time, developed a computational tool that comes with a guarantee on its reliability when reconstructing the DNA sequence of organisms, thus enabling a more streamlined process for reconstructing and studying genomic sequences.

The work, lead by Dr. Niranjan Nagarajan, Assistant Director of Computational and Mathematical Biology at the GIS, was reported in the November 2011 issue of the Journal of Computational Biology.

The genomic study of life (plants and animals alike) is based on computational tools that can first piece together the DNA sequence of these organisms, a process called genome assembly, that is similar to solving a giant puzzle or putting together the words in a book from a shredded copy. Due to the sheer scale of this challenge, existing approaches for genome assembly rely on heuristics and often result in incorrect reconstructions of the genome. The work reported here represents the first algorithmic solution for genome assembly that provides a quality guarantee and scales to large datasets. A new and improved implementation for this algorithm called Opera is now freely available at sourceforge.net/projects/operasf/ and has been used at the GIS for successfully assembling large plant and animal genomes.

The assembled genome of an organism forms the basis for a range of downstream biological investigations and serves as a critical resource for the research community. The draft human genome, for example, was

obtained at the expense of billions of dollars, serves as a fundamental resource for biomedical research and is, in fact, still being refined. Improved assembly tools thus serve to generate the most complete and accurate draft genomes that can be reconstructed from the data, avoiding mis-assembly related dead-ends for downstream research as well as minimizing the painstaking effort needed to refine and correct a draft assembly.

"Genetic studies of organisms of interest for human health (such as those causing infectious diseases), agriculture, animal husbandry and other areas of the bio-economy, such as biofuels, are driven by the availability of draft genome sequences, said Dr Nagarajan. "This research describes a novel computational approach to reconstruct more complete and accurate draft genomes. From an algorithmic perspective, Opera demonstrates the utility of a clear optimization function and an exact algorithm derived from a parametric complexity analysis in providing a robust solution to a seemingly intractable problem."

Mihai Pop, Associate Prof, Department of Computer Science; and Interim Director, Center for Bioinformatics and Computational Biology at the University of Maryland said: "Opera is an important advance in genome assembly algorithms – currently it is the best stand-alone genome scaffolder available in the community. In Opera, Dr. Nagarajan's team has introduced a rigorous theoretical framework for genome scaffolding as well as a practical implementation that achieves remarkable performance. These results are impressive given the substantial research in the field over the past 30 years, as well as the numerous developments spurred in recent years by advances in sequencing technologies."

  More information: The research findings described in the press release can be found in the November 2011 issue of *Journal of Computational Biology* under the title "Opera: Reconstructing Optimal

Genomic Scaffolds with High-Throughput Paired-End Sequences".

Provided by Agency for Science Technology and Research (A*STAR)