A Divide-and-Conquer Algorithm for Large-Scale De Novo Transcriptome Assembly through Combining Small Assemblies from Existing Algorithms

Sing-Hoi Sze
Department of Computer Science and Engineering
Department of Biochemistry & Biophysics
Texas A&M University
College Station, TX 77843
Email: shsze@cse.tamu.edu

Jonathan J. Parrott
Department of Entomology
Texas A&M University
College Station, TX 77843
Email: jonathan.parrott@tamu.edu

Aaron M. Tarone
Department of Entomology
Texas A&M University
College Station, TX 77843
Email: tamlucilia@tamu.edu

Abstract—While high-throughput sequencing has facilitated studies of entire transcriptomes in non-model organisms along with differential expression analysis, it has become increasingly difficult to perform de novo transcriptome assemblies due to the large amount of libraries, which include many experimental conditions or developmental stages with replicated experiments.

Although transcriptome assembly algorithms that can process a large amount of data are available, algorithms that have higher accuracy are generally very memory-intensive, thus they can only be used to construct small assemblies. We consider a divide-and-conquer strategy that allows these algorithms to be utilized, by subdividing a large RNA-Seq data set into small libraries and assembling each individual library in parallel, while employing a merging algorithm to combine the small assemblies into a large transcriptome.

The merging algorithm starts by picking a subset of high quality transcripts to form a transcriptome while preferring longer transcripts, which are more highly expressed and better assembled. A de Bruijn graph is constructed to help extend some of these transcripts at the two ends when there are no ambiguities. To reduce redundancy, shorter transcripts with all their corresponding nodes in the de Bruijn graph covered by longer transcripts are removed.

We validate our algorithm by performing Schizosaccharomyces pombe, Drosophila melanogaster and Arabidopsis thaliana transcriptome assemblies using publicly available RNA-Seq libraries. When compared to an existing algorithm that returns a single assembly directly, our algorithm achieves comparable or better accuracy in most cases while having a higher number of translocated transcripts.

We demonstrate the utility of our algorithm by assembling a large set of 84 Cochliomyia macellaria RNA-Seq libraries that we have generated in our lab, which is about 248 G in size.