Abstract

Dramatic advances in DNA sequencing technology have made it possible to study microbial environments by direct sequencing of environmental DNA samples. Yet, due to the huge volume and high data complexity, current de novo assemblers cannot handle large metagenomic datasets or fail to perform assembly with acceptable quality. This paper presents the first parallel solution for decomposing the metagenomic assembly problem without compromising the post-assembly quality. We transform this problem into that of finding weakly connected components in the de Bruijn graph. We propose a novel distributed memory algorithm to identify the connected subgraphs, and present strategies to minimize the communication volume. We demonstrate the scalability of our algorithm on a soil metagenome dataset with 1.8 billion reads. Our approach achieves a runtime of 22 min using 1280 Intel Xeon cores for a 421 GB uncompressed FASTQ dataset. Moreover, our solution is generalizable to finding connected components in arbitrary undirected graphs.