SWAP-Assembler: A Scalable De Bruijn Graph Based Assembler for Massive Genome Data

Jintao Meng [1,2], Bingqiang Wang [2], Yanjie Wei [1,*], Shengzhong Feng [1,*], Jiefeng Cheng [1], Pavan Balaji [3]

[1]: Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, P.R. China.

[2]: Beijing Genomics Institutes Shenzhen, P.R. China.

[3]: Mathematics and Computer Science Division, Argonne National Laboratory, USA

[*]: Dr. Wei and Prof. Feng are corresponding authors.

Emails: Jintao Meng (jt.meng@siat.ac.cn); Bingqiang Wang (wangbingqiang@genomics.cn); Shengzhong Feng (sz.feng@siat.ac.cn); Jiefeng Cheng (jf.cheng@siat.ac.cn); Pavan Balaji (balaji@mcs.anl.gov)

Sequencing species with large genome can produce Tara bytes data, and the de bruijn graph constructed from these data - in some cases having ten billions of vertices and edges - poses challenges to genome assembly problem. This paper presents a multi-step bi-directed graph (MSG) to abstract the standard genome assembly (SGA) problem. With MSG, SGA can be decomposed into several edge merging operations, and this operation and the multi-step semi-extended edges are proved to be a semi-group. Afterwards a small world asynchronous parallel model (SWAP), which can automatically detect and make use of the locality of computation and communication in semi-group to maximize potential parallelism, is proposed for this type of computation. With MSG and SWAP, SWAP-assembler is developed, the scalability test shows that it can scale up to 1024 cores with improved performance, the 2008 Asian (YanHuang) genome can be assembled in 2 hours, which is 6 times faster than SOAPdenovo on one server with 32 cores, and about 24 times faster than ABySS with 1024 cores.