In the field of genome assembly research where assemblers are dominated by de-Bruijn graph-based approaches, string graph based assembly approach is getting more attention because of its ability to losslessly retain information from sequence data. Despite the advantages provided by a string graph in repeat detection and in maintaining read coherence, high computational cost for constructing a string graph hinders its usability for genome assembly. Even though different algorithms have been proposed over last decade for string graph construction, efficiency is still a challenge due to the demand for processing large amount of sequence data generated by NGS technologies. Therefore, in this thesis, we provide a novel, linear time and alphabet-size-independent algorithm SOF which uses the property of irreducible edges and transitive edges to efficiently construct string graph from an overlap graph. Experimental results show that SOF is at least 2 times faster than the string graph construction algorithm provided in SGA, one of the most popular string graph based assembler, while maintaining almost same memory footprint as SGA. Moreover, availability of SOF as a subprogram in SGA assembly pipeline will give a user facilities to access the preprocessing and postprocessing steps for genome assembly provided in SGA.

Major: Computer Science

Educational Career:
Bachelor’s of Computer Science and Engineering, BS, 2013, University of Dhaka
Master’s of Computer Science and Engineering, MS, 2016, University of Dhaka

Committee in Charge:
Shibu Yooseph, Chair, Computer Science
Shaoyjie Zhang, Computer Science
Sharma Valliyil Thankachan, Computer Science

Approved for distribution by Shibu Yooseph, Committee Chair, on March 20, 2019.

The public is welcome to attend.