Workshop: Transcript assembly through a d^2-based MST approach

Y. Zhang[1], D.M. Rao[1], J. Mueller[4], M. Ozden[1], C. Liang[1,2], and J.E. Karro[1,3]
Departments of [1]Computer Science, [2]Botany, [3]Microbiology and [4]Information Technology Services
Miami University
Oxford, Ohio U.S.A.
Email: karroje@muohio.edu

***Abstract*—PEACE and EAST, together forming a pipeline for the clustering and assembly of EST transcript fragments, make use of a novel d^2-based minimum spanning tree algorithm that results in quickly generated, high quality assemblies when applied to the results of Sanger sequencing. The tool is considerably more robust to sequencing error than competing assembly fragments, and has shown to be effective when applied to next generation technologies producing reads of moderate size (e.g. 454).**

***Keywords*-transcript; assembly; clustering; next generation**

I. Introduction

We present PEACE + EAST, an open-source user-friendly pipeline for the high-throughput De novo clustering (PEACE) and assembly (EAST) of gene transcript fragments, applicable to the outputs of both Sanger sequencing and Next Generation sequencing technologies. Both tools are built on the novel use of minimum spanning trees in combination with the d^2 alignment-free sequence distance measure. By using d^2 to identify overlap, a parallelized implementation of Prims algorithm that allows us to induce clusters from the results MSTs, and an exploration of the tree structure to infer assembly information within the cluster, we achieve high quality scores with a low runtime, allowing us to tackle large clustering problems.

The PEACE+EAST pipeline has been extensively compared with several well-known assembly tools (including the WCD clustering tool and the Cap3, TGICL, Velvet, and MIRA assembly tools) on a number of sequencing technologies (including Sanger, 454, Illumina, and hybrid sets derived from these technologies). We find that it outperforms all tools, both in terms of quality and runtime, when applied to Sanger sequences (Figure 1), it outperforms all tools in terms of quality when applied to 454 sequences averaging 250 bp (Figure 2), is not competitive with Velvet on Illumina sequences, and it has mixed results on the hybrid sets outperforming competing tools when dealing with higher coverage sets or data subjected to higher sequencing error rates. Further, we find our tool pipeline considerably more robust to sequencing error, achieving high-quality assemblies when other tools fail to overcome high base-call error rates.

In short, the PEACE+EAST toolset, available with a user-friendly GUI at http://www.peace-tools.org/, is both more versatile and considerably more robust to sequencing error than competing tools. While the performance of EAST does suffer significantly from the presence of extremely short sequences, there is potential to both tailor the algorithm to increase quality and implement a parallel version to improve runtime. We believe this serves as a proof-of-concept for the d^2 MST based strategy, showing that it has the potential to produce high quality alignments for both Sanger and Next Generation sequence sets.

978-1-61284-852-5/11/$26.00 ©2011 IEEE