Parahaplo: A Parallel Tree-based Solver for Haplotype Assembly on CPU-GPU Systems

Robert Clucas
University of the Witwatersrand
School of Electrical and Information Engineering
Johannesburg, South Africa
robert.clucas@students.wits.ac.za

Sasha Naidoo
University of the Witwatersrand
School of Electrical and Information Engineering
Johannesburg, South Africa
sasha.naidoo@students.wits.ac.za

***Abstract*—Haplotyping is an important problem in bioinformatics, and can significantly impact areas such as disease diagnosis and drug discovery. Organisms such as humans have two copies of each chromosome, and haplotyping partitions genotype calls between the two chromosomes. The Minimum error correction (MEC) formulation of the haplotype assembly (HA) problem is NP-hard and involves correcting the minimum number of SNPs to infer the haplotypes from a set of aligned reads. Most existing solutions either make assumptions to reduce complexity, or require impractical computation times. In this work, we present Parahaplo, a parallel tree-based solver for the general case HA problem where columns are both heterozygous and homozygous, and reads may contain multiple gaps. Input matrices obtained from DNA sequencing technology are typically extremely large, up to hundreds of Gigabytes, and hence cannot fit into GPU memory. To overcome this, Parahaplo uses many-core CPUs to split the input matrices into uncorrelated sub-blocks, which are then represented as a binary tree, where each node is a position in the haplotype. Additionally, the CPUs find the root node of each tree, as well as the correlation between the haplotype positions. The trees for each sub-block are solved in parallel on the GPUs using a parallel, breadth first branch-and-bound algorithm. The breadth first search utilises the massively-parallel nature of GPUs and ensures the tree is always balanced, minimizing the number of underutilized cores and limiting thread synchronization. Highly parallel bounding operators are designed to efficiently determine the upper and lower bound, allowing nodes which cannot provide an optimal solution to be pruned early in the search, limiting the size of the search tree.**

***Keywords*-Haplotype; CPU; GPU; Tree; Branch; Bound;**

I. Introduction

Parahaplo is a parallel tree-based solver for the haplotype assembly problem, designed for large-scale hetrogenous systems ...

A. Subsection Heading Here

1) Subsubsection Heading Here:

II. Type style and Fonts

III. Conclusion

The conclusion goes here. this is more of the conclusion

Acknowledgment