Shenghua Chen
Prof. T. Andrew Binkowski
MPCS 56430
May 24th, 2022

Distributed Framework for Gene Finding using Open-MPI

1. Introduction

Genes are parts of DNA whose job is to make specific proteins that play a key role in the structure and function of the body. In computational biology, gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes. The rapid advancement in generation sequencing techniques demands further development in the DNA sequencing approaches to keep up with the pace. Among them is the development of an approach and algorithm that deploys distributed computational resources more effectively and efficiently in gene finding. The objective of the current work is to develop a framework for Gene Finding using Open-MPI.

2. Background

Several computational frameworks for the identification of protein-coding genes exist. The frameworks have introduced incremental features, causing rapid progression of genomic DNA. For example, FINDER is a gene annotator pipeline tasked with automating the process of downloading short reads, aligning them, and using the assembled transcripts to generate gene annotations (Banerjee et al., 2021). Another example is the mGene – an accurate SVM-based gene finding with an application to nematode genomes conceptualized by (Schweikert et al., 2009). mGene works by combining the flexibility of generalized hidden Markov models with the predictive power of machine learning approaches, such as SVM in an unprecedented fashion, producing unparalleled accuracy in gene prediction. Chothia, & Durbin (2002) developed GAZE, a generic framework for the integration of gene-prediction data by dynamic programming. It remains the case, however, that the inflexibility (single algorithm) and scalability (single node) limitations of the above approaches in automating the identification of gene structures are yet to be solved.

3. Methodology/Approach

To address the two drawbacks of the existing methods: the limitation of running on single nodes and their gene-specific nature, a novel approach is proposed. The goal of the proposed approach is develop a flexible framework that can distribute on multiple nodes and it can accept user defined customized algorithms. Figure 1 summarizes the working of the proposed framework.

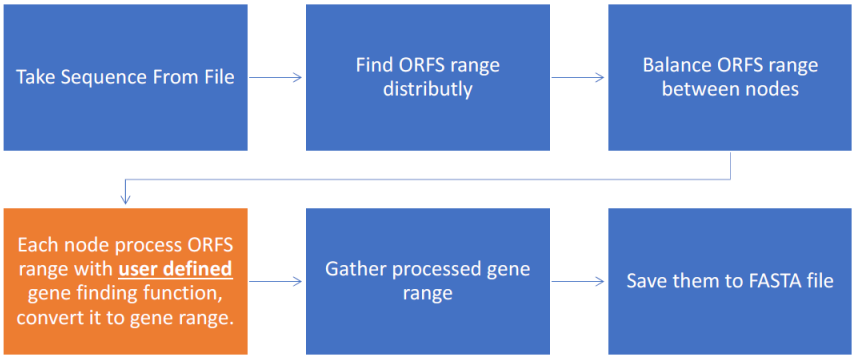

Figure 1: A concept for distributed framework for gene finding using Open-MPI

The program takes in the sequence file for each node; hence, every node has a complete copy of the sequence. This means that each node will be processing a part of sequence. For example, node 1 are going to find possible ORFS that start with location [0, 99], while node 2 will find possible ORFS that start with the location [100,199]. Because the proposed system deploys openMP, the above step will execute the task multi-threadly to output ORFS ranges in different nodes. Therefore, there is need to verify if the outputted ORFS is a gene. Checking the ORFS is usually the most tasking job, and to eliminate the unnecessary overhead, we balanced the ORFS ranges between the nodes before they were processed. The ORFS proceed for processing using specified gene-finding functions. The User defined gene finding functions can provided with user complied dynamic linked library. This process outputs genes in multiple nodes, which the program gathers and aggregates into a single node and saves the range to sequence in a FASTA file.

4. Results

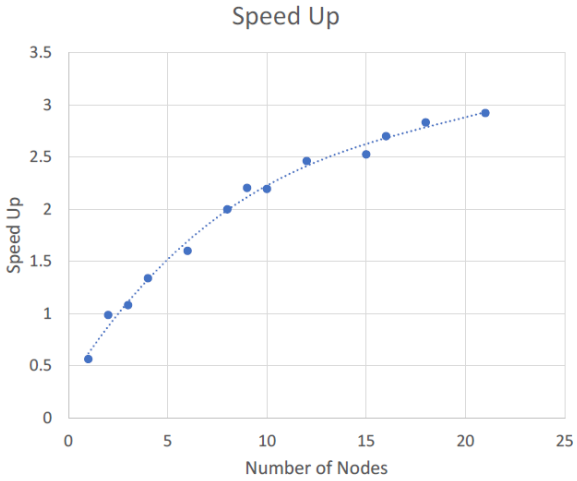

Figure 2: Performance of the approach

The program was tested on a slurm cluster on google cloud with each node having 2 threads. Speed up is based on complexity of user-defined gene finding function. This step is fully parallelizable. Speedup is defined as: $Speedup = \frac{T_{\text{single\_node}}}{T_{mutiple\_nodes}}$. Because getting gene evaluation from user-defined function was fully parallelizable, taking longer in this part would reduce the percentage of time taken by non-parallelizable, increasing the speedups. With highly complex gene finding function, speedup factor can be over 4 when there are 5 nodes.

5. Discussion

This paper contributes to the ongoing progression in genomic DNA by providing a flexible distributed framework for gene finding using open-MPI. The framework presented will ease deployment of gene finding job, reducing the turnaround time. The proposed framework eliminates the limitations of running on single nodes and gene-specific nature of current approaches that hinders scalability. Future studies should explore solutions based on edge computing and heterogeneous computing.

Works Cited

Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P., & Andorf, C. M. (2021). FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics*, *22*(1), 1–26. doi:10.1186/s12859-021-04120-9

Howe, K. L., Chothia, T., & Durbin, R. (2002). GAZE: A genetic framework for the integration of gene-prediction data by dynamic programming. *Genome Research*, *12*(9), 1418–1427. doi:10.1101/gr.149502

Schweikert, G., Zien, A., Zeller, G., Behr, J., Dieterich, C., Cheng, S. O., … Rätsch, G. (2009). mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, *19*(11), 2133–2143. doi:10.1101/gr.090597.108