International Research Journal of Biochemistry and Bioinformatics Vol. 12(6) pp. 1-3, December, 2022
Available online http://www.interesjournals.org/IRJBB

Mini Review

Multiple DNA/RNA Sequence Alignment and Phylogenetic Tree-building Programme called SaAlign for Extremely Huge Datasets and Extremely Lengthy Sequences

Jung Cook*

College of Computer Science, Sichuan University, Chengdu 610064, Sichuan, PR China

*Corresponding Author's E-mail: cookjung@rediff.com

Received: 02-Dec-2022, Manuscript No. IRJBB-22-84547; **Editor assigned:** 05-Dec-2022, PreQC No. IRJBB-22-84547 (PQ); **Reviewed:** 19-Dec-2022, QC No. IRJBB-22-84547; **Revised:** 24-Dec-2022, Manuscript No. IRJBB-22-84547 (R); **Published:** 31-Dec-2022, DOI: 10.14303/2250-9941.2022.37

Abstract

Multiple DNA/RNA sequence alignment is a crucial bioinformatics foundational technique, particularly for the creation of phylogenetic trees. The volume of bioinformatics data is continuously growing as a result of advancements in DNA-sequencing, necessitating the continuous iteration of several tools. Bioinformatics software is needed to analyse the mitochondrial genomes of various people and species, thus its performance has to be improved. We used longest common substring techniques to optimise a dynamic programming solution for the alignment of extremely big datasets and extremely lengthy sequences (Liou et al., 2013). The Multiple DNA/RNA Sequence Alignment Tool Based on Suffix Tree (SaAlign), which aligns sequences of diverse lengths, some exceeding 300 kb (kilobases), was shown to save time and computing space on extremely large test DNA datasets. It performed better than the available technical instruments, such as MAFFT and HAlign-II. MAFFT completed the necessary tasks for mitochondrial genome datasets with a small number of sequences; however it was unable to handle extremely large mitochondrial genome datasets due to core dump errors. In order to maximise the spatial and temporal efficiency, we construct a multiple DNA/RNA sequence alignment tool based on the Center Star technique and apply the suffix array algorithm (Siniscalco et al., 2008). These days, as whole-genome research and NGS technologies gain traction, it is important to save computing resources for labs. This programme is extremely important in these areas, particularly for the study of plants' whole mitochondrial genome (Tzouvelekis et al., 2013).

INTRODUCTION

Similarities between sequences are used to build phylogenetic trees for biological studies during the processing of biological data. The data processing scale has expanded from Mega Byte (MB) and Giga Byte (GB) to Terabyte (TB), PB, and even EB and ZB in recent years due to the dramatic growth in next-generation sequencing findings. For instance, millions of sequence reads are examined in metagenomics investigations to identify the functional or taxonomic composition of microbial samples taken from the environment. Therefore, multi-sequence analysis and phylogenetic tree construction tools must operate at a higher level. There are several software programmes for sequence alignment analysis available online. Sequence alignments may be conducted on a multi-thread workstation and for a particular context using current state-of-the-art programmes like MAFFT, PASTA (>200,000 sequences MSA), ProbPFP (PHMM model refined by particle swarm optimization), and Minimap2 (designed for nanopore sequencing). Comparative genomics, cladistics, bioinformatics, and other areas also make heavy use of a variety of phylogenetic tree building software tools. For Unix-like operating systems, MAFFT is software for multiple sequence alignment. It provides a number of different alignment techniques, such as L-INS-I and FFT NS-2. MAFFT has undergone several modifications and enhancements thanks to contributions from participants all around the

world, and it is currently the most widely used programme for DNA and protein sequence alignments. However, it takes a lot of time to use MAFFT to align large, unrelated sequences. If the spatial complexity of Needleman-Wunsch is O (n2), then those unrelated sequences also demand additional memory when employing Dynamic programming methods. The user may still pick and control the algorithms using MAFFT's terminal-based user interface. Distributed computing frameworks, including Hadoop and Spark, are used by some software packages and are gaining popularity over time. Sequential SparkBWA, which makes use of the BWA technique to increase processing efficiency in the Spark cluster, made multi-node computing possible (Ghaedi et al., 2013). Many faster methods utilising Spark have been proposed for the global alignment of DNA sequences and the construction of phylogenetic trees, such as HAlign and HAlign-II, which can effectively build phylogenetic trees from numerous biological sequences and provide user-friendly web servers through high-performance and distributed computing infrastructures. These quicker methods do, however, have significant drawbacks.

Pairwise alignment of DNA sequences using the Needleman-Wunsch method demonstrates considerable spatial complexity, which precludes whole-genome comparisons in favour of just satellite RNA, viral DNA, and genetic segments for eukaryotes. Even if genome-wide comparisons were possible, they would need extremely memory-intensive workstations. Additionally, the code is written in Java, which is slower in performing sequence comparisons than software written in C (Huang et al., 2014).

DISCUSSION

In this instance, we created an MSA tool based on the centre star approach. We employed a suffix array to speed up the MSA of very near DNA sequences. Although it worked well, it couldn't manage scalable data or genomes. It functioned in tandem with Spark, a free parallel programming framework. Studies have demonstrated that SaAlign can handle more than 30,000 kb of sequencing data. In contrast to other studies, the experimental datasets employed here had a rising number of sequences since non-repeating DNA sequences were added instead of the initial sequence set being repeated (Williams 2003). The main objective of our work was to improve the MSA's effectiveness and capacity for handling massive volumes of data. As a result, we focused on cutting down the execution time for massive data. When the number of sequences was not very high, MAFFT and HAlign-II worked well for the alignment of short DNA sequences (fungal ITS). Additionally, additional running time was saved as the number of nodes grew. Due to the improvement of the algorithm based on the LCS method of the suffix array used in double-sequence comparisons, SaAlign was more efficient at processing lengthy sequences than HAlign-II (Dahlin et al., 2004). Additionally, managing massive amounts of files is made easier by leveraging the Spark distributed architecture. SaAlign can now evaluate extremely huge datasets with very lengthy sequences with greater time. Additionally, we assessed the accuracy using SPS. Additionally, phylogenetic trees with bootstrap values over 70% also showed good accuracy. We calculated the SaAlign acceleration ratio and looked into the effects of sequence length and number (Kreda et al., 2001). The acceleration ratio was high when the average sequence length and node length of the sequence dataset were both low. As the number of nodes rose, the acceleration ratio of the sequence dataset with a large number of sequences also increased. When the estimation nodes were the same and the acceleration ratio was large, the series length was longer. Amdal's law states that acceleration will not be optimal if a section of the programme cannot be parallelized. The transmission overhead time was reduced for the dataset with fewer sequences because it produced a shorter centre sequence with the same average sequence length. Additionally, compared to other datasets with two nodes, the dataset with fewer sequences benefited more from acceleration. With more nodes and a high level of parallelization, the anticipated acceleration became comparably noticeable (DeMaio et al., 2009). Overhead scheduling has a decreasing impact on the average operating time. As a result, there was a bigger acceleration ratio for sequence datasets with large numbers of sequences when the number of nodes was high. For datasets with the same number of sequences, the time saved with lengthy sequences was much larger than with central sequence transmission and scheduling. The short sequence dataset's acceleration ratio was significantly affected by the centre sequence's propagation and scheduling time. Additionally, as the number of nodes expanded, the acceleration ratio dropped, even to less than 1. For sequences with an average length of 500 bp, clustering computations had minimal effect on synchronisation and did nothing to increase calculation efficiency. Even while MSA algorithms employ phylogenetic trees as guidance, phylogenetic tree construction techniques frequently need MSA findings as input results (Newman et al., 1999). Large unaligned DNA sequences have been the subject of several free MSA phylogenetic tree building techniques, many of which are based on LCSs. The efficiency of the sequence alignment was significantly increased by using the suffix array to calculate the LCS. Since protein sequences are frequently short, the detected LCSs have no effect on performance. However, it is possible to find reasonably lengthy LCSs for the analysis of metagenomic sequences, which would enhance computing performance. Protein sequences wouldn't benefit as much from this technique as DNA sequences would.

REFERENCES

1. Liou TG, Raman SM, Cahill BC (2013). Lung transplantation for chronic obstructive pulmonary disease. Transplant Res Risk Manage. 5: 1-20.

2. Siniscalco D, Sullo N, Maione S, Rossi F (2008). Stem cell therapy:

the great promise in lung disease. Therapeutic Advances in Respiratory Disease. 2: 173-177.

3. Tzouvelekis A, Laurent G, Bouros D (2013). Stem cell therapy in chronic obstructive pulmonary disease. Seeking the Prometheus effect. Current Drug Targets. 14: 246-252.
4. Ghaedi M, Calle EA, Mendez JJ (2013). Human iPS cell-derived alveolar epithelium repopulates lung extracellular matrix. The J of Clinical Investig. 123: 4950-4962.
5. Huang SXL, Islam MN (2014). Efficient generation of lung and airway epithelial cells from human pluripotent stem cells. Nat Biotechnol. 32: 84-91.
6. Williams MC (2003). Alveolar type I cells: molecular phenotype and development. Annu Rev Physiol. 65: 669-695.
7. Dahlin K, Mager EM, Allen L (2004). Identification of genes differentially expressed in rat alveolar type I cells. Am J Respir Cell Mol Bio. 31: 309-316.
8. Kreda SM, Gynn MC, Fenstermacher DA, Boucher RC, Gabriel SE (2001). Expression and localization of epithelial aquaporins in the adult human lung. Am J Respi Cell Mol Bio. 24: 224-234.
9. DeMaio L, Tseng W, Balverde Z (2009). Characterization of mouse alveolar epithelial cell monolayers. Am J Physi Lung Cell Mole Physio. 296: 1051-1058.
10. Newman GR, Campbell L, Von Ruhland C, Jasani B, Gumbleton M (1999). Caveolin and its cellular and subcellular immunolocalisation in lung alveolar epithelium: implications for alveolar epithelial type I cell function. Cell Tissue Res. 295: 111-120.