Massive Multiple Sequence Alignment of 16S Bacterial Ribosomal RNAs Using ClustalW-Message Passing Interface (MPI) Based on Beowulf Linux System

Hyon Chang Kim[1], Yong Beom Seo[2], Ji Hwan Song[2], Dong Soon Choi[3], Churl K. Min[3], Han Jip Kim[1*]

[1]Department of Biological Sciences, [2]Division of Information & Computer Engineering, [3]Department of Molecular Science & Technology, Ajou University

Suwon, South Korea

**Corresponding author: Hanjip Kim (genetics@ajou.ac.kr)*

Abstract

We have built a Debian-Beowulf computer cluster consisting of 15 computational nodes, each equipped with 15 AMD Opteron 64 bit microprocessors. Local Area Multicomputer (LAM) – Message Passing Interface (MPI) was used as a portable high-performance implementation for MPI. More than 2,000 bacterial 16S ribosomal RNAs (rRNAs) were multiply aligned using ClustalW-MPI. Systematic sequence comparison provided several sequences with a very high degree of homology despite their different origins of species. These highly conservative sequences were collected as candidate sequences for drug targets of ribosomal antibiotics.

Key word: multiple sequence alignment, ClustalW-MPI, 16S ribosomal RNA, ribosomal antibiotics

1. Introduction

Gleaning meaningful information from the enormous amounts of biological data is one of the most important tasks in biology today. Currently, several web-based interfaces offer services to help researchers find and mine useful data of their own. However, these interfaces suffer from limited usage especially when the amounts of data become large. For instance, the input for ClustalW web services from the European Bioinformatics Institute is limited to a maximum of 500 sequences or to a 10MB file. This restraint reduces researchers' options for their experiments.

Due to the advances in computer technology, it is now possible to construct high-performance computers at a relatively low cost. We have built a Debian-Beowulf computer cluster consisting of 15 computational nodes, each equipped with 15 AMD Opteron 64 bit microprocessors. One master node was connected to 14 slave nodes with a gigabit Ethernet switch. LAM – MPI was used as a portable high-performance implementation for MPI and more than 2,000 bacterial 16S rRNAs were multiply aligned using ClustalW-MPI.

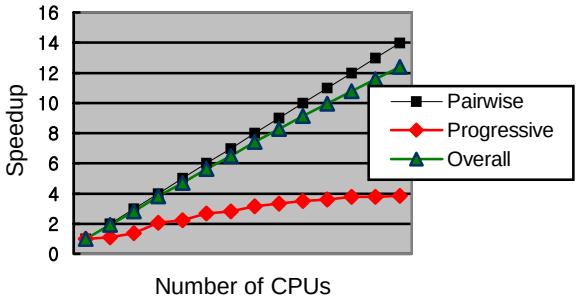
These rRNA genes are highly repeated multiple copies, whereas most proteins in bacteria are encoded in single-copy genes [1]. Due to the fact that conventional antibiotics are vulnerable to a point mutation which may critically affect the drug targets of bacteria, ribosomal antibiotics interacting with rRNA are much less affected by a point mutation which occurs every 10^7 ~ 10^8 generations. Because the frequency of the same resistance mutation in several copies of rRNA genes becomes extremely low [1]. This makes ribosomal antibiotics a highly efficient weapon since it is generally very difficult to find defenses against them [1]. By utilizing computer resources and biological data, we were able to gather highly conserved ribosomal sequences from more than 2,000 different strains. These unique but well conserved bacterial ribosomal sequences could very well enable us to develop new antibiotics targeting 16S rRNA.

2. Methods

The Beowulf [2] system was selected in order to obtain high-performance computing ability for massive multiple alignments while maintaining a low cost. The system contains relatively inexpensive AMD Opteron processors that are able to handle numerical operations

such as floating point computation at high speeds. Expenses were greatly reduced since computational nodes are conducted by diskless booting, and therefore, do not require any I/O installation such as hard disks or video cards. We used a gigabit switch to maximize the speed of the input and output of data. The LAM Version 7.1.1 (stable) [3] was downloaded and was installed to provide MPI for parallel processors. The system performance test was carried out using the High-Performance Linpack (HPL) benchmark [4] with the GOTO libraries [5].

ClustalW-MPI, a distributed and parallel implementation of ClustalW was then installed. The speed-up test was performed using 100 DNA sequences. The first step, which is the pairwise alignment of ClustalW, involves calculating a distance matrix between each pair of sequences and all elements of the distance matrix are independent [6]. Therefore, efficient parallelization was achieved as the number of CPU processors increased as shown in Figure 1. The second step is a guide-tree generation and decides the topology of the progressive alignment [6]. This step is not shown because the effective speedup was not observed due to the data dependency problem. The third step, which is a progressive step, is less affected by the parallel process of ClustalW-MPI compared to the pairwise alignment. The overall speedup increases almost proportionally to the number of CPUs, showing that the pairwise alignment takes up



most of the ClustalW-MPI process.

Figure 1. Speedups for the ClustalW-MPI results of the 100-sequence data

The 16S rRNA sequences from 2,128 bacterial strains were collected from Genbank. Several sequence segments showed very high homology despite their different origins. We selected these highly conservative sequences, and compared them with human rRNA sequences to eliminate any possible interactions between ribosomal antibiotics and the human rRNAs. The remaining unique bacterial origin sequences are now undergoing test in vivo, to determine the best targets for ribosomal antibiotics.

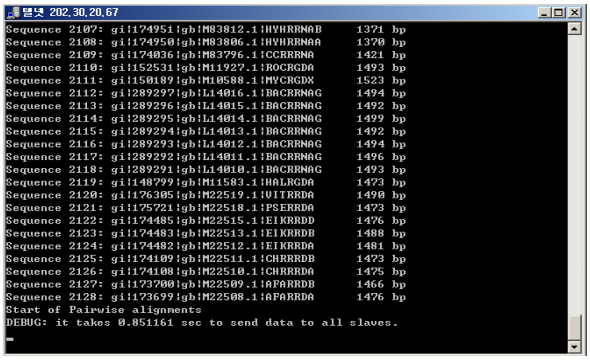


Figure 2. Multiple sequence alignment of 2128 bacterial rRNAs

3. References

[1] A. S. Mankin, “Ribosomal Antibiotics”, Molecular Biology, 35, 509-520, 2001

[2] Becker, Donald J., Thomas Sterling, Daniel Savarse, John E. Dorband, Udaya A. Ranawak, Charles V. Packer, “Beowulf: A Parallel Workstation For Scientific Computation”, Proceeding, International Conference on parallel Processing, 1995

[3] LAM/MPI Parallel Computing// http://www.lam-mpi.org/7.1/download.php

[4] A. Petitet, R. C. Whaley, J. Dongarra, A. Cleary HPL-A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers// http://netserv2.chg.ru/pub/prog/netlib/benchmark/hpl/index

[5] High-Performance BLAS by Kazushige Goto// http://www.cs.utexas.edu/users/flame/goto/

[6] Kuo-Bin Li, “ClustalW-MPI: ClustalW analysis using distributed and parallel computing”, Bioinformatics Application Note 19, 1585-1586, 2003