APPLICATION NOTE **Genomics & Informatics** Vol. 6(3) 157-159, September 2008

A Simple and Fast Web Alignment Tool for Large Amount of Sequence Data

Yong Seok Lee*,¶ and Jeongsu Oh¶

Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea

Abstract

Multiple sequence alignment (MSA) is the most important step for many of biological sequence analyses, homology search, and protein structural assignments. However, large amount of data make biologists difficult to perform MSA analyses and it requires much computational time to align many sequences. Here, we have developed a simple and fast web alignment tool for aligning, editing, and visualizing large amount of sequence data. We used a cluster server installed ClustalW-MPI using web services and message passing interface (MPI). It also enables users to edit multiple sequence alignments for manual editing and to download the input data and results such as alignments and phylogenetic tree.

Availability: A web services is freely available at http://www.koreanbio.org/ClustalwMPI/.

Keywords: alignment, ClustalW-MPI, large data, Jalview, parallel process

Introduction

The multiple alignment of biological sequences has a long history and is used to identify regions of homology that shows functional, structural, or evolutionary relationships between the sequences. Many multiple sequence alignment (MSA) programs have been development by bioinformatists in the past. ClustalW (Thompson *et al*., 1994) is perhaps the most widely used MSA program due to its speed. Also, as a single standalone algorithm, it is easy to run in a computer especially in UNIX and Linux environments. Recently, massively large sequence data are generated due to the technical advancement of sequencing. For aligning large number of sequences, several parallel alignment programs have been reported (Ebedes and Datta, 2004; Kleinjung *et al*., 2002; Li 2003; Mikhailov *et al*., 2001). Kleinjumg *et al*. (2002) is based on the so-called a parallel progressive alignment strategy without a guide tree. Mikhailov *et al*. (1993) designed a method for shared-memory multiprocessor machines. Usually, these machines have commodity parallel architectures and are expensive. While Ebedes and Datta (2004) and Li (2003) reported a parallelization of the ClustalW on a workstation cluster using the message passing interface (MPI; MPI site, 2008). It has a smaller network bandwidth and longer message latency. However, single or multiple CPU ClustalW programs are usually not suitable for most biologists who are not used to UNIX systems.

Web based applications have many advantages in that they can reach large number of users. There is no need to install programs, and it is easier to update and manage than to maintain desktop applications (Shin *et al*., 2008; Woo *et al*., 2007). However, some web applications are dependent on specific web browsers, so they are incompatible with some web browsers. When a web alignment tool becomes more useful to operate if it is implemented following the W3C recommended XHTML 1.0 Transitional DTD standard and does not use technologies that are dependent on specific web browsers. This is one way to make a web alignment tool more compatible with many web browsers.

We have developed a user-friendly a web based alignment tool based on ClustalW-MPI program. It is standard and easy to maintain. This web tool will help researchers to carry out multiple sequence alignment with a large number of input accompanies by a viewer and an editing function. It also enables users to download the results and do basic analyses such as building trees and sequence clustering.

Features and Results

In order to use alignment tools, most advanced users use UNIX or Linux commands and options directly in a console window. It is very inconvenient to use. It also can cause frequent mistakes. A web alignment tool can be executed through a GUI environment on the web page by selecting commands and options. Our web alignment tool has the following features; input, downloadable output, and visualization.

Users input multiple sequences in the web alignment

¶Yong Seok Lee and Jeongsu Oh contributed equally to this work.

*Corresponding author: E-mail dolsemtl@kribb.re.kr
Tel +82-42-879-8511, Fax +82-42-879-8519

Accepted 9 September 2008

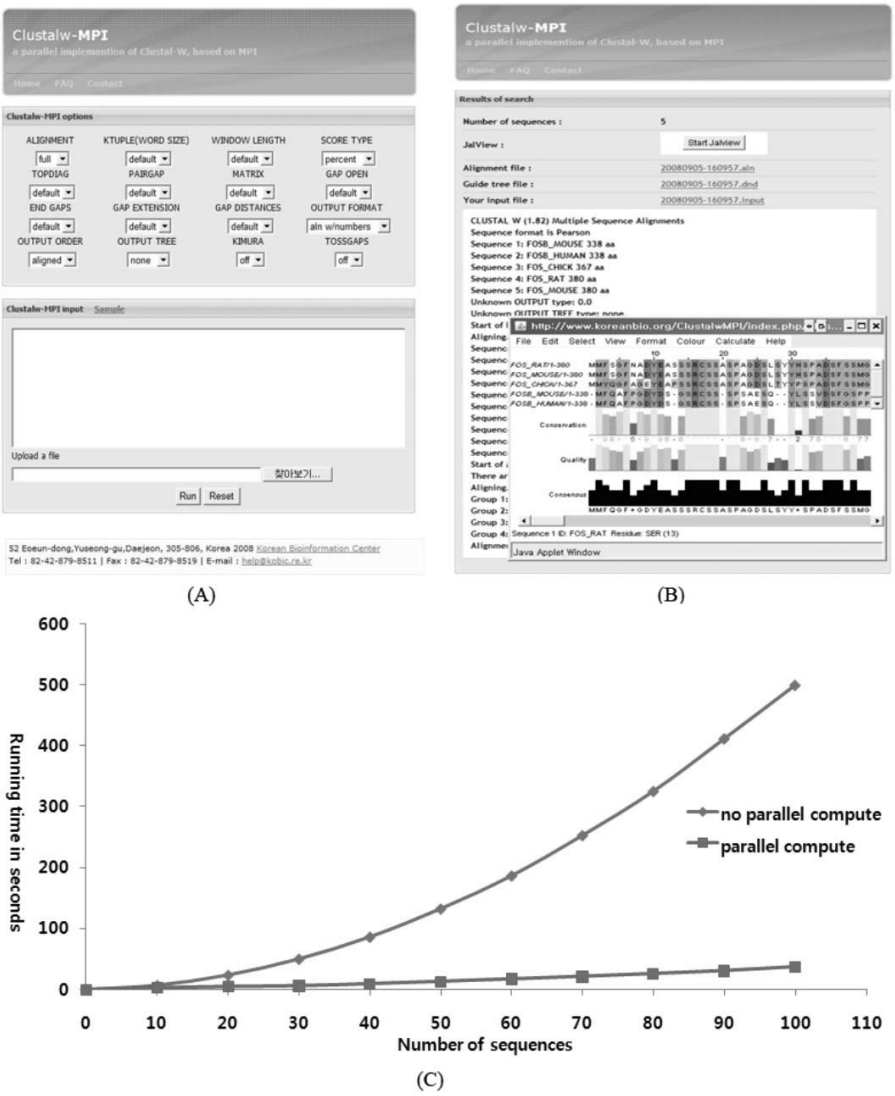

Fig. 1. (A) An example main input page. Users can paste in or upload sequences to carry out the alignment program. (B) An example main output page. (C) Comparison of running time and speed-ups between parallel and non-parallel systems. We ran the alignment program with mitochondrial COX1 gene on the ten widely varying test data with 1537 bases. X-axis indicates the number of input sequences. Y-axis indicates the running time in seconds. We showed the results for two running time comparison with a parallel system against a non-parallel of ClustalW-MPI.

tool first (Fig. 1A). It supports two types of input data: 1) copy & paste and 2) file uploading. Users set the options as for ClustalW in the option panel. The web alignment tool provides all the options for ClustalW.

By clicking the run button to activate ClustalW-MPI after users finish inputting the sequence and setting the options, the users will see a wait dialog window. At this point, the web server sends the input sequence data and options that are set by the users to a CPU cluster that is installed with ClustalW-MPI. The CPU cluster machine receives request from the web part to execute ClustalW-MPI, and then returns the result to the web server. All the communication between the web server and the CPU cluster machine is through the web services technology. Web alignment tool offers three downloadable types of the outputs after finishing each job: Alignment file, Guide tree file, and Input file (Fig. 1B). In the web alignment tool, users can download files easily through the web page. The web alignment tool shows the same output printed in consol screen.

Web alignment tool also offers a Jalview applet for editing and visualizing the outputs on the result web page by clicking start Jalview button on the result web page (Fig. 1B). Jalview is widely used multiple sequence alignment editor in the world. This applet enables users to edit, view, and perform a basic analysis such as drawing a tree and the removal of sequences, exporting the results, and so on.

System and Implementation

The web alignment tool is divided into two parts: 1) web part and 2) ClustalW-MPI part. The web part is constructed by CodeIgniter php framework (CodeIgniter

site, 2008) as a web server and web services client. It runs on the Linux operating system. The web part sends the request to a CPU cluster server installed with ClustalW-MPI, upon receiving a request from a user on the web browser. The ClustalW-MPI part is constructed by java 6.0 and xfire SOAP framework (XFIRE site, 2008) as web-services to run ClustalW-MPI. It runs on a cluster machine installed with rocks-cluster which has 20 nodes (Rocksclusters site, 2008) in KOBIC (Korean Bioinformation Center). When the ClustalW-MPI part is invoked from the web part, it executes ClustalW-MPI, and then returns the result to the web part.

Performance of Web Alignment Tool

Our test was run with the mitochondrial COX1 gene. We showed the result in Figure 1. Ten widely varying inputs from ten to 100 sequences are shown in X-axis. Y-axis indicates the running time of an individual input. One input was run in a common web alignment tool without a parallel CPU system. The other was run in our web alignment tool in a parallel CPU system using web services and an MPI system. We found that the running time increased exponentially in a single CPU system, while the running time of all the varying size inputs was below than 100 seconds in the parallel CPU system that runs ClustalW-MPI.

Acknowledgements

This research was supported by a grant from the Korean Research Institute of Bioscience and Biotechnology (KRIBB) Research Initiative Program.

Reference

Ebedes, J., and Datta, A. (2004). Multiple sequence alignment in parallel on a workstation cluster. *Bioinformatics* 20, 1193-1195.

Kleinjung, J., Douglas, N., and Heringa, J. (2002). Parallelized multiple alignment. *Bioinformatics* 18, 1270-1271.

Li, K.B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics* 19, 1585-1586.

Mikhailov, D., Cofer, H., and Gomperts, R. (2001). Performance optimization of ClustalW: parallel ClustalW, HT Clustal, and MULTICLUSTAL. *White papers*, Silicon Graphics, Mountain View, CA.

Shin, J., Park, H., Ahn, Y., Cho, D., Kim, J., Kee, M., Kim, S., Lee, J., and Kim, S. (2008). GTVseq: a web-based genotyping tool for viral sequences. *Genomics & Informatics* 6, 54-56.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22, 4673-4680.

Woo, T., Kim, Y., Kwon, J., and Seo, J. (2007). RepWeb: a web-based search tool for repeat-related literatures. *Genomics & Informatics* 5, 89-91.

Websites

CodeIgniter (2008). *http://www.coldscripts.com/*
XFIRE site. (2008). *http://xfire.codehaus.org/*
Rocksclusters site. (2008). *http://www.rocksclusters.org*
MPI site. (2008). *http://www-unix.mcs.anl.gov/mpi/*