SinicView: An Interactive Visualization Tool for Comparison of Multiple Sequence Alignment Results

Arthur Chun-Chieh Shih, D.T. Lee*, Laurent Lin, Chin-Lin Peng
Shiang-Heng Chen, Chun-Yi Wong, Meng-Yuan Chou, and Tze-Chang Shiao
Institute of Information Science and Genomics Research Center, Academia Sinica, Taiwan

Abstract

In the initial stage of sequence analysis, a biologist is first faced with the questions about how to choose the best tool to align sequences of interest and how to analyze and visualize the alignment results, and then with the question about whether unaligned regions produced by the tool are indeed not homologous or are just results due to inappropriate alignment tools or scoring systems used. In this paper, we present a versatile alignment visualization system, called SinicView, (for Sequence-aligning INnovative and Interactive Comparison VIEWer), which allows users to efficiently compare and evaluate assorted alignment results obtained by different tools. SinicView calculates similarity of the alignment outputs under a sliding window using the sum-of-pairs method and provides scoring profiles of each set of aligned sequences. Combined with the annotations information, the user can visually compare alignment results either in graphic scoring profiles or in plain text format of the aligned nucleotides. With SinicView, users can use their own data sequences to compare various alignment tools or scoring systems and select the most suitable one to perform alignment and sequence analysis. SinicView is avaiable for free download from http://biocomp.iis.sinica.edu/.

1. Introduction

Study of comparative genomics demonstrates its power to help biologists identify novel conserved and functional regions in genomes[3]. Based on the comparison of cross-species genomic sequences, biologists can understand the evolutionary relationship of genomic regions among species, discover conserved regions between different genomes, vertebrate genomes, discover regulatory motifs and promoters, or identify potential conserved non-genic sequences (CNGs).

*To whom correspondence should be addressed. E-mail:dtlee@iis.sinica.edu.tw.

To align these genomic sequences, several efficient tools have been proposed, such as MLAGAN[2], MAVID [1], and so on. However, the majority of these tools usually fail to generate consistent results especially in aligning divergent cross-species sequences. Therefore, comparisons of the alignment tools using a set of benchmarking sequences have been conducted in recent years.

Although these comparison results provide a fair evaluation of several popular alignment tools, the user usually does not know for sure whether those unaligned regions are indeed non-homologous or just due to inappropriate alignment tools or scoring systems used. Thus, the user may have to try different tools or scoring systems to evaluate the correctness and accuracy of alignment results in the initial stage of sequence analysis. Thus, it is desirable and most useful to have a visualization system that provides a direct and efficient method and can assist users to cross compare and inspect alignment results obtained by different multiple sequence alignment (MSA) tools, especially in the initial stage of sequence analysis.

In recent years a number of visualization tools have been released in the public domain. The VISTA-related tools, including mVISTA, rVISTA, GenomeVISTA, and PhyloVISTA, are among the famous ones that provide users with novel graphical user interfaces to view alignment results from different viewpoints have been developed for several years (http://genome.lbl.gov/vista/index.shtml). PipMaker, zPicture, and ECR Browser are also popular visualization tools for sequence or genomes alignment results. All of these tools are web-based with friendly user interfaces, and allow users to easily visualize alignment results with annotations. However, these tools are limited solely to single alignment results. The capability of simultaneously comparing multiple results from different alignment tools or different scoring systems is notably lacking.

2. Result

We present a versatile alignment visualization system, SinicView (Sequence-aligning INnovative and Interactive

Comparison VIEWer), which enables users to efficiently compare and evaluate assorted alignment results obtained by different tools. SinicView calculates similarity of the alignment outputs under a sliding window using the sum-of-pairs method and provides scoring profiles of each set of aligned sequences. Users can visually compare alignment results either in graphic scoring profiles or in plain text format of the aligned nucleotides. In addition, the information about alignment gaps and sequence annotations is also presented. The real-time juxtaposition of the visualization results from different MSA programs would bring more insights into the evaluation process. With SinicView, users can use their own sequences to survey and compare various multiple alignment tools and thus to unveil their merits (and shortcomings). Moreover, the cross-tools comparison can provide users more confidence in their final alignment results especially for those unaligned regions.

There are three viewing sections in SinicView: Global View, Detailed View, and Information View (including annotations and gaps.) The Global View section shows the whole percent identity plots that calculate the sum-of-pairs scores based on one specified reference sequence. In the Detailed View section, the panels show the whole percent identity plots of different alignment results individually. The Information View section containing annotation and gap information is stacked beneath the Detailed View section. SinicView also provides some global comparison charts that can assist biologists to choose the best alignment result among those produced by the programs under consideration. Each of the aligners is denoted by a pre-defined color with the "performance color" label right next to the name of the tool. SinicView is implemented entirely in Java language to ensure portability across major platforms and is accessible with a web browser and Internet connection.

2.1. An example

SinicView offers a series of manipulative and navigational controls, such as zooming, shifting, and gap/annotation toggling. As shown in Fig. 1, SinicView displays the alignment results obtained by three different MSA methods: ClustalW, MAVID, and MLAGAN. The input sequences contain orthologous regions around the SCL gene in five vertebrate species: human, mouse, chicken, pufferfish and zebrafish, and the human sequence is selected as the reference base. Users can manually input numerical values or click on the highlighted colored region in the Global View section that specify the zooming or shifting factors in a drag-and-drop fashion. Generally speaking, the highest conserved region located at 30kp of human sequence is all well aligned by these three tools. But the highest identical rates of the alignment by ClustalW are lower than those by either MLAGAN or MAVID.

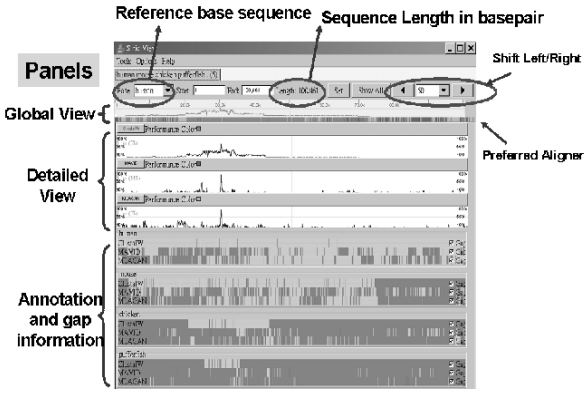

Figure 1. The screenshot shows the user interface of SinicView.

3. Conclusions

Deluged by increasing completed genomic sequences, biologists have encountered a challenge of aligning more and much longer sequences from divergent species. Thus, the need to align longer sequences, like mega base-pair sequences or even genomics-scale sequences, and evaluate the alignment results becomes more urgent. In this paper, we have presented a visualization tool for comparison of MSA programs. With a standard simple protocol for the input/output format, users can easily upload their own alignment programs to SinicView. The performance and capabilities of SinicView depend on the system's internal memory. In a 64M RAM JAVA environment, SinicView can load and visualize several mega bases alignment results. Users can easily perform sequence alignment by employing multiple alignment tools and visualize the results on the fly by SinicView. For a more detailed description of SinicView the reader is referred to http://biocomp.iis.sinica.edu.tw.

This work was supported by the National Science Council of Taiwan under the grants No. NSC-92-3112-B-001-018-Y, NSC-92-3112-B-001-021-Y, NSC-93-3112-B-001-018-Y, and NSC93-3112-B-001-023-Y.

References

[1] N. Bray and L. Pachter. Mavid: constrained ancestral alignment of multiple sequences. *Genome Res*, 14:693–9, 2004.

[2] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Res*, 13:721–31, 2003.

[3] W. Miller, K. D. Makova, A. Nekrutenko, and R. C. Hardison. Comparative genomics. *Annu Rev Genomics Hum Genet*, 5:15–56, 2004.