中央研究院
資訊科學研究所
Institute of Information Science, Academia Sinica • Taipei, Taiwan, ROC

TR-IIS-05-005

SinicView: A Visualization Environment for Comparisons of Multiple Sequence Alignment Tools

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

June 2005 || Technical Report No. TR-IIS-05-005

http://www.iis.sinica.edu.tw/LIB/TechRept.htm

SinicView: A Visualization Environment for Comparisons of Multiple Sequence Alignment Tools

Arthur Chun-Chieh Shih[1], D.T. Lee[*1,2], Laurent Lin[1], Chin-Lin Peng[2]
Shiang-Heng Chen[1], Chun-Yi Wong[1], Meng-Yuan Chou[1], and Tze-Chang Shiao[1]

[1]Institute of Information Science and [2]Genomics Research Center
Academia Sinica, Taipei, 115, Taiwan

Running Head

SinicView: A tool for Alignments Comparison

Keywords

Comparative genomics, multiple sequence alignment, sequence alignment visualization, web-based collaborative environment

*To whom correspondence should be addressed.

Abstract

Motivation:

Deluged by completed genomic sequences, the need to align longer sequences becomes more urgent, and many more tools have thus become available. In the initial stage of sequence analysis, a biologist usually faces 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. Although several systematic evaluations of multiple sequence alignment programs have been proposed, they may not provide a standard-bearer for most biologists because those unaligned regions in these evaluations are never discussed. Thus, a tool that allows cross comparison of the alignment results obtained by different tools could help a biologist evaluate their correctness and accuracy.

Results:

In this paper, we present a versatile alignment visualization system, called SinicView, (for Sequence-aligning INnovative and Interactive Comparison VIEWer), which allows the user 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. The user can visually compare alignment results either in graphic scoring profiles or in plain text format of the aligned nucleotides along with the annotations information. 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. We illustrate the capabilities of our visualization system by comparing alignment results obtained by ClustalW, MLAGAN, MAVID, and MULTIZ, respectively.

Contact: dtlee@iis.sinica.edu.tw

Availability: http://biocomp.iis.sinica.edu.tw/

INTRODUCTION

With exponentially increasing genomic sequences available in the public domain (Gibbs, Weinstock, et al., 2004; Hillier, Miller, et al., 2004; Lander, Linton, et al., 2001; Venter, Adams, et al., 2001; Waterston, Lindblad-Toh, et al., 2002), study of comparative genomics demonstrates its power to help biologists identify novel conserved and functional regions in genomes (Dubchak and Frazer, 2003; Frazer, Elnitski, et al., 2003; Heilig, Eckenberg, et al., 2003; Miller, Makova, et al., 2004). 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, such as yeast species genomes (Kellis, Patterson, et al., 2003), metazoan genomes (Ureta-Vidal, Ettwiller, et al., 2003), vertebrate genomes (Thomas, Touchman, et al., 2003), and mammalian genomes (Brudno, Poliakov, et al., 2004), discover regulatory motifs in the yeast (Cliften, Sudarsanam, et al., 2003) and human promoters (Xie, Lu, et al., 2005) or identify potential conserved non-genic sequences (CNGs) (Dermitzakis, Reymond, et al., 2003).

However, genomic sequences can be megabase long and thus the traditional sequence alignment tools based on dynamic programming would not work efficiently due to their time and space complexities, which are of the order of the product of the lengths of the input sequences. To better tackle this problem, several tools for genomic sequence alignment have been proposed, such as pairwise sequence aligners like MUMmer (Delcher, Kasif, et al., 1999), GS-Aligner (Shih and Li, 2003), Avid (Bray, Dubchak, et al., 2003) and LAGAN (Brudno, Do, et al., 2003), and multiple sequence alignment (MSA) programs like MultiPipMaker (Schwartz, Elnitski, et al., 2003), MULTIZ (Blanchette, Kent, et al., 2004), MLAGAN (Brudno, Do, et al., 2003), MAVID (Bray and Pachter, 2004), and MUSCLE (Edgar, 2004; Edgar, 2004). These alignment tools, however, are heuristics based and do not provide any indication of how far they are from an optimal solution. The majority of these tools usually fail to generate consistent results especially in aligning divergent cross-species sequences. Consequently the more alignment tools there are available in the public domain, the more confusion it creates for users to decide which tool is most suitable to align their sequences. Therefore, comparisons of the alignment tools using a set of benchmarking sequences have been conducted in recent years (Karplus and Hu, 2001; Pollard, Bergman, et al., 2004; Raghava, Searle, et al., 2003).

Although these comparison results provide a fair evaluation of several popular alignment tools, the conclusions may not be directly applicable to users' sequences. Furthermore 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. Consequentially, if some homologous regions are unaligned, the estimated evolution distances of these sequences may be inaccurate and therefore the constructed phylogenetic trees may be incorrect. Facing this problem, 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. On the other hand, new alignment tools are released continually. Users may want to compare these newly released tools with those that they are most familiar with. 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 MSA tools especially at 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 are among the famous ones that have been developed for several years (http://genome.lbl.gov/vista/index.shtml). mVISTA is a set of programs for comparing DNA sequences from two or more species up to

megabases long and visualize these alignments with annotation information (http://genome.lbl.gov/vista/mvista/submit.shtml). rVISTA (regulatory Vista) combines database searches for transcription factor binding sites with a comparative sequence analysis (Loots and Ovcharenko, 2004; Loots, Ovcharenko, et al., 2002). GenomeVISTA compares users' sequences with several whole genome assemblies (Couronne, Poliakov, et al., 2003; Frazer, Pachter, et al., 2004). Phylo-VISTA analyzes multiple DNA sequence alignments of sequences from different species while considering their phylogenic relationships (Shah, Couronne, et al., 2004). In general, the VISTA family of tools provides users with a novel graphical user interface (GUI) to view alignment results from different viewpoints. In addition to the VISTA family, PipMaker (Schwartz, Elnitski, et al., 2003; Schwartz, Zhang, et al., 2000), zPicture (Ovcharenko, Loots, et al., 2004), and ECR Browser (Ovcharenko, Nobrega, et al., 2004) 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.

In this article, 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.

METHOD

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. By observing the graphical results, it is much more intuitive and straightforward to judge the consistency of the alignment results. When the sliding window is less than 100 base

pairs, the Detailed View section will switch from the curve-based plot to the display of the detailed alignments in a colored text format. The Information View section containing annotation and gap information is stacked beneath the Detailed View section. SinicView also provides several global comparison charts that can assist biologists to choose the best alignment result among those produced by the programs under consideration. SinicView is implemented entirely in Java language to ensure portability across major platforms and is accessible with a web browser and Internet connection. The main features of SinicView are summarized as follows:

1. Visualization of the similarity distribution of alignment results in a curve-based graphic format;
2. Generation of the comparison stacked-bar and pie charts, which shows the distribution of the identical rates among various alignment programs for benchmarking purposes;
3. Inclusion of a versatile manipulative functionality (gap-display toggling, drag-and-drop zooming/shifting, and graph/text display toggling);
4. Visualization of annotation information and display of the phylogenetic tree provided by users;
5. Visualization of detailed text alignments results;
6. Capability to export the visualization results to portable image files.

In what follows, we will introduce the characteristics and functionality of SinicView in more detail.

Manipulative Operations in SinicView

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. The input sequences contain orthologous regions around the Stem Cell Leukemia (SCL) gene in five vertebrate species: human, mouse, chicken, pufferfish and zebrafish. The buttons and text-field boxes of manipulative functions are located on top of the frame. 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. When the highlighted region is clicked and dragged, the equivalent of a shift action will be performed and the display region can be resized by adjusting the edge of the highlighted area.

SinicView can display more than one alignment result obtained by different alignment programs (either pairwise or multiple ones.) The assorted mixed-color span under the Global View panel shows among the alignment tools used the preferred aligner, which generates comparatively better results on the spot. Each of the aligners is denoted by a pre-defined color with the “performance color” label right next to the name of the tool.

Multi-panel Functionality in SinicView

In the Detailed View section, the Percent Identity Plot (PIP) panels show, from top to bottom, the similarity curves of the alignment results obtained by different programs, along with the names of the alignment tools. In the Information View section, the Gap & Annotation panels (in pink and gray) display the information of annotations, which is provided by users, and gaps of aligned sequences. The information and similarity ratios can also be displayed as the current scan-line (i.e. cursor) moves. The boxes in maroon denote the annotation area and the horizontal line represents the original sequences interleaved with inserted gaps (light gray areas.) The gap display can be toggled on or off via the checkbox on the right.

Because different alignment results are usually of different lengths, it is not plausible to compare these results base-pair by base-pair. In SinicView, therefore, we let users select one of input sequences as a *reference* and then calculate the sum-of-pair scores of each base pair in the reference within a fixed window. For example, each alignment result in the PIP panels at the scan-line position corresponds to human sequence, selected as the reference in Fig. 1. When the user selects different sequences as the reference, SinicView can demonstrate the variations between the PIP curves of the alignment results, for example, mouse in Fig. 2(a) and chicken in Fig. 2(b).

Comparison Chart Capability

The “Comparison Charts” function under the “Tools” menu provides two types of figures for quick and easy comparison of alignment results in statistics. The stacked bar chart, in Fig. 3, illustrates the distribution of the identical rates with the threshold over 40%. The pie chart displays the distribution of the identical rates from 0 to 100 percent based upon a selected alignment program. The statistics on which these charts are based can also be displayed in a tabulated text form.

Text-mode comparison

SinicView also provides a plain-text view of the alignment results in the Detailed View section when the sliding window size is less than 100 aligned base pairs. As shown in Fig. 4, the plain-text alignment results replace the percent identity curves and the fully identical bases in a column are labeled in red blocks. Thus, users can check the correctness of detailed alignment results base pair by base pair.

Installation and execution of the standalone SinicView

The applet version can be accessed via any Internet-enabled browsers with Java Runtime Environment (JRE), including Java virtual machine, thus, saving the hassle of installation and choosing the right platform. However, the ease of running SinicView on-the-go cannot accommodate the bandwidth requirement in case of huge

amount of sequence data involved. Hence, we also design a standalone version of SinicView, which is wrapped in JRE and can be executed offline.

The installation procedure of the standalone SinicView is quite straightforward. After the execution of the Setup.exe file, the main program will be installed. Upon launching the standalone SinicView, the user will be prompted three options. The first two are to read user's Phylogenetic Tree files and MSA results from the local disk. The source can be from multiple files and these data are optional.

RESULTS

In what follows, we will introduce three different examples to demonstrate how SinicView can assist users to analyze alignment results in the initial stage of sequence comparison. The total alignment lengths in the examples 1, 2, and 3 are several thousands, hundreds of thousands, and millions of base pairs, respectively. The conservations of the aligned sequences are also different in each example.

Example 1: OPN1LW

Opsins are membrane proteins related to the protein moiety of the photoreceptive molecule rhodopsin; they typically act as light sensors in animals (Terakita, 2005). The visual opsins can be subdivided into cone opsins and rhodopsin. The first cone-opsin sequences isolated were those of the human blue (ultraviolet/ violet sensitive, UVS/VS) (OPN1SW), green and red cone opsins (both longwave sensitive, LWS) (OPN1MW, OPN1LW) (Nathans, Thomas, et al., 1986). The LWS opsin shares the four conserved introns with the rodand UVS/VS cone-opsin, but has an additional fifth intron in a 5' prime position to the four conserved introns (Bellingham, Wells, et al., 2003). Not all species possess all three opsins; for example, cats, dogs, and goats possess red and blue opsins, while pigs, rabbits, and deer have green and blue opsins (Terakita, 2005). Therefore, it is of great interest to study the molecular evolution of color vision genes (Deloukas, Earthrowl, et al., 2004; Terakita, 2005).

From the Ensembl Genome Browser (http://www.ensembl.org/), we downloaded the red cone opsin genes of human (OPN1LW, Ensembl gene no: ENSG00000102076, length: 14,543 bp), dog (OPSR_CANFA, Ensembl gene no: ENSCAFG00000019441, length: 12,865 bp), and zebrafish (opn1lw2, Ensembl gene no: ENSDARG00000028107, length: 5,156 bp) and the green opsin cone gene of mouse (Opn1mw, Ensembl gene no: ENSMUSG00000031394, length: 23,289 bp). Besides, we also downloaded the annotation files. We aligned these sequences by using three MSA tools: ClustalW (Thompson, Higgins, et al., 1994), MAVID (Bray and Pachter, 2004) and MLAGAN (Brudno, Do, et al., 2003). Using the human sequence as the reference, Fig. 5(a) shows the PIP curves of three alignment results in the Detailed View section. It is quite obvious that five exons are well and consistently aligned by

MAVID and MLAGAN but only two exons are aligned by ClustalW. In the stacked-bar chart, the identical rates, over 40%, of the aligned regions by ClustalW are all lower than those by both MAVID and MLAGAN, as shown Fig. 5(b). Thus, if the user directly used the alignment by ClustalW to calculate the substitution rates, the estimated result would be much higher than those obtained by the other tools when no alignment refinement is further performed. Therefore in this case the result obtained by ClustalW is not reliable because several important conserved regions are not aligned. The user should then use the results obtained by either MAVID or MLAGAN for further analysis.

Example 2: **SCL (Stem Cell Leukemia) gene**

The Stem Cell Leukemia (SCL) gene plays a critical role in normal processes that, when disrupted, can result in leukemia. The *SCL* gene, also known as *tal-1*, encodes a basic helix-loop-helix transcription factor that is pivotal for the normal development of all hematopoietic lineages, and is highly conserved between mammals and zebrafish (Barton, Gottgens, et al., 1999; Gottgens, Barton, et al., 2002). Previous analyses of the SCL genes in five vertebrate genomes, including human, mouse, chicken, pufferfish, and zebrafish, have revealed that the SCL promoter/enhancer motifs are conserved in all five species (Gottgens, Barton, et al., 2002). The alignment and visualization tools used in their analyses included BLAST (Altschul, Madden, et al., 1997), PipMaker (Schwartz, Zhang, et al., 2000), and DiAlign (Lenhof, Morgenstern, et al., 1999). Shah et al. (2004) realigned these gene regions in five species by a pairwise alignment tool, LAGAN (Brudno, Do, et al., 2003), and demonstrated the alignment result by Phylo-VISTA (Shah, Couronne, et al., 2004). In this paper, we also downloaded these sequences and realigned them by the multiple alignment tools: ClustalW, MAVID and MLAGAN. The lengths of the human, mouse, chicken, pufferfish, and zebrafish sequences are approximately 100 kb, 65 kb, 67 kb, 22 kb, and 8 kb, respectively.

Fig. 6(a) shows the global view of the results obtained by three alignment tools using the human one as the reference. 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. Moreover, the total quantity and quality of the result obtained by MLAGAN seems better than those by both ClustalW and MAVID, as shown in Fig. 6(b).

Interestingly, when we selected the zebrafish sequence as the reference, the result obtained by ClustalW shows the highest conserved region located at around 27.5kbp whereas those of both MAVID and MLAGAN show it at around 45.89k bp, as shown

in Fig. 6(c). The comparison reveals that the region at around 27.5 kbp in the zebrafish sequence will be assumed the homologous region by ClustalW. But according to MAVID and MLAGAN, the homologous regions are located at around 45.89k bp rather than 27.5k bp. This ambiguous result may be caused by segmental duplication in the sequences and by difference in alignment strategy. In this case, more advanced or further inspections should be performed to either check the detailed alignment results in both regions or realign these sequences by using other pairwise or local alignment tools.

Example 3: **The greater *CFTR* region**

The cystic fibrosis transmembrane conductance regulator (CFTR) gene is responsible for the cystic fibrosis disorder that spans approximately 190 k bp of genomic DNA and consists of 27 exons (McCarthy and Harris, 2005). The greater CFTR region is defined as a genomic segment of about 1.8 M bp on human chromosome 7q31.3 containing the CFTR gene and nine other genes, including TES1, CAV1, CAV2, MET, CAPZA2, ST7, WNT2, GASZ, and CORTBP2 (Thomas, Touchman, et al., 2003). The comparative analysis of this region in 13 vertebrate species has been reported in Thomas et al., 2003 (Thomas, Touchman, et al., 2003) in which the alignment tool used was BlastZ on PipMaker Web server (Schwartz, Zhang, et al., 2000). In this paper, we downloaded the sequences of four mammalian species, including human, baboon, dog, and mouse, from the NIH Intramural Sequencing Center (NISC) Website (http://www.nisc.nih.gov/data/20020612_Target1_0051/). The lengths of these sequences are from 1.0 M bp to 1.5M bp. Since ClustalW could not produce alignment result even in several weeks, we realigned these sequences only by MLAGAN, MAVID, and MULTIZ (Blanchette, Kent, et al., 2004). The total number of bases of the final alignment results, including gaps, is approximately 24 M bp.

Figs. 7(a) and 7(b) show the global PIP curves and their detailed views of three alignment results, respectively. In general, most of high identity regions are well and consistently aligned by these three programs. But those of low identities are not reported by MULTIZ because this program is based on the local alignment results by BlastZ. The stacked-bar chart shows the qualities of these alignment results where the average identical rates for MULTIZ are somewhat better than those for MLAGAN and MAVID as shown in Fig. 7(c). However, some conserved regions between these sequences are not aligned by MULTIZ but identified by MLAGAN and MAVID. Thus, we may wonder whether a better alignment result can be generated by hybridization of these alignment tools.

DISCUSSION

Comparative approach for alignment validity

As the comparison results revealed by SinicView, the alignments of either short or long sequences using different MSA tools are usually inconsistent. We begin to wonder whether the computational results obtained by different tools may in fact lead to different findings. For example, based on our experiments, ClustalW usually misaligns some homologous regions. Moreover, the identical rates of the alignment results are often lower than those obtained by the other programs. Although it was not originally developed to align longer sequences, ClustalW is currently the paragon alignment tool for many users to align their sequences especially in calculating evolutionary distances and construction of phylogenetic trees. For identification of alignment correlation, additional checks of alignment validity by using different scoring systems and different tools have been recognized in the literature (Aparicio, Chapman, et al., 2002). Thus, a cross comparison approach along with visualization could provide an efficient and easy way for general users to verify and validate the alignment results as to whether the aligned regions are reasonable and whether those unaligned regions are indeed non-homologous.

Comparative environment to promote new alignment tools

It is not easy to promote newly developed tools because users usually cannot directly compare the new tools with the traditional ones. With SinicView, users can compare the alignment results obtained by different tools and select an appropriate one for further analysis. Thus, if the new tool can align more regions than those by the old ones and can also indicate their statistical significances, it will be welcomed and better received by the community. We would like to make this tool available to the community of computational biologists. In addition to helping the user find a most appropriate alignment tool to use, SinicView may also be used to check whether previously obtained alignment results by different tools are worth a re-investigation, and see if this revisit of alignment results would lead to different conclusions.

Further possible modifications for SinicView

The capability of fine-tuning parameters relevant to the alignment process will be made available in a user-friendly interface. Furthermore, the ability to allow plug-ins of more alignment programs, in addition to the currently pre-selected ones, such as ClustalW, MAVID, MLAGAN, and GS-Aligner, will inevitably broaden the usage of SinicView. The issue of the compatibility of the input and output formats for each alignment tool also needs to be resolved. For example, both MAVID and MLAGAN require the phylogenetic tree data as input, but ClustalW does not. The ordering of the outputs of these aforementioned tools is usually switched without notice. Thus, to be

able to work under a unified comparison framework requires further processing of these outputs. Besides, identifying a standard-bearer mechanism is still a challenge in entrusting existing alignment programs. So far, we have used the “sum-of-pairs” method to define the “identical rate” in each alignment result. In the future, we may provide other criteria for users to use to measure their alignment results.

CONCLUSION

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 genome-scale sequences, and evaluate the alignment results becomes more urgent. In this paper, we have presented a visualization tool for comparison of multiple sequence alignment programs. With a standard simple protocol for the input/output format, it is quite easy for users to upload their own alignment programs to SinicView. The performance of SinicView depends 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.

ACKNOWLEDGEMENTS

We thank Dr. F.-C. Chen and Dr. H.-K. Tsai for valuable discussions. 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 NSC93-3112-B-001-023-Y and NSC 93-2752-E-002-005-PAE, and by the Institute of Information Science, and the Genomics Research Center of Academia Sinica in Taiwan.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes, *Science*, **297**, 1301-1310.

Barton, L.M., Gottgens, B. and Green, A.R. (1999) The stem cell leukaemia (SCL) gene: a critical regulator of haemopoietic and vascular development, *Int J*

Biochem Cell Biol, **31**, 1193-1207.

Bellingham, J., Wells, D.J. and Foster, R.G. (2003) In silico characterisation and chromosomal localisation of human RRH (peropsin)--implications for opsin evolution, *BMC Genomics*, **4**, 3.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. and Miller, W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner, *Genome Res*, **14**, 708-715.

Bray, N., Dubchak, I. and Pachter, L. (2003) AVID: A global alignment program, *Genome Res*, **13**, 97-102.

Bray, N. and Pachter, L. (2004) MAVID: constrained ancestral alignment of multiple sequences, *Genome Res*, **14**, 693-699.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A. and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA, *Genome Res*, **13**, 721-731.

Brudno, M., Poliakov, A., Salamov, A., Cooper, G.M., Sidow, A., Rubin, E.M., Solovyev, V., Batzoglou, S. and Dubchak, I. (2004) Automated whole-genome multiple alignment of rat, mouse, and human, *Genome Res*, **14**, 685-692.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A. and Johnston, M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting, *Science*, **301**, 71-76.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. and Dubchak, I. (2003) Strategies and tools for whole-genome alignments, *Genome Res*, **13**, 73-80.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. (1999) Alignment of whole genomes, *Nucleic Acids Res*, **27**, 2369-2376.

Deloukas, P., Earthrowl, M.E., Grafham, D.V., Rubenfield, M., French, L., Steward, C.A., Sims, S.K., Jones, M.C., Searle, S., Scott, C., et al. (2004) The DNA sequence and comparative analysis of human chromosome 10, *Nature*, **429**, 375-381.

Dermitzakis, E.T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C. and Antonarakis, S.E. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs), *Science*, **302**, 1033-1035.

Dubchak, I. and Frazer, K. (2003) Multi-species sequence comparison: the next frontier in genome annotation, *Genome Biol*, **4**, 122.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, **5**, 113.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and

high throughput, *Nucleic Acids Res*, **32**, 1792-1797.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I. and Hardison, R.C. (2003) Cross-species sequence comparisons: a review of methods and available resources, *Genome Res*, **13**, 1-12.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I. (2004) VISTA: computational tools for comparative genomics, *Nucleic Acids Res*, **32**, W273-279.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution, *Nature*, **428**, 493-521.

Gottgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R. and Green, A.R. (2002) Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci, *Genome Res*, **12**, 749-759.

Heilig, R., Eckenberg, R., Petit, J.L., Fonknechten, N., Da Silva, C., Cattolico, L., Levy, M., Barbe, V., de Berardinis, V., Ureta-Vidal, A., et al. (2003) The DNA sequence and analysis of human chromosome 14, *Nature*, **421**, 601-607.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution, *Nature*, **432**, 695-716.

Karplus, K. and Hu, B. (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set, *Bioinformatics*, **17**, 713-720.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements, *Nature*, **423**, 241-254.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001) Initial sequencing and analysis of the human genome, *Nature*, **409**, 860-921.

Lenhof, H.P., Morgenstern, B. and Reinert, K. (1999) An exact solution for the segment-to-segment multiple sequence alignment problem, *Bioinformatics*, **15**, 203-210.

Loots, G.G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites, *Nucleic Acids Res*, **32**, W217-221.

Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I. and Rubin, E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites, *Genome Res*, **12**, 832-839.

McCarthy, V.A. and Harris, A. (2005) The CFTR gene and regulation of its expression, *Pediatr Pulmonol*.

Miller, W., Makova, K.D., Nekrutenko, A. and Hardison, R.C. (2004) Comparative genomics, *Annu Rev Genomics Hum Genet*, **5**, 15-56.

Nathans, J., Thomas, D. and Hogness, D.S. (1986) Molecular genetics of human color vision: the genes encoding blue, green, and red pigments, *Science*, **232**, 193-202.

Ovcharenko, I., Loots, G.G., Hardison, R.C., Miller, W. and Stubbs, L. (2004) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles, *Genome Res*, **14**, 472-477.

Ovcharenko, I., Nobrega, M.A., Loots, G.G. and Stubbs, L. (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes, *Nucleic Acids Res*, **32**, W280-286.

Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E. and Eisen, M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA, *BMC Bioinformatics*, **5**, 6.

Raghava, G.P., Searle, S.M., Audley, P.C., Barber, J.D. and Barton, G.J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy, *BMC Bioinformatics*, **4**, 47.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C. and Miller, W. (2003) MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences, *Nucleic Acids Res*, **31**, 3518-3524.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences, *Genome Res*, **10**, 577-586.

Shah, N., Couronne, O., Pennacchio, L.A., Brudno, M., Batzoglou, S., Bethel, E.W., Rubin, E.M., Hamann, B. and Dubchak, I. (2004) Phylo-VISTA: interactive visualization of multiple DNA sequence alignments, *Bioinformatics*, **20**, 636-643.

Shih, A.C. and Li, W.H. (2003) GS-Aligner: a novel tool for aligning genomic sequences using bit-level operations, *Mol Biol Evol*, **20**, 1299-1309.

Terakita, A. (2005) The opsins, *Genome Biol*, **6**, 213.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions, *Nature*, **424**, 788-793.

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.

Ureta-Vidal, A., Ettwiller, L. and Birney, E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes, *Nat Rev Genet*, **4**, 251-262.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001) The sequence of the human genome, *Science*, **291**, 1304-1351.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002) Initial sequencing and comparative analysis of the mouse genome, *Nature*, **420**, 520-562.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals, *Nature*, **434**, 338-345.

The legends of figures

Fig. 1: The screenshot shows the user interface of SinicView. The alignment result is of the SCL gene regions in human, mouse, chicken, pufferfish, and zebrafish. Three alignment results of five sequences aligned by ClustalW, MAVID, and MLAGAN are shown.

Fig. 2: Different from the result shown in Fig. 1, the SinceView shows the snapshots with the references being (a) mouse and (b) chicken respectively. The curve-based plots in blue color are the percent identity rates of ClustalW, MAVID, and MLAGAN, respectively (from the top down.)

Fig. 3: The Tools Menu functions. Two comparison charts can be generated by SinicView: the stacked-bar chart illustrates the proportion comparison of cross alignment results and the pie chart shows the proportion of different identical rates of an individual alignment result. The complete data of the charts are tabulated on the left.

Fig. 4: The detailed text display of the different alignment results. The matched identical sequences are labeled in red blocks. Interestingly, all three results do not contain consistent matching alignments in this case.

Fig. 5. The cross comparison of three alignment results of four LW/MW opsin sequences by SinicView. (a) The PIP curves show the whole scale comparison while the human one is selected as the reference. (b) Comparison of the results in

stacked-bar chart.

Fig. 6: (a) The comparison of three alignment results by SinicView while using the human sequence as the reference. (b) The stacked-bar chart generated by SinicView illustrates the proportion comparison of cross alignment results. (c) Using zebrafish as the reference, the highest conserved region (around 62%) produced by ClustalW concentrates around at 27.5kbp. However, there are discrepancies between the result of ClustalW and those of MAVID and MLAGAN.

Fig. 7. The cross comparison of three alignment results by SinicView. (a) The whole scale PIP curves using the human one as reference. (b) The detailed view of (a). (c) Comparison of the results in the stacked-bar chart. (d) Comparison of the results in the pie charts.

Fig.1.

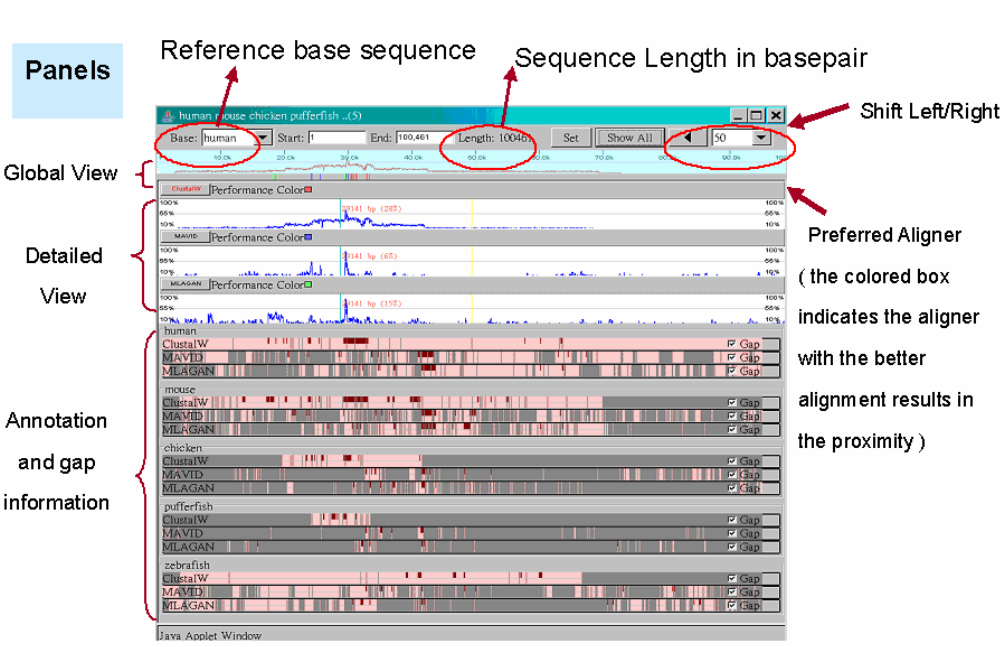
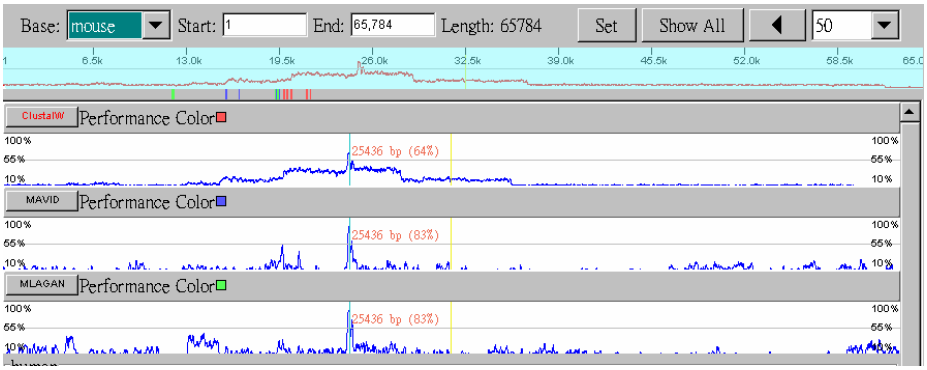
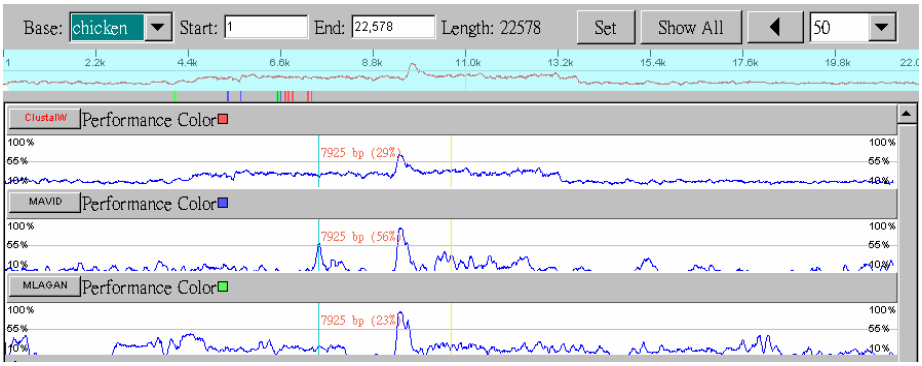

Panels
Reference base sequence
Sequence Length in basepair
Shift Left/Right
Global View
Detailed View
Annotation and gap information
Preferred Aligner (the colored box indicates the aligner with the better alignment results in the proximity)
Performance Color
human
mouse
chicken
pufferfish
zebrafish
ClustalW
MAVID
MLAGAN
Gap
Java Applet Window

Fig.2.


Base: mouse
Start: 1
End: 65,784
Length: 65784
Set
Show All
50
ClustalW Performance Color
25436 bp (64%)
MAVID Performance Color
25436 bp (83%)
MLAGAN Performance Color
25436 bp (83%)

(a)


Base: chicken
Start: 1
End: 22,578
Length: 22578
Set
Show All
50
ClustalW Performance Color
7925 bp (29%)
MAVID Performance Color
7925 bp (56%)
MLAGAN Performance Color
7925 bp (23%)

(b)

Fig. 3.

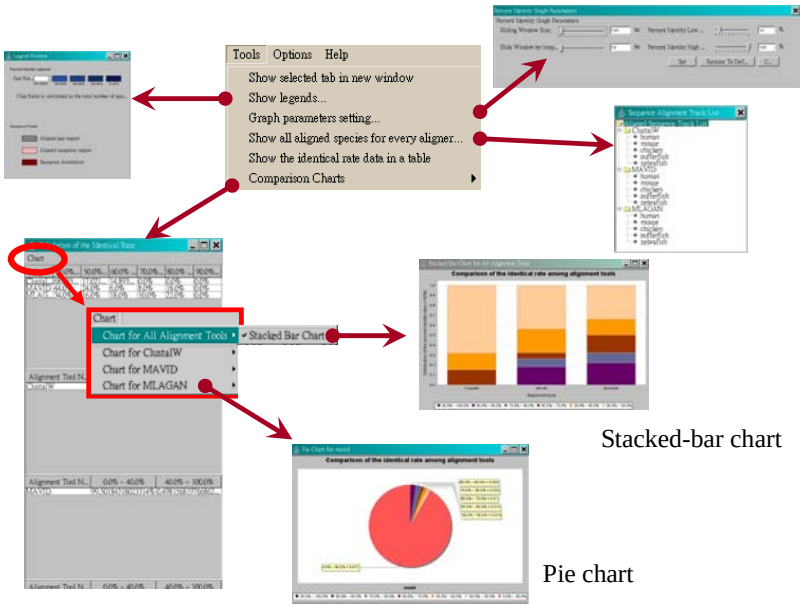

Tools
Options
Help
Show selected tab in new window
Show legends...
Graph parameters setting...
Show all aligned species for every aligner...
Show the identical rate data in a table
Comparison Charts
Chart
Chart for All Alignment Tools
Chart for ClustalW
Chart for MAVID
Chart for MLAGAN
Stacked Bar Chart
Stacked-bar chart
Pie chart

Fig. 4.

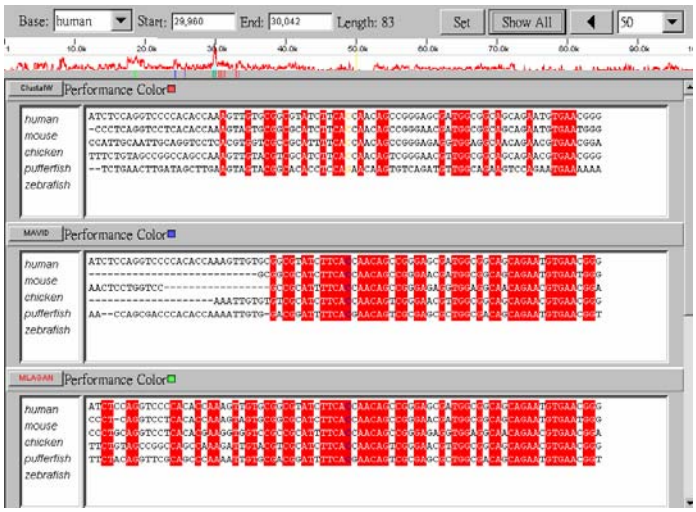
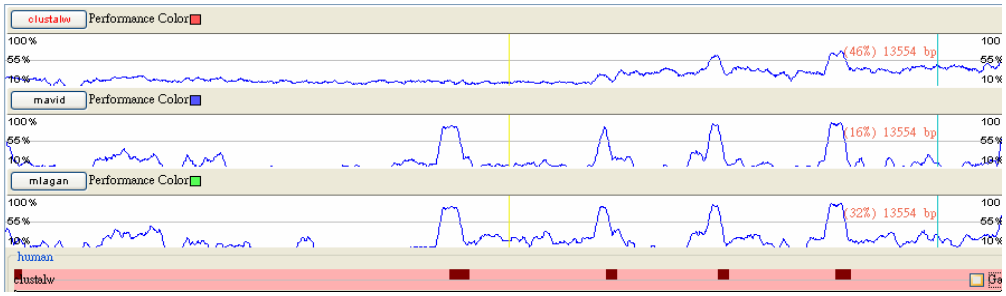
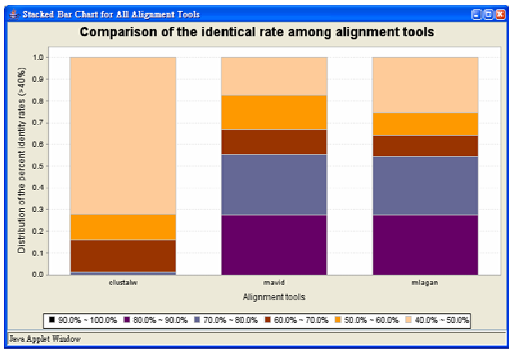

Base: human
Start: 29,960
End: 30,042
Length: 83
Set
Show All
ClustalW Performance Color
MAVID Performance Color
MLAGAN Performance Color
human
mouse
chicken
pufferfish
zebrafish

Fig. 5.

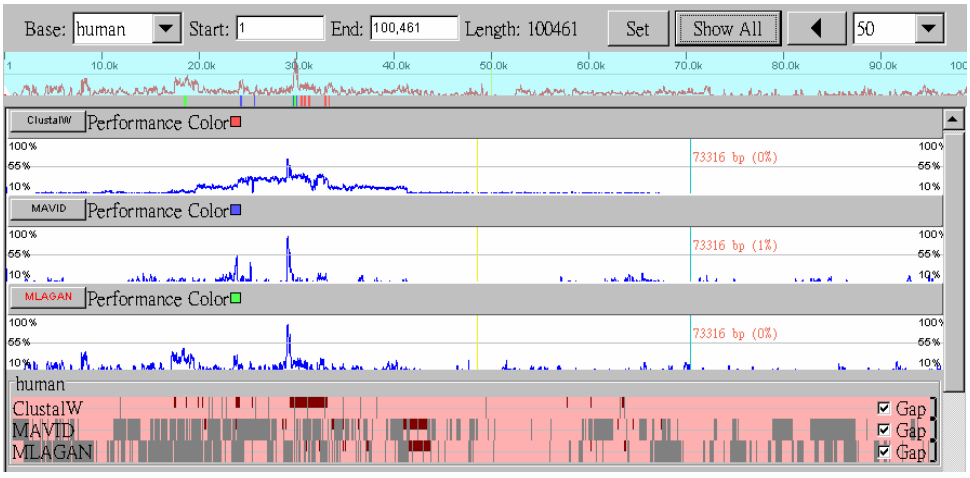

clustalw
Performance Color
100 %
55 %
10 %
(46%) 13554 bp
mavid
Performance Color
(16%) 13554 bp
mlagan
Performance Color
(32%) 13554 bp
human
clustalw

(a)

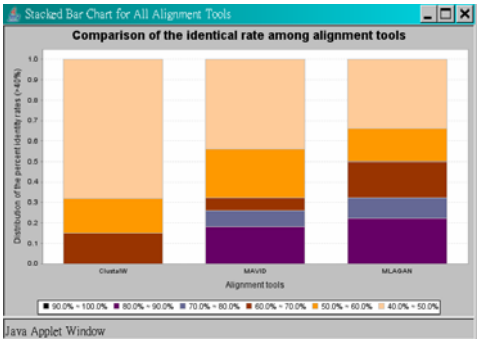

Stacked Bar Chart for All Alignment Tools
Comparison of the identical rate among alignment tools
Distribution of the percent identity rates (>40%)
clustalw
mavid
mlagan
Alignment tools
90.0% ~ 100.0%
80.0% ~ 90.0%
70.0% ~ 80.0%
60.0% ~ 70.0%
50.0% ~ 60.0%
40.0% ~ 50.0%
Java Applet Window

(b)

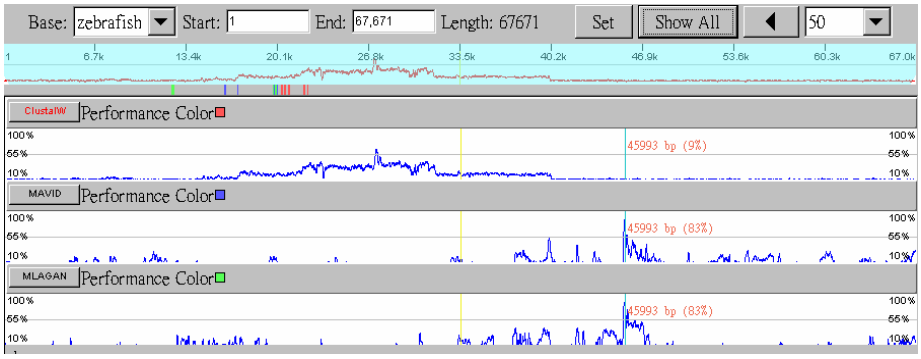
Fig. 6.


Base: human
Start: 1
End: 100,461
Length: 100461
Set
Show All
50
ClustalW Performance Color
73316 bp (0%)
MAVID Performance Color
73316 bp (1%)
MLAGAN Performance Color
73316 bp (0%)
human
ClustalW
MAVID
MLAGAN
Gap
Gap
Gap

(a)

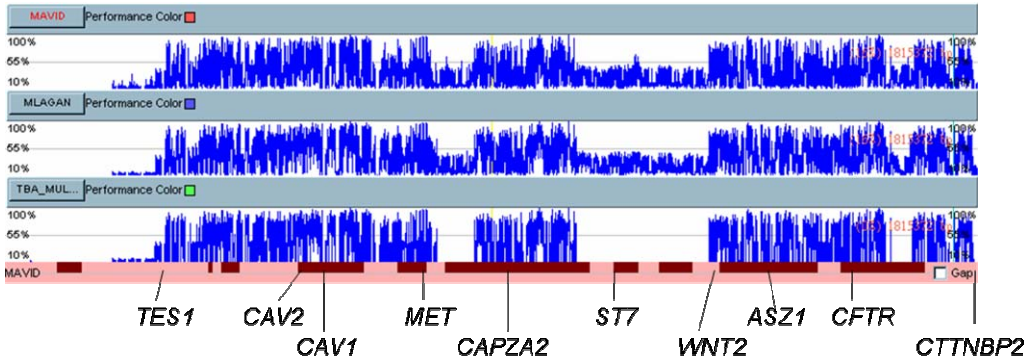

Stacked Bar Chart for All Alignment Tools
Comparison of the identical rate among alignment tools
Distribution of the percent identity rates (+42%)
ClustalW
MAVID
MLAGAN
Alignment tools
90.0% - 100.0%
80.0% - 90.0%
70.0% - 80.0%
60.0% - 70.0%
50.0% - 60.0%
40.0% - 50.0%
Java Applet Window

(b)

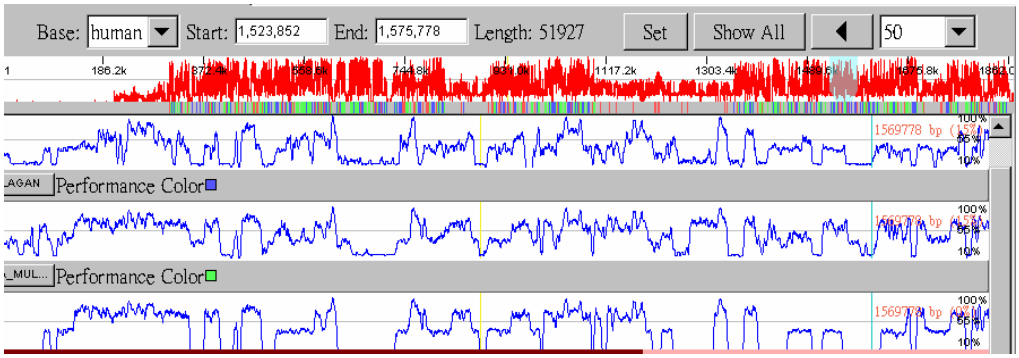

Base: zebrafish
Start: 1
End: 67,671
Length: 67671
Set
Show All
50
ClustalW Performance Color
45993 bp (9%)
MAVID Performance Color
45993 bp (83%)
MLAGAN Performance Color
45993 bp (83%)

(c)

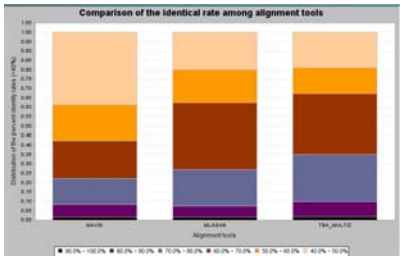
Fig. 7.


MAVID
Performance Color
MLAGAN
TBA_MUL...
100%
55%
10%
TES1
CAV2
CAV1
MET
CAPZA2
ST7
WNT2
ASZ1
CFTR
CTTNBP2

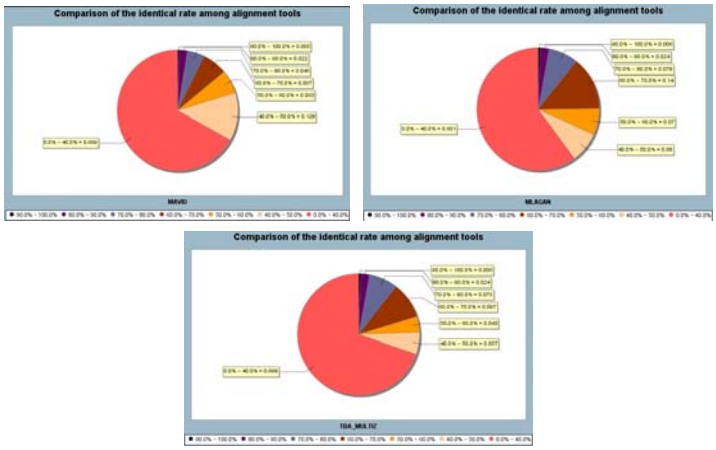
(a)


Base: human
Start: 1,523,852
End: 1,575,778
Length: 51927
Set
Show All
50
Performance Color
Performance Color

(b)


Comparison of the identical rate among alignment tools

(c)


Comparison of the identical rate among alignment tools
Comparison of the identical rate among alignment tools
Comparison of the identical rate among alignment tools

(d)