BMC Bioinformatics

Software

SinicView: A visualization environment for comparisons of multiple nucleotide sequence alignment tools

Arthur Chun-Chieh Shih[1], DT Lee*[1,2], Laurent Lin[1], Chin-Lin Peng[2], Shiang-Heng Chen[1], Yu-Wei Wu[1], Chun-Yi Wong[1], Meng-Yuan Chou[1], Tze-Chang Shiao[1] and Mu-Fen Hsieh[1]

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

Email: Arthur Chun-Chieh Shih - arthur@iis.sinica.edu.tw; DT Lee* - dtlee@ieee.org; Laurent Lin - laurent@iis.sinica.edu.tw; Chin-Lin Peng - coolpon@gate.sinica.edu.tw; Shiang-Heng Chen - shiangheng@gmail.com; Yu-Wei Wu - karlon@iis.sinica.edu.tw; Chun-Yi Wong - robinw@iis.sinica.edu.tw; Meng-Yuan Chou - mychou@iis.sinica.edu.tw; Tze-Chang Shiao - supera@iis.sinica.edu.tw; Mu-Fen Hsieh - mufen@tamu.edu

* Corresponding author

Published: 02 March 2006

*BMC Bioinformatics*2006, **7**:103 doi:10.1186/1471-2105-7-103

Received: 25 September 2005
Accepted: 02 March 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/103

Abstract

Background: Deluged by the rate and complexity of completed genomic sequences, the need to align longer sequences becomes more urgent, and many more tools have thus been developed. In the initial stage of genomic sequence analysis, a biologist is usually faced with the questions of 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 of whether poorly aligned 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 (MSA) programs have been proposed, they may not provide a standard-bearer for most biologists because those poorly aligned regions in these evaluations are never discussed. Thus, a tool that allows cross comparison of the alignment results obtained by different tools simultaneously 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 nucleotide alignment results obtained by different tools. SinicView calculates similarity of the alignment outputs under a fixed 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. We illustrate the capabilities of our visualization system by comparing alignment results obtained by MLAGAN, MAVID, and MULTIZ, respectively.

Conclusion: 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 in the initial stage of sequence analysis.

Background

With exponentially increasing genomic sequences available in the public domain [1-5] comparative genomics demonstrates its power to help biologists identify novel conserved and functional regions in genomes [6-9]. 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 [10], metazoan genomes [11], vertebrate genomes [12], and mammalian genomes [13], discover regulatory motifs in the yeast [14] and human promoters [15] or identify potential conserved non-genic sequences (CNGs) [16].

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. To better tackle this problem, several tools for genomic sequence alignment have been proposed, such as pairwise sequence aligners like MUMmer [17], GS-Aligner [18], Avid [19] and LAGAN [20], and multiple sequence alignment (MSA) programs like T-COFFEE[21], MAFFT [22], MultiPipMaker [23], MULTIZ [24], MLAGAN [20], MAVID [25], and MUSCLE [26,27]. These alignment tools, however, are heuristics based and do not provide any indication of how far they are from an optimal solution. The comparisons of alignment tools using a set of benchmarking sequences have also been conducted in recent years [28-30]. We found that the majority of these tools usually fail to generate consistent results especially in aligning divergent cross-species sequences. As a result, 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.

Although the comparison results in [28-31] provide some evaluations 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 poorly aligned regions produced by the alignment tools are indeed non-homologous or just due to inappropriate tools or scoring systems used. Consequently, 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. These tools can be roughly divided into two categories: integrated genome/sequence browser and individual alignment result visualization. In the former category, such as UCSC ENCODE project [32,33], UCSC human genome browser [34], *Ensembl* [35], ECR Browser [36,37], users can view alignment results mapped onto the sequenced genomes. Some of these browsers also provide registered users to submit alignment results and see the conservation regions between different genomes. In the latter category, the tools are developed to visualize individual alignment results. The VISTA-related tools are among the famous ones that have been developed for several years [38]. 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 [39]. rVISTA (regulatory Vista) combines database searches for transcription factor binding sites with a comparative sequence analysis [40,41]. GenomeVISTA compares users' sequences with several whole genome assemblies [42,43]. Phylo-VISTA analyzes alignments of multiple DNA sequences from different species while considering their phylogenetic relationships [44]. 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 [23,45], and zPicture [46] 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 parameters of a scoring system, such as changing match rewards or mismatch penalties, 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 for the present calculates similarity of the alignment outputs under a fixed window using the sum-of-pairs method and provides scoring profiles of each set of aligned sequences. Other scoring matrices, such as EMBOSS DNA scoring matrix [47] and YASS [48], are also provided in SinicView for users to select. Besides, users can also upload their preferable scoring matrices to calculate the scoring profile

Figure 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.

curves. 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 poorly aligned regions.

Implementation

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-pair 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 automatically switch from the curve-based plot to the display of the detailed alignments in a colored text format where identical characters are shown. 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 scoring distribution of alignment results in a curve-based graphic format;

2. Generation of the comparison charts using stacked-bar and pie charts, which shows the distribution of the identical rates among various alignment programs for benchmarking purposes;

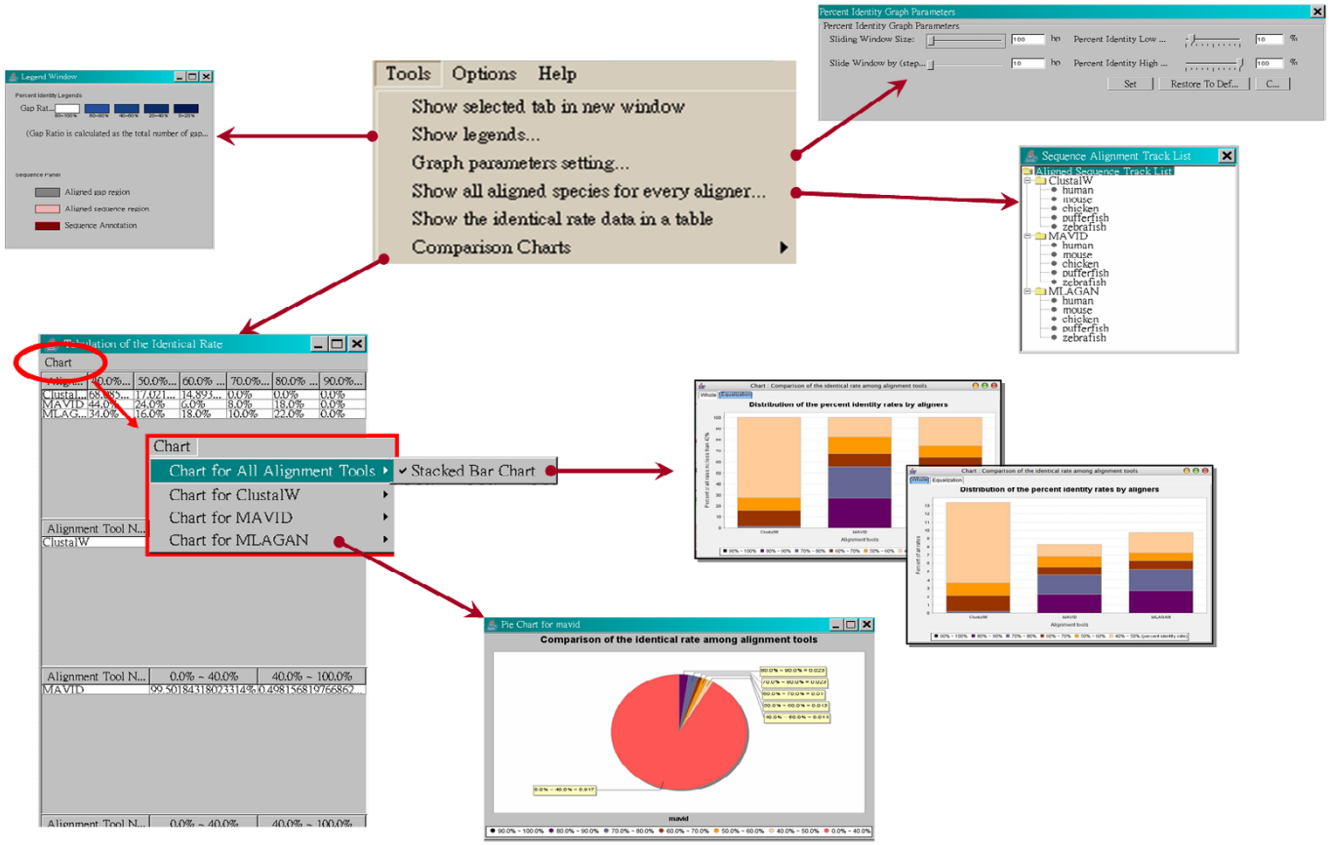

Figure 2
The tools menu functions. Two comparison charts can be generated by SinicView: the stacked-bar chart illustrates the proportion comparison of cross alignment results and the pie chart shows the proportion of different identical rates of an individual alignment result. The complete data of the charts are tabulated on the left.

3. Inclusion of a versatile manipulative functionality (gap-display toggling, drag-and-drop zooming/shifting, and graphic/text display toggling);

4. Visualization of annotation information and display of the phylogenetic trees provided by users in which the drawing tree program uses the ATVtree [49];

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 Figure 1, SinicView displays the alignment results obtained by three different MSA methods. The input sequences contain the 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 specifies 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.

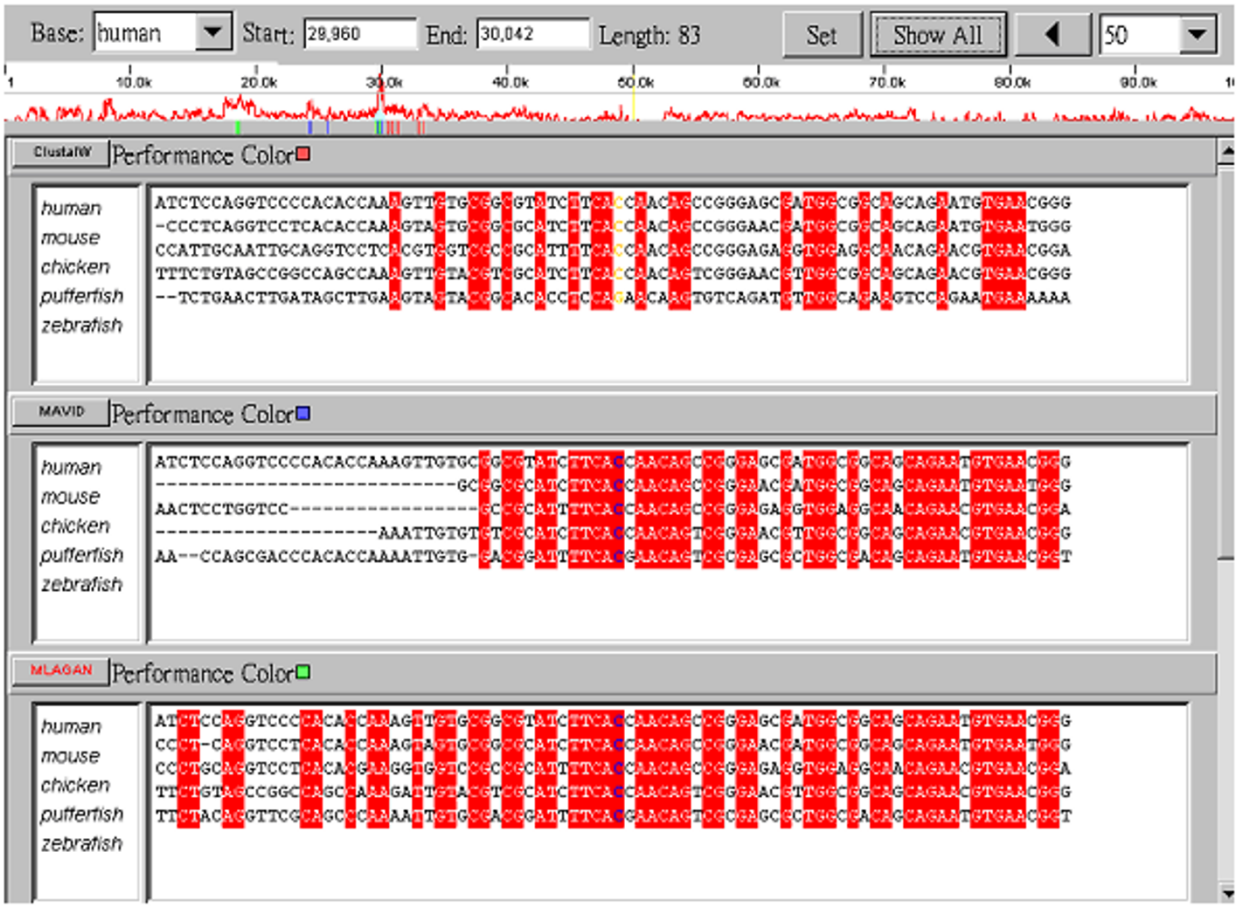

Figure 3
The detailed text display of the different alignment results. The matched identical sequences are labeled in red blocks. Interestingly, all three results do not contain consistent matching alignments in this case.

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 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 Figure 1. When the user selects different sequences as the

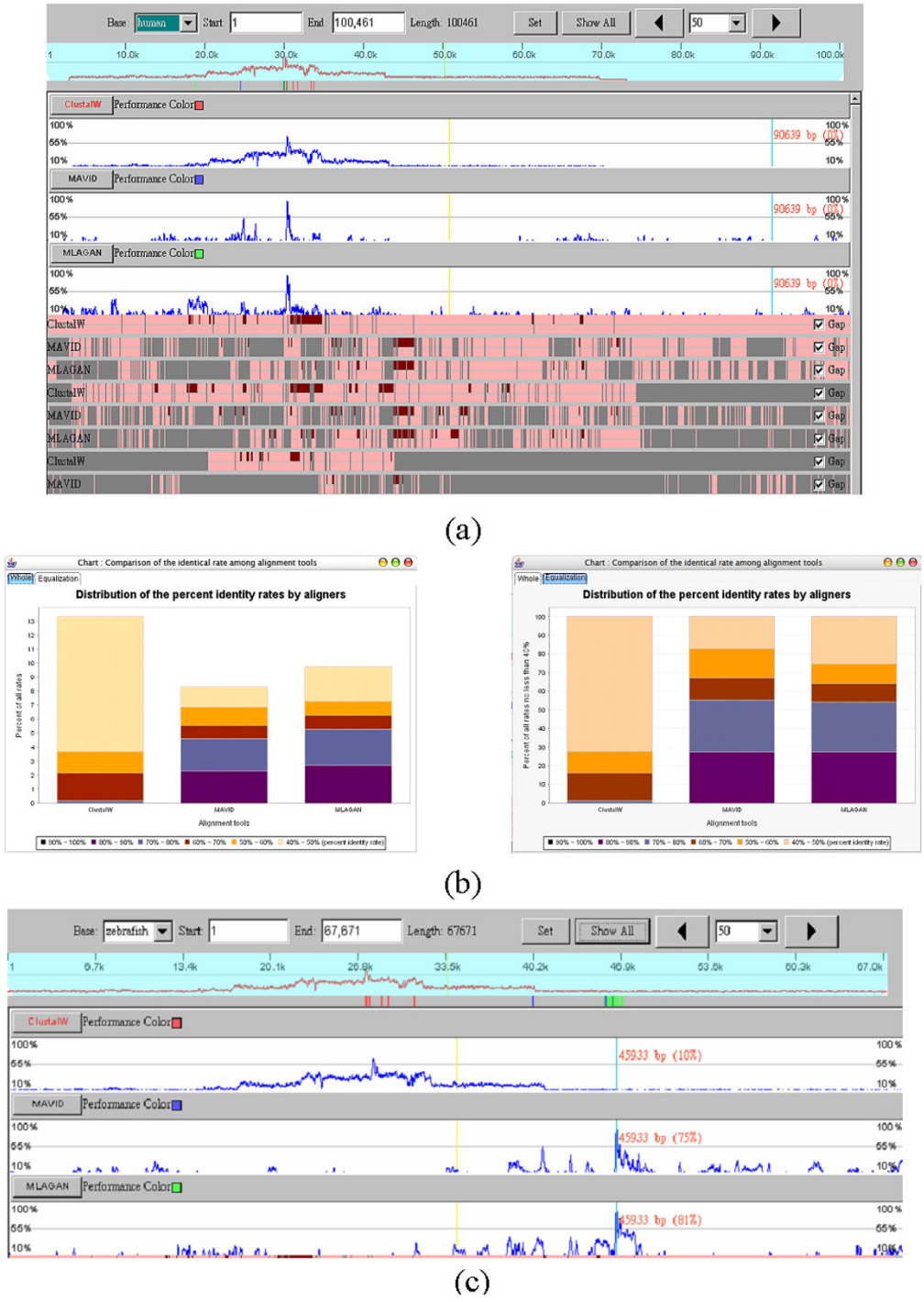

Figure 4

The comparison of different alignment results of SCL gene regions. (a) The comparison of three alignment results by SinicView while using the human sequence as the reference. (b) The whole (non-equalization) and equalization stacked-bar charts 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 27.5 k bp. However, there are discrepancies between the result of ClustalW and those of MAVID and MLAGAN.

reference, SinicView can demonstrate the variations between the PIP curves of the alignment results.

Visualization of SinicView: comparison chart and text-mode comparison

The functionality under the "Tools" menu, called "Comparison Charts", offers two types of charts for quick-and-easy evaluation of the alignment quality. The stacked bar chart, in Figure 2, illustrates the distribution of the identical rates with the threshold over 40%. The pie chart, on the other hand, 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.

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 Figure 3, 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 JRE (Java Runtime Environment)-enabled browsers with Internet connection, thus making the installation and choosing the right platform hassle-free. 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 have also implemented a standalone application of SinicView, which is wrapped in JRE, for off-line use.

The execution procedure of the standalone SinicView is quite straightforward. Upon launch, the user will be prompted three options. The first two are to read user's Phylogenetic Tree files, an option, and MSA results from the local disk.

Results

In what follows, we will introduce two examples to demonstrate how SinicView can assist users to analyze alignment results in the initial stage of sequence comparison. The total alignment lengths in both of the examples are few hundreds of thousands of base pairs and several millions of base pairs, respectively. The conservations of the aligned sequences are different in each example. More examples can be found in [50].

Example 1: 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 [51,52]. 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 [51]. The alignment and visualization tools used in their analyses included BLAST [53], PipMaker [45], and DiAlign [54]. Shah et al. (2004) realigned these gene regions in five species by a pairwise alignment tool, LAGAN [20], and demonstrated the alignment result by Phylo-VISTA [44]. 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.

Figure 4(a) shows the global view of the results obtained by three alignment tools using the human sequence as the reference. Generally speaking, the highest conserved region located at 30 k bp 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 of the result obtained by MLAGAN is better than those by both ClustalW and MAVID while the quantity of the result obtained by ClustalW is better than those by the others, as shown in Figure 4(b). Interestingly, when we selected the zebrafish sequence as the reference, the result obtained by ClustalW shows the highest conserved region located at around 27.5 k bp whereas those by both MAVID and MLAGAN show it at around 45.89 k bp, as shown in Figure 4(c). The comparison reveals that the region at around 27.5 k bp 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 45.89 k bp rather than at 27.5 k 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 2: 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 [55]. 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 [12]. The

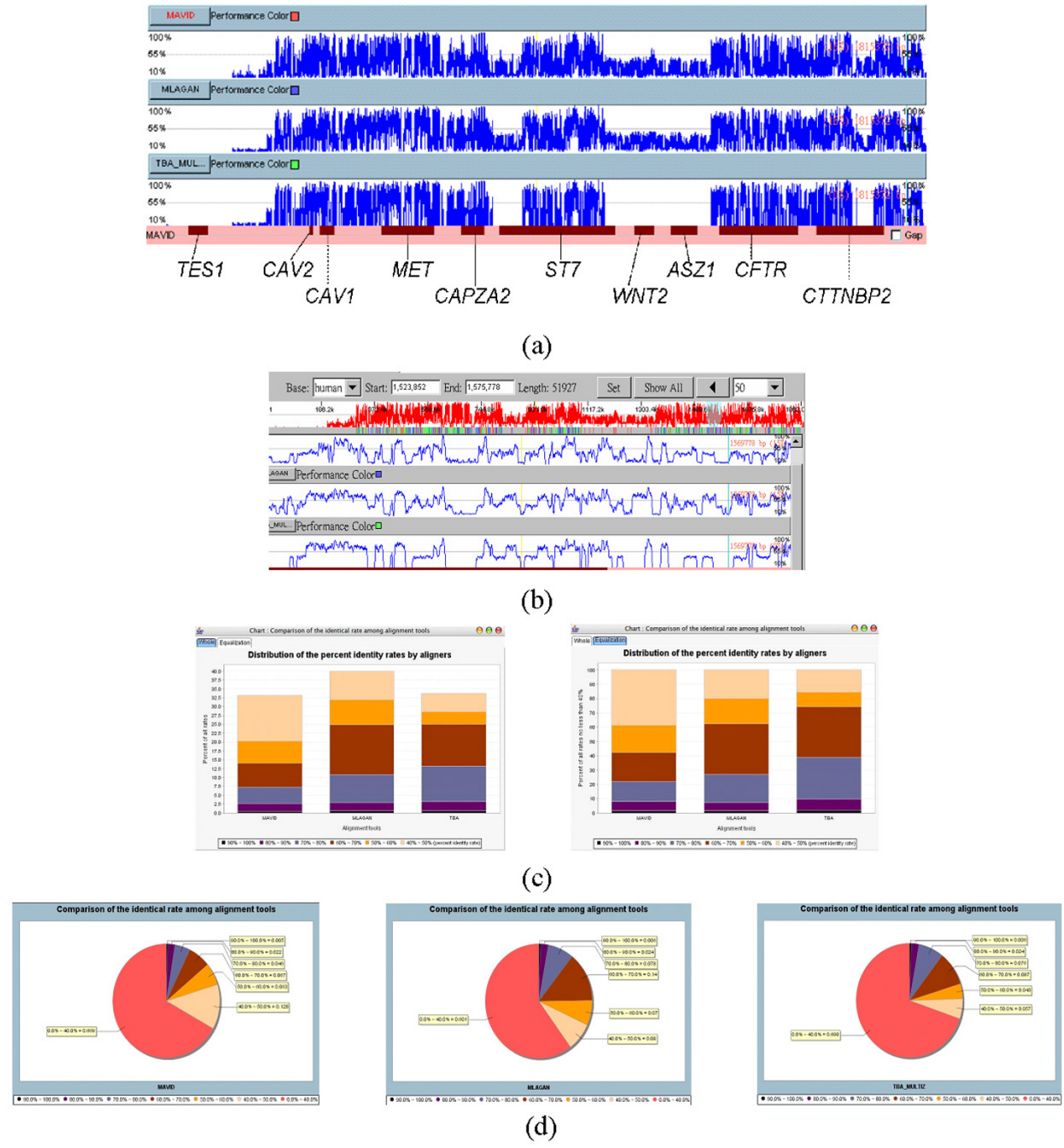

Figure 5
The comparison of different alignment results of great CFTR gene regions. The cross comparison of three alignment results by SinicView. (a) The whole scale PIP curves using the human one as reference. (b) The detailed view of (a). (c) Comparison of the results in the whole and equalization stacked-bar charts. (d) Comparison of the results in the pie charts.

comparative analysis of this region in 13 vertebrate species has been reported in Thomas et al., 2003 [12] in which the alignment tool used was BlastZ on PipMaker Web server [45]. 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 [56]. However, the original sequences had been updated in other genome

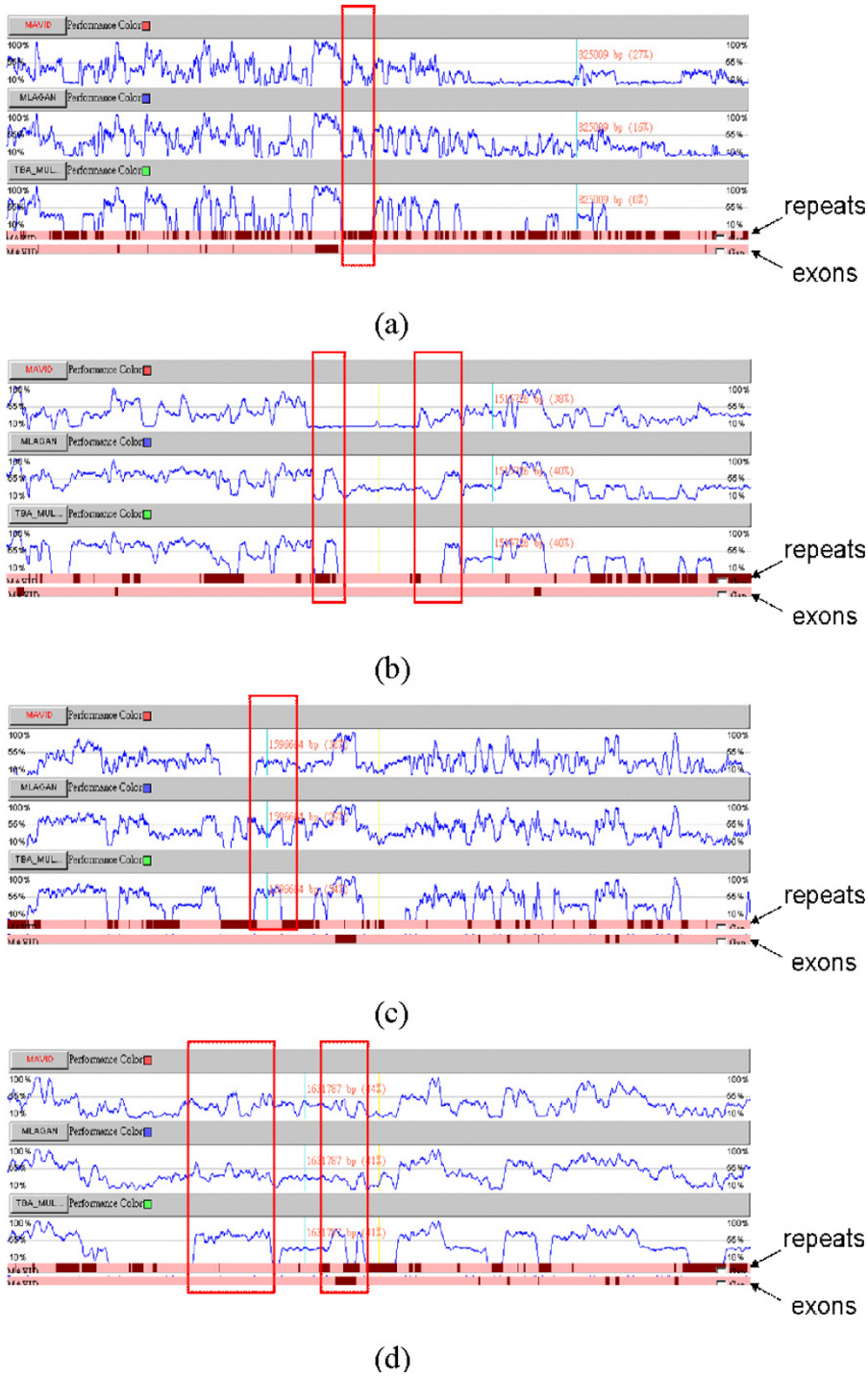

Figure 6
The detailed comparison of Example 2. The detailed comparison of different alignment results of great CFTR gene regions at different intervals. (a) From 786,112 bp to 836,774 bp. (b) From 1,500,792 bp to 1,523,689 bp. (c) From 1,583,342 bp to 1,621,404 bp. (d) From 1,623,603 bp to 1,644,063 bp.

browsers. Thus, we eventually downloaded the last versions of these sequences from the UCSC Genome Browser. The lengths of these sequences are from 1.0 M bp to 1.5 M bp. We realigned these sequences by MLAGAN, MAVID, and TBA (kernel: MULTIZ) [24] and the total number of bases of the final alignment results, including

Table 1: The test results of the applet version and standalone application of SinicView on different platforms and OS's

	Applet		Standalone Application	
	Specification (Applet)	Status	Specification (Application)	Status
Sun OS	OS : Sun OS 5.7 Sparc JVM : java_1.4.2_08	**OK**	OS : Sun OS 5.7 Sparc JVM : java_1.4.2_08	**OK**
Mac OS	OS : Mac OS 10.4.2 Tiger JVM : java_1.4.2_08 java_1.5 update 4 Browser : Safari 2.0	**OK**	OS : Mac OS Tiger 10.4.2 java_1.5 update 4 java_1.5 update 4	**OK**
Linux/Unix	OS : Linux Fedora Core 3 JVM : java_1.4.2_08 Browser : Mozilla Firefox 1.0.2	**OK**	OS : Linux Fedora Core 3 JVM : java_1.4.2_08	**OK**
Windows	OS : Windows XP Service Pack 2 JVM : java_1.4.2_08 java_1.5 update 4 Browser : Internet Explorer 6.0 Mozilla Firefox 1.0.4	**OK**	OS : Windows XP Service Pack 2 JVM : java_1.4.2_08 java_1.5 update 4	**OK**

gaps, are approximately 12 M bp, 11 M bp, and 7.5 M bp, respectively.

Figures 5(a) and 5(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 not as high identities are not reported by TBA because the kernel of this program, MULTIZ, is based on the local alignment results by BlastZ. As shown in Figure 5(c), the stacked-bar charts show the quality and the quantity of these alignment results where the average identical rates for TBA are somewhat better than those for MLAGAN and MAVID although the total number of aligned conserved regions for MLAGAN is larger than those for the others.

For comparisons of these alignments from a functional viewpoint, we downloaded the annotation of the human sequence, including exons and repeats, from the Ensembl Genome Browser [35]. The detailed comparisons of the alignment results by different aligners demonstrated that the alignments of noncoding regions are often inconsistent. But for the coding regions, the alignment results by different aligners seem consistent and well-aligned.

Figures 6(a)–(b) show the detailed alignment results at four different intervals. In Figure 6(a), we find that some conserved regions are not aligned by TBA but identified by MLAGAN and MAVID. This region is annotated by repeats and implies that some repetitive elements were inserted into these sequences of their common ancestor. However, this conserved insertion event could not be observed by using TBA. Although the kernel of TBA, MULTIZ, is known not to align regions with repetitive elements, we still find that some other regions with repetitive elements are aligned by this program, as shown in Figure 6(b).

Generally speaking, the regions aligned by TBA usually have higher identical rates than by others. As the frames shown in red in Figures 6(c) and 6(d), the alignment of these regions by TBA seems superior to those by others. However, the kernel of TBA, MULTIZ, usually neglects to align the regions with low conservations. Thus, some lowly conserved regions may not be aligned by TBA.

Since each alignment tool has its own advantage and reveals different alignment results, we therefore wonder whether a better alignment result can be generated by hybridization of these alignment tools.

Loading performance and platforms test

SinicView is implemented totally in Java. Theoretically, it should be portable across different operating systems (OSs) and platforms. To demonstrate interoperability on real cases, we tested the applet and application versions of SinicView on different platforms and OSs. As shown in Table 1, both versions of SinicView seem to perform well. Thus, users can use either the applet version or the standalone application of SinicView, according to their requirements.

Besides, we also tested the loading performance of SinicView. Because the performance of an applet on the Web is strongly dependent on the network bandwidth and traffic, the estimation of loading time may not be a fair comparison. Thus, in this part we only estimated the loading performance of the standalone application of SinicView.

In general, the loading performance of a Java application is dependent on the memory heap size. The default values of the initial heap size and the maximum size of a Java Virtual Machine (java_1.4.2 version or higher) are 4 M (mega) bytes and 64 M bytes, respectively. These values

Table 2: The loading performance of standalone SinicView The loading time of standalone SinicView by different sizes of input data and initial and maximum memory heap sizes. The default value for the initial JVM heap size is 4 M bytes; maximum is 64 M bytes. For the maximum 64 M byte heap size, the standalone SinicView can handle up to approximately 11 M byte alignment data. The maximum value of the input data size is linear in the maximum heap size. We observe that the initial heap memory size has little impact on the loading time. This result was benchmarked on a 3 GHz Pentium4 PC with 1 GB RAM.

Input data size (bytes)	Loading Time (sec) Java Application Virtual Machine Memory Heap Size, Initial/Max (M Bytes)				
	64 MB/64 MB	128 MB/128 MB	64 MB/256 MB	128 Mb/256 MB	256 MB/256 MB
0.5 M	4	4	4	4	4
1 M	6	7	8	8	7
5 M	28	27	27	26	26
10 M	59	53	56	55	55
20 M	NA	104	107	105	106
40 M	NA	NA	214	212	212

NA: Not available.

can be adjusted by the following command in the terminal mode:

java -Xms64m -Xmx128m -jar SinicView.jar,

where the parameters Xms64m and Xmx128m represent that the initial heap size is 64 M bytes and the maximum size is 128 M bytes, respectively. Thus, we used different input data sizes, initial heap sizes, and the maximum sizes to estimate the loading time of SinicView. As shown in Table 2, using the default maximum heap size, 64 M bytes, the standalone SinicView can handle up to approximately 11 M bytes alignment data. If the maximum size is set up to 256 M bytes, the loading ability of input data size could be over several dozens of mega bytes. Moreover, Table 2 shows that the maximum data size is dependent on the maximum heap size and the loading times are linearly dependent on the sizes of input data. All performance test results were benchmarked on a 3 GHz Pentium4 PC with 1 GB RAM.

Discussion

Repetitive elements in sequence alignments

The eukaryotic genome is usually characterized by the presence of repetitive DNA consisting of nucleotide sequences of various lengths and compositions that occur from a few times to thousands of times in the genome either in tandem or in a dispersed fashion[57]. The repetitive fractions can be classified into two types of repeated families: localized and dispersed [57,58]. Localized repetitive sequences usually occur as tandem arrays and they are called tandem repetitive DNA. Dispersed repetitive sequences are dispersed throughout the genome. In addition, there are moderately repetitive sequences, which are usually transposable elements or processed pseudogenes and are usually dispersed over the genome. *Alu* is the largest family of interspersed mobile elements (~300 bp) and propagated to more than one million copies in primate genomes. This type of repeat has been inserted into these genomes within the last 65 million year period [58]. Because this type of repetitive elements only appears in the primate genomes, when we align homologous sequences of primate and non-primate genomic sequences, these *Alu* inserted regions should not be aligned. However, other interspersed elements may possibly have been inserted into the ancestral sequence of mammalians. The regions of these repeats may be able to align together between the sequences of different mammalians, as shown in Example 2. However, these regions in the alignment results by different aligners are inconsistent. Since these repetitive elements in sequences could be detected by RepeatMasker [59], the poorly aligned regions may have to be checked whether they belong to repetitive elements.

Comparative approach for alignment validity

As the comparison results using SinicView show, the alignments of sequences using different MSA tools are inconsistent. We begin to wonder whether the computational results obtained by different tools may in fact lead to different findings. For identification of alignment correlation, a need for additional checks of alignment validity by using different tools and scoring systems has been recognized in the literature [60]. 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 poorly aligned regions are indeed non-homologous.

How to decide on a "good" alignment result

Except evaluation of the alignment quality by comparison charts in SinicView, how to decide on a good alignment with biological meanings may need much more experiences and knowledge. Sometimes, this judgment depends also on what kind of the biological problems users want

to study. Here, we suggest some general rules for users to judge the alignments by biological meanings.

In the coding regions, a triplet of adjacent nucleotides constitutes a codon. Usually, the first two nucleotides are identical between the two sequences and allow the third one to be either identical or different. Thus, when the partial alignment results reveal the two-out-of-three regularity for each triplet, it may imply that the aligned regions are potential coding regions. This alignment result should be more biologically meaningful than those without the two-out-of-three regularity.

From molecular evolutionary viewpoint, nature prefers inserting or deleting considerable consecutive nucleotides together to interspersed individual nucleotides [57]. Thus, an alignment with consecutive gaps would be better than those with interspersed gaps.

If one of the alignment sequences has been annotated, the information is definitely useful for users to judge the alignment results by different aligners.

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 SinicView 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 enhancements 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, in addition to what have been already provided in SinicView.

Conclusion

Deluged by the increasing number of 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 64 M 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. More information can be found at [50].

Availability and requirements

Project name: 1. Development of Novel Large-scale Sequence Alignment and Visualization Tools and Their Applications to Bioinformatics

2. Development of a web-based personalized research environment for study of computational and evolutionary genomics

Project home page: http://biocomp.iis.sinica.edu.tw

Operating system(s): Window XP, Sun OS 5.7 Sparc, Mac OS 10.4.2 Tiger, and Linux Fedora Core 3

Programming language: Java

Other requirements: Java 1.4.2 or higher

License: Any restrictions to use by non-academics: free downloads and usage for academics only.

List of abbreviations

SinicView: Sequence-aligning INnovative and Interactive Comparison VIEWer

JRE: Java Runtime Environment

SCL: Stem Cell Leukemia

CFTR: Cystic Fibrosis Transmembrane Conductance Regulator

Authors' contributions

Arthur Chun-Chieh Shih and D.T. Lee contributed the original idea, developed the system organization, and drafted the paper. Laurent Lin supervised the system implementation and also drafted some parts of the paper. Chin-Lin Peng, Yu-Wei Wu, Chun-Yi Wong, Meng-Yuan Chou, and Tze-Chang Shiao implemented the codes. Shiang-Heng Chen and Mu-Fen Hsieh implemented some partial codes before leaving their positions.

Acknowledgements

We thank Dr. Feng-Chin Chen and Dr. Huai-Kuang Tsai for valuable discussions and Mr. Hung-Yi Chen for his assistance in organizing some alignment results. We also thank the anonymous reviewers for their comments and suggestions that help improve the presentation of this paper. 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, NSC-94-2213-E-001-029, 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

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291(5507):**1304-1351.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409(6822):**860-921.
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J,

Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420(6915):**520-562.

4. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, Dodgson JB, Chinwalla AT, Cliften PF, Clifton SW, Delehaunty KD, Fronick C, Fulton RS, Graves TA, Kremitzki C, Layman D, Magrini V, McPherson JD, Miner TL, Minx P, Nash WE, Nhan MN, Nelson JO, Oddy LG, Pohl CS, Randall-Maher J, Smith SM, Wallis JW, Yang SP, Romanov MN, Rondelli CM, Paton B, Smith J, Morrice D, Daniels L, Tempest HG, Robertson L, Masabanda JS, Griffin DK, Vignal A, Fillon V, Jacobbson L, Kerje S, Andersson L, Crooijmans RP, Aerts J, van der Poel JJ, Ellegren H, Caldwell RB, Hubbard SJ, Grafham DV, Kierzek AM, McLaren SR, Overton IM, Arakawa H, Beattie KJ, Bezzubov Y, Boardman PE, Bonfield JK, Croning MD, Davies RM, Francis MD, Humphray SJ, Scott CE, Taylor RG, Tickle C, Brown WR, Rogers J, Buerstedde JM, Wilson SA, Stubbs L, Ovcharenko I, Gordon L, Lucas S, Miller MM, Inoko H, Shiina T, Kaufman J, Salomonsen J, Skjoedt K, Wong GK, Wang J, Liu B, Wang J, Yu J, Yang H, Nefedov M, Koriabine M, Dejong PJ, Goodstadt L, Webber C, Dickens NJ, Letunic I, Suyama M, Torrents D, von Mering C, Zdobnov EM, Makova K, Nekrutenko A, Elnitski L, Eswara P, King DC, Yang S, Tyekucheva S, Radakrishnan A, Harris RS, Chiaromonte F, Taylor J, He J, Rijnkels M, Griffiths-Jones S, Ureta-Vidal A, Hoffman MM, Severin J, Searle SM, Law AS, Speed D, Waddington D, Cheng Z, Tuzun E, Eichler E, Bao Z, Flicek P, Shteynberg DD, Brent MR, Bye JM, Huckle EJ, Chatterji S, Dewey C, Pachter L, Kouranov A, Mourelatos Z, Hatzigeorgiou AG, Paterson AH, Ivarie R, Brandstrom M, Axelsson E, Backstrom N, Berlin S, Webster MT, Pourquie O, Reymond A, Ucla C, Antonarakis SE, Long M, Emerson JJ, Betran E, Dupanloup I, Kaessmann H, Hinrichs AS, Bejerano G, Furey TS, Harte RA, Raney B, Siepel A, Kent WJ, Haussler D, Eyras E, Castelo R, Abril JF, Castellano S, Camara F, Parra G, Guigo R, Bourque G, Tesler G, Pevzner PA, Smit A, Fulton LA, Mardis ER, Wilson RK: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432(7018):**695-716.

5. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R, Celera , Holt RA, Adams MD, Amanatides PG, Baden-Tillson H, Barnstead M, Chin S, Evans CA, Ferriera S, Fosler C, Glodek A, Gu Z, Jennings D, Kraft CL, Nguyen T, Pfannkoch CM, Sitter C, Sutton GG, Venter JC, Woodage T, Smith D, Lee HM, Gustafson E, Cahill P, Kana A, Doucette-Stamm L, Weinstock K, Fechtel K, Weiss RB, Dunn DM, Green ED, Blakesley RW, Bouffard GG, De Jong PJ, Osoegawa K, Zhu B, Marra M, Schein J, Bosdet I, Fjell C, Jones S, Krzywinski M, Mathewson C, Siddiqui A, Wye N, McPherson J, Zhao S, Fraser CM, Shetty J, Shatsman S, Geer K, Chen Y, Abramzon S, Nierman WC, Havlak PH, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Li B, Liu Y, Qin X, Cawley S, Worley KC, Cooney AJ, D'Souza LM, Martin K, Wu JQ, Gonzalez-Garay ML, Jackson AR, Kalafus KJ, McLeod MP, Milosavljevic A, Virk D, Volkov A, Wheeler DA, Zhang Z, Bailey JA, Eichler EE, Tuzun E, Birney E, Mongin E, Ureta-Vidal A, Woodwark C, Zdobnov E, Bork P, Suyama M, Torrents D, Alexandersson M, Trask BJ, Young JM, Huang H, Wang H, Xing H, Daniels S, Gietzen D, Schmidt J, Stevens K, Vitt U, Wingrove J, Camara F, Mar Alba M, Abril JF, Guigo R, Smit A, Dubchak I, Rubin EM, Couronne O, Poliakov A, Hubner N, Ganten D, Goesele C, Hummel O, Kreitler T, Lee YA, Monti J, Schulz H, Zimdahl H, Himmelbauer H, Lehrach H, Jacob HJ, Bromberg S, Gullings-Handley J, Jensen-Seaman MI, Kwitek AE, Lazar J, Pasko D, Tonellato PJ, Twigger S, Ponting CP, Duarte JM, Rice S, Goodstadt L, Beatson SA, Emes RD, Winter EE, Webber C, Brandt P, Nyakatura G, Adetobi M, Chiaromonte F, Elnitski L, Eswara P, Hardison RC, Hou M, Kolbe D, Makova K, Miller W, Nekrutenko A, Riemer C, Schwartz S, Taylor J, Yang S, Zhang Y, Lindpaintner K, Andrews TD, Caccamo M, Clamp M, Clarke L, Curwen V, Durbin R, Eyras E, Searle SM, Cooper GM, Batzoglou S, Brudno M, Sidow A, Stone EA, Venter JC, Payseur BA, Bourque G, Lopez-Otin C, Puente XS, Chakrabarti K, Chatterji S, Dewey C, Pachter L, Bray N, Yap VB, Caspi A, Tesler G, Pevzner PA, Haussler D, Roskin KM, Baertsch R, Clawson H, Furey TS, Hinrichs AS, Karolchik D, Kent WJ, Rosenbloom KR, Trumbower H, Weirauch M, Cooper DN, Stenson PD, Ma B, Brent M, Arumugam M, Shteynberg D, Copley RR, Taylor MS, Riethman H, Mudunuri U, Peterson J, Guyer M, Felsenfeld A, Old S, Mockrin S, Collins F: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428(6982):**493-521.

6. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: **Cross-species sequence comparisons: a review of methods and available resources.** *Genome Res* 2003, **13(1):**1-12.

7. Dubchak I, Frazer K: **Multi-species sequence comparison: the next frontier in genome annotation.** *Genome Biol* 2003, **4(12):**122.

8. Heilig R, Eckenberg R, Petit JL, Fonknechten N, Da Silva C, Cattolico L, Levy M, Barbe V, de Berardinis V, Ureta-Vidal A, Pelletier E, Vico V, Anthouard V, Rowen L, Madan A, Qin S, Sun H, Du H, Pepin K, Artiguenave F, Robert C, Cruaud C, Bruls T, Jaillon O, Friedlander L, Samson G, Brottier P, Cure S, Segurens B, Aniere F, Samain S, Crespeau H, Abbasi N, Aiach N, Boscus D, Dickhoff R, Dors M, Dubois I, Friedman C, Gouyvenoux M, James R, Mairey-Estrada B, Mangenot S, Martins N, Menard M, Oztas S, Ratcliffe A, Shaffer T, Trask B, Vacherie B, Bellemere C, Belser C, Besnard-Gonnet M, Bartol-Mavel D, Boutard M, Briez-Silla S, Combette S, Dufosse-Laurent V, Ferron C, Lechaplais C, Louesse C, Muselet D, Magdelenat G, Pateau E, Petit E, Sirvain-Trukniewicz P, Trybou A, Vega-Czarny N, Bataille E, Bluet E, Bordelais I, Dubois M, Dumont C, Guerin T, Haffray S, Hammadi R, Muanga J, Pellouin V, Robert D, Wunderle E, Gauguet G, Roy A, Sainte-Marthe L, Verdier J, Verdier-Discala C, Hillier L, Fulton L, McPherson J, Matsuda F, Wilson R, Scarpelli C, Gyapay G, Wincker P, Saurin W, Quetier F, Waterston R, Hood L, Weissenbach J: **The DNA sequence and analysis of human chromosome 14.** *Nature* 2003, **421(6923):**601-607.

9. Miller W, Makova KD, Nekrutenko A, Hardison RC: **Comparative genomics.** *Annu Rev Genomics Hum Genet* 2004, **5:**15-56.

10. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423(6937):**241-254.

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

12. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED: **Comparative analyses of multi-species sequences from targeted genomic regions.** *Nature* 2003, **424(6950):**788-793.

13. Brudno M, Poliakov A, Salamov A, Cooper GM, Sidow A, Rubin EM, Solovyev V, Batzoglou S, Dubchak I: **Automated whole-genome multiple alignment of rat, mouse, and human.** *Genome Res* 2004, **14(4):**685-692.

14. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in Saccharomyces genomes by phylogenetic footprinting.** *Science* 2003, **301(5629):**71-76.

15. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434(7031):**338-345.

16. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE: **Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs).** *Science* 2003, **302(5647):**1033-1035.

17. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: **Alignment of whole genomes.** *Nucleic Acids Res* 1999, **27(11):**2369-2376.

18. Shih AC, Li WH: **GS-Aligner: a novel tool for aligning genomic sequences using bit-level operations.** *Mol Biol Evol* 2003, **20(8):**1299-1309.
19. Bray N, Dubchak I, Pachter L: **AVID: A global alignment program.** *Genome Res* 2003, **13(1):**97-102.
20. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13(4):**721-731.
21. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302(1):**205-217.
22. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30(14):**3059-3066.
23. Schwartz S, Elnitski L, Li M, Weirauch M, Riemer C, Smit A, Green ED, Hardison RC, Miller W: **MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences.** *Nucleic Acids Res* 2003, **31(13):**3518-3524.
24. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14(4):**708-715.
25. Bray N, Pachter L: **MAVID: constrained ancestral alignment of multiple sequences.** *Genome Res* 2004, **14(4):**693-699.
26. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5):**1792-1797.
27. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5(1):**113.
28. Karplus K, Hu B: **Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set.** *Bioinformatics* 2001, **17(8):**713-720.
29. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ: **OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy.** *BMC Bioinformatics* 2003, **4(1):**47.
30. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB: **Benchmarking tools for the alignment of functional noncoding DNA.** *BMC Bioinformatics* 2004, **5(1):**6.
31. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB: **Correction: Benchmarking tools for the alignment of functional noncoding DNA.** *BMC Bioinformatics* 2004, **5:**73.
32. **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306(5696):**636-640.
33. **ENCODE project** [http://genome.ucsc.edu/ENCODE/encode.html]
34. **UCSC human genome browser** [http://genome.ucsc.edu/cgi-bin/hgGateway]
35. **Ensembl** [http://www.ensembl.org/index.html]
36. **ECR Browser** [http://ecrbrowser.dcode.org/]
37. Ovcharenko I, Nobrega MA, Loots GG, Stubbs L: **ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes.** *Nucleic Acids Res* 2004, **32(Web Server):**W280-286.
38. **VISTA** [http://genome.lbl.gov/vista/index.shtml]
39. **mVISTA** [http://genome.lbl.gov/vista/mvista/submit.shtml]
40. Loots GG, Ovcharenko I: **rVISTA 2.0: evolutionary analysis of transcription factor binding sites.** *Nucleic Acids Res* 2004, **32(Web Server):**W217-221.
41. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12(5):**832-839.
42. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I: **Strategies and tools for whole-genome alignments.** *Genome Res* 2003, **13(1):**73-80.
43. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics.** *Nucleic Acids Res* 2004, **32(Web Server):**W273-279.
44. Shah N, Couronne O, Pennacchio LA, Brudno M, Batzoglou S, Bethel EW, Rubin EM, Hamann B, Dubchak I: **Phylo-VISTA: interactive visualization of multiple DNA sequence alignments.** *Bioinformatics* 2004, **20(5):**636-643.
45. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker – a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10(4):**577-586.
46. Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L: **zPicture: dynamic alignment and visualization tool for analyzing conservation profiles.** *Genome Res* 2004, **14(3):**472-477.
47. **EMBOSS** [http://emboss.sourceforge.net/]
48. **YASS** [http://yass.loria.fr/interface.php]
49. **ATVtree** [http://www.genetics.wustl.edu/eddy/atv/]
50. **Computational Genomics Lab** [http://biocomp.iis.sinica.edu.tw/]
51. Gottgens B, Barton LM, Chapman MA, Sinclair AM, Knudsen B, Grafham D, Gilbert JG, Rogers J, Bentley DR, Green AR: **Transcriptional regulation of the stem cell leukemia gene (SCL) – comparative analysis of five vertebrate SCL loci.** *Genome Res* 2002, **12(5):**749-759.
52. Barton LM, Gottgens B, Green AR: **The stem cell leukaemia (SCL) gene: a critical regulator of haemopoietic and vascular development.** *Int J Biochem Cell Biol* 1999, **31(10):**1193-1207.
53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
54. Lenhof HP, Morgenstern B, Reinert K: **An exact solution for the segment-to-segment multiple sequence alignment problem.** *Bioinformatics* 1999, **15(3):**203-210.
55. McCarthy VA, Harris A: **The CFTR gene and regulation of its expression.** *Pediatr Pulmonol* 2005.
56. **NIH Intramural Sequencing Center (NISC)** [http://www.nisc.nih.gov/data/20020612_Target1_0051/]
57. Li W-H: **Molecular Evolution.** Sunderland, MA: Sinauer Press; 1997.
58. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity.** *Nat Rev Genet* 2002, **3(5):**370-379.
59. Smit AF, Green P: **RepeatMasker.** [http://ftp.genome.washington.edu/].
60. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S: **Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.** *Science* 2002, **297(5585):**1301-1310.