Phylo-VISTA: Interactive Visualization of Multiple DNA Sequence Alignments

Nameeta Shah[1,*], Olivier Couronne[2,*], Len A. Pennacchio[2], Michael Brudno[3], Serafim Batzoglou[3], E. Wes Bethel[2], Edward M. Rubin[2], Bernd Hamann[1, 2] and Inna Dubchak[2]

[1]University of California, Davis, [2]Lawrence Berkeley National Laboratory, [3]Stanford University
[*]Corresponding authors

Summary

Motivation

The power of multi-sequence comparison for biological discovery is well established. The need for new capabilities to visualize and compare cross-species alignment data is intensified by the growing number of genomic sequence datasets being generated for an ever-increasing number of organisms. To be efficient these visualization algorithms must support the ability to consistently accommodate a wide range of evolutionary distances in a comparison framework based upon phylogenetic relationships.

Results

We have developed Phylo-VISTA, an interactive tool for analyzing multiple alignments by visualizing a similarity measure for multiple DNA sequences. The complexity of visual presentation is effectively organized using a framework based upon interspecies phylogenetic relationships . The phylogenetic organization supports rapid, user-guided interspecies comparison. To aid in navigation through large sequence datasets, Phylo-VISTA leverages concepts from VISTA that provide a user with the ability to select and view data at varying resolutions. The combination of multiresolution data visualization and analysis, combined with the phylogenetic framework for interspecies comparison, produces a highly flexible and powerful tool for visual data analysis of multiple sequence alignments.

Availability

Phylo-VISTA is available at http://www-gsd.lbl.gov/phylovista. It requires an Internet browser with Java Plug-in 1.4.2 and it is integrated into the global alignment program LAGAN at http://lagan.stanford.edu.

Contact

phylovista@lbl.gov

1 Introduction

Large-scale genome sequencing efforts are producing an abundance of sequence data for an increasing number of organisms. Comparative analysis of DNA sequences from multiple species is a powerful strategy for identifying functional elements such as genes and their regulatory sequences (Frazer et al., 2003). This approach is based on the assumption that functionally important elements evolve more slowly than nonfunctional genomic regions due to selective constraints. For instance, the comparison of relatively distant species, such as human and mouse, has revealed conservation among a significant fraction of mammalian genes and other functional elements in these organisms (Waterston et al., 2002). In addition, "*phylogenetic shadowing"* (Boffelli et al., 2003) has led to the discovery of primate-specific regulatory elements by extensive sequence comparisons of numerous primate species. Several efforts are ongoing to sequence and analyze targeted genomic regions for conservation across many evolutionarily diverse species (for example, for human, chimp, baboon, mouse, rat, cow, pig, dog, cat, chicken, pufferfish and zebrafish (Cooper et al., 2003)).

Recent developments in local and global alignment methods have allowed scientists to perform genomic comparisons between multiple species on a megabase scale. BLASTZ (Schwartz et al., 2003a) and PatternHunter (Ma et al., 2002) are local alignment techniques that can be used for the comparison of whole vertebrate genome assemblies (Waterston et al., 2002). In addition, efficient global alignment programs such as LAGAN (Brudno et al., 2003), AVID (Bray et al., 2003), and Mummer (Delcher et al., 2002) provide pairwise global comparison of very large genomic regions. LAGAN and AVID also produce multiple alignments of megabase-scale sequences (Brudno et al., 2003; Bray and Pachter, 2003). Recently developed computational schemes use a combined local/global alignment method to quickly identify all regions of homology between two entire genomes and provide global alignment of these sequences (Couronne et al., 2003).

Several publicly available tools exist for visualization of long pairwise DNA alignments. PIPMaker (Schwartz et al., 2000; Elnitski et al., 2002) generates a highly detailed plot of a local alignment as a series of dots and dashes representing the degree of conservation between the base and a second orthologous sequence. VISTA (Dubchak et al., 2000; Mayor et al., 2000) presents comparative data in the form of a curve to display the level of sequence conservation in a predefined window of a global alignment. SynPlot (Göttgens et al., 2001) also utilizes a global alignment and a curve plot, but each visualization technique is presented in a different display. All three tools are intended for visualizing pairwise alignments as well as multiple pairwise alignments on the same scale.

An important consideration in multiple species sequence alignment is phylogeny. Phylogenetic trees have been used extensively in creating alignments. For instance, progressive pairwise alignment techniques use a precomputed phylogenetic tree as a “guide” to indicate the order in which multiple sequences should be aligned (Brudno et al., 2003). While there are tools for visualizing phylogenetic trees and calculating trees based on an alignment (see http://evolution.genetics.washington.edu/phylip.html), no tool exists for visualizing sequence alignment data while taking phylogeny of the sequences into account.

Although global alignment algorithms have pitfalls in working with regions containing small duplications and inversions, the first local multiple alignment algorithms have just appeared (Schwartz et al., 2003b) and large-scale comparison is yet to be done. Still lacking are algorithms for

visualization and analysis of multiple aligned sequences to support conservation analysis across species. Furthermore, the need for algorithms to universally incorporate a wide range of evolutionary distances creates a substantial challenge.

Our work is motivated by the need to perform visual data analysis of sequence data across an arbitrary number of species, along with the need to perform such analysis at varying levels of resolution to accommodate the explosive growth in the size of sequence data. Our system, called "Phylo-VISTA" (short for Phylogenetic VISTA) supports visualization of multi-species sequence comparison using phylogenetic trees as a framework to guide the display and analysis of conservation levels across tree nodes. Phylo-VISTA supports interactive visual analysis of pre-aligned multi-species sequences by performing the following functions: (1) display of a multiple alignment sequences with the associated phylogenetic tree; (2) computation of a similarity measure over a user-specified window for any node of the tree; (3) visualization of the degree of sequence conservation by a line plot; and (4) presentation of comparative data together with annotations.

2 Approach

We used the successful VISTA concept (Dubchak et al., 2000; Mayor et al., 2000) as basis for the visualization of multiple alignments along with an associated phylogenetic tree. For pairwise comparison, VISTA requires a user to select one of the sequences as the *base sequence*. To create a VISTA plot, the user moves a window over an alignment, and VISTA calculates the percent-identity between the base sequence and the aligned sequence over a window surrounding each basepair. The x-axis represents the base sequence, and the y-axis represents percent-identity. The alignment data is projected on the base sequence, and annotations are also presented in the plots. VISTA displays the size and location of gaps in the aligned sequence. Loss of information about the gaps in the base sequence and corresponding data of other sequences results by using one sequence as "the" base. For multiple alignment, VISTA produces plots for each sequence against the base sequence. One of the main limitations of such multiple-pairwise plots is that it's designed only to detect regions that are conserved in the base sequence. Therefore, Phylo-VISTA uses the entire multiple alignment as a base. As a result, Phylo-VISTA is also capable of displaying location and length of gaps in all sequences. In addition, Phylo-VISTA provides annotations beyond those associated with a single base sequence. Multi-species plots allow a user to analyze desirable features in a single visualization (e.g., to view and analyze gaps and annotations of all sequences being compared). A sum of weighted pairwise similarity measures is used for comparing more than two sequences.

3 Phylo-VISTA Scoring Method

Phylo-VISTA assumes that the given data is a multiple alignment file in *multi-fasta* format. In addition, it utilizes as input the phylogenetic tree (Figure 1.A) that was applied in the progressive alignment phase of a multiple alignment algorithm such as in Clustalw (Thompson et al., 1994) or generated for the alignment after it was built (Swofford et al., 1996). Currently, Phylo-VISTA requires as input a pairwise phylogenetic tree (e.g. ((human mouse) chicken)) which is used for generating similarity plots and computing the similarity measure.

Phylo-VISTA aims to highlight the similarity of genomic sequences over an entire phylogeny. Navigating through a phylogenetic tree allows a user to identify regions conserved among subtrees within a multiple alignment, for example, the subtree associated with mammals. Consequently, we have adopted a scoring scheme that takes into account similarity across nodes of a given rooted

phylogenetic tree. Each leaf node in the Phylo-VISTA tree represents a sequence in the alignment. Each internal node corresponds to a similarity plot. This plot indicates the average percent-identity over a window between sets of sequences from the left and right subtrees of the node. Similarity between sequences from the same subtree is ignored. More formally, the similarity value of a node X at position k in the alignment is defined as:

$$S_k = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} B_{i,j} D_{i,j}}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} B_{i,j}},$$

where

S_k is the similarity at the k^{th} position in the alignment,
n is the number of leaf nodes that are descendents of node X,
$B_{i,j}$ is the Boolean value for sequence pairs i and j, and
$D_{i,j}$ is the distance between sequences i and j at the kth position.

The $D_{i,j}$ value is defined as

$$D_{i,j} = \begin{cases} 1, & \text{if sequences i and j have the same base pair at position k in the alignment} \\ 0, & \text{otherwise} \end{cases}.$$

The Boolean values $B_{i,j}$ are defined as

$$B_{i,j} = \begin{cases} 0, & \text{if there is a path from i to j that does not include X} \\ 1, & \text{otherwise} \end{cases}.$$

We consider an example involving three species: human, mouse, and chicken. A phylogenetic tree is shown in Figure 1.A.

Position	*1 2 3 4 5...*
Human	TAA-C...
Mouse	GAAA-...
Chicken	TAT--...

The similarity measure for the node human-mouse-chicken at position one is

$$S_1 = \frac{B_{\text{human, mouse}} D_{\text{human, mouse}} + B_{\text{human, chicken}} D_{\text{human, chicken}} + B_{\text{mouse, chicken}} D_{\text{mouse, chicken}}}{B_{\text{human, mouse}} + B_{\text{human, chicken}} + B_{\text{mouse, chicken}}}.$$

Since human and mouse are in the same subtree, there exists a path between them in the phylogenetic tree, and therefore $B_{\text{human, mouse}}$ is zero. As every path from human to chicken and mouse to chicken includes the human-mouse-chicken node, $B_{\text{human, chicken}}$ and $B_{\text{mouse, chicken}}$ are both equal to one. In the example above, for the first position, the values of $D_{\text{human, mouse}}$ and $D_{\text{mouse, chicken}}$ are zero, and the value of $D_{\text{human, chicken}}$ is one. Thus, the value of S_1 is 0.5. Similarly, S_2 has the value one, and S_3 has the value zero.

This scoring scheme was motivated by the desire to produce similarity plots not dominated by conservation in closely related species. In the above example, the human-mouse-chicken similarity plot will not be dominated by human-mouse similarities. Human-mouse similarities are shown in the plot corresponding to the human-mouse node in the tree. Thus, we compute similarity between different subtrees. The time complexity for the computation of this similarity measure is $O(N^2)$, where N is the number of sequences. Our implementation takes $O(N)$ time by exploiting the fact that sequence data consists only of a small number of characters.

4 Components

The Phylo-VISTA layout consists of four main components (Figure 1), which are described in more detail in the next subsections.

4.1 Phylogenetic Tree

Phylogenetic tree is provided as an input to Phylo-VISTA along with multiple sequence alignment data. Figure 1.A shows a sample phylogenetic tree used for the alignment of five sequences (human, mouse, chicken, pufferfish, and zebrafish). In Phylo-VISTA, each internal node (shown in black) represents a similarity plot for all the sequences that are descendents of that node. Thus, peaks in the plot indicate regions of the ancestral sequence conserved among its descendents. A user can select a subset of internal nodes to choose which alignments to show.

4.2 Sequence Traversal Panel

This panel contains a traversal bar for each of the sequences, and an additional global bar for the alignment (Figure 1.B). The red rectangle indicates the currently selected region of each of the sequences. A user can move and resize the rectangle on the bar of the sequence of interest, and choose the size of the region for generating plots. When selecting a region in one sequence, the corresponding aligned regions in the other sequences are selected automatically (Figure 1.B). User-supplied annotations are displayed above the bar. Below the bar of each sequence, a narrow strip shows how the sequence is distributed across the alignment.

4.3 Similarity Plots

A similarity plot visually represents conservation among a given set of sequences based on the similarity measure described in Section 3. Similarity plots are defined for every selected node in the tree (Figure 1.B). The x-axis represents the alignment at the selected node, and the y-axis represents percent-similarity. A user selects a subregion of the alignment data using sequence traversal panel. In the selected region, Phylo-VISTA computes the similarity score for each basepair within the region. User-supplied annotations for all the sequences along with the gaps are displayed beneath each plot. Gaps are shown as gray rectangles. When gaps exist in all the sequences for a given plot the entire plot area is shaded in gray. As the x-axis represents the alignment, rather than actual sequences, the basepair number is shown for all sequences on the left-hand side of the plot. The plots can be viewed at varying resolution to facilitate visualization of sequences of arbitrary lengths (Figure 2).

4.4 Text Window

The "Text window" allows a user to view a selected region of alignment in text format. The text is color-coded such that conserved DNA sequence motifs are highlighted. Black represents base pairs that are similar in all sequences in the alignment (Figure 1.C).

5 Complexity

To support interactive visualization, Phylo-VISTA must be computationally efficient. The time-critical step involves the computation of sequence similarity. As mentioned in Section 3, the similarity calculation at each base pair takes *O(N)* time, where N is the number of sequences. The algorithm for calculating a similarity plot for an entire alignment thus requires *O(NL)* time, where L is the length of the alignment. The maximum number of plots is the number of internal nodes in the phylogenetic tree, which is N-1. In practice, the number of plots is limited by the display area and can be considered constant. Phylo-VISTA stores an entire multiple alignment dataset in memory. Thus, Phylo-VISTA has *O(NL)* time complexity and *O(NL)* memory complexity.

6 Example: Analysis of Multiple Alignment of the Stem Cell Leukemia Region

To demonstrate the use of Phylo-VISTA for multi-species DNA sequence alignments, we have examined the stem cell leukemia (SCL) gene interval (Göttgens et al. 2002). The SCL gene encodes a transcription factor that plays a crucial role in the formation and development of blood cells in bone marrow (hematopoiesis) and in embryonic formation and differentiation of the vascular system (vasculogenesis) . The expression pattern of this gene is highly conserved throughout vertebrates, from mammals to teleost fish . Previous comparative analysis of five vertebrate SCL loci (considering human, mouse, chicken, pufferfish, and zebrafish) revealed five DNA sequence motifs in the SCL promoter/enhancer that are conserved in all five species. These five conserved motifs are known to be essential for the appropriate expression pattern of SCL (Göttgens et al. 2002 and references therein).

We have applied Phylo-VISTA on a LAGAN (Brudno et al., 2003) multiple alignment of the SCL region, consisting of orthologous sequences corresponding to approximately 100 kb (kilo-basepairs) of human, 65 kb of mouse, 22 kb of chicken, 8 kb of pufferfish, and 67 kb of zebrafish within the SCL region. After aligning all five species, the length of the resulting multiple alignment equaled approximately 150 kb. As an exercise to show the utility of Phylo-VISTA, we extracted regulatory motifs in the promoter region of the SCL gene from this multiple alignment data. Figure 2.A shows the similarity plot of the entire alignment for the node human-mouse-chicken-pufferfish-zebrafish in the phylogenetic tree shown in Figure 1.A. The annotations for all the sequences are shown below the plot. Blue rectangles indicate exons and gray rectangles indicate gaps. Figure 2.B shows the Phylo-VISTA result obtained when zooming in on the region with peaks (highlighted by an oval in Figure 2.A). A peak (shown by an oval in Figure 2.B) is visible in front of exon 1 (promoter region), and this peak indicates conservation in the promoter region among all species. By repeatedly applying the zoom operator, the plot shown in Figure 2.C is obtained. Reducing the sliding window width yields the similarity plot shown in Figure 2.D. The size of the selected region consists of only 39 basepairs, and sequence motifs in the text window can be examined.

Figure 1.C shows a part of the conserved promoter/enhancer region of all sequences in text format. The basepairs that are conserved in all sequences are highlighted in black. The highlighted motif

AATGAATCATTT is a known SKN-1 cis-regulatory site (Göttgens et al. 2002). The other two motifs GCCAAAT (CS1, Conserved Sequence 1) and ATAATGG (CS2, Conserved Sequence 2) were identified in earlier comparative analysis efforts (Göttgens et al. 2002). All three motifs are known to be binding sites for transcription factors responsible for regulating the expression of SCL (Göttgens et al. 2002).

We stress that Phylo-VISTA is very convenient for biological analysis since it simultaneously presents pattern of conservation among all subtrees of species on the same scale. Analysis of conservation of the same SCL loci described in the study of Göttgens (Göttgens, 2002) provides a good example for comparison. In order to visualize conservation pattern across five species (human, mouse, chicken, pufferfish, and zebrafish) as well as subsets of those five species, three individual plots are presented: (i) multi-pairwise VISTA plots with the mouse sequence serving as reference; (ii) VISTA plots of three species (chicken, pufferfish and zebrafish) with the chicken sequence serving as a reference; and (iii) a SynPlot showing conservation between pufferfish and zebrafish. In general, the number of possible subsets for a given set of sequences is exponential. In practice, it is not necessary to examine all possible subsets. Phylo-VISTA automates the selection of these subsets by using a pairwise phylogenetic tree as a guide and thus limits the number of subsets to N-1, where N is the number of sequences. For the same SCL example and a given phylogenetic tree, Phylo-VISTA allows a user to look at four plots (human-mouse-chicken-pufferfish-zebrafish, human-mouse-chicken, human-mouse, pufferfish-zebrafish) simultaneously at the same scale. A user can compare subsets of sequences using a "one-step technique" to obtain the same information concerning conservation among the five species. Therefore, Phylo-VISTA supports a much more direct and efficient analysis.

7 Conclusions and Future Work

Phylo-VISTA is a new interactive visualization and analysis tool for aligned genome sequences for multiple species. Its novel capabilities include:

1. visualization of multiple alignments at varying levels of resolution;
2. simultaneous visualization of alignments of different subsets of given sequences at the same scale, where subsets are defined by the internal nodes on the phylogenetic tree;
3. interactive specification of processing and visualization parameters, like sliding window width and percent-similarity cutoff;
4. simultaneous display of gaps and gene annotations for all sequences in a multiple alignment; and
5. multiresolution browsing that allows a user to begin with visualizing several thousand basepairs as a similarity plot and then drill down to few basepairs that can be viewed in a text format.

Phylo-VISTA is modular and can accommodate different similarity measures. For example, the Boolean values in the current measure can be substituted by weights based on evolutionary distance between species. We plan to integrate Phylo-VISTA with a search engine for transcription factor binding sites. Limited display area and limited display resolution are physical restrictions we must consider when developing interactive sequence data exploration methods for the comparison of several hundred sequences, each one consisting of several million basepairs. We plan to develop additional innovative visualization techniques to address these issues.

References

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. and Rubin, E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome, *Science*, **299**, 1391-1394.

Bray, N., Dubchak, I. and Pachter, L. (2003) AVID: A Global Alignment Program, *Genome Research*, **13**, 97-102.

Bray N. and Pachter L. (2003) MAVID multiple alignment server, *Nucleic Acids Research*, **31**, 3525-3526.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., NISC Comparative Sequencing Program, Green, E.D., Sidow, A. and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA, *Genome Research*, **13**, 721-731.

Cooper, G.M., Brudno, M., Green, E.D., Batzoglou, S., Sidow, A. (2003) Estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Research*, **13(5)**, 813-820.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E.M., Pachter, L. and Dubchak, I. (2003) Strategies and Tools for Whole Genome Alignments, *Genome Research*, **13**, 73-80.

Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Research*, **30**, 2478-2483.

Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M. and Frazer, K.A. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons, *Genome Research***, 10**, 1304-1306.

Elnitski, L., Riemer, C., Petrykowska, H., Florea, L., Schwartz, S., Miller, W. and Hardison, R. (2002) PipTools: A Computational Toolkit to Annotate and Analyze Pairwise Comparisons of Genomic Sequences, *Genomics*, **80**, 681-690.

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 Research***, 13**, 1-12.

Göttgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R. and Green, A.R. (2001) Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences, *Genome Research*, **11**, 87-97.

Göttgens, 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 Research*, **12**, 749-759.

Ma, B., Tromp, J. and Li, M. (2002) PatternHunter: faster and more sensitive homology search, *Bioinformatics*, **18**, 440-445.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L. and Dubchak, I. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length, *Bioinformatics***, 16**, 1046-1047.

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 Research*, **10**, 577-586.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-Mouse Alignments with BLASTZ, *Genome Research*, **13**, 103-107.
Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., Miller, W. (2003) MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.*, **31**, 3518-24.

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

Waterston, R.H. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**,520-562.

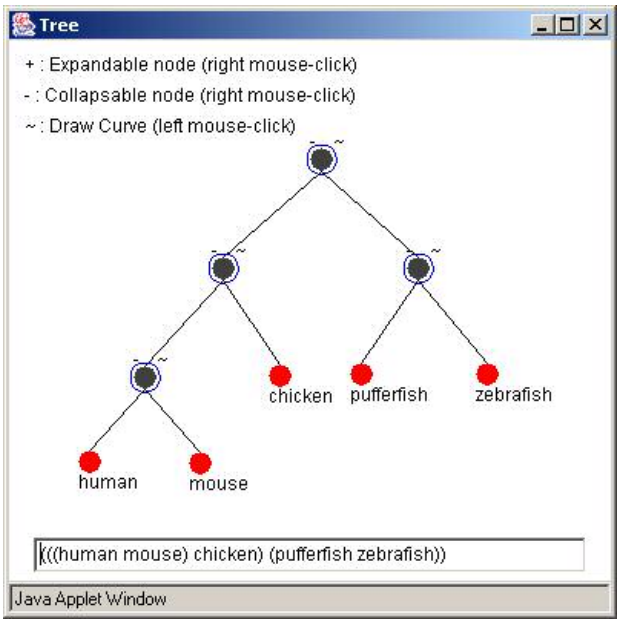
Acknowledgements

We are grateful to biologists from the Genome Sciences Department of Lawrence Berkeley National Laboratory for their support and inspiration and to Alex Poliakov for helpful discussions.

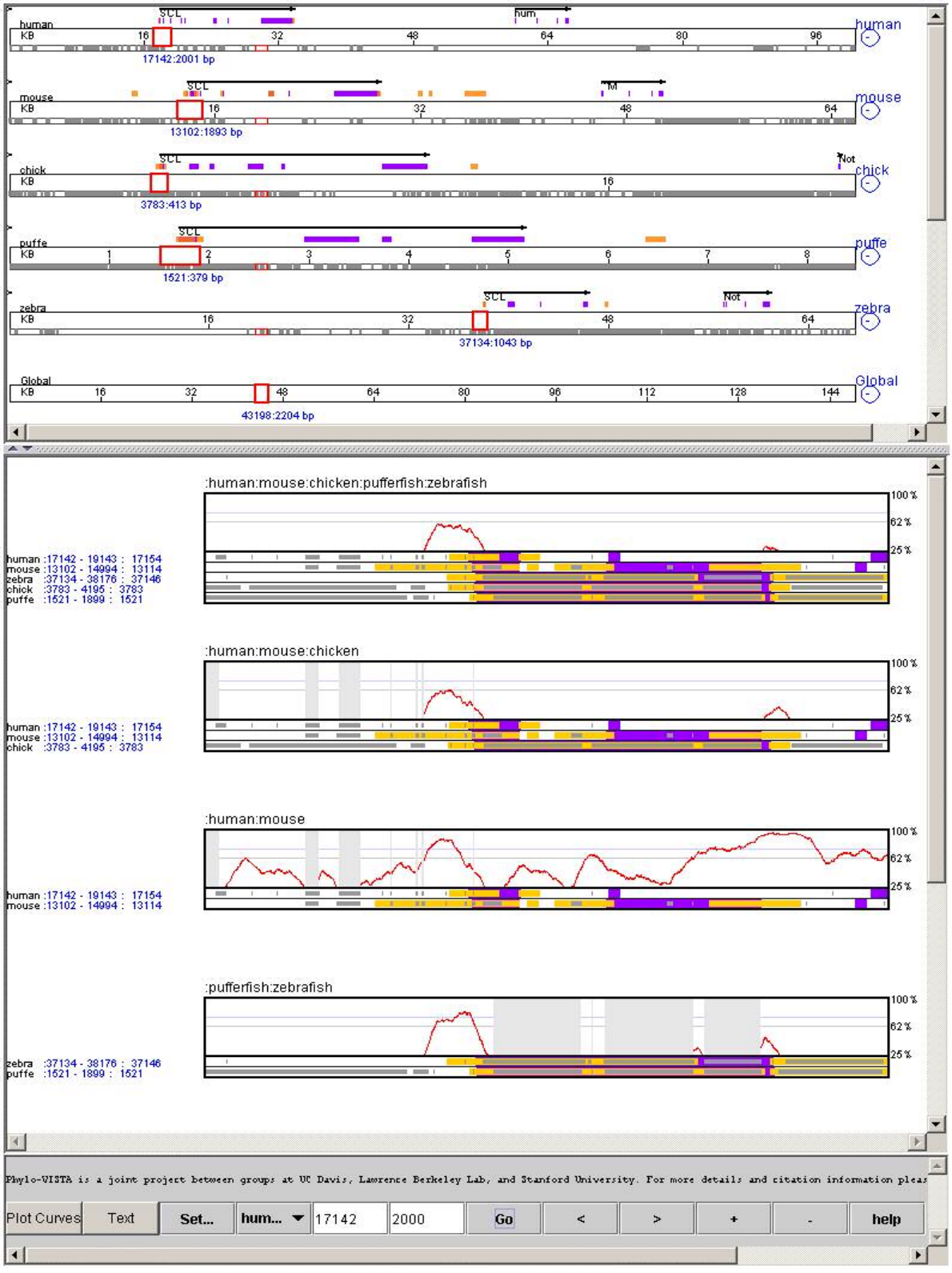
This work was supported through the Laboratory Directed Research and Development (LDRD) program at Lawrence Berkeley National Laboratory under the grant of the US Department of Energy contract DE-AC03-76SF00098. The project was performed in collaboration between the Center for Image Processing and Integrated Computing (CIPIC), University of California, Davis and the Genome Sciences Department and the National Energy Research Scientific Computing (NERSC) center, Lawrence Berkeley National Laboratory. In part, the project was also supported by the National Science Foundation under contract ACI 9624034 (CAREER Award), through the Large Scientific and Software Data Set Visualization (LSSDSV) program under contract ACI 9982251, and through the National Partnership for Advanced Computational Infrastructure (NPACI). The project was partially supported by a Program for Genomic Applications grant from the National Heart Lung and Blood Institute. Michael Brudno was supported by an NSF graduate fellowship.

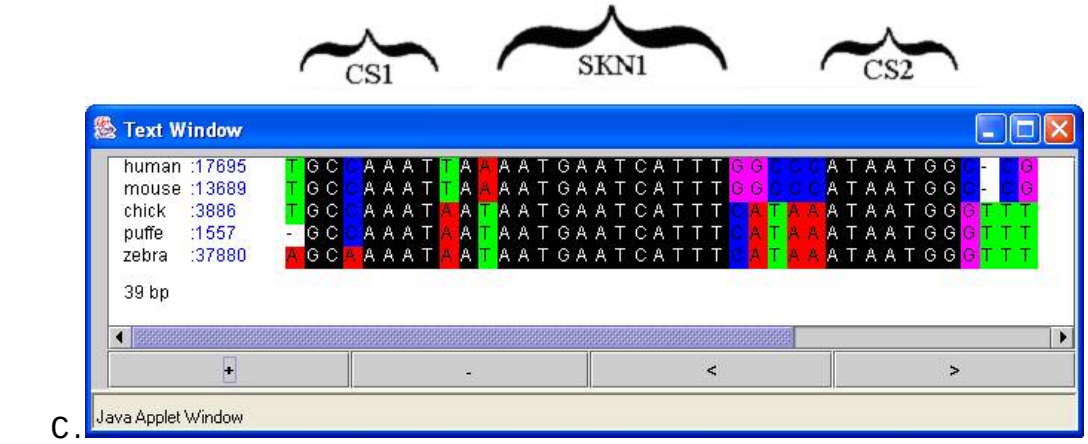
Captions:

Figure 1. Phylo-VISTA output
Figure 2. Visualization of a multiple alignment dataset, consisting of human, mouse, chicken, pufferfish, and zebrafish data - stem cell leukemia (SCL) regions being analyzed

A. 
Tree
+ : Expandable node (right mouse-click)
- : Collapsable node (right mouse-click)
~ : Draw Curve (left mouse-click)
chicken
pufferfish
zebrafish
human
mouse
((human mouse) chicken) (pufferfish zebrafish))
Java Applet Window

B.


human
KB
17142:2001 bp
mouse
13102:1893 bp
chick
3783:413 bp
puffe
1521:379 bp
zebra
37134:1043 bp
Global
43198:2204 bp
:human:mouse:chicken:pufferfish:zebrafish
:human:mouse:chicken
:human:mouse
:pufferfish:zebrafish
100%
62%
25%
Phylo-VISTA is a joint project between groups at UC Davis, Lawrence Berkeley Lab, and Stanford University. For more details and citation information pleas
Plot Curves
Text
Set...
hum...
17142
2000
Go
<
>
+
-
help



C.

Figure 1. Phylo-VISTA output.

Visualization of the alignment of about 100,000 basepairs of human stem cell leukemia (SCL) region, considering mouse, chicken, pufferfish, and zebrafish sequences.

A. Phylogenetic tree.
   In this pairwise phylogenetic tree, all sequences in the alignment are represented by red leaf nodes. Each black node represents a similarity plot for all the descendent leaf nodes. The selected node is circled, representing a similarity plot for human, mouse, and chicken.

B. Similarity plots for the selected nodes of the phylogenetic tree in part B.
   The sequence traversal panel for the alignment of SCL regions in human, mouse, chicken, pufferfish, and zebrafish sequences. A bar is shown for each sequence. A red rectangle shows the selected region in each sequence. A black arrow on the top of each bar indicates a gene. The blue rectangles are exons, and yellow rectangles show conserved features supplied by a user. These annotations show that the selected region is upstream of the SCL gene in all sequences. The numbers below each bar denote the starting position and the size of the selected region in the corresponding sequence. For example, the starting position in the human sequence is 17142, and the size of the selected region is 2001 basepairs. A narrow strip below each bar shows the distribution of the sequence on the alignment scale. The initial part of the zebrafish sequence does not align with any other sequence, leading to the gaps in all the other sequences in the initial part of the alignment.
   Similarity plots corresponding to the black nodes of the phylogenetic tree are shown in A. The height in the line plot corresponds to percent-similarity. Minimum conservation is set to 25%. Below the line plots the annotations for each sequence are provided. Gray rectangles indicate gaps, blue rectangles represent exons, and yellow rectangles show user-supplied conserved features. The plot area shaded in gray indicates the presence of gaps in all the sequences of that plot. The text on the left side of the annotations shows the name of the sequence, its selected start and end positions, and the current cursor position. The peak visible in all plots corresponds to a region conserved in all sequences.

C. Part of the alignment in color-coded text.
   The text window shows the selected part of the sequences in text format. Each basepair is shown in a different color. The window shows the starting position of the selected region in a sequence. Basepairs conserved in all sequences are highlighted in black. This figure shows the promoter/enhancer of the SCL gene. Three conserved motifs (CS1, SKN-1 and CS2) are highlighted. These three motifs are binding sites for transcription factors that are known to be

essential for the appropriate expression pattern of SCL. The SKN-1 motif is a known binding site. The CS1 and CS2 motifs were discovered by using multiple alignment.

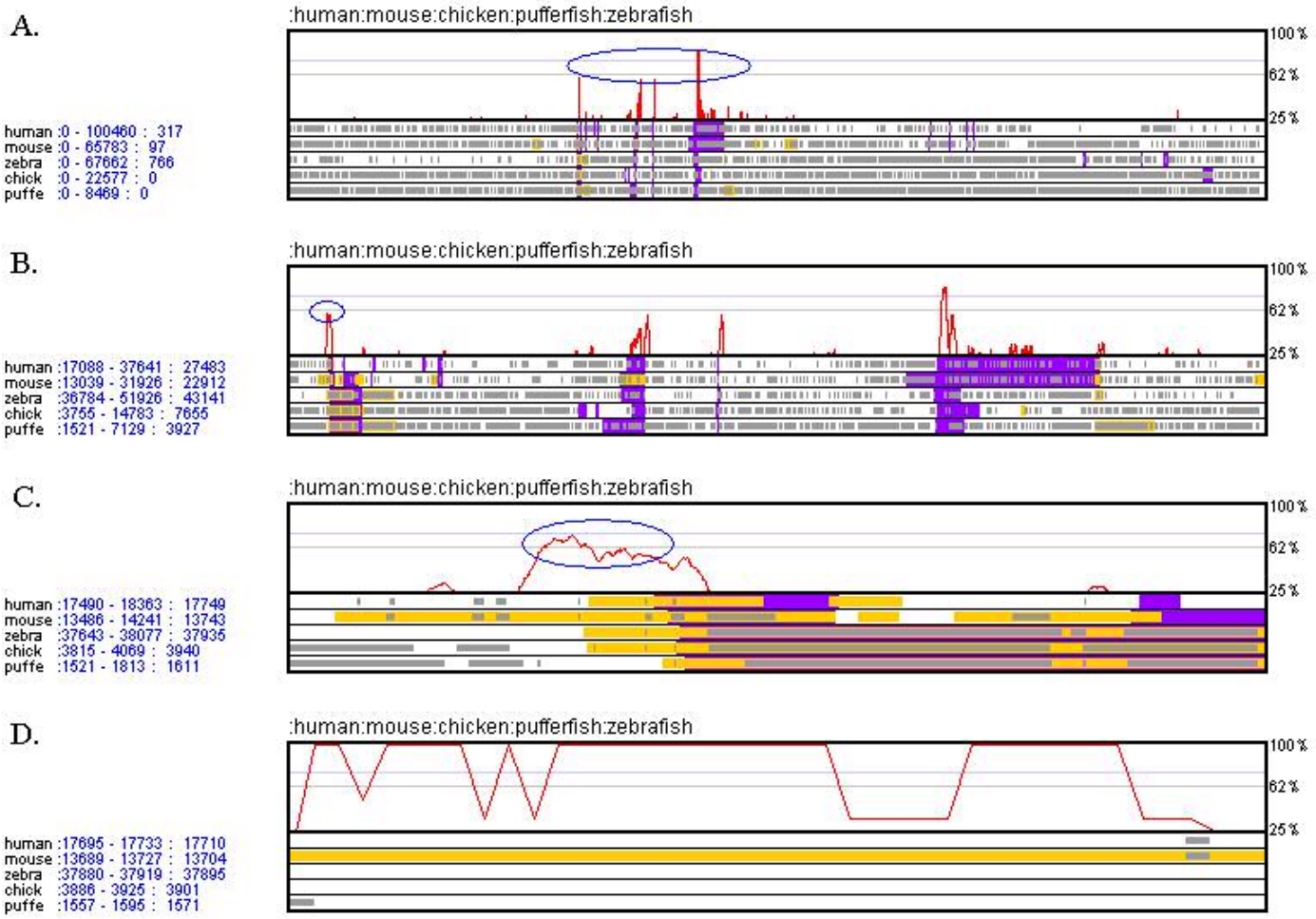


Figure 2. Visualization of a multiple-alignment data set consisting of human, mouse, chicken, pufferfish, and zebrafish SCL regions.

A. Bird's eye view of the alignment consisting of about 150,000 basepairs. Peaks indicate conservation. The region selected for applying the zoom operator is shown by an oval.
B. A peak (oval) exists upstream of exon 1 of the SCL gene. The zoom operator is applied to the peak.
C. Region without gaps selected for zooming.
D. Conserved region seen at high resolution, using a window width of one. This plot documents that motifs are conserved at a level of 100%. The corresponding text is shown in Figure 1.C.