Phylo-VISTA: An Interactive Visualization Tool for 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], Inna Dubchak[2]

[1]University of California at Davis, [2]Lawrence Berkeley National Laboratory, [3]Stanford University

[*]Corresponding author

Summary

Motivation.

The power of multi-sequence comparison for biological discovery is well established. As sequence data from a growing list of organisms is generated and multi-alignment software for large sequences are becoming available there is a strong need for computational strategies to visually analyze comparison across various species. To be efficient these visualization algorithms require the ability to universally handle a wide range of evolutionary distances in the framework of their phylogeny relationship.

Results. We have developed Phylo-VISTA, an interactive tool for analyzing multiple alignments by visualizing the similarity of DNA sequences among multiple species while considering their phylogenic relationships. Features include a broad spectrum of resolution parameters for examining the alignment and the ability to easily compare any subtree of sequences within a complete alignment dataset. Phylo-VISTA uses VISTA concepts that have been successfully applied previously to a wide range of comparative genomics data analysis problems.

Availability

Phylo-VISTA is an interactive java applet available for downloading at http://graphics.cs.ucdavis.edu/~nyshah/Phylo-VISTA.

It is also available on-line at http://www-gsd.lbl.gov/phylovista and is integrated with the global alignment program LAGAN at http://lagan.stanford.edu.

Contact

phylovista@lbl.gov

1. Introduction

Large-scale genome sequencing efforts have produced an abundance of sequence data for a growing list 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 (Loots et al., 2000; Dubchak et al., 2000; Göttgens et al., 2001; Pennacchio et al., 2001; Frazer et al., 2003). This approach is based on the assumption that functionally important elements evolve more slowly than nonfunctional genomic regions. 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 (Hardison et al., 1997; Loots et al., 2000; Batzoglou et al., 2000; Pennacchio et al., 2001, Waterston et al., 2002). In addition, "*phylogenetic shadowing*" (Boffelli et al., 2003) has led to the discovery of primate-specific regulatory elements by deep 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, mouse, chicken, pufferfish, and zebrafish (Göttgens et al., 2002) and for human, chimp, baboon, mouse, rat, cow, pig, dog, cat, chicken, pufferfish and zebrafish (Brudno 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., 2003) and PatternHunter (Ma el al., 2002) are local alignment techniques applied to the comparison of whole vertebrate genome assemblies (Waterston et al., 2002). In addition, efficient global alignment programs such as Mummer (Delcher et al., 2000, 2002), AVID (Bray et al., 2003), and LAGAN (Brudno et al., 2003) provide pair-wise global comparison of very large genomic regions. AVID and LAGAN also produce multiple alignments of megabase-scale sequences (http://baboon.math.berkeley.edu/mavid/; Brudno et al., 2003). Recently developed computational schemes use a combined local/global alignment method to quickly identify all regions of homology between several entire genomes and provide a detailed alignment of these sequences (Couronne et al., 2003). 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.

Several publicly available visualization tools exist for long pair-wise 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 levels 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 in a different display. All three tools can be used for visualizing pairwise alignments as well as multiple pairwise alignments on the same scale (examples are provided in Elnitski et al., 2002, Göttgens et al., 2001, Göttgens et al., 2002, Frazer et al., 2003).

An important consideration during multiple species sequence alignments 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 (Thompson et al., 1994, Brudno et al., 2003, Edgar and Sjolander, 2003). Phylogenetic trees are also useful for calculating proper substitution matrices for an alignment (Henikoff and Henikoff, 1992) and in regulatory element identification (Blanchette and Tompa, 2002).

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 phylogenetic trees into account.

We have developed a tool called "Phylo-VISTA" (short for Phylogenetic VISTA) to address this need. Phylo-VISTA provides visualization of multi-species sequence comparison by using phylogenetic trees as a guide to display and analyze the level of conservation across tree nodes.

Phylo-VISTA supports interactive visual analysis of prealigned multi-species sequences by performing the following functions: (1) display of a multiple alignment with the associated phylogenetic tree; (2) computation of a measure of similarity 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 available annotations.

2. Approach

We applied the successful VISTA concepts (Dubchak et al., 2000; Mayor et al., 2000) to the visualization of multiple alignments considering an associated phylogenetic tree. In order to achieve this goal we developed several extensions to VISTA. For pairwise comparison, VISTA requires a user to select one of the sequences as the *base sequence*. A VISTA plot is created by moving a window over an alignment and calculating 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. Annotations for a base sequence are presented in the plots as well. VISTA displays the size and location of gaps in the aligned sequence, but for the base sequence only their location can be displayed. Thus, using one sequence as a base results in loss of information. Therefore, Phylo-VISTA uses the entire multiple alignment as a base in the x-axis. Similar realization was used in Synplot for pairwise alignments (Göttgens et al., 2001). As a result, the tool is capable of displaying location and length of gaps in all sequences. In addition, to visualize all available data for each sequence, Phylo-VISTA can provide annotations beyond 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. The modularity of our program allows other, more advanced measures to be added.

3. Description of Phylo-VISTA

Phylo-VISTA assumes that the given data is a multiple alignment file in *multi-fasta* format. In addition, it takes as input the phylogenetic tree that presumably was applied in a progressive alignment phase of a multiple-alignment algorithm (such as in LAGAN, Brudno et al., 2003). These trees are used in Phylo-VISTA for generating similarity plots and computing the similarity measure.

The two most widely used methods for scoring multiple alignments are the "sum-of-pairs" method, where the score at every position is the sum of substitution scores in all pairs of sequences, and "consensus scoring", where a "consensus" letter is chosen at every position, and substitutions are penalized relative to the consensus. It is possible to combine the two methods. LAGAN, for instance, uses the sum-of-pairs model for scoring substitutions, and a consensus model, scaled appropriately, for scoring gaps. Several other methods for scoring multiple alignments have been suggested (Notredame, et al 2000; Holmes & Bruno, 2001).

Phylo-VISTA aims to highlight the similarity of genomic sequences over an entire phylogeny. 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 pairs 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 for a node X at position k in the alignment is defined to be:

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

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

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

In this sum, 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 (the circled node in Figure 1.A) at position one is

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

Because human and mouse are on the same subtree there exists a path between them in the phylogenetic tree and therefore, $B_{\text{human, mouse}}$ is null. 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 similarity measure ensures that the not genuine peaks resulting from the strong human-mouse similarity are eliminated.

A user of Phylo-VISTA can navigate through multi alignment data by performing these operations:

1. Selecting nodes in the phylogenetic tree to view the similarity plots for the corresponding subtrees.
2. Selecting a level of resolution of the plot by specifying a certain region from any of the sequences for display. Phylo-VISTA allows a user to analyze sequence data at multiple levels of detail from the curve of similarity to the actual alignment and the identification of conserved short motifs. We support this functionality by allowing a user to view the entire alignment and apply a “zoom” operator in a region of interest, up to the level of the individual basepairs.

4. Components

The Phylo-VISTA layout consists of four main components (Figure 1), which are described as follows:

1. Phylogenetic tree

Figure 1.A shows a sample phylogenetic tree used for the alignment of five sequences (human, mouse, chicken, pufferfish, and zebrafish). Each black node represents a similarity plot for all the sequences that are descendents of that node. The user can modify the tree if required.

2. Sequence traversal panel

This panel contains a collapsible traversal bar for each of the sequences, and an additional global bar for the alignment (Figure 2). The red rectangle indicates the currently selected region of each of the sequences. The 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 2). Each bar displays user-supplied annotations. Below the bar of each sequence, a narrow strip shows how the sequence is distributed across the alignment.

3. Similarity plots

A similarity plot is defined for each selected node in the tree (Figure 3). The x-axis represents the alignment projected to the subtree rooted at the selected node, and the y-axis represents percent-similarity. Similar to VISTA, the plot is produced by sliding a window of user-specified length over the alignment and calculating the similarity score at each basepair in that window. Below each plot are shown user-supplied annotations for all the sequences along with the gaps. 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. Because the x-axis represents the alignment, and not any of the 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 allowing a user to visualize sequences of arbitrary lengths (Figure 4).

4. Text window

In the text window sequences can be viewed in text format (Figure 5). The text is color-coded such that conserved DNA sequence motifs are highlighted. Black represents complete identity, while other colors are used to indicate identical bases between a subset of species.

To support interactive visualization, Phylo-VISTA is required to be computationally efficient. The time-critical step involves the computation of sequence similarity. The algorithm for calculating similarity requires $O(N^2L)$ time, where N is the number of sequences and L is the length of the alignment. Phylo-VISTA has $O(N^2L)$ time complexity and $O(NL)$ memory complexity. The Phylo-VISTA Java applet, being a thin client, does not require a user to have access to a high-end computer.

Example: analysis of a 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 (Begley and Green 1999; Orkin et al. 1999). 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) (Porcher et al. 1996; Robb et al. 1996). The expression pattern of this gene is highly conserved throughout vertebrates from mammals to teleost fish (Green et al. 1992; Kallianpur et al. 1994; Gering et al. 1998; Liao et al. 1998; Mead et al. 1998; Sinclair et al. 1999; Drake and Fleming 2000). Previous comparative analysis of five vertebrate SCL loci (considering human, mouse, chicken, pufferfish, and zebrafish) has 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).

We have applied Phylo-VISTA on a LAGAN multiple alignment of the SCL region, consisting of about 100 kb human, 65 kb mouse, 22 kb chicken, 8 kb pufferfish, and 67 kb zebrafish sequence data within the SCL region. After aligning all five species the length of the resulting multiple alignment equaled approximately 150 kb. Figure 4.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 4.B shows the Phylo-VISTA result obtained when zooming in the region with peaks (highlighted by an oval in Figure 4.A). A peak (shown by an oval in Figure 4.B) is visible in front of exon 1. By repeatedly applying the zoom operator, the plot shown in Figure 4.C is obtained. Reducing the sliding window width yields the similarity plot shown in Figure 4.D. The sequence traversal panel for this stage is shown in Figure 2. It can be seen that the promoter/enhancer region of SCL is selected for all sequences. The size of the selected region consists of only 39 basepairs, and sequence motifs in the text window can be examined.

Figure 5 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 (Lecointe et al. 1994; Bockamp et al. 1995, 1997, 1998; Sinclair et al. 1999). The other two motifs "GCCAAAT" (CS1, Cleavage signal-1 protein) and "ATAATGG" (CS2, Calsyntenin-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).

4. Conclusions and Future Work

Phylo-VISTA is a new interactive visualization and analysis tool for aligned sequences of multiple species. Its current functionality includes:

1. Visualization of multiple alignments at various levels of resolution

2. Visualization of alignments of any branch of a given phylogenetic tree connecting the aligned sequences
3. Adjustment of sliding window width and percent-similarity cutoffs for the representation of conserved regions (an important feature when dealing with distant versus close species)
4. Visualization of gaps and gene annotations for all species
5. Ability to view the alignment at the text level, assisting with the identification of sequence motifs

Phylo-VISTA is a modular program and can support, in principle, 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.

4. Availability

Phylo-VISTA is implemented as a Java applet. It is available online at http://www-gsd.lbl.gov/phylovista. It can be downloaded together with a help manual from http://graphics.cs.ucdavis.edu/~nyshah/Phylo-VISTA. The required input is an alignment file in multi-fasta format, a phylogenetic tree, and optional annotation files for all sequences in GFF format. Phylo-VISTA is also integrated with the multiple alignment program LAGAN at http://lagan.stanford.edu.

References

Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research,* **10,** 950–958.

Begley, C.G. and Green, A.R. (1999) The SCL gene: from case report to critical hematopoietic regulator. *Blood*, **93**, 2760-2770.

Blanchette, M. and Tompa, M. (2002) Discovery of Regulatory elements by a computation method for phylogenetic footprinting, *Genome Research,* **12**, 739-748.

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.

Bockamp, E.O., McLaughlin, F., Murrell, A.M., Göttgens, B., Robb, L., Begley, C.G., and Green, A.R. (1995) Lineage-restricted regulation of the murine SCL/TAL-1 promoter. *Blood,* **86,** 1502-1514.

Bockamp, E.O., McLaughlin, F., Göttgens, B., Murrell, A.M., Elefanty, A.G. and Green, A.R. (1997) Distinct mechanisms direct SCL/tal-1 expression in erythroid cells and CD34 positive primitive myeloid cells. *J. Biol. Chem.,* **272,** 8781-8790.

Bockamp, E.O., Fordham, J.L., Göttgens, B., Murrell, A.M., Sanchez, M.J., and Green, A.R. (1998) Transcriptional regulation of the stem cell leukemia gene by PU.1 and Elf-1. *J. Biol. Chem,* **273,** 29032-29042.

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

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.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E.M., Pachter, L. and Dubchak, I. (2002) 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.

Drake, C.J. and Fleming, P.A (2000) Vasculogenesis in the day 6.5 to 9.5 mouse embryo. *Blood,* **95,** 1671-1679.

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.

Edgar, RC, Sjolander, K . (2003) Simultaneous sequence alignment and tree construction using hidden Markov models. Pac Symp Biocomput. 180-91.

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.

Gering, M., Rodaway, A.R.F., Göttgens, B., Patient, R.K. and Green, A.R. (1998) The SCL gene specifies haemangioblast development from early mesoderm. *EMBO J.,* **17,** 4029-4045.

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.

Green, A.R., Lints, T., Visvader, J., Harvey, R., and Begley, C.G. (1992) SCL is co-expressed with GATA-1 in haemopoietic cells but is also expressed in developing brain. *Oncogene*, **7**, 653-660.

Hardison, R.C., Oeltjen, J. and Miller, W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome, *Genome Research*, **7**, 959-966.

Henikoff,S. and Henikoff,J.G. (1992) Amino Acid Substitution Matrices from Protein Blocks, *Proceedings of the National Academy of Sciences*, **89**,10915-10919.

Holmes, I and Bruno, W.J. (2001) Evolutionary HMMs: A Bayesian Approach to Multiple Alignment. *Bioinformatics*, **17**, 803-820

Kallianpur, A.R., Jordan, J.E., and Brandt, S.J. (1994) The SCL/TAL-1 gene is expressed in progenitors of both the hematopoietic and vascular systems during embryogenesis. *Blood,* **83,** 1200-1208.

Lecointe, N., Bernard, O., Naert, K., Joulin, V., Larsen, C.J., Romeo, P.H., and Mathieu-Mahul, D. (1994) GATA- and SP1-binding sites are required for the full activity of the tissue-specific promoter of the tal-1 gene. *Oncogene*, **9,** 2623-2632.

Liao, E.C., Paw, B.H., Oates, A.C., Pratt, S.J., Postlethwait, J.H., and Zon, L.I. (1998) SCL/Tal-1 transcription factor acts downstream of *cloche* to specify hematopoietic and vascular progenitors in zebrafish. *Genes & Dev.,* **12,** 621-626.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M.

and Frazer, K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, *Science*, **288**, 136-140.

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

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.

Mead, P.E., Kelley, C.M., Hahn, P.S., Piedad, O. and Zon, L.I. (1998) SCL specifies hematopoietic mesoderm in *Xenopus* embryos. *Development,* **125:** 2611-2620.

Orkin, S.H., Porcher, C., Fujiwara, Y., Visvader, J., and Wang, L.C. (1999) Intersections between blood cell development and leukemia genes. *Cancer Res.* **59,** 1784-1787.

Notredame, C., Higgins, D. G. and J. Heringa. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.,* **302**, 205-217.

Pennacchio, L.A., Olivier,M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M. and Rubin, E.M. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing, *Science*, **294**, 169-173.

Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F.W., and Orkin, S.H. (1996) The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell,* **86,** 47-57.

Robb, L., Elwood, N.J., Elefanty, A.G., Köntgen, F., Li, R., Barnett, L.D., and Begley, C.G. (1996) The SCL gene product is required for the generation of all hematopoietic lineages in the adult mouse. *EMBO J.,* **15,** 4123-4129.

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.

Sinclair,A.M., Gottgens,B., Barton,L.M., Stanley,M.L., Pardanaud,L., Klaine,M., Gering,M., Bahn,S., Sanchez,M., Bench,A.J., Fordham,J.L., Bockamp,E., Green, A.R. (1999) Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. *Dev. Biol.,* **209**, 128-142.

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 LBNL for support and inspiration, to Anu Chakicherla for writing the Web interface for the tool, 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), UC 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. M.B. was supported by an NSF graduate fellowship.

Captions:

Figure 1. Phylo-VISTA output

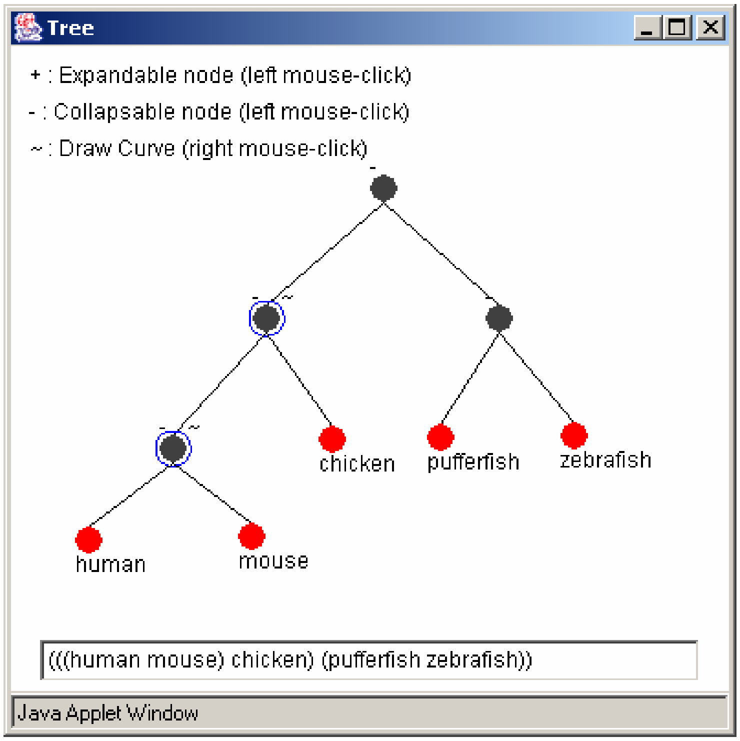
Figure 2. Sequence traversal panel

Figure 3. Similarity plots

Figure 4. Visualization of a multiple alignment dataset, consisting of human, mouse, chicken, pufferfish, and zebrafish data - stem cell leukemia (SCL) regions being analyzed

Figure 5. Text window

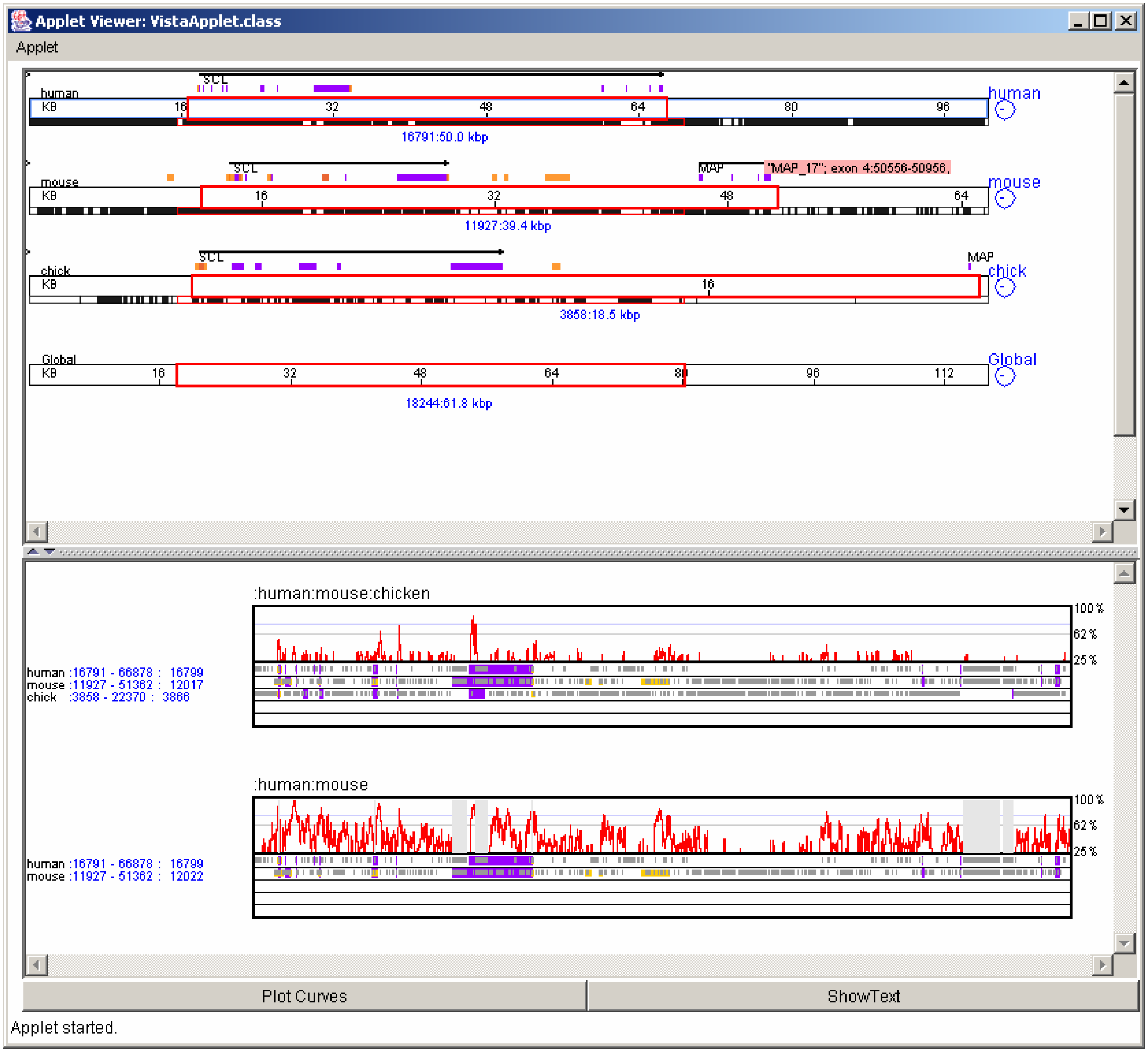
A.



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

(((human mouse) chicken) (pufferfish zebrafish))

B.

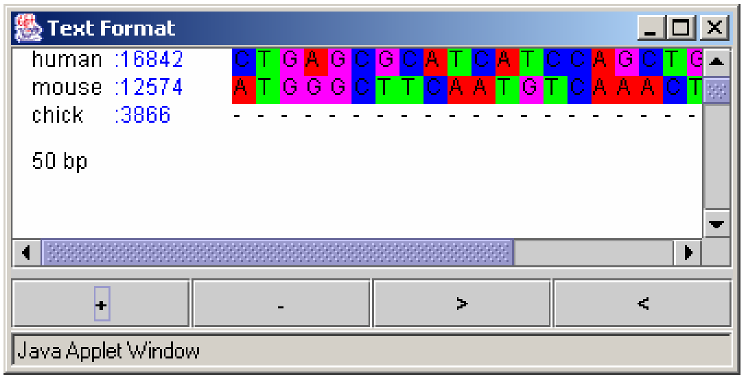


```
human :16791 - 66878 :  16799
mouse :11927 - 51362 :  12017
chick  :3858 - 22370 :  3866
```

```
human :16791 - 66878 :  16799
mouse :11927 - 51362 :  12022
```

Applet started.

C.



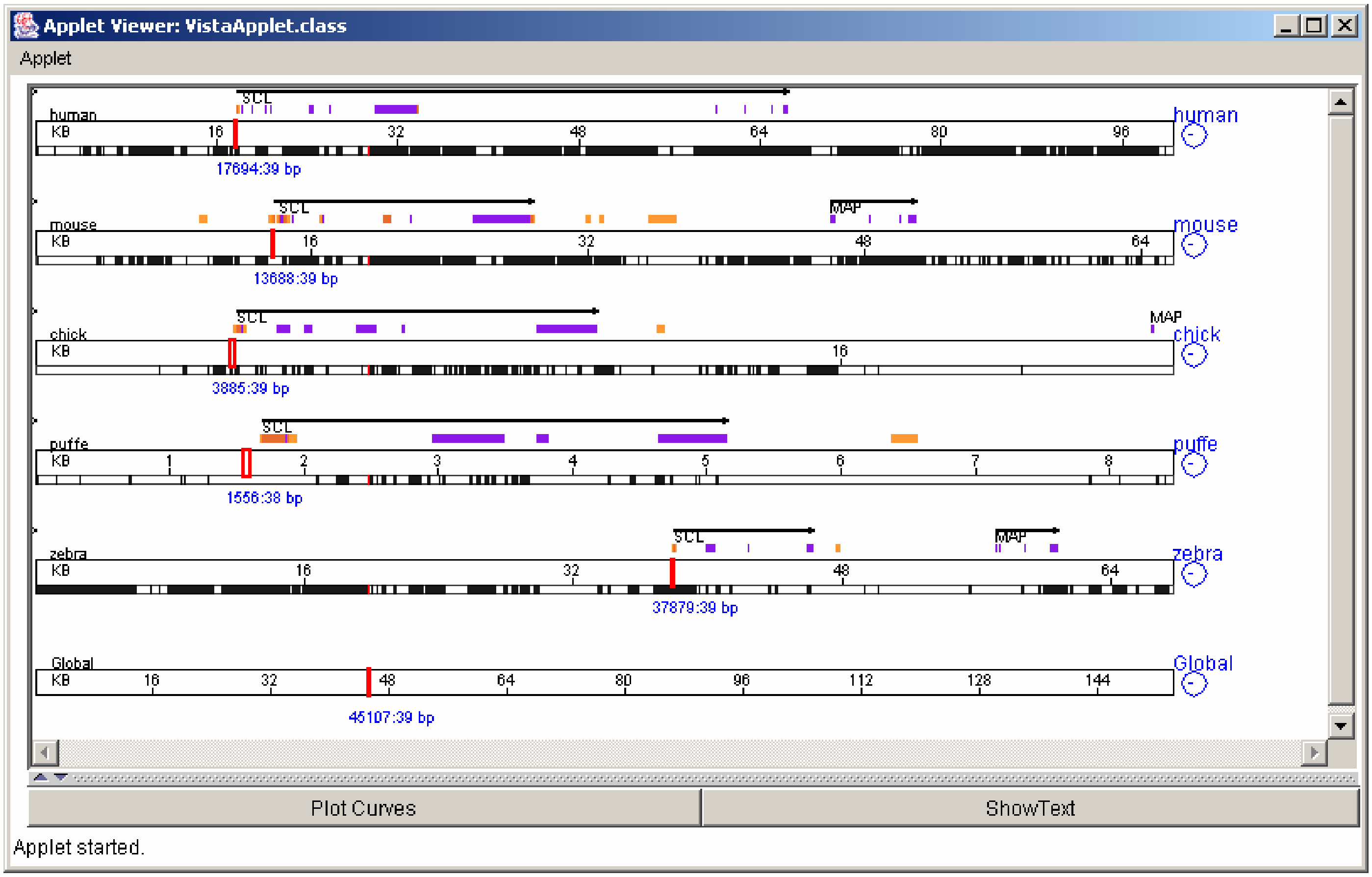
```
human  :16842   C T G A G C G C A T C A T C C A G C T G
mouse  :12574   A T G G G C T T C A A T G T C A A A C T
chick  :3866    - - - - - - - - - - - - - - - - - - - -

50 bp
```

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.

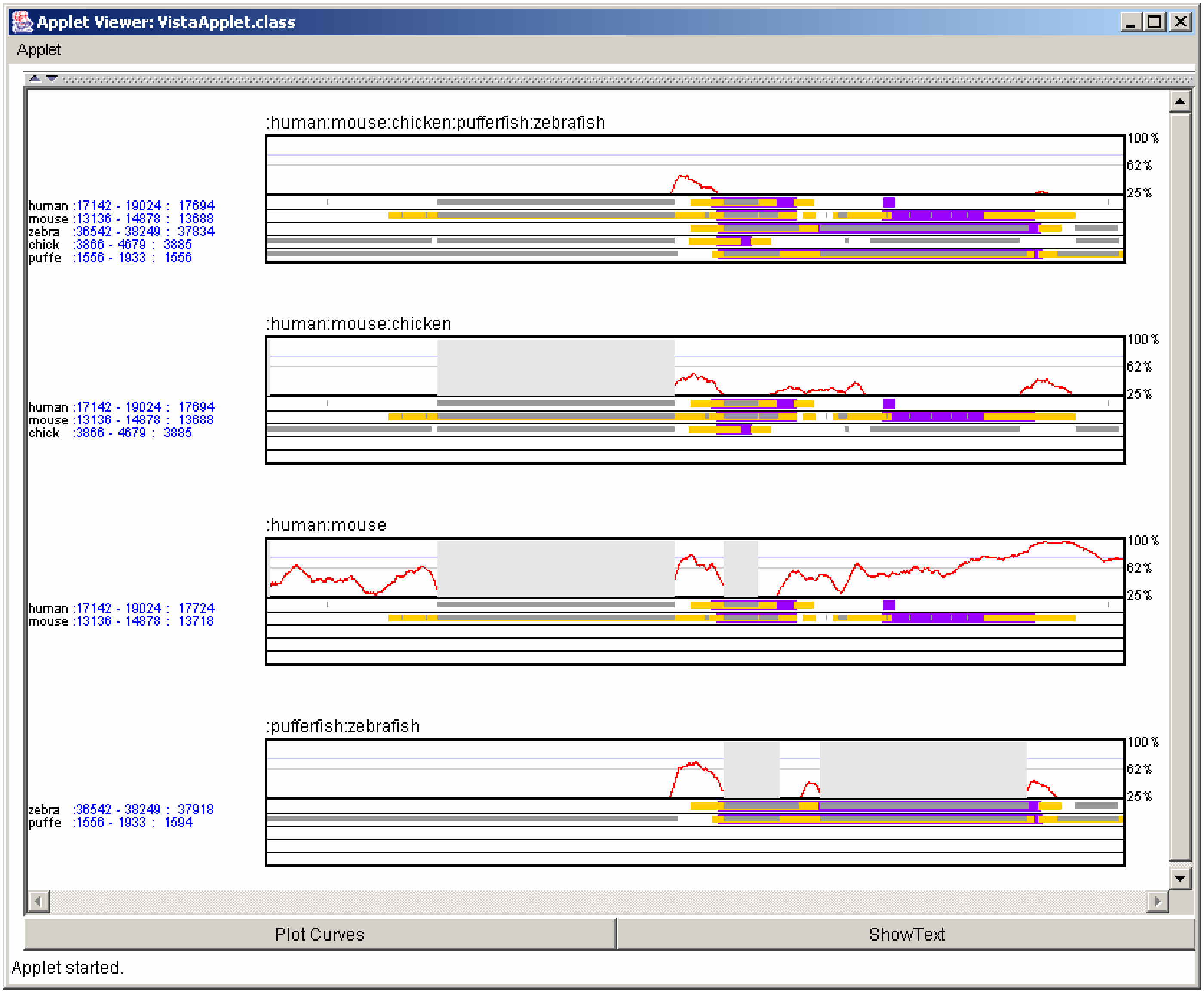
A. Phylogenetic tree. In this pairwise phylogenetic tree, all sequences in the alignment are represented by red leaf nodes. Each black node represents a similarity plot for all the descendent leaf nodes. The selected node is circled, representing a similarity plot for human, mouse, and chicken.
B. Similarity plots (details shown in Figure 3) for the selected nodes of the phylogenetic tree in part B. The sequence traversal panel (details shown in Figure 2) shows the bars for the human, mouse, and chicken sequences. The bars for pufferfish and zebrafish are not shown, as they were not selected in the tree.
C. Part of the alignment in color-coded text (details shown in Figure 4).



Applet started.

Figure 2. Sequence traversal panel.

A sequence traversal panel for the alignment of stem cell leukemia (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 the 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 17694, and the size of the selected region is 39 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.



```
human :17142 - 19024 :  17694
mouse :13136 - 14878 :  13688
zebra  :36542 - 38249 :  37834
chick  :3866 - 4679 :  3885
puffe  :1556 - 1933 :  1556

human :17142 - 19024 :  17694
mouse :13136 - 14878 :  13688
chick  :3866 - 4679 :  3885

human :17142 - 19024 :  17724
mouse :13136 - 14878 :  13718

zebra  :36542 - 38249 :  37918
puffe  :1556 - 1933 :  1594
```

Figure 3. Similarity plots.

Similarity plots corresponding to the black nodes of the phylogenetic tree shown in Figure 1.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 given. Gray rectangles indicate gaps, blue rectangles represent exons, and yellow rectangles show user-supplied conserved features. 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 indicates a region conserved in all sequences.

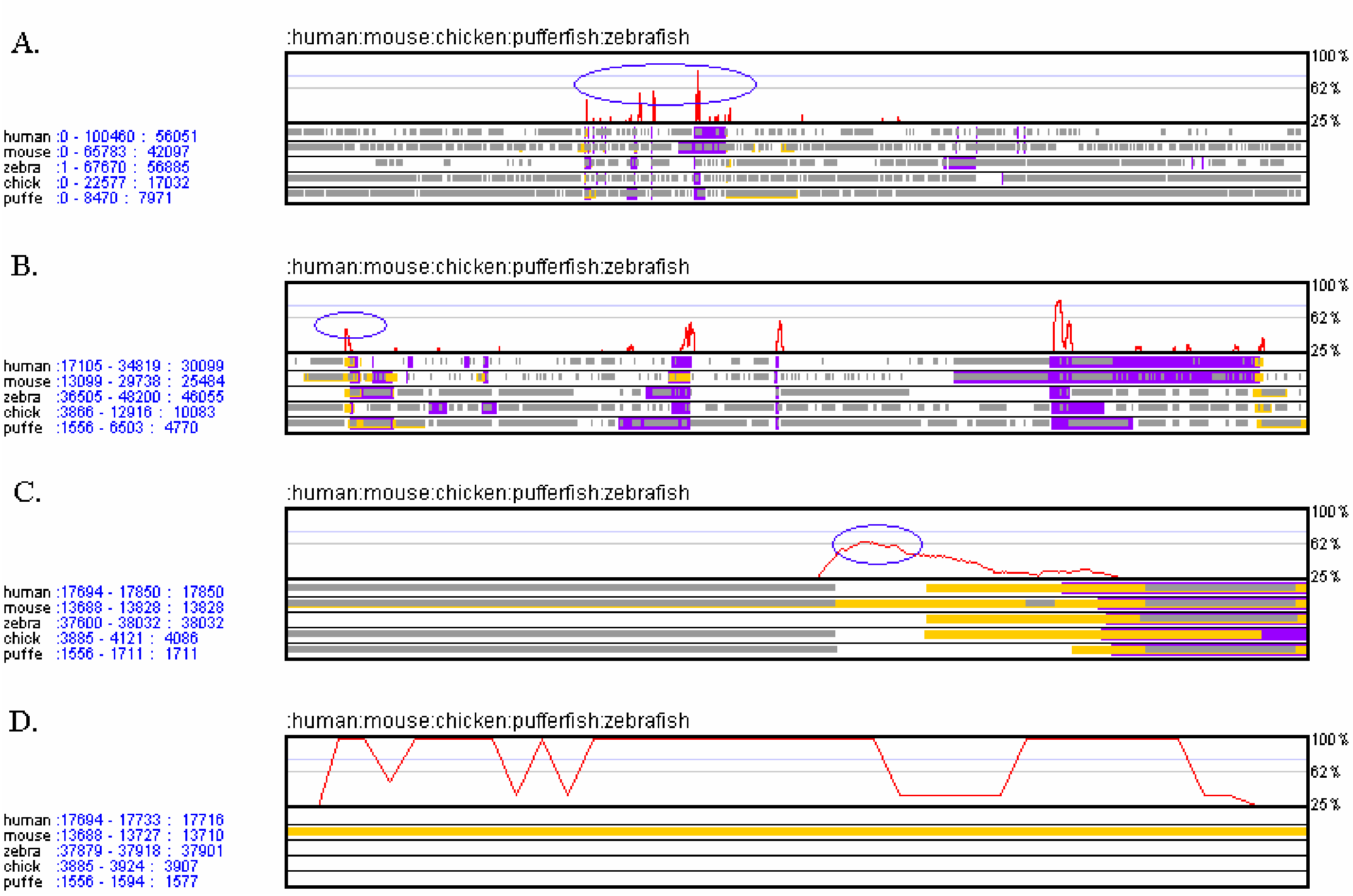
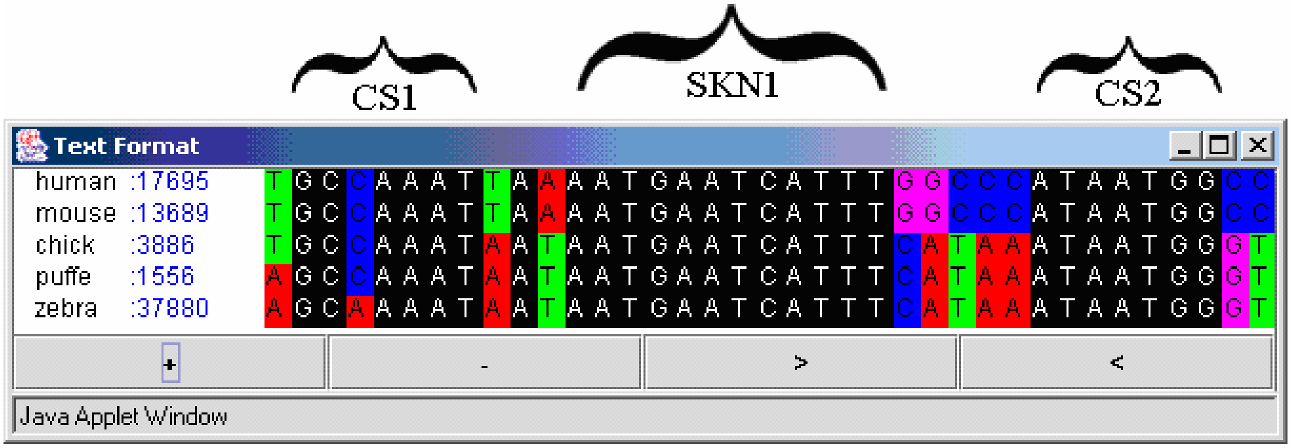


Figure 4. Visualization of a multiple-alignment data set consisting of human, mouse, chicken, pufferfish, and zebrafish SCL regions.

A. Bird's eye view of the alignment consisting of about 150000 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 SCL. 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 sequence traversal panel is shown in Figure 2. The corresponding text is shown in Figure 2.



```
human  :17695   T G C C A A A T T A A A A T G A A T C A T T T G G C C C A T A A T G G C C
mouse  :13689   T G C C A A A T T A A A A T G A A T C A T T T G G C C C A T A A T G G C C
chick  :3886    T G C C A A A T A A T A A T G A A T C A T T T C A T A A A T A A T G G G T
puffe  :1556    A G C C A A A T A A T A A T G A A T C A T T T C A T A A A T A A T G G G T
zebra  :37880   A G C A A A A T A A T A A T G A A T C A T T T C A T A A A T A A T G G G T
```

Figure 5. Text window.

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.