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, 2] and Inna Dubchak[2]

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

Abstract

We have developed Phylo-VISTA (Shah et al., 2003)*, an interactive software tool for analyzing multiple alignments by visualizing a similarity measure for DNA sequences of multiple species. The complexity of visual presentation is effectively organized using a framework based upon inter-species phylogenetic relationships. The phylogenetic organization supports rapid, user-guided inter-species comparison. To aid in navigation through large sequence datasets, Phylo-VISTA provides a user with the ability to select and view data at varying resolutions. The combination of multi-resolution data visualization and analysis, combined with the phylogenetic framework for inter-species comparison, produces a highly flexible and powerful tool for visual data analysis of multiple sequence alignments.*

Availability - http://www-gsd.lbl.gov/phylovista/

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. 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)).

An important consideration in multiple species sequence alignment is phylogeny. Phylogenetic analysis is widely used to study evolution. A phylogenetic tree is an acyclic graph representing a series of hypothesis about evolutionary events. Phylogenetic trees have been used extensively in creating alignments. For instance, progressive pairwise alignment techniques use a pre-computed phylogenetic tree as a "guide" to indicate the order in which multiple sequences should be aligned (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 phylogeny of the sequences into account.

Approach

We used the successful VISTA concept (Dubchak et al., 2000; Mayor et al., 2000) as a basis for the visualization of multiple alignments along with an associated phylogenetic tree. In order to visualize multiple alignment data, we developed several extensions to VISTA. 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. VISTA can show annotations only for the base sequence, and gap information can be displayed only for the sequences other than the base. Thus, using one sequence as base results in loss of information about the gaps in the base sequence and corresponding data of other sequences. Therefore, Phylo-VISTA uses the entire multiple alignment as a base. As a result, Phylo-VISTA is capable of displaying location and length of gaps in all sequences. In addition, to visualize all available data for each sequence, 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. The modularity of our program allows one to add other, more advanced measures.

Phylo-VISTA Scoring Method

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 sets of sequences from the left and right subtrees of the node. Similarity between sequences from the same subtree is ignored.

Components

The Phylo-VISTA layout consists of four main components:

1. **Phylogenetic Tree** - 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.
2. **Sequence Traversal Panel** - This panel contains a traversal bar that can be collapsed 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.
3. **Similarity Plots** - A similarity plot visually represents conservation among a given set of sequences based on the similarity measure described in the previous section. Similarity plots are defined for every selected node in the tree (Figure 1.B). The x-axis represents the alignment

projected to the subtree rooted at the selected node, and the y-axis represents percent-similarity. As in VISTA, the user selects a subset of the alignment data using a sliding window. 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 1.B).

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).

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. 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;
2. interactive specification of processing and visualization parameters, like sliding window width and percent-similarity cutoff;
3. simultaneous display of gaps and gene annotations for all sequences in a multiple alignment; and
4. multi-resolution 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. (This capability helps a user to identify conserved motifs.)

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 for this challenge.

References

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

Brudno, M., *et al*. (2003) LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA, *Genome Research*, **13**, 721-731.

Dubchak, I., *et al*. (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.

Frazer, K.A, *et al*. (2003) Cross-species Sequence Comparisons: A Review of Methods and Available Resources. *Genome Research***, 13**, 1-12.

Göttgens, B., *et al*. (2002) Transcriptional regulation of the stem cell leukemia gene (SCL)--comparative analysis of five vertebrate SCL loci. *Genome Research*, **12**, 749-759.

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

Mayor, C., *et al*. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length, *Bioinformatics***, 16**, 1046-1047.

Shah, N., Couronne, O., Pennacchio, L. A., Brudno, M., Batzoglou, S., Bethel E. W., Rubin, E. M., Hamann B., and Dubchak, I. (2003) Phylo-VISTA: An Interactive Visualization Tool for Multiple DNA Sequence Alignments, LBNL Technical Report LBNL-52539.

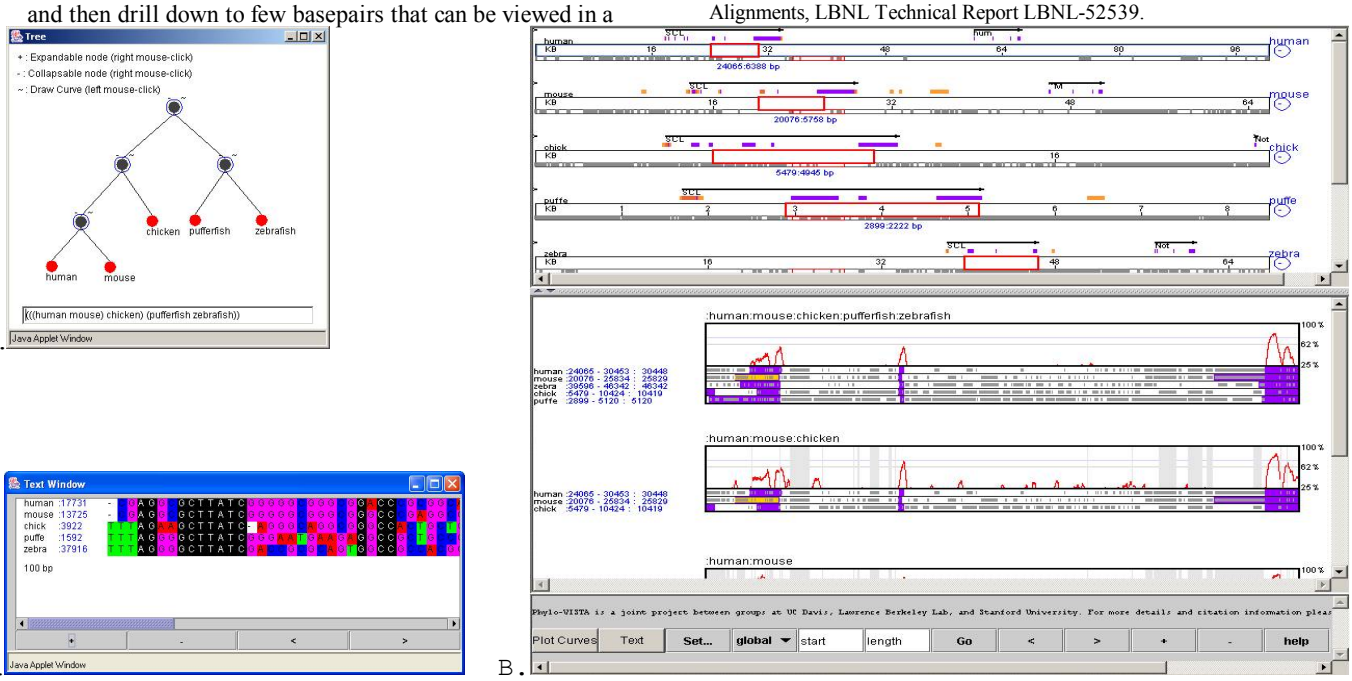

Figure 1. Phylo-VISTA output. Visualization of the alignment of about 100,000 basepairs of human stem cell leukemia (SCL) region, considering mouse, chicken, pufferfish, and zebrafish sequences.

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. Sequence Traversal Panel and Similarity Plots - The sequence traversal panel shows the bars for the human, mouse, chicken, pufferfish and zebrafish sequences.

C. Text Window - Part of the alignment in color-coded text.