BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 11 2000
Pages 1046–1047

VISTA: visualizing global DNA sequence alignments of arbitrary length

Chris Mayor[1], Michael Brudno[1], Jody R. Schwartz[2], Alexander Poliakov[2], Edward M. Rubin[2], Kelly A. Frazer[2], Lior S. Pachter[3,*] and Inna Dubchak[1,*]

[1]National Energy Research Scientific Computing Center, [2]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and [3]Department of Mathematics University of California at Berkeley, Berkeley, CA 94720, USA

Received on June 1, 2000; revised on July 19, 2000; accepted on August 2, 2000

Abstract

Summary: *VISTA is a program for visualizing global DNA sequence alignments of arbitrary length. It has a clean output, allowing for easy identification of similarity, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution. It is currently available on the web, thus allowing for easy access by all researchers.*

Availability: *VISTA server is available on the web at http://www-gsd.lbl.gov/vista. The source code is available upon request.*

Contact: *vista@lbl.gov*

Motivation

Alignment of genomic sequence from different organisms is becoming an increasingly powerful method in biology, and is being used for many purposes. Comparative sequence analysis has enabled identification of regulatory non-coding regions (Loots *et al.*, 2000; Dubchak *et al.*, 2000), and location of coding exons (Batzoglou *et al.*, 2000) using purely computational means. Visual front-ends are necessary to make the process of viewing alignments intuitive and easy and to facilitate discovery of conserved sequences for functionally significant regions.

For short alignments, dot plots (Gibbs and McIntyre, 1970; Sonnhammer and Durbin, 1996) have proven to be very useful, allowing for efficient visualization of repeats, rearrangements, and conservation. At the same time dot plots are less adept at displaying longer alignments, when the length of the sequences becomes larger than the resolution of the displaying medium. The Alignment Service package (Resenchuk and Blinov, 1995) was designed to align long sequences and to visualize resulting multiple alignments. This tool was applied to the annotation of open reading frames in viral genomes. Unfortunately this tool does not have a web interface and is hard to obtain. The earlier study (Date *et al.*, 1993) shows visual presentation of sequence variability along the alignment as a graph. For longer alignments, the only widely available visualization tool is PIPMaker (Schwartz *et al.*, 2000). PIPMaker generates a highly detailed plot of a local alignment as a series of dots and dashes representing the levels of conservation between the base sequence and clones from the second sequence.

Nevertheless, none of the currently available visualization methods combine the following critical features: (1) a clear, configurable output; (2) the ability to visualize several global alignments on the same scale; (3) the use of a continuous curve to represent the level of identity; (4) the ability to visualize alignments of up to several megabases; (5) effective handling of gaps in the alignment; (6) available source code. The *VISTA* program contains all of the aforementioned features.

Features

The *VISTA* plot (Figure 1) is based on moving a user-specified window over the entire alignment and calculating the percent identity over the window at each base pair. The x-axis represents the base sequence; the y-axis represents the percent identity. If the user supplies an annotation file (see below), genes and exons are marked above the plot. The direction of genes is indicated by an arrow, while the coding exons and UTRs are marked with rectangles of different color. Conserved regions (defined below) are highlighted under the curve, with red indicating a conserved non-coding region and blue indicating a conserved exon. Conserved UTRs are colored turquoise. The colors can be modified by the user.

A conserved region is defined with percentage and length cutoffs. Conserved segments with percent identity x and length y are defined to be regions in which

*To whom correspondence should be addressed.

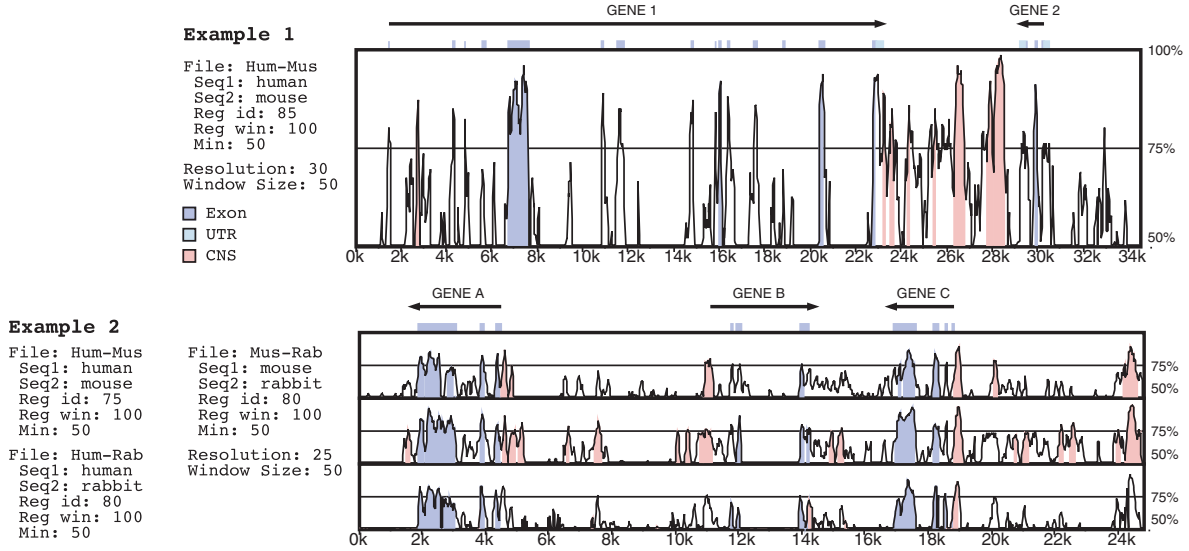

Fig. 1. Two sample *VISTA*s. Part (a) shows a two-way comparison between human and mouse sequences. Part (b) shows a fragment of a three-way comparison with human, mouse, and rabbit sequences.

every contiguous subsegment of length y was at least $x\%$ identical to its paired sequence. These segments are merged to define the conserved regions.

VISTA can be configured for visualizing alignments of various lengths by changing several parameters: the number of pages on which the output appears, the number of frames per page, the window size, and the resolution at which the alignment is plotted. *VISTA* allows one to easily create figures for various documents. For simplicity it is also possible to specify only a subset of these parameters, with the rest being automatically calculated. *VISTA* also supports simultaneous visualization of several related alignments (Figure 1b). This is particularly useful for long genomic sequences for which easy-to-use tools for multiple alignments and its graphic depiction are absent at the present time.

VISTA is implemented as an automatic server located at http://www-gsd.lbl.gov/vista web site. The input to *VISTA* consists of three parts. (1) One or more global alignments in one of the standard formats. If unaligned sequences are provided, they are aligned using the GLASS (Pachter, 1999; Batzoglou *et al.*, 2000) alignment program, which is specifically designed for the global alignment of large genomic regions. (2) An annotation file for the base sequence, in either the Sanger Centre's GFF format (http://www.sanger.ac.uk/Software/formats/GFF/), or another, simpler format described on our web page. (3) Any of the parameters (such as the coloring criteria or length of output) described previously. A more complete manual for using *VISTA* is available on the web site.

A *VISTA* plot is created by a Java program which outputs PDF files using the retepPDF (http://www.retep.org.uk/pdf/) library. *VISTA* uses time and memory linear in the length of the input sequences.

Acknowledgements

This work has been supported by the following grants: US Department of Energy contract DE-AC03-76SF00098, NIH GM-5748202 (KAF).

References

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

Date,S., Kulkarni,R., Kulkarni,B., Kulkarni-Kale,U. and Kolaskar,A.S. (1993) Multiple alignment of sequences on parallel computers. *CABIOS*, **9**, 397–402.

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

Gibbs,A.J. and McIntyre,G.A. (1970) The diagram, a method for comparing sequences. *Eur. J. Biochem.*, **16**, 1–11.

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.

Pachter,L. (1999) Domino tiling, gene recognition and mice, *PhD Thesis*, MIT.

Resenchuk,M. and Blinov,V.M. (1995) ALIGNMENT SERVICE: creation and processing of alignemnt of sequences of unlimited length. *CABIOS*, **11**, 7–11.

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

Sonnhammer,E.L.L. and Durbin,R. (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, 1–10.