Volume Visualization of Multiple Alignment of Large Genomic DNA

Nameeta Shah[1,2], Scott E. Dillard[1], Gunther H. Weber[1,2], and Bernd Hamann[1,2]

[1] Institute for Data Analysis and Visualization (IDAV), Department of Computer Science, One Shields Avenue, University of California, Davis, CA 95616-8562, U.S.A. {nyshah, sedillard, ghweber, bhamann}@ucdavis.edu

[2] Visualization Group, National Energy Research Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, U.S.A.

Summary. Genomes of hundreds of species have been sequenced to date, and many more are being sequenced. As more and more sequence data sets become available, and as the challenge of comparing these massive "billion basepair DNA sequences" becomes substantial, so does the need for more powerful tools supporting the exploration of these data sets. Similarity score data used to compare aligned DNA sequences is inherently one-dimensional. One-dimensional (1D) representations of these data sets do not effectively utilize screen real estate. As a result, tools using 1D representations are incapable of providing informatory overview for extremely large data sets. We present a technique to arrange 1D data in 3D space to allow us to apply state-of-the-art interactive volume visualization techniques for data exploration. We demonstrate our technique using multi-millions-basepair-long aligned DNA sequence data and compare it with traditional 1D line plots. The results show that our technique is superior in providing an overview of entire data sets. Our technique, coupled with 1D line plots, results in effective multi-resolution visualization of very large aligned sequence data sets.

Key words: Multiple alignment, Hilbert curve, volume visualization.

1 Introduction

The human genome consists of about three billion basepairs, of which only a small percentage is well-understood. In order to decipher the rest of the genome, and to understand general principles of genome structure and function, biologists look for overrepresented patterns in the genomes. Another approach to understanding genetic code is through comparison of genomes, or parts of genomes, of different species. Although many techniques for visualization of DNA sequences have been developed and various tools exist for

visualization of alignment data, there exists a need for visualization techniques that can handle very large data sets.

1.1 Multiple Alignment

```
Human     AATTCCGATGGGAACTACTGGATC-CGG
Chimp     AATTCCAATGGGAA--ACTGGATCCCGG
Mouse     AAATCCG---GAAACCACTGG----AGG
Rat       A--TCCG---GAAACCACTGG----AGG

All       63166631116266316666611102 66
          (6-100%, 3-50%, 2-33%, 1-17%)

Primates  11111101111111001111111 0111
          (1-100%, 0-0%)

Rodents   10011110001111111111 10000111
          (1-100%, 0-0%)
```

Fig. 1. Multiple alignment of four species: human, chimp, mouse and rat, and sum-of-pairs similarity scores for three different plots. (A sum-of-pair similarity score is computed by adding one for every basepair consisting of the same base, considering all possible distinct species pair.)

Currently, biologists compare genomes of many species at different evolutionary distances by examining multiple alignments [7]. A multiple alignment is a set of sequences in a "rectangular arrangement," where each row consists of one sequence padded by gaps, such that the columns highlight similarity/conservation between positions (http://www.cryst.bbk.ac.uk/BCD/bcdgloss.html). Figure 1 shows an example of a multiple alignment of four sequences, where the characters A, T, C and G represent the bases adenine, thymine, cytosine and guanine, respectively. Below the alignment we show similarity scores for the multiple alignment. The similarity score for each column shows the level of conservation among sequences, considering all possible pairwise comparisons of characters (six in the case of four species). Different schemes are used to calculate similarity scores, including *entropy, sum-of-pairs, weighted sum-of-pairs, parsimony, etc.* (http://lepo.it.da.ut.ee/∼mremm/kurs/multali.htm). We assume that similarity scores are provided as input for our visualization purposes.

1.2 Related Work

Comparison of biological sequences is a very important aspect of genome research. Fractal-based visualization using chaos automata and iterated function

systems [1] and space-filling curves [12, 4] have been used for identifying patterns, similarities and dissimilarities in biological sequences. PATTVision is a 3D visualization tool that uses texture mapping for viewing patterns in multiple sequences [23]. Arc diagrams have been used to visualize shared patterns among sequences [22]. These methods are effective for relatively small-sized sequences consisting of up to few thousand basepairs. Alignment is one of the most extensively used techniques for comparing DNA sequences. With larger sequences being aligned, text-based alignment viewers are inadequate, and alignment visualization using line-based glyphs in 3D space have been developed as a response [5, 6]. Currently, several tools for the visualization of alignment data are publicly available. One highly popular and successful tool is VISTA [16]. VISTA represents the level of conservation between species as a curve calculated by sliding a window of predefined size over the given alignment and computing the average similarity score over the window. VISTA shows pairwise similarity scores. Phylo-VISTA [19] extends the VISTA concept to the visualization of multiple (more than two) alignments. Other commonly used tools like MultiPipMaker [18] and SynPlot [9] also use 1D line or dot plots. Various genome browsers [26, 14, 11] also use textual display and 1D line plots to show alignment scores. SequenceJuxtaposer [21] uses "focus+context" techniques to provide interactive multiresolution navigation of sequences up to two million basepairs in total.

1.3 Motivation

With advancements in sequencing technology, increases in computational power, and the development of better computational methods, it is now possible to align several-million-basepair-long sequences. It is clear that the need exists, or will exist in the very near future, to develop new visualization techniques to support the interactive, visual exploration for such extremely large sequence data. The basic principle or the *Visual Information Seeking Mantra* is [20]:

Overview first, zoom and filter, then details-on-demand.

As the size of a typical alignment reaches several million basepairs, all existing techniques for visualization of alignment data fail to provide a good overview of an entire data set that will highlight regions of interest for further focus. Our work is driven by the need for visually presenting large 1D data sets in their entirety such that interesting features of the data for more detailed exploration are clearly visible.

Earlier work by Wong *et al.* [27] used space-filling *Hilbert curves* to transform sequential data into 2D space. This transformation allows one to display one million basepairs using a 1000×1000 pixel image. Application of digital image processing filters to such images reveals interesting patterns in the data. This work motivated us to represent multiple alignment data in 3D

space, embedded in a fixed volume by using 3D space-filling curves. This approach allows us to apply various volume visualization techniques to render the data. We use hardware-accelerated volume rendering with maximal intensity projection [10] for visualization. In general, choosing a transfer function for volume rendering is not a trivial task. In our application, however, we specify a transfer function based on parameters relevant from the perspective of the driving biological problem.

2 Our Approach

With current multiple alignment algorithms, million-basepair alignments are becoming increasingly common. Techniques are required to examine such large data sets that contain more data points than there are screen pixels. "Focus+context" approaches have been effectively utilized in such cases where the most important data is given more screen space and is displayed in more detail than the rest of the data, which is still displayed at lower-resolution to provide context. This method is not very helpful when a user might need to examine the complete data set to locate interesting features in the data. Our goal is to make better use of the screen real estate to provide more screen area per data element.

2.1 From 1D Sequential Data to 3D Volume Data

A naïve approach for arranging sequential data in 3D space is using a scanline traversal in 2D planes and stacking these planes in perpendicular direction. In a 64^3 volume grid, for example, such an arrangement will place the 0^{th} position (mapped to (0, 0, 0) in 3D space) and 4096^{th} position (mapped to (0, 0, 1) in 3D space) next to each other. When two positions that are distant in the 1D sequence happen to be adjacent in a 3D arrangement, interpretation of data/visuals is difficult. To mitigate this problem, a 3D arrangement that maximizes spatial coherence should be used. Work by Keim et al. [13] and Voorhies [25] has shown that a Hilbert curve-based mapping is among the most coherent space-filling-curve-based approaches. Coherence is defined as the amount by which neighboring pixels (voxels in our case) are at sequential positions on the curve [27]. Figure 4 shows a 3D Hilbert curve. We map a score at position i in a multiple alignment to a position in 3D space using the algorithm described by Max [15].

2.2 Volume-based Visualization

Once the 1D alignment data is tranformed to volume data, we apply volume rendering to the data. In the following, we describe the details of volume-based visualization.

Color channels

Consider a multiple alignment of sequences of four species: human, chimp, mouse and rat, see Figure 1. Biologists are interested in

- conserved features in all four species,
- primate-specific features, and
- rodent-specific features.

These features can be identified by considering the following three different similarity plots and comparing them:

- a similarity plot for all four species,
- a similarity plot for human and chimp (primates), and
- a similarity plot for mouse and rat (rodents).

In this scenario, we can take advantage of three color channels, red, green and blue, to visualize three similarity plots: We map the first similarity plot to the red channel, the second plot to the green channel and the third plot to the blue channel. This approach allows a user to compare common and distinct features of all plots. Different colors highlight different features of the data. For example, blue represents a primate-specific feature absent in rodents, green represents a rodent-specific feature absent in primates, and cyan represents a feature specific to both primates and rodents. Further, white (the combination of all colors) represents a region conserved among all species.

Transfer function

A separate transfer function is associated with each color channel. We use a linear transfer function like the one shown in Figure 2. We use a user-defined similarity score threshold, below which everything is rendered transparently. The slope of the function is adjustable.

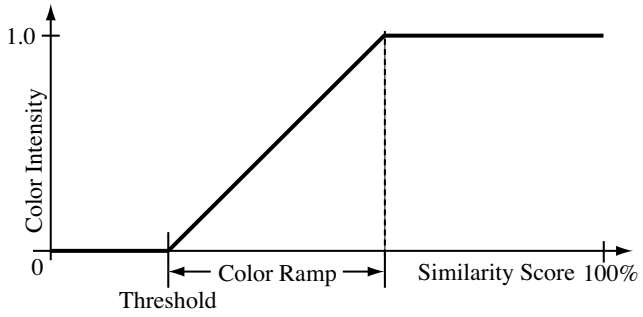

Fig. 2. Transfer function.

Maximal intensity projection

The Hilbert curve transforms a 1D sequence into a 3D volume that is just as large in Z direction as in X and Y directions. When this volume is projected on to 2D screen space, there is a substantial overlap of data. If the data is rendered fully opaquely, then the frontmost layer will be the only visible output, occluding the rest of the data. A back-to-front rendering allows for alpha blending and transparency, so that uninteresting areas of the data are drawn transparently and do not occlude anything. The user can freely rotate the data, in order to avoid arrangements in which interesting features occlude each other. But it is possible that a feature in the data may be completely surrounded by opaque voxels, so that rotation alone will not reveal it (Figure 7.C). We employ a maximal intensity projection that maps the intensity of a voxel to its screen depth. This depth is then used as input to the Z-buffer algorithm, common to most graphics hardware available today. This method guarantees that highly conserved regions in the DNA are always displayed before less conserved regions. Data is still occluded, but it is assumed that the user is more interested in regions of high conservation rather than regions of less conservation (Figure 7.D). In situations where this assumption is false, the user may invert the data for display, which is effectively a minimal intensity projection.

2.3 Annotations

For analysis of multiple alignment data, biologists often need additional biological information, such as information concerning a known gene model. For example, it is desirable to show the start and end coordinates of a gene, exons (the protein coding part of a gene), etc. This information can be provided in the form of annotations next to genome sequences. Thus, in addition to displaying similarity plots of a multiple alignment, displaying annotations is a major aspect of genomic data visualization. We draw annotations as "pipes" following the 3D Hilbert curve (Figure 4). As the size of an annotation grows larger so does the number of pipes we draw. This method may clutter the display and slow down interaction. We handle this problem by using a multiresolution Hilbert curve. Figure 3 shows a Hilbert curve drawn in 2D space using two resolutions. The black curve is a high-resolution curve with 15 line segments; the red curve is a lower-resolution version with just three segments. The image on the left-hand side in Figure 4 shows annotations drawn using a multiresolution Hilbert curve approach. The image on the right-hand side uses the same annotations drawn as a high-resolution Hilbert curve. We draw lower-resolution curves as pipes with large diameters to show that they cover a greater volume. We plot overlapping annotations by drawing multiple Hilbert curves with a slight offset (Figure 4). As a result of our volume-based visualization approach annotations can also be displayed at a higher resolution. Consider a 2097151-basepair-long genome sequence with three annotations of

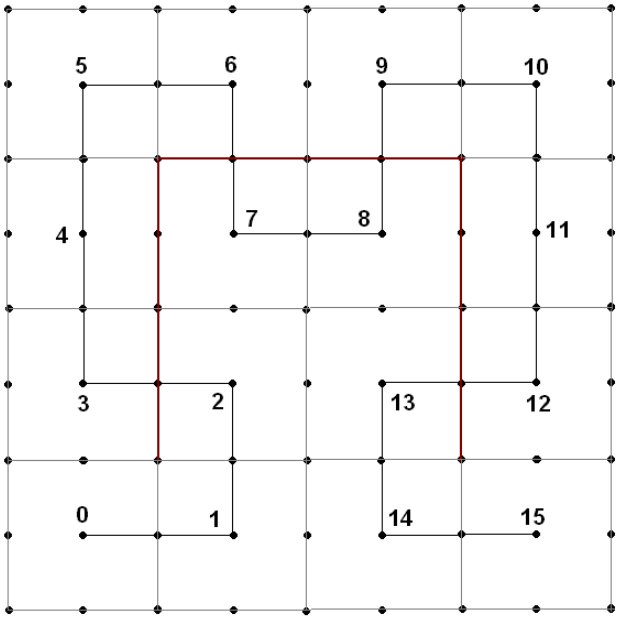

Fig. 3. Multiresolution Hilbert curve embedded in 2D space. The black curve is a high-resolution Hilbert curve; a lower-resolution version is shown in red.

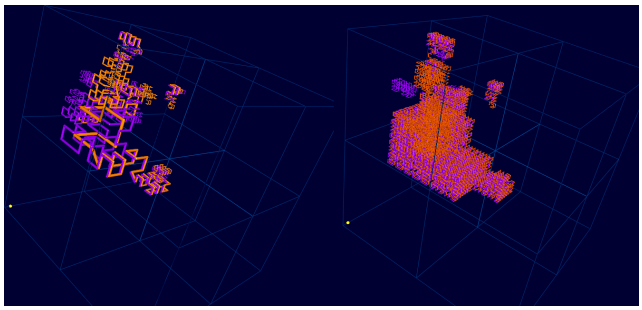

Fig. 4. Annotations. Left: annotations drawn using multiresolution Hilbert curve; right: annotations drawn using high-resolution Hilbert curve. Purple and orange pipes represent two different sets of annotations.

sizes 63, 54 and 63 basepairs and separated by 565 and 11260 basepairs, respectively. The first two annotations, when displayed using a traditional 1D plot at 1000-pixel resolution, would occupy the same pixel position on the screen, and the two annotations would be indistinguishable. Using our 3D visualization method all three annotations can be seen distinctly (Figure 5).

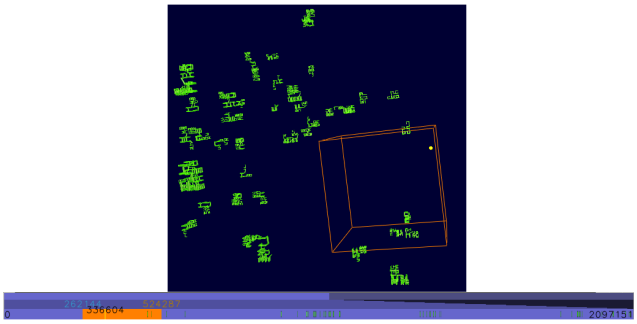

Fig. 5. Effective use of screen real estate. Three annotations of sizes 63, 54 and 63 separated by 565 and 11260 basepairs, respectively, on 2097151-basepair-long sequence are distinctly visible in 3D space (in orange cube). But the first two merge in a 1D representation (in orange rectangle).

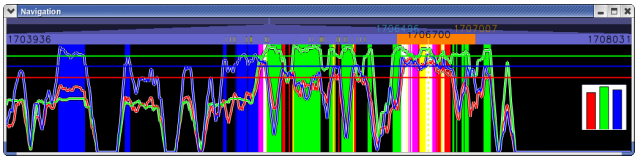

Fig. 6. Control for navigation on original 1D data sequence. The upper bar corresponds to the complete sequence, and the lower bar corresponds to the displayed portion of a sequence. A highlighted region in the upper bar represents the displayed portion of the complete sequence and is connected to the lower bar, showing the correspondence of the displayed sequence portion to the complete sequence. The highlighted region in the lower bar denotes the selected octant. Yellow rectangles are the annotations. 1D plots for all three channels are drawn as red, blue and green curves. The background color is the same as the corresponding color of the voxel in the volume. The similarity score for all three channels is shown at the marker position using a bar graph as shown in the square on the right-hand side.

3 Implementation

The adaptation of volume rendering techniques to modern commodity graphics hardware has made it easy and affordable to achieve interactive framerates. The massive parallelism of a modern GPU is well suited to evaluating the relatively simple transfer function millions of times per second, while the SIMD architecture allows for three channels of data to be processed as quickly as one. Using an NVidia 5900 FX graphics processor, our implementation displays 15 frames/sec for a 64^3 data set ($\sim$260,000 basepairs), where each frame consists of 800×800 pixels.

3.1 Volume Renderer

We render images using view-perpendicular slices with a back-to-front ordering approach. This approach allows us to incorporate transparency into the transfer function. We use a fragment program to evaluate the transfer function [2]. This program compares the value at each voxel with the three thresholds (for the three color channels.) If a value is above the threshold, that color channel is activated for that voxel. If no value is above the threshold, the voxel is transparent rather than black. The color ramp (see figure 2) can be increased to yield a smooth transition between states. In this case, values near the threshold will be only slightly visible while values much greater than the threshold will clearly stand out. The maximal intensity projection is evaluated within the fragment program. A specific color channel, or the sum of all three, is mapped to the Z-buffer.

3.2 User Interface

We provide a user interface for our volume renderer that allows a biologist to explore intuitively large alignment data. We describe the specifics in the following.

Volume Rendering Controls

A user can choose thresholds for three color channels, which will automatically set the transfer function. Any of the three channels, or the sum of all three, can be chosen for the maximal intensity projection. A box filter of variable size can be applied to the similarity score data for smoothing.

Annotation Controls

Multiple sets of annotations can be loaded and displayed with our prototype system. User-defined colors are associated with each set of annotations. In addition, the diameters of the pipes can be adjusted.

Navigation Controls

A Hilbert curve is *fractal* in nature [17]. As a result, a data string embedded in a 3D volume can be organized using an octree-like structure. For navigation purposes, a currently displayed sequence portion, which always has the shape of a cube, is divided into octants. A user can then select an octant for zooming in. It is also possible to shift the selected octant by half of its size in both direction. To provide a context for the currently displayed sequence portion, the bounding box of the volume corresponding to the complete sequence can be displayed.

A 3D representation of inherently 1D data poses problems for navigation in 3D space. A user must know the 1D position of the underlying (sequence) data, but one can lose one's orientation when navigating in 3D space. We tackle this problem by using a 1D representation of the sequence shown next to the volume display. This representation allows a biologist to keep track of the position within the sequence in a more traditional fashion. This additional display consists of two bars, see Figure 6, where the upper bar corresponds to the complete sequence and the lower bar corresponds to the displayed portion of a sequence. A highlighted region in the upper bar represents the displayed portion of the complete sequence and is connected to the lower bar, indicating the correspondence of the displayed sequence portion to the complete sequence.

Annotations are also displayed using rectangles on the lower bar. Selecting an octant for zooming in can be accomplished by clicking on the corresponding octant in the volumetric image, or by clicking on the corresponding part of the 1D bar representing the displayed portion of the sequence. Zooming out leads a user to the next lower level of detail. The 1D sequence representation can also be used to move a marker through the 3D display volume. This marker is symbolized by a vertical line in the 1D sequence display. Dragging this line by using a mouse moves the marker through the volume along the Hilbert curve. We show the similarity score for all three channels at the marker position using a bar graph as shown in the square on the right-hand side of Figure 6. We also show 1D plots for all three channels and we use the color in the volume as background for the line plot.

4 Results

We have applied our method to multiple alignment datasets that were created using MLAGAN [3]. We have used these two test datasets:

1. Stem Cell Leukemia (SCL) dataset.
   The SCL dataset is a multiple alignment DNA sequence data set of sequences from five species: human, mouse, chicken, pufferfish and zebrafish. All sequences contain the SCL gene. The alignment consists of 150000 basepairs. These sequences were aligned in order to discover regulatory elements of the SCL gene. Regulatory elements are short DNA sequences consisting between 6 and 12 basepairs. They are generally found in a region in front of a gene called *promoter*. The underlying assumption is that they are conserved in evolutionary distant species because regulatory elements are functionally important.
2. Cystic Fibrosis Transmembrane Conductance (CFTR) dataset.
   The CFTR dataset is a multiple alignment DNA data set of sequences from 12 species: human, chimp, baboon, cat, dog, cow, pig, mouse, rat, chicken, fugufish and zebrafish. The sequences are from the region containing a gene coding the CFTR regulator, and nine other genes. The

alignment is four million basepairs long. This alignment can help biologists with the discovery of regulatory elements as well as their identification of subclass-specific features.

We compare volume-based visualizations of these datasets with 1D similarity plots and SequenceJuxtaposer. Most of the currently available tools including genome browsers like UCSF, Ensembl and NCBI use line plots and yield similar results to the 1D plots we compare our results with. We are not aware of any other tools that can handle well alignment data as large as four million basepairs.

Figure 7 shows the visualization of the SCL dataset. We visualize three similarity plots: the similarity plot for all five species, which is mapped to the red channel; the similarity plot for human-mouse-chicken, which is mapped to the green channel; and the similarity plot for the two fish species, which is mapped to the blue channel. The resolution of the displayed volume is 64^3 (Figure 7.C, D). Yellow pipes show exons (protein-coding parts of a gene) for the human sequence. In Phylo-VISTA plots, exons are shown as purple bars below the plot. Two bars exist: an upper one showing exons of the human sequence and a bottom one showing exons of the mouse sequence (Figure 7.B). A box filter with a width of 50 was used to smooth data for all three similarity scores. A threshold of 25% was used for all plots. We also generated a visualization of the SCL dataset using SequenceJuxtaposer [21] (Figure 7.A).

Red lines in Figure 7.A show differences among the five sequences. It is difficult to find regions of high similarity looking at this picture. White spots in the volume-rendered image (Figure 7.D) indicate regions of high conservation in all three similarity plots. These spots are seen as peaks in all three corresponding Phylo-VISTA plots (Figure 7.B). Green spots are conserved regions in the human-mouse-chicken plot that are absent in the fish plot. Similarly, blue spots are fish-specific conserved regions. The conserved regions seen in these images contain the regulatory elements of the SCL gene [8]. In order to compare three 1D plots a user has to inspect them by eye and determine whether there are peaks in different plots at the same position. This analysis approach may create problems, especially when one pixel represents the similarity score for more than one column in a multiple alignment. In the case of the volume-rendered image, a user needs to consider only the color to compare all three plots at once. Of course, the same color scheme can be applied to 1D plots, and this is show in Figure 7.E. One can see that the four white peaks are barely visible in this plot whereas the corresponding four white spots in the volume-rendered image are strikingly visible. One of the issues in volume rendering is occlusion as can be seen in Figure 7.C, where the high similarity regions are hidden inside the volume. We handle this problem by using maximal intensity projection, as a result of which the high-similarity regions show through (Figure 7.D).

Figure 8 shows a visualization of the CFTR dataset. This figure shows only one plot for the similarity among all 12 species, mapped to the red

channel. The resolution of the volume is 256^3. The sequence data fills 25% of the volume. White pipes show genes, and purple pipes indicate the exons of the human sequence. The data was smoothed using a box filter of width 50. The threshold was 25%. The correspondence between the 3D and 1D plots is illustrated by thick gray lines. The middle image shows the entire dataset. Exons and their conservation are much more clearly and distinctly visible in the 3D plot than in the 1D plot. The red spots in the first half of the dataset indicate conservation among all species. The top and the bottom images show zoomed-in views from different viewpoints of the same circled part of the middle image. The circled part in the top image indicates conservation of the first exon of a gene and the promoter region. The image to its right shows a zoomed-in view. The conserved region seen in this right-most top image contains regulatory elements of CAV2 gene [24]. Similarly, the bottom images show the conservation of regulatory elements of the CAPZA2 gene [24].

Figures 9 and 10 show the visualization of the CFTR dataset. Three different similarity plots are used in the volume visualization. The similarity plot for primates (human, chimp, baboon), artiodactyls (cow, pig) and carnivores (cat, dog) are mapped to the red channel. The green channel is used to display the similarity plot for primates, and the blue channel for showing the similarity plot of carnivores and artiodactyls. Figure 9.A shows the entire dataset. We use different thresholds for different channels: 70% for red, 90% for green and 80% for blue. The 3D visualization reveals many more features when compared to the 1D plot. Larger primate-specific (green) and artiodactyls-carnivores-specific (blue) regions can be identified from both 3D and 1D plots. In the 3D plot, we can see a white spot in an otherwise blue-green region (shown with the red arrow). This feature of the dataset is not visible in the 1D plot (indicated by the black arrow). As we zoom-in further (Figure 9.B, C), the spot remains visible in the 3D plots but not in the 1D plots. In Figure 10.D, we start seeing a white line in the 1D plot, which becomes more distinctly visible when we zoom-in further (Figure 10.E). This conserved region is in a noncoding region that is far away from any gene but can potentially be a distant regulatory element. Thus, our 3D representation allowed us to detect an interesting feature immediately that would have been missed by looking at just 1D plots.

5 Conclusions

We have presented a volume-based visualization technique for analyzing multiple alignment data. Our results demonstrate that 3D representations and visualizations of genome data are quite effective and utilize 3D display space efficiently. As a result, we can convey information more compactly, especially for billion-basepair sequence data.

Although developed for a particular biological application, our method can be applied to other kinds of massive sequential 1D data sets. Other volume-

based visualization techniques, like isosurfacing or plane slicing, etc. could also be used when appropriate for a given application.

6 Acknowledgements

This work was supported by the National Science Foundation under contracts ACI 9624034 (CAREER Award), through the Large Scientific and Software Data Set Visualization (LSSDSV) program under contract ACI 9982251, through the National Partnership for Advanced Computational Infrastructure (NPACI) and a large Information Technology Research (ITR) grant; the National Institutes of Health under contract P20 MH60975-06A2, funded by the National Institute of Mental Health and the National Science Foundation; by the Director, Office of Science, U.S. Department of Energy under contract DE-AC03-76SF00098; and the Lawrence Berkeley National Laboratory through a Laboratory Directed Research Development (LDRD) project. We thank Chris Co and Oliver Kreylos for their helpful suggestions. We also thank the members of the Visualization and Graphics Research Group at the Institute for Data Analysis and Visualization (IDAV) at the University of California, Davis, and the members of the Genome Sciences Department and the NERSC Visualization Group of the Lawrence Berkeley National Laboratory for their support.

References

1. Dan Ashlock and Jim Golden. *Ch.11 Evolutionary Computation and Fractal Visualization of Sequence Data*. Morgan Kaufmann, 2002.
2. Bob Beretta, Pat Brown, Matt Craighead, Cass Everitt, Evan Hart, Jon Leech, Bill Licea-Kane, Bimal Poddar, Jeremy Sandmel, Jon Paul Schelter, Avinash Seetharamaiah, and Nick Triantos. GL_ARB_fragment_program Specification. Online OpenGL Extension Registry, August 2003.
3. Michael Brudno, Chuong Do, Gregory Cooper, Michael F. Kim, Eugene Davydov, Eric D. Green, Arend Sidow, and Serafim Batzoglou. Lagan and multilagan: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4):721–731, 2003.
4. Hsuan T. Chang, Neng-Wen Lo, Wei C. Lu, and Chung J. Kuo. Visualization and comparison of DNA sequences by use of three-dimensional trajectories. In *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003*, pages 81–85, Adelaide, Australia, 2003.
5. Ed Huai-hsin Chi, Phillip Barry, Elizabeth Shoop, John Carlis, Ernest Retzel, and John Riedl. Visualization of biological sequence similarity search results. In Gregory M. Nielson and Deborah Silver, editors, *Proceedings of IEEE Visualization 1995*, IEEE Visualization, Annual Conference Series, pages 44–51, Atlanta, USA, 1995. IEEE, IEEE Computer Society Press.

6. Ed Huai-hsin Chi, John Riedl, Elizabeth Shoop, John V. Carlis, Ernest Retzel, and Phillip Barry. Flexible information visualization of multivariate data from biological sequence similarity searches. In Roni Yagel and Gregory M. Nielson, editors, *Proceedings of IEEE Visualization 1996*, IEEE Visualization, Annual Conference Series, pages 133–140, San Francisco, USA, 1996. IEEE, IEEE Computer Society Press.
7. Kelly A. Frazer, Laura Elnitski, Deanna M. Church, Inna Dubchak, and Ross C. Hardison. Cross-species sequence comparisons: A review of methods and available resources. *Genome Research*, 13(1):1–12, 2003.
8. Berthold Göttgens, Linda M. Barton, Michael A. Chapman, Angus M. Sinclair, Bjarne Knudsen, Darren Grafham, James G.R. Gilbert, Jane Rogers, David R. Bentley, and Anthony R. Green. Transcriptional regulation of the stem cell leukemia gene (scl) comparative analysis of five vertebrate scl loci. *Genome Research*, 12(5):749–759, 2002.
9. Berthold Göttgens, James G R. Gilbert, Linda M. Barton, Darren Grafham, Jane Rogers, David R. Bentley, and Anthony R. Green. 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(1):87–97, 2001.
10. W Heidrich, M McCool, and J. Stevens. Interactive maximum projection volume rendering. In Gregory M. Nielson and Deborah Silver, editors, *Proceedings of IEEE Visualization 1995*, IEEE Visualization, Annual Conference Series, pages 11–18, Atlanta, USA, 1995. IEEE, IEEE Computer Society Press.
11. T. Hubbard, D. Barker, E. Birney1, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C.Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E.Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.
12. H. Joel Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.
13. Daniel A. Keim, Mihael Ankerst, and Hans-Peter Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In Gregory M. Nielson and Deborah Silver, editors, *Proceedings of IEEE Visualization 1995*, IEEE Visualization, Annual Conference Series, pages 279–288, Atlanta, Georgia, 1995. IEEE, IEEE Computer Society.
14. W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.
15. Nelson L. Max. Visualizing Hilbert curves. IEEE visualization 1998. In David S. Ebert, Holly Rushmeier, and Hans Hagen, editors, *Proceedings of IEEE Visualization 1998*, IEEE Visualization, Annual Conference Series, pages 447–450, North Carolina, USA, 1998. IEEE, IEEE Computer Society Press.
16. C Mayor, M Brudno, J R. Schwartz, A Poliakov, E M. Rubin, Kelly A. Frazer, Lior S. Pachter, and Inna Dubchak. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047, 2000.
17. Hans Sagan. *Space-Filling Curves*. Springer-Verlag, 1994.

18. S Schwartz, L Elnitski, M Li, M Weirauch, C Riemer, A Smit, E D. Green, R C. Hardison, W Miller, and NISC Comparative Sequencing Program. Multipipmaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Research*, 31(13):3518–3524, 2003.
19. Nameeta Shah, Olivier Couronne, Len A. Pennacchio, M Brudno, Serafim Batzoglou, E W. Bethel, E M. Rubin, Bernd Hamann, and Inna Dubchak. Phylovista: interactive visualization of multiple DNA sequence alignments. *Bioinformatics*, 20(5):636–643, 2004.
20. Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages 1996*, pages 336–343. IEEE Computer Society, 1996.
21. James Slack, Kristian Hildebrand, Tamara Munzner, and Katherine St. John. Sequencejuxtaposer: Fluid navigation for large-scale sequence comparison in context. In Robert Giegerich and Jens Stoye, editors, *German Conference on Bioinformatics*, volume 53 of *LNI*. GI, 2004.
22. Rhazes Spell, Rachael Brady, and Fred Dietrich. BARD: A visualization tool for biological sequence analysis. In Tamara Munzner and Stephen North, editors, *IEEE Symposium on Information Visualization, 2003*, pages 219–226. IEEE Computer Society, 2003.
23. Praveen Thiagarajan and Guang Gao. Visualizing biosequence data using texture mapping. In Pak Chung Wong and Keith Andrews, editors, *IEEE Symposium on Information Visualization, 2002*, pages 103–109. IEEE Computer Society Press, 2002.
24. J W. Thomas, J W. Touchman, R W. Blakesley, G G. Bouffard, S M. Beckstrom-Sternberg, E H. Margulies, M Blanchette, A C. Siepel, P J. Thomas, J C. Mcdowell, B Maskeri, N F. Hansen, M S. Schwartz, R J. Weber, W J. Kent, D Karolchik, T C. Bruen, R Bevan, D J. Cutler, S Schwartz, L Elnitski, J R. Idol, A B. Prasad, S.-Q Lee-Lin, V V. B. Maduro, T J. Summers, M E. Portnoy, N L. Dietrich, N Akhter, K Ayele, B Benjamin, K Cariaga, C P. Brinkley, S Y. Brooks, S Granite, X Guan, J Gupta, P Haghighi, S.-L Ho, M C. Huang, E Karlins, P L. Laric, R Legaspi, M J. Lim, Q L. Maduro, C A. Masiello, S D. Mastrian, J C. Mccloskey, R Pearson, S Stantripop, E E. Tiongson, J T. Tran, C Tsurgeon, J L. Vogt, M A. Walker, K D. Wetherby, L S. Wiggins, A C. Young, L.-H Zhang, K Osoegawa, B Zhu, B Zhao, C L. Shu, P J. De Jong, C E. Lawrence, A F. Smit, A Chakravarti, D Haussler, P Green, W Miller, and E D. Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(14):788–793, 2003.
25. Douglas Voorhies. Space-filling curves and a measure of coherence. In James Arvo, editor, *Graphics Gems II*, Graphics Gems, pages 26–30. Academic Press, 1991.
26. David L. Wheeler, Colombe Chappey, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, Gregory D. Schuler, Tatiana A. Tatusova, and Barbara A. Rapp. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 28(1):10–14, 2000.
27. Pak Chung Wong, Kwong Kwok Wong, Harlan Foote, and Jim Thomas. Global visualization and alignment of whole bacterial genomes. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):361–377, 2003.

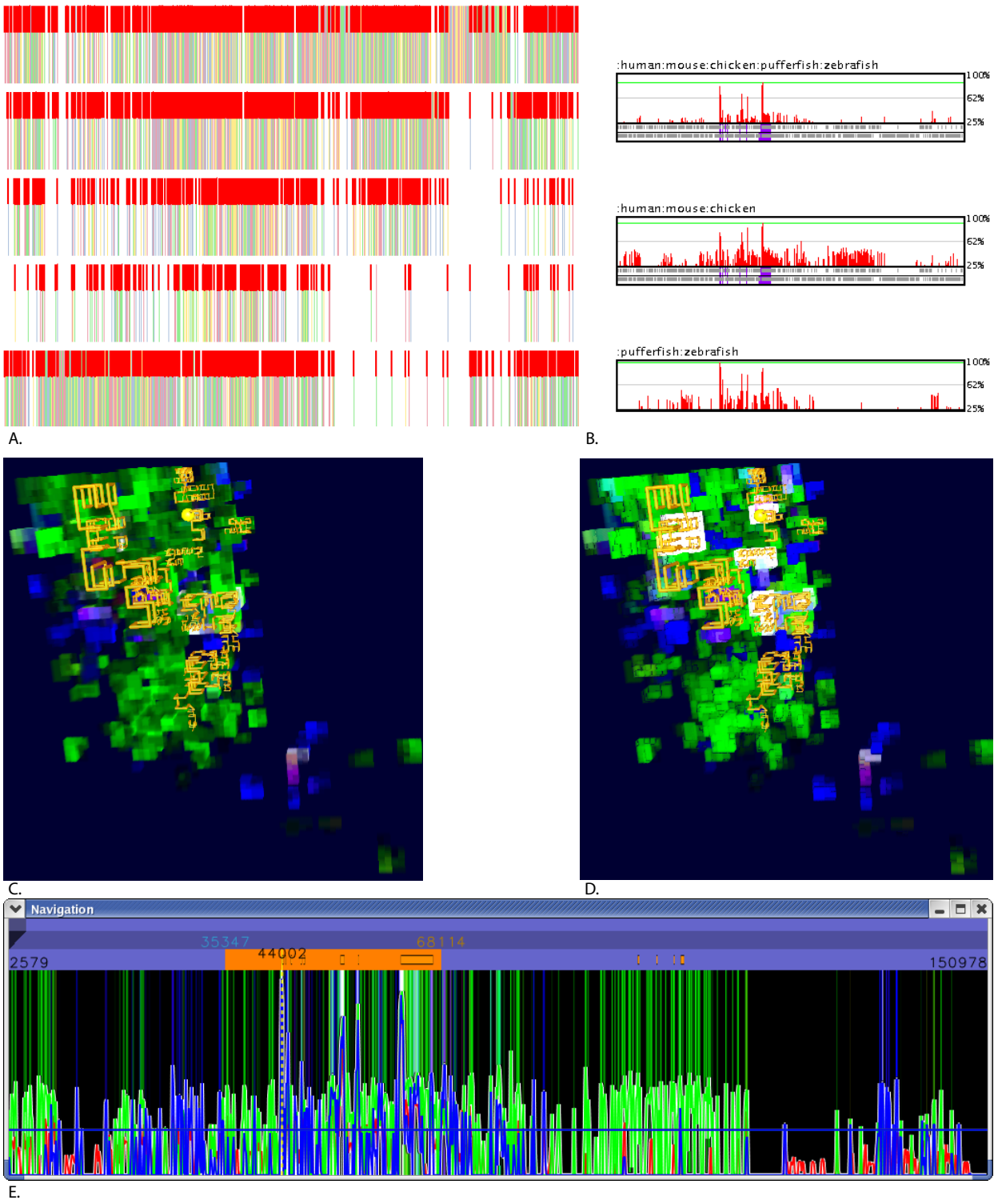

Fig. 7. Using volume rendering to discover a regulatory region for the SCL gene: **A.** Red lines show differences among the five sequences. This image was generated using SequenceJuxtaposer. **B.** 1D plots created with Phylo-VISTA. **C.** The volume-rendered image with yellow pipes showing annotations. The 3D representation occludes some of the interesting features of the data **D.** The white regions in the volume-rendered image correspond to regions that are highly conserved in all sequences. The volume-rendered representation using maximal intensity projection allows users to detect these regions instantaneously, without the need to compare multiple plots. **E.** 1D line representation obtained by overlaying all three similarity plots. White lines show conserved regions.

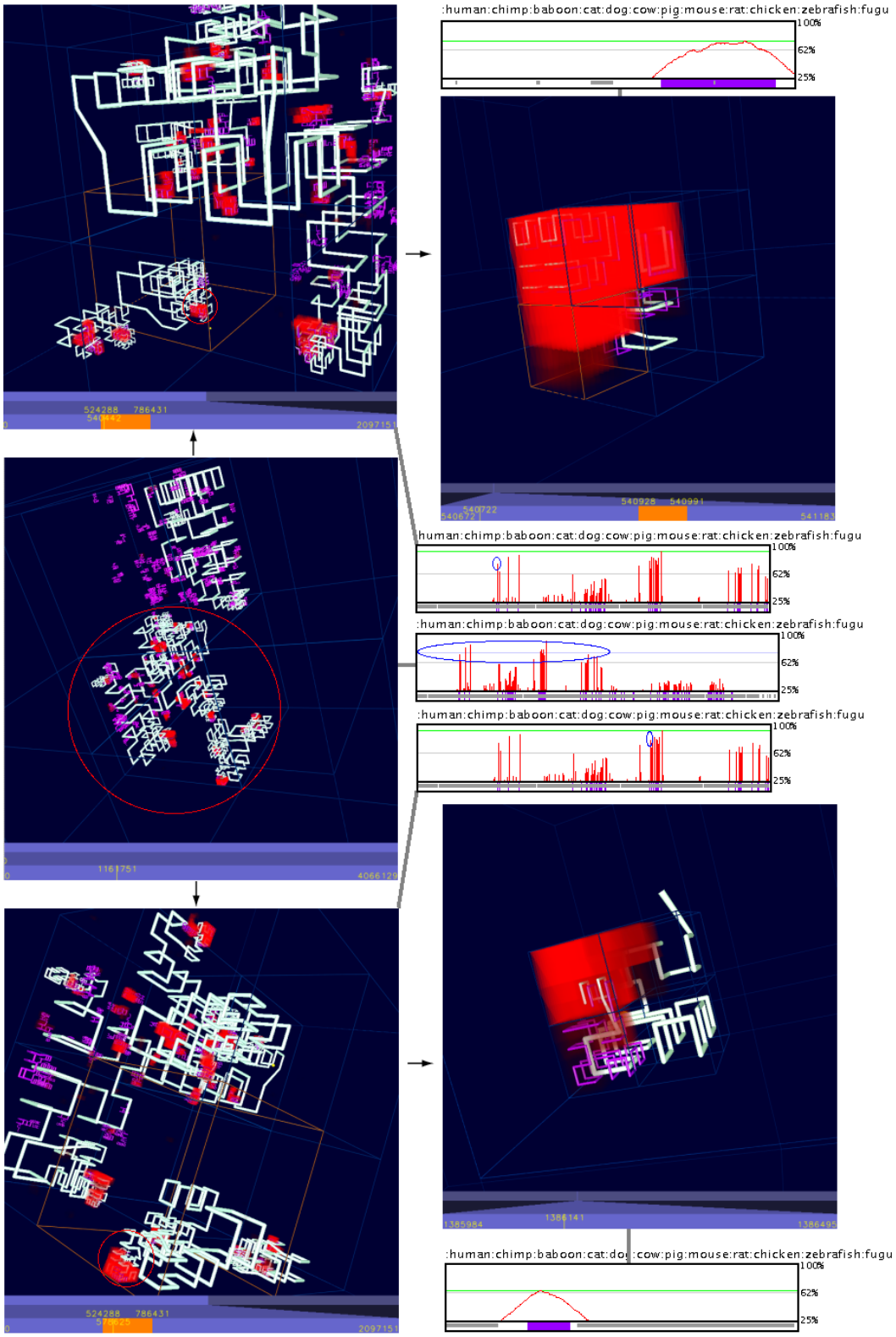

Fig. 8. Visualizing the CFTR data set: The red spots indicate similarities among all species in a region corresponding to the CAPZA2 gene. In the volume-rendered image, exons and their conservation are much more clearly and distinctly visible than in the 1D plot.

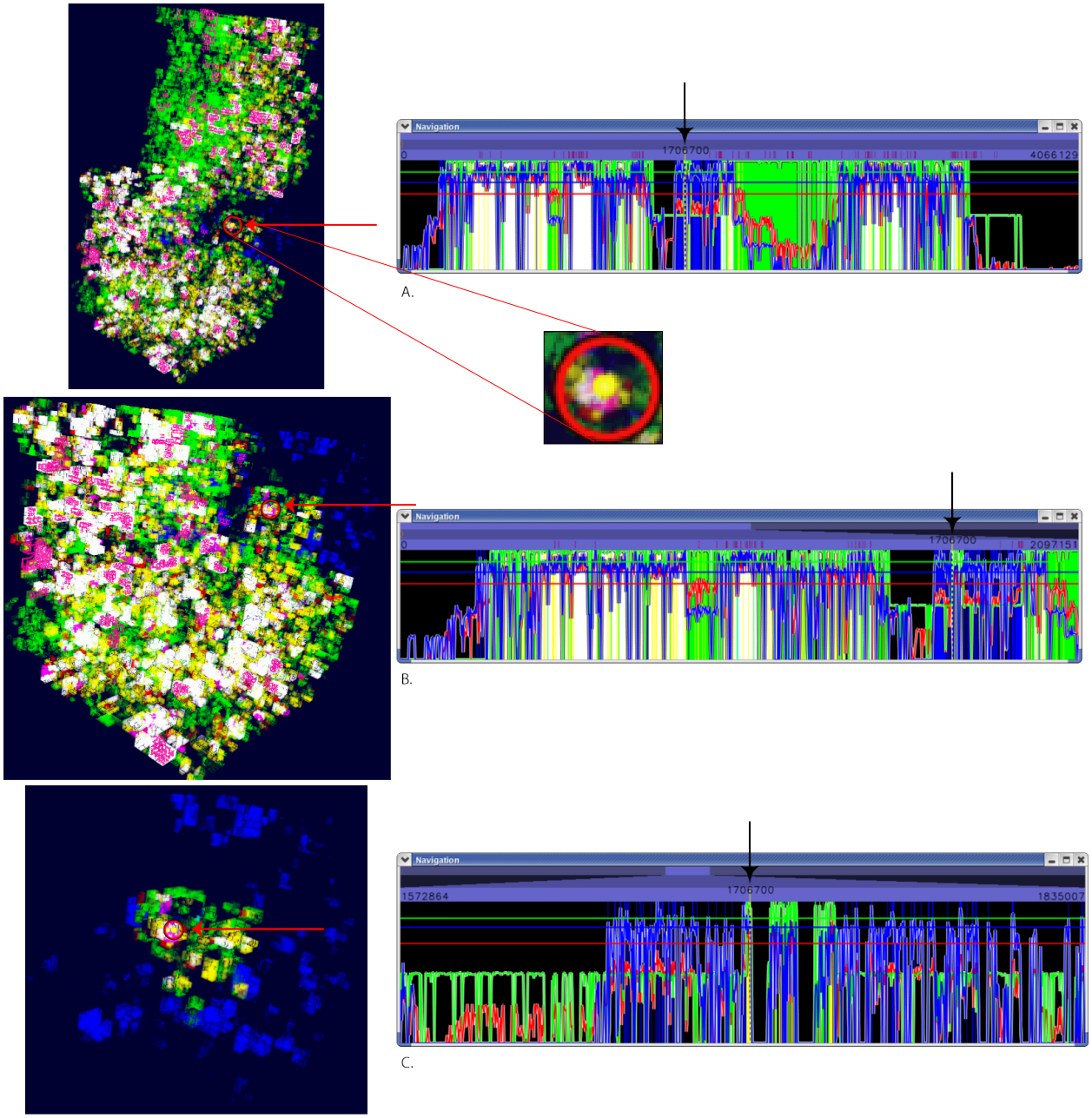

Fig. 9. Visualizing three different similarity scores for the CFTR data set. Larger primate-specific (green) and artiodactyls-carnivores-specific (blue) regions can be seen in both 3D and 1D plots. **A.** A white spot in an otherwise blue-green region (indicated by the red arrow). The black arrow indicates the corresponding region in 1D plot with the feature not visible. **B, C.** Zoomed-in views with the feature in the 3D plots being invisible in 1D plots. (Figure continued on next page)

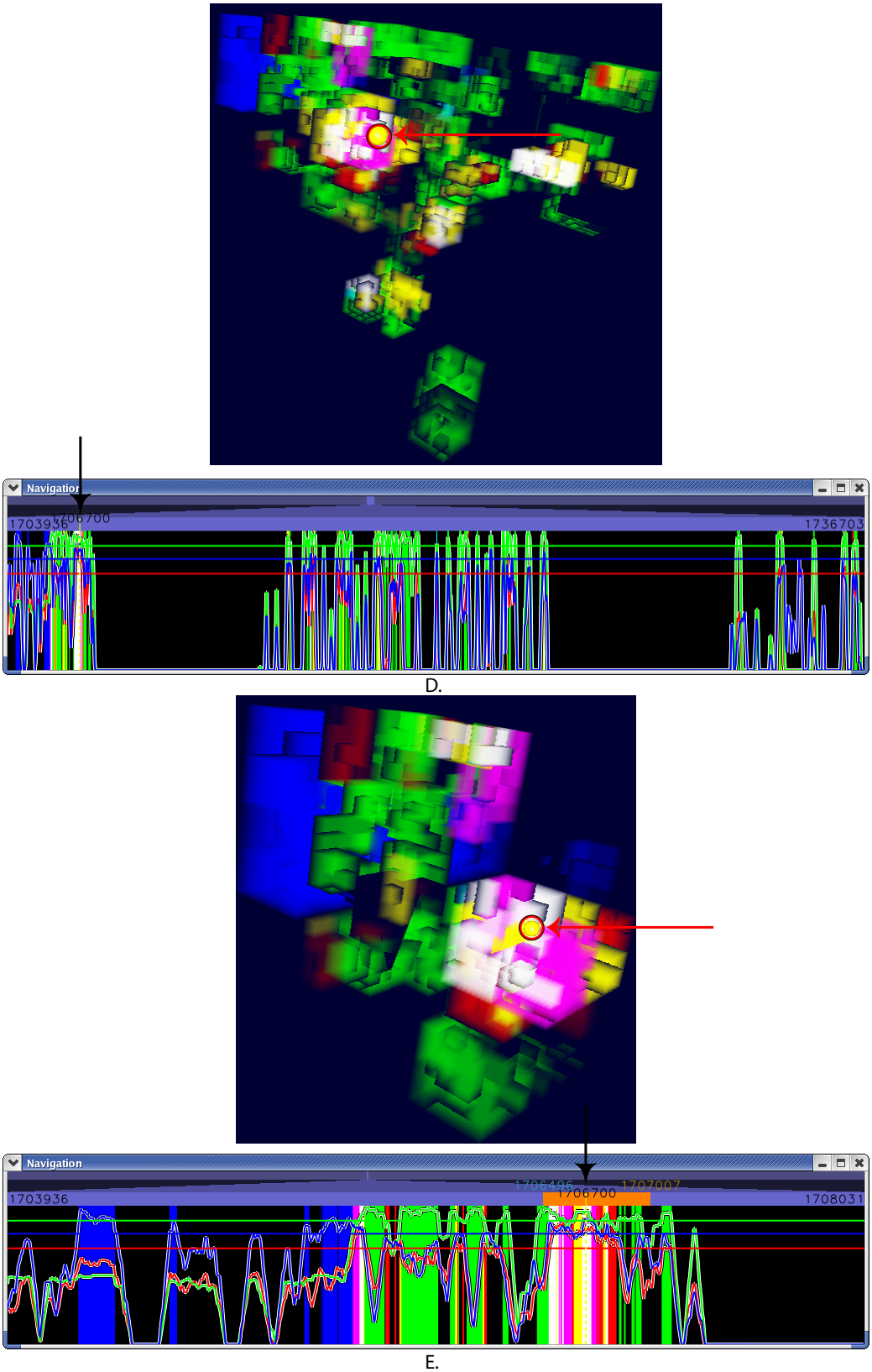

Fig. 10. Zoomed-in views of three different similarity scores for the CFTR data set. **D.** A feature visible in the 3D plot is visible as a thin, white line in the corresponding 1D plot. **E.**A feature visible in the 3D plot is clearly visible in the 1D plot after zooming-in.