Supporting Detail-in-Context for the DNA Representation, H-Curves

M. L. Lantin [*] M. S. T. Carpendale [†]

School of Computing Science
Simon Fraser University

Abstract

This paper presents a tool for the visual exploration of DNA sequences represented as H-curves [7]. Although very long sequences can be plotted using H-curves, micro-features are lost as sequences get longer. We present a new three-dimensional distortion algorithm to allow the magnification of a sub-segment of an H-curve while preserving a global view of the curve. This is particularly appropriate for H-curves as they provide useful visual information at several resolutions. Our approach also extends the current possibilities of detail-in-context viewing in 3D. It provides a non-occluding, orthogonal technique that preserves uniform scaling within regions and maintains geometric continuity between regions.

1 INTRODUCTION

As you read this paper, sequencing projects around the world are generating massive amounts of DNA sequence data. As of June 1998 [5], the Genbank database contained 1,622,000,000 base pairs in 2,356,000 sequence records. The analytic challenges posed by this vast quantity of data can be usefully addressed through visualization tools. Ideally such tools will increase researchers' ability to analyze the information contained in these records, enabling them to more readily identify regions of interest such as repetitive subsequences within large DNA sequences.

The motivation for this work has arisen from observing biologists at work with existing text and visual representations, from published requests for improved visual access [7], and from personal communication from scientists in the field. The importance of understanding genetic sequences coupled with the difficulty of working with such immense volumes of data underscores the urgent need for supportive visual tools [1].

We examine augmenting the usability of H-curves [7], a powerful visual representation designed explicitly for DNA sequence data. H-curves have been used to display entire genomes and to detect features in sequences such as a change in the DNA template-strand transcribed, overlapping genes and repetitive sub-sequences. They are also suited to comparing global features among sequences, as sequences from similar genome families are expected to have similar codon biases [6]. This paper presents an orthogonal 3D distortion viewing tool with which a specific area of the H-curve can be magnified without losing the global setting in which it is embedded, or interfering with the positional information that is integral to H-curves.

2 DNA REPRESENTATIONS

Providing good visual access to extremely long sequence information such as DNA data is a challenge. The most straight-forward way of simply listing the nucleotides as text from beginning to end limits the number of nucleotides that can be reasonably displayed simultaneously to a few hundred making it difficult to identify distinguishing characteristics. This type of representation is so uniform as to hide all distinguishing features of the sequence being displayed.

Line diagrams such as those used in ACEDB [3] (Figure 1) are a relatively compact and effective way of displaying annotated sequence information. However, due to their linear form, they are not suitable for visual identification of unannotated sequence features.

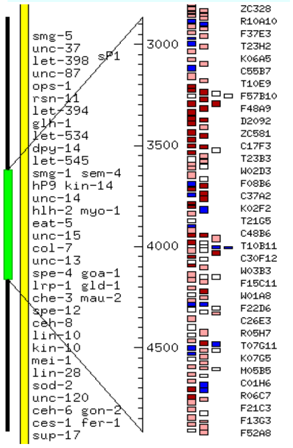


Figure 1: ACEDB line diagram of a section of map sequence I

G-curves were introduced by Hamori [7, 6] as a five-dimensional graphical representation of DNA sequences. The first four dimensions correspond to the nucleotides A, G, C, T respectively, and the fifth to the location of a particular nucleotide within the sequence. To plot the sequence in 5-space one starts at $(0,0,0,0,0)$ and for each nucleotide in the sequence moves one unit in the fifth dimension and one unit in the corresponding dimension of the nucleotide. For example, the sequence ACT would generate the following sequence of points:

$$\begin{array}{c}(0,0,0,0,0)\\(1,0,0,0,1)\\(1,0,1,0,2)\\(1,0,1,1,3)\end{array}$$

To provide a more readable 3D display, a 5D G-curve can be mapped to a 3D H-curve. For an H-curve, each type of nucleotide is assigned a base vector. Starting at position $(0, n, 0)$ [1], where n is the number of nucleotides in the sequence being mapped, the base vectors corresponding to each nucleotide are added sequentially to the curve. The base vectors are assigned such that they all point downward towards a different corner of the xz plane (Figure 2).

Using the base vectors shown in Figure 2, the H-curve corresponding to the G-curve example above would consist of the points:

[*] lantin@cs.sfu.ca

[†] carpenda@cs.sfu.ca

[1] In the 3D space used for this paper, the y axis runs up-down, the x axis right-left and the z axis perpendicular to the page.

$$(0,3,0)$$
$$(1,2,1)$$
$$(0,1,0)$$
$$(1,0,-1)$$

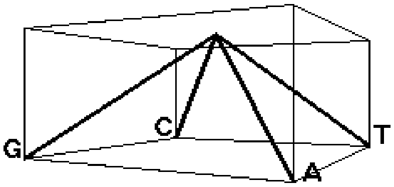


Figure 2: Base vectors

The visual representation of the same H-curve is shown in Figure 3. Even though the nucleotides can be recognized from orientation alone, they are also colour coded for ease of recognition; A is yellow, T is green, C is cyan and G is magenta (see colour plate Figures 7(a) and 8).

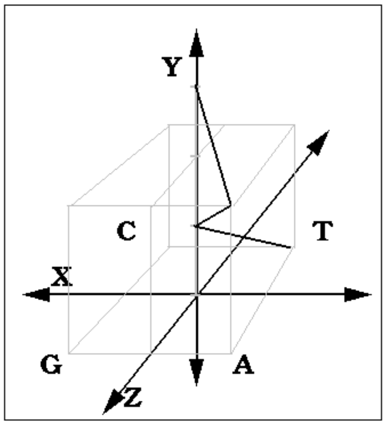


Figure 3: H-curve for the sequence ACT

Once plotted the H-curve can be rotated or the base vectors reassigned to provide different views. A 2D projection of the H-curve is equivalent to a change of alphabet (Figure 4). For example, using the base vector assignment shown in Figure 2, a projection onto the yz plane changes the alphabet of the sequence from $\{A, G, C, T\}$ to $\{AG, CT\}$ which is equivalent to the biological concepts of Purines and Pyrimidines. H-curves mirror the regularities and irregularities of a given DNA sequence. Long stretches of purines followed by pyrimidines, for example, are easily identifiable as sharp turns in the curve and repeats are visually apparent as repeating curve segments. The end point of the curve (and any segment within the curve), is an important indicator of the relative frequency of nucleotides in the sequence.

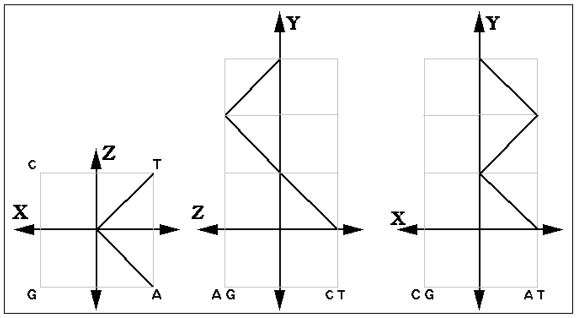
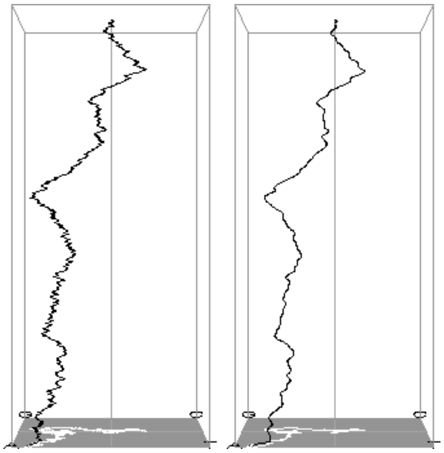


Figure 4: The three projections for H-curve of the sequence ACT



(a) H-curve

(b) Smoothed H-curve

Figure 5: The H-curve representation of the Bovine Gene for epsilon (2) beta-globin (BTEBGL2). First 885 nucleotides are displayed.

To subdue local fluctuations in the H-curve when displaying large sequences, it is possible to smooth the H-curve by an averaging technique. For a given smoothing factor, w, the vector at position i in the curve is replaced with the average of the base vectors of positions $i - w$ through $i + w$. That is, for every section of $2w + 1$ nucleotides being averaged, the vector formed by the start and end points of the section is normalized by $2w + 1$ and used as the new base vector. This has been shown to reduce the visual clutter of local detail while preserving the essential long-range characteristics of the H-curve [7](Figure 5).

3 AUGMENTING VISUAL ACCESS

There are several salient properties of the H-Curve DNA representation. They are a visually comprehensible 3D representation of 5D data. They make innovative use of relative position to encode the additional dimensions, creating a representation that can be usefully read in all of its 2D projections as well as its 3D form. Rotating an H-curve changes which of the DNA sequence aspects are revealed. When presented compactly, they are capable of displaying very large sequences and reveal global features. When presented in high resolution they reveal individual residues in subsequences. As a visual representation H-curves hold full data information while still allowing drill-down for explicit details. For instance, at high resolutions the vector angle encodes the nucleotide type, and relative position indicates sequential information.

Providing visual drill-down to the details in a representation without losing the global features corresponds to the goals of focus and context or distortion viewing techniques. However, as H-curves reveal information both compressed and magnified, our intention is to provide a distortion technique that goes beyond setting readable details in context by presenting combined displays that maintain

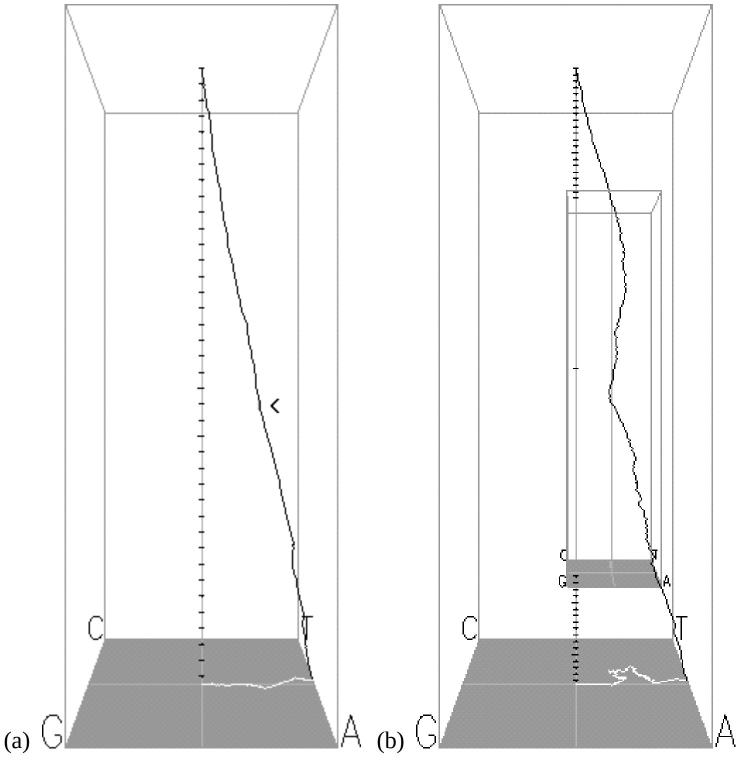


Figure 6: (a) H-curve region marked but unzoomed. The zoom box is located at the arrow and is almost indistinguishable. (b) The H-curve representation of the human beta globin region on chromosome 11. Nucleotides 1-38505 displayed. The zoomed region starts at position 20759 and extends 542 nucleotides. The tick marks on the main Y axis are representative of the scale. There are 1000 nucleotides between each tick mark.

the useful visual characteristics of H-curves at each resolution. To achieve this goal there are several aspects of the representation that must be respected. Orthogonality [13], or left/right, up/down ordering must be preserved. Also, in order to maintain the ability to read the position of segments in relation to each other it is important to preserve the connectivity between them. As H-curves are readable at several resolutions uniform scaling must be maintained with regions. Further, orientation, or individual vector angles must be maintained, as they encode nucleotide type and provide global features like drift.

While there has been extensive investigation of this type of viewing for 2D representation, (for surveys see [8, 10, 11]) There has been much less investigation of applying distortion ideas for 3D representations. Semnet [4], the first approach for 3D representations, makes use of filtering which is not appropriate for our goals. Mitra's approach, intended for aircraft maintenance diagrams [9], creates filtered exploded views which would interfere with connectivity making H-curves unreadable. The 3D visual access technique described in [2] makes use of non-linear distortion techniques to address occlusion problems in 3D displays. This technique would also cause unwanted separation, and further, the elegant sparse display of H-curves does not pose an occlusion problem. While 3D Magic Lenses [14] are capable of providing local magnification, they operate as a 3D insert thereby lose of context and connectivity between magnified and non-magnified sections. What is desired is a 3D detail in context technique that utilizes a step magnification/compression function and provides geometric continuity between the sections of differing but constant scale. More closely related to these goals are the orthogonal approaches [1, 12, 13], however, they have only been developed for 2D representations and have not considered preserving connectivity and actual start and end positioning between regions of differing scale.

4 METHOD

To provide a distortion viewing environment appropriate for H-curves, we use the familar approach of compressing the unzoomed segments to absorb the extra space used by the zoomed region. However, the distortion algorithm presented here preserves the essential features of the H-curve such as the end point and the position of the curve relative to the central axis.

For this discussion, the space used a by an H-curve segment refers to the displacement of the end point relative to the start point of the segment. From this point onward, it will be referred to as simply the displacement of the segment. The transformation is achieved by scaling the individual components of the base vectors such that the scaled displacement of the unzoomed region added to the magnified displacement of the zoomed region, is equal to the displacement of the original H-curve (Equation 1). This satisfies the constraint that the end point of the H-curve remain the same. The distortion is applied independently and uniformly in each dimension, depending on the amount of displacement to be absorbed.

$$\begin{aligned} sf &= \text{scale factor} \\ zf &= \text{zoom factor} \\ uzdisp &= \text{unzoomed displacement} \\ zdisp &= \text{zoomed displacement} \\ uzdisp * (1 - sf) &= zdisp * (zf - 1) \qquad (1) \end{aligned}$$

The scaling factor is limited to the range .1 and 2 in all dimensions. We find that this permits sufficient local zooming while preserving essential features of the unzoomed segments. A negative scaling factor would reverse displacement causing incorrect readings, a scaling factor too close to zero would cause excessive flattening.

Because H-curves can have quite irregular patterns, it is essential to provide visual cues to the location and extent of the zooming and distortion. The location of the zoomed curve segment is marked on the H-curve by a box surrounding the segment, similar to the one surrounding the whole H-curve (Figure 6(b)). The zoom factors for each dimension are controlled from a dialog window, while the screen is continuously updated to reflect user input. As the user moves the location or extent of the zoomed segment, the maximum zooming factors are updated to reflect the newly calculated constraints.

5 DISCUSSION AND FUTURE WORK

The local zooming feature of this tool has been tested using sequences ranging from 200 to 80,000 nucleotides and revealed sequence characteristics that may have otherwise been hard to find. The box surrounding the magnified region acts as a non-occluding 3D lens. By providing appropriate compensating compression, our tool retains all context and maintains sequence connectivity while displaying more than one level of resolution. Moving the self-adjusting magnifying box along the curve allows for interactive exploration.

For additional visual functionality, 3D smoothing or sharpening lens could be useful. Automatic smoothing could be provided for the individual segments of the H-curve, particularly those that have been compressed. The smoothing factor needed could be estimated by calculating how many line segments are occupying a single pixel on the screen and setting the smoothing factor such that no more than one line segment is drawn per pixel. This would be useful especially at low resolutions where the local detail combined with the distortion obscures the global characteristics of the curve.

As a starting point towards supporting H-curves for viewing annotated sequences, we colour H-curve regions according to nucleotides (Colour Plate: Figure 7(a)) and exons (Colour Plate: Figure 7(b)). We note that the use of hyper-linked markers along the curve to point out other sequence features, and further a mechanism to add markers with associated annotations would be useful additions.

6 CONCLUSION

We present a 3D distortion technique especially suited to H-curves, preserving those aspects of H-curves that biologists have identified as particularly valuable for identification of unannotated sequences features at differing resolutions. In contrast to traditional zooming methods which offer single resolution views, or local zooming with occlusion, our method provides a controllable 3D magnification lens which uses automatic compression to prevent occlusion. This distortion technique is capable of displaying H-curves at more than one resolution without disrupting continuity, global positioning and drift. In conjunction with traditional zooming techniques, we believe that this significantly enhances the utility of H-curves. We look forward to working further with molecular biologists to provide an increasingly enhanced DNA sequence exploration environment.

ACKNOWLEDGMENTS

This work was supported in part by a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada. SGI hardware resources were provided by the Graphics and Multimedia Research Lab at Simon Fraser University. The authors are grateful for helpful and insightful suggestions by Gabor Melli and David Cowperthwaite.

References

[1] L. Bartram, A. Ho, J. Dill, and F. Henigman. The continuous zoom: A constrained fisheye technique for viewing and navigating large information spaces. In *UIST'95: Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 207–216. ACM Press, 1995.

[2] M. S. T. Carpendale, D. J. Cowperthwaite, and F. D. Fracchia. Extending distortion viewing from 2D to 3D. *IEEE Computer Graphics and Applications*, 17(4):42–51, July/August 1997.

[3] Richard Durbin and Jean Thierry Mieg. A Caenorhabditis elegans database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk or ncbi.nlm.nih.gov, 1991-.

[4] K. M. Fairchild, S. E. Poltrock, and G. W. Furnas. Semnet: Three-dimensional graphic representation of large knowledge bases. In *Cognitive Science and its Applications for Human-Computer Interaction*, pages 201–234. Lawerence Erlbaum Associates, 1988.

[5] NIH: National Center for Biotechnology Information. Genbank overview. http://www.ncbi.nlm.nih.gov/Web/Genbank/, June 1997.

[6] Eugene Hamori. Visualization of biological information encoded in DNA. In Clifford A. Pickover and Stuart K. Tewksbury, editors, *Frontiers of Scientific Visualization*, volume 3 of *Scientific Visualization*, chapter 4, pages 90–121. Wiley-Interscience, 1994.

[7] Eugene Hamori and John Ruskin. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *The Journal of Biological Chemistry*, 258(2):1318–1327, January 1983.

[8] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM ToCHI*, 1(2):126–160, 1994.

[9] D. A. Mitra. A fisheye presentation strategy: Aircraft maintenance data. In *Human-Computer Interaction - INTERACT '90*, pages 875–880, 1990.

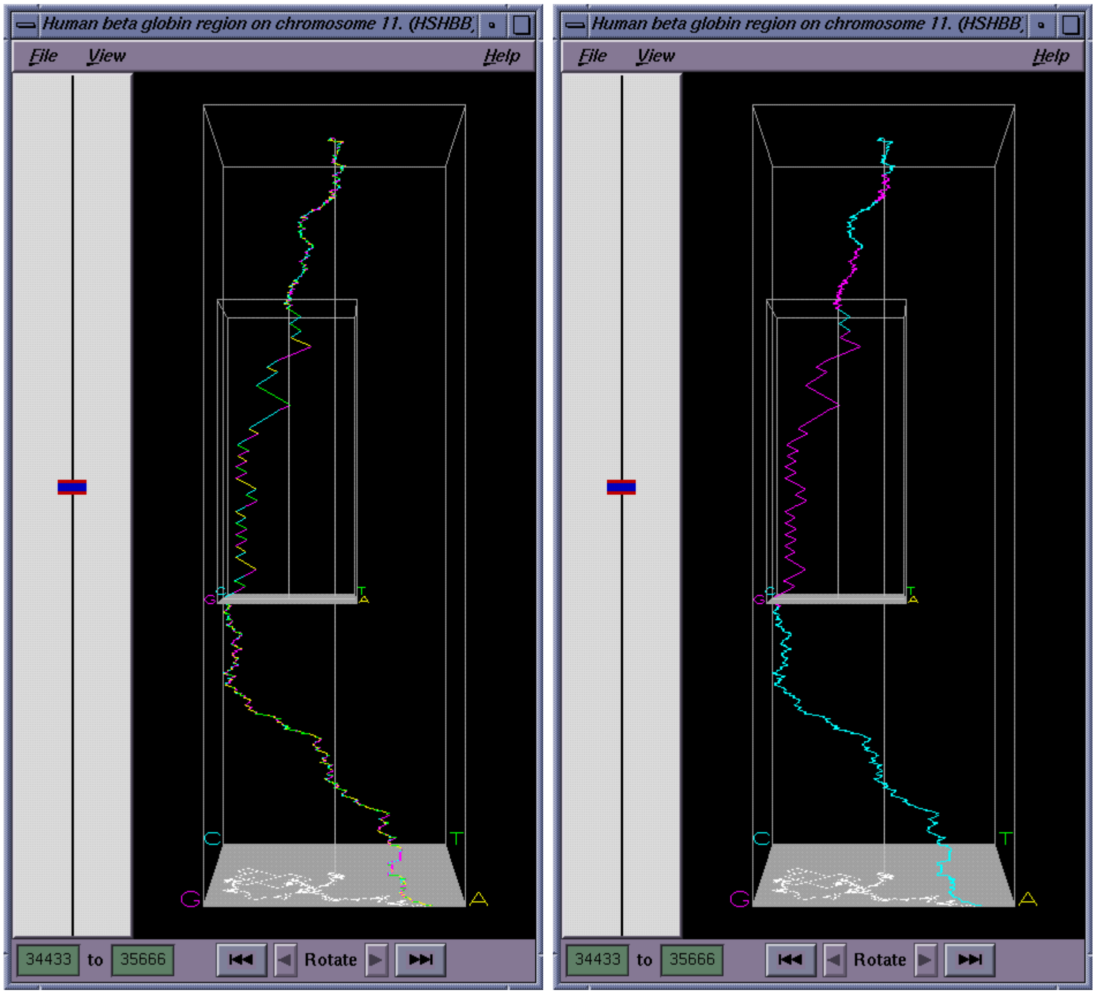
[10] E. Noik. A space of presentation emphasis techniques for visualizing graphs. In *Proceedings of Graphics Interface '94*, pages 225–233, May 1994.

[11] R. Spence. A taxonomy of graphical presentation. Information Engineering Section report 93/3, Imperial College of Science, Technology and Medicine, 1993.

[12] R. Spence and M. Apperly. Data base navigation: an office enviroment for the professional. *Behaviour and Information Technology*, 1(1):43–54, 1982.

[13] M. A. Storey and H. A. Müller. Graph layout adjustment strategies. In *Graph Drawing '95*, pages 487–499, 1995.

[14] John Viega, Matthew J. Conway, George Williams, and Randy Pausch. 3D magic lenses. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 51–58. ACM, ACM Press, November 1996.


Human beta globin region on chromosome 11. (HSHBB
File
View
Help
G
C
T
A
Rotate
Human beta globin region on chromosome 11. (HSHBB
File
View
Help
G
C
T
A
Rotate

34433 to 35666

34433 to 35666

(a) Coloured by residue

(b) Coloured by exon

Figure 7: H-curve of human beta globin on chromosome 11 (nucleotides 34433-35666 shown). Zoomed region is 50 nucleotides long and is magnified approximately tenfold.

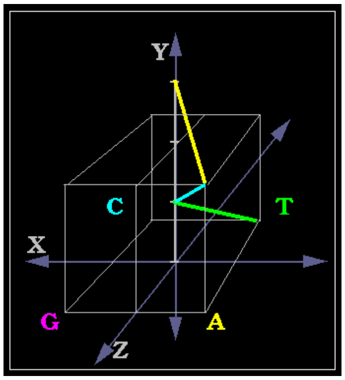


Figure 8: The H-Curve for the sequence ACT with the nucleotides colour-coded. A is yellow, C is cyan, and G is magenta.