Computer Visualization of Long Genomic Sequences

Dachywan Wu, Department of Computer Science
James Robergé, Department of Computer Science
Douglas J. Cork, Department of Biology
Bao Gia Nguyen, Department of Mathematics
Thom Grace, Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616

Abstract

Human beings find it difficult to analyze local and global oligonucleotide patterns in the linear primary sequences of a genome. In this paper, we present a family of iterated function systems (IFS) that can be used to generate a set of visual models of a DNA sequence. A new visualization function, the W-curve, that is derived from this IFS family is introduced. Using W-curves, a user can readily compare subsequences within a long genomic sequence — or between genomic sequences — and can visually evaluate the effect of local variations (mutations) upon the global genomic information content.

1 Introduction

The advent of biological sequencing technologies has led to a substantial growth in the quantity of biological sequencing information available to researchers [1]. This rapid proliferation of new sequences has, in turn, dramatically increased the need for fresh approaches to the analysis of genomic sequences. The types of analyses required range in complexity from the straightforward task of locating a specific six-base restriction site in a DNA sequence, to the highly complex task of identifying regions of similarity (or interest) both within and between sequences.

The emergence of large databases of genomic sequences [2, 3] have spawned efforts in which scientists examine classes of genomes (e.g., genomic data from the HIV viruses) searching for "interesting" nucleotide patterns. The biological and statistical significance of these patterns is then assessed, and hypotheses developed based upon the more significant patterns.

Traditionally, DNA sequences have been represented as long strings composed of the letters A, C, G, and T (for Adenine, Cytosine, Guanine, and Thymine, respectively). While this representation of genomic data is compact, human analysis of the patterns and structures in a genomic sequence represented in this form is poor at best. Long strings of characters do not readily convey important statistical properties (e.g., characteristic nucleotide ratios of certain subsequences, sudden changes in prevailing nucleotide compositions at certain locations, and so forth). More importantly, the recognition, recollection, and comparison of long character subsequences is nearly impossible for most people [4].

Algorithmic techniques for comparing genomic sequences based upon this representation scheme have focused upon the application of methods for measuring the similarity of pairs of genomic sequences [5]. Existing methods include matching specific subsequences in sequence pairs [6, 7], determining the minimal changes needed to transform one sequence into another [8], and measuring the degree of difference (distance) between sequences (or equivalently, the degree of similarity between them) [9]. Effective application of these algorithmic measures is greatly complicated by the sheer length of genomic sequences. In addition, the occurrence of small variations throughout genomic sequences make it difficult to develop measures that accurately account for variations in both the nature and the alignment of the subsequences in long genomic sequences.

Rather than encoding genomic data as strings of letters, this information can be represented in a more visual form. The intent in adopting a more visual representation is to allow a scientist to apply the extensive pattern analysis infrastructure inherent in the human visual system to the search (and evaluation) of nucleotide patterns.

Hamori and Ruskin [4] introduced the H-curve as

1070-2385/93 $3.00 © 1993 IEEE

just such a visual representation for studying genomic data. Their representation scheme maps each of the letters A, C, T, and G to a specific three-dimensional vector. By concatenating vectors head to tail, the sequence of letters encoding a genome can be transformed into a curve (or walk) in $\mathbb{R}^3$. Using this visualization technique, analysis of the structures and patterns within (and between) DNA sequences can be done in terms of the shapes of the resulting curves (Fig. 4 and 5).

The chaos game representation developed by Jeffrey [10, 11] uses an iterated function system (IFS) [12, 13] to map the letters corresponding to a DNA sequence to a set of points in $\mathbb{R}^2$. The resulting point set is not joined by line segments (as in the H-curve), but rather is displayed in its totality, with variations in the resulting pattern of points revealing properties of the corresponding genomic sequence (Fig. 1).

In this paper, we introduce a family of iterated function systems that can be used in the visualization of genomic data. This family includes both the H-curve and chaos game representation outlined above. More importantly, it includes a broad range of additional visual representation schemes. We will examine one such scheme, the W-curve, and show that it provides a three-dimensional representation that is compact, and that can be used to visually identify patterns at both the sequence (global) and subsequence (local) level.

2 An IFS for modeling DNA sequences

A DNA sequence of length n can be represented using a character sequence

$$\boldsymbol{\lambda} = \lambda_1, \lambda_2, \ldots, \lambda_n,$$

where $\lambda_i \in \{A, C, G, T\}$ for each $i = 1, 2, \ldots, n$. By associating each letter in $\{A, C, G, T\}$ with a two-dimensional vector using the following mapping

$$\begin{array}{ll} A: & \delta_A = (-1, -1), \\ C: & \delta_C = (-1,\ 1), \\ G: & \delta_G = (\ 1, -1), \\ T: & \delta_T = (\ 1,\ 1), \end{array}$$

the character sequence $\boldsymbol{\lambda}$ can be mapped to the sequence of vectors $\boldsymbol{\xi}$

$$\boldsymbol{\xi} = \xi_1, \xi_2, \ldots, \xi_n,$$

where $\xi_i \in \{\delta_A, \delta_C, \delta_G, \delta_T\}$ for each $i = 1, 2, \ldots, n$. This vector sequence can then be used as the basis for the family of iterated function systems $\mathcal{F}$

$$\mathcal{F} : \mathbb{R}^2 \times \{\delta_A, \delta_T, \delta_G, \delta_C\} \rightarrow \mathbb{R}^2$$

defined by

$$\begin{aligned} X_i &= \mathcal{F}(X_{i-1},\ \xi_i) \\ &= \alpha X_{i-1} + \beta \xi_i, \end{aligned} \tag{1}$$

where ξ_i is defined as above, $X_0 = (0, 0)$, and $\alpha, \beta \in \mathbb{R}$.

The sequence of values $\{X_i | i \in \mathbb{N}, 1 \leq i \leq n\}$ can be visualized in several ways: as a set of points in $\mathbb{R}^2$, as a set of concatenated vectors in $\mathbb{R}^2$, or as a set of points/vectors $\{(X_i, i) | i \in \mathbb{N}, 1 \leq i \leq n\}$ in $\mathbb{R}^3$. Note that if X_i denotes the point $(x, y) \in \mathbb{R}^2$, then (X_i, i) denotes the point $(x, y, i) \in \mathbb{R}^3$.

Varying α and β produces the various members of IFS family $\mathcal{F}$. In this paper, we will examine three members of this family

$$\begin{array}{ll} \mathcal{F}_1: & \alpha = \beta = 1, \\ \mathcal{F}_2: & \alpha = \beta = 1/2, \\ \mathcal{F}_3: & \alpha = \beta = 1/k. \end{array}$$

We begin by showing that the first two members of this group correspond to existing visualization techniques — the H-curve and chaos game representation, respectively. We then introduce the third member, the W-curve, and analyze its properties as a visualization technique in terms of the first two members.

2.1 IFS $\mathcal{F}_1$: H-curve

Hamori and Ruskin [4] define the H-curve in the following manner. Let z denote the position number of a nucleotide in an arbitrary DNA sequence $\boldsymbol{\lambda}$ of length n, and let $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ denote the unit vectors along the x, y, and z axes, respectively. If we define

$$g(z) = \begin{cases} \mathbf{i} + \mathbf{j} + \mathbf{k} & \text{if } \lambda_z = A \\ \mathbf{i} - \mathbf{j} + \mathbf{k} & \text{if } \lambda_z = T \\ -\mathbf{i} - \mathbf{j} + \mathbf{k} & \text{if } \lambda_z = C \\ -\mathbf{i} + \mathbf{j} + \mathbf{k} & \text{if } \lambda_z = G \end{cases}$$

and let

$$h(z) = \sum_{i=1}^{z} g(i)$$

then the three-dimensional curve traced out by $h(z)$ as z increases from 1 to n is the H-curve of $\boldsymbol{\lambda}$. Expanding $h(z)$ once, we see that

$$h(z) = \sum_{i=1}^{z} g(i)$$

$$= \sum_{i=1}^{z-1} g(i) + g(z)$$
$$= h(z-1) + g(z).$$

If we let $\alpha = \beta = 1$, then Equation (1) becomes

$$X_z = X_{z-1} + \xi_z,$$

Note that ξ_z is the projection of $g(z)$ onto the xy-plane and that the projection of $g(z)$ along the z-axis is the unit vector $\mathbf{k}$. Thus, the curve described by $\{(X_z, z) | z \in \mathbb{N}, 1 \leq z \leq n\}$ in xyz-space is precisely the H-curve.

2.2 IFS $\mathcal{F}_2$: Chaos game representation

The chaos game representation of a DNA sequence, introduced by Jeffrey [10, 11], can be described as follows. Starting from the origin of the xy-plane, iteratively mark n points $\{X_i | i \in \mathbb{N}, 1 \leq i \leq n\}$ using the rule that a point X_i is the midpoint of the line segment connecting points X_{i-1} and ξ_i, where ξ_i is defined as above.

Simple geometry reveals that

$$X_i = X_{i-1} + \frac{1}{2}(\xi_i - X_{i-1}),$$

which is the same as Equation (1) with $\alpha = \beta = 1/2$. Thus, $\{X_i | i \in \mathbb{N}, 1 \leq i \leq n\}$ is the chaos game representation of the sequence $\boldsymbol{\lambda} = \lambda_1, \lambda_2, \ldots, \lambda_n$.

2.3 IFS $\mathcal{F}_3$: W-curve

Having seen that these existing techniques for visualizing genomic sequences are instances of the IFS family described by Equation (1), we wondered whether there exist other members of this family that might produce new insights into the patterns and properties of genomic sequences.

Another member is the set of W-curves. A W-curve is the set of points in $\mathbb{R}^3$ $\{(X_i, i) | i \in \mathbb{N}, 1 \leq i \leq n\}$ that is produced by restricting α and β in Equation (1) to values in the set $\{\frac{1}{k} | k \in \mathbb{N}, k > 1\}$. The patterns in the zy- or zx-projection of a W-curve that is produced using large values of k have a characteristic W-shaped form, thus the name W-curve (Fig. 6 and 7).

2.4 Properties of the W-curve

Reflecting their membership in IFS family $\mathcal{F}$, W-curves have a combination of the properties of H-curves and the chaos game representation. In the case of W-curves, this combination results in a visual representation that allows a user to easily inspect and compare both local and global features in long genomic sequences.

Like an H-curve, the z-coordinate of a point on a W-curve is the position number of the corresponding nucleotide in the defining genomic sequence. Thus, an zy- or zx-projection of a W-curve provides an ordered view of the nucleotides in the sequence (Fig. 2a). This ordering is important in that it allows a user to easily compare subsequences by moving along the z-axis in either of these projections (that is, by moving from left to right in Figure 2a). Note that the points in the chaos game representation are not ordered in any visually meaningful manner.

Unlike H-curves, W-curves are constrained to lie within a distance of one from the z-axis. As a result, W-curves are in the space of $\Psi^2 \times \mathbb{N}$, where Ψ is the open interval $(-1, 1)$, as opposed to H-curves which are in the space of $\mathbb{R}^2 \times \mathbb{N}$. The constrained space occupied by W-curves makes the resulting display more compact, and thus more manageable.

Like the chaos game representation, each point X_i on a W-curve encodes the history of the sequence from the first nucleotide to the nucleotide corresponding to point X_i. By simple substitution into Equation (1), we obtain

$$\begin{aligned} X_i &= \beta\xi_i + \alpha X_{i-1} \\ &= \beta\xi_i + \alpha(\beta\xi_{i-1} + \alpha X_{i-2}) \\ &= \beta\xi_i + \beta\alpha\xi_{i-1} + \alpha^2(\beta\xi_{i-2} + \alpha X_{i-3}). \end{aligned}$$

Iterating this calculation further we derive

$$X_i = \beta(\xi_i + \alpha\xi_{i-1} + \alpha^2\xi_{i-2} + \ldots + \alpha^{i-1}\xi_1) + \alpha^i X_0. \quad (2)$$

Scaling Equation (2) by

$$\hat{X}_i = \left(\frac{X_i - \alpha^i X_0}{\beta}\right)$$

yields

$$\hat{X}_i = \xi_i + \alpha\xi_{i-1} + \alpha^2\xi_{i-2} + \ldots + \alpha^{i-1}\xi_1. \quad (3)$$

When $\alpha \in \mathbb{N}$ and $\alpha > 1$, Equation (3) is the expansion of $\hat{X}_i$ in terms of the base α number system. Similarly, when $\alpha \in \{\frac{1}{k} | k \in \mathbb{N}, k > 1\}$, equation (3) becomes

$$\hat{X}_i = \xi_i + \frac{1}{k}\xi_{i-1} + \frac{1}{k^2}\xi_{i-2} + \ldots + \frac{1}{k^{i-1}}\xi_1, \quad (4)$$

which is the expansion of $\hat{X}_i$ in terms of the base k number system. In both cases, the coordinates of $\hat{X}_i$ uniquely determine the subsequence

$\lambda_1, \lambda_2, \ldots, \lambda_i$. Hence, it is possible to compare two different (sub)sequences by examining the difference between the last point on the W-curves for the (sub)sequences. Note that this property is not true of H-curves where different sequences can result in the same endpoint — the sequences AAATT and TATAA will both produce the endpoint $(-1, -1, 5)$, for instance.

As a practical matter, the finite resolution of numeric calculations and display devices limit this "historical" effect to those nucleotides that occur in a local neighborhood prior to a given nucleotide. Since $0 < \alpha < 1$ in a W-curve,

$$\lim_{i \to \infty} \alpha^i = 0.$$

Thus, $\exists\ m \in \mathbb{N}$ such that

$$\alpha^i \approx 0,\ \forall\ i > m,$$

and therefore, by Equation (2),

$$X_i \approx \beta(\xi_i + \alpha\xi_{i-1} + \alpha^2\xi_{i-2} + \ldots + \alpha^m\xi_{i-m}), \quad (5)$$

for some $m \in \mathbb{N}$.

Examining this equation, we see that X_i depends heavily upon only the preceding $m + 1$ symbols. In other words, X_i is not affected by the nucleotides that lie a significant distance prior to λ_i. As a result, sequences that differ because of local mutations will produce similar W-curves, an important property when comparing two similar, but not identical, genomic sequences. Consider the W-curves shown in Figure 2. The curve shown in Figure 2b was produced by inserting 60 adenine nucleotides (the region labeled 1) and 60 guanine nucleotides (the region labeled 2) into the curve shown in Figure 2a. Note how rapidly the modified W-curve returns to the subsequence patterns present in the original curve.

Finally, similar subsequences with vastly different histories can be visually compared by rotating the user's viewpoint with respect to the W-curve(s) that contain the subsequence. This property allows a user to identify occurrences of a particular subsequence within a long genomic sequence simply by rotating the subsequence (or the sequence) and inspecting the result. It can be shown [14] that rotations in the range $(-\pi/2, \pi/2]$ are sufficient to visually align any pair of identical subsequences.

3 Results

We have developed a visualization system for the Silicon Graphics IRIS that incorporates the three iterated function systems described in the preceding section. Using this system, a scientist can select a genomic sequence and investigate the properties and patterns in its H-curve, chaos game, and W-curve representations.

The W-curve, in particular, provides heretofore unavailable insights into the properties of genomic sequences. Figure 1 shows the chaos game representation of a pair of DNA sequences (HIVBIU and EMBLHU85, respectively). Despite the fact that these are very different sequences, their chaos game representations appear very similar. On the other hand, the differences between these sequences is starkly evident in the zy-projections of their W-curves (where k=200) shown in Figure 3.

Note also the similarities between the subsequences labeled 1 and 2 in Figure 3b. These subsequences point out how readily one can detect the repetition of a subsequence using a W-curve. This is true even when one has no a priori knowledge as to which subsequences may recur within a given genome.

We are actively engaged in the development of biological hypotheses based upon the insights that we have derived from these representations. Clearly, there are many more members of IFS family $\mathcal{F}$, as well as many other IFS families. Our hope is that these systems will provide scientists with additional insights into the patterns and properties of genomic sequences.

References

[1] N. G. Core, E. W. Edmiston, J. H. Saltz, and R. M. Smith, "Supercomputers and biological sequence comparison algorithms," *Computers and Biomedical Research*, no. 22, pp. 497–515, 1989.

[2] H. S. Bilofsky and C. Bruks, "The GenBank® genetic sequence data bank," *Nucleic Acids Research*, vol. 16, no. 5, pp. 1861–1863, 1988.

[3] G. N. Cameron, "The EMBL data library," *Nucleic Acids Research*, vol. 16, no. 5, pp. 1865–1867, 1988.

[4] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *The Journal of Biological Chemistry*, vol. 258, pp. 1318–1327, July 1983.

[5] E. C. Tyler, M. R. Horton, and P. R. Krause, "A review of algorithms for molecular sequence comparison," *Computers and Biomedical Research*, no. 24, pp. 72–69, 1991.

[6] S. Karlin, M. Morris, G. Ghandour, and M.-Y. Leung, "Efficient algorithm for molecular sequence analysis," *Proceedings National Academy of Sciences*, vol. 85, pp. 841–845, February 1988.

[7] M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, no. 147, pp. 159–197, 1981.

[8] M. S. Waterman, "Efficient sequence alignment algorithms," *Journal of Theoretical Biology*, no. 108, pp. 333–337, 1984.

[9] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, no. 162, pp. 705–708, 1982.

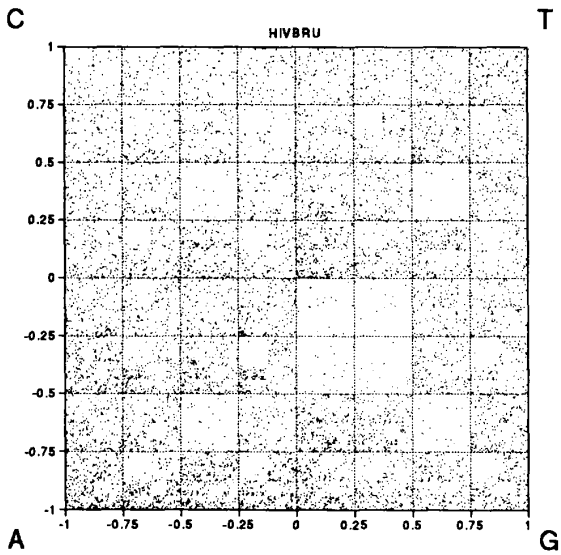
[10] H. J. Jeffrey, "Chaos games representation of genetic sequences," *Nucleic Acids Research*, vol. 18, no. 8, pp. 2163–2170, 1990.

[11] H. J. Jeffrey, "Chaos game visualization of sequences," *Computer & Graphics*, vol. 16, no. 1, pp. 25–33, 1992.

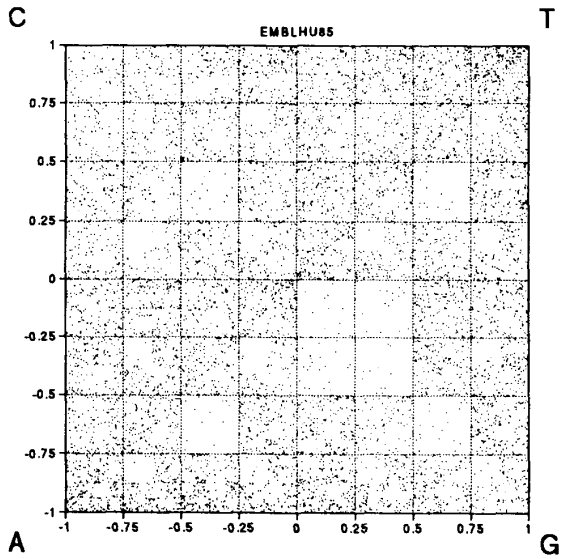
[12] M. F. Barnsley, *Fractals Everywhere*. Academic Press, San Diego, 1988.

[13] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*. Springer-Verlag, New York, 1992.

[14] D. Wu. Technical Report, Department of Computer Science, Illinois Institute of Technology, 1992.

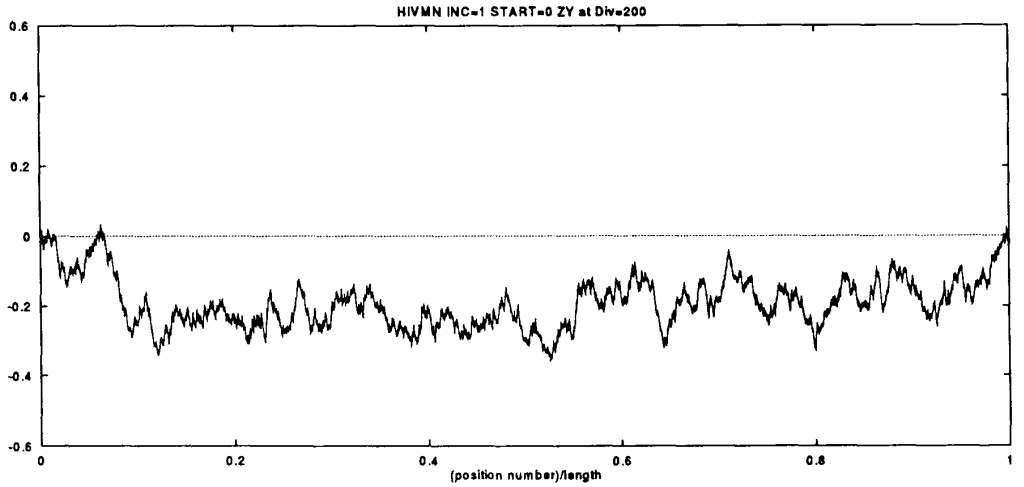


(a) **HIVBRU**

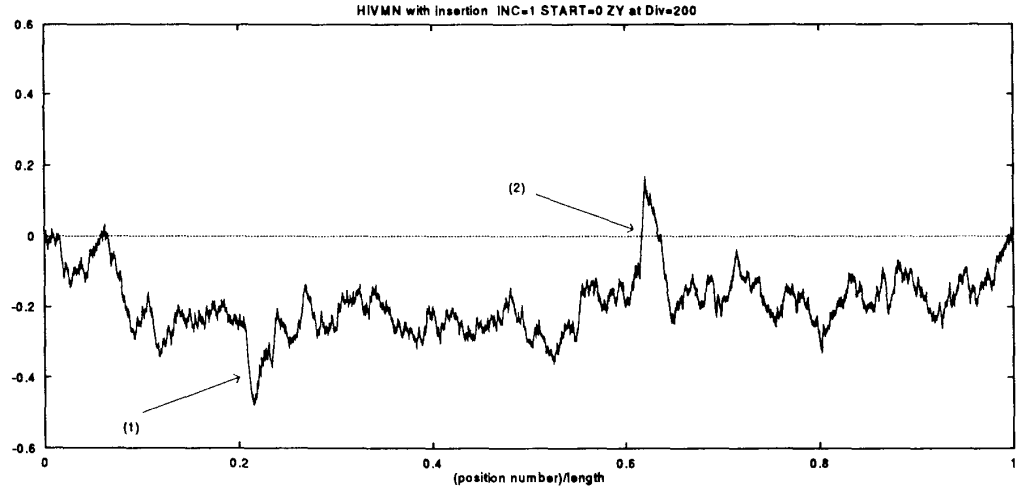


(b) **EMBLHU85**

Fig. 1. Chaos game representation of two DNA sequences.

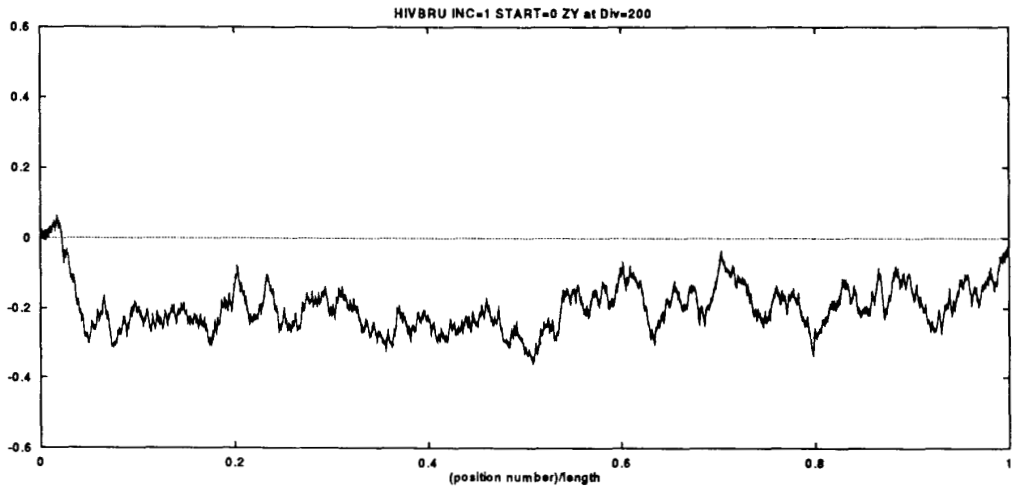


(a) HIVMN

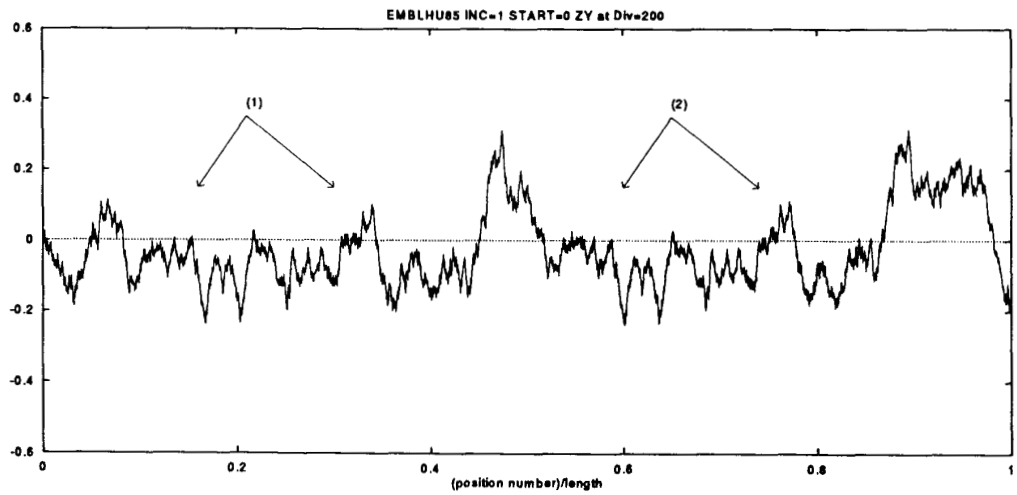


(b) HIVMN with insertions

Fig. 2. ZY-projections of a pair of W-curves.



(a) HIVBRU



(b) EMBLHU85

Fig. 3. ZY-projections of a pair of W-curves.

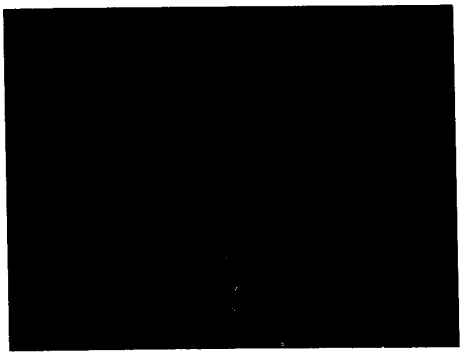


Fig. 4. H-curve.

Fig. 5. Close-up view of an H-curve showing individual nucleotides.

Fig. 6. W-curve.

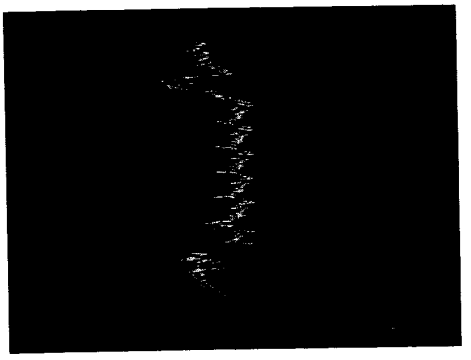


Fig. 7. Subset of a W-curve showing a repetitive subsequence.

(See color plates, p. CP-33.)

Computer Visualization of Long Genomic Sequences, D. Wu et al., pp. 308-315.

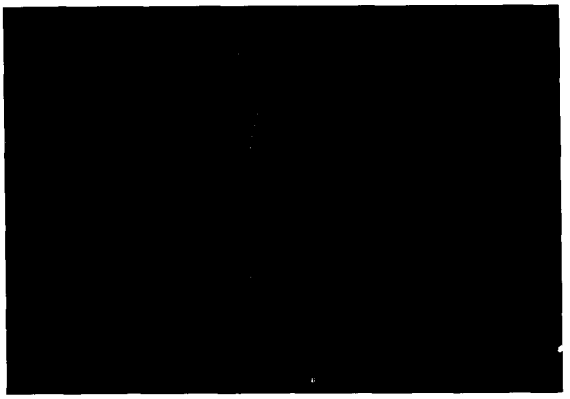


Figure 4: H-curve.

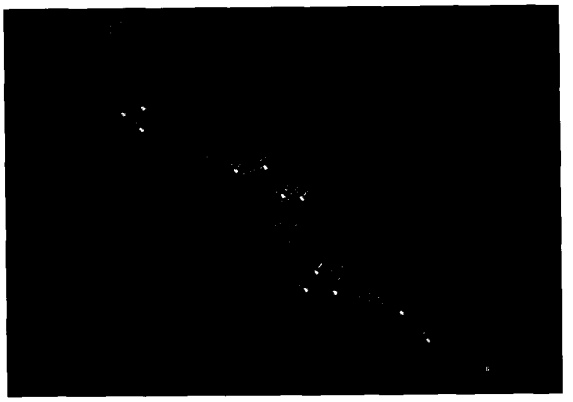


Figure 5: Close-up view of an H-curve showing individual nucleotides.

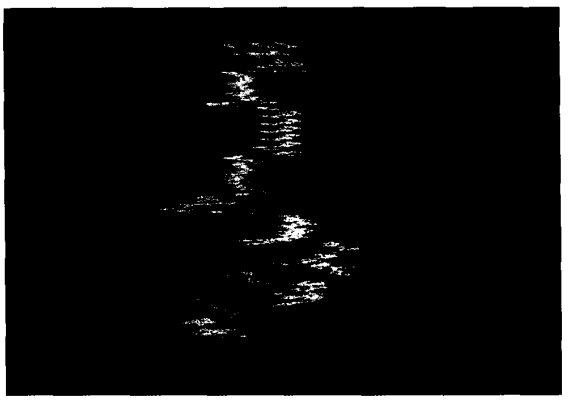


Figure 6: W-curve.

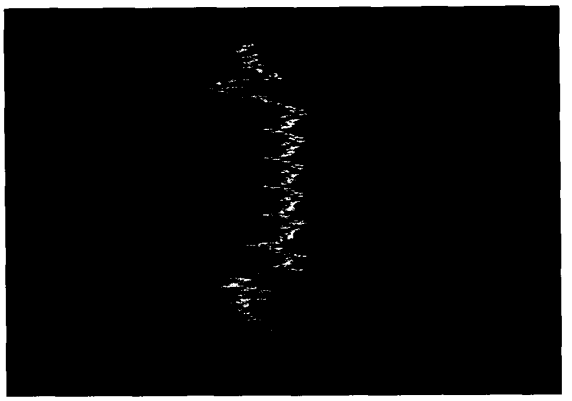


Figure 7: Subset of a W-curve showing a repetitive subsequence.