
Geometric global genomic informatics for RNA viruses

Douglas J. Cork* and Dachywan Wu

Department of Biological, Chemical and Physical Sciences,
Illinois Institute of Technology,
Chicago, IL 60616, USA
E-mail: douglascork@gmail.com
*Corresponding author

Michael J. North

Center for Complex Adaptive Agent Systems Simulation,
Argonne National Laboratory,
Decision and Information Sciences Division,
9700 S. Cass Avenue,
Argonne, IL 60439-4867 USA
and
Computation Institute,
University of Chicago,
Chicago, IL 60637 USA
E-mail: north@anl.gov

Abstract: The W-curve is in the class of autoregressive, geometrical algorithms that will permit bioinformaticists to visualise repeating/and or shifting patterns embedded within the whole genome. Herein, we inspect whole HIV-1 genomes. Conventional string based phylogenetic trees are insufficient to visualise the 3-D information content of an HIV-1 genome.

String based analysis cannot visualise, as can a typical W-curve based treatment, the subtlety of quasi-species shifting between different HIV-1 genomes. Herein we have conveyed some of the similar patterns embedded in different HIV-1 genomes. It is hoped that conventional string-based phylogenetic trees will enjoy some reinterpretation of the definition of quasi-species of the HIV-1 genome, in light of HIV-1 shift experiment performed with W-curves.

Keywords: genetic sequence visualisation; HIV-1; W-curve; autoregressive algorithms; geometric global genomics; genomic information content; RNA viruses.

Reference to this paper should be made as follows: Cork, D.J., Wu, D. and North, M.J. (2008) 'Geometric global genomic informatics for RNA viruses', *Int. J. Medical Engineering and Informatics*, Vol. 1, No. 2, pp.227–249.

Biographical notes: Douglas J. Cork is a Professor at the Department of Biological, Chemical and Physical Sciences at the Illinois Institute of Technology in Chicago, Illinois, USA. He completed his PhD from University of Arizona at Tuscon. He researches computational biology and bioremediation of chlorinated aromatic and sulphur containing toxic compounds.

D. Wu graduated from the Computer Sciences Dept., IIT with a PhD and has worked for a variety of CS companies in California.

Michael North graduated with an MS degree from the Dept. of Computer Sciences, IIT.

1 Introduction

Determining the homology of genetic sequences is presently a crucial question in the field of biology. The standards for estimating genetic homology, based on Blast, Clustal, Phylodraw and many other phylogeny tools that could be used from sources such as PHYLIP and PAUP, may be inadequate or ambiguous under many circumstances. Variations of the W-curve algorithm offer a tool to study overall genetic sequences in reasonable periods of time.

This review describes the standard W-curve, introduces the generalised W-curve, and proves some of their topological properties. This review then demonstrates that W-curve algorithms can be combined with the Smith-Waterman algorithm to address many of the key concerns raised with BLAST and the Smith-Waterman algorithm alone. As an integral part of this demonstration, HIV-1 is used as an example to show the usefulness of the W-curve approach.

2 Genetic homology in computational biology

2.1 *The challenge*

Determining the homology of genetic sequences is presently a crucial question in the field of computation microbiology (Ayala, 1976; Calvin, 1969; Eigen, 1993; Mindel, 1991; Miyamoto and Cracraft, 1991; Pendick, 1993; Sidow and Wilson, 1991; Waterman et al., 1991). The answers to large number of biologically important questions depend on the use of methods for finding genetic homology. These critical questions include the elucidation of relative lineages and the prediction of organism traits (Ayala, 1976; Calvin, 1969; Waterman et al., 1991). While a wide variety of techniques currently exist to study genetic homology, yet the most effect tend to only be applicable to short genetic sequences (Pearson and Miller, 1992; Waterman et al., 1991). Given that modern biotechnologists have begun to study sequences with lengths in the ten of millions of bases, more effective and traceable techniques will be required (Bishop and Waldholtz, 1990; Slightom et al., 1991; Waterman et al., 1991; Wills, 1991).

2.2 *Genetic homology*

Biological homology can be defined as the existence of similarities derived from a shared evolutionary history (Hale and Margham, 1991). Organisms that exhibit homology with one another are said to be homologous (Hale and Margham, 1991). Genetic homology between two or more genomic subsequences can be defined as subsequences similarity arising from common ancestry (Miyamoto and Cracraft, 1991). This relationship of common ancestry, also known as orthology, between homologous subsequences provides

the key to understanding the importance of determining genetic homology (Miyamoto and Cracraft, 1991).

2.3 *Relative lineages and traits*

Since genetic homology helps to define the evolutionary relationships between organisms, it can be used to clarify their relative lineages and to predict their traits. For most biologists, this information can be employed for taxonomic purposes and applied to a limited extent to trait predict. For many epidemiologists, this information can be at least as valuable as guide to the past and probable future dispersion patterns for the organisms in question. While several divergent schools of taxonomy exist, most taxonomic biologists apply knowledge of homology between genetic sequences to classify organisms into appropriate taxa (Campbell, 1990.) All three broad schools of taxonomy, phenetics, cladistics, and classical evolutionary taxonomy imply knowledge of genetic homology to some degree (Campbell, 1990).

Phenetics, which presently has few strict supporters, focuses on an organisms phenotype or observable characteristics. These observable characteristics are produced by interactions between the organisms environment and its genetic composition. For particularly simple organisms, such as viruses, phenotypes and genotypes are in many respects equivalent. Therefore, knowledge of genetic homology can be useful for directly observing the phenotypes of very simple organisms such as viruses. Cladistics focuses on the classification of organisms based on the branching of their traits from common ancestors. The level of difference between individual organisms is generally considered to be irrelevant and only the event of branching itself is regarded as important (Hale and Margham, 1991). This approach makes the most extensive use of genetic homology to discover the event of branching between various types of organisms (Campbell, 1990).

Classical evolutionary taxonomy focuses on producing reasonable compromises between phenetics and cladistics. In particular, attempts are made to accommodate both observable characteristics and genealogical information (Campbell, 1990). While the need to account for observable traits causes some knowledge of genetic homology to be ignored in favour of more directly observable traits, the use of genetic homology is widespread and expected to increase.

The successful classification of organisms into taxa is expected to yield new knowledge about both the current interrelationships between organisms and possible future interrelationships. Recent discoveries of unexpected genetic homology between species have produced some surprising and controversial reclassification of species (Ayala, 1976; Calvin, 1969; Campbell, 1990). Ongoing research relying upon genetic homology in viruses such as HIV-1 has yielded numerous unexpected results. Furthermore, since genetic homology is defined as resulting from orthology, we can see that the traits encoded in the given subsequences are likely to be related. These relationships can be sometimes be useful in predicting the traits expressed by a given organisms, especially for the relatively simple case of viruses.

3 Current state of the art

3.1 Blast

A huge variety of techniques exist which are based on the statistical use of subsequence matching algorithms (Waterman et al., 1991). In general, these techniques attempt to derive statistical measures of homology based on a series of Smith-Waterman scores or other scoring methods from selected portions of the genomes of interest (Altschul et al., 1990; National Center for Biological Information, 1995). Unfortunately, there is little agreement as to the types of statistical measures to be used and even in the best cases the measures chosen are often regularly revised (Altschul et al., 1990; National Center for Biological Information, 1995).

An example of genome statistics based system is the Basic Local Alignment Search Tool (BLAST) from the National Center for Biological Information (NCBI) (Altschul et al., 1990; National Center for Biological Information, 1995). BLAST is one of the most widely used of the genomic statistics implementations and is considered the state of the art in homology search tools. It attempts to find matches for input sequences in a given database of target sequences. The database can either be a large generic sequences list supplied directly by the NCBI or it can be a custom sequence list developed for a particular application. Despite being accepted as a de facto industry standard, it still has employed a variety of different statistical techniques in each of its most recent versions (Altschul et al., 1990; National Center for Biological Information, 1995). These statistical methods are generally designed to complement the diverse set of subsequence matching tools used.

In general, the BLAST system runs alignments on pairs of subsequences matches are then aggregated to according to one of several statistical functions and are then reported once these cross a specified threshold of significance. Unfortunately, this system requires the maximum length of the pairs to be specified along with a large variety of other parameters and at several different statistical metrics in its most recent versions alone (Altschul et al., 1990; National Center for Biological Information, 1995) (look for most recent Altschul). Even the NCBI itself considers the algorithm heuristic at best and can only offer guesses as to what matching tools and statistics should be used in a given case (National Center for Biological Information, 1995). Needless to say, determining homology for long sequences in an unambiguous way under such conditions can be difficult at best, especially when coupled to conventional tree-building tools, such as found in Phylip.

3.2 The Smith-Waterman algorithm

The Smith-Waterman algorithm is currently one of the primary techniques used in the high precision local alignment of genetic sequences (Los Alamos National Laboratory, 1994; Waterman et al., 1991). Furthermore, it can be incorporated into BLAST searches.

The Smith-Waterman algorithm itself essentially attempts to match all pairs of contiguous subsequences within two given genomic sequences (Smith and Waterman, 1981). This matching process views tile two input sequences as strings composed of the letters A for adenine, C for cytosine, G for guanine and T for thymine. From these strings matches for all contiguous subsequences are produced and stored in a matrix.

Following the description of the algorithm given in the original paper by Smith and Waterman, we define the two input genomic sequences as $A = a_1, a_2, \dots, a_n$ and $B = b_1, b_2, \dots, b_m$ (Smith and Waterman, 1981). Elements of these input sequences are matched according to a measure of similarity given by the function $s(a_i, b_j)$ and deletions of k elements from a string are given a weight W_k . The $n + 1$ by $m + 1$ matrix, H , giving the measure of the similarity between subsequences ending in a_i and b_j is defined by $H_{i,j} = \max\{H_{i-1,j-1} + s(a_i, b_j), \max_{k \geq 1}\{H_{i-k,j} - W_k\}, \max_{l \geq 1}\{H_{i,j-l} - W_l\}, 0\}$ for $1 \leq i \leq n$, $1 \leq j \leq m$

$$H_{k,0} = H_{0,1} = 0$$

and for $0 \leq k \leq n$, $0 \leq l \leq m$.

The matrix elements, $H_{i,j}$ are defined so that if the given endpoints a_i and b_j are related then they are assigned a similarity value given by $H_{i-1,j-1} + s(a_i, b_j)$. If there is a deletion of k bases ending at a_i then the value $H_{i-k,j} - W_k$ is assigned while a deletion of l bases ending at b_j has the value $H_{i,j-l} - W_l$ assigned. Finally, if there is no similarity then zero is assigned. As Smith and Waterman note, this check against zero is needed only if $s(a_i, b_j)$ can take negative values (Smith and Waterman, 1981).

The H matrices produced by the Smith-Waterman algorithm are characterised by Table 1 (Smith and Waterman, 1981). The columns and rows marked with a Δ are used as place holders to insure that $H_{i,j}$ is well defined when $i = 0$ or $j = 0$. In this example, two input sequences ATG and ATCGT are compared for similarity using the function $s(a_i, b_j) = 1$ when $a_i = b_j$ and -0.3 otherwise. This definition of $s(a_i, b_j)$ was originally chosen by Smith and Waterman under the assumption that each of the four bases is equally likely in a given position so that the average of this metric should be zero over a large number of random sequences (Smith and Waterman, 1981). As is noted in their original paper, the weight for deletions, W_k , must be greater than or equal to the difference between the value of $s(a_i, b_j)$ for a match and for a mismatch (Smith and Waterman, 1981). This must be true to prevent mismatches from being preferred over deletions. The function suggested by Smith and Waterman and adopted in this example is $W_k = 1.0 + 1/3k$.

Table 1 A Smith-Waterman alignment matrix

	Δ	A	T	G
Δ	0.0	0.0	0.0	0.0
A	0.0	1.0	0	0
T	0.0	0	2.0	0.7
C	0.0	0	0.7	1.7
G	0.0	0	0	1.7
T	0.0	0	1.0	0.3

Once the matrix for a given pair of sequences is produced, the endpoint of the subsequences with maximum homology is given by the indices for the maximal element in H , $\max(H_{i,j})$. These endpoints are extended to produce their associated subsequences by stepping backward through the $H_{i,k}$ values used in the computation of the maximal matrix element $\max(H_{i,j})$. This backtracking is continued until a contributing element with a value of 0.0 is found. The result is the pair of subsequences with maximal homology along with their alignment. To find the next best match, the backtracking

procedure is repeated beginning with the next largest element of H among the elements not part of previous subsequences (Smith and Waterman, 1981). In the example, 2.0 is the maximal element and the backtracking includes the element $H_{1,1} = 1.0$. As can be expected, the resulting common subsequence is AT.

Considering the construction of the $n + 1$ by $m + 1$ matrix H , it can be seen that the time complexity of the Smith-Waterman algorithm must be at least $\mathcal{O}(m + n)$ (Baase, 1988). Furthermore, since most sequences to be checked for homology are of similar length, n , we have the time complexity as $\mathcal{O}(n^2)$. A set of example sequence lengths is shown in Table 2 and taken from data found in the references by Bishop and Waldholtz (1990), Campbell (1990) and Wills (1991). Given the large genomic sequences that may need to be matched for homology, it becomes clear that simply running the Smith-Waterman algorithm may be impractical if not impossible in general. Therefore, there is a need to select relatively small subsequences from a given set of sequences for Smith-Waterman analysis. However, the selection of the subsequences itself poses difficult questions.

Table 2 Sequence length and example time complexity

<i>Sequence</i>	<i>Approximate length n</i>	<i>Example time complexity $\mathcal{O}(n^2)$</i>
Codon	3	9
Common DNA probe	200	40,000
HIV genome	9,000	81,000,000
Human gene	100,000	1×10^{10}
Human chromosome	65,000,000	4×10^{15}
Human genome	3×10^9	9×10^{18}

Clearly, the relatively high time complexity of the Smith-Waterman algorithm makes it highly unsuitable for the alignment of complete genomic sequences. As such, current researchers are compelled to restrict their analysis to limited subsequences of a given set of complete sequences. This choice of subsequences to align is absolutely critical to the results produced by the matching algorithm since the closeness of the matches is solely determined by the input data. Thus, effective homology studies require reliable and efficient subsequence selection tools. However, as we have seen from BLAST, effective subsequence selection is in itself a very difficult question (Altschul et al., 1990; National Center for Biological Information, 1995; Waterman et al., 1991).

Several methods are commonly used to select subsequences for Smith-Waterman analysis. These methods tend to involve special case arguments, statistical analysis, or some combination of these two.

The methods employing special case arguments typically depend on particular knowledge or belief about genomic properties. As an example, some recent HIV-1 studies have attempted to select subsequences based on known rates of mutation. The subsequences with the highest and lowest rates of mutation were chosen for comparison between viruses under the belief that their combination would give the best overall homology results. A variety of other authors have suggested that completely different techniques are likely to produce more meaningful results for general analysis (Altschul et al., 1990; National Center for Biological Information, 1995; Los Alamos National Laboratory, 1994). Furthermore, recent findings about the actual rates of HIV-1 mutation

have cast doubt on the original estimates of mutation rates (Ho et al., 1995; Wain-Hobson, 1995; Wei et al., 1995). Other methods of special case subsequence selection have produced similarly ambiguous results (Altschul et al., 1990; Sternberg, 1992; National Center for Biological Information, 1995) (use website reference).

As might be expected, subsequence selection based on statistical analysis has been arguably more consistent than that for special case selection. However, as in the case with BLAST, there currently is little if any agreement on the statistical methods to be employed in a given selection process.

We can see that the special case knowledge methods replace the difficult subsequence selection question with the equally difficult question of which knowledge to use. Similarly, the statistical methods replace the subsequence selection question with the difficult question of which statistical measures to use. In practice both of these methods lead to continuous ambiguities in the choice of subsequences. In addition to debates over the sometimes arbitrary choices of subsequences, this ambiguity can lead to a lack of experimental repeatability, especially in light of new evidence.

4 The W-curve algorithms

4.1 The standard W-curve algorithm

The standard W-curve, as developed by Wu et al. (1993), represents DNA sequences using the character input, $A = a_1, a_2, \dots, a_i, \dots, a_n$, like that for the Smith-Waterman algorithm (Wu et al., 1993). However, the W-curve uses a single input and produces graphical output.

Typically the input sequences are arranged so that the elements count from the 5' end of a given strand to the 3' end. Each element of this sequence is then assigned a vector in two dimensional Euclidean space according to the mapping,

$$A: \delta_A = (-1, -1)$$

$$C: \delta_C = (-1, 1)$$

$$G: \delta_G = (1, -1)$$

$$T: \delta_T = (1, 1)$$

to produce a sequence of vectors $\xi = \xi_1, \xi_2, \dots, \xi_i, \dots, \xi_n$ such that $\xi \in \{\delta_A, \delta_C, \delta_G, \delta_T\}$ for $1 \leq i \leq n$ (Wu et al., 1993). A family of iterated function systems, IFS, is then defined by the autoregressive relation:

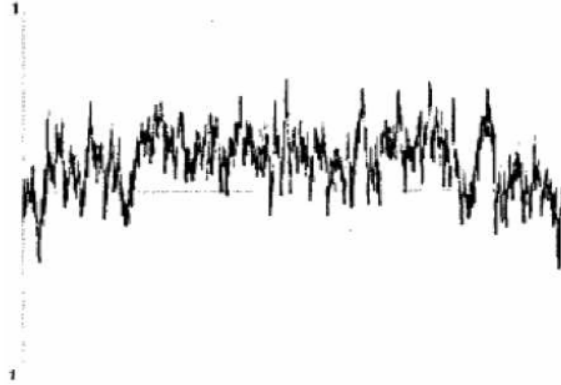
$$X_i = \alpha X_{i-1} + \beta \xi_i \quad (1)$$

for $1 \leq i \leq n$ and constants α and β (Wu et al., 1993). As elaborated by Wu and his colleagues, the choice $\alpha = \beta = 1$ produces the *H*-curve and the choice $\alpha = \beta = 1/2$ produces the chaos game representation, and the choice $\alpha = \beta = k^{-1}$ produces the standard W-curve (Wu et al., 1993).

In order to plot the standard W-curve, points in the \mathbb{R}^3 of the form $\{(X_i, i) | i \in \mathbb{N}, 1 \leq i \leq n\}$ are plotted along with connecting straight lines (Wu et al., 1993).

Graphical results of the standard W-curve algorithm for the complete genome of HIV-1 BRU are shown in Figure 1. From this figure the typical output of the standard W-curve can be seen.

Figure 1 A standard W-curve for HIV-1 BRU



Wu and his colleagues have proven a variety of useful W-curve properties (Wu, 1992; Wu et al., 1993). Included in these properties are the fact that the components of the X_i vectors remain within the interval $(-1, 1)$ for all i , that the effect of a change to single base at i has a rapidly diminishing effect on X , for $j > i$, that each endpoint of a standard W-curve uniquely determines its sequence, and that the alignment of any two sequences can be achieved with rotations $\phi \in (-\pi/2, \pi/2]$ (Wu, 1992; Wu et al., 1993). In addition, it is interesting to note that the entropy of the sample fragment, as defined by Shannon, can be roughly estimated from the rates of change observable in the graph (MacMillian, 1953; Shannon, 1948, 1951) (entropy as defined by applied bionformatics article Vol. 1). Also, Wu and others have noted that base pair biases can be immediately observed (Wu et al., 1993).

4.2 The generalised W-curve algorithm

The generalised W-curve is defined by the autoregressive relation:

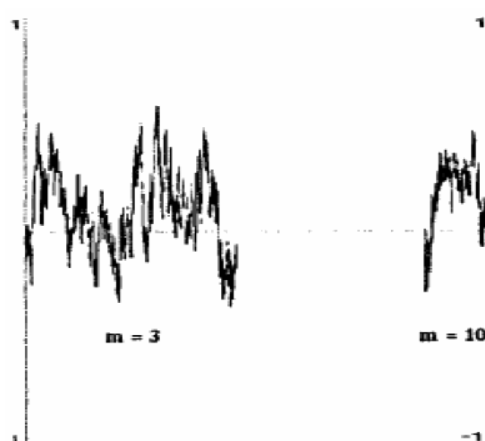
$$X_j = \alpha X_{j-1} + \sum_{i=0}^{m-1} \frac{\beta_i}{m} \xi_{mj+i} \quad (2)$$

where $\xi = \xi_1, \xi_2, \dots, \xi_i, \dots, \xi_n$ is the sequence of vectors ξ_i defined for the standard W-curves, m is a integer aggregation constant, $\beta_i \in \{k^{-1} | k \in \mathbb{N}\}$, $1 \leq j \leq [n/m]$ and $X_0 \equiv (0, 0)$. The resulting vectors, X_j are plotted as points $(X_{j,i})$ connected by straight lines. Clearly, the standard W-curves are the subset of the generalised W-curves with $m = 1$. The definition of the generalised W-curves is motivated by the need to examine very long genomic sequences. Specifically, the generalised W-curves can be used to more easily represent long sequences by assigning values to m that are greater than unity. These choices of m allow long sequences to be more easily viewed by causing the resulting

curves to be more compact and by causing them to have more variation in the xy-plane compared to a standard W-curve.

Graphical results from the generalised W-curve algorithm for the complete genome of HIV-1 BRU are shown in Figure 2. From this figure the typical output of the general W-curve for $m = 3$ and $m = 10$ can be compared to the standard W-curve output shown in Figure 1.

Figure 2 Two general W-curves for HIV-1 BRU



Many choices of the parameter m are possible, but it is expected that $m = 3$ may be commonly chosen since three base pairs occur in a codon or coding region (Hale and Margham, 1991). Other excellent choices of m may include values corresponding to the average lengths of specialised coding regions such as promoters, operators, structural genes and whole operons (Hale and Margham, 1991).

Several of the properties of the standard W-curve are shared by the generalised W-curves. These properties include the fact that the components of the X_i vectors remain within the interval $(-1, 1)$ for all i , that the effect of a change to single base at i has a rapidly diminishing effect on X_j for $j > i$, and that any two sequences can be aligned using rotations $\phi \in (-\pi/2, \pi/2]$.

The proof that the vector components of the generalised W-curve are constrained remain within the interval $(-1, 1)$ follows from their definition and the concepts presented by Wu and his colleagues (Wu et al., 1993). Specifically, for a given j , the maximum value possible for a vector component in the resultant of the sum of the ξ_i is simply $m/m = 1$. From the definition of the W-curve family we have $\alpha = k-1 \leq 1/2$, $\beta = k-1 \leq 1/2$ so that if the components of X_{j-1} are in $(-1, 1)$ then the components of X_j are also in Willis (1991) $(-1, 1)$. However, $X_0 \equiv (0, 0)$ so that the components of X_j must be in $(-1, 1)$ for all $i \leq j$.

The effect of a change to single base at i can be seen to have a rapidly diminishing effect on X_j for $j \gg i$ employing an argument identical to that given by Wu et al. (1993). The central notion is that $\lim_{j \rightarrow \infty} k^{-j} = 0$ so that the contribution from a given term dies off with increasing j (Wu et al., 1993).

The fact that any two generalised W-curve sequences can be aligned using rotations $\phi \in (-\pi/2, \pi/2]$ is identical to the result (Wu et al., 1993). The bounding of the vector components on the interval $(-1, 1)$ causes the resulting arguments to be identical.

4.3 Time complexity and the W-curve algorithms

As we have seen previously, the Smith-Waterman algorithm has an $\mathcal{O}(n^2)$ time complexity. Fortunately, it can be seen the general W-curve algorithms require a single pass over an input sequence of length n . Thus, the W-curve algorithms have a $\mathcal{O}(n)$ or linear time complexity (Baase, 1988).

In practice, the time needed to generate either standard or general W-curve plots is small enough to be completed in an interactive fashion on ordinary personal computers. Microsoft Windows, Microsoft DOS, and Silicon Graphics versions of the standard W-curve algorithm are available. Furthermore, Microsoft Windows and Microsoft DOS versions of the generalised W-curve algorithm have been created as part of the research for this paper. In addition, both James Ruscio and Jamys Kirk have completed an open source version for Linux in 2002, available on bioinformatics.org. Both versions of these programs have displayed significant increases in speed for $m > 1$. Due to the much smaller amounts of time required for the W-curve algorithms, they may be applied to interactively inspect genomic sequences that would be prohibitively large for the Smith-Waterman algorithm alone.

5 Topological properties of W-curves

5.1 Standard and generalised W-curves

We have seen that the standard W-curves are defined by the autoregressive relation in equation 4.1 for $1 \leq i \leq n$ and $\alpha = \beta = k-1$ (Wu et al., 1993). Each of the ξ_i is a vector that represents either Adenine (A), Cytosine (c), Guanine (G) or Thymine (T). This representation scheme is taken to count from the 5' end of the given strand to the 3' end. We have also seen that the generalised W-curve is defined by the autoregressive relation in equation 4.2 where ξ_i are the vectors defined for the standard W-curves. m is an integer constant, $\beta_i \in \{k^{-1} | k \in \mathbb{N}\}$, and $1 \leq j \leq \lceil n/m \rceil$. In order to simplify later work, we will denote a generalised W-curve $\gamma(z)$ as

$$\gamma(z) = (z-i)(X_{i+1} - X_i) + X_i \quad (3)$$

for some $i \in \mathbb{N}$ such that $i \leq z < i+1$. This notation emphasises the W-curves definition as the curve constructed by connecting the points (X_j, i) with straight lines.

5.2 Homotopic mapping of the generalised W-curves

In order to find a set of homotopic mapping for the generalised W-curves, we need to find continuous mapping between pairs of W-curves (Patterson, 1959). Therefore, we

need to construct continuous mapping of the form, $E_{\delta,\gamma}(C, z)$ for $C \in [0, 1]$, between any two generalised W-curves $\delta(z)$ and $\gamma(z)$ such that $E_{\delta,\gamma}(0, z) = \delta$ and $E_{\delta,\gamma}(1, z) = \gamma$ (Munkres, 1984).

In order to construct the continuous mapping $E_{\gamma,\delta}(C, z)$ as specified, we can first construct a continuous mapping $E_{\gamma,e}(C, z)$ from any W-curve to a unique W-curve e . For convenience, define $e(z)$ to be the W-curve with $\beta = 0$. Thus, e is simply the z -axis. Using this special W-curve, we can construct a continuous mapping, $E_{\gamma,e}(C, z)$ for $C \in [0, 1]$, between any generalised W-curve $\gamma(z)$ and $e(z)$ by defining,

$$E_{\gamma,e}(C, z) \equiv (1 - C)\gamma(z) + (C)e(z) = (1 - C)\gamma(z)$$

for $C \in [0, 1]$.

Clearly, $E_{\gamma,e}(0, z) = \gamma(z)$ and $E_{\gamma,e}(1, z) = e(z)$. Furthermore, we can see that $E_{\gamma,e}(C, z)$ produces a generalised W-curve for each value of C since all of the points between $z = i$ and $z = i + 1$ on the resulting curve have coordinates governed by $(1 - C)\gamma(z) = (1 - C)[(z - i)(X_{i+1} - X_i + X_i)]$ and thus lie on the straight line segment between $((1 - C)X_i, i)$ and $((1 - C)X_{i+1}, i + 1)$ (Anton and Rorres, 1987; Kelley, 1955; Mendelson, 1962). Next, define $E_{e,\gamma}(C, z) \equiv E_{\gamma,e}(1 - C, z)$ so that we can write,

$$E_{\gamma,\delta}(C, z) \equiv \begin{cases} E_{\gamma,e}(2C, z) & \text{when } 0 \leq C \leq \frac{1}{2} \\ E_{e,\gamma}(2C - 1, z) & \text{when } \frac{1}{2} \leq C \leq 1 \end{cases} \quad (4)$$

As expected we have $E_{\gamma,\delta}(0, z) = \gamma(z)$ and $E_{\gamma,\delta}(1, z) = \delta(z)$. Furthermore, since the components $E_{\gamma,e}(C, z)$ and $E_{e,\gamma}(C, z)$ have been shown to be continuous mappings with $\lim_{c \rightarrow +1/2} E_{\gamma,e}(C, z) = e(z) = \lim_{c \rightarrow -1/2} E_{e,\gamma}(C, z)$, we have $E_{\gamma,\delta}(C, z)$ as a continuous mapping as well (Gaughan, 1987). This set then constitutes a set of homotopic mappings for the generalised W-curves (Lang, 1958). We have thus shown that the mapping given by equations here for equations here constitutes a set of homotopic mappings from the generalised W-curves. Clearly, the mappings equations here have the same form for all z and therefore constitute a set of ordinary homotopic mappings for the generalised W-curves (Patterson, 1959). In addition to defining a set of ordinary homotopic mappings for the generalised W-curves, we can see that $E_{\gamma,\delta}(C, z)$ also constitutes a relative homotopic mapping (Patterson, 1959). This can be seen from the fact that $(X_0, 0)$ is a fixed point of the generalised W-curves. Thus, since $\gamma(0) = \delta(0)$ we have at $E_{\gamma,\Psi} \sim E_{\delta,\Psi}$, $\text{rel}(X_0, 0)$ for any W-curve $\Psi(z)$ (Patterson, 1959).

5.3 Homotopy types of the generalised W-curves

In order to find spaces, L , of the same homotopy type as the space of generalised W-curves, W , we need to find continuous mappings g and h such for $g: W \rightarrow L$ and

$h: L \rightarrow W$, we have $h \circ g: W \rightarrow W$ homotopic to the identity mapping $i: W \rightarrow W$ and $g \circ h: L \rightarrow L$ homotopic to the identity mapping $i': L \rightarrow L$ (Patterson, 1959).

Let L be the space of all line segments of length l parallel to the z -axis, We have previously demonstrated the existence of a continuous mapping, $E_{\gamma,\delta}(C, z)$ such that $E_{\gamma,e}(C, z)$ where e is a line segment of length l parallel to the z -axis. Therefore, define $g \equiv E_{\gamma,e}(C, z)$.

Next, define $h \equiv E_{\iota,e}(C, z)$ for $0 \leq z \leq 1$ and $E_{\iota,e}(C, z)$ as defined earlier. This defines h as a continuous map between each point $e(z)$ on the line segment e and any point $\iota(z)$ on the line segment ι for $C \in [0, 1]$.

Given the above definitions, we have $h \circ g: W \rightarrow W$ and $h \circ g: L \rightarrow L$. For a given W -curve $\gamma(z)$ we have $h \circ g(\gamma(z)) = e(z)$. This can be seen to be homotopic to the identity mapping $i: W \rightarrow W$ since we have already demonstrated the existence of the continuous mapping $E_{\gamma,e}(C, z)$ between $\gamma(z)$ and $e(z)$. Conversely, for $\iota(z)$ we have $g \circ h(\iota(z)) = e(z)$. This can be seen to be homotopic to the identity mapping $i': L \rightarrow L$ since we have already demonstrated the existence of the continuous mapping $E_{\gamma,e}(C, z)$ between $\iota(z)$ and $e(z)$. Thus, we can see that the W -curves are of the same homotopy type as the defined space of line segments (Patterson, 1959).

In addition to being of the same homotopy type as the defined set of line segments L , the W -curves can also be seen to be of the same homotopy type as the set of points in the \mathbb{R}^2 plane. This can be seen since continuous mappings exist between the points in \mathbb{R}^2 and the space of line segments L . Denoting these mappings as $r: \mathbb{R}^2 \rightarrow L$ and $s: L \rightarrow \mathbb{R}^2$, we can define $h' \equiv h \circ r: \mathbb{R}^2 \rightarrow W$ and $g' \equiv r \circ g: W \rightarrow \mathbb{R}^2$.

Given the above definitions, we have $h' \circ g': W \rightarrow W$ and $h' \circ g': \mathbb{R}^2 \rightarrow \mathbb{R}^2$. For a given W -curve (z) we have $h' \circ g'(z) = e(z)$. This can be seen to be homotopic to the identity mapping $i: W \rightarrow W$ since we have already demonstrated the existence of the continuous mapping ($E_{\gamma,e}(C, z)$) between $\gamma(z)$ and $e(z)$. Conversely, for a given point ι we have $g \circ h(\iota) = (0, 0)$. This can be seen to be homotopic to the identity mapping $i'': \mathbb{R}^2 \rightarrow \mathbb{R}^2$ since there clearly is a continuous mapping in \mathbb{R}^2 between ι and $(0, 0)$. Thus, we can see that the W -curves are of the same homotopy type as the set of points in the \mathbb{R}^2 plane (Patterson, 1959).

5.4 *The homotopy groups of the generalised W -curves*

The homotopy group for generalised W -curves can be seen to be the identity group, Z_1 , by considering the set of all closed paths with beginning and ending point given by an arbitrary point X . Let $[a]$ be one such path and $[b]$ be another and define $[a][b]$ (Patterson, 1959). However, all closed paths $[ab]$ are homotopic the null path at X so that this congruency becomes, $[b] = [a] = [a][b]$. Thus, both $[a]$ and $[b]$ must be identity elements,

showing that the fundamental group at X is Z_1 (Gallian, 1990). Furthermore, since X is an arbitrary point, we can consider Z_1 to be the fundamental group for the generalised W-curves without ambiguity. This can also be seen from the fact that the W-curves form an arc-wise connected topological space (Patterson, 1959). Since the fundamental group is the identity group, it is of course trivially commutative.

We have seen that the fundamental homotopy group for the generalised W-curves at any point X is Z_1 . This gives the one-dimensional homotopy group of the generalised W-curves as Z_1 . Defining $\pi_n(W, X)$ to be the n -dimensional homotopy group of the generalised W-curves at the point X , we have $\pi_1(W, X) = Z_1$ (Whitehead, 1966).

For $n > 2$ we can also see that $\pi_n(W, X) = Z_1$ by considering the two homotopy classes $[a]$ and $[b]$. By definition, $[ab] = [a][b]$. However, as in the case of $n = 1$, we can note that all closed surfaces $[ab]$ are homotopic the null path at X so that this congruency becomes, $[b] = [a] = [a][b]$ (Patterson, 1959). As before, both $[a]$ and $[b]$ must be identity elements so that the n -dimensional homotopy group at X is Z_1 .

6 The W-curve and Smith-Waterman algorithms

6.1 W-curve screening for Smith-Waterman analysis

The Smith-Waterman algorithm is clearly superior in finding the closer alignment match between any two given genomic subsequences, if the required computational resources are available. However, it can be seen from the Smith-Waterman algorithm's $\mathcal{O}(n^2)$ time complexity that the computational resources needed to compare whole genomic sequences, or even large subsequences, are prohibitive. However, the much quicker, linear time W-curve algorithms can be interactively applied to even the largest sequences. Considering these comparisons, it becomes clear that a system which combines the individual properties of these algorithms might be very effective.

By combining the W-curve algorithms with the Smith-Waterman algorithm, an analysis technique with some of the best properties of both can be formulated. This combination of algorithms should hopefully retain the linear computational time of the W-curves and permit the detailed subsequence matching of the Smith-Waterman algorithm.

A new technique with the required properties involves the use of W-curves to screen a set of genomic sequences for possible matches followed by the use of the Smith-Waterman algorithm on the identified subsequences. The result can thus produce subsequence alignments with the high quality of the Smith-Waterman algorithm while having the linear time and reduced ambiguity characteristic of the W-curve algorithms. Since the first phase of the W-curve Screening process uses the W-curve algorithm to perform the initial sequence screening, the overall time complexity must be at least that of the W-curves themselves, $\mathcal{O}(n)$. If during the second phase we chose at most m subsequences of maximum length l , then the overall time complexity for this phase is that of the Smith-Waterman algorithm acting on m sequences of length l , $\mathcal{O}(ml^2)$. However, if we chose $l \leq (n/m)^{1/2}$ then we have the order as, $\mathcal{O}(ml^2) = \mathcal{O}(m[(n/m)^{1/2}]^2) = \mathcal{O}(n)$. Thus the overall order will be $\mathcal{O}(n) + \mathcal{O}(n) = \mathcal{O}(n)$ (Baase, 1988). Table 3 lists the

approximate lengths of sequences that can be given Smith-Waterman analysis for selected values of m and n .

Table 3 Example W-curve screening parameters

m (bp)	n (bp)	l (bp)
10	10,000	31
10	1,000,000	316
10	100,000,000	3162
100	10,000	10
100	1,000,000	100
100	100,000,000	1,000

By using the W-curve algorithm subsequence selection ambiguity can be greatly reduced. The typical users of such a system, biologists, are comfortable with visual classifications of large sets of objects so they may tend to feel similar comfort with the W-curve system. Furthermore, the subsequence selection process can become very rapid compared to current techniques for large genomes.

6.2 Case study: HIV-1

The use of the W-curve screening technique can be demonstrated using the example case of the HIV-1 BRU, HIV-1 MAL and HIV-1 MN viral quasi-species. These quasi-species were investigated using generalised W-curve analysis for this paper. Additional analysis was performed on W-curve shift experiment data originally collected by Dachywan Wu and Douglas Cork of the Illinois Institute of Technology. The independent confirmation of the original results was achieved by successfully repeating the original results using the new generalised W-curve system. All of the W-curve experiments for this paper employed the HIV-1 BRU, HIV-1 MAL, and HIV-1 MN isolates. Various parts of the genomes were investigated, many of which code for specific viral traits (Los Alamos National Laboratory, 1994; National Institutes of Health, 1991). The viral genomic sequences and the related viral background information cited in this paper was taken from the Los Alamos National Laboratory Human Retroviruses and AIDS 1994 report prepared by the Theoretical Biology and Biophysics Group (Los Alamos National Laboratory, 1994). The data was accessed through an internet database system maintained by Los Alamos National Laboratory. Complete genomes were available from the Los Alamos National Laboratory database and were examined in all cases. All three of these viruses were originally extracted from human patients. The HIV-1 BRU isolate is a 9229 base pair ss-RNA virus originally known as LAV-1 (Los Alamos National Laboratory, 1994). The genomic sequence used for this paper originates at the RNA cap site and continues on to the final base, number 9229 (Los Alamos National Laboratory, 1994). According to the information available from the Los Alamos National Laboratory, this virus was originally known as LAV (Los Alamos National Laboratory, 1994). It was also known as LAV-1 to differentiate it from HIV-2 or LAV-2 (Los Alamos National Laboratory, 1994). Clones of this virus which are infectious have been created by Keith Peden of the Molecular Biology and Genetics Department of Johns Hopkins University School of Medicine (Los Alamos National Laboratory, 1994). Another related infectious clone, HIVNL43, [has beginning at its 3' end half a clone of HIV-1 BRU strain (Los

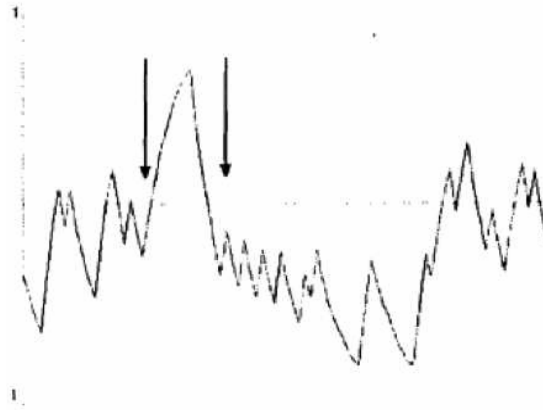
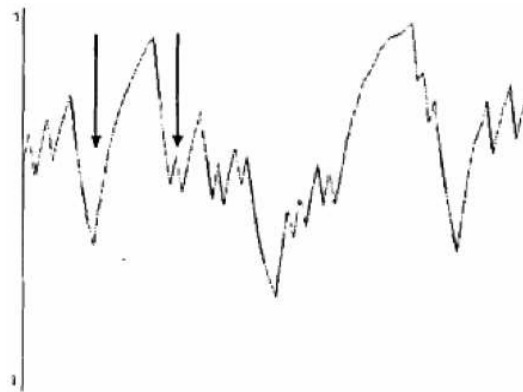
Alamos National Laboratory, 1994)]. For purposes of locating this information for future research, HIV-1 BRU has been assigned accession number K02013 by Los Alamos National Laboratory.

The HIV-1 MAL isolate is a 9229 base pair ss-RNA virus originally isolated from two AIDS patients in Africa (Los Alamos National Laboratory, 1994). As with HIV-1 BRU, the genomic sequence used for this paper originates at the RNA cap site and continues on to the final base (Los Alamos National Laboratory, 1994). According to the information available from Los Alamos National Laboratory, HIV-1 MAL has resisted efforts to create a clone which is infectious as of 1994 (Los Alamos National Laboratory, 1994). For purposes of locating this information for future research, Los Alamos National Laboratory has assigned HIV-1 MAL accession number K03456 (Los Alamos National Laboratory, 1994).

The HIV-1 MN isolate is a 9738 base pair ss-RNA virus originally isolated in 1984 from a pediatric patient with AIDS (Los Alamos National Laboratory, 1994). The genomic sequence used for this paper originates at the genomes' left end and continues on to the final base, number 9738 (Los Alamos National Laboratory, 1994). According to the information available from Los Alamos National Laboratory, HIV-1 MN is non-infectious (Los Alamos National Laboratory, 1994). For purposes of locating this information for future research, Los Alamos National Laboratory has assigned accession number MI7449 10 HIV-1 MN.

In the generalised W-curve experiments the HIV-1 BRU and HIV-1 MAL genomes were investigated using the Microsoft Windows version of the generalised W-curve system. Sequences of varying lengths, several divisors β_i , and different rotational angles were examined for a range of m values. Matches with varying qualities were found.

An example match with $m = 100$ and $\beta_i = 6^{-1}$ for all $i \in N$ is given in Figure 3 and Figure 4. Figure 3 shows the generalised W-curve for HIV-1 BRU from base 331 to base 9000 projected in the xz-plane. Figure 4 shows the generalised W-curve for HIV-1 MAL from base 321 to base 9000 projected parallel to the z-axis with a rotational angle of 185 degrees from the xz-plane. In the HIV shift experiments as originally performed by Dachywan Wu and D. Cork, and independently verified for this paper, selected portions of the HIV-1 BRU, HIV-1 MAL and HIV-1 MN genomes were visualised and compared using W-curve representations. These representations allowed differential comparisons to be made within and among the various known HIV strains. While primarily focusing on similarities in W-curve representations, these differential comparisons also allowed differences to be studied. Many parts of the given HIV-1 genome were investigated. The final selection of the portions to view for the HIV shift experiments involved a codon based arithmetic sequence. For the purposes of the W-curve shift experiments, codons were modeled as two non-coding bases coupled with a single coding base. Thus, the shift experiments combined bases into groups of three with one active member. From with these groups of three, a single base position was consistently chosen for W-curve display. The result was to select every third base for display once a starting location of base one, two, or three was selected. Thus, a modular congruency was established such that the equation $n = (b \bmod 3)$ for $n \in \{0, 1, 2\}$ is satisfied for each base position b shown. The resulting base position sequences are of the forms $\{1, 4, 7, 10, \dots\}$, $\{2, 5, 8, 11, \dots\}$ and $\{3, 6, 9, 12, \dots\}$.

Figure 3 HIV-1 BRU from bp 331 to bp 9000 for $m = 100$ **Figure 4** HIV-1 MAL from bp 321 to bp 9000 for $m = 100$ 

In real codons, groups of three bases are aggregated to code for specific amino acids (Hale and Margham, 1991). While triplets of four bases can yield up to $4^3 = 64$ unique combinations, only 20 amino acids are commonly produced (Hale and Margham, 1991). The result is that codons display degeneracy with several combinations coding for the same amino acid (Hale and Margham, 1991). During the actual creation of amino acids, messenger RNA carries one of 64 codes from the DNA to transfer RNA molecules (Campbell, 1990). Thus, the three base pairs from the original DNA codon are mapped into one out of the 20 different types of transfer RNA (Campbell, 1990). The transfer RNA can properly map the 64 unique codes into only 20 different types of molecules due to a special case pairing rule called wobble (Campbell, 1990). Wobble occurs because the 5' end of a transfer RNA anticodon can bond to several types of bases in its final, 3', position (Campbell, 1990). This variable bonding of the last position in the codon can be due to the presence of the base inosine or uracil (Campbell, 1990). The base inosine rarely occurs in normal DNA, but in transfer RNA it can bond to U, C, or A (Campbell, 1990). The base U in the 5' position can hydrogen bond with either G or A (Campbell, 1990). The variable nature of these bonds causes the final base pair in some transfer RNA molecules to act as wild cards that can accept more than one base pair (Campbell, 1990).

As an example, consider the transfer RNA molecule CC1 written from the 3' to the 5' end (Campbell, 1990). This molecule can bond with GGA, GGC, or GGU, all of which produce glycine (Campbell, 1990).

The selection of a congruency modulo three was chosen as a first approximation to real codon structures. As can be seen from the results of the shift experiment, this approximation has been highly useful for identifying repeated patterns (Figures 5–13.)

For a given value of n and a given modulo, the notion selecting of bases by modular congruency is equivalent to defining a generalised W-curve such that m is equal to the modulo and $\beta_i = 0$ when $n = (i \bmod m)$. These choices of m and β_i can be augmented with selected deletions and other related manipulations if desired. As can be seen from Figures 5 through 12, pairs of repeated patterns can be observed after performing such a shift experiment on these HIV-1 isolates.

Figures 5 and 6 have the approximate beginning and ending points of their repeated sequences marked with arrows. Figure 5 shows a projection of the HIV-1 BRU virus from a view perpendicular to the yz -plane. Base pairs beginning at zero and ending near 9229 are shown. As discussed previously, congruencies modulo three were used to select the base pairs to be shown. In the case of Figure 5, congruencies to one were selected. Comparing this to the region marked with arrows in Figure 6, a surprising similarity can be observed. Figure 6 is a projection of the HIV-1 MAL virus from a position perpendicular to the yz -plane, Base pairs from zero to 9229 are displayed. For this plot congruency to zero modulo three was employed.

Figure 5 HIV-1 BRU from bp 0 to bp 9229 with $(i \bmod 3) = 2$

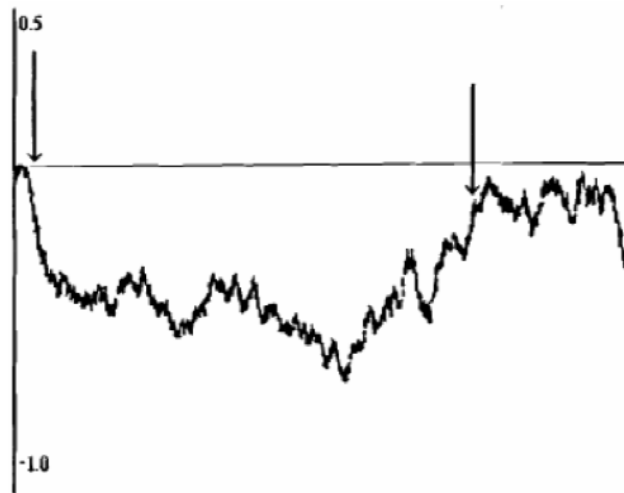
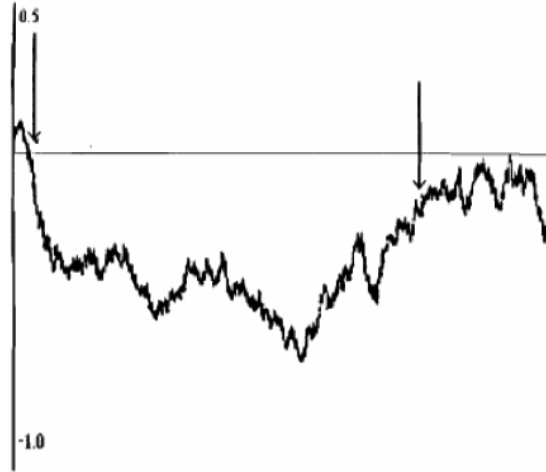


Figure 6 HIV-1 MAL from bp 0 to bp 9229 with $(i \bmod 3) = 0$



Figures 7 and 8 have the approximate initial and final points of their repeated sequences marked with arrows. Figure 7 shows a projection of the HIV-1 MN virus from a view perpendicular to the xz -plane. Base pairs beginning at zero and ending with the 3' end of the strand are displayed. Congruency to one modulo three is shown. Comparing this to the region marked with arrows in Figure 6, a decided similarity can be observed.

Figure 7 HIV-1 MN from bp 0 to bp 9738 with $(i \bmod 3) = 1$

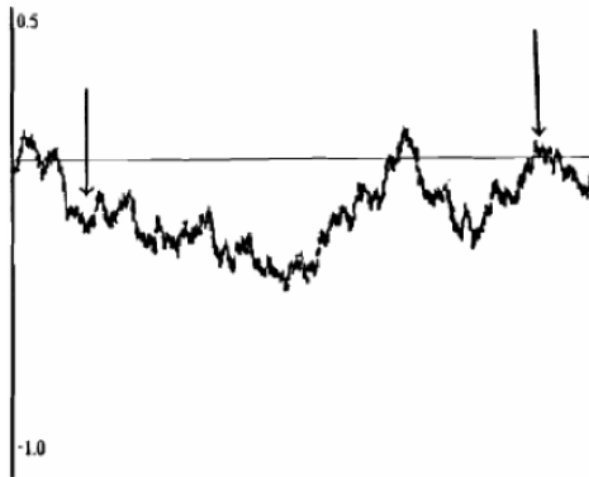
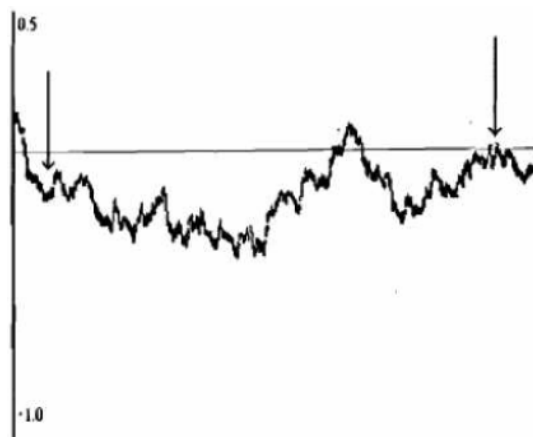


Figure 8 HIV-1 BRU from bp 0 to bp 9229 with $(i \bmod 3) = 0$ 

Figures 9 and 10 have the approximate opening and closing points of their repeated sequences again marked with arrows. Figure 9 shows a projection of the HIY-1 MAL virus from a view perpendicular to the xz -plane. Base pairs beginning at zero and ending with the 3' end of the strand are shown. Congruency two modulo three is displayed. Comparing this to the region marked with arrows in Figure 10, a marked similarity again can be observed. Figure 10 is a projection of the HIV-1 BRU virus from a position perpendicular to the yz -plane, Base pairs from zero to the 3' end of the strand are again shown. Herein congruency is zero modulo three. The significant similarities visible in Figures 5 through 10 suggest that many of the viruses studied should be classified as part of the same group. However, most of these strains are currently considered to be members of divergent clusters in string-based phylogenetic trees. Thus, the results of the W-curve shift experiments indicate that the current HIV-1 classifications may need to be revised. This revision may be especially important since HIV-1 classification schemes directly impact a variety of methods employed in the treatment of AIDS. A third experiment conducted by Wu and Cork involved the creation of a simulated mutation of HIV-I BRU by adding a base in a single position. Some of the results from these experiments are shown in Figures 11 and 12. Again, the approximate starting and stopping points of the repeated sequences are marked with arrows. Figure 11 shows a projection of the HIV-1 BRU virus with a single string of nucleotide A added in position 4615 from a view perpendicular to the xz -plane. Base pairs beginning at zero and ending with the 3 end of the strand are displayed.

Comparing this to the region marked with arrows in Figure 12, a noticeable similarity can be observed. Figure 12 is a HIV-1 BRU virus projection from a position perpendicular to the xz -plane. Base pairs from zero to the end of the strand are again shown.

Figure 9 HIV-1 MAL from bp 0 to bp 9229 with $(i \bmod 3) = 2$

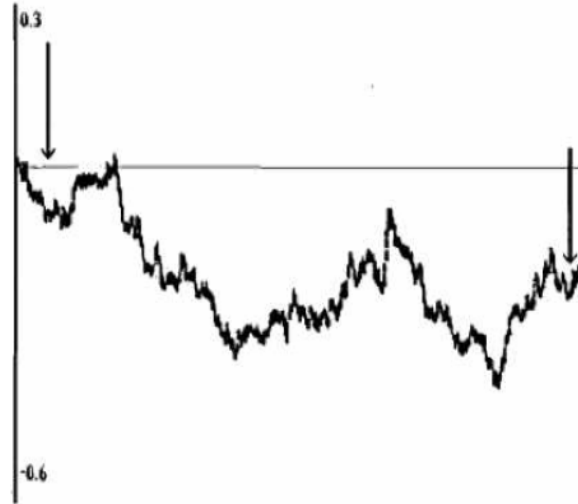


Figure 10 HIV-1 BRU from bp 0 to bp 9229 with $(i \bmod 3) = 0$

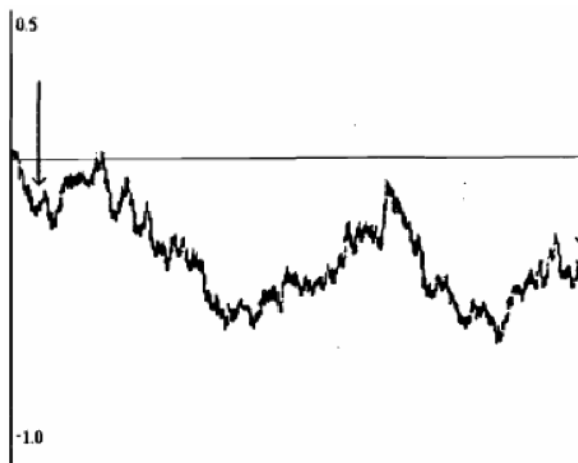


Figure 11 HIV-1 BRU+A from bp 4615 to bp 9229

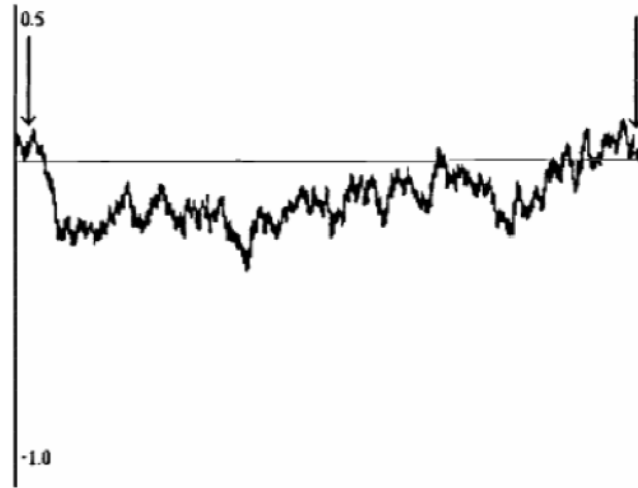
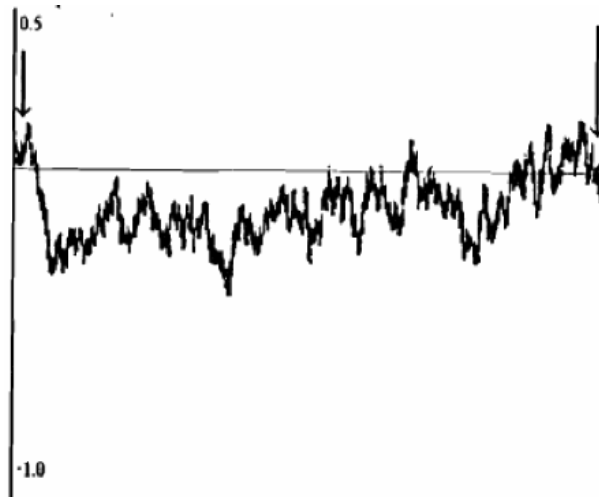


Figure 12 HIV-1 BRU from bp 0 to bp 9229



7 Summary

The W-curve algorithms offer the potential to greatly reduce the time requirements inherent in unambiguous selection of genomic subsequences. Furthermore, details of base pair groupings including codons, promoters, operators, structural genes, and whole operons can be easily considered. These advantages make W-curve screening for Smith-Waterman analysis especially useful for rapid investigation of large genomic sequences. Results from applying the W-curve indicate that current HIV-1 classifications critical to the treatment of AIDS may need to be revised. Further work includes the implementation of a convenient interface which integrates the W-curve and

Smith-Waterman algorithms into a single package as an improvement on the highly divergent interfaces currently available separately and the continued investigation of the vast array of information produced by W-curve analysis.

Acknowledgements

We acknowledge the cooperation of Dr. James Kenevan and his colleagues. Special thanks to Dachywan for planting and growing the seeds that made this work possible. The authors would also like to acknowledge the many resources generously provided by the faculty and fellow students of the Illinois Institute of Technology. Some of these resources, including a computer system integral to this research, were provided under Department of Defense Army Research Office Award DAAHO43GO212 as part of the High Performance Computing Initiative. Dr. Cork thanks Yacin Nadij, Dept. of Computer Sciences, IIT, for editing the paper.

References

- Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. and Lipman, D.J. (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, Vol. 215, pp.403–410.
- Anton, H. and Rorres, C. (1987) *Elementary Linear Algebra with Applications*, John Wiley & Sons, New York City, New York.
- Ayala, F.J. (1976) *Molecular Evolution*, Sinauer Associates, Sunderland, Massachusetts.
- Baase, S. (1988) *Computer Algorithms: Introduction to Design and Analysis*, 2nd ed., Addison-Wesley Publishing, Reading, Massachusetts.
- Bishop, J.E. and Waldholtz, M. (1990) *Genome*, Simon and Schuster, New York City, New York.
- Calvin, M. (1969) *Chemical Evolution: Molecular Evolution Towards the Origin of Living Systems on the Earth and Elsewhere*, Clarendon Press, Oxford, UK.
- Campbell, N.A. (1990) *Biology*, Benjamin/Cummings Publishing, Redwood City, California.
- Eigen, M. (1993) 'Viral quasispecies', *Scientific American*, Vol. 269, pp.42–49.
- Gallian, J.A. (1990) *Contemporary Abstract Algebra*, 2nd ed, D.C. Health and Company. Lexington, Massachusetts.
- Gaughan, E.D. (1987) *Introduction to Analysis*, Brooks/Cole Publishing, Pacific Grove, California.
- Hale, W.G. and Margham, J.P. (1991) *The Harper Collins Dictionary of Biology*, Harper Collins. New York City, New York.
- Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M. and Markowitz, M. (1995) 'Rapid turnover of plasma viremia and CD4 lymphocytes in HIV-1 infection', *Nature*, Vol. 373, pp.123–126.
- Kelley, J.L. (1955) *General Topology*, Van Nostrand Reinhold, New York City, New York.
- Lang, S. (1958) *Introduction to Algebraic Geometry*, John Wiley & Sons, New York.
- Los Alamos National Laboratory (1994) *Human Retroviruses and AIDS 1994*, Theoretical Biology and Biophysics Group, Los Alamos, New Mexico.
- McMillian, B. (1953) 'The basic theorems of information theory', in D. Slepian (Ed.): *Key Papers in the Development of Information Theory*, IEEE Press, New York City, New York, pp.57–80.
- Mendelson, B. (1962) *Introduction to Topology*, Allyn and Bacon, Boston, Massachusetts.

- Mindell, D.P. (1991) 'Aligning DNA sequences: homology and phylogenetic weighing', in Michael M. Miyamoto and Joel Cracraft, (Eds.): *Phylogenetic Analysis of DNA Sequences*, Oxford University Press, Oxford, UK, pp.73–1S9.
- Miyamoto, M.M. and J. Cracraft (1991) 'Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics', in Michael M Miyamoto and Joel Cracraft (Eds.): *Phylogenetic Analysis of DNA Sequences*, Oxford University Press, Oxford, UK, pp 3–17.
- Munkres, J.R. (1984) *Elements of Algebraic Topology*, Addison-Wesley Publishing, Redwood City, California.
- National Center for Biological Information (1995) *BLAST Manual*, National Center for Biological Information, Washington, D.C.
- National Institutes of Health (1991) 'Evolutions: the HIV-1 genome', *The Journal of NIH Research*, Vol. 3, p.119.
- Patterson, E.M. (1959) *Topology*, 2nd ed., Oliver and Boyd, Edinburgh, UK.
- Pearson, W.R. and Miller, W. (1992) 'Dynamic programming algorithms for biological sequence comparison', in Ludwig Brand and Michael L. Johnson (Eds.): *Methods in Enzymology*, Academic Press, San Diego, California, pp.575–600.
- Pendick, D. (1993) *New Method May Speed Gene Searches*, Science News, Vol. 143, p.294.
- Sidow, A. and Wilson, A.C. (1991) 'Compositional statistics evaluated by computer simulations', in Michael M. Miyamoto and Joel Cracraft (Eds.): *Phylogenetic Analysis of DNA Sequences*, Oxford University Press, Oxford, UK, pp.129–146.
- Shannon, C.E. (1948) 'A mathematical theory of communication', in David Slepian (Ed.): *Key Papers in the Development of Information Theory*, IEEE Press, New York City, New York, pp.5–29.
- Shannon, C.E. (1951) 'Prediction and entropy of printed english', in David Slepian (Ed.): *Key Papers in the Development of Information Theory*, IEEE Press, New York City, New York, pp.42–46.
- Slightom, J.L., Siemieniak, D.R. and Sieu, C. (1991) 'DNA sequencing: strategy and methods to directly sequence large DNA molecules', in Michael M. Miyamoto and Joel Cracraft, *Phylogenetic Analysis of DNA Sequences*, Oxford University Press, Oxford, UK, pp.18–44.
- Smith, T.F. and Waterman, M.S. (1981) 'Identification of common molecular subsequences', *Journal of Molecular Biology*, Vol. 147, pp.195–197.
- Sternberg, S. (1992) 'HIV comes in five family groups', *Science*, Vol. 256, pp.966.
- Wain-Hobson, S. (1995) 'Virological mayhem', *Nature*, Vol. 373, pp.102.
- Waterman, M.S., Joyce, J. and Eggert, M. (1991) 'Computer alignment of sequences', in Michael M. Miyamoto, and Joel Cracraft (Eds.): *Phylogenetic Analysis of DNA Sequences*, Oxford University Press, Oxford, UK, pp.59–72.
- Wei, X., Ghosh, S.K., Taylor, M.E., Johnson, V.A., Emini, E.A., Deutsch, P., Lifson, J.D., Bonhoeffer, S., Nowak, M.A., Hahn, B.H., Saag, M.S. and Shaw, G.M. (1995) 'Viral dynamics in human immunodeficiency virus type I infection', *Nature*, Vol. 373, pp.117–122.
- Wills, C. (1991) *Exons, Introns and Talking Genes*, HarperCollins Publishers, New York City, New York.
- Whitehead, G.W. (1966) *Homotopy Theory*, MIT Press, Cambridge, Massachusetts.
- Wu, D. (1992) *Technical Report*, Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois
- Wu, D., Roberge, J., Cork, D.J., Nguyen, B.G. and Grace, T. (1993) 'Computer visualization of long genomic sequences', *Proceedings of the Conference on Visualization*, IEEE Press, New York City, New York, CP 33, pp.308–314.