Published in final edited form as:
J Comput Biol. 1996 ; 3(3): 345–360.

Over- and Underrepresentation of Short DNA Words in Herpesvirus Genomes

MING-YING LEUNG[1], **GENEVIEVE M. MARSH**[2], and **TERENCE P. SPEED**[3]

[1]Division of Mathematics and Statistics, University of Texas at San Antonio, San Antonio, Texas 78249

[2]Mexican American Treatment Effectiveness Research Center, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78284

[3]Department of Statistics, University of California, Berkeley, California 94720

Abstract

The relative abundance and rarity of DNA words have been recognized in previous biological studies to have implications for the regulation, repair, and evolutionary mechanisms of a genome. In this paper, we review several different measures of abundance and rarity of DNA words, including *z*-scores, representation ratios, and cross-ratios, that have appeared in the recent literature, and examine the concordance among them using the human cytomegalovirus genome sequence. We then rank all words of length $k = 2, \ldots, 5$ of seven herpesvirus genomes according to their abundance, as measured by one of the *z*-scores based upon a stationary Markov model of order $k - 2$. Using a simple metric on the ranks of 2-words of the seven herpesvirus sequences, we construct an evolutionary tree. Several 3-words are observed to be consistently over- or underrepresented in all seven herpesviruses. Furthermore, clusters of some of the most over- and underrepresented 4- and 5-words in the genomes are identified with functional sites such as the origins of replication and regulatory signals of individual viruses.

Keywords

DNA sequence; word count; Markov chain; *z*-score; herpesviruses

INTRODUCTION

A short DNA segment comprising *k* nucleotide bases will be called a *k*-word (e.g., *AA* and *TC* are 2-words, *ATA* and *GCT* are 3-words). The present study focuses on the relative abundance of short ($k \leq 5$) DNA words, which has been implicated in the molecular structure and stability of DNA, as well as in recombination, replication, regulation, and repair activities (see, e.g., McClelland, 1985; Bhagwat and McClelland, 1992; Burge *et al*., 1992; Merkl *et al*., 1992; Karlin and Ladunga, 1994). Although the abundance or rarity of a word can be directly measured by its relative frequency, unequal base composition and

Address reprint requests to: Ming-Ying Leung, Division of Mathematics and Statistics, University of Texas at San Antonio, San Antonio, TX 78249, leung@minuet.utsa.edu.

interdependence of neighboring bases in DNA (Nussinov, 1981; Blaisdell, 1985) preclude a meaningful comparison of relative abundance of different words by frequency alone. Instead, DNA sequence analysts often prefer to assess a *k*-word by comparing its observed frequency to that expected under a probability model that reflects the composition of the data sequence in its representation of shorter (length ≤ *k* − 1) words.

Following the terminology of Phillips *et al*. (1987a,b), we call a word over- (or under) represented if it is observed more (or less) frequently than expected, under some specified probability model. While it is generally accepted that the over-and underrepresentation of short words are statistical aspects of DNA sequences with potential biological relevance, the methods of quantifying over- and underrepresentation are quite varied. In this paper, we first describe one representation measure, the *z*-score, based on a Markov chain model. Several other representation measures are then reviewed and compared with one another. Finally, we use the *z*-score to explore evolutionary relationships, family commonalities, and individual genome features in seven herpesviruses.

STATISTICAL MEASURES OF OVER- AND UNDERREPRESENTATION

For any *k*-word *W* in a DNA sequence, a standardized frequency can be defined by

$$z(W)=\frac{N_W-E(N_W)}{\sqrt{V(N_W)}} \quad (1)$$

where N_W is the observed count of word *W*, while $E(N_W)$ and $V(N_W)$ are, respectively, the mean and variance of N_W, these being calculated under a probability model for random sequences which resembles the data sequence in the occurrence frequencies of all words with lengths 1, …, *k*−1. For example, if *W* is a 2-word, we can use an independent and identically distributed (i.i.d.) sequence model with probabilities equal to the relative frequencies of the bases (1-words) in the data sequence. For 3-words, we can use a Markov chain model with transition probabilities estimated from the base counts and dinucleotide (2-word) counts.

If $E(N_W)$ and $V(N_W)$ are known, then under rather general conditions the statistic $z(W)$ is asymptotically normally distributed with zero mean and unit variance as the sequence length $n \to \infty$, assuming the validity of the model in question (see Billingsley, 1995 for the Markov chain case). In practice $E(N_W)$ and $V(N_W)$ are seldom known, but will be estimated from the DNA sequence under study. When we replace these quantities by consistent estimates $\hat{E}(N_W)$ and $\hat{V}(N_W)$, the quantity $[N_W-\hat{E}(N_W)]/\sqrt{\hat{V}(N_W)}$ will in general no longer be asymptotically distributed as a standard normal. This is because the differences $N_W - E(N_W)$ and $\hat{E}(N_W) - E(N_W)$ will typically have the same asymptotic rate, and so their difference is not properly standardized by $\sqrt{\hat{V}(N_W)}$ (see Prum *et al*., 1995, Waterman, 1995, chapter 12).

z-scores in Markov models I

A stationary Markov chain model for a DNA sequence is specified by a matrix of transition probabilities $p(a, b)$, $a, b \in \mathfrak{A} = \{A, G, C, T\}$ with a stationary probability distribution $[\pi(a), a \in \mathfrak{A}]$ satisfying $\sum_{a \in \mathfrak{A}} \pi(a)p(a, b) = \pi(b)$, $b \in \mathfrak{A}$. It has been long known that under such a model, a word $W = w_1w_2\ldots w_{k-1}w_k$ appears in a sequence of length n with a frequency N_W whose expectation is

$$E(N_W)=(n-k+1)\pi(w_1)p(w_1,w_2)\ldots p(w_{k-1},w_k)$$

A natural consistent estimate of this quantity is

$$\hat{E}(N_W)=\frac{\prod_{i=1}^{k-1}N_{w_iw_{i+1}}}{\prod_{i=2}^{k-1}N_{w_i}} \quad (2)$$

which has been used in many studies of word abundance in DNA sequences.

It is only relatively recently that an exact expression for the variance $V(N_W)$ has been given by Kleffe and Borodovsky (1992). As the formula is quite lengthy, we refer the readers to their original paper. $V(N_W)$ too has a natural "plug-in" estimator $\hat{V}(N_W)$ based upon the maximum likelihood estimates $p(\hat{a}, b) = N_{ab}/N_a$ and the consistent estimates $\hat{\pi}(a) = N_a/n$. Applying these one can calculate what we term the *maximum likelihood* plug-in *z*-score

$$z^L(W)=\frac{N_W-\hat{E}(N_W)}{\sqrt{\hat{V}(N_W)}} \quad (3)$$

For long sequences like the herpes genomes where $n > 10^5$, the distribution of $z^L(W)$ should be normally distributed with mean zero. Indeed, the $q - q$ plot (Venables and Ripley, 1994) of the z^L-scores for any specific 3-word in 100 simulated Markov sequences against the standard normal distribution does have a straight line appearance (e.g., Fig. 1a shows the $q - q$ plot for the word *TTG*). Furthermore, when the z^L-scores of *all* 3-words in a simulated sequence are considered together, they also appear to be normally distributed, as demonstrated by Fig. 1b. These suggest that $z^L(W)$ may still be useful in identifying over-or underrepresented words even though we know that it does not have an asymptotic standard normal distribution as $n \to \infty$.

In general, a representation measure for k-words requires a Markov model of order $k - 2$ to represent words with a length $\leq k - 1$ according to their observed frequencies. For $k > 3$, this higher order Markov chain can be converted to a first-order chain by expanding the state space. For instance, the 4-word *CTAG* in a second-order Markov sequence is the same as the 3-word (*CT*)(*TA*)(*AG*) in a first-order sequence with the 16 member alphabet {*AA*, *AC*, …, *TT*}. The 16 × 16 transition probability matrix will have elements $p(ab, cd)$, which are nonzero only when $b = c$. The z^L-scores can be obtained in the same way by applying Eqs. (1)–(3) to the expanded chain.

It is already well known that DNA sequence data do not agree well with any homogeneous Markov chain model (see, e.g., Karlin and Brendel, 1993; Pevzner, 1992). Thus it is not entirely surprising to see that the z^L-scores in the human cytomegalovirus (HCMV), one of the herpesviruses in our data set, are distributed quite differently from those of the simulated Markov sequences. Figure 2 shows that 14 out of 16 2-words and 40 out of 64 3-words have z^L-scores outside the range of the 100 simulated values. The $q - q$ plots in Figure 3a–c also indicate a nonnormal distribution for the totality of z^L-scores. The z^L-score distribution for 5-words appears to be much closer to normal (Fig. 3d), suggesting that as k increases, the $k - 2$nd-order Markov chain model may fit better with the data sequence.

Despite the nonconformity of DNA to homogeneous Markov models, extremely over- or underrepresented words found by statistical analyses on *E. coli* DNA data using such models have provided valuable qualitative insights in biological investigations (e.g., see Phillips *et al*., 1987a, b; Merkl *et al*., 1992). It would, therefore, seem appropriate to use $z^L(W)$ as an index of over- and underrepresentation to provide a ranking of the abundance of W among words of equal length in a DNA sequence.

z-scores in Markov models II

A Markov model for DNA restricted to those sequences preserving the original dinucleotide counts was first considered by Cowan (1991), making use of a formula proved by Whittle (1955). For a (first-order) Markov sequence $(X_1, \ldots, X_n)$ we denote by **N** the 4×4 matrix

$$\mathbf{N} = \begin{pmatrix} N_{AA} & N_{AC} & N_{AG} & N_{AT} \\ N_{CA} & N_{CC} & N_{CG} & N_{CT} \\ N_{GA} & N_{GC} & N_{GG} & N_{GT} \\ N_{TA} & N_{TC} & N_{TG} & N_{TT} \end{pmatrix}$$

of dinucleotide counts in the sequence. Next we let **S** denote the class of all sequences of length n with the same initial base X_1 and the same matrix **N** of 2-word counts.

Whittle (1955) proved that under the first-order Markov model, the conditional distribution of $(X_1, \ldots, X_n)$ given X_1 and **N** is uniform on the class **S**, and he derived a formula for the number of sequences in **S**. Cowan (1991) obtained the conditional expectation of N_W given **S**, showing that $E(N_W|X_1, \mathbf{N})$ can be calculated relatively straightforwardly in terms of Whittle's formula. Prum *et al*. (1995) gave an expression for the conditional $V(N_W|X_1, N)$ and ascertained the asymptotic standard normal distribution of the quantity

$$z^C(W) = \frac{N_W - E(N_W|X_1, \mathbf{N})}{\sqrt{V(N_W|X_1, \mathbf{N})}}$$

as $n \to \infty$ under the first-order Markov model. Schbath *et al*. (1995) further showed that by expanding the Markov chain states as previously discussed, it is possible to calculate $z^C(W)$

for any *k*-word *W*, conditional upon the matrix **N** of counts of (*k* − 1)-words and the first (*k* − 2)-word of the sequence.

When *n* is large, the standard normal approximation for the distribution of $z^C(W)$ would, in principle, allow us to set criteria for discerning whether a word is significantly over- or underrepresented. For instance, at significance level α, a conservative critical value can be set at

$$c_k = \mathbf{\Phi}^{-1}[1-\alpha/(2\cdot 4^k)]$$

where Φ is the standard normal distribution function. This critical value comes from the simple union bound

$$Pr\{\max|z^C(W)|>c_k\} \le 4^k\cdot 2\cdot[1-\mathbf{\Phi}(c_k)]$$

the maximum being taken over all *k*-words *W*.

With α = 0.05, c_k for *k* = 2, 3, 4, and 5 are respectively 2.96, 3.36, 3.73, and 4.06. When the z^C-scores for HCMV are compared with these critical values, we find that the majority of the 2-, 3-, and 4-words would be deemed significantly over- or underrepresented! So, statistical significance of the z^C-scores in DNA data may not indicate anything unusual. The unexpectedly large number of statistically significant z^C-scores for the 2- to 4-words is, most likely, due to the lack of fit of the homogeneous Markov model with the data mentioned before. We have already seen that the (3rd-order) Markov model fits better as *k* increases to 5. Indeed, the percentage of 5-words with absolute z^C-scores exceeding c_5 reduces to 4.5%.

It is worth noting that the asymptotic analysis of Cowan's formula by Prum *et al*. (1995) shows that as $n \to \infty$,

$$E(N_W|X_1,\mathbf{N}) = \frac{\prod_{i=1}^{k-1}N_{w_iw_{i+1}}}{\prod_{i=2}^{k-1}N_{w_i}}[1+O(N^{-1})] = \hat{E}(N_W)[1+O(n^{-1})]$$

where $\hat{E}(N_W)$ is given by Eq. (2) above. Consequently, the appropriate normalizing factor for $N_W - \hat{E}(N_W)$ is $\sqrt{V(N_W|X_1,\mathbf{N})}$. One can replace $V(\hat{N}_W)$ in expression (3) by $V(N_W|X_1, \mathbf{N})$ and the resulting z^L will have an asymptotic standard normal distribution. We choose not to do so for the present application mainly because of the excessive amount of time required to compute the conditional variances (it takes more than 16 h on a Silicon Graphics Indy station to compute $V(N_W|X_1, \mathbf{N})$ for the 3-, 4-, and 5-words of the HCMV sequence); and in all the test cases we have tried, the resulting word rankings are almost identical to the original.

z-scores in Markov models III

Prum *et al*. (1995) introduce a slightly different kind of *z*-score, which in essence tests whether the last base in the word $W = w_1 w_2 \ldots w_k$ appears with a frequency that is consistent with a Markov chain of the given order. The statistic used is a normalized version of the difference

$$N_W - \frac{N_{W[-1]} N_{w_{k-1} w_k}}{N_{w_{k-1}}}$$

where $W^{[-1]} = w_1 \ldots w_{k-1}$ is the word W with its last letter deleted. The asymptotics they develop for this difference make use of a martingale central limit theorem. We therefore call the corresponding abundance measure the *martingale* *z*-score, and denote it by z^M. The full expression for z^M is

$$z^M(W) = \frac{1}{\sqrt{n}\hat{\sigma}_W} \left(N_W - N_{W^{[-1]}} \frac{N_{w_{k-1} w_k}}{N_{w_{k-1}}} \right)$$

where $\hat{\sigma}_W^2$ is any consistent estimate of the limiting variance given in Prum *et al*. (1995). In that article, we can also find a proof that $z^M(W)$ is asymptotically standard normal as $n \to \infty$, assuming the correctness of the Markov chain model.

Other representation measures in Markov models

The representation of short words in DNA sequences has also been examined using two rather different measures, both based upon stationary Markov chain models for the entire sequence.

Merkl *et al*. (1992) use the simple ratio of the observed to the expected frequency of occurrence of W

$$r(W) = \frac{N_W}{E(N_W)}$$

which we will call the *representation ratio*. In practice $E(N_W)$ is replaced by a consistent estimator $\hat{E}(N_W)$ of $E(N_W)$, typically expression (2) above or another expression asymptotically equivalent to it. We note that Cuticchia *et al*. (1992) have also used ratios, and no doubt others have done so as well. As pointed out by Prum *et al*. (1995), for longer and hence rarer words, the variance will often be quite close to the mean, and in such cases the ratio $r(W)$ and the *z*-score z^L will differ by unity. (Self-overlap is also an issue here, see Prum *et al*., 1995.)

Phillips *et al*. (1987a,b) fit stationary Markov chain models of various orders to examine the abundance of *k*-words, $k = 2, \ldots, 6$. As a measure of abundance, they use signed likelihood

ratio "residuals," i.e., the separate terms in a likelihood-ratio test of the fit of the Markov chain model. For the word *W* this takes the form

$$s(W)=\text{sign}[N_W-\hat{E}(N_W)]\sqrt{e^2}$$

where

$$e^2=\begin{cases} 2\left[N_W \log\left(\frac{N_W}{\hat{E}(N_W)}\right)-(N_W-\hat{E}(N_W))\right] & \text{if } N_W>0 \\ 2\hat{E}(N_W) & \text{if } N_W=0 \end{cases}$$

Cross-ratios

Burge *et al*. (1992) define the term "odds ratio" for examining word representation in DNA sequences. Their expressions are more closely related to what are called cross-product ratios, or simply *cross-ratios* in the literature on categorical data (Agresti, 1990) rather than odds ratios. We will, therefore, use the term cross-ratio in the sequel. For a 2-word *ab* the corresponding cross-ratio is

$$t_{ab}=\frac{nN_{ab}}{N_aN_b} \quad (4)$$

while for 3-words *abc* and 4-words *abcd* the cross-ratios are

$$t_{abc}=\frac{N_{abc}N_aN_bN_c}{nN_{ab}N_{bc}N_{a+c}}$$

and

$$t_{abcd}=\frac{nN_{abcd}N_{ab}N_{bc}N_{cd}N_{a+c+}N_{a++d}N_{+b+d}}{N_{abc}N_{bcd}N_{a+cd}N_{ab+d}N_aN_bN_cN_d}$$

In these expressions a subscript + denotes the sum over all possible bases in that position, e.g., $N_{a+c} = \Sigma_{x\in\mathcal{B}} N_{axc}$.

These statistics appear to be inspired by expressions in the theory of log-linear models (see Agresti, 1990), which estimate interaction parameters within multidimensional contingency tables, but they are not modified in any way to take into account the sequential nature of DNA data. The 2-word cross-ratios (t_{ab}) have clear interpretations as measures of dependence, and can be considered against an i.i.d. model for the sequence of bases. By contrast, the 3-word measures (t_{abc}) are reminiscent of estimates of 3-way interaction in contingency tables, and the so-called "perfect" 3-way tables (π_{abc}) of probabilties, for which

$$\pi_{abc}=\frac{\pi_{ab}+\pi_{a+c}\pi_{+bc}}{\pi_{a++}\pi_{+b+}\pi_{++c}}$$

(Darroch, 1962). The main features of interest in such tables are (1) the absence of 3-way interaction, and (2) a compatibility between the 2-way parameters in the 3-way table and the analogous parameters in the corresponding 2-way subtable. However, neither of these issues seems to be relevant to frequencies generated by sequence data, at least not without some modification. Similar remarks apply to the 4-word measures (t_{abcd}).

Concordance among different representation measures

Using the HCMV sequence as an example, Table 1 gives the Spearman rank correlation coefficients between various pairs of the representation measures we have discussed. They are calculated in the following way: For k = 2, 3, 4, 5, every k-word is assigned a rank from 1 to 4^k by each of two different representation measures, and the rank correlation coefficient between these sets of 4^k ranks is then calculated using Spearman's formula (see, e.g., Daniel, 1995, chapter 13).

The Spearman rank correlation coefficients in Table 1 indicate that concordance among various representation measures are rather high, with the exception of the cross-ratios t for 3- and 4-words. To better visualize the inconsistency, we make a scatter plot (Fig. 4a) of z^L-ranks against t-ranks for the 4-words in HCMV. The random scattering of points is not observed in similar plots for more concordant pairs of statistics such as z^L and r (see Fig. 4b).

Fig. 4a shows that a fair number of words are classified overrepresented by t but underrepresented by z^L, and vice versa. In fact, a total of 19 words are ranked above the upper quartile by one measure but below the lower quartile by the other. Two extreme examples are *AGTA* (t-rank 14, z^L-rank 207) and *TCTG* (t-rank 212, z^L-rank 22). These observations demonstrate that different statistical measures can lead to quite different conclusions regarding the abundance of a particular word. Such differences may affect any biological inferences being made.

Pevzner *et al*. (1989a) note that the effect of the self-overlapping structure of words on their frequencies of occurrence cannot be adequately reflected by its expected count. For example, in an i.i.d. sequence with equal probabilities of observing any of the four bases, $E(N_{AA}) = E(N_{AC})$. However, while N_{AA} = 4 in the 5-base sequence "*AAAAA*," it requires a sequence of length at least 8 to observe *AC* four times. In contrast, the variance $V(N_W)$ does take the self-overlapping structure of W into consideration. This can be easily seen in both the unconditional and conditional calculations for a Markov chain model where terms involving words formed by overlapping W with another copy of itself are involved in the formulas of the second moments of N_W (Kleffe and Borodovsky, 1992; equation (7), Prum *et al*., 1995; Theorem 4). For this reason, z-scores may be preferred over the representation ratio r, the residual s, as well as the cross-ratio t.

The *z*-scores z^L, z^C, and z^M will produce almost identical word rankings as they are highly concordant (see Table 1 and also Prum *et al*., 1995 for a comparison of z^C with z^M). Among the three statistics, z^C is the only one derived from a model that faithfully reflects the shorter word composition of the data sequence. This statistic is perhaps the best among all the *z*-scores for 2- and 3-words. However, its computation for 4- and 5-words, which involves numerous determinant evaluations, is still rather slow on most of today's computers. Both z^L and z^M can be computed very efficiently, giving them an advantage over z^C. For the present application, we have chosen to use z^L since the conceptual simplicity of the maximum likelihood estimates can probably be better appreciated by biologists than the martingale formulation of z^M.

OBSERVATIONS ON SEVEN HERPESVIRUS GENOMES

The genome of a herpesvirus consists of a single DNA molecule that is packed into a dense donut-shaped core covered by an icosahedral capsid and the envelope. Between the capsid and envelope is a compartment called the tegument, which is composed of structural and regulatory viral proteins. A virus infects the host by introducing its genome into a suitable host cell. Inside the cell, herpesviruses may stay dormant for most of the time, and only become harmful after entering a lytic cycle in which they grow and replicate thousands of copies. Several human herpesviruses like herpes simplex, varicella-zoster, Epstein–Barr, and cytomegalovirus are associated with life threatening diseases such as AIDS and various cancers (Labrecque *et al*., 1995; Vital *et al*., 1995).

Herpesviruses have been classified into the α, β, and γ subfamilies according to biological properties such as range of hosts, types of infected cells, etc. (Cann, 1993, chapter 3). In this section, we analyze the full DNA sequences of seven herpesviruses. Table 2 lists their sizes, hosts, C + G contents, and classifications. The z^L-scores of 2-, 3-, 4-, and 5-words are used to explore sequence features that may be of biological interest. We are able to construct an evolutionary tree, and find several over- and underrepresented words common to all seven herpesviruses in our data set, which may possibly generalize to other large DNA molecules. Furthermore, clusters of some extremely over- or underrepresented words in individual genomes are located and identified with functional sites of these viruses.

Evolutionary tree by 2-word representation

Table 3 gives the z^L-scores of all 2-words and their corresponding ranks. We now convert this list of ranks into a set of pairwise distances with a view to clustering the viruses. More fully, if $r = (r_{AA}, r_{AC}, \ldots, r_{TT})$ is the vector of ranks of one virus, and $r' = (r'_{AA}, r'_{AC}, \ldots, r'_{TT})$ is that of a second, then we define the distance between these two viruses as the quantity

$$d = \frac{1}{16}(|r_{AA} - r'_{AA}| + |r_{AC} - r'_{AC}| + \cdots + |r_{TT} - r'_{TT}|)$$

This gives a matrix of distances between each pair of viruses. We then apply an agglomerative hierarchical clustering algorithm (Johnson and Wichern, 1992, chapter 12), as implemented in the S-plus "hclust" function (Statistical Sciences Inc., 1993), on these

pairwise distances. The result is the tree depicted in Figure 5a. In this tree, members within the α and γ subfamilies are grouped close together. The β virus HCMV is classified closer to the α family than the γ. HSI, whose classification is uncertain, turns out to be distant from all others. Unlike evolutionary trees inferred on the basis of the DNA sequences of specific genes or their amino acid sequences, this tree is founded on characteristics of entire genomes.

For comparison, we also construct a tree (Fig. 5b) in the same manner using the ranks of 2-word frequencies instead of the z^L-ranks. This grouping seems to express the similiarity in C + G content of the genomes (e.g., the two viruses low in C + G are grouped close together) rather than their biological classification. This confirms the value of word representation measures for revealing properties of genomes that are not necessarily indicated by the word frequencies directly.

The relationship among the seven genomes displayed in Fig. 5a agrees well with that implied by the star diagram presented in Karlin *et al*. (1994, Fig. 1). In that paper, the evolution of herpesviruses is studied using symmetrized cross-ratios for a large collection of herpes sequences, including those in our data set. The symmetrized ratios are the same as those defined in Eq. (4) above, except that the word frequency N_W is replaced by $(N_W + N_{W'})/2$, where W' is the reverse complement of the word W (e.g., if W is *AC*, W' will be *GT*). The symmetrization is necessary because their data set contains some DNA fragments of unknown relative orientation from one of the two strands of the genome. Given the high concordance of t with z^L for 2-words, and the similarity in base composition in the two strands of DNA as noted by Rogerson (1991) and Fickett *et al*. (1992), such agreement is not unexpected.

Pervasively over- and underrepresented 3-words

Common trends in 2-word representation of the herpesviruses and other DNA have been noted and discussed in a number of studies (see, e.g., Burge *et al*., 1992; Cardon *et al*., 1994). Since our results using 2-word z^L-scores are very similar, we do not repeat them here. Among the 3-words, six of them, namely *AAA*, *TTT*, *AGA*, *TCT*, *ATA*, *TAT*, are consistently ranked above the upper quartile (AUQ), and one other, *ACT* is consistently below the lower quartile (BLQ) for all seven genomes. No such consistency is found using word frequencies. The six AUQ words form three complementary pairs. *AGT*, the reverse complement of the BLQ word *ACT*, is BLQ for all except one (HSE1) of the viruses, where it is ranked 18. Table 4 displays the z^L-scores and ranks of these words.

Had the words been ranked independently and randomly in each individual virus, it would be very unusual to find any 3-word ranked AUQ or BLQ in all seven genomes (the probability is less than 1%). Is it possible that these AUQ and BLQ words are in some sense characteristic of the herpesviruses? Alternatively, could this representation pattern be a general property of DNA sequences? Table 5 indicates the z^L-rank of these words in various DNA and RNA molecules with respect to the three quartiles. The patterns of over- and underrepresentation of these words are similar in the other large DNA molecules found in eukaryotic cells, including their viral, mitochondrial, and chloroplast genomes. Some

contrary indications are seen with the prokaryotic phages and plasmids, as well as the short DNA and single-stranded viruses.

The pervasive overrepresentation of *AAA* had been reported by Nussinov (1980) and Brendel *et al*. (1986). They suggested the possibility of a connection with the molecular structure of DNA. In their study of *E. coli* data, Phillips *et al*. (1987b) proposed an explanation for the overrepresentation of *AAA* by codon usage (i.e., the preferential use of triplets of DNA bases to code for amino acids). It is also interesting to note that the words *TAT*, *ATA*, and *AAA* are frequently present in promoter sequences of herpesviruses (Wagner, 1991; Stinski *et al*., 1991), eukaryotic cells (Bucher and Trifonov, 1986), and even bacterial sequences (Waterman, 1989). These observations indicate that the overrepresentation of those three 3-words in the herpesviruses may be a consequence of some general biological properties of large DNA molecules.

On the other hand, Pevzner *et al*. (1989b) demonstrated that DNA words of the poly-*A*/*T* and poly-*C*/*G* forms are "nonstationary" in the sense that their occurrence frequencies fluctuate from one region of the genome to another. These authors suggested that deviations of occurrence frequencies of poly-*A*/*T* and poly-*C*/*G* words from their supposed expected values might not necessarily have any "biological significance," but could simply be attributable to the zonal structure of DNA, which was not accounted for by a homogeneous probability model. Four of our AUQ words, namely *AAA*, *TTT*, *ATA*, and *TAT*, would fit the poly-*A*/*T* characterization. However, the remaining AUQ words *AGA* and *TCT*, and the BLQ words *ACT* and *AGT*, alternating in *A*/*T* and *C*/*G* nucleotides, are neither poly-*A*/*T* nor poly-*C*/*G*. So far, we have not been able to relate any possible biological properties with the overrepresentation of *AGA* and *TCT* or the underrepresentation of *ACT* and *AGT* in the herpesviruses.

Clusters of extremely over- or underrepresented words

No distinctive across-family commonalities are shown among the z^L-scores for 4- and 5-words of the herpes genomes. Nevertheless, interesting studies relating over- and underrepresented words in an individual genome to its biological properties (e.g., the possible connection between the extreme underrepresentation of *CTAG* in *E. coli* and its DNA repair mechanism proposed by Merkl *et al*., 1992) has prompted us to look at those 4- and 5-words with outlying z^L-scores. We pick out the six most over- and underrepresented 4-words, and eight such extremal 5-words in each virus to examine their distributions in the genome. Their frequencies along different parts of the DNA sequence can be visualized conveniently by sliding window plots (with window lengths about 0.5% of the genome sizes) like those in Figure 6.

Over 80% of these extremal 4- and 5-words appear to be distributed fairly uniformly throughout the entire DNA sequence (see Fig. 6a for a typical sliding window plot). A number of them, however, exhibit very unusual clusters (Fig. 6b–n). These clusters drew our attention to 15 repetitive regions of the herpes genomes, whose locations are listed in Table 6. These segments either harbor numerous copies of a word that is extremely avoided in the rest of the genome, or are the main contributing factors to the extreme overrepresentation of a word.

Much information about the four human herpesviruses, especially HSV1 and EBV, can be found in the biological literature (e.g., McGeoch and Schaffer, 1993; Farrell, 1993; Masse *et al*., 1992; Kornberg and Baker, 1992; and the annotations of these genome sequences in their GenBank files). As a result, we are able to associate some of the sequence segments in Table 6 with known functional sites. For example, the clusters of *GAGGA* and *CCGCT* in HSV1 are found in the same locations as the repetitive "a sequences" responsible for the processing and packaging of newly synthesized viral DNA. The clusters of *CCCGC* in EBV and *CCGG* in HCMV are found around their lytic origins of replication. Other clusters are found in genes coding for tegument proteins and replication origin binding proteins.

CONCLUDING REMARKS

We have investigated several different representation measures as quantifiers of the relative abundance of short DNA words. They are shown to be highly concordant with one another for 2-words. With 3-and 4-words, the cross-ratio *t* gives a representation ranking rather different from other representation measures. Such contrasts may allow information encoded in the genome sequences to be interpreted from different perspectives.

In this study of the herpesviruses, we have employed homogeneous Markov models for entire DNA molecules. Since our purpose is to obtain an overview of the word representation characteristics of these genomes, we have not attempted to model their zonal structure (e.g., fluctuation of base composition, periodicity in the genes) in detail. However, this exploration into the herpes sequences, and the comparison of our observation with results from similar studies on other DNA data, have suggested several specific questions for further research. For example, it would be of interest to find out how much of the consistent over- or underrepresentation of the AUQ and BLQ words identified in these herpes genomes can be explained by codon usage in the genes, whether this representation pattern extends to long DNA sequences in eukaryotic cells, and whether viral origins of replication and DNA packaging signal sequences are often associated with clusters of highly over- or underrepresented words. Answers to these questions should help to relate word representation characteristics with the biological properties of a genome, and serve as statistical guidelines for identifying functional sites in other large DNA sequences.

Acknowledgments

We thank Karl Broman for his helpful comments on the manuscript. This work was supported in part by grants from the National Science Foundation (DMS 9113527), the Texas Higher Education Coordinating Board Advanced Research Program (010115-005), and the San Antonio Area Foundation.

References

Agresti, A. Categorical Data Analysis. John Wiley; New York: 1990.

Bhagwat AS, McClelland M. DNA mismatch correction by very short patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. Nucl Acids Res. 1992; 20(7):1663–1668. [PubMed: 1579457]

Billingsley, P. Probability and Measure. 3. John Wiley; New York: 1995.

Blaisdell BE. Markov Chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. J Mol Evol. 1985; 21:278–288. [PubMed: 6443131]

Brendel V, Beckmann JS, Trifonov EN. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. J Biomol Struct Dyn. 1986; 4(1):11–21. [PubMed: 3078230]

Bucher P, Trifonov EN. Compilation and analysis of eukaryotic POL II promoter sequences. Nucl Acids Res. 1986; 14:10009–10026. [PubMed: 3808945]

Burge C, Campbell AM, Karlin S. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA. 1992; 89:1358–1362. [PubMed: 1741388]

Cann, AJ. Principles of Molecular Virology. Academic Press; San Diego, CA: 1993.

Cardon LR, Burge C, Clayton DA, Karlin S. Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci USA. 1994; 91:3799–3803. [PubMed: 8170990]

Cowan R. Expected frequencies of DNA patterns using Whittle's formula. J Appl Prob. 1991; 28:886–892.

Cuticchia AJ, Ivarie R, Arnold J. The application of Markov chain analysis to oligonucleotide frequency prediction and physical mapping of Drosophila melanogaster. Nucl Acids Res. 1992; 20:3651–3657. [PubMed: 1641330]

Daniel, WW. Biostatistics: A Foundation for Analysis in the Health Sciences. 6. John Wiley; New York: 1995.

Darroch JN. Interactions in multi-factor contingency tables. J R Statist Soc B. 1962; 24:251–263.

Dembo A, Karlin S. Poisson approximations for r-scan processes. Ann Appl Prob. 1992; 2(2):329–357.

Farrell, PJ. Epstein-Barr virus. In: O'Brien, SJ., editor. Genetic Maps, Sixth Edition, Book 1, Viruses. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 1993. p. 1.120-1.133.

Fickett J, Torney D, Wolf D. Base compositional structure of genomes. Genomics. 1992; 13:1056–1064. [PubMed: 1505943]

Johnson, RA.; Wichern, DW. Applied Multivariate Statistical Analysis. 3. Prentice Hall; Englewood Cliffs, NJ: 1992.

Karlin S, Brendel V. Patchiness and correlations in DNA sequences. Science. 1993; 259(5095):677–680. [PubMed: 8430316]

Karlin S, Cardon LR. Computational DNA sequence analysis. Annu Rev Microbiol. 1994; 48:619–654. [PubMed: 7826021]

Karlin S, Ladunga I. Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA. 1994; 91:12932–12936. [PubMed: 7809149]

Karlin S, Mocarski ES, Schachtel GA. Molecular evolution of herpesviruses: Genomic and protein sequence comparisons. J Virol. 1994; 68(3):1886–1902. [PubMed: 8107249]

Kleffe J, Borodovsky M. First and second moment of counts of words in random texts generated by Markov chains. Comp Appl Biosci. 1992; 8(5):433–441. [PubMed: 1422876]

Kornberg, A.; Baker, TA. DNA Replication. 2. W. Freeman Co; New York: 1992.

Labrecque LG, Barnes DM, Fentiman IS, Griffin BE. Epstein-Barr virus in epithelial cell tumors: A breast cancer study. Cancer Res. 1995; 55(1):39–45. [PubMed: 7805038]

Leung MY, Schachtel GA, Yu HS. Scan statistics and DNA sequence analysis: The search for an origin of replication in a virus. Nonlinear World. 1994; 1:445–471.

Masse MJ, Karlin S, Schachtel GA, Mocarski ES. Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. Proc Natl Acad Sci USA. 1992; 89:5246–5250. [PubMed: 1319057]

McClelland M. Selection against dam methylation sites in the genomes of DNA of enterobacteriophages. J Mol Evol. 1985; 21:317–322. [PubMed: 6443311]

McGeoch, DJ.; Schaffer, PA. Herpes simplex virus. In: O'Brien, SJ., editor. Genetic Maps, Sixth Edition, Book 1, Viruses. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, NY: 1993. p. 1.147-1.156.

Merkl R, Kröger M, Rice P, Fritz HJ. Statistical evaluation and biological interpretation of nonrandom abundance in the *E. coli* K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair. Nucl Acids Res. 1992; 20(7):1657–1662. [PubMed: 1579456]

Nussinov R. Strong adenine clustering in nucleotide sequences. J Theor Biol. 1980; 85:285–291. [PubMed: 7431956]

Nussinov R. Nearest neighbor nucleotide patterns. Structural and biological implications. J Biol Chem. 1981; 256(16):8458–8462. [PubMed: 6943145]

Pevzner PA. Nucleotide sequences versus Markov models. Comput Chem. 1992; 16:103–106.

Pevzner PA, Borodovsky MY, Mironov AA. Linguistics of nucleotide sequences I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. J Biomol Struct Dyn. 1989a; 6(5):1013–1026. [PubMed: 2531596]

Pevzner PA, Borodovsky MY, Mironov AA. Linguistics of nucleotide sequences II: Stationary words in genetic texts and the zonal structure of DNA. J Biomol Struct Dyn. 1989b; 6(5):1027–1038. [PubMed: 2531597]

Phillips G, Arnold J, Ivarie R. Mono- through hexanucleotide composition of the Escherichia coli genome: A Markov chain analysis. Nucl Acids Res. 1987a; 15:2611–2626. [PubMed: 3550699]

Phillips G, Arnold J, Ivarie R. The effect of codon usage on the oligonucleotide composition of the E. coli genome and identification of over- and underrepresented sequences by Markov chain analysis. Nucl Acids Res. 1987b; 15:2627–2638. [PubMed: 3550700]

Prum B, Rodolphe F, De Turckheim E. Finding words with unexpected frequencies in DNA sequences. J R Statist Soc B. 1995; 57(1):205–220.

Rogerson AC. There appear to be conserved constraints on the distribution of nucleotide sequences in cellular genomes. J Mol Evol. 1991; 32:24–30. [PubMed: 1901365]

Schbath S, Trum B, De Turckheim E. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J Comp Biol. 1995; 2(3):417–437.

Statistical Sciences Inc. S-plus Reference Manual, Version 3.2. Vol. 1. StatSci., a division of MathSoft, Inc; Seattle: 1993.

Stinski, MF.; Maline, CL.; Hermiston, TW.; Liu, B. Regulation of human cytomegalovirus transcription. In: Wagner, EK., editor. Herpesvirus Transcription and Its Regulation. CRC Press; Boca Raton, FL: 1991. p. 245-260.

Venable, WN.; Ripley, BD. Modern Applied Statistics with S-plus. Springer-Verlag; New York: 1994.

Vital C, Monlun E, Vital A, Martin-Negrier ML, Cales V, Leger F, Longy-Boursier M, Le Bras M, Bloch B. Concurrent herpes simplex type 1 necrotizing encephalitis, cytomegalovirus ventriculoencephalitis and cerebral lymphoma in an AIDS patient. Acta Pathol. 1995; 89(1):105–108.

Wagner, EK. Herpesvirus transcription - general aspects. In: Wagner, EK., editor. Herpesvirus Transcription and Its Regulation. CRC Press; Boca Raton, FL: 1991. p. 1-15.

Waterman, MS. Consensus patterns in sequences. In: Waterman, MS., editor. Mathematical Methods for DNA Sequences. CRC Press; Boca Raton, FL: 1989. p. 93-115.

Waterman, MS. Introduction to Computational Biology. Chapman and Hall; New York: 1995.

Whittle P. Some distribution and moment formulae for the Markov chain. J R Statist Soc B. 1955; 17:235–242.

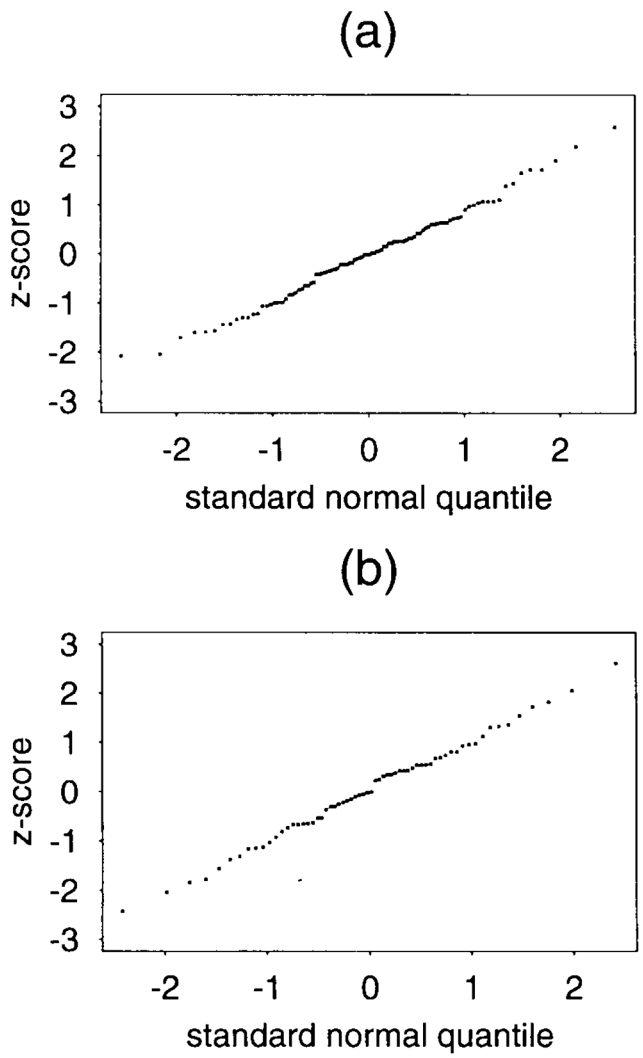

FIG. 1.

(a) Normal $q - q$ plot of z^L-scores of the 3-word *TTG* in 100 simulated Markov DNA sequences each with 229,354 bases and transition probabilities estimated from the human cytomegalovirus (HCMV) by maximum likelihood. (b) Normal $q - q$ plot of z^L-scores of all 64 3-words in one of the simulated Markov sequences.

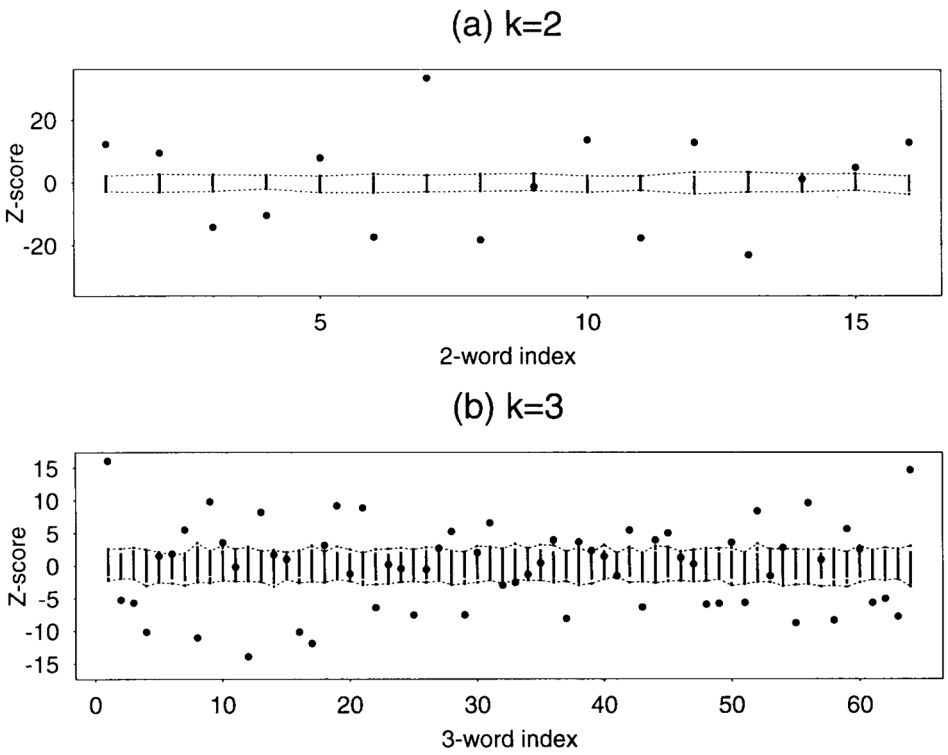

FIG. 2.

The z^L-scores of all (a) 2-words and (b) 3-words in lexicographical order for the 100 simulated Markov DNA sequences. The broken curves envelope the simulated z^L-scores. Solid dots are the z^L-scores of the HCMV sequence.

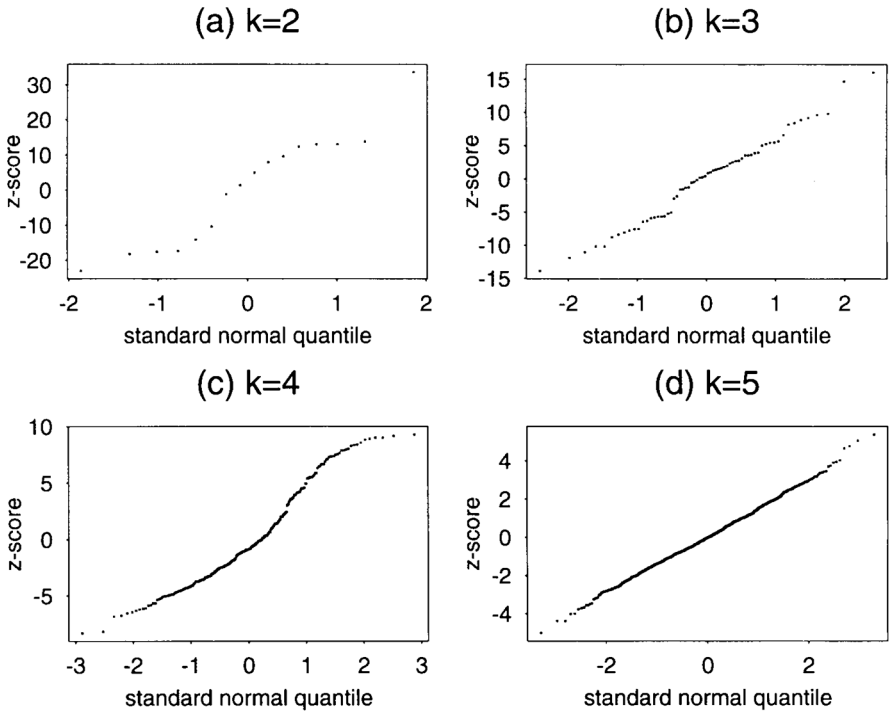

FIG. 3.

Normal $q - q$ plots of the z^L-scores of the HCMV sequence for all 2-, 3-, 4-, and 5-words.

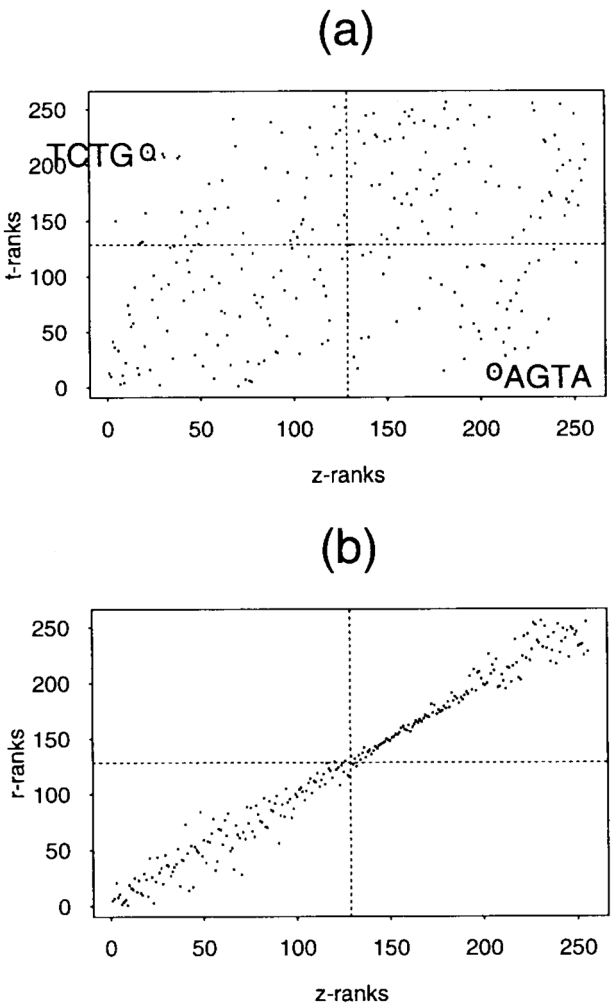

FIG. 4.

Scatter plots of 4-word ranks in HCMV ordered by (a) z^L-score versus cross-ratio t (the two words showing most severe difference in z^L- and t-rankings are identified); (b) z^L-score versus representation ratio r.

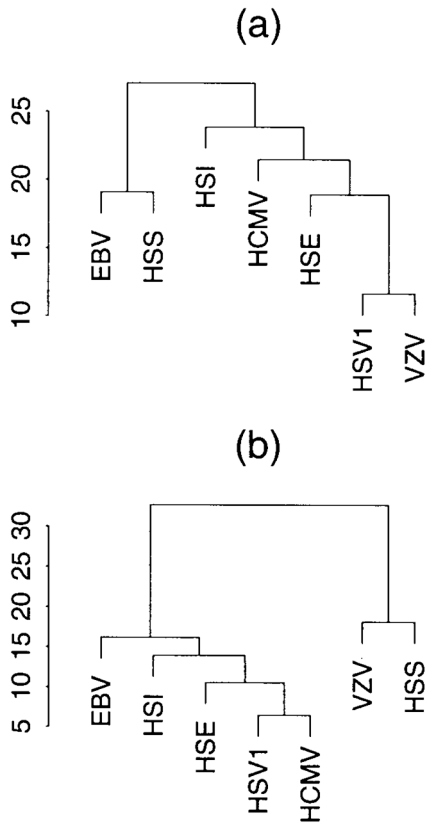

FIG. 5.

Tree (a), derived from the 2-word z^L-ranks, tends to group the viruses in the same family together. In contrast, the 2-word frequency based tree (b) reflects more of the similarity in base composition of the viruses.

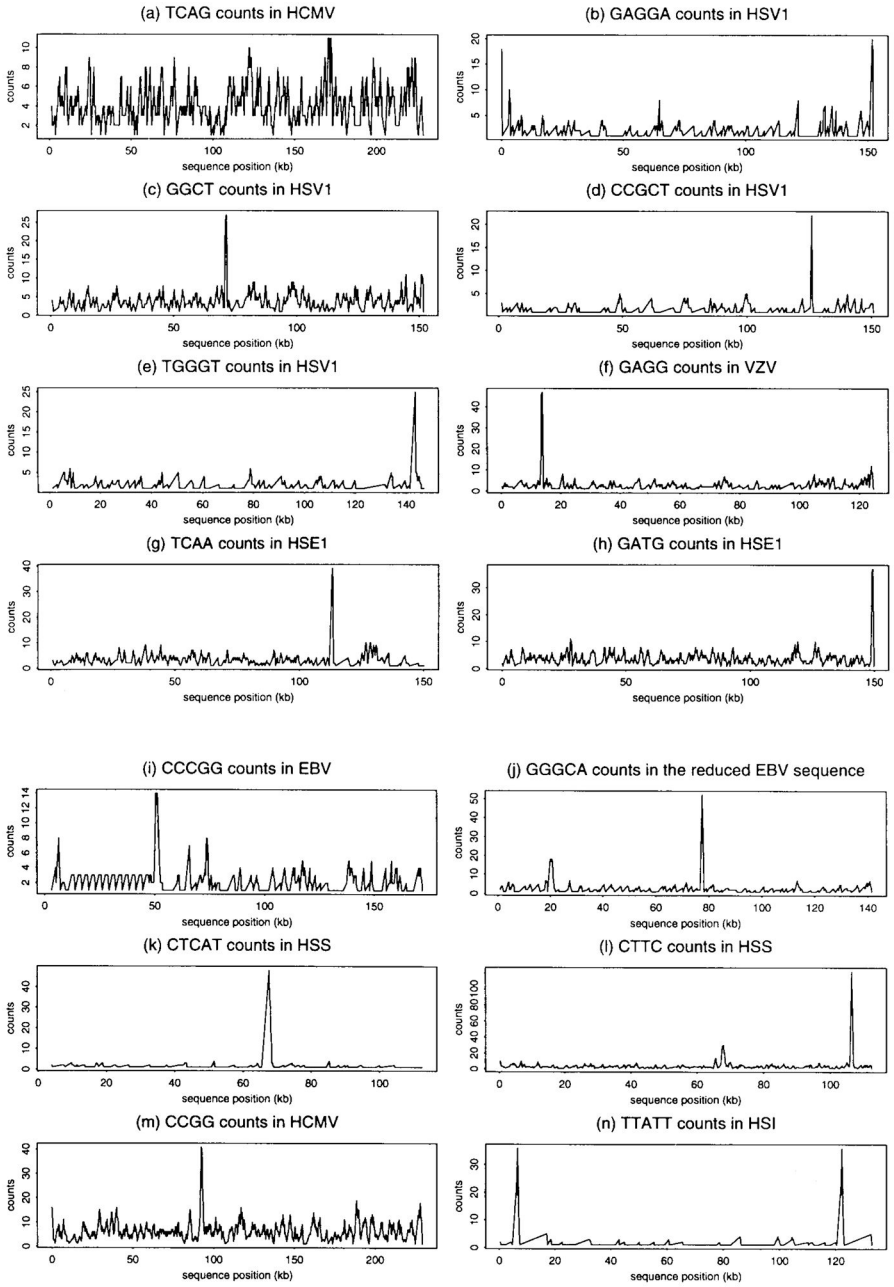

FIG. 6.

Word counts in a sliding window of length equal to 0.5% of the genome (rounded to the nearest hundred) are kept at each occurrence of the word. A typical sliding window plot of the extremal 4- and 5-words looks like graph (a). Graphs (b) through (n) show those extremal words with unusual clusters. These clusters are characterized to be significant ($p < 0.001$) by the *r*-scan statistics (Dembo and Karlin, 1992). An explanation of the application of scan statistics in evaluating clusters of special sequence patterns in a genome is given by Leung *et al*. (1994).

Table 1

Rank Correlation Coefficients for the HCMV Sequence between Different Measures of Representation[a]

Statistics	r	s	t	z^C	z^L	z^M
(a) 2-words below diagonal, 3-words above diagonal						
r	.	.99	.67	.99	.99	.99
s	1.00	.	.68	1.00	1.00	1.00
t	1.00	1.00	.	.68	.67	.68
z^C	.99	.99	.99	.	1.00	1.00
z^L	.98	.98	.98	.95	.	1.00
z^M	—	—	—	—	—	.
(b) 4-words below diagonal, 5-words above diagonal						
r	.	1.00	—	.99	.99	.99
s	.99	.	—	1.00	1.00	1.00
t	.37	.35	.	—	—	—
z^C	.99	1.00	.34	.	1.00	1.00
z^L	.99	1.00	.35	1.00	.	1.00
z^M	.99	1.00	.34	1.00	1.00	.

[a] Representation ratio (*r*), signed likelihood residual (*s*), cross-ratio (*t*), conditional *z*-score (z^C), MLE-based *z*-score (z^L), martingale *z*-score (z^M). A — indicates that one or both statistics are not calculated.

Table 2

Summary of the Seven Herpesviruses Analyzed

Name and abbreviation	Size (kb)[a]	Host	Type	C + G content (%)
Herpes simplex 1 (HSV1)	152	Human	α	68
Varicella-zoster (VZV)	125	Human	α	46
Equine herpes simplex 1 (HSE1)	150	Horse	α	57
Cytomegalovirus (HCMV)	229	Human	β	57
Epstein–Barr (EBV)	172	Human	γ	60
Herpes saimiri (HSS)	113	Monkey	γ	34
Ictalurid herpes (HSI)	134	Channel catfish	?	56

[a]Size of genomes measured in kilobases (kb).

Table 3

The z^L-Scores (Ranks) for All 2-Words in Seven Herpesviruses

Word	HSV1	VZV	HSE	HCMV	EBV	HSS	HSI
AA	15.12 (16)	11.20 (12)	10.10 (15)	12.38 (12)	5.07 (9)	8.59 (11)	−0.40 (8)
AC	−0.20 (7)	5.95 (10)	−0.79 (8)	9.56 (11)	−26.54 (2)	−2.28 (8)	5.51 (11)
AG	−14.82 (3)	−28.60 (2)	−5.54 (5)	−14.19 (5)	27.82 (16)	18.48 (14)	−19.79 (3)
AT	2.49 (11)	6.35 (11)	−5.79 (4)	−10.39 (6)	−7.27 (6)	−27.88 (2)	15.83 (14)
CA	−0.03 (8)	1.58 (9)	0.16 (10)	8.00 (10)	19.43 (13)	21.29 (15)	−1.92 (6)
CC	8.57 (13)	12.13 (13)	1.52 (11)	−17.30 (4)	27.25 (15)	−6.55 (6)	3.28 (9)
CG	1.96 (10)	12.87 (14)	−0.70 (9)	33.58 (16)	−67.39 (1)	−41.91 (1)	15.15 (13)
CT	−17.54 (1)	−28.75 (1)	−1.68 (7)	−18.24 (2)	10.24 (11)	17.60 (12)	−18.85 (4)
GA	−2.49 (6)	−9.97 (4)	−4.55 (6)	−1.13 (7)	−4.38 (7)	−6.70 (5)	29.24 (15)
GC	−14.94 (2)	−5.36 (6)	7.22 (14)	13.79 (15)	−17.48 (4)	17.93 (13)	−38.49 (1)
GG	8.73 (14)	14.06 (15)	2.32 (12)	−17.60 (3)	22.01 (14)	−7.47 (4)	4.14 (10)
GT	3.23 (12)	−1.75 (7)	−6.56 (3)	12.98 (14)	−10.70 (5)	−1.01 (9)	5.60 (12)
TA	−14.33 (4)	−7.25 (5)	−7.89 (2)	−23.05 (1)	−23.41 (3)	−27.23 (3)	−28.85 (2)
TC	1.46 (9)	−15.83 (3)	−9.13 (1)	1.40 (8)	2.17 (8)	−5.22 (7)	30.27 (16)
TG	−3.13 (5)	−0.58 (8)	2.88 (13)	4.99 (9)	11.57 (12)	21.96 (16)	−2.46 (5)
TT	14.27 (15)	17.03 (16)	11.80 (16)	12.96 (13)	6.19 (10)	8.30 (10)	−1.10 (7)

J Comput Biol. Author manuscript; available in PMC 2014 June 30.

Table 4

The z^L-Scores (Ranks) of Those AUQ and BLQ 3-Words Common to All Seven Herpesviruses[a]

Virus	AAA	TTT	AGA	TCT	ATA	TAT	ACT	AGT
HSV1	10.36(64)	10.35(63)	5.42(52)	6.04(54)	8.65(60)	8.99(61)	−7.19(8)	−8.98(5)
VZV	4.97(54)	3.52(50)	8.26(62)	8.90(64)	6.36(56)	8.40(63)	−4.32(16)	−6.30(6)
HSE1	11.03(63)	11.96(64)	6.03(55)	6.11(56)	6.14(57)	9.18(61)	−4.25(15)	−3.98(18)
HCMV	16.06(64)	14.64(63)	9.79(62)	9.62(61)	8.17(57)	8.36(58)	−11.04(3)	−13.89(1)
EBV	9.56(62)	9.00(61)	10.43(63)	10.97(64)	5.91(54)	4.23(51)	−8.17(4)	−7.93(5)
HSS	5.38(57)	4.99(56)	6.51(59)	6.04(58)	6.96(62)	6.76(60)	−4.90(12)	−5.21(9)
HSI	10.78(61)	9.78(59)	5.18(50)	6.92(56)	8.13(57)	11.43(63)	−5.81(15)	−6.12(13)

[a]The only exception is *AGT*, which is ranked 18 in HSE1.

Table 5

Ranking of z^L-Scores of the AUQ/BLQ Words of the Herpesviruses in Various Nucleic Acid Molecules[a]

Genome	AAA	TTT	AGA	TCT	ATA	TAT	ACT	AGT
Long DNA molecules in eukaryotic cells and their organelles and viruses								
HCMV	++	++	++	++	++	++	− −	− −
Vacinna virus	++	++	++	++	++	+	− −	− −
Adenovirus	++	++	++	++	+	++	−	− −
Yeast chromosome III	++	+	++	++	++	++	−	−
Tobacco chloroplast	+	++	++	++	++	++	−	−
Liverwort mitochrondrion	++	++	+	+	++	++	− −	− −
Prokaryotic phages and plasmids								
E. coli phage λ	++	++	−	+	++	++	−	+
Mycobacterial phage L5	−−	− −	−	−	−	− −	−	− −
C. burnetii plasmid QpH1	+	++	−	+	++	++	−	− −
Birmingham IncP-α plasmid	−	−	−	− −	+	+	− −	− −
Short DNA viruses								
Human papilloma	++	++	++	++	−	++	− −	− −
SV40	++	++	− −	− −	−	++	−	+
Human polyoma	++	++	−	++	++	++	− −	+
Human hepatitis B	+	+	++	+	+	++	−	− −
Single-stranded viruses								
Mouse minute virus	+	++	++	−	+	+	+	−
Encephalitis	++	+	++	− −	+	+	−	− −
Measles	−	− −	−	−	−	++	−	−
HIV	+	+	++	+	+	++	+	+

[a]Here, ++ (respectively − −) indicates AUQ (respectively BLQ) ranking, and + (respectively −) indicates a ranking between the 2nd and 3rd (respectively 1st and 2nd) quartiles.

Table 6

Sequence Segments Containing Significant Clusters of Extremal 4-and 5-Words

Genome	Word	z^L (rank)	Cluster location	Feature
HSV1	*GAGGA*	3.86 (1019)	1–552 151709–152260	Terminal segments (the “a sequence”) containing signals for processing/packaging nascent DNA
	GGCT	−5.52 (4)	71078–72077	Complement in UL36[a], encoding a tegument protein
	CCGCT	−3.00 (6)	125983–126782	Contains inverted complement of the “a sequence”
	TGGGT	4.29 (1024)	143674–144473	
VZV	*GAGG*	6.38 (256)	13904–14504	In ORF11[a], possibly encoding a tegument protein
HSE1	*TCAA* *GATG*	5.64 (252) 5.85 (253)	113353–114152 149036–150223	
HCMV	*CCGG*	−6.86 (3)	92756–93855	In oriLyt[b]
EBV	*CCCGC*	−5.05 (1)	50659–51558	Part of repetitive region upstream of oriLyt[b]
EBV[c]	*GGGCA*	3.90 (1024)	108239–108938	In EBNA1[a] gene, encoding an oriP[b] binding protein
HSS	*CTCAT* *CTTC*	3.90 (1020) 9.82 (256)	67575–68174 106513–107112	
HSI	*TTATT*	6.82 (1024)	6585–7284 122255–122954	

[a] Conventional names designated to coding segments of genomes. UL36 refers to the segment 71052–80543 on the complementary strand of HSV, which codes for a very large tegument protein. ORF11 is the segment from 13590 to 16049 of VZV. It is homologous to UL47 of HSV, which is also a tegument protein gene. The segment 107950–109875 of EBV codes for the protein EBNA1, which is involved in regulating the viral DNA replication process.

[b] Conventional names designated to various origins of replication.

[c] The EBV genome contains 11.6 copies of a 3072 base segment repeated tandemly. The word *GGGCA* is not among the extremal words of EBV, but is extremal only when 10 copies of the repeating unit (which makes up > 17% of the genome) are removed. The cluster location of *GGGCA* has been adjusted to give its corresponding position in the original genome.