Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses

David S. H. Chew¹,*; Kwok Pui Choi¹,² and Ming-Ying Leung³,⁴

¹Department of Mathematics and ²Department of Statistics and Applied Probability, National University of Singapore, Singapore, ³Bioinformatics Program and ⁴Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968, USA

Received May 30, 2005; Revised June 27, 2005; Accepted August 15, 2005

ABSTRACT

Many empirical studies show that there are unusual clusters of palindromes, closely spaced direct and inverted repeats around the replication origins of herpesviruses. In this paper, we introduce two new scoring schemes to quantify the spatial abundance of palindromes in a genomic sequence. Based on these scoring schemes, a computational method to predict the locations of replication origins is developed. When our predictions are compared with 39 known or annotated replication origins in 19 herpesviruses, close to 80% of the replication origins are located within 2% of the genome length. A list of predicted locations of replication origins in all the known herpesviruses with complete genome sequences is reported.

INTRODUCTION

Early studies (1,2) have reported that the nucleotide sequences around replication origins of certain herpesviruses have complex repetitive structures of closely spaced direct and inverted repeats. A palindrome is a special case of inverted repeats where a segment of nucleotides is immediately followed by its reverse complement. A high concentration of palindromes around replication origins has been found in these herpesviruses.

Herpesviruses utilize two different types of replication origins during lytic and latent infections. For each type of origins, the count and locations in the genome vary from one kind of herpesvirus to another. Most herpesviruses have one to two copies of latent and lytic origins. Presence of palindromes around replication origins is prevalent in both latent and lytic types (1–5).

As the central step in the reproduction of herpesviruses, viral DNA replication has been the target for a number of anti-herpesvirus drugs (e.g. acyclovir). Understanding the molecular mechanisms involved in DNA replication is of great importance in further developing strategies to control the growth and spread of viruses (6–8). Since replication origins are regarded as major sites for regulating genome replication, labor-intensive laboratory procedures have been used to search for replication origins (9–11).

With the increasing availability of genomic DNA sequence data, one way that may save time and resources would be to scan the viral genome sequence for the expected sequence features by a computer program before an experimental search for replication origins is launched. Masse et al. (3) first used this computational approach to predict the replication origin oriLyt on the human cytomegalovirus (HCMV) and then confirmed it by experimentation. In that computational analysis, one of the sequence features being scanned for in the genome sequence is the presence of a high concentration of palindromes of length 10 or above clustering within a window of 1000 bases.

A palindrome reads exactly the same from the 5’ end to the 3’ end on both strands of DNA (see Figure 1 for example). More precisely, we can define a palindrome to be a word pattern of the form \(b_1 \ldots b_L b_L' \ldots b_1'\), where \(b'\) is the complement of base \(b\) and \(L\) is the half-length of the palindrome. We call the letter \(b_L\) the left-center and \(b_L'\) the right-center of the palindrome.

The length of the palindrome in Figure 1 is 10 and \(L = 5\).

Figure 1. A palindrome of length 10.

⁴To whom correspondence should be addressed. Tel: +65 6847 1653; Fax: +65 6779 5452; Email: matchewd@nus.edu.sg

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org
Palindromes play important roles as protein-binding sites in DNA replication processes [(12), Chapter 1]. The local 2-fold symmetry created by the palindrome provides a binding site for DNA-binding proteins which are often dimeric in structure. Such double binding markedly increases the strength and specificity of the binding interaction [(13), Chapter 8]. High concentration of palindromes around replication origins is generally attributed to the reason that the initiation of DNA replication typically requires the binding of an assembly of enzymes to these DNA sequences. Helicase is an example of these enzymes known to bind with the initiation site, locally unwind the DNA helical structure, and pull apart the two complementary strands. This explanation is consistent with the observation of AT-rich regions, believed to facilitate the complementary strands. This explanation is consistent with unwinding, in replication origin domains of the genome (5).

Leung et al. (14) describe how an evaluation criterion, based on the scan statistics (15,16), is developed for assessing palindrome clusters by modeling the occurrences of palindromes in the genome as points randomly sampled from the unit interval according to the uniform distribution. By identifying windows on the genome sequence containing statistically significant clusters of palindromes, the scan statistics, in principle, provide a method to predict likely locations of replication origins. This criterion, however, essentially assesses a window of the genome by only the counts of palindrome contained in it, regardless of the actual extent of the palindrome lengths. This drawback has led to missing some replication origins which contain one extremely long palindrome rather than a cluster of moderately long ones. In the present paper, we propose two new schemes for evaluating palindrome clusters and use the rankings of these evaluation criteria to predict the replication origins in the herpesviruses. By checking with known replication origins reported either in published literature or GenBank annotations, we assess the accuracy of the new prediction schemes. These assessments demonstrate that there is a substantial improvement over the original scan statistics criterion.

In Methods section, we describe the main steps of the prediction method and three scoring schemes. The first scoring scheme, called the palindrome count scheme (PCS), is essentially the scan statistics method first described by Leung et al. (14), and further discussed in the articles of Leung and Yamashita (17), and Leung et al. (4). Two new scoring schemes, namely, the palindrome length scheme (PLS) and the base-pair weighted scheme (BWS) are introduced as measures of palindrome clusters. In Results and Discussion section, we report the results of applying these scoring schemes to predict the locations of replication origins for 39 fully sequenced herpesviruses, and compare the prediction accuracies in terms of sensitivity and positive predictive value. A few concluding remarks are given in the final section.

METHODS

We propose a computational method to identify regions of a genome which harbor unusual clusters of palindromes. This, in turn, becomes the basis of our method to predict replication origins for the herpesviruses. Table 1 presents the viruses to be analyzed. The data set comprises all complete genome sequences of the herpesvirus family downloaded from GenBank at the NCBI web site in April 2005. For each virus, we list its abbreviation, accession number, sequence length and the relative frequencies of the four nucleotide bases in the genome (see Table 1).

Our method for predicting replication origins consists of four basic steps: (i) locate palindromes at or above a prescribed length; (ii) choose a scoring scheme for palindromes; (iii) compute a score for each window of the genome according to the chosen scoring scheme; and (iv) select regions with high scores.

Step (i): Locating palindromes at or above a prescribed length

As very short palindromes occur frequently by chance, a parameter, L, needs to be chosen where palindromes of length below 2L will not be considered in the analysis. Leung et al. (4) propose a procedure, which is based on bench-marking with the well-studied HCMV virus, for the choice of L. This choice takes into account the length of the sequence, as well as the base frequencies in the genome. Using this criterion, L is chosen to be 6 for the BoHV1, BoHV5, CeHV1, HSV1, HSV2 and SHV1 sequences and 5 for the other sequences. Once the minimal palindrome length has been chosen, the sequences are run through the palindrome program, which is part of EMBASSS [European Molecular Biology Open Software Suite, (18)], to extract the palindrome positions and lengths. Each of these palindromes will be assigned a score according to a scoring scheme chosen in the next step. Note that although it is possible for one palindrome to contain a shorter one in it (e.g. the length 12 palindrome ACCGTGCACGGT contains the length 10 palindrome CCGTGCACGG), EMBASSS automatically discards the shorter redundant palindrome and report only the longest one.

Step (ii): Choosing a scoring scheme for palindromes

Three schemes for scoring palindromes are described. In all of them, any palindrome of length less than 2L will always get a score 0.

(i) **Palindrome count score** (PCS): In this scoring scheme, a palindrome is given a score 1 when its length is at or above 2L.

(ii) **Palindrome length score** (PLS): A palindrome of length 2s ≥ 2L is given a score s/L. For example, if we let L = 5, a palindrome of length 10 will get a score of 1, while one of length 24 will get a score of 2.4.

(iii) **Base-pair weighted score** of order m (BWS_m).

The idea behind BWS is that a higher score should be given to rarer palindromes, namely those which have lower probabilities to occur by chance. We assess the probability of occurrence of a particular palindrome based on Markov type sequence models [(19),Chapter 3]. Here M denotes the order of the Markov chain. Then, we take the negative logarithm of the probability of a palindrome to give it a positive score which is higher when the probability is lower.

We give a simple example of calculating the BWS0 score. In the Markov model with order m = 0, the letters in the sequence are independent of each other. A palindrome containing respectively n_A, n_C, n_G, n_T of A, C, G and T occurs with probability 

\[ p_A^{n_A} p_C^{n_C} p_G^{n_G} p_T^{n_T} \]

where \( p_A, p_C, p_G, p_T \) are the relative base frequencies in the sequence. The BWS0 score of such a palindrome will be the negative logarithm of this probability,
which is equal to $-n_A \log p_A + n_C \log p_C + n_G \log p_G + n_T \log p_T$.

Consider two palindromes: $CAGTACGTG$ and $TTTTTAAAAA$ in a very $CG$-rich genome, say, with relative base frequencies $p_A = 0.1$ and $p_C = p_G = 0.4$. The latter palindrome is much less likely to occur than the former, and accordingly should receive a higher score to reflect its rarity compared with the former. Indeed, the calculated scores of the two palindromes turn out to be $14.7$ for the former and $14.4$ for the latter.

**Step (iii): Computing the window score**

The score of a window in the genome is simply the total of the scores of all the palindromes occurring in this window. A palindrome is considered in the window if its left-center is at or before the $w$th bases and so on. Because of the way the sliding windows are constructed, the length of the last window is usually shorter than $w$.

**Step (iv): Selecting regions with significant palindrome clusters**

For the PCS, regions that harbor statistically significant clusters of palindromes are identified using the scan statistics criterion as described in Leung et al. (14). As the criteria for statistical significance for PLS and BWS have not yet been established, we use a non-parametric approach where a fixed number of top scoring windows are chosen as the predicted locations of replication origins. It is well known that herpesviruses have multiple replication origins. However, there does not appear to be any obvious rule to determine the number of top scoring windows that one should take. Based on sensitivity and positive predictive value consideration (defined below), we find that using the top 3–5 ranked windows for prediction works well for the herpesviruses.

**RESULTS AND DISCUSSION**

**Scan statistics method versus the new scoring schemes**

To compare and contrast the two new scoring schemes with the scan statistics method, now called PCS, the sliding window plots for HCMV and HSV1 using PCS, PLS and BWS score schemes are displayed in Figure 2. In each plot, the scores of
the windows are plotted against the position of the window. For HCMV, the highest scoring window is the same for all three schemes. This window corresponds to the oriLyt of the HCMV identified by Masse et al. (3). For HSV1, however, the plot of the PCS look rather different from those of the PLS and BWS. The highest scoring window in each of PLS and BWS corresponds to the oriL, and the two next highest peaks are close to the two oriS. In contrast, the PCS fails to locate any significant clusters of palindromes.

Table 2 shows the top 3 scoring windows for each of the 39 viruses under both the PLS and BWS schemes. The numbers in the table indicate the middle positions of the windows. In cases where two or more high scoring windows are close to one another, only one of them is picked to represent the region that gave the high scores. We adopt the practice that when a certain high scoring window is chosen, the neighboring 8 windows both to the left and to the right of it will not be considered subsequently. Rows that are shaded indicate that the particular viruses have known replication origins either from literature or from annotation. Underlined entries denote that the particular viruses have known replication origins either from literature or from annotation. Underlined entries denote that the middle positions of the windows which are within 2 map units (a map unit, abbreviated mu, is 1% of the genome length) of known replication origins. Shaded rows without any underlined entries show that the computational method fails to predict the known origins of replication. Finally, rows that are not shaded denote those viruses whose origins of replication are not known, as far as we know. Table 3 lists the regions with significant clusters of palindromes as found by the PCS scheme.

**Prediction accuracy**

We next examine the correspondence between the locations of these high scoring windows and those of the known replication origins. From Genbank sequence entries, annotations and literature, we are able to compile a list of 39 known replication origins for some of the viruses in our dataset. Table 4 shows the distance between each known origin from the nearest significant palindrome cluster for PCS, or the nearest high scoring window for PLS and BWS1 if the center of the cluster or window is within 2 mu of the origin. Otherwise a ‘—’ is entered. The distance is calculated from the mid-point of the window to the mid-point of the closest replication origin. Clearly, Table 4 shows that both PLS and BWS present a substantial improvement in the prediction accuracy of replication origins. For the PLS and BWS, we have used the top 3 scoring windows for each virus to construct this table.

Prediction accuracy of the different schemes can be quantified by two commonly accepted measures: sensitivity and positive predictive value (PPV). In our context, sensitivity is the percentage of known origins that are close to the regions suggested by the prediction; and positive predictive value is the percentage of identified regions that are close to the known origins.

Figure 3 shows the performance of the various schemes. For the PLS and BWS1, the sensitivity and positive predictive value using 1–10 top scoring windows are given in percentages. Results from BWS0 and BWS2 are also obtained (data not shown). Their prediction accuracies are close to but slightly less than that of BWS1. Note that as the number of windows increases, we gain in sensitivity but at the same time lose in positive predictive value. The highest sensitivities attained by PLS and BWS1 are 67 and 79%, respectively. The highest positive predictive values for both schemes are 47%.

**Difference between PLS and BWS**

Note that both PLS and BWS take the length of the palindromes into account, as longer palindromes have lower
probability of occurrence than shorter ones. Moreover, the BWS takes into account the base and word frequencies which affect the probability of occurrence of the palindrome. Consider, for example, the BWS0 score

\[ \text{BWS0} = n_A \log p_A + n_C \log p_C + n_G \log p_G + n_T \log p_T \]

can be viewed as a weighted sum, with weights according to the negative logarithms of the base frequencies. If the base probabilities are all equal, the BWS0 will reduce to \((\log 4)(n_A + n_C + n_G + n_T)\) which is equal to \((\log 4) \times \text{Length of palindrome and hence is equivalent to the PLS.}\]

In essence, the BWS includes more information about the sequence in its prediction and so we expect it to give better prediction accuracy. Our results show that this is indeed true. When we choose to use 3 or more top ranking windows, the BWS performs better than the PLS in terms of (higher) sensitivity and positive predictive value.

Suspecting that the probability of occurrence of palindromes might not be well estimated on the basis of a global base and word frequencies, we also try calculating palindrome probabilities using the base and word frequencies of those at the local window rather than those of the entire genome. Figure 4 shows the sensitivity and positive predictive values of the local BWS of order 0, 1 and 2. We use BWS\(^m\) (Local) to represent the local version of BWS of order \(m\). According to these results, the local version still does not perform any better than BWS1.

Further improvement of the algorithm

While our results show that using PLS and BWS with the ranking approach clearly outperforms the PCS, we have to note that the PCS is the only scheme where a rigorous statistical significance criterion, based on the probability distribution of the scan statistics, is currently available. The probability distributions of the maximal window scores with PLS and BWS have yet to be established. We have some preliminary results on approximating the distributions of the window score under PLS by compound Poisson distribution.

### Table 2. High scoring windows of PLS and BWS\(^1\)

| Virus      | PLS rankings | BWS\(^1\) rankings |
|------------|--------------|---------------------|
|            | 1            | 2                   | 3                   | 1            | 2                   | 3                   |
| AIHV1      | 113701       | 32701               | 123301              | 113701       | 123301              | 32701               |
| AhHV3      | 99001        | 54751               | 97001               | 99251        | 97001               | 54751               |
| BoHV1      | 113401       | 124501              | 103801              | 113401       | 124501              | 87301               |
| BoHV4      | 30251        | 54751               | 72251               | 54751        | 72251               | 54751               |
| BoHV5      | 19201        | 78001               | 107401              | 18901        | 113401              | 129601              |
| CaHV3      | 116201       | 133351              | 23101               | 116201       | 133351              | 23101               |
| CCMV       | 91201        | 207001              | 177001              | 177001       | 177001              | 177001              |
| CeHV1      | 133001       | 149451              | 61601               | 133001       | 149451              | 61601               |
| CeHV15     | 8001         | 34801               | 138801              | 8001         | 34801               | 138801              |
| CeHV2      | 129501       | 144201              | 61601               | 129501       | 144201              | 61601               |
| CeHV7      | 18601        | 93601               | 15601               | 18601        | 106601              | 121801              |
| CeHV8      | 161151       | 147401              | 198001              | 161151       | 147401              | 198001              |
| EV         | 76001        | 53201               | 127601              | 76001        | 53201               | 127601              |
| EHV1       | 116201       | 146651              | 47601               | 116201       | 146651              | 47601               |
| EHV2       | 6301         | 54001               | 173251              | 6301         | 173251              | 6301                |
| EHV4       | 105351       | 142801              | 3851                | 105351       | 142801              | 3851                |
| GaHV1      | 41651        | 68601               | 99751               | 68601        | 41651               | 99751               |
| GaHV2      | 160801       | 801                 | 137601              | 801          | 137601              | 801                 |
| GaHV3      | 158801       | 138401              | 11201               | 138401       | 11201               | 138401              |
| HCMV       | 94051        | 196351              | 77001               | 94051        | 196351              | 77001               |
| HCMV-M     | 175451       | 94051               | 153451              | 175451       | 94051               | 153451              |
| HHV6       | 30101        | 8051                | 110601              | 8051         | 30101               | 110601              |
| HHV6B      | 90401        | 69201               | 132801              | 90401        | 132801              | 90401               |
| HHV7       | 133351       | 9351                | 127401              | 9351         | 127401              | 9351                |
| HHV8       | 23401        | 119401              | 15001               | 23401        | 119401              | 15001               |
| HSV1       | 62301        | 129851              | 148401              | 62301        | 129851              | 148401              |
| HSV2       | 73551        | 7351                | 119701              | 73551        | 119701              | 73551               |
| IcHV1      | 55501        | 9301                | 89701               | 89701        | 9301                | 89701               |
| MCMV       | 92951        | 142451              | 200201              | 92951        | 142451              | 200201              |
| MeHV1      | 5601         | 117951              | 11551               | 5601         | 117951              | 11551               |
| MMRV       | 132601       | 3301                | 117601              | 132601       | 117601              | 132601              |
| MuhV4      | 99251        | 26251               | 62001               | 99251        | 26251               | 62001               |
| OsHV1      | 21001        | 144001              | 185001              | 21001        | 144001              | 185001              |
| PSHV1      | 130401       | 151601              | 26801               | 130401       | 151601              | 26801               |
| RC MV       | 75901        | 110551              | 83601               | 75901        | 110551              | 83601               |
| SaHV1      | 105751       | 112501              | 27751               | 105751       | 112501              | 27751               |
| SHV1       | 38151        | 19151               | 11551               | 38151        | 19151               | 11551               |
| ThHV       | 134101       | 103001              | 50401               | 134101       | 103001              | 50401               |
| VZV        | 119401       | 110101              | 100501              | 119401       | 110101              | 100501              |

The numbers in the table indicate the middle positions of the windows. Rows that are shaded indicate that the particular viruses have known replication origins either from literature or from annotation. Underlined entries denote the middle positions of the windows which are within 2 map units (i.e. 2% of the genome length) of known replication origins.
The compound Poisson distribution is motivated from a marked Poisson process point of view. The occurrence of a palindrome of length $2L$ and above is modeled by a Poisson process (4), and the actual length of this palindrome is modeled by a geometric distribution.

On closer examination of the known replication origins in this set of genome sequences, we notice that some of the origins missed by this prediction algorithm are actually rather long approximate palindromes. They are missed because we choose to consider only the perfect palindromes. For example, in HSV2, allowing just one error would have let us pick up a 136 base long approximate palindrome centered at 62 930, which is where the reported replication origin is located. If we include these approximate palindromes in our consideration, the sensitivity can be further increased.

**CONCLUDING REMARKS**

It is mentioned in the introduction that palindromes are merely one type of sequence features known to be associated with replication origins. Other frequently observed characteristics around replication origins include clustering of closely spaced direct and inverted repeats, as well as high AT content. We have actually examined each of these other types of sequence features and found that none of them, when used alone on our data set, reaches the same level of prediction accuracy offered by the BWS. However, it is likely that the prediction accuracy can be further improved by appropriately incorporating them in the prediction scheme. In fact, several replication origins in BoHV4, EHV4 and HSV2 which are not identified by any of PCS, PLS or BWS can be easily detected by the high local AT content around them. Exactly in what way all the different

**Table 3. Regions with significant clusters of palindromes as found by the PCS**

| Virus      | Region                                                   |
|------------|----------------------------------------------------------|
| AlHV1      | 113 456–113 759                                         |
| AdHV3      | 95 350–100 098                                          |
| BoHV1      | 77 155–77 168, 102 895–106 948, 113 462–113 636, 124 582–124 756, 131 268–135 221 |
| CalHV3     | 21 899–23 918, 115 406–117 660, 133 180–133 587         |
| CCMV       | 88 376–93 639, 206 555–207 582                           |
| CeHV1      | 112 833–113 219                                         |
| CeHV8      | 147 015–147 280, 158 953–164 225                         |
| CeHV15     | 5182–10 840, 32 483–36 810, 137 852–139 781, 150 277–152 289 |
| EBV        | 6772–11 675, 49 460–54 858                               |
| EHV1       | 115 125–119 096, 144 064–148 035                         |
| EHV2       | 4911–9106, 147 228–147 250, 171 785–175 980             |
| GaHV3      | 10 409–11 952, 104 965–105 067, 121 153–123 174, 138 321–138 935, 158 536–159 150 |
| HCMV       | 90 515–95 115, 195 962–196 203                           |
| HCMV-M     | 90 881–96 835, 175 177–176 003, 201 246–201 487         |
| HHV6b      | 88 469–94 716                                           |
| HHV7       | 124 985–128 653                                          |
| HHV8       | 21 913–23 705                                           |
| MCMV       | 92 621–93 412, 142 118–142 186                           |
| MeHV1      | 116 644–116 667                                         |
| MMRV       | 3464–3517, 130 148–132 723                               |
| MuHV4      | 96 755–105 094                                          |
| P7HV1      | 128 677–131 155, 151 017–153 495                         |
| RCMV       | 74 134–76 485, 118 126–118 854                           |
| SHV1       | 36 683–41 606                                           |
| THV        | 10 089–11 213                                           |

The compound Poisson distribution is motivated from a marked Poisson process point of view. The occurrence of a palindrome of length $2L$ and above is modeled by a Poisson process (4), and the actual length of this palindrome is modeled by a geometric distribution.

On closer examination of the known replication origins in this set of genome sequences, we notice that some of the origins missed by this prediction algorithm are actually rather long approximate palindromes. They are missed because we choose to consider only the perfect palindromes. For example, in HSV2, allowing just one error would have let us pick up a 136 base long approximate palindrome centered at 62 930, which is where the reported replication origin is located. If we include these approximate palindromes in our consideration, the sensitivity can be further increased.

**CONCLUDING REMARKS**

It is mentioned in the introduction that palindromes are merely one type of sequence features known to be associated with replication origins. Other frequently observed characteristics around replication origins include clustering of closely spaced direct and inverted repeats, as well as high AT content. We have actually examined each of these other types of sequence features and found that none of them, when used alone on our data set, reaches the same level of prediction accuracy offered by the BWS. However, it is likely that the prediction accuracy can be further improved by appropriately incorporating them in the prediction scheme. In fact, several replication origins in BoHV4, EHV4 and HSV2 which are not identified by any of PCS, PLS or BWS can be easily detected by the high local AT content around them. Exactly in what way all the different

**Figure 3.** Sensitivity and positive predictive values of the PLS and BWS. In our context, sensitivity is the percentage of known origins that are close to the regions suggested by the prediction; and positive predictive value is the percentage of identified regions that are close to the known origins. The sensitivity and positive predictive values of the PCS are 15 and 25, respectively.

**Figure 4.** Sensitivity and positive predictive values of local BWS.
sequence features should be combined to produce the optimal prediction results is the subject of an ongoing investigation.

While it is encouraging to see that close to 80% of replication origins can be predicted using a palindrome-based scoring scheme like BWS, we have also noted that the positive predictive value is rather low whenever the corresponding sensitivity exceeds 50%. This means that a substantial percentage of the high-scoring windows do not correspond to confirmed replication origins. On closer examination of these high scoring windows which are not replication origins, some of them turn out to be regulatory sequences such as transcription factor binding sites. So far, we have not made use of palindromes to predict regulatory sites, but this would be an important area to explore.

Our prediction scheme is geared towards herpesviruses and still needs to be tested on other DNA viruses. There are a few other methods proposed for prediction of replication origins for bacterial, archaean and yeast genomes (20–23). These methods, which are based on DNA asymmetry, flanking sequence similarity, z-curves, might be adapted to work on viral DNA as well.

Finally, we note that these endeavors to accurately predict replication origins has motivated several interesting and challenging mathematical problems about random letter sequences and probability distributions of patterns on them. We are now dealing with palindromes only but there will be a stream of similar problems about direct and inverted repeats that calls for efforts from the mathematical scientists.

ACKNOWLEDGEMENTS

We would like to thank the editor and two anonymous reviewers for helpful comments and suggestions. Kwok Pui Choi was supported by BMRC grant BMRC01/1/21/19/140 and National University of Singapore ARF Research grant R-146-000-068-112; and Ming-Ying Leung by NIH grants 5S06-GM08012-34 and RCMI 2G13-RR008124. Funding to pay the Open Access publication charges for this article was provided by NIH grant 5S06-GM08012-34.

Conflict of interest statement. None declared.

Table 4. Prediction performance of various scoring schemes, PLS and BWS, based on top 3 scoring windows

| Virus   | Known ORIs/Names | PCS    | PLS    | BWS1   |
|---------|------------------|--------|--------|--------|
| BoHV1   | 111 080–111 300 (oriS) | 1.75 μm | 1.63 μm | 1.63 μm |
| BoHV4   | 126 918–127 138 (oriS) | 1.61 μm | 1.87 μm | 1.87 μm |
| BoHV5   | 97 143–98 850 (oriLyt) | —      | —      | —      |
| CeHV1   | 113 206–113 418 (oriLyt) | —      | —      | 0.06 μm |
| CeHV2   | 129 595–129 807 (oriLyt) | —      | —      | 0.07 μm |
| CeHV5   | 61 592–61 789 (oriL) | —      | 0.057 μm | 0.057 μm |
| CeHV7   | 61 795–61 992 (oriL) | —      | 0.18 μm | 0.18 μm |
| EBV     | 132 795–132 796 (oriS1) | —      | 0.13 μm | 0.13 μm |
| EHV1    | 132 998–132 999 (oriS2) | —      | 0.0016 μm | 0.0016 μm |
| EHV4    | 149 425–149 426 (oriS2) | —      | 0.016 μm | 0.016 μm |
| GaHV1   | 149 628–149 629 (oriS1) | —      | 0.11 μm | 0.11 μm |
| HCMV    | 61 445–61 542 (oriL) | —      | 0.07 μm | 0.07 μm |
| HHV6    | 129 452–129 623 (oriS) | —      | 0.02 μm | 0.02 μm |
| HHV6b   | 144 386–144 557 (oriS) | —      | 0.17 μm | 0.17 μm |
| HSV1    | 7315–9312 (oriP) | —      | contains ori | 0.41 μm |
| HSV2    | 52 589–53 581 (oriLyt) | —      | contains ori | 0.067 μm |
| HSV3    | 126 187–126 338 | —      | —      | —      |
| HSV4    | 73 900–73 919 (oriL) | —      | —      | —      |
| HSV5    | 119 462–119 481 (oriS) | —      | —      | —      |
| HSV6    | 138 568–138 587 (oriS) | —      | —      | —      |
| HSV7    | 24 738–25 005 (oriL) | —      | —      | —      |
| HCMV    | 93 201–94 646 (oriLyt) | —      | contains ori | 0.055 μm |
| HHV6b   | 67 617–67 993 (oriLyt) | —      | —      | —      |
| HHV6    | 68 740–69 581 (oriLyt) | —      | 0.024 μm | —      |
| HHV7    | 66 685–67 298 | —      | —      | —      |
| HSV1    | 62 475 (oriL) | —      | 0.11 μm | 0.11 μm |
| HSV2    | 131 999 (oriS) | —      | 1.41 μm | 1.41 μm |
| HSV3    | 146 235 (oriS) | —      | 1.42 μm | 1.42 μm |
| HSV4    | 62 930 (oriL) | —      | —      | —      |
| HSV5    | 132 760 (oriS) | —      | —      | —      |
| HSV6    | 148 981 (oriS) | —      | —      | —      |
| HSV7    | 75 666–78 970 (oriLyt) | —      | overlaps ori | 0.62 μm |
| SHV1    | 63 848–63 908 (oriL) | —      | —      | —      |
| SHV2    | 114 393–115 009 (oriS) | —      | —      | —      |
| SHV3    | 129 593–130 209 (oriS) | —      | —      | —      |
| VZV     | 75 666–78 970 (oriLyt) | —      | overlaps ori | 0.62 μm |
| VZV     | 110 087–110 350 | —      | 0.094 μm | 0.094 μm |
| VZV     | 119 547–119 810 | —      | 0.22 μm | 0.22 μm |

The table shows the distance between each known origin from the nearest significant palindrome cluster for PCS, or the nearest high scoring window for PLS and BWS, if the center of the cluster or window is within 2 μm of the origin. For example, one of the top 3 scoring windows under the PLS (and BWS) for RCMV is 0.62 map unit away from the RCMV oriLyt.
REFERENCES

1. Weller, S.K., Spadaro, A., Schaffer, J.E., Murray, A.W., Maxam, A.M. and Schaffer, P.A. (1985) Cloning, sequencing, and functional analysis of oriL, a herpes simplex virus type 1 origin of DNA synthesis. Mol. Cell. Biol., 5, 930–942.

2. Reisman, D., Yates, J. and Sugden, B. (1985) A putative origin of Replication of plasmids derived from Epstein–Barr virus is composed of two cis-acting components. Mol. Cell. Biol., 5, 1822–1832.

3. Masse, M.J., Karlin, S., Schachtel, G.A. and Mocarski, E.S. (1992) Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. Proc. Natl Acad. Sci. USA, 89, 5246–5250.

4. Leung, M.Y., Choi, K.P., Xia, A. and Chen, L.H.Y. (2005) Nonrandom clusters of palindromes in herpesvirus genomes. J. Computat. Biol., 12, 331–354.

5. Lin, C.L., Li, H., Wang, Y., Zhu, F.X., Kudchodkar, S. and Yuan, Y. (2003) Kaposi’s sarcoma-associated Herpesvirus lytic origin (oriLyt)-dependent DNA replication: identification of the ori-Lyt and association of K8 bZip protein with the origin. J. Virol., 77, 5578–5588.

6. Delecluse, H.J. and Hammerschmidt, W. (2000) The genetic approach to the Epstein–Barr virus: from basic virology to gene therapy. J. Clin. Pathol. Mol. Pathol., 53, 270–279.

7. Hartline, C.B., Harden, E.A., Williams-Aziz, S.L., Kushner, N.L., Brideau, R.J. and Kern, E.R. (2005) Inhibition of herpesvirus replication by a series of 4-oxo-dihydroquinolines with viral polymerase activity. Antiviral Res., 65, 97–105.

8. Villarreal, E.C. (2003) Current and potential therapies for the treatment of herpesvirus infections. Prog. Drug Res., 60, 263–307.

9. Zha, Y., Huang, L. and Anders, D.G. (1998) Human cytomegalovirus oriLyt sequence requirements. J. Virol., 72, 4989–4996.

10. Newton, C.S. and Theis, J.F. (2002) DNA replication joins the revolution: whole genome views of DNA replication in budding yeast. BioEssays, 24, 300–304.

11. Deng, H., Chu, J.T., Park, N. and Sun, R. (2004) Identification of cis sequences required for lytic DNA replication and packaging of murine gammaherpesvirus 68. J. Virol., 78, 9123–9131.

12. Kornberg, A. and Baker, T.A. (1992) DNA Replication, 2nd edn. W. Freeman, New York.

13. Creighton, T.E. (1993) Proteins. W.H. Freeman, New York.

14. Leung, M.Y., Schachtel, G.A. and Yu, H.S. (1994) Scan statistics and DNA sequence analysis: the search for an origin of replication in a virus. Nonlinear World, 1, 445–471.

15. Glaz, J. (1989) Approximations and bounds for the distribution of the scan statistics. J. Am. Statist. Assoc., 84, 560–566.

16. Dembo, A. and Karlin, S. (1992) Poisson approximations for r-scan processes. Ann. Appl. Probab., 2, 329–357.

17. Leung, M.Y. and Yamashita, T.E. (1999) Applications of the scan statistic in DNA sequence analysis. In Glaz, J. and Balakrishnan, N. (eds), Scan Statistics and Applications. Birkhauser Publishers, Boston, pp. 269–286.

18. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genetics, 16, 276–277.

19. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.

20. Breier, A.M., Chatterji, S. and Cozzarelli, N.R. (2004) Prediction of Saccharomyces cerevisiae replication origins. Genome Biol., 5, R22.

21. Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R. and Tomb, J.F. (1998) Skewed oligomers and origins of replication. Gene, 217, 57–67.

22. Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M.R. and Cebrat, S. (2004) Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res., 16, 3781–3791.

23. Zhang, R. and Zhang, C.T. (2004) Identification of replication origins in archael genomes based on the Z-curve method. Archaea, 1.