Research article

AT excursion: a new approach to predict replication origins in viral genomes by locating AT-rich regions

David SH Chew*1, Ming-Ying Leung2 and Kwok Pui Choi1,3

Address: 1Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore, 2Department of Mathematical Sciences and Bioinformatics Program, The University of Texas at El Paso, TX 79968, USA and 3Department of Mathematics, National University of Singapore, Singapore 117543, Singapore

Email: David SH Chew* - david.chew@nus.edu.sg; Ming-Ying Leung - mleung@utep.edu; Kwok Pui Choi - stackp@nus.edu.sg

* Corresponding author

Abstract

Background: Replication origins are considered important sites for understanding the molecular mechanisms involved in DNA replication. Many computational methods have been developed for predicting their locations in archaeal, bacterial and eukaryotic genomes. However, a prediction method designed for a particular kind of genomes might not work well for another. In this paper, we propose the AT excursion method, which is a score-based approach, to quantify local AT abundance in genomic sequences and use the identified high scoring segments for predicting replication origins. This method has the advantages of requiring no preset window size and having rigorous criteria to evaluate statistical significance of high scoring segments.

Results: We have evaluated the AT excursion method by checking its predictions against known replication origins in herpesviruses and comparing its performance with an existing base weighted score method (BWS1). Out of 43 known origins, 39 are predicted by either one or the other method and 26 origins are predicted by both. The excursion method identifies six origins not predicted by BWS1, showing that the AT excursion method is a valuable complement to BWS1. We have also applied the AT excursion method to two other families of double stranded DNA viruses, the poxviruses and iridoviruses, of which very few replication origins are documented in the public domain. The prediction results are made available as supplementary materials at [1]. Preliminary investigation shows that the proposed method works well on some larger genomes too.

Conclusion: The AT excursion method will be a useful computational tool for identifying replication origins in a variety of genomic sequences.

Background

Recent advances in biotechnology have rendered sequencing a complete genome routine. With the increasing availability of DNA sequences, computational methods to predict likely locations of important functional sites before experimental search are highly valuable because the computational predictions can often help design finely tuned experiments to find these functional sites in shorter time with less labor and fewer resources. Replication origins, which are places on the DNA molecules where replication processes are initiated, are considered important sites for understanding the molecular mecha-
nisms involved in DNA replication. For some viruses with double stranded DNA (dsDNA) genomes in particular, detailed knowledge of their replication processes have had significant impact in developing effective strategies to control the growth and spread of viruses (see, for example, [2]).

A number of computational methods have been developed for predicting replication origins in bacterial, archaeal, and eukaryotic genomes. All these algorithms exploit certain characteristic sequence features found around the replication origins. For example, Lobry [3] employs the GC skew plot to predict replication origins and terminus in bacterial genomes. The skew \((G-C)/(G+C)\), where \(G\) and \(C\) respectively stand for the percentages of guanine and cytosine bases in a sliding window, switches polarity in the vicinity of the replication origin and terminus, with the leading strand manifesting a positive skew. Salzberg et al. [4] predict the replication origins for a number of bacterial and archaeal genomes by identifying some 7-mers and/or 8-mers whose orientation is preferentially skewed around the replication origins. Zhang and Zhang [5] use the Z-curve method successfully to identify several replication origins in bacterial and archaeal genomes. The Z-curve of any given DNA sequence is a three-dimensional curve which uniquely represents the sequence so that unusual sequence compositional features, such as those around a replication origin, can sometimes be visually recognized. Mackiewicz et al. [6] propose three methods, based on DNA asymmetry, the distribution of DnaA boxes and dnaA gene location, were applied to identify the putative replication origins in 112 bacterial chromosomes. They find that DNA asymmetry is the most universal method of putative oriC identification and better prediction can be achieved when the method is applied together with others.

For eukaryotic DNA, Breier et al. [7] develop the Oriscan algorithm to predict replication origins in the S. cerevisiae genome by searching for sequences similar to a training set of 26 known yeast origins pinpointed by site-directed mutagenesis. Oriscan uses both the origin recognition complex binding site and its flanking regions to identify candidates, and then ranks potential origins by their likelihood of activity. More recently, wavelet based multi-scale analysis of DNA strand asymmetries have also been developed [8,9] for detecting mammalian DNA replication origins.

It is important to note that a prediction method designed for one kind of genomes may not necessarily work well on others because the differences in DNA replication mechanisms in different organisms naturally lead to differences in sequence features around their replication origins. One would not expect that the prediction methods designed for bacterial, archaeal, and eukaryotic genomes can be applied directly to viral genomes and produce accurate results. Indeed, when we attempted to use the above algorithms on some herpesviruses genomes with known replication origins like those listed in Table 5 of [10], a variety of difficulties were encountered. For instance, no clear cut switches of polarity were observed in the GC skew plot. No definitive peaks can be visually identified from the Z-curves as potential replication origins of the viruses. When we mined for DnaA boxes [6] in the herpesviruses, just one cluster of DnaA boxes was observed, but it is not near to any known replication origins. Information about origin recognition complex binding sites for herpesvirus genomes, needed for applying Oriscan, are not readily available. While the method based on oligomers skew [4] is designed to work for genomes with single replication origins, the herpesviruses and many other dsDNA viruses contain multiple replication origins in their genomes.

Computational prediction of replication origins, based on the observation of a high concentration of palindromes around the origins, for dsDNA viral genomes was first attempted by Masse et al. [11] on the human cytomegalovirus. Leung et al. [10] formalize the procedure by laying down the mathematical foundation to justify the use of scan statistics for identifying statistically significant palindrome clusters. The location of such palindrome clusters are then taken to be the likely locations of replication origins in herpesviruses. Viewing the scan statistics approach as equivalent to counting the palindromes in sliding windows, Chew et al. [12] offer two more refined schemes of quantifying palindrome concentration to improve the sensitivity of the prediction. One of these schemes, namely the base weighted scheme \((BWS)\), which scores each palindrome according to how rarely it is expected to occur in a nucleotide sequence generated randomly as a first order Markov chain, is found to be the most sensitive for the herpesviruses.

Because of the lack of strong family-wide sequence similarities around the origins, the above prediction methods designed for relatively large and complex dsDNA viruses like the herpesviruses with over 100,000 base pairs in the genomes are based on various sequence statistics rather than the actual nucleotide sequences around replication origins.

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. It has been documented in various studies (e.g. [11,13,14]) that the nucleotide sequences around the replication origins are specific to the individual viruses. Yet the presence of clus-
The AT excursion method can be assessed. Our results demonstrate that the AT excursion method not only can compare with but can also complement the BWS, predictions very well. Having established that AT excursion method is a credible prediction tool, our third objective is to use it for predicting likely replication origin locations for two other families of dsDNA viruses, namely the poxviruses and iridoviruses of which very few replication origins are documented in the public domain. To demonstrate the generality of the AT excursion approach, we also apply it to several larger genomes.

Methods

We adopt the score-based excursion approach [19] to identify segments of a genome having high AT concentration. This, in turn, forms the basis of our proposed 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 March 2006. For each virus, we list its abbreviation, accession number, sequence length, and AT percentages.

Score-based sequence analysis

Score-based sequence analysis is a powerful and yet flexible tool to identify segments of a biological (DNA, RNA or amino acids) sequence containing high concentration of residues of interest according to the users' objectives. One assigns high positive scores to residues of interest, high negative scores to contrasting residues and low or zero scores for the rest. Using various score schemes, Karlin and his collaborators applied this approach with success to gene finding, identification of transmembrane protein segments, and DNA-binding domains. For details and other applications, see, for example, [20-22] and the references therein.

Our interest in this paper is to identify segments of genomic sequences with high AT content. Towards this end, we label bases C or G as "strongly bonding" base S; bases A or T as "weakly bonding" base W. Under this label, S bases (i.e., C or G) are given a score of s and W bases (i.e., A or T) a score of w. The scores s and w will be specified below. We next model the genomic sequence as a realization of a sequence of independent and identically distributed random variables, X_1, X_2, ..., X_n (where n is the genome length), taking values in {s, w}. If the ith base is labeled as W, X_i is given a score w otherwise X_i = s. We let p := P(X_i = s) and P(X_i = w) = 1 - p (denoted by q). The parameter p is naturally estimated by the CG percentage in the genome. An additional constraint needed to be imposed on the choice of s and w is that the expected score per base \( \mu = ps + qw \) has to be negative. This condition prevents favoring long segments to be high scoring segments. A moment's reflection shows that we can always standardize one of the scores to be 1. Here we let w = 1 and choose s to be a negative integer (integer-value choice due to a technical reason as pointed out after equation (3)) so that the expected score per base, \( \mu = ps + qw \) is close to the value

Lin et al [15] have observed that in some herpesvirus genomes, the nucleotide sequences around replication origins are richer in A and T bases. This is not surprising because DNA replication typically requires the binding of an assembly of enzymes (e.g., helicases) to locally unwind the DNA helical structure, and pull apart the two complementary strands (see Chapter 1 in [16,17]). Higher AT content around the origins makes the two complementary DNA strands bond less strongly to each other. This facilitates the two strands to be pulled apart and initiate the replication process. Indeed, Segurado et al. [18] have used a sliding window approach to find "islands" within the *Schizosaccharomyces pombe* genome that have high AT content. They measure base composition using sliding windows of different sizes and find that AT content of windows in regions containing replication origins is significantly higher than those that do not.

Chew et al. [12] have also reported using sliding windows of AT percentages on herpesviruses. Using windows with top AT percentages they are able to predict 65% of replication origins in their dataset. Moreover, this method has successfully identified four origins not predicted by BWS1, suggesting that the AT percentages may be a useful parameter to evaluate statistical significance of high scoring segments [20-22].

There are three main objectives in this paper. First, we shall develop the AT excursion method as a possible alternative to existing approaches for replication origin prediction in DNA sequences. Second, we shall assess the performance of AT excursion in comparison with the prediction results of BWS1 on a data set of currently known performance of AT excursion in comparison with the pre-

Diction results of BWS1 on a data set of currently known performance of AT excursion in comparison with the pre-

- terms of direct or inverted repetitive sequences, including palindromes, is quite common in both types of origins in many members of the herpesvirus family (see [12] and references therein).

- Lin et al [15] have observed that in some herpesvirus genomes, the nucleotide sequences around replication origins are richer in A and T bases. This is not surprising because DNA replication typically requires the binding of an assembly of enzymes (e.g., helicases) to locally unwind the DNA helical structure, and pull apart the two complementary strands (see Chapter 1 in [16,17]). Higher AT content around the origins makes the two complementary DNA strands bond less strongly to each other. This facilitates the two strands to be pulled apart and initiate the replication process. Indeed, Segurado et al. [18] have used a sliding window approach to find "islands" within the *Schizosaccharomyces pombe* genome that have high AT content. They measure base composition using sliding windows of different sizes and find that AT content of windows in regions containing replication origins is significantly higher than those that do not.

- Chew et al. [12] have also reported using sliding windows of AT percentages on herpesviruses. Using windows with top AT percentages they are able to predict 65% of replication origins in their dataset. Moreover, this method has successfully identified four origins not predicted by BWS1, suggesting that the AT percentages may be a useful parameter to evaluate statistical significance of high scoring segments [20-22].
Table 1: The list of herpesviruses to be analyzed.

| Virus                                      | Abbrev. | Accession | Length | AT%  |
|--------------------------------------------|---------|-----------|--------|------|
| Alcelaphine herpesvirus 1                  | alhv1   | NC_002531 | 130608 | 53   |
| Ateline herpesvirus 3                      | athv3   | NC_001987 | 108409 | 63   |
| Bovine herpesvirus 1                       | bohv1   | NC_001847 | 135301 | 28   |
| Bovine herpesvirus 4                       | bohv4   | NC_007665 | 108873 | 59   |
| Bovine herpesvirus 5                       | bohv5   | NC_005261 | 138390 | 25   |
| Callitrichine herpesvirus 3                | calhv3  | NC_004367 | 149696 | 51   |
| Cercopithecine herpesvirus 1               | cehv1   | NC_004812 | 156789 | 26   |
| Cercopithecine herpesvirus 2               | cehv2   | NC_006560 | 150715 | 24   |
| Cercopithecine herpesvirus 8               | cehv8   | NC_006150 | 221454 | 51   |
| Cercopithecine herpesvirus 9               | cehv7   | NC_002686 | 124138 | 59   |
| Cercopithecine herpesvirus 15              | cehv15  | NC_006146 | 171096 | 38   |
| Cercopithecine herpesvirus 16              | cehv16  | NC_007653 | 156487 | 24   |
| Cercopithecine herpesvirus 17              | mmrv    | NC_003401 | 133719 | 47   |
| Equid herpesvirus 1                        | ehv1    | NC_001491 | 150224 | 44   |
| Equid herpesvirus 2                        | ehv2    | NC_001650 | 184427 | 43   |
| Equid herpesvirus 4                        | ehv4    | NC_001844 | 145597 | 50   |
| Gallid herpesvirus 1                       | gahv1   | NC_006623 | 148687 | 52   |
| Gallid herpesvirus 2                       | gahv2   | NC_002229 | 174077 | 56   |
| Gallid herpesvirus 3                       | gahv3   | NC_002577 | 164270 | 46   |
| Human herpesvirus 1                        | hsv1    | NC_001806 | 152261 | 32   |
| Human herpesvirus 2                        | hsv2    | NC_001798 | 154746 | 30   |
| Human herpesvirus 3                        | vzv     | NC_001348 | 124884 | 54   |
| Human herpesvirus 4                        | ebv     | NC_007605 | 171823 | 41   |
| Human herpesvirus 5                        | hcmv    | NC_001347 | 230287 | 43   |
| (AD169)                                    |         |           |        |      |
| Human herpesvirus 5                        | hcmv-m  | NC_006273 | 235645 | 42   |
| (Merlin)                                   |         |           |        |      |
| Human herpesvirus 6                        | hhv6    | NC_001664 | 159321 | 58   |
| Human herpesvirus 68                       | hhv6b   | NC_000899 | 162114 | 58   |
| Human herpesvirus 7                        | hhv7    | NC_001716 | 153080 | 63   |
| Human herpesvirus 8                        | hhv8    | NC_003409 | 137508 | 47   |
| Ictalurid herpesvirus 1                    | ichtv   | NC_001493 | 134226 | 43   |
| Meleagrid herpesvirus 1                    | mehv1   | NC_002641 | 159160 | 52   |
| Murid herpesvirus 1                        | mcnv    | NC_004065 | 230278 | 41   |
| Murid herpesvirus 2                        | rcnv    | NC_002512 | 230138 | 39   |
| Murid herpesvirus 4                        | muhv4   | NC_001826 | 119450 | 53   |
| Macaca fuscata                             | mfrv    | NC_007016 | 131217 | 48   |
| rhadinovirus                               |         |           |        |      |
| Ostreid herpesvirus 1                      | oshv1   | NC_005881 | 207439 | 61   |
| Ovine herpesvirus 2                        | ohv2    | NC_007646 | 135135 | 47   |
| Pongine herpesvirus 4                      | ccmv    | NC_003521 | 241087 | 38   |
| Psittacid herpesvirus 1                    | pshv1   | NC_005264 | 163025 | 39   |
| Saimiriine herpesvirus 2                   | sahv2   | NC_001350 | 112930 | 65   |
| Suid herpesvirus 1                         | shv1    | NC_006151 | 143461 | 26   |
| Tupaiid herpesvirus 1                      | thv     | NC_002794 | 195859 | 34   |
of -0.5 (where we adopt Karlin's choice of expected value as in [21]). In other words, \( w := 1 \) and

\[
x := \left\lfloor \frac{\mu - qw}{p} \right\rfloor,
\]

where \( \mu = -0.5 \) and \( \lfloor \cdot \rfloor \) denotes the integer floor function.

**Excursions and their values**

We next compute the cumulative scores and seek to identify segments of the genome that have significantly high scores. As we are only interested in segments with positive additive scores, we reset our cumulative scores to zero whenever it becomes negative.

The excursion scores \( E_i \)'s are defined recursively as

\[
E_0 = 0, \quad E_i = \max\{E_{i+1} + X_i, 0\}, \quad \text{for} \quad 1 \leq i \leq n.
\]

Using this recursive definition, we are able to construct "excursions" for each of the genomes. An excursion starts at a point \( i \) where \( E_i \) is zero and ends at \( j > i \) where \( E_j \) is the very next zero. The score then stays at zero until it first becomes positive again for the start of the next excursion. The value of an excursion is defined to be the peak score during the course of that particular excursion.

**Distribution of the Maximal Aggregate Score**

For each value of \( x \), the maximal aggregate score

\[
M_n = \max_{1 \leq k \leq n} E_k
\]

satisfies

\[
P\left( M_n > \frac{\ln n}{\lambda^*} + x \right) = 1 - \exp\{-K^* e^{-\lambda^* x}\},
\]

where \( \lambda^* \) is the unique positive solution to the equation

\[
E(e^{\lambda X_1}) = \frac{pe^{\delta} + qe^{iw}}{\lambda^*} = 1 \quad \text{and} \quad K^* \text{ is a parameter given by an explicit series expansion (See [23]).}
\]

When \( X \) is an integer-valued variable of span \( \delta \), we have a simpler expression for \( K^* \) ([23]):

\[
\exp\{-K^*_+ e^{-\lambda^* x}\} \leq \liminf_{n \to \infty} P\left( M_n - \frac{\ln n}{\lambda^*} < x \right) \leq \limsup_{n \to \infty} P\left( M_n - \frac{\ln n}{\lambda^*} < x \right) \leq \exp\{-K^*_+ e^{-\lambda^* x}\},
\]

where

\[
K^- = \frac{\lambda^* \delta}{e^{\lambda^*} - 1}, \quad K^+ = \frac{\lambda^* \delta + K^*}{1 - e^{\lambda^*} - 1}.
\]

For the simple score scheme with values \{ -m, ..., -1, 0, 1 \} occurring with probabilities \{ \( p_m \), ..., \( p_1 \), \( p_0 \), \( p_1 \) \} we have,

\[
K = (e^{-\lambda^*} - e^{-2\lambda^*}) E(Xe^{\lambda X}).
\]

We can set the left hand side of Equation (2) to some predetermined significance level, say \( P = 0.05 \) or \( 0.01 \), and solve for \( x \). A segment with score exceeding \( M_p = \frac{\ln n}{\lambda^*} + x \) is then said to be significant at the 100P\% level.

In this paper, we use \( K \) in place of \( K^* \) in Equation (2) for a "conservative" estimate of the probability and \( K \) for a "generous" one.

We use Equation (2) with \( P = 0.05 \) and \( P = 0.01 \) to get \( M_{0.05} \) and \( M_{0.01} \) respectively. If the value of an excursion exceeds the critical value \( M_{0.05} \) (or \( M_{0.01} \)), then the segment from the beginning of the excursion up to the base where the peak value is realized is said to be a high-scoring segment (HSS) significant at the 5\% (or 1\%) level.

**HSS Selection**

For each of the genomic sequences listed in Table 1, we obtain a set of HSS, significant at the 5\% (or 1\%) level. In each set of HSS, it is common to find several of them located close to one another. We thus apply a filtering procedure so that, if this happens, we shall only select one of several neighboring excursions as a representative for that part of the genome. In fact, we first sort all the HSS according to their aggregate scores. Starting with the one with the highest value, say segment A, we 'discard' neighboring HSS that are within 2 map units of it. After that, we pick among the rest (not including segment A and the discarded HSS), the HSS with the next highest value, say segment B, and repeat the process. Only the representative segments A, B, and so forth, will be used in replication origin prediction.

**Results and Discussion**

**HSS Tables and Excursion Plots**

Table 2 lists the HSS for each herpesvirus in Table 1. We have also tried locating high-scoring segments by running the excursions from the 3' end to 5' end of the genome. The results obtained are not much different from the "vanilla" version (i.e., from 5' to 3').

For visualizing the locations of the selected HSS relative to the entire genome, the excursion plot is a convenient tool. The excursion plot of the Human Herpesvirus 3 (vzv) is
Table 2: Herpesviruses: HSS at 5% level using the conservative bound.

| Virus | Start | Peak | Value | Virus | Start | Peak | Value |
|-------|-------|------|-------|-------|-------|------|-------|
| alhv1 | 1204  | 1370 | 54    | ebv   | 11854 | 11950| 45    |
|       | 32478 | 32850| 48    |       | 77111 | 77150| 24    |
|       | 113630| 113684| 46    |       | 43158 | 42335| 23    |
|       | 85923 | 85992| 45    | ehv1  | 20348 | 20431| 47    |
|       | 72999 | 73115| 44    |       | 134195 | 134276| 36    |
|       | 125691| 125726| 31    |       | 65055 | 65126| 35    |
| ahv3  | 8827  | 8892 | 40    |       | 99301 | 99374| 34    |
| bohv1 | 100410| 100484| 26    |       | 11034 | 11141| 32    |
|       | 109702| 109730| 25    |       | 105796 | 105862| 30    |
|       | 128487| 128515| 25    |       | 73653 | 73746| 27    |
|       | 16593 | 16626 | 21    |       | 113818 | 113849| 25    |
|       | 113720| 113738| 18    |       | 149310 | 149341| 25    |
|       | 124479| 124497| 18    |       | 110314 | 110352| 23    |
|       | 29    | 45   | 16    |       | 128924 | 128992| 23    |
|       | 58542 | 58569| 15    | ehv2  | 160281 | 160518| 102   |
| bohv4 | 60687 | 60826| 35    |       | 86522 | 86622| 76    |
| bohv5 | 68440 | 68507| 49    |       | 53843 | 54012| 61    |
|       | 113549| 113583| 28    |       | 140661 | 140826| 57    |
|       | 129429| 129463| 28    |       | 4580  | 4655 | 51    |
|       | 592   | 616  | 21    |       | 171454 | 171529| 51    |
|       | 86191 | 86215| 21    |       | 95342 | 95440| 50    |
|       | 102074| 102106| 17    |       | 10772 | 10820| 48    |
|       | 92511 | 92535| 15    |       | 39983 | 39977| 48    |
|       | 120935| 120959| 15    |       | 177646 | 177694| 48    |
|       | 59921 | 59938| 14    |       | 113310 | 113399| 47    |
|       | 17408 | 17433| 13    |       | 134709 | 134772| 45    |
|       | 41883 | 41899| 13    |       | 166114 | 166207| 42    |
|       | 147310| 147384| 31    |       | 45831 | 45965| 41    |
|       | 50872 | 50973| 50    |       | 15443 | 15482| 39    |
|       | 158344| 158701| 45    |       | 19722 | 19845| 39    |
|       | 95375 | 95603| 39    |       | 182317 | 182356| 39    |
|       | 3519  | 3602 | 35    |       | 153977 | 154145| 36    |
|       | 24084 | 24156| 33    |       | 123321 | 123362| 35    |
|       | 182982| 183136| 31    |       | 147222 | 147341| 35    |
|       | 14314 | 14370| 23    |       | 34816 | 34884| 29    |
|       | 177170| 177247| 23    |       | 76380 | 76454| 29    |
|       | 180941| 180975| 22    |       | 103167 | 103223| 29    |
|       | 147310| 147384| 20    |       | 64344 | 64402| 25    |
|       | 116723| 116836| 53    |       | 786   | 831 | 24    |
|       | 92092 | 92118| 26    | ehv4  | 109852 | 110086| 60    |
|       | 61680 | 61700| 20    |       | 19878 | 19943| 50    |
|       | 132785| 132805| 20    |       | 132383 | 132462| 49    |
|       | 149415| 149435| 20    |       | 105284 | 105365| 48    |
|       | 52055 | 52075| 17    |       | 23895 | 24016| 43    |
|       | 42984 | 43006| 16    |       | 3984  | 4110  | 42    |
|       | 11389 | 11407| 15    |       | 73340 | 73509| 37    |
|       | 24415 | 24441| 14    |       | 98849 | 98930| 33    |
|       | 11965 | 12011| 28    |       | 46612 | 46674| 32    |
|       | 114927| 114988| 19    |       | 10630 | 10697| 31    |
|       | 92913 | 92940| 23    |       | 58833 | 58906| 31    |
|       | 62970 | 62991| 21    |       | 82616 | 82701| 31    |
|       | 133468| 133489| 21    |       | 127230 | 127351| 31    |
|       | 149813| 149834| 21    |       | 112929 | 112967| 29    |
|       | 8303  | 8331 | 20    |       | 145082 | 145120| 29    |
|       | 118685| 118713| 20    | gahv1 | 24852 | 24890| 30    |
|       | 53056 | 53100| 18    | gahv2 | 106724 | 106811| 35    |
|       | 25423 | 25473| 16    | gahv3 | 11168 | 11198| 27    |
|       | 1717  | 1736 | 15    |       | 122384 | 122414| 27    |
Table 2: Herpesviruses: HSS at 5% level using the conservative bound. (Continued)

|                | HSS                       | p-value     |
|----------------|---------------------------|-------------|
| cehv2          | 7681                      | 0.028       |
|                | 7738                      | 0.031       |
| hcmv           | 3402                      | 0.034       |
|                | 3542                      | 0.037       |
| cehv7          | 8617                      | 0.038       |
|                | 86296                     | 0.035       |
| cehv8          | 149643                    | 0.039       |
|                | 149720                    | 0.036       |
| hcmv-m         | 3798                      | 0.042       |
|                | 3939                      | 0.045       |
| mfrv           | 128046                    | 0.046       |
|                | 128640                    | 0.042       |
| hhv6           | 130410                    | 0.043       |
|                | 130501                    | 0.040       |
| mfrv           | 123938                    | 0.044       |
|                | 124238                    | 0.047       |
| hhv6b          | 139482                    | 0.045       |
|                | 139569                    | 0.043       |
| ohv2           | 115365                    | 0.047       |
|                | 115545                    | 0.049       |
| hhv8           | 136287                    | 0.048       |
|                | 136704                    | 0.046       |
| mhhv4          | 6000                      | 0.049       |
|                | 6037                      | 0.047       |
| hshv1          | 24625                     | 0.051       |
|                | 24685                     | 0.054       |
| mhhv4          | 73292                     | 0.052       |
|                | 73460                     | 0.051       |
| hshv2          | 5584                      | 0.053       |
|                | 5628                      | 0.055       |
| mhhv4          | 161395                    | 0.057       |
|                | 161505                    | 0.056       |
Table 2: Herpesviruses: HSS at 5% level using the conservative bound. (Continued)

| Entry | Value | HSS | Replication Origin |
|-------|-------|-----|--------------------|
| ichv1 | 81195 | 81220 | 16 |
|       | 99337 | 99370 | 15 |
|       | 6068  | 6290  | 81 |
|       | 121738| 121960| 81 |
|       | 104134| 104399| 70 |
|       | 17065 | 17333 | 58 |
|       | 132735| 133003| 58 |
| rcmv  | 150923| 151612| 92 |
| 451   | 726   | 50   |
| 116121| 116396| 50   |
| 60752 | 60845 | 30   |
| 42919 | 43007 | 28   |
| 20109 | 20187 | 24   |
| 10016 | 10063 | 23   |
| 125686| 125733| 23   |
| mehv1 | NIL   |  |
| 155163| 156341| 125 |
|       | 161228| 161391| 40 |
|       | 115543| 115640| 37 |
|       | 102865| 102960| 35 |
|       | 79497 | 79573 | 34 |
|       | 15628 | 15724 | 33 |
|       | 144170| 144290| 33 |
|       | 73525 | 73579 | 27 |
|       | 39209 | 39248 | 24 |
|       | 92997 | 93026 | 24 |
|       | 219239| 219282| 22 |

Entries in italics are significant at 1% too.

presented in Figure 1, where the AT excursion values are plotted against the bases along the genome. The general appearance of Figure 1 is typical of the excursion plots for all the herpesviruses analyzed. In the case of vzv, three peaks with excursion values exceeding the 5% significance level are observed. Two of these peaks are close to the centers of the only two known replication origins of vzv (see Table 3).

Prediction Performance
The high-scoring segments are checked against known replication origins in herpesviruses to evaluate their performance as a prediction tool. Table 3 lists all the known replication origins for the herpesviruses in Table 1. These origins are reported either in published literature or GenBank annotations. For each replication origin, we list the HSS (at 5% level) closest to it. For this table we had used the "conservative" estimate for the value of $K^*$ (See Equations (2) and (3)). When the peak of an HSS is less than 2 map units (one map unit is one percent of the genome length) away from the center of a replication origin, we say that our method has correctly predicted that particular
replication origin. From Table 3, we see that of the 43 replication origins known, compiled from literature or annotations, 32 of them are close to HSS that have been identified.

We had also tried using the "generous" estimate for $K^*$ at the 5% and 1% level of significance. Table 4 gives a summary of the performance of our prediction scheme when those bounds were used. The first two columns of the table gives the sensitivity level and positive prediction value of our scheme. Sensitivity refers to the percentage of replication origins predicted by our method, and PPV (positive predictive value) the proportion of HSS that correctly predict replication origins. APD (average predictive distance), given in map units (± one standard deviation), shows the average of the distances (in map units) between the center of each replication origin and the HSS that predicts it. Note that the APD values say that on average, when a prediction by an HSS is successful, the replication origin is about 0.35 map units away from it. We have also done some simple analysis of the location of the center of each replication origin with respect to the HSS closest to it. We count the number of times the center of replication origin falls within the left, right or center of the HSS. The columns %L, %R, and %C in Table 4 give these proportions. Our results show that the origin falls within the center of the HSS half the time.

**Comparison with Other Approaches**

How does the AT excursion method compare with the sliding window approach using palindrome based scoring schemes previously presented in [12]? Since the BWS$_1$ scheme has been shown to perform best among the various palindrome based schemes, we have examined the numbers of replication origins correctly predicted by AT excursion and by BWS$_1$. The results are summarized in Figure 2.

The majority of the 43 known origins in the herpesviruses listed in Table 1 are predicted by both methods and most of the remaining ones are predicted by one method or the other. Only four of the origins fail to be predicted by either method. This suggests that the AT excursion method and the BWS$_1$ scheme complement each other very well.

There are certain advantages in the AT excursion approach over BWS$_1$. First, AT excursion does not require any sequence specific parameters to be prescribed by the user. It is window size free because it does not require any sliding window to measure AT concentration. Moreover, while the palindrome based methods require the specification of a minimal palindrome length before the analysis can be carried out, no such parameter is needed for AT excursion. Second, the AT excursion method is statistically

| Table 3: Prediction results at 5% level using the conservative bound. |
|---|
| **Virus** | **Ori Center** | **Nearest HSS** | **Start** | **Peak** | **Value** | **Prediction** |
| bohv1 | 111190 | 109702 | 109730 | 25 | Yes |
| bohv1 | 127028 | 128487 | 128515 | 25 | Yes |
| bohv4 | 97966.5 | 60687 | 60826 | 35 | No |
| bohv5 | 113312 | 113549 | 113583 | 28 | Yes |
| bohv5 | 129701 | 129429 | 129463 | 28 | Yes |
| cehv1 | 61690.5 | 61680 | 61700 | 20 | Yes |
| cehv1 | 61893.5 | 61680 | 61700 | 20 | Yes |
| cehv1 | 132795.5 | 132785 | 132805 | 20 | Yes |
| cehv1 | 132998.5 | 132785 | 132805 | 20 | Yes |
| cehv1 | 149425.5 | 149415 | 149435 | 20 | Yes |
| cehv1 | 149628.5 | 149415 | 149435 | 20 | Yes |
| cehv16 | 62981 | 62970 | 62991 | 21 | Yes |
| cehv16 | 133479 | 133468 | 133489 | 21 | Yes |
| cehv16 | 149824 | 149813 | 149834 | 20 | Yes |
| cehv2 | 61493.5 | 61483 | 61503 | 20 | Yes |
| cehv2 | 129537.5 | 129527 | 129547 | 20 | Yes |
| cehv2 | 144471.5 | 144461 | 144481 | 20 | Yes |
| cehv7 | 109636.5 | 86167 | 86296 | 37 | No |
| cehv7 | 118622.5 | 86167 | 86296 | 37 | No |
| ebv | 8313.5 | 11854 | 11950 | 45 | No |
| ebv | 40797 | 43158 | 43325 | 23 | Yes |
| ebv | 143825.5 | 77111 | 77150 | 24 | No |
| ehv1 | 126262.5 | 128924 | 128992 | 23 | Yes |
| ehv4 | 73909.5 | 73340 | 73509 | 37 | Yes |
| ehv4 | 119471.5 | 112929 | 112967 | 29 | No |
| ehv4 | 138577.5 | 132383 | 132462 | 49 | No |
| gahv1 | 24871.5 | 24852 | 24890 | 30 | Yes |
| hcmv | 93923.5 | 96685 | 96824 | 34 | Yes |
| hhv6 | 67805 | 130410 | 130501 | 59 | No |
| hhv6b | 69160.5 | 132997 | 133163 | 62 | No |
| hhv7 | 66991.5 | 128589 | 128984 | 70 | No |
| hsv1 | 62475 | 62465 | 62485 | 20 | Yes |
| hsv1 | 131999 | 131990 | 132008 | 18 | Yes |
| hsv1 | 146235 | 144115 | 144142 | 18 | Yes |
| hsv2 | 62930 | 62919 | 62939 | 17 | Yes |
| hsv2 | 132760 | 132691 | 132711 | 17 | Yes |
| hsv2 | 148981 | 146600 | 146631 | 19 | Yes |
| rcnv | 77318 | 24072 | 24108 | 21 | No |
| shv1 | 63878 | 63862 | 63892 | 24 | Yes |
| shv1 | 114701 | 114686 | 114715 | 20 | Yes |
| shv1 | 129901 | 129607 | 129636 | 20 | Yes |
| vzw | 110218.5 | 110195 | 110227 | 32 | Yes |
| vzw | 119678.5 | 119669 | 119701 | 32 | Yes |

For each replication origin, we list the high-scoring segment (at 5% level) closest to it. When the peak of a high-scoring segment is less than 2 map units away from the center of a replication origin, we say that our method has correctly predicted that particular replication origin.
based, as the probabilistic distribution has already been established [20-22]. This allows the statistical significance for HSS be evaluated easily.

We also note that the more elaborate AT excursion approach performs better than the simpler procedure of measuring the percentage of A and T bases on a sliding window in terms of number of correct predictions and the proximity of these predictions to the true origins. Out of the 43 known replication origins for the herpesviruses in Table 1, 32 are correctly predicted by AT excursion but only 28 by AT sliding window plot. Furthermore, the box-plots of the predictive distances (Figure 3) of the AT excursion approach suggests that the predictions given by the AT excursion approach are much closer to known replication origins as compared to those of the AT sliding window plot approach. (In fact, the predictive distances of the AT excursion approach compared to that of the PLS and BWS1 approaches mentioned in [12] are observably shorter. See Figure 3.) This suggests that the excursion values might more correctly capture the essence of A/T abundance variation along genomic sequences.

Table 4: Prediction Performance Summary.

| Significance | Sensitivity | PPV  | APD         | %L  | %R  | %C  |
|--------------|-------------|------|-------------|-----|-----|-----|
| 5% (C)       | 74%         | 22%  | 0.34 ± 0.57 | 16% | 31% | 53% |
| 5% (G)       | 86%         | 17%  | 0.35 ± 0.53 | 24% | 30% | 46% |
| 1% (C)       | 67%         | 25%  | 0.31 ± 0.52 | 14% | 34% | 52% |
| 1% (G)       | 74%         | 18%  | 0.34 ± 0.57 | 16% | 31% | 53% |

(C) indicates that the "Conservative" bound is used while (G) indicates that the "Generous" bound is used. Sensitivity refers to the percentage of replication origins predicted by our method, and PPV (positive predictive value) the proportion of HSS that correctly predict replication origins. APD (average predictive distance), given in map units (± one standard deviation), shows the average of the distances between the center of each replication origin and a HSS that predicts it in map units. %L, %R and %C count the number of times the center of replication origin falls within the left, right or center of the HSS.

Herpesvirus Replication Origins Alignment and Motif Finding

One might ask whether or not the nucleotide sequences around replication origins in various viruses of the same family share sufficient similarities so that the origins can be identified by sequence alignments and motif finding techniques. We therefore extracted the nucleotide sequences of the known herpesvirus origins according to their documented locations for closer examination. These sequences are available as supplementary materials on the companion website. A multiple alignment using CLUSTAL W [24] and motif searches using MEME and MAST [25,26] have been conducted for the herpesvirus origin sequences. No significant sequence similarity or common

Figure 2
Predictions of AT excursion and BWS1. In this figure, the set A consists of origin replications predicted by the AT excursion method and B consists of those predicted by the BWS1 method. $A \cap B = \{\text{cehv7, cehv72, ehv41, hsv2}, \text{hsv22, hsv23}\}, A \cap B = \{\text{cehv16, cehv162, ebv1}, \text{ebv2, hhv6, hhv66, rcmv}\}, (A \int B) = \{\text{bohv4, ehv43, ehv44, hhv7}\}$. The rest of the replication origins (26 of them) are predicted by both methods. Note that for viruses with several known replication origins, such as the hsv2, which has three (see Table 3), we denote the replication origins as hsv21, hsv22, hsv23, etc.

Figure 3
Predictive Distances for PLS, BWS1, AT-swp and AT excursion. These boxplots show the predictive distances for PLS, BWS1, AT-swp and AT excursion.
motif pattern across all the origin sequences has been found, agreeing with the findings of [11,13,14].

What if we first classify these nucleotide sequences according to some classification schemes, will the members within each class share noticeable sequence similarities? We classified the origins according to (i) the sub-family of the virus (herpesviruses are classified into the alpha, beta, and gamma sub-families by their biological properties [27]), (ii) the type of origin (i.e., whether the origin is a oriL, oriLyt or oriS). We ran MEME and MAST separately on the sequences in each sub-family/type of origins to detect common motif patterns. From the outputs under classification (i), we note that the origins from the alpha sub-family can be further divided into two groups. Each group has a common motif pattern across its members. For the beta and gamma sub-families, no distinct patterns can be found. However, the rcmv and ebv origins contain many repeat patterns. For classification (ii), we find that both the oriL. oriLyt origins contain sequence motifs common to a number of their members. No motif was found for oriS sequences. The results of our motif search are made available in the supplementary materials.

Although our investigations are preliminary, the motifs found in these subsets of herpesvirus genomes may suggest new information that can be incorporated into the replication origin prediction procedures.

Other Families of Viruses
Aside from the herpesviruses, we have also applied the AT excursion method to search for HSS in the poxviruses and iridoviruses. These two viral families are chosen because, like the herpesviruses, they are large, complex dsDNA viruses with no RNA stage. Their genome lengths are also similar in magnitude to those of the herpesviruses.

Poxviruses infect a large variety of animal species that gather in swarms and herds (e.g., mosquitoes, cows). Smallpox is a major disease caused by the variola virus, a member of the poxvirus family. Smallpox was eradicated in 1977 by preventive inoculations with cowpox or vaccinia viruses through the dedicated efforts of the World Health Organization and many individuals. In the recent few years, as the threat of the variola virus being used as a biological weapon is raised, there is growing interest in further studying poxviruses for biodefense purposes [28,29]. Iridoviruses are found in a variety of fish, amphibians, and reptiles. Some iridoviruses have been associated with serious diseases (e.g., viral erythrocytic necrosis of salmonids), while others have only been found in apparently healthy animals (e.g., goldfish iridovirus). Iridovirus infection is considered a serious concern in modern aquaculture, fish farming, and wildlife conservation [30].

Amongst these two families, only one genome, namely the Chilo iridescent virus, has documented replication origin locations [31]. Our method has correctly predicted one of these locations. Due to the lack of confirmed origin locations, prediction accuracy cannot be tested on these families. Nevertheless, our predictions may assist researchers to investigate these viruses experimentally to identify and confirm the exact locations of replication origins in their genomes. We have, therefore, made our prediction results available at [1].

AT excursion applied to larger genomes
To gauge whether the AT excursion approach can potentially be generalized to predict replication origins for non-viral genomes, we apply it to several archaeal and bacterial genomes which have been previously analyzed. From [4,5,32] we are able to compile a list of 15 known or suggested replication origins (11 known, 4 suggested). Using the AT excursion method, we manage to correctly predict 9 of the replication origins (6 known, 3 suggested). Although our studies are preliminary, the results show that the AT excursion method can work reasonably well even on larger genomes.

Conclusion
This paper introduces the AT excursion method to quantify local AT abundance in genomic sequences. The simple and intuitive idea of locating regions with high AT content as potential replication origin sites proves to be effective in identifying several replication origins not previously predicted. This shows that the AT excursion approach is a valuable addition to existing prediction tools. However, we have also observed that quite a number of the statistically significant HSS found by AT excursions are not close to replication origins. Whether these HSS correspond to other important functional sites in the genomic sequences remains an interesting question to be investigated.

The availability of statistical significance criteria and the independence of ad hoc parameters like the minimal palindrome length and sliding window size make the AT excursion method particularly easy to apply to those viral genomes where no replication origin information in similar and related genomes is available. On the other hand, if such information is available, the AT excursion method is not capable of taking advantage of it. To address this issue, machine learning approaches (e.g., neural networks and support vector machines), which better allow us to use knowledge in related genomes, are currently being explored. We anticipate that a combination of score based statistics with machine learning approaches will provide a highly accurate prediction tool set for replication origins.
Authors’ contributions
DC participated in the design of the study and performed the data and statistical analysis. KPC and MYL conceived the study, and participated in its design and coordination. All authors contributed to writing, reading and approving the final manuscript.

Acknowledgements
We would like to thank the editor and the anonymous reviewers for helpful comments and suggestions which improve this paper. Kwo Pui Choi was supported by National University of Singapore ARF Research grant R-155-000-051-112, and Ming-Ying Leung by NH grants 506GM08012-35, 5G12RR008124-11 and 3T34GM008048-2051.

References
1. Supplementary Materials [http://www.stat.nus.edu.sg/~stadb/Authors.html]
2. Delecluse HJ, Hammerschmidt W: The genetic approach to the Epstein-Barr virus from basic virology to gene therapy. Mol Pathol 2000, 53(5):270-279.
3. Lobry JR: Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 1996, 13(5):660-665.
4. Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF: Skewed oligomers as markers of DNA replication origins. Gene 1998, 217(1-2):557-67.
5. Zhang R, Zhang CT: Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea 2005, 1(5):335-346.
6. Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S: Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res 2004, 32(13):3781-3791 [http://dx.doi.org/10.1093/nar/gkh699].
7. Brewer AM, Chattarji S, Cozzarelli NR: Prediction of Saccharomyces cerevisiae replication origins. Genome Biol 2004, 5(4):R22 [http://dx.doi.org/10.1001/ [186].jpg.2004.5.4.22].
8. Brodie of Brodie EB, Nicolay S, Touchon M, Audit B, d’Aubenton Carafa Y, Thermes C, Arneodo A: From DNA sequence analysis to modeling replication in the human genome. Phys Rev Lett 2005, 94(24):248103.
9. Touchon M, Nicolay S, Audit B, Brodie of Brodie EB, d’Aubenton Carafa Y, Arneodo A: Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci USA 2005, 102(28):9836-9841 [http://dx.doi.org/10.1073/pnas.0500771102].
10. Leung MY, Choi KP, Xia A, Chen LHY: Nonrandom clusters of palindrome cores in herpesvirus genomes. J Comput Biol 2005, 12(3):331-354 [http://dx.doi.org/10.1089/cmb.2005.12.331].
11. Machaj D, Kornberg EB, Mocarski ES: Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. Proc Natl Acad Sci USA 1992, 89(12):5246-5250.
12. Chew DSH, Choi KP, Leung MY: Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. Nucleic Acids Res 2005, 33(15e134) [http://dx.doi.org/10.1093/nar/gni135].
13. Haenzel PT, Lietman PS, Gibson W, Hayward GS: Identification of the lytic origin of DNA replication in human cytomegalovirus by a novel approach utilizing ganciclovir-induced chain termination. J Virol 1990, 64(12):6184-6195.
14. C. Chan H, Krenitsky DM, Dewhurst S: Stringent structural and sequence requirements of the human herpesvirus 6B lytic-phase origin of DNA replication. J Gen Virol 1997, 78( Pt 5):1125-1129.
15. Lin CL, Li H, Wang Y, Zhu FX, Kudchodkar S, Yuan Y: Kapoor’s sarcoma-associated herpesvirus lytic origin (ori-Lyt)-dependent DNA replication: identification of the ori-Lyt and association of K8 bZIP protein with the origin. J Virol 2003, 77(10):5578-5588.
16. Kornberg A, Baker TA: DNA Replication 2nd edition. New York, New York: WH Freeman and Company; 1992.
17. Braich D, Kornberg A: A model for initiation at origins of DNA replication. Cell 1988, 54(7):915-918.
18. Segurodo M, de Luis A, Antequera F: Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. EMBO J 2003, 22(11):467-4680.
19. Karlin S, Altschul SF: Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990, 87(6):2264-2268.
20. Karlin S, Brendel V: Chance and statistical significance in protein and DNA sequence analysis. Science 1992, 257(5066):39-49.
21. Karlin S: Statistical studies of molecular sequences: score-based methods. Phil Trans R Soc Lond B 1994, 344(1310):391-402.
22. Karlin S: Statistical signals in bioinformatics. Proc Natl Acad Sci USA 2005, 102(38):3355-3362 [http://dx.doi.org/10.1073/pnas.0502804102].
23. Karlin S, Dembo A, Kawabata T: Statistical composition of high-scoring segments from molecular sequences. Ann Statist 1990, 18(2):571-581.
24. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994, 22(22):4673-4680.
25. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology Volume 2. Menlo Park, California: AAAI Press; 1994:28-36.
26. Bailey TL, Gribskov M: Combining evidence using p-values: application to sequence homology searches. Bioinformatics 1998, 14:48-54.
27. Cann AJ: Principles of Molecular Virology 4th edition. San Diego: Elsevier Academic Press; 2005.
28. Henderson DA: The looming threat of bioterrorism. Science 1999, 283(5406):1279-1282.
29. Miller SE: Bioterrorism and electron microscopic differentiation of poxviruses from herpesviruses: dos and don’ts. Ultrastruct Pathol 2003, 27(3):133-140.
30. Tsai CT, Ding JW, Wu MH, Wu MF, Guo JC, Chang CY: Complete genome sequence of the grouper iridovirus and comparison of genomic organization with those of other iridoviruses. J Virol 2005, 79(20):10200-10203 [http://dx.doi.org/10.1128/JVI.79.10.20200-10203.2005].
31. Chandermann M, Schnitzler P, Rosen-Wolf A, Raab K, Sonntag KC, Darai G: Identification and mapping of origins of DNA replication within the DNA sequences of the genome of insect iridovirus type 6. Virus Genes 1992, 4:15-32.
32. Freeman M, Plasterer TN, Smith TF, Mohr SC: Patterns of genome organization in bacteria. Science 1998, 279(5358):1827a [http://www.sciencemag.org].