INFORMS Journal on Computing
Vol. 16, No. 4, Fall 2004, pp. 331–340
ISSN 0899-1499 | EISSN 1526-5528 | 04 | 1604 | 0331

DOI 10.1287/ijoc.1040.0087

Palindromes in SARS and Other Coronaviruses

David S. H. Chew
Department of Mathematics, National University of Singapore, Singapore 117543, Singapore, matchewd@nus.edu.sg

Kwok Pui Choi
Departments of Mathematics, and of Statistics and Applied Probability, National University of Singapore, Singapore 117543, Singapore, matckp@nus.edu.sg

Hans Heidner
Department of Biology, University of Texas at San Antonio, San Antonio, Texas 78249, USA, hheidner@utsa.edu

Ming-Ying Leung
Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas 79968, USA, mleung@utep.edu

With the identification of a novel coronavirus associated with the *severe acute respiratory syndrome* (SARS), computational analysis of its RNA genome sequence is expected to give useful clues to help elucidate the origin, evolution, and pathogenicity of the virus. In this paper, we study the collective counts of palindromes in the SARS genome along with all the completely sequenced coronaviruses. Based on a Markov-chain model for the genome sequence, the mean and standard deviation for the number of palindromes at or above a given length are derived. These theoretical results are complemented by extensive simulations to provide empirical estimates. Using a z score obtained from these mathematical and empirical means and standard deviations, we have observed that palindromes of length four are significantly underrepresented in all the coronaviruses in our data set. In contrast, length-six palindromes are significantly underrepresented only in the SARS coronavirus. Two other features are unique to the SARS sequence. First, there is a length-22 palindrome TCTTTAACAAGCTTGTTAAAGA spanning positions 25962–25983. Second, there are two repeating length-12 palindromes TTATAATTATAA spanning positions 22712–22723 and 22796–22807. Some further investigations into possible biological implications of these palindrome features are proposed.

Key words: Markov chain; palindrome counts; simulation; RNA viral genome; severe acute respiratory syndrome
History: Accepted by Harvey J. Greenberg, Guest Editor; received August 2003; accepted January 2004.

1. Introduction

In March 2003, a novel coronavirus associated with the *severe acute respiratory syndrome* (SARS) was identified. The outbreak of SARS in different parts of the world, causing hundreds of deaths, has initiated much international effort that includes clinical, epidemiologic, and laboratory investigations with the aim of controlling the spread of the virus (Bloom 2003, Marra et al. 2003, Ruan et al. 2003, Rota et al. 2003). Although the world was cleared of new SARS cases by July 2003, the pursuit for a thorough understanding of the origin, evolution, and pathogenicity of this deadly virus continues.

With the availability of the complete genome sequence of the SARS and several other coronaviruses in public databases (e.g., GenBank), it is possible to do a computational analysis of the viral genome, looking for unusual genome sequence features either unique to the SARS virus or common to the coronavirus family. Such information can give clues to the origin, natural reservoir, and evolution of the virus. It may contribute to the studies of the immune response to this virus and the pathogenesis of SARS-related disease (Rota et al. 2003).

Statistical and experimental studies of palindromes in the other classes of viral genomes, such as the double stranded DNA viruses, bacteriophages, retroviruses, etc., have been performed (Cain et al. 2001, Dirac et al. 2002, Hill et al. 2003, Karlin et al. 1992, Leung et al. 2002, Rocha et al. 2001, among others). These studies have suggested that palindromes might be involved in the viral packaging, replication, and defense mechanisms. Unlike these well-studied viruses involved in fatal diseases such as AIDS and various cancers, the coronaviruses have not received as much attention until the recent outbreak of SARS.

In the present study, we focus our attention on palindromes in the positive-stranded RNA genomes of coronaviruses. In accordance with GenBank convention, we represent an RNA sequence as a string of letters from the alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The four letters respectively stand for the RNA bases adenine,

cytosine, guanine, and uracil. The letters A and T are complementary to each other because adenine and uracil form hydrogen bonds with each other. The same applies to C and G. A palindrome is a symmetrical word such that when it is read in the reverse direction, it is exactly the complement of itself. For example, ACGT is a palindrome of length four. A palindrome is necessarily even in length because the middle base in any odd-length nucleotide string cannot be identical to its complement.

Several points are worth noting from this initial exploratory analysis of palindromes in the coronavirus genome sequences: (1) The palindrome counts in the coronavirus genomes seem lower than what would be expected from random sequences. (2) The SARS virus contains an exceptionally long palindrome with 22 nucleotide bases. This is the longest among all palindromes observed in the coronaviruses. (3) There are two copies of a length-12 palindrome situated within 100 bases of each other in the SARS genome. This is not observed in the other coronaviruses.

Whether or not these palindrome-related features have any biological relevance will, of course, have to rely on careful laboratory investigations by the virologists. At this stage, however, it would be only reasonable to assess whether these features can indeed be considered statistically unusual when compared to random-sequence models. Our observations call for investigations into the probability distributions of palindrome counts, lengths, and locations in a random sequence. This paper will focus only on the palindrome counts, leaving the others for future studies.

In the next section, the mathematical formulas for the theoretical mean and variance for the number of palindromes at or above a prescribed length are derived based on a Markov-chain random-sequence model. Section 3 summarizes the computational results in comparing palindrome counts of the coronavirus genomes to the random-sequence models. In §4, we propose some biological questions that may be investigated in relation to these observed nonrandom features. A few concluding remarks are given in §5.

2. Palindrome Counts in Markov-Chain Models

The main objective of this paper is to assess whether the palindrome counts in the coronavirus genomes are observed more (or less) frequently than expected, under some specified probability models. We model the genome sequence as a realization of a sequence of random variables $\xi_1, \xi_2, \ldots, \xi_n$ taking values in $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and n is the genome length. Throughout, we will assume that either

(i) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ are independent and identically distributed (M0); or

(ii) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ form a stationary Markov chain of order one (M1).

For studying DNA words of length k, one can choose to use Markov chains of order up to the maximum order of $k-2$ as the sequence model. A higher-order Markov chain will better fit the data sequence, but at the same time the number of parameters in the model increases exponentially. In this study, we carried out some simulations using the second-order Markov-chain model (M2). The computation takes much longer, but the z scores obtained gave the same interpretation as that of the M1 model. We therefore content ourselves with the M0 and M1 models for our analysis of palindromes of length four and above.

We are interested in deriving the mean and standard deviation of the random variable X_L, total number of palindromes of length at least $2L$ under the M0 and M1 sequence models. This will help quantify the extent of deviation of the observed palindrome counts in the coronavirus genome from the expected counts under the specified probability model. For $L \le k \le n-L$, define

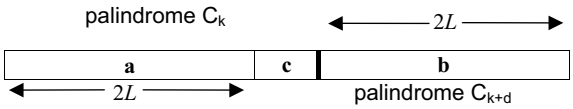
$$I_k = \begin{cases} 1 & \text{if the } k\text{th base is the left center of a palindrome of length} \ge 2L \\ 0 & \text{otherwise} \end{cases}.$$

We say that a palindrome *occurs* at k when $I_k = 1$. Therefore, $X_L = \sum_{k=L}^{n-L} I_k$. Note that the distribution of I_k depends only on the joint distribution of $(\xi_{k-L+1}, \ldots, \xi_{k+L})$. Under the M0 or M1 model, the joint distribution of $(\xi_{k-L+1}, \ldots, \xi_{k+L})$ is independent of k. Hence $\mathbb{P}[I_k = 1]$ is a constant in k. Similarly $\mathbb{P}[I_j = 1,\ I_k = 1]$ depends only on $|j-k|$. Therefore, for $L \le k \le n-L$ and $1 \le d \le n-L-k$, we define

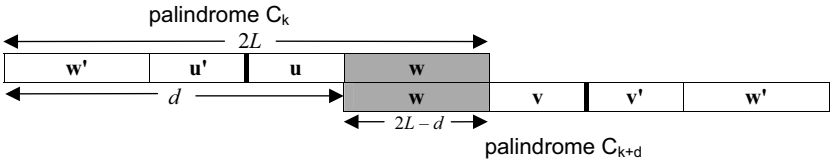
$$\gamma(0) := \mathbb{P}[I_k = 1] \quad \text{and} \quad \gamma(d) := \mathbb{P}[I_k = 1,\ I_{k+d} = 1].$$

The expressions of $\gamma(0)$ and $\gamma(d)$ are crucial to calculating the mean and variance of X_L (see Proposition 3 below). Lemma 1 (respectively, Lemma 2) deals with the computation of $\gamma(0)$ and $\gamma(d)$ under the M1 (respectively, M0) sequence model. Indeed, we will deduce Lemma 2 from Lemma 1.

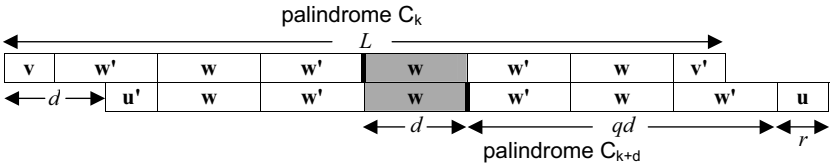
Throughout, we use b' to denote the complementary base of b, and $\mathbf{w}'$ the inversion (i.e., the complementary word read in reverse) of the word $\mathbf{w}$. There are quite a few details to work out all the possible overlap cases because the overlap structures depend on the relative sizes of d (the extent of overlap) and $2L$ (the cutoff length of a palindrome). However, there are only two basic patterns in the overlap. In the first pattern (as illustrated by Figure 1b), the shaded segment, due to the complimentary requirement of



(a) $d \geq 2L$. Here the palindromes C_k and C_{k+d} do not overlap and **c** denotes the segment between them.



(b) $L \leq d < 2L$. Here **w** denotes the common segment of palindromes C_k and C_{k+d}. And **w** determines the left end and right end of C_k and C_{k+d}.



(c) $1 \leq d < L$ with q as quotient when L is divided by d and r the remainder. The shaded segment determines the rest of both palindromes

Figure 1 Overlapping Structures of Palindromes C_k and C_{k+d} for Different Values of d
Note. (a), (b), and (c) are drawn with different scales.

a palindrome, will uniquely determine the left and right ends of C_k and C_{k+d}. And in the other pattern (as illustrated by Figure 1c), the shaded segment will determine the rest of both palindromes. In Figure 1a, even though palindromes C_k and C_{k+d} do not actually overlap (i.e., $d \geq 2L$), the occurrence of a palindrome at k will still have an effect on the probability that a palindrome will occur at $k+d$ under the M1 sequence model. Lemma 1 provides expressions of $\gamma(d)$ under all possible situations.

LEMMA 1. *Suppose the genome sequence is modeled as a stationary Markov chain of order one with stationary distribution* $\pi := (\pi(A), \pi(C), \pi(G), \pi(T))$. *For* $a, b \in \mathcal{A}$ *and* $m \geq 1$, *let* $P(a,b)$ *and* $P^{(m)}(a,b)$ *respectively denote the transition probability and the m-step transition probability from base a to base b.*

(a) *We have*

$$\gamma(0) = \sum_{b_1,\ldots,b_L \in \mathcal{A}} \pi(b_1) P(b_L, b'_L) \prod_{j=1}^{L-1} \left[P(b_j, b_{j+1}) P(b'_{j+1}, b'_j) \right]. \tag{1}$$

(b) *For* $d \geq 1$, *we have the following three cases*:

(i) $d \geq 2L$:

$$\gamma(d) = \sum_{\substack{1 \leq i \leq L \\ a_i, b_i \in \mathcal{A}}} \pi(a_1) P(a_L, a'_L) P(b_L, b'_L) P^{(d-2L+1)}(a'_1, b_1) \cdot \prod_{j=1}^{L-1} \left[P(a_j, a_{j+1}) P(a'_{j+1}, a'_j) P(b_j, b_{j+1}) P(b'_{j+1}, b'_j) \right].$$

(ii) $L \leq d < 2L$:

$$\gamma(d) = \sum_{b_1,\ldots,b_d \in \mathcal{A}} \pi(b'_L) P(b'_1, b_1) P(b_d, b'_d) \prod_{j=1}^{d-1} P(b_j, b_{j+1}) \cdot \prod_{l=1}^{L-1} \left[P(b'_{l+1}, b'_l) P(b'_{d-L+l+1}, b'_{d-L+l}) \right].$$

(iii) $1 \leq d < L$: *we let* $L = qd + r$.

$$\gamma(d) = \sum_{b_1,\ldots,b_d \in \mathcal{A}} K_{r,d}(b_1, \ldots, b_d) \left[P(b_d, b'_d) \prod_{j=1}^{d-1} P(b'_{j+1}, b'_j) \right]^{q+1} \cdot \left[P(b'_1, b_1) \prod_{j=1}^{d-1} P(b_j, b_{j+1}) \right]^{q},$$

where

$$K_{r,d}(b_1, \ldots, b_d) = \begin{cases} \pi(b_{d-r+1}) P(b'_1, b_1) \prod_{j=1}^{r-1} P(b_j, b_{j+1}) \\ \quad \cdot \prod_{j=d-r+1}^{d-1} P(b_j, b_{j+1}) & r \geq 2 \\ \pi(b_{d-r+1}) P(b'_1, b_1) & r = 1 \\ \pi(b'_d)/P(b_d, b'_d) & r = 0 \end{cases}.$$

PROOF. (a) Note that a palindrome of length at least $2L$ is of the form $b_1 \cdots b_L b'_L \cdots b'_1$ where $b_1, \ldots, b_L \in \mathcal{A}$. Therefore,

$$\gamma(0) = \sum_{b_1,\ldots,b_L \in \mathcal{A}} \mathbb{P}\left[b_1 \cdots b_L b'_L \cdots b'_1 \right].$$

Because

$$\mathbb{P}\big[b_1\cdots b_L b'_L\cdots b'_1\big] = \pi(b_1)\Bigg[\prod_{j=1}^{L-1} P(b_j, b_{j+1})\Bigg] \cdot P\big(b_L, b'_L\big)\Bigg[\prod_{j=1}^{L-1} P\big(b'_{j+1}, b'_j\big)\Bigg],$$

(1) follows immediately after rearranging terms.

(b) To compute the overlap probability $\gamma(d)$, i.e., the probability that there are palindromes at k and $k+d$, we call the stretch of bases $\xi_{k-L+1}\cdots\xi_{k+d+L}$ the *span* of palindromes C_k and C_{k+d}.

For (i) $d \geq 2L$: The span $\mathbf{s}$ of the two palindromes C_k and C_{k+d} is of the form $\mathbf{acb}$ where $\mathbf{a} = a_1\cdots a_L a'_L\cdots a'_1$, $\mathbf{c} = c_1\cdots c_{d-2L}$, and $\mathbf{b} = b_1\cdots b_L b'_L\cdots b'_1$. Hence,

$$\begin{aligned}\gamma(d) &= \sum_{\mathbf{a},\mathbf{c},\mathbf{b}} \mathbb{P}[\mathbf{s}] = \sum_{\mathbf{a},\mathbf{b}}\sum_{\mathbf{c}} \mathbb{P}[\mathbf{a}]\,\mathbb{P}\big[\mathbf{c}b_1 \mid a'_1\big]\,\mathbb{P}[\mathbf{b} \mid b_1] \\ &= \sum_{\mathbf{a},\mathbf{b}} \mathbb{P}[\mathbf{a}] P^{(d-2L+1)}(a'_1, b_1)\,\mathbb{P}[\mathbf{b} \mid b_1].\end{aligned}$$

Hence (i) follows immediately from

$$\mathbb{P}[\mathbf{a}] = \pi(a_1)\Bigg[\prod_{j=1}^{L-1} P(a_j, a_{j+1})\Bigg] P\big(a_L, a'_L\big)\Bigg[\prod_{j=1}^{L-1} P\big(a'_{j+1}, a'_j\big)\Bigg];$$

and

$$\mathbb{P}[\mathbf{b} \mid b_1] = \Bigg[\prod_{j=1}^{L-1} P(b_j, b_{j+1})\Bigg] P\big(b_L, b'_L\big)\Bigg[\prod_{j=1}^{L-1} P\big(b'_{j+1}, b'_j\big)\Bigg].$$

For (ii) $L \leq d < 2L$: Refer to Figure 1(b), let $\mathbf{w} = b_{d-L+1}\cdots b_L$ denote the common segment of palindromes C_k and C_{k+d}. Assuming $d > L$, let $\mathbf{u} = b_1\cdots b_{d-L}$ and $\mathbf{v} = b_{L+1}\cdots b_d$; we can represent $C_k = \mathbf{w}'\mathbf{u}'\mathbf{u}\mathbf{w}$ and $C_{k+d} = \mathbf{w}\mathbf{v}\mathbf{v}'\mathbf{w}'$ where $b_1, \ldots, b_d \in \mathcal{A}$. Therefore,

$$\begin{aligned}\gamma(d) &= \sum_{b_1,\ldots,b_d\in\mathcal{A}} \mathbb{P}[\mathbf{w}'\mathbf{u}'\mathbf{u}\mathbf{w}\mathbf{v}\mathbf{v}'\mathbf{w}'] \\ &= \sum_{b_1,\ldots,b_d\in\mathcal{A}} \mathbb{P}\big[b'_L\cdots b'_1 b_1\cdots b_d b'_d\cdots b'_{d-L+1}\big].\end{aligned}$$

Writing it out in terms of the initial distribution and transition probabilities, we have proved (ii) for $d > L$. The case for $d = L$ is similar: Take $\mathbf{u}$ and $\mathbf{v}$ as null words and proceed as in the case $d > L$.

To prove (iii), we consider the case $r \geq 1$ first. This time, let $\mathbf{w} = b_1\cdots b_d$ denote the first d bases to the right of the center of C_k and to the left of the center of C_{k+d}. Let $\mathbf{u} = b_1\cdots b_r$ and $\mathbf{v} = b_{d-r+1}\cdots b_d$, respectively denote the first and last r bases of $\mathbf{w}$. Figure 1(c) displays the necessary structure in C_k and C_{k+d} for both of them to be palindromes when $q = 3$. If q is odd, then the span of C_k and C_{k+d} is of the form $\mathbf{v}\underbrace{\mathbf{w}'\mathbf{w}}_{1}\cdots\underbrace{\mathbf{w}'\mathbf{w}}_{q}\mathbf{w}'\mathbf{u}$. Therefore,

$$\begin{aligned}\gamma(d) = \sum_{b_1,\ldots,b_d\in\mathcal{A}} \mathbb{P}\big[&b_{d-r+1}\cdots b_d \underbrace{b'_d\cdots b'_1 b_1\cdots b_d}_{1}\cdots \\ &\underbrace{b'_d\cdots b'_1 b_1\cdots b_d}_{q} b'_d\cdots b'_1 b_1\cdots b_r\big].\end{aligned} \tag{2}$$

If q is even, then the span of C_k and C_{k+d} is changed accordingly to the form $\mathbf{u}'\underbrace{\mathbf{w}\mathbf{w}'}_{1}\cdots\underbrace{\mathbf{w}\mathbf{w}'}_{q}\mathbf{w}\mathbf{v}'$ and

$$\begin{aligned}\gamma(d) = \sum_{b_1,\ldots,b_d\in\mathcal{A}} \mathbb{P}\big[&b'_r\cdots b'_1 \underbrace{b_1\cdots b_d b'_d\cdots b'_1}_{1}\cdots \\ &\underbrace{b_1\cdots b_d b'_d\cdots b'_1}_{q} b_1\cdots b_d b'_d\cdots b'_{d-r+1}\big].\end{aligned} \tag{3}$$

By making the one-to-one transformation in the summation, $b_1 \to b'_d, \ldots, b_d \to b'_1$, and we can see that both sums on the RHS of (2) and (3) are the same. So without loss of generality, we compute $\gamma(d)$ under the assumption that q is odd. The crucial step is then to calculate the probability of the span of C_k and C_{k+d}, and part (iii) will follow immediately from summing over all possible $b_1, \ldots, b_d$. We first consider $r \geq 2$, then

$$\begin{aligned}&\mathbb{P}\big[b_{d-r+1}\cdots b_d \underbrace{b'_d\cdots b'_1 b_1\cdots b_d}_{1}\cdots \\ &\underbrace{b'_d\cdots b'_1 b_1\cdots b_d}_{q} b'_d\cdots b'_1 b_1\cdots b_r\big] \\ &\quad= \pi(b_{d-r+1}) P\big(b'_1, b_1\big)\Bigg[\prod_{j=1}^{r-1} P(b_j, b_{j+1})\Bigg] \\ &\qquad\cdot\Bigg[\prod_{j=d-r+1}^{d-1} P(b_j, b_{j+1})\Bigg]\Bigg[P\big(b_d, b'_d\big)\prod_{j=1}^{d-1} P\big(b'_{j+1}, b'_j\big)\Bigg]^{q+1} \\ &\qquad\cdot\Bigg[P\big(b'_1, b_1\big)\prod_{j=1}^{d-1} P(b_j, b_{j+1})\Bigg]^{q}.\end{aligned} \tag{4}$$

For $r = 1$, (4) becomes

$$\begin{aligned}&\mathbb{P}\big[b_d \underbrace{b'_d\cdots b'_1 b_1\cdots b_d}_{1}\cdots\underbrace{b'_d\cdots b'_1 b_1\cdots b_d}_{q} b'_d\cdots b'_1 b_1\big] \\ &\quad= \pi(b_d) P\big(b'_1, b_1\big)\Bigg[P\big(b_d, b'_d\big)\prod_{j=1}^{d-1} P\big(b'_{j+1}, b'_j\big)\Bigg]^{q+1} \\ &\qquad\cdot\Bigg[P\big(b'_1, b_1\big)\prod_{j=1}^{d-1} P(b_j, b_{j+1})\Bigg]^{q}.\end{aligned}$$

If $r = 0$, reasoning similar to the above leads us to consider just the case q is odd. However, the span of C_k and C_{k+d} becomes (one can take $\mathbf{u}$ and $\mathbf{v}$ as

empty words) $\underbrace{\mathbf{w}'\mathbf{w}}_{1}\cdots\underbrace{\mathbf{w}'\mathbf{w}}_{q}\mathbf{w}'$. And hence,

$$\mathbb{P}\big[\underbrace{b'_d\cdots b'_1b_1\cdots b_d}_{1}\cdots\underbrace{b'_d\cdots b'_1b_1\cdots b_d}_{q}b'_d\cdots b'_1\big]$$
$$=\frac{\pi(b'_d)}{P(b_d,b'_d)}\left[P(b_d,b'_d)\prod_{j=1}^{d-1}P(b'_{j+1},b'_j)\right]^{q+1}\cdot\left[P(b'_1,b_1)\prod_{j=1}^{d-1}P(b_j,b_{j+1})\right]^{q}.\quad\square$$

Under the M0 model, the stationary distribution $\pi=(p_A,p_C,p_G,p_T)$, and the transition probabilities $P(a,b)=p_b$ and $P^{(m)}(a,b)=p_b$ for any $a,b\in\mathcal{A}$, $m\ge1$. Substituting these into Lemma 1(a) and (i) and (ii) of Lemma 1(b) immediately gives us the corresponding parts in Lemma 2 below. Part (iii) of Lemma 1(b) can be simplified further according to how big the remainder r is in relation to d. We shall omit the details. In this way, we have deduced the following Lemma 2, which was first proved in Leung et al. (2002).

Lemma 2. *Suppose the genome sequence is modeled as M*0 *and let*

$$\theta:=2(p_Ap_T+p_Cp_G).$$

(a) *We have*
$$\gamma(0)=\theta^L.$$

(b) *For* $d\ge1$, *we have the following four cases*:
(i) $d\ge2L$:
$$\gamma(d)=\theta^{2L};$$
(ii) $L\le d<2L$:
$$\gamma(d)=\theta^{2(d-L)}[p_Ap_T(p_A+p_T)+p_Cp_G(p_C+p_G)]^{2L-d};$$
when $1\le d<L$ *we let* $L=qd+r$ *where* $0\le r<d$, *and consider two subcases according to how big the remainder r is in relation to d*.
(iii) $1\le d<L$ *and* $0\le r<(d+1)/2$:
$$\gamma(d)=\big[2\big((p_Ap_T)^{q+1}+(p_Cp_G)^{q+1}\big)\big]^{2r}\cdot\big[(p_Ap_T)^q(p_A+p_T)+(p_Cp_G)^q(p_C+p_G)\big]^{d-2r};$$
(iv) $1\le d<L$ *and* $(d+1)/2\le r<d$:
$$\gamma(d)=\big[2\big((p_Ap_T)^{q+1}+(p_Cp_G)^{q+1}\big)\big]^{2(d-r)}\cdot\big[(p_Ap_T)^{q+1}(p_A+p_T)+(p_Cp_G)^{q+1}(p_C+p_G)\big]^{2r-d}.$$

Proposition 3. *With the* I_k*'s as defined at the beginning of* §2, *the total number of palindromes of length at least* 2L *is given by* $X_L:=\sum_{k=L}^{n-L}I_k$. *And hence*,

$$\lambda_L:=E(X_L)=(n-2L+1)\gamma(0)$$

and

$$\sigma_L^2:=\mathrm{Var}(X_L)=(n-2L+1)\gamma(0)(1-\gamma(0))+2\sum_{d=1}^{n-2L}(n-2L+1-d)[\gamma(d)-\gamma(0)^2],$$

where $\gamma(0)$ *and* $\gamma(d)$ *are given as in Lemma* 2 *under the M*0 *sequence model, and Lemma* 1 *under M*1 *sequence model*.

Proof. The first equation follows immediately from taking expectations on both sides of $X_L:=\sum_{k=L}^{n-L}I_k$, and

$$\begin{aligned}\sigma_L^2&=\sum_{j=L}^{n-L}\mathrm{Var}(I_j)+2\sum_{j=L}^{n-L-1}\sum_{k=j+1}^{n-L}\mathrm{Cov}(I_j,I_k)\\&=(n-2L+1)\gamma(0)(1-\gamma(0))+2\sum_{j=L}^{n-L-1}\sum_{d=1}^{n-L-j}\big[\mathbb{P}[I_j=1,\ I_{j+d}=1]-\gamma(0)^2\big]\\&=(n-2L+1)\gamma(0)(1-\gamma(0))+2\sum_{d=1}^{n-2L}(n-2L+1-d)[\gamma(d)-\gamma(0)^2].\quad\square\end{aligned}$$

3. Palindrome Counts in Coronaviruses

The derived means and variances under the M0 and M1 sequence models enable us to assess whether the observed palindrome count in a genome is too abundant or rare. The z score defined in (5) below is a modification of a generally accepted measure of over (or under)representation of a DNA word. For $L\ge2$, a standardized frequency under the assumption of the M1 sequence model is defined as

$$z_{\mathrm{M1}}=\frac{X_L-\mu_{\mathrm{M1}}}{\sigma_{\mathrm{M1}}},\tag{5}$$

where X_L is the observed number of palindromes of length at least $2L$, and μ_{M1} and σ_{M1} denote its expected value and standard deviation, respectively. (For simplicity, we do not indicate the dependence of μ and σ on L.) The corresponding z score is defined similarly for the M0 sequence model. When L is small compared with the genome length n, X_L is a sum of weakly dependent random indicators I_k and it is therefore well approximated by a normal distribution. Indeed, if we let $X_L^{(j)}$ denote the number of occurrences of the jth palindrome in the genome, then the count vector $(X_L^{(1)},X_L^{(2)},\ldots,X_L^{(4^L)})$ will converge to a multivariate normal distribution as $n\to\infty$ (see Theorem 12.5 in Waterman 1995). And hence $X_L=\sum_{1\le j\le4^L}X_L^{(j)}$ will converge to a normal distribution as $n\to\infty$. For $L=2$ or 3, and n in the range 30,000, we expect that the distribution of the z scores will

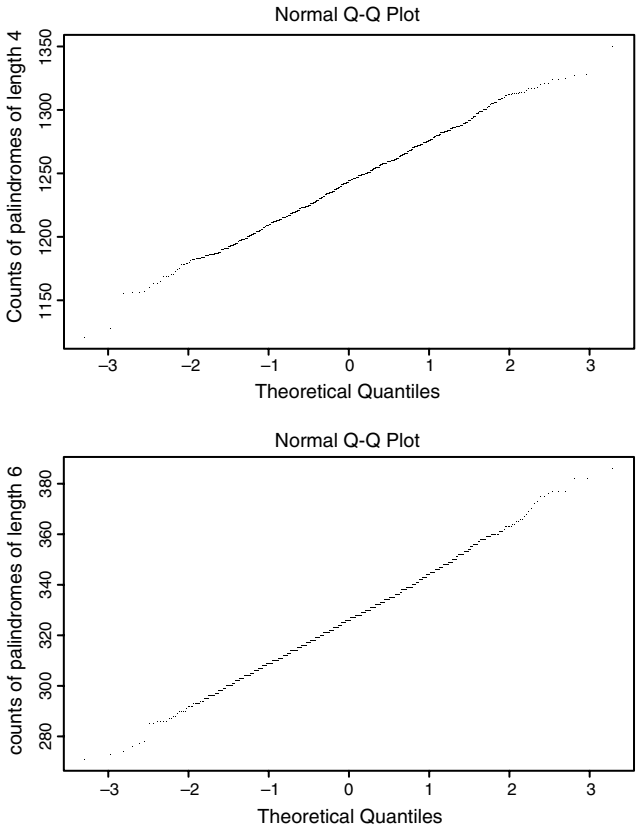


Figure 2 **Normal Q-Q Plots of Counts of Palindromes of Length Four (Top) and Six (Bottom) in the 1,000 Random Sequences Under the M1 Model for the SARS Genome**

be approximately standard normal. The near-straight lines in the Q-Q plots in Figure 2 confirmed that this is the case. This motivates our definition: The count is said to be *over* (or *under*)*represented*, if the z score is greater than 1.645 or less than -1.645, respectively (i.e., in the upper or lower 5% of a standard normal distribution, as commonly used in one-tailed hypothesis tests in biological experiments). However, it should be emphasized that these cutoff z score values can only be considered as a convenient statistical guideline to help bring out interesting observations rather than a strict criterion to lead to a definitive conclusion.

We compute the z scores of the genomes in the following data set: It is composed of seven coronaviruses with complete genome sequences and four other RNA viruses. For some coronaviruses, the genome sequences of multiple strains of the same virus are available. Only one strain is included in our data set because their genomes are very similar. Four other RNA viruses outside the coronavirus family are included in the data set. Two of these (the rubella virus and the equine arteritis virus) have positive-stranded RNA genomes like the coronaviruses, one (rabies virus) has a negative-stranded RNA genome, and the remaining one (HIV) is a retrovirus. Table 1 lists the names of the viruses, abbreviations, GenBank accession numbers, genome lengths, and base compositions of the seven coronaviruses and the other four RNA viruses. Table 2 displays the z scores for counts of palindromes of length four and above under the M0 and M1 models.

Table 2 indicates that there is a general avoidance of palindromes of length four and above in the coronavirus genomes. A natural question that follows is whether palindromes of a given exact length are also underrepresented in these viruses.

To answer this question, one would need the mean ν and standard deviation τ for the count Y_L of palindromes of exact length $2L$. It is easy to obtain the mean because $\nu = E(Y_L) = E(X_L) - E(X_{L+1})$. The standard deviation of Y_L can be derived with suitable modification of the method of proofs in Lemmas 1 and 2, but the expression obtained is rather lengthy due to an increase in the overlapping structures. Instead, we adopt an alternative approach to estimate the standard deviation by simulation, which at the same time serves to validate our derived means and standard deviations. This approach has a further advantage of giving us the empirical distributions, and Figure 2 shows that for small values of L, the distributions are well approximated by normal distributions.

Table 1 **List of Seven Coronaviruses and Four Other RNA Viruses to be Analyzed**

Name	Abbrev.	Accession	Length	Base composition
SARS coronavirus Urbani	SARS	AY278741	29,727	(0.28, 0.20, 0.21, 0.31)
Avian infectious bronchitis virus	AIBV	NC_001451.1	27,608	(0.29, 0.16, 0.22, 0.33)
Bovine coronavirus	BCoV	NC_003045.1	31,028	(0.27, 0.15, 0.22, 0.36)
Human coronavirus 229E	HCoV	NC_002645.1	27,317	(0.27, 0.17, 0.22, 0.35)
Murine hepatitis virus	MHV	NC_001846	31,357	(0.26, 0.18, 0.24, 0.32)
Porcine epidemic diarrhea virus	PEDV	NC_003436.1	28,033	(0.25, 0.19, 0.23, 0.33)
Transmissible gastroenteritis virus	TGV	NC_002306.2	28,586	(0.29, 0.17, 0.21, 0.33)
Rubella virus	RUV	NC_001545.1	9,755	(0.15, 0.39, 0.31, 0.15)
Equine arteritis virus	EAV	NC_002532.2	12,704	(0.21, 0.26, 0.26, 0.27)
Rabies virus	RV	NC_001542.1	11,932	(0.29, 0.22, 0.23, 0.26)
Human immunodeficiency virus 1	HIV-1	NC_001802.1	9,181	(0.36, 0.18, 0.24, 0.22)

Table 2 *z* **Scores for Counts of Palindromes of Length Four and Above**

Virus	Counts	μ_{M0} (σ_{M0})	μ_{M1} (σ_{M1})	z_{M0}	z_{M1}
SARS	1,554	1,981.0 (43.4)	1,687.6 (40.3)	−9.83	−3.32
AIBV	1,578	1,896.6 (42.8)	1,675.3 (38.2)	−7.45	−2.54
BCoV	1,886	2,115.6 (45.4)	2,007.5 (45.5)	−5.06	−2.67
HCoV	1,451	1,843.6 (42.2)	1,567.6 (37.0)	−9.30	−3.15
MHV	1,793	2,006.6 (43.8)	1,911.3 (41.4)	−4.88	−2.86
PEDV	1,457	1,781.6 (41.2)	1,578.8 (38.3)	−7.87	−3.18
TGV	1,610	1,993.9 (43.8)	1,695.6 (38.9)	−8.76	−2.20
RUV	868	793.2 (28.0)	845.6 (28.3)	2.67	0.79
EAV	672	784.3 (27.2)	710.4 (25.8)	−4.13	−1.49
RV	559	758.0 (26.7)	564.3 (23.0)	−7.45	−0.23
HIV-1	475	551.9 (23.1)	480.2 (21.9)	−3.33	−0.24

For each virus in Table 1, 1,000 random sequences were generated for both the M0 and M1 models using scripts written in the R language (http://www.r-project.org/). The sequences are run through the *palindrome* program which is part of EMBOSS (European Molecular Biology Open Software Suite, Rice et al. 2000) to extract the palindrome positions and length. Each output is then read by R again and the counts of palindromes of various length are tabulated.

Tables 3 and 4 present the counts of palindromes of exact length four, six, and eight, along with their expected values ν, estimated standard deviations $\hat{\tau}$, and z scores. Based on the z scores, Tables 3 and 4 indicate that length-four palindromes are significantly underrepresented across the coronavirus family under both the M0 and M1 sequence models. However, for length-six palindromes, SARS is the only member of the coronavirus family that shows underrepresentation under the M1 sequence model. For length eight or above, no distinct patterns are observed.

For palindromes of length four and above, it is possible to fit higher-order Markov models to the genome sequence. For example, the second-order Markov-chain model that takes the base, dinucleotide, as well as trinucleotide composition into account, can be used to calculate the z scores. We simulated 1,000 random sequences with the M2 model, but the results did not differ much from the M1 model.

As the EMBOSS *palindrome* program provides us with a detailed listing of all occurrences of palindromes of length four and above, we are able to notice two unique features in SARS. First, the SARS sequence contains a long palindrome of length 22, the longest among all palindromes observed in the coronaviruses. Second, there are two identical, length-12 palindromes situated within 100 bases of each other in the SARS genome. These are not observed in the other coronaviruses. Although contributing little to the total palindrome counts, these three palindromes appear unusual enough to warrant further study of their possible biological roles, as discussed in the next section.

4. Discussion

Various statistical assessments of unusual abundance and rarity of individual words, including individual palindromes, in nucleotide sequences have been done using random-sequence models in a number of previous studies (Karlin et al. 1992; Merkl and Fritz 1996; Rocha et al. 1998, 2001; Schbath et al. 1995, to name just a few). The present study, however, aims at investigating the unusual abundance and rarity of palindromes collectively rather than individually. The mathematical results in §2 provide a directly computable formula to give a single z score for all palindromes with a given minimal length. We hope the exploratory results in this paper will serve as a basis for more detailed investigations to see how palindromes might be involved in important biological mechanisms of the coronaviruses.

There are two random sequence models M0 and M1 used in this paper. Because M1 can take the genome dinucleotide compositions into consideration while M0 cannot, M1 is preferred over M0. Comparatively, the z scores under M1 are less extreme than those

Table 3 *z* **Scores for Palindromes of Various Lengths Under the M0 Model**

	Length-four palindromes			Length-six palindromes			Length-eight palindromes		
	Counts	ν_{M0} ($\hat{\tau}_{M0}$)	z_{M0}	Counts	ν_{M0} ($\hat{\tau}_{M0}$)	z_{M0}	Counts	ν_{M0} ($\hat{\tau}_{M0}$)	z_{M0}
SARS	1,144	1,469.6 (36.9)	−8.82	284	379.4 (19.4)	−4.92	90	97.9 (9.7)	−0.82
AIBV	1,142	1,399.5 (37.5)	−6.87	320	366.8 (18.6)	−2.52	91	96.1 (9.9)	−0.52
BCoV	1,360	1,563.2 (40.4)	−5.03	389	408.2 (20.4)	−0.94	98	106.6 (10.7)	−0.80
HCoV	1,054	1,364.7 (36.9)	−8.42	287	354.5 (18.9)	−3.57	82	92.1 (9.8)	−1.03
MHV	1,328	1,499.0 (38.0)	−4.50	340	379.2 (19.5)	−2.01	82	95.9 (9.9)	−1.41
PEDV	1,079	1,332.5 (36.5)	−6.94	274	335.9 (18.5)	−3.35	79	84.7 (9.2)	−0.62
TGV	1,180	1,467.3 (38.4)	−7.48	306	387.5 (19.7)	−4.14	85	102.3 (9.8)	−1.77
RUV	610	567.0 (22.8)	1.89	167	161.7 (12.6)	0.42	68	46.1 (6.9)	3.17
EAV	479	589.4 (23.8)	−4.64	145	146.4 (12.3)	−0.12	36	36.4 (6.1)	−0.06
RV	407	567.0 (23.7)	−6.75	102	142.9 (12.4)	−3.30	38	36.0 (5.9)	0.34
HIV-1	347	416.6 (20.1)	−3.46	89	102.1 (10.2)	−1.29	34	25.0 (4.8)	1.87

Table 4 z **Scores for Palindromes of Various Lengths Under the M1 Model**

	Length-four palindromes			Length-six palindromes			Length-eight palindromes		
	Counts	ν_{M1} $(\hat{\tau}_{M1})$	z_{M1}	Counts	ν_{M1} $(\hat{\tau}_{M1})$	z_{M1}	Counts	ν_{M1} $(\hat{\tau}_{M1})$	z_{M1}
SARS	1,144	1,242.7 (33.4)	−2.96	284	327.3 (18.0)	−2.41	90	86.5 (9.4)	0.37
AIBV	1,142	1,229.8 (35.4)	−2.48	320	326.9 (17.8)	−0.39	91	87.0 (9.4)	0.42
BCoV	1,360	1,476.5 (37.2)	−3.13	389	390.4 (19.5)	−0.07	98	103.4 (9.8)	−0.55
HCoV	1,054	1,146.9 (34.5)	−2.69	287	307.6 (17.4)	−1.18	82	82.7 (8.9)	−0.08
MHV	1,328	1,421.3 (37.8)	−2.47	340	364.3 (18.8)	−1.29	82	93.5 (9.8)	−1.17
PEDV	1,079	1,169.8 (34.5)	−2.63	274	302.9 (17.5)	−1.65	79	78.6 (9.1)	0.05
TGV	1,180	1,239.5 (34.0)	−1.75	306	333.2 (18.4)	−1.48	85	89.8 (9.7)	−0.49
RUV	610	604.3 (24.5)	0.23	167	172.5 (13.8)	−0.40	68	49.2 (6.9)	2.72
EAV	479	529.6 (22.5)	−2.25	145	134.8 (11.3)	0.91	36	34.3 (5.7)	0.30
RV	407	415.2 (19.1)	−0.43	102	109.8 (10.4)	−0.75	38	28.9 (5.3)	1.71
HIV-1	347	358.3 (18.7)	−0.60	89	91.0 (9.6)	−0.21	34	23.1 (4.5)	2.42

of M0. M1 is therefore more conservative in declaring the palindrome counts in a genome to be significantly different from those in random sequences. We shall base our discussion of the results on M1 whenever possible.

The counts of palindromes of length at least four in each coronavirus analyzed are significantly lower than expected (see Table 2). As the palindrome length increases to six and above, the underrepresentation of palindromes no longer holds across the family (theoretical z scores under M1 range from −1.66 to 0.46). This suggests that there is a family-wide avoidance of palindromes of exact length four in the coronaviruses, which is confirmed by the empirical z scores for exact-length palindromes in Tables 3 and 4. With this knowledge, a thorough examination of the relative abundance of individual length-four palindromes, conditional on the total length-four palindrome count is called for. We are in the process of setting up such a study.

Although the underrepresentation of length-four palindromes is observed for all of the coronaviruses in our data set that include members from all three antigenic groups (Marra et al. 2003), this underrepresentation is not universally true in all RNA viruses, as demonstrated by the other RNA viruses outside the coronavirus family. While it is conceivable that palindrome underrepresentation is just a characteristic of the common ancestor of the coronaviruses, it is worth noting that the characteristic is preserved in the family despite the reputation for RNA viruses to be nature's swiftest evolvers (Worobey and Holmes 1999). So far, we cannot find any previous report of underrepresentation of short palindromes in RNA viruses with eukaryotic hosts. However, avoidance of short palindromes in some bacterial and phage DNA genomes has been reported in several studies (Karlin et al. 1992; Merkl and Fritz 1996; Rocha et al. 1998, 2001, among others). The phenomenon is generally explained in relation to the defense mechanisms of the bacterial and phage genomes, protecting themselves against being destroyed by restriction enzymes capable of cutting up DNA molecules at certain palindromic sites. It will be interesting to investigate whether there is any possible interaction of the short palindromes in the coronavirus genomes with the immune system of the host cells that might have detrimental effects on the survival of the virus.

Length-six palindromes are found significantly underrepresented only in SARS but not in the other six coronaviruses (see Table 4). Would this avoidance of length-six palindromes in the SARS genome offer a protective effect on the virus, making it comparatively more difficult to be destroyed and contributing to the rapid spread and the severity of the disease? This will be an interesting point to observe as we seek to learn more about the SARS virus.

Among all palindromes found in the seven coronaviruses genomes we analyzed, the longest one resides in SARS. It is composed of the 22 bases TCTTTAACAAGCTTGTTAAAGA spanning positions 25962–25983. Because the probability distribution of palindrome lengths has not been rigorously obtained, we can only attempt a rough estimation, based on the simple M0 sequence model, of observing a length-22 palindrome in a genome with base composition like that of SARS. It has been demonstrated in Leung et al. (2002) that for larger values of L (say ≥ 5), we may approximate the counts of palindromes at or above length $2L$ by a Poisson random variable with parameter λ equal to the expected count. We therefore have $\mathbb{P}[\text{maximal palindrome length} \geq 22] = \mathbb{P}[X_{11} \geq 1]$, which can be approximated by the corresponding Poisson probability with $\lambda_{11} = E(X_{11}) = 0.01008$ by Proposition 3. This Poisson probability is equal to $1 - e^{-\lambda_{11}}$, about 1%.

Knowing that this long palindrome is quite unlikely to occur by chance, one would logically ask the question of whether it plays any particular functional role. According to the classification of open reading frames

(ORFs) encoding potential nonstructural proteins of the SARS virus (Rota et al. 2003, Table 1), this palindrome occurs in the overlapping region of the two ORFs designated X1 and X2. Due to the location of this palindrome, it is tempting to speculate that it might be involved in some secondary structures serving similar purposes like those of a pseudoknot, which is typically found at frame-shift locations in overlapping coding sequences (Giedroc et al. 2000). One would have to perform a detailed secondary structure prediction on this part of the SARS and other coronavirus genomes before further suggestions can be made. The methods and tools used by Qin et al. (2003) to predict the secondary structure in another part of the SARS virus genome (around the packaging-signal sequence) are likely to be applicable here as well.

Another feature unique to SARS is the occurrence of two repeating length-12 palindromes TTATAATTATAA spanning positions 22712–22723 and 22796–22807, all within 100 bases of the genome in the coding sequence of the surface-spike glycoprotein, which is important for virus entry and virus-receptor interactions (Yu et al. 2003). Both copies begin on the third position of a codon. Three amino acids Tyr-Asn-Tyr are coded by the second through tenth bases of the palindrome. No such repeating palindromes are observed in the corresponding glycoprotein-coding sequences for any of the other six coronaviruses. Probabilistic assessment of close repeating palindromes occurring in random sequences has yet to be formulated mathematically or estimated by simulation. (The method of Robin and Daudin 1999 can be used to assess the probability that a given palindrome repeats itself in close proximity.) If such an observation is found to be unlikely to occur by chance, then these repeating palindromes might be tested for potential regulatory functions. Large palindromes present in single-stranded RNA have the inherent ability to form double-stranded stem structures through the formation of intramolecular base pairs; thus, it is possible that these sequences form secondary RNA structures in the genomic RNA and in one or more subgenomic RNAs of the SARS virus. In many of the single-stranded RNA viruses, stem structures play important regulatory roles in genome replication or gene expression. It should be possible to investigate potential regulatory roles of these repeated length-12 palindromes by engineering silent mutations within these sequences such that the encoded protein is not altered but the palindromes and putative secondary structures are lost.

5. Concluding Remarks

While we hope that there will never be another outbreak of SARS, we believe that detailed analysis of the SARS genome sequence can help generate useful information for understanding the biology of the coronaviruses and perhaps other RNA viruses in general. This first exploration about palindromes in the coronavirus family generates many questions to be investigated in greater detail mathematically, computationally, as well as biologically.

Closely related to palindromes is the sequence feature of close inversion, which is a palindrome with its two halves separated by a short stretch of intervening nucleotides. These close inversions are well known to form stem-loop and other secondary structures involved in the viral recombination and packaging process (Rowe et al. 1997, Qin et al. 2003). We anticipate that a set of interesting and challenging questions in random-sequence models will again emerge from the analysis of close inversions.

Acknowledgments

K. P. Choi was supported by BMRC Grant BMRC01/1/21/19/140 and M. Y. Leung by NIH Grants S06GM08194-23 and S06GM08194-24 and NSF Grant DUE9981104.

References

Bloom, B. R. 2003. Lessons from SARS. *Science* **300** 701.

Cain, D., O. Erlwein, A. Grigg, R. A. Russell, M. O. McClure. 2001. Palindromic sequence plays a critical role in human foamy virus dimerization. *J. Virology* **75** 3731–3739.

Dirac, A. M., H. Huthoff, J. Kjems, B. Berkhout. 2002. Requirements for RNA heterodimerization of the human immunodeficiency virus type 1 (HIV-1) and HIV-2 genomes. *J. General Virology* **83** 2533–2542.

Giedroc, D. P., C. A. Theimer, P. L. Nixon. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J. Molecular Biol.* **298** 167–185.

Hill, M. K., M. Shehu-Xhilaga, S. M. Campbell, P. Poumbourios, S. M. Crowe, J. Mak. 2003. The dimer initiation sequence stem-loop of Human Immunodeficiency Virus Type 1 is dispensable for viral replication in peripheral blood mononuclear cells. *J. Virology* **77** 8329–8335.

Karlin, S., C. Burge, A. M. Campbell. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* **20** 1363–1370.

Leung, M. Y., K. P. Choi, A. Xia, L. H. Y. Chen. 2002. Nonrandom clusters of palindromes in herpesvirus genomes. IMS preprint series 2002-2, Institute for Mathematical Sciences, National University of Singapore, Singapore.

Marra, M. A., S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton, R. L. Roper. 2003. The genome sequence of the SARS-associated coronavirus. *Science* **300** 1399–1404.

Merkl, R., H. J. Fritz. 1996. Statistical evidence for a biochemical pathway of natural, sequence-targeted G/C to C/G transversion mutagenesis in *Haemophilus influenzae Rd. Nucleic Acids Res.* **24** 4146–4151.

Qin, L., B. Xiong, C. Luo, Z. M. Guo, P. Hao, J. Su, P. Nan, Y. Feng, Y. X. Shi, X. J. Yu, X. M. Luo, K. X. Chen, X. Shen, J. H. Shen, J. P. Zou, G. P. Zhao, T. L. Shi, W. Z. He, Y. Zhong, H. L. Jiang, Y. X. Li. 2003. Identification of probable genomic packaging signal sequence from SARS-CoV genome by bioinformatics analysis. *Acta Pharmacologica Sinica* **24** 489–496.

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

Robin, S., J. J. Daudin. 1999. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.* **36** 179–193.

Rocha, E. P., A. Danchin, A. Viari. 2001. Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.* **11** 946–958.

Rocha, E. P., A. Viari, A. Danchin. 1998. Oligonucleotide bias in *Bacillus subtilis*: General trends and taxonomic comparisons. *Nucleic Acids Res.* **26** 2971–2980.

Rota, P. A., M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Penaranda, B. Bankamp, K. Maher, M. H. Chen, S. Tong, A. Tamin, L. Lowe, M. Frace, J. L. DeRisi, Q. Chen, D. Wang, D. D. Erdman, T. C. Peret, C. Burns, T. G. Ksiazek, P. E. Rollin, A. Sanchez, S. Liffick, B. Holloway, J. Limor, K. McCaustland, M. Olsen-Rasmussen, R. Fouchier, S. Gunther, A. D. Osterhaus, C. Drosten, M. A. Pallansch, L. J. Anderson, W. J. Bellini. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300** 1394–1399.

Rowe, C. L., J. O. Fleming, M. J. Nathan, J. Y. Sgro, A. C. Palmenberg, S. C. Baker. 1997. Generation of coronavirus spike deletion variants by high-frequency recombination at regions of predicted RNA secondary structure. *J. Virology* **71** 6183–6190.

Ruan, Y. J., C. L. Wei, A. E. Ling, V. B. Vega, H. Thoreau, S. T. Su, J. M. Chia, P. Ng, K. P. Chiu, L. Lim, T. Zhang, K. P. Chan, L. E. Oon, M. L. Ng, S. Y. Leo, L. F. P. Ng, E. C. Ren, L. W. Stanton, P. M. Long, E. T. Liu. 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* **361** 1779–1785.

Schbath, S., B. Prum, E. de Turckheim. 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.* **2** 417–437.

Waterman, M. S. 1995. *Introduction to Computational Biology*. Chapman & Hall, New York.

Worobey, M., E. C. Holmes. 1999. Evolutionary aspects of recombination in RNA viruses. *J. General Virology* **80** 2535–2543.

Yu, X. J., C. Luo, J. C. Lin, P. Hao, Y. Y. He, Z. M. Guo, L. Qin, J. Su, B. S. Liu, Y. Huang, P. Nan, C. S. Li, B. Xiong, X. M. Luo, G. P. Zhao, G. Pei, K. X. Chen, X. Shen, J. H. Shen, J. P. Zou, W. Z. He, T. L. Shi, Y. Zhong, H. L. Jiang, Y. X. Li. 2003. Putative hAPN receptor binding sites in SARS-CoV spike protein. *Acta Pharmacologica Sinica* **24** 481–488.