Palindromes in SARS and other Coronaviruses

Ming-Ying Leung
Department of Mathematical Sciences
University of Texas at El Paso
El Paso, TX 79968-0514

Outline:
- Coronavirus genomes
- Palindromes
- Mean and Variance of palindrome counts
- Under-representation of short palindromes
- A long palindrome in SARS
**DNA and RNA**

DNA is deoxyribonucleic acid, made up of 4 nucleotide bases Adenine, Cytosine, Guanine, and Thymine.

RNA is ribonucleic acid, made up of 4 nucleotide bases Adenine, Cytosine, Guanine, and Uracil.

For uniformity of notation, all DNA and RNA data sequences deposited in GenBank are represented as sequences of A, C, G, and T. The bases A and T form a complementary pair, so are C and G.
**Palindrome**: A string of nucleotide bases that reads the same as its reverse complement. A palindrome must be even in length.

E.g. A palindrome of length 10.

```
5' ..... GCAATATTGC .....3'
```

Note that for a palindrome of length $2L$, the $i$th and the $(2L-i+1)$st base must be complementary to each other.

We say that the palindrome occurs at position $j$ when it is centered between positions $j$ and $j + 1$. 
Palindromes counts in random nucleotide sequences

Define the indicator random variable

\[ I_j = \begin{cases} 
1 & \text{if palindrome of length } \geq 2L \text{ occurs at base } j \\
0 & \text{otherwise} 
\end{cases} \]

Then

\[ X_L = \sum_{j=L}^{n-L} I_k \]

is the total count of palindromes of length at least \(2L\) in a sequence of length \(n\).
Mean and variance of palindrome counts

\[ \mu_L = E(X_L) = E\left(\sum_{j=L}^{n-L} I_k\right) = (n - 2L + 1)E(I_L) \]

\[ \sigma^2_L = \text{var}(X_L) = \sum_{j=L}^{n-L} \text{var}(I_j) + 2 \sum_{j=L}^{n-L-1} \sum_{k=j+1}^{n-L} \text{cov}(I_j, I_k) \]

If we let

\[ \gamma(0) = P(I_j = 1) \text{ for } L \leq j \leq n - L \]
\[ \gamma(d) = P(I_j = 1, I_{j+d} = 1) \text{ for } 1 \leq d \leq n - L - j \]

then

\[ E(I_j) = \gamma(0) \]
\[ \text{var}(I_j) = \gamma(0)(1 - \gamma(0)) \]
\[ \text{cov}(I_j, I_{j+d}) = \gamma(d) - \gamma(0)^2 \]
Mean and variance of palindrome counts (cont’d)

\[ \mu_L = E(X_L) = (n - 2L + 1)\gamma(0) \]

\[ \sigma^2_L = \text{var}(X_L) \]

\[ = \sum_{j=L}^{n-L} \text{var}(I_j) + 2 \sum_{j=L}^{n-L-1} \sum_{k=j+1}^{n-L} \text{cov}(I_j, I_k) \]

\[ = (n - 2L + 1)\gamma(0)(1 - \gamma(0)) \]

\[ + 2 \sum_{d=1}^{n-2L} (n - 2L + 1 - d) \left[ \gamma(d) - \gamma(0)^2 \right] \]
How to find the $\gamma$’s?

Under a Markov sequence model, Chew et al. (2004, to appear in *INFORMS Journal of Computing*) have obtained computable formulas for the $\gamma$’s, expressed in terms of the transition and stationary probabilities of the Markov chain. These can be estimated by the observed base frequencies and dinucleotide frequencies.

Let’s look at a special case, namely the i.i.d. random sequence model where the nucleotide bases are generated independently with probability $p_A, p_C, p_G, p_T$. 
Finding $\gamma(0)$ for the i.i.d. sequence model

$$\gamma(0) = P(I_j = 1) = [2(p_A p_T + p_C p_G)]^L$$

\begin{array}{ccccccc}
  j - L + 1 & j & j + 1 & j + L \\
  b_1 & b_2 & \ldots & b_L & b_{L+1} & \ldots & b_{2L-1} & b_{2L} \\
\end{array}
Finding $\gamma(d)$ for the i.i.d. sequence model:

**Case 1: $d \geq 2L$**

Case 2: $L \leq d \leq 2L$

Case 3: $1 \leq d < L$
The $z$-score

If $\mu$ and $\sigma$ are mean and variance of the palindrome counts under a certain random model, the $z$-score

$$z = \frac{X_L - \mu}{\sigma}$$

is a measure of over- or under-representation of palindromes in the sequence. For small $L$, the $z$-score is approximately normally distributed.
Normal Q–Q Plot

counts of palindromes of length 6

Theoretical Quantiles
**$z$-Scores for Counts of Palindromes of Length 4 or Longer**

<table>
<thead>
<tr>
<th>Virus</th>
<th>Counts</th>
<th>$\mu(\sigma)$</th>
<th>$z$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARS</td>
<td>1554</td>
<td>1687.6 (40.3)</td>
<td>-3.32</td>
</tr>
<tr>
<td>AIBV</td>
<td>1578</td>
<td>1675.3 (38.2)</td>
<td>-2.54</td>
</tr>
<tr>
<td>BCoV</td>
<td>1886</td>
<td>2007.5 (45.5)</td>
<td>-2.67</td>
</tr>
<tr>
<td>HCoV</td>
<td>1451</td>
<td>1567.6 (37.0)</td>
<td>-3.15</td>
</tr>
<tr>
<td>MHV</td>
<td>1793</td>
<td>1911.3 (41.4)</td>
<td>-2.86</td>
</tr>
<tr>
<td>PEDV</td>
<td>1457</td>
<td>1578.8 (38.3)</td>
<td>-3.18</td>
</tr>
<tr>
<td>TGV</td>
<td>1610</td>
<td>1695.6 (38.9)</td>
<td>-2.20</td>
</tr>
<tr>
<td>RUV</td>
<td>868</td>
<td>845.6 (28.3)</td>
<td>0.79</td>
</tr>
<tr>
<td>EAV</td>
<td>672</td>
<td>710.4 (25.8)</td>
<td>-1.49</td>
</tr>
<tr>
<td>RV</td>
<td>559</td>
<td>564.3 (23.0)</td>
<td>-0.23</td>
</tr>
<tr>
<td>HIV-1</td>
<td>475</td>
<td>480.2 (21.9)</td>
<td>-0.24</td>
</tr>
</tbody>
</table>
All the z-scores of the coronaviruses are below -1.645, the 5\textsuperscript{th} percentile of the standard normal, suggesting that palindromes of length 4 or longer are underrepresented in the coronavirus family. This is not true for all RNA viruses.

It would be of interest to investigate the representation of palindromes at exact lengths 4, 6, 8,\ldots For each virus sequence, 1000 Markov sequences are simulated to estimate the mean and standard deviation of palindrome counts at various exact lengths. For short palindromes, the z-scores are roughly normally distributed, as demonstrated by Q-Q plots.
### z-Scores for Palindromes of Various Exact Lengths

<table>
<thead>
<tr>
<th>Virus Name</th>
<th>Length 4</th>
<th></th>
<th>Length 6</th>
<th></th>
<th>Length 8</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Counts</td>
<td>z-score</td>
<td>Counts</td>
<td>z-score</td>
<td>Counts</td>
<td>z-score</td>
</tr>
<tr>
<td>SARS</td>
<td>1144</td>
<td>-2.96</td>
<td>284</td>
<td>-2.41</td>
<td>90</td>
<td>0.37</td>
</tr>
<tr>
<td>AIBV</td>
<td>1142</td>
<td>-2.48</td>
<td>320</td>
<td>-0.39</td>
<td>91</td>
<td>0.42</td>
</tr>
<tr>
<td>BCoV</td>
<td>1360</td>
<td>-3.13</td>
<td>389</td>
<td>-0.07</td>
<td>98</td>
<td>-0.55</td>
</tr>
<tr>
<td>HCoV</td>
<td>1054</td>
<td>-2.69</td>
<td>287</td>
<td>-1.18</td>
<td>82</td>
<td>-0.08</td>
</tr>
<tr>
<td>MHV</td>
<td>1328</td>
<td>-2.47</td>
<td>340</td>
<td>-1.29</td>
<td>82</td>
<td>-1.17</td>
</tr>
<tr>
<td>PEDV</td>
<td>1079</td>
<td>-2.63</td>
<td>274</td>
<td>-1.65</td>
<td>79</td>
<td>0.05</td>
</tr>
<tr>
<td>TGV</td>
<td>1180</td>
<td>-1.75</td>
<td>306</td>
<td>-1.48</td>
<td>85</td>
<td>-0.49</td>
</tr>
<tr>
<td>RUV</td>
<td>610</td>
<td>0.23</td>
<td>167</td>
<td>-0.40</td>
<td>68</td>
<td>2.72</td>
</tr>
<tr>
<td>EAV</td>
<td>479</td>
<td>-2.25</td>
<td>145</td>
<td>0.91</td>
<td>36</td>
<td>0.30</td>
</tr>
<tr>
<td>RV</td>
<td>407</td>
<td>-0.43</td>
<td>102</td>
<td>-0.75</td>
<td>38</td>
<td>1.71</td>
</tr>
<tr>
<td>HIV-1</td>
<td>347</td>
<td>-0.60</td>
<td>89</td>
<td>-0.21</td>
<td>34</td>
<td>2.42</td>
</tr>
</tbody>
</table>
Observation
1. Length 4 palindromes are under-represented across the coronavirus family.
2. Length 6 palindromes are most under-represented in SARS.

Conjecture for a possible biological explanation:
Avoidance of short palindromes might have a protective effect on the coronavirus genomes against the immune system of the host cells.
A long palindrome in SARS

TCTTTAACAAGCTTTGTTAAAGA

Positions: 25962-25983 (22 bases)

- Longest palindrome found in all 7 coronavirus genomes.
- The next longest palindrome in SARS is 14 bases long.
- Found in the overlapping region of two open reading frames designated X1 and X2 by Rota et al. (2003), or orf 3 and orf 4 by Marra et al. (2003). We are currently investigating whether this long palindrome is involved in the mechanisms for frame-shifting in these overlapping orf’s.
Acknowledgments

Collaborators
David Chew (National University of Singapore)
Kwok Pui Choi (National University of Singapore)
Hans Heidner (University of Texas at San Antonio)

Funding Support
NIH S06GM08194-23 and S06GM08194-24
NSF DUE9981104
Singapore BMRC 01/21/19/140