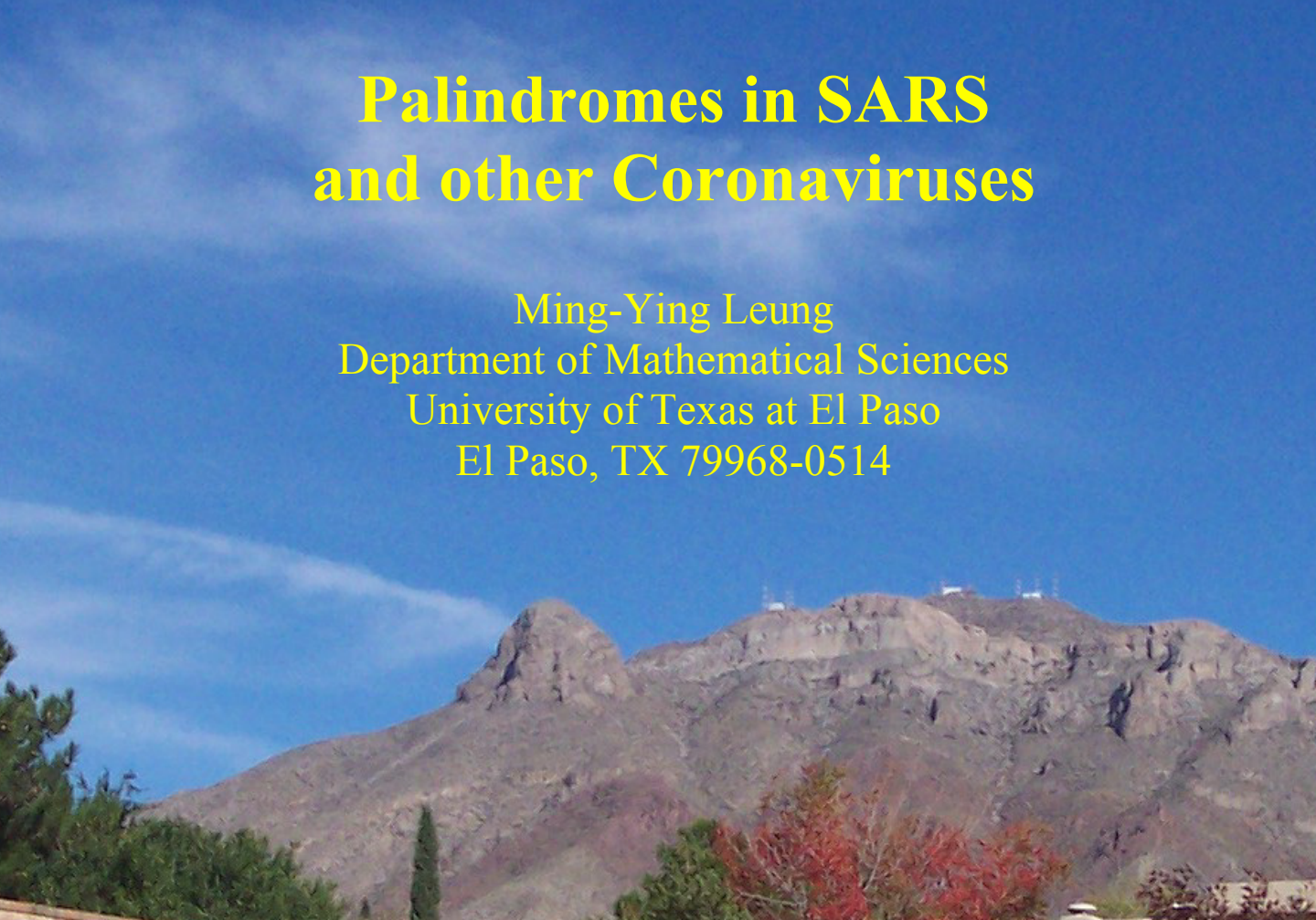
Palindromes in SARS and other Coronaviruses

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

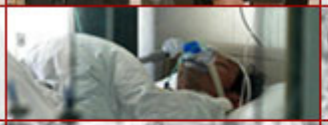


Outline:

- Coronavirus genomes
- Palindromes
- Mean and Variance of palindrome counts
- Under-representation of short palindromes
- A long palindrome in SARS

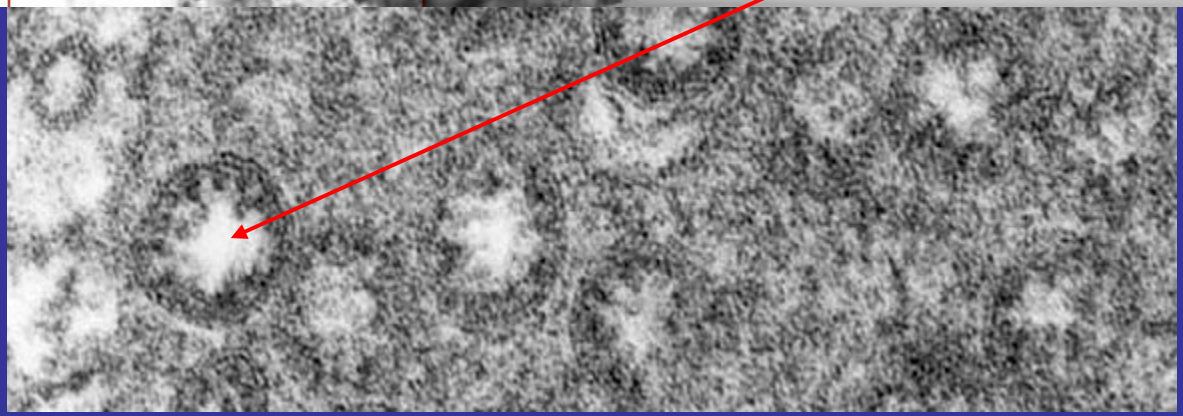
Next

Severe acute respiratory syndrome (SARS) was first detected late last year and is believed to have originated in southern China. The mysterious disease is apparently resistant to standard treatments and has put health authorities worldwide in a spin to find ways to curb the intensifying outbreak.

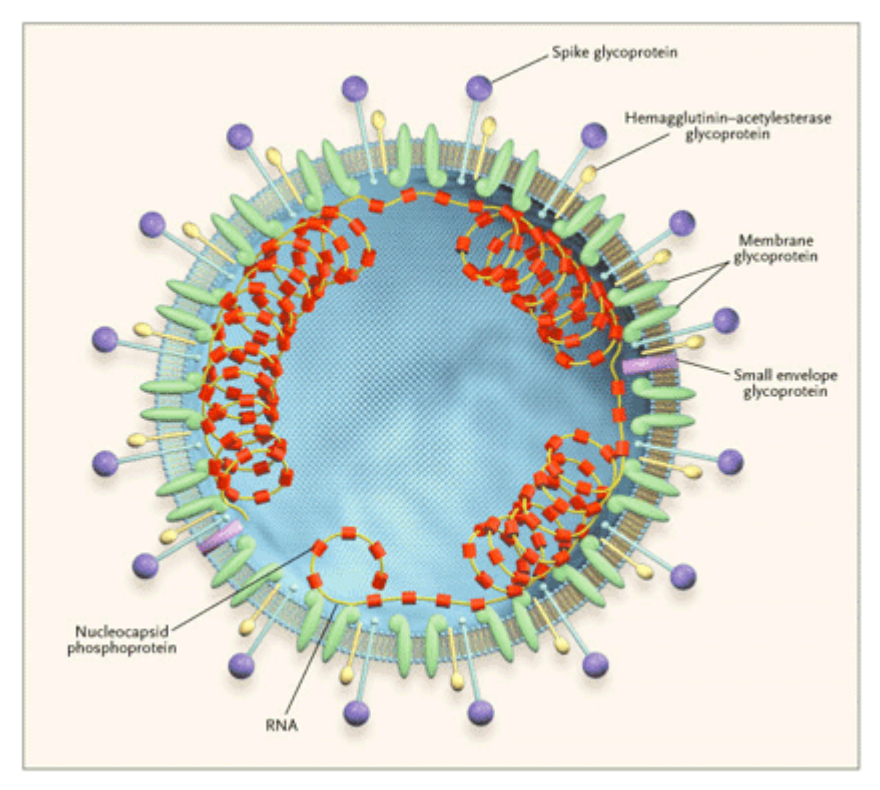


AP Photos/World Health Organistion/CNN Graphics

SARS Viral Particles



SARS Virus

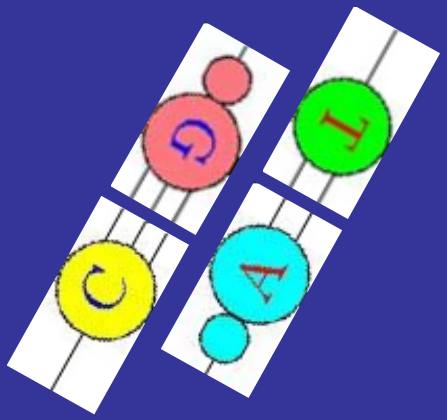

Spike glycoprotein
Hemagglutinin–acetylesterase glycoprotein
Membrane glycoprotein
Small envelope glycoprotein
Nucleocapsid phosphoprotein
RNA

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.

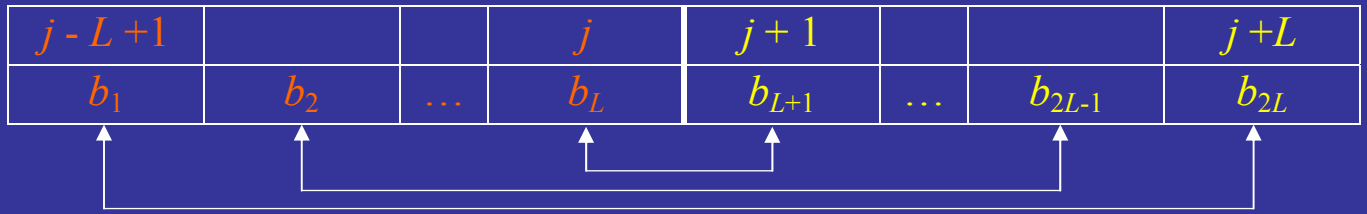


Palindrome: A string of nucleotide bases that reads the same as its reverse complement. A palindrome must be even in length.

E.g. A palindrome of length 10.

5’ ….. GCAATATTGC …..3’

Note that for a palindrome of length $2L$, the ith 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.

Palindrome 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

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

$$\sigma_L^2 = \mathrm{var}(X_L) = \sum_{j=L}^{n-L} \mathrm{var}(I_j) + 2\sum_{j=L}^{n-L-1}\sum_{k=j+1}^{n-L} \mathrm{cov}(I_j, I_k)$$

If we let

$$\gamma(0) = P(I_j = 1) \text{ for } L \le j \le n-L$$

$$\gamma(d) = P(I_j = 1, I_{j+d} = 1) \text{ for } 1 \le d \le n-L-j$$

then

$$E(I_j) = \gamma(0)$$

$$\mathrm{var}(I_j) = \gamma(0)(1-\gamma(0))$$

$$\mathrm{cov}(I_j, I_{j+d}) = \gamma(d) - \gamma(0)^2$$

Mean and variance of palindrome counts (cont'd)

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

$$\sigma_L^2 = \mathrm{var}(X_L)$$

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

$$= (n-2L+1)\gamma(0)\big(1-\gamma(0)\big)$$

$$+ 2\sum_{d=1}^{n-2L} \big(n-2L+1-d\big)\big[\gamma(d)-\gamma(0)^2\big]$$

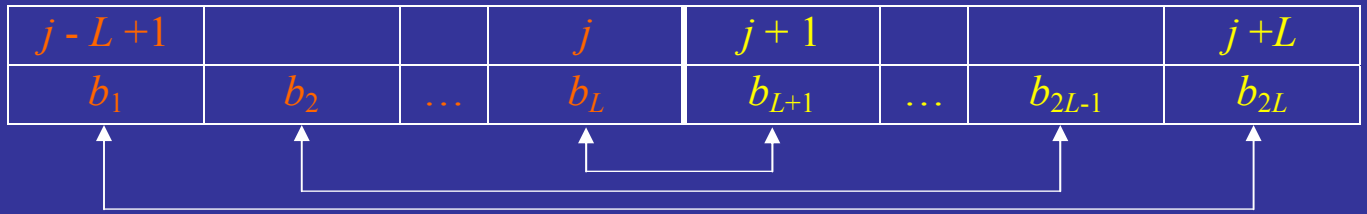
How to find the γ's?

Under a Markov sequence model, Chew *et al.* (2004, to appear in *INFORMS Journal of Computing*) have obtained computable formulas for the γ'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,.

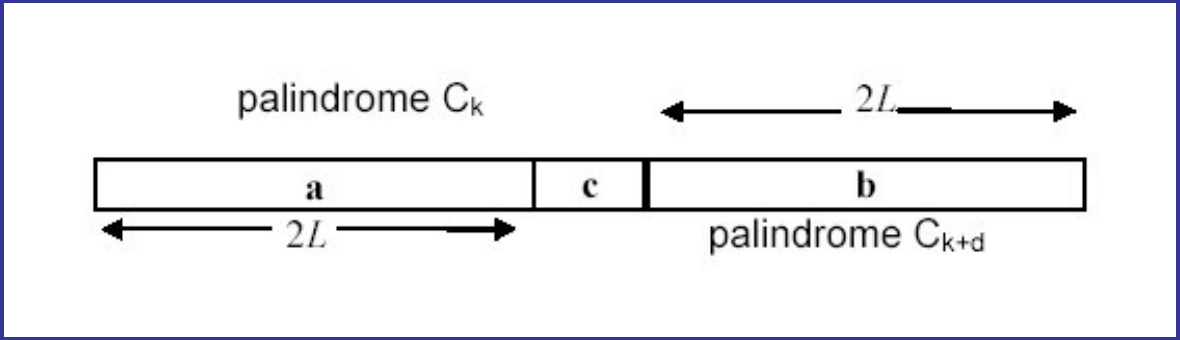
Finding $\gamma(0)$ for the i.i.d. sequence model

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

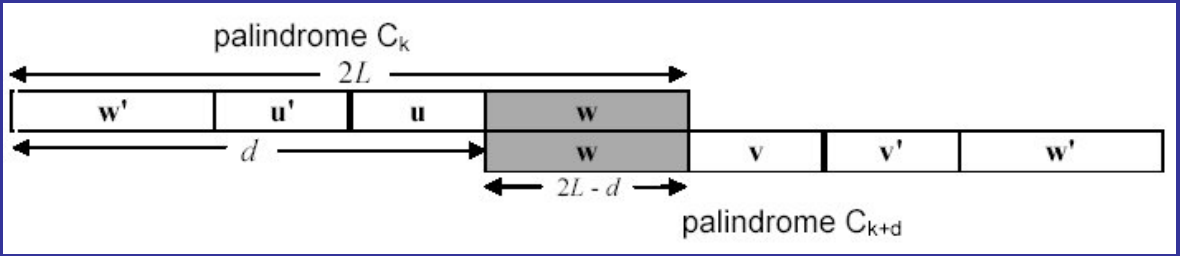


Finding $\gamma(d)$ for the i.i.d. sequence model:

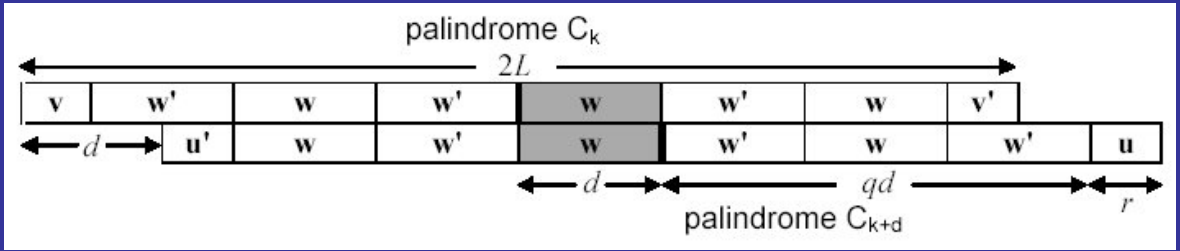
Case 1: $d \geq 2L$



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



Case 3: $1 \leq d < L$

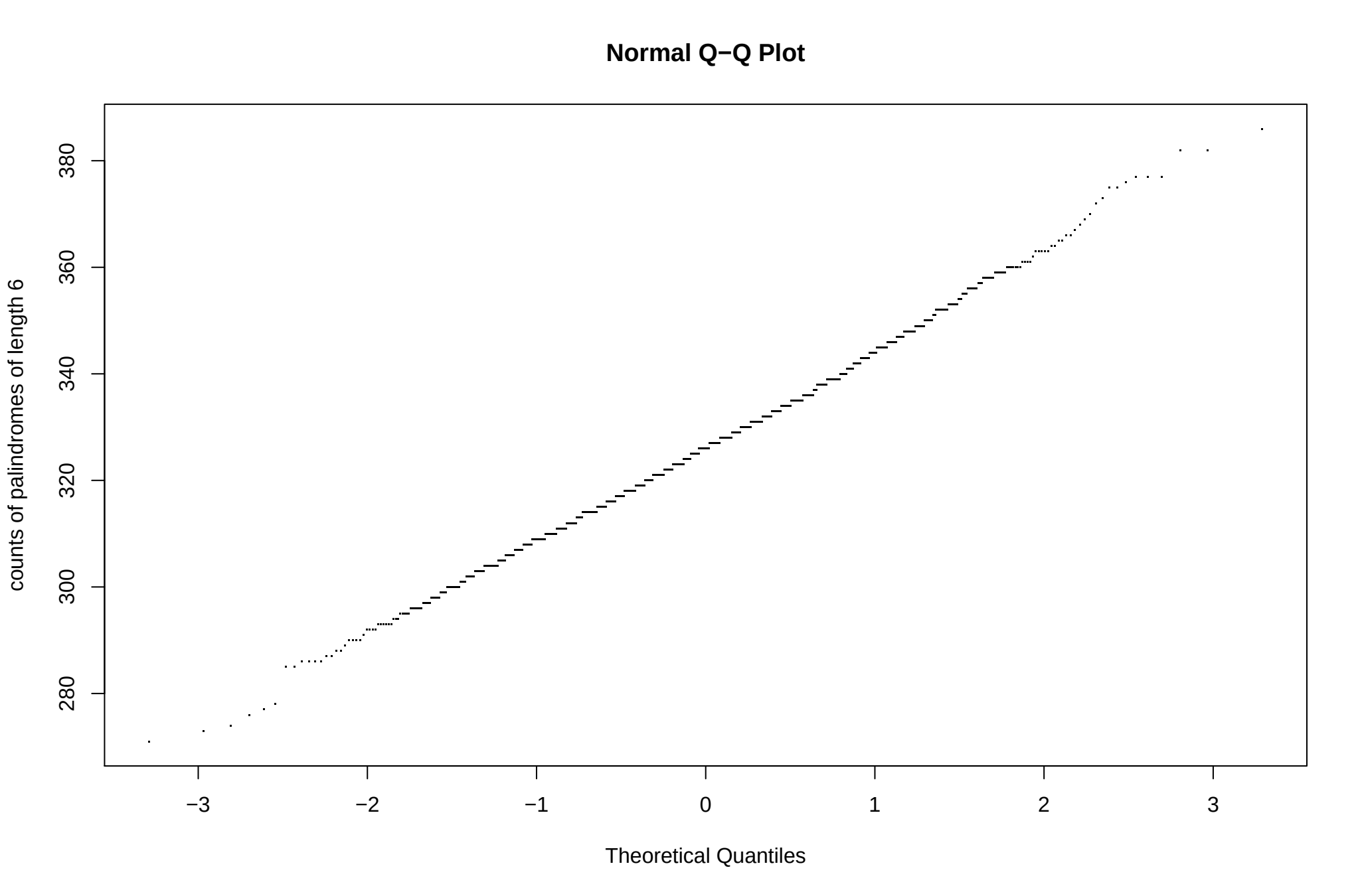


The *z*-score

If μ and σ 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
280
300
320
340
360
380
−3
−2
−1
0
1
2
3
Theoretical Quantiles

z-Scores for Counts of Palindromes of Length 4 or Longer

Virus	*Counts*	*μ(σ)*	*z-score*
SARS	1554	1687.6 (40.3)	-3.32
AIBV	1578	1675.3 (38.2)	-2.54
BCoV	1886	2007.5 (45.5)	-2.67
HCoV	1451	1567.6 (37.0)	-3.15
MHV	1793	1911.3 (41.4)	-2.86
PEDV	1457	1578.8 (38.3)	-3.18
TGV	1610	1695.6 (38.9)	-2.20
RUV	868	845.6 (28.3)	0.79
EAV	672	710.4 (25.8)	-1.49
RV	559	564.3 (23.0)	-0.23
HIV-1	475	480.2 (21.9)	-0.24

All the *z*-scores of the coronaviruses are below -1.645, the 5^{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,… 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

Virus	Length 4		Length 6		Length 8	
Name	Counts	*z*-score	Counts	*z*-score	Counts	*z*-score
SARS	1144	-2.96	284	-2.41	90	0.37
AIBV	1142	-2.48	320	-0.39	91	0.42
BCoV	1360	-3.13	389	-0.07	98	-0.55
HCoV	1054	-2.69	287	-1.18	82	-0.08
MHV	1328	-2.47	340	-1.29	82	-1.17
PEDV	1079	-2.63	274	-1.65	79	0.05
TGV	1180	-1.75	306	-1.48	85	-0.49
RUV	610	0.23	167	-0.40	68	2.72
EAV	479	-2.25	145	0.91	36	0.30
RV	407	-0.43	102	-0.75	38	1.71
HIV-1	347	-0.60	89	-0.21	34	2.42

z-Scores for Palindromes of Various Exact Lengths

Virus	Length 4		Length 6		Length 8	
Name	Counts	*z*-score	Counts	*z*-score	Counts	*z*-score
SARS	1144	-2.96	284	-2.41	90	0.37
AIBV	1142	-2.48	320	-0.39	91	0.42
BCoV	1360	-3.13	389	-0.07	98	-0.55
HCoV	1054	-2.69	287	-1.18	82	-0.08
MHV	1328	-2.47	340	-1.29	82	-1.17
PEDV	1079	-2.63	274	-1.65	79	0.05
TGV	1180	-1.75	306	-1.48	85	-0.49
RUV	610	0.23	167	-0.40	68	2.72
EAV	479	-2.25	145	0.91	36	0.30
RV	407	-0.43	102	-0.75	38	1.71
HIV-1	347	-0.60	89	-0.21	34	2.42

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

TCTTTAACAAGCTTGTTAAAGA

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