Searching for unusual clusters of palindromes and close inversions in the SARS virus genome

June 4, 2003

Kwok-Pui Choi
Department of Mathematics, NUS
Department of Statistics and Applied Probability

joint work with
Louis H Y Chen, NUS
David Chew, NUS
Ming-Ying Leung, UTSA

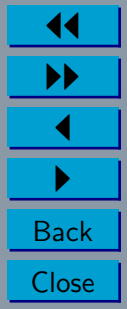


Back
Close

Acknowledgement:

- David Chew for technical assistance in the preparation of this talk.

- helpful discussion with
  - —Louxin Zhang and Hock-Peng Chan
  - —some researchers in the Genome Institute of Singapore, particularly, PK.



Outline of the talk

I. Introduction

II. A tiny bit of biology

III. Palindromes and close Inversions

IV. Measuring clusters of palindromes and close inversions in the SARS genome

V. Which clusters *surprise* us?

—some mathematical and statistical analysis

VI. Some observations

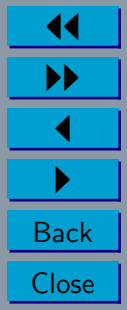
I. Introduction

The SARS virus genome is

- a single stranded RNA
- positive stranded
- 29,711 bases
- `AT` rich:

$$(\texttt{A, C, G, T}) = (0.285, 0.200, 0.208, 0.307)$$

The following slide is taken from the SARS genome,
an isolate (SIN2774), one of the primary contacts from the index case
Basepairs from 20,001 to 23,500



GACCAGCACAAGCTAGCGTCAATGGAGTCACATTAATTGGAGAATCAGTAAAAACACAGTTTAACTACTTTAAGAAAGTAGACGGCATTATTCAACAGTT
GCCTGAAACCTACTTTACTCAGAGCAGAGACTTAGAGGATTTTAAGCCCAGATCACAAATGGAAACTGACTTTCTCGAGCTCGCTATGGATGAATTCATA
CAGCGATATAAGCTCGAGGGCTATGCCTTCGAACACATCGTTTATGGAGATTTCAGTCATGGACAACTTGGCGGTCTTCATTTAATGATAGGCTTAGCCA
AGCGCTCACAAGATTCACCACTTAAATTAGAGGATTTTATCCCTATGGACAGCACAGTGAAAAATTACTTCATAACAGATGCGCAAACAGGTTCATCAAA
ATGTGTGTGTTCTGTGATTGATCTTTTACTTGATGACTTTGTCGAGATAATAAAGTCACAAGATTTGTCAGTGATTTCAAAAGTGGTCAAGGTTACAATT
GACTATGCTGAAATTTCATTCATGCTTTGGTGTAAGGATGGACATGTTGAAACCTTCTACCCAAAACTACAAGCAAGTCAAGCGTGGCAACCAGGTGTTG
CGATGCCTAACTTGTACAAGATGCAAAGAATGCTTCTTGAAAAGTGTGACCTTCAGAATTATGGTGAAAATGCTGTTATACCAAAAGGAATAATGATGAA
TGTCGCAAAGTATACTCAACTGTGTCAATACTTAAATACACTTACTTTAGCTGTACCCTACAACATGAGAGTTATTCACTTTGGTGCTGGCTCTGATAAA
GGAGTTGCACCAGGTACAGCTGTGCTCAGACAATGGTTGCCAACTGGCACACTACTTGTCGATTCAGATCTTAATGACTTCGTCTCCGACGCAGATTCTA
CTTTAATTGGAGACTGTGCAACAGTACATACGGCTAATAAATGGGACCTTATTATTAGCGATATGTATGACCCTAGGACCAAACATGTGACAAAAGAGAA
TGACTCTAAAGAAGGGTTTTTCACTTATCTGTGTGGATTTATAAAGCAAAAACTAGCCCTGGGTGGTTCTATAGCTGTAAAGATAACAGAGCATTCTTGG
AATGCTGACCTTTACAAGCTTATGGGCCATTTCTCATGGTGGACAGCTTTGTTACAAATGTAAATGCATCATCATCGGAAGCATTTTTAATTGGGGCTA
ACTATCTTGGCAAGCCGAAGGAACAAATTGATGGCTATACCATGCATGCTAACTACATTTTCTGGAGGAACACAAATCCTATCCAGTTGTCTTCCTATTC
ACTCTTTGACATGAGCAAATTTCCTCTTAAATTAAGAGGAACTGCTGTAATGTCTCTTAAGGAGAATCAAATCAATGATATGATTTATTCTCTTCTGGAA
AAAGGTAGGCTTATCATTAGAGAAAACAACAGAGTTGTGGTTTCAAGTGATATTCTTGTTAACAACTAAACGAACATGTTTATTTTCTTATTATTTCTTA
CTCTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCAAGCTCCTAATTACACTCAACATACTTCATCTATGAGGGGGGTTTACTA
TCCTGATGAAATTTTTAGATCAGACACTCTTTATTTAACTCAGGATTTATTTCTTCCATTTTATTCTAATGTTACAGGGTTTCATACTATTAATCATACG
TTTGGCAACCCTGTCATACCTTTTAAGGATGGTATTTATTTTGCTGCCACAGAGAAATCAAATGTTGTCCGTGGTTGGGTTTTTGGTTCTACCATGAACA
ACAAGTCACAGTCGGTGATTATTATTAACAATTCTACTAATGTTGTTATACGAGCATGTAACTTTGAATTGTGTGACAACCCTTTCTTTGCTGTTTCTAA
ACCCATGGGTACACAGACACATACTATGATATTCGATAATGCATTTAATTGCACTTTCGAGTACATATCTGATGCCTTTTCGCTTGATGTTTCAGAAAAG
TCAGGTAATTTTAAACACTTACGAGAGTTTGTGTTTAAAAATAAAGATGGGTTTCTCTATGTTTATAAGGGCTATCAACCTATAGATGTAGTTCGTGATC
TACCTTCTGGTTTTAACACTTTGAAACCTATTTTTAAGTTGCCTCTTGGTATTAACATTACAAATTTTAGAGCCATTCTTACAGCCTTTTCACCTGCTCA
AGACATTTGGGGCACGTCAGCTGCAGCCTATTTTGTTGGCTATTTAAAGCCAACTACATTTATGCTCAAGTATGATGAAAATGGTACAATCACAGATGCT
GTTGATTGTTCTCAAAATCCACTTGCTGAACTCAAATGCTCTGTTAAGAGCTTTGAGATTGACAAAGGAATTTACCAGACCTCTAATTTCAGGGTTGTTC
CCTCAGGAGATGTTGTGAGATTCCCTAATATTACAAACTTGTGTCCTTTTGGAGAGGTTTTTAATGCTACTAAATTCCCTTCTGTCTATGCATGGGAGAG
AAAAAAAATTTCTAATTGTGTTGCTGATTACTCTGTGCTCTACAACTCAACATTTTTTTCAACCTTTAAGTGCTATGGCGTTTCTGCCACTAAGTTGAAT
GATCTTTGCTTCTCCAATGTCTATGCAGATTCTTTTGTAGTCAAGGGAGATGATGTAAGACAAATAGCGCCAGGACAAACTGGTGTTATTGCTGATTATA
ATTATAAATTGCCAGATGATTTCATGGGTTGTGTCCTTGCTTGGAATACTAGGAACATTGATGCTACTTCAACTGGTAATTATAATTATAAATATAGGTA
TCTTAGACATGGCAAGCTTAGGCCCTTTGAGAGAGACATATCTAATGTGCCTTTCTCCCCTGATGGCAAACCTTGCACCCCACCTGCTCTTAATTGTTAT
TGGCCATTAAATGATTATGGTTTTTACACCACTACTGGCATTGGCTACCAACCTTACAGAGTTGTAGTACTTTCTTTTGAACTTTTAAATGCACCGGCCA
CGGTTTGTGGACCAAAATTATCCACTGACCTTATTAAGAACCAGTGTGTCAATTTTAATTTTAATGGACTCACTGGTACTGGTGTGTTAACTCCTTCTTC
AAAGAGATTTCAACCATTTCAACAATTTGGCCGTGATGTTTCTGATTTCACTGATTCCGTTCGAGATCCTAAAACATCTGAAATATTAGACATTTCACCT
TGCTCTTTTGGGGGTGTAAGTGTAATTACACCTGGAACAAATGCTTCATCTGAAGTTGCTGTTCTATATCAAGATGTTAACTGCACTGATGTTTCTACAG
CAATTCATGCAGATCAACTCACACCAGCTTGGCGCATATATTCTACTGGAAACAATGTATTCCAGACTCAAGCAGGCTGTCTTATAGGAGCTGAGCATGT
CGACACTTCTTATGAGTGCGACATTCCTATTGGAGCTGGCATTTGTGCTAGTTACCATACAGTTTCTTTATTACGTAGTACTAGCCAAAAATCTATTGTG

- Entire genome will take us 10 slides
- What can be learned from this sequence?
- Illustrated in Sean Eddy's web page





Back
Close

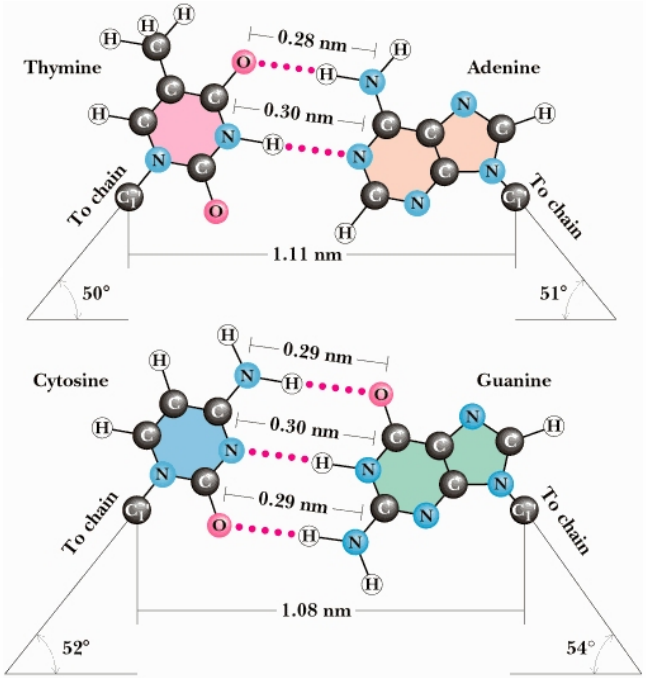
```
   1 GATCCTTGTAGATTTTGAATTTGAAGTTTTTTCTCATTCCAAAACTCTGTGATCTGAAAT 60
  61 AAAATGTCTCAAAAAsean eddy's home page.TATTTATCAGTTATGGTTTTCAA 120
 121 AATTTTCTGACATACCGTTTTGCTTCTTTTTTTCTCATCTTCTTCAAATATCAATTGTGA 180
 181 TAATCTGAcontact info.AATTTCTTTTCCTTTTTCTTTTTCCAACAACTCCAGTGA 240
 241 GAACTTTTGAATATCTTCAAGTGACTTCACCACATCAGAAGGTGTCAACGATCTTGTGAG 300
 301 AACATCGAATGAAGATAlab home page.GTTACAGTTTTTCCTCCGACAATTCCTGA 360
 361 TTTACGAACATCTTCTTCAAGCATTCTACAGATTTCTTGATGCTCTTCTAGGAGGATGTT 420
 421 GAAATCCGAAGTTGobtaining reprints & lab publications.GGATCCGA 480
 481 TTCAGATGGACGACCTGGCAGTCCGAGAGCCGTTCGAAGGAAAGATTCTTGTGAGAGAGG 540
 541 CGTGAAhmmer.AGGGTATAGGTTCTTCTTCAGATTCATATCACCAACAGTTTGAATATC 600
 601 CATTGCTTTCAGTTGAGCTTCGCATACACGACCAATTCCTCCAACCTAAAAAATTATCTA 660
 661 GGTAAAACTAGAAGGTTATGCTTTAATAGTCTCACCTTACGAATCGGTAAATpfam.AAA 720
 721 AACTCCATAATCGCGTTTTTATCATTTTCTAACACATATTGACCATTTGGTTTGTTCAAA 780
 781 TCAGAAinfernal.GCGAGCATAAAGTTAGATGCGATTCCAGCAGAACATGTTAATCCC 840
 841 GTGAGTTGTTCAACTCGAAATCGAATTTCTCGAACAGCCTCCTCTCGTCCAGTTCCGAAC 900
 901 TCCACATGGTCGTAGTAGATTTTCCGCGATTTTTCGCATTTTGGACAGATrfam.CTTCG 960
 961 ATTTTCAAGTCTTCCAAAGTATTTTCATTCTCGTCGAAACGGGGTAACCAACATGGACAA 1020
1021 TCbiological sequence analysis.CTAGTAAGCAAATAGTTTTTTGTTAATAA 1080
1081 TCAAATCTAAATCACTAACTTTTTTCTGTATTACTTGCCACATAGTCTGTCAAATCTATA 1140
1141 AATGCbio5495/bme537 computational molecular biology.TTTGTGAA 1200
1201 AATTGGCGACTGACTTTAGTGTATTTAGGGTAATTTCCTGGAACAATCGTTAGACTCGGA 1260
1261 CAAAGTTTATTTGAGATGAAGthe fine print.CGGACTCCAAAACGGCGAGCCAAG 1320
1321 TAGTTGGATGTGCTCTGAAAGAATAGATTTAAAGCTTTTCCGAAATCGAAAATTTACTTT 1380
1381 TCAAATTGAGTTAGGTGCTTACCAGCATTGCCGATGAGCCTACmy links.AACTGTTC 1440
1441 TCAGTGCAGGATTATCTCTCATTTCAACTGCGGCAAAATAAGCATCCATATCTATACAAA 1500
1501 CACAGTCTCTTGATAAATCTCTAGATGATTCCAGTTTCATCTCAAGATTCTCCATCTGAA 1560
```

Back
Close

Many approaches to learn from this sequence: Via

- homology search
- phylogenetic analysis
- under/over-represented of some specific words in the genome, such as `CTAG`, a palindrome itself.
- our approach is via a general class of word patterns: palindromes and close inversions
- etc ...

- 4 DNA/RNA nucleotides: `A`, `C`, `G` and `T/U`



- Helices formed by complementary base pairing:

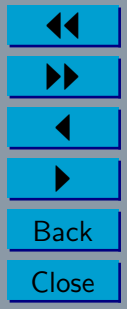
Watson-Crick pairs:

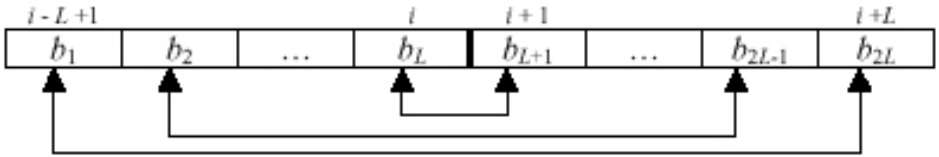
$$\texttt{A} \leftrightarrow \texttt{T/U} \quad \text{(2 hydrogen bonds)}$$
$$\texttt{C} \leftrightarrow \texttt{G} \quad \text{(3 hydrogen bonds)}$$

Both pairs are almost coplanar; almost always stacked onto other base pairs in an RNA structure.

Wobble pairs:

$$\texttt{G} \leftrightarrow \texttt{U} \quad \text{(some extent)}$$
$$\texttt{G} \leftrightarrow \texttt{A} \quad \text{(more rarely)}$$





palindrome of length $2L$

Example:

...GTAAC|GTTAC...

is a **palindrome** of length 10.

Example:

...GTAAC|nnnn|GTTAC...

is a **close inversion** of stem length 5 and loop length 4.


GACCAGCACAAGCTAGCGTCAATGGAGTCACATTAATTGGAGAATCAGTAAAAACACAGTTTAACTACTTTAAGAAAGTAGACGGCATTATTCAACAGTT
GCCTGAAACCTACTTTACTCAGAGCAGAGACTTAGAGGATTTTAAGCCCAGATCACAAATGGAAACTGACTTTCTCGAGCTCGCTATGGATGAATTCATA
CAGCGATATAAGCTCGAGGGCTATGCCTTCGAACACATCGTTTATGGAGATTTCAGTCATGGACAACTTGGCGGTCTTCATTTAATGATAGGCTTAGCCA
AGCGCTCACAAGATTCACCACTTAAATTAGAGGATTTTATCCCTATGGACAGCACAGTGAAAAATTACTTCATAACAGATGCGCAAACAGGTTCATCAAA
ATGTGTGTGTTCTGTGATTGATCTTTTACTTGATGACTTTGTCGAGATAATAAAGTCACAAGATTTGTCAGTGATTTCAAAAGTGGTCAAGGTTACAATT
GACTATGCTGAAATTTCATTCATGCTTTGGTGTAAGGATGGACATGTTGAAACCTTCTACCCAAAACTACAAGCAAGTCAAGCGTGGCAACCAGGTGTTG
CGATGCCTAACTTGTACAAGATGCAAAGAATGCTTCTTGAAAAGTGTGACCTTCAGAATTATGGTGAAAATGCTGTTATACCAAAAGGAATAATGATGAA
TGTCGCAAAGTATACTCAACTGTGTCAATACTTAAATACACTTACTTTAGCTGTACCCTACAACATGAGAGTTATTCACTTTGGTGCTGGCTCTGATAAA
GGAGTTGCACCAGGTACAGCTGTGCTCAGACAATGGTTGCCAACTGGCACACTACTTGTCGATTCAGATCTTAATGACTTCGTCTCCGACGCAGATTCTA
CTTTAATTGGAGACTGTGCAACAGTACATACGGCTAATAAATGGGACCTTATTATTAGCGATATGTATGACCCTAGGACCAAACATGTGACAAAAGAGAA
TGACTCTAAAGAAGGGTTTTTCACTTATCTGTGTGGATTTATAAAGCAAAAACTAGCCCTGGGTGGTTCTATAGCTGTAAAGATAACAGAGCATTCTTGG
AATGCTGACCTTTACAAGCTTATGGGCCATTTCTCATGGTGGACAGCTTTGTTACAAATGTAAATGCATCATCATCGGAAGCATTTTTAATTGGGGCTA
ACTATCTTGGCAAGCCGAAGGAACAAATTGATGGCTATACCATGCATGCTAACTACATTTTCTGGAGGAACACAAATCCTATCCAGTTGTCTTCCTATTC
ACTCTTTGACATGAGCAAATTTCCTCTTAAATTAAGAGGAACTGCTGTAATGTCTCTTAAGGAGAATCAAATCAATGATATGATTTATTCTCTTCTGGAA
AAAGGTAGGCTTATCATTAGAGAAAACAACAGAGTTGTGGTTTCAAGTGATATTCTTGTTAACAACTAAACGAACATGTTTATTTTCTTATTATTTCTTA
CTCTCACTAGTGGTAGTGACCTTGACCGGTGCACCACTTTTGATGATGTTCAAGCTCCTAATTACACTCAACATACTTCATCTATGAGGGGGGTTTACTA
TCCTGATGAAATTTTTAGATCAGACACTCTTTATTTAACTCAGGATTTATTTCTTCCATTTTATTCTAATGTTACAGGGTTTCATACTATTAATCATACG
TTTGGCAACCCTGTCATACCTTTTAAGGATGGTATTTATTTTGCTGCCACAGAGAAATCAAATGTTGTCCGTGGTTGGGTTTTTGGTTCTACCATGAACA
ACAAGTCACAGTCGGTGATTATTATTAACAATTCTACTAATGTTGTTATACGAGCATGTAACTTTGAATTGTGTGACAACCCTTTCTTTGCTGTTTCTAA
ACCCATGGGTACACAGACACATACTATGATATTCGATAATGCATTTAATTGCACTTTCGAGTACATATCTGATGCCTTTTCGCTTGATGTTTCAGAAAAG
TCAGGTAATTTTAAACACTTACGAGAGTTTGTGTTTAAAAATAAAGATGGGTTTCTCTATGTTTATAAGGGCTATCAACCTATAGATGTAGTTCGTGATC
TACCTTCTGGTTTTAACACTTTGAAACCTATTTTTAAGTTGCCTCTTGGTATTAACATTACAAATTTTAGAGCCATTCTTACAGCCTTTTCACCTGCTCA
AGACATTTGGGGCACGTCAGCTGCAGCCTATTTTGTTGGCTATTTAAAGCCAACTACATTTATGCTCAAGTATGATGAAAATGGTACAATCACAGATGCT
GTTGATTGTTCTCAAAATCCACTTGCTGAACTCAAATGCTCTGTTAAGAGCTTTGAGATTGACAAAGGAATTTACCAGACCTCTAATTTCAGGGTTGTTC
CCTCAGGAGATGTTGTGAGATTCCCTAATATTACAAACTTGTGTCCTTTTGGAGAGGTTTTTAATGCTACTAAATTCCCTTCTGTCTATGCATGGGAGAG
AAAAAAAATTTCTAATTGTGTTGCTGATTACTCTGTGCTCTACAACTCAACATTTTTTTCAACCTTTAAGTGCTATGGCGTTTCTGCCACTAAGTTGAAT
GATCTTTGCTTCTCCAATGTCTATGCAGATTCTTTTGTAGTCAAGGGAGATGATGTAAGACAAATAGCGCCAGGACAAACTGGTGTTATTGCTGATTATA
ATTATAAATTGCCAGATGATTTCATGGGTTGTGTCCTTGCTTGGAATACTAGGAACATTGATGCTACTTCAACTGGTAATTATAATTATAAATATAGGTA
TCTTAGACATGGCAAGCTTAGGCCCTTTGAGAGAGACATATCTAATGTGCCTTTCTCCCCTGATGGCAAACCTTGCACCCCACCTGCTCTTAATTGTTAT
TGGCCATTAAATGATTATGGTTTTTACACCACTACTGGCATTGGCTACCAACCTTACAGAGTTGTAGTACTTTCTTTTGAACTTTTAAATGCACCGGCCA
CGGTTTGTGGACCAAAATTATCCACTGACCTTATTAAGAACCAGTGTGTCAATTTTAATTTTAATGGACTCACTGGTACTGGTGTGTTAACTCCTTCTTC
AAAGAGATTTCAACCATTTCAACAATTTGGCCGTGATGTTTCTGATTTCACTGATTCCGTTCGAGATCCTAAAACATCTGAAATATTAGACATTTCACCT
TGCTCTTTTGGGGGTGTAAGTGTAATTACACCTGGAACAAATGCTTCATCTGAAGTTGCTGTTCTATATCAAGATGTTAACTGCACTGATGTTTCTACAG
CAATTCATGCAGATCAACTCACACCAGCTTGGCGCATATATTCTACTGGAAACAATGTATTCCAGACTCAAGCAGGCTGTCTTATAGGAGCTGAGCATGT
CGACACTTCTTATGAGTGCGACATTCCTATTGGAGCTGGCATTTGTGCTAGTTACCATACAGTTTCTTTATTACGTAGTACTAGCCAAAAATCTATTGTG

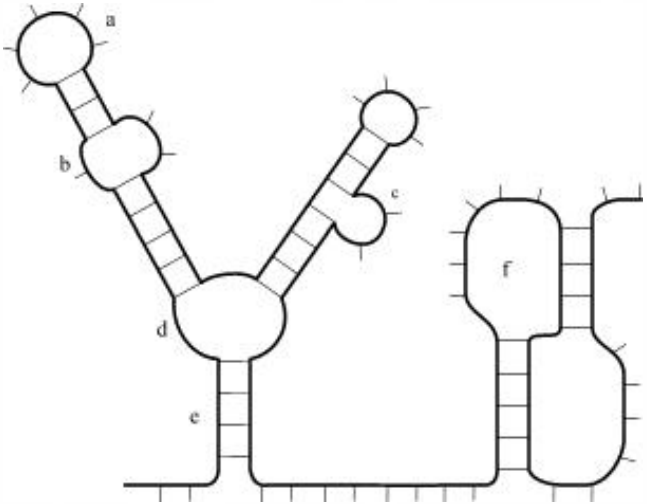
Back
Close

Palindromes are involved in a variety of biological processes:

- recognition sites for bacterial restriction enzymes mostly palindromic
- crucial roles in gene regulation and DNA replication processes
- appears related with DNA-protein binding
- ...

Close inversion:

- provides the possibility of forming hairpins and many other secondary structures
- Hairpins increase/decrease the rate of translation of mRNA dependent on its location.



by Peter De Rijk

stems: contiguous stacked base pairs are called stems;

loops: single stranded substrings of nucleotides bounded by base pairs in a stem;

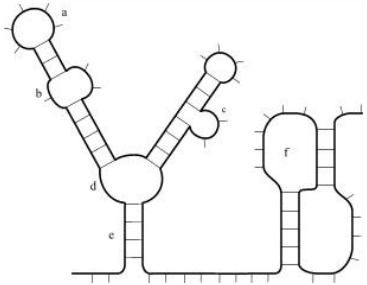
hairpins (or called stem-loops): simple substructures consisting of a simple stem and loop;

bulges: single stranded bases occurring within one side of a stem;

bulge loops: single stranded bases occurring within both sides of a stem;

multi-branched loops: loop from which 3 or more stems radiate;

psuedoknots: non-nested base pairs



by Peter De Rijk

Regions rich in palindromes/close inversions are interesting biologically



Back
Close

IV. Clusters of palindromes/close inversions

- How do we measure the clustering of close inversions or palindromes?

- We propose two classes of measures:
  - A. Window based
    1. Count
    2. Affine score
    3. Log-prob score
  - B. Maximal segmental score and Excursion plot
    –adaptation of Karlin's approach

- Will apply these measures on the SARS genome sequence analysis

A1. Count

Advantages of this scoring scheme:

- It is simple and intuitive.

• By a duality relation, it is related to r-scan statistic, which is window free.

• Some successes:

– In Leung, Schachtel and Yu (1994), predicting the origin of replication/enhancer element in HCMV

– In Leung, Choi, Xia and Chen (2002), regions rich in palindromes:

· overlaps transcriptional regulators in EHV1

· contains the origins of replications in EBV

· close proximity of origins of replications in BHV1

However, it fails to predict origin of replication in HSV1

–it has a very *long* palindrome of length 144

A2. Affine score–to capture not just counts, but their lengths

- Choose a window size
- Identify all the palindromes (maximally extended) in the window
- score for each palindrome is defined as:

$$\frac{\text{length of one stem}}{L}$$

- score for the window is the

$$\text{total of scores}$$

A3. Log-prob score–to capture length as well as the fact that

- Not all palindromes are 'equal'
  –in the sense of their likelihoods to appear in the genome.

Some are rarer than others

- For example, in a AT rich genome, the palindrome

`AGTTA|TAACT`

occurs more frequently than

`GACCG|CGGTC`

- we define the score of a palindrome with numbers of `A` bases, N_A, and so on ...

$$\sum_{a \in \mathcal{A}} (-\log p_a) \times N_a$$

where p_A is the proportion of A in the genome.

- Can see it as a weighted sum of the basepair compositions of the palindromes–generalizing the affine score.

B. Maximal segmental score and Excursion plot

$$\cdots \boxed{k^{\text{th}} \text{ base}} \cdots$$

We score each basepair, we score it as

$$X_k := \begin{cases} c & \text{if } k\text{th base inside a palindrome of length } \geq 2L \\ -1 & \text{otherwise} \end{cases}$$

where c is to be suitably chosen such that $EX_k < 0$.

Define $E_0 = 0$, and inductively, for $1 \leq k \leq n$,

$$E_k = \max\{E_{k-1} + X_k, 0\}.$$

- Plot $\{(k, E_k) : 1 \leq k \leq n\}$–excursion plot.

- The excursion plot captures
  - −length of palindromes;
  - −proximity of neighboring palindromes;
  - −length of palindromes;
  - −window size free.

Choice of c. Let

$$\psi := P(X_k = 1).$$

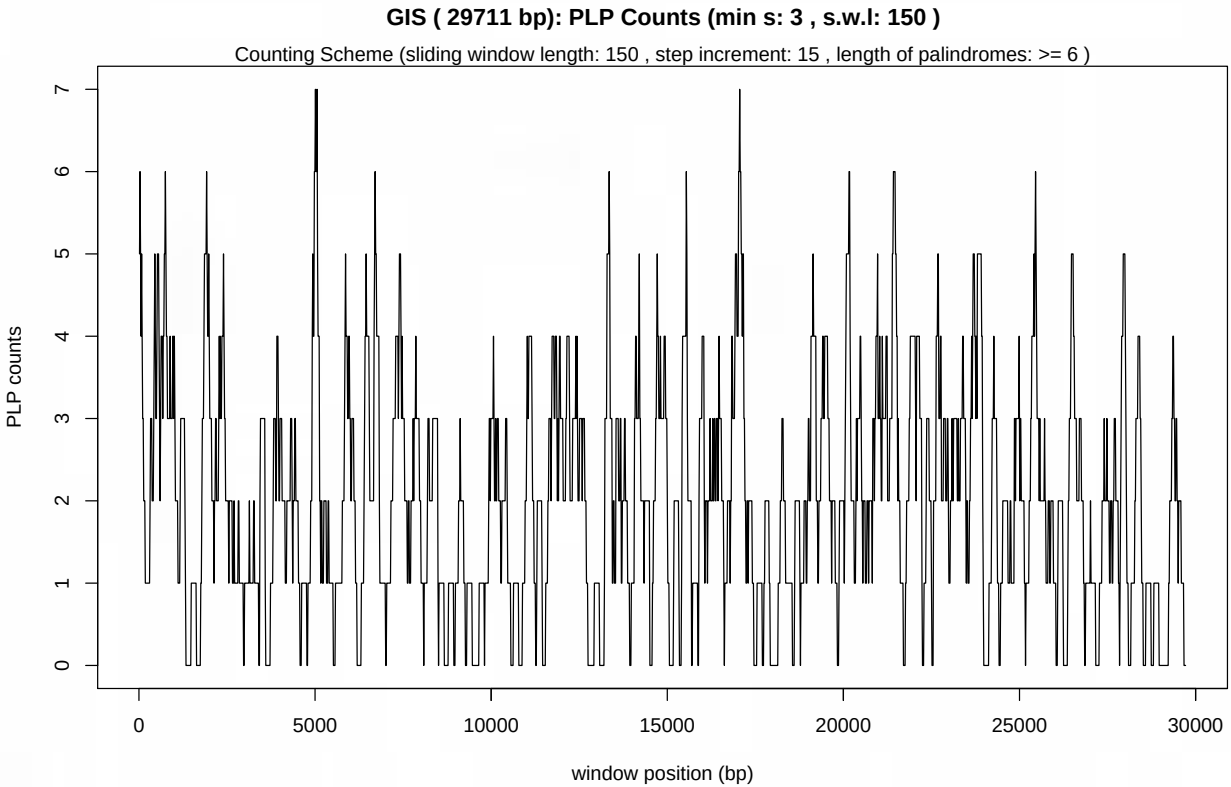
$EX_k < 0$ imposes

$$1 \leq c < 1/\psi - 1.$$

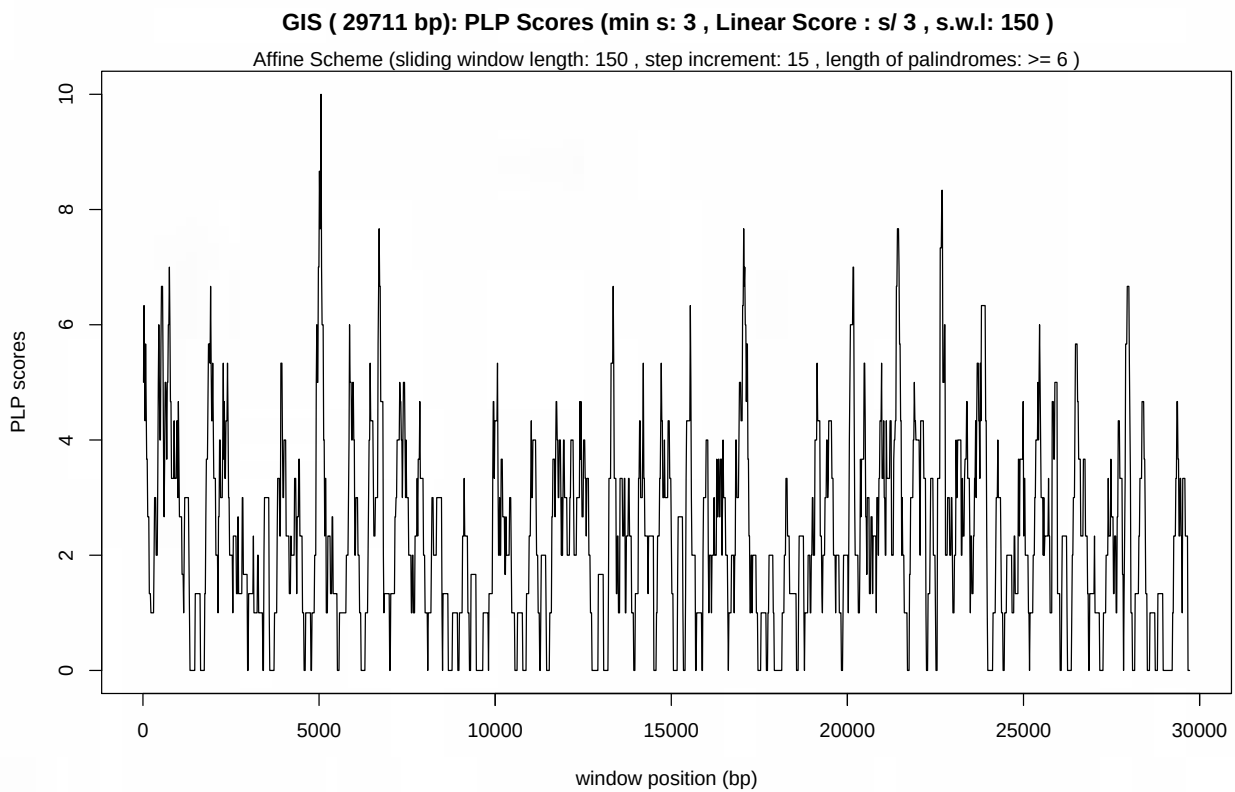
Let p_U be an upper bound (given by a Bonferroni-type inequality), we *propose* to choose

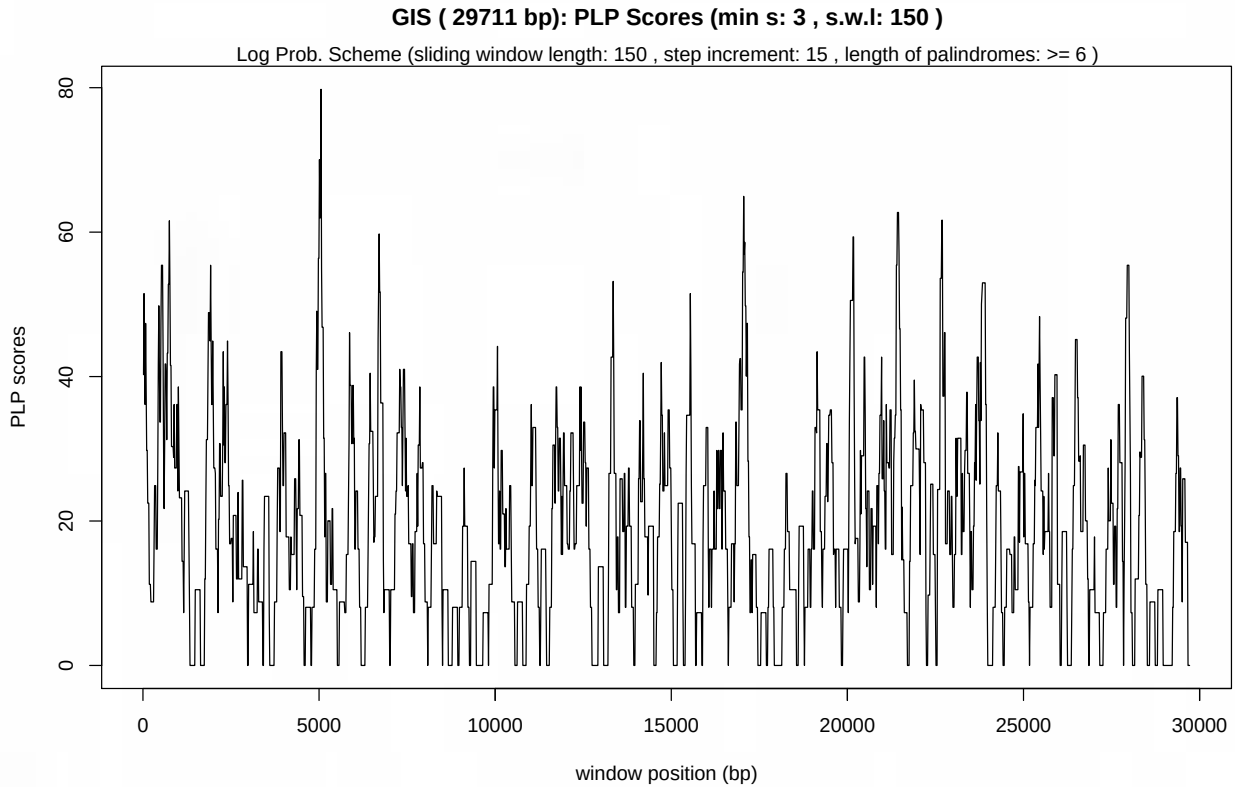
$$c := 1/p_U - 1.$$

- p_U is computable from the base composition in the SARS genome.

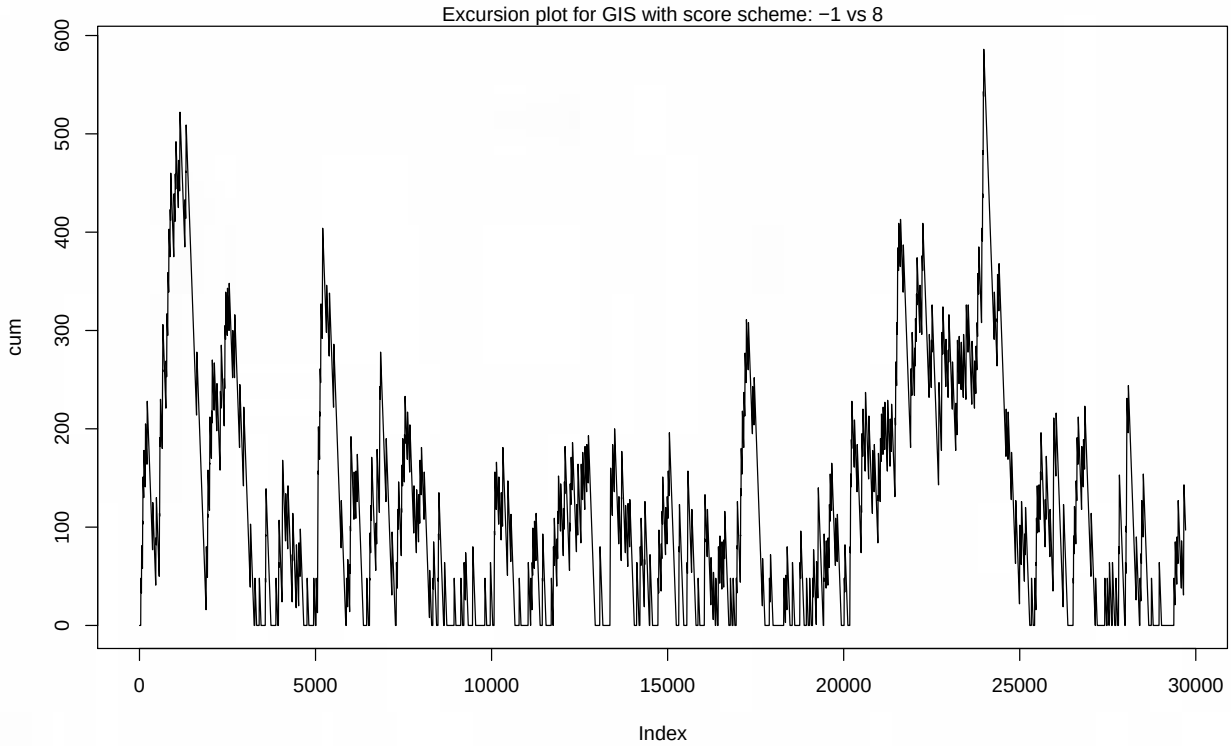

GIS (29711 bp): PLP Counts (min s: 3 , s.w.l: 150)
Counting Scheme (sliding window length: 150 , step increment: 15 , length of palindromes: >= 6)
PLP counts
0
1
2
3
4
5
6
7
0
5000
10000
15000
20000
25000
30000
window position (bp)

Back
Close

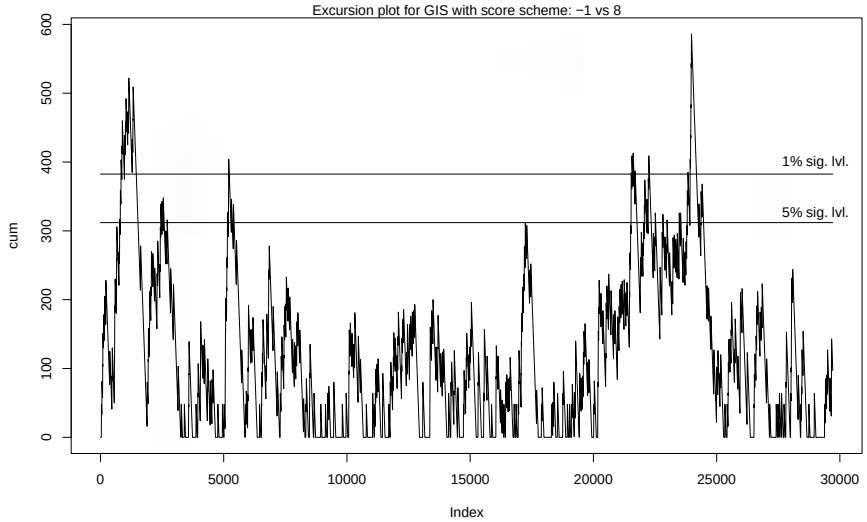

GIS (29711 bp): PLP Scores (min s: 3 , Linear Score : s/ 3 , s.w.l: 150)
Affine Scheme (sliding window length: 150 , step increment: 15 , length of palindromes: >= 6)
PLP scores
0
2
4
6
8
10
0
5000
10000
15000
20000
25000
30000
window position (bp)


GIS (29711 bp): PLP Scores (min s: 3 , s.w.l: 150)
Log Prob. Scheme (sliding window length: 150 , step increment: 15 , length of palindromes: >= 6)
PLP scores
0
20
40
60
80
0
5000
10000
15000
20000
25000
30000
window position (bp)

Back
Close


Excursion plot for GIS with score scheme: −1 vs 8
cum
0
100
200
300
400
500
600
0
5000
10000
15000
20000
25000
30000
Index

V. Which clusters *surprise* us?



- These two cut-offs at 1% and 5% are based on Karlin's work:
- He assumed X_k's being independent and identically distributed.
- Need to work out our case: X_k's are dependent.

Back
Close

For class A measures of clustering:

• Chen-Stein's method for the Poisson process approximation showed that

• if the basepairs are IID, then occurrences of palindromes can be approximated by a Poisson process.

Leung, Choi, Xia and Chen (2002).

In progress:

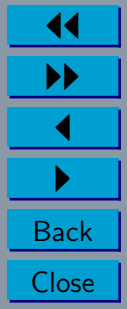
For affine/Log-prob cases: We proceed as

• Given the window size and there are m such windows covering the entire sequence.

• For the kth window: score is S_k.
• For large c, want to compute

$$P(\max_{1\le k\le m} S_k \ge c).$$

• Due to overlap, compound Poisson approximation (Barbour, Chen and Loh, 1992) should be more useful:



Let

$$I_k := I(S_k \geq c), \quad 1 \leq k \leq m$$

and

$$W = I_1 + \cdots + I_m.$$

Want to approximate

$$\begin{aligned} P(\max_{1 \leq k \leq m} S_k \geq c) &= P(W \geq 1) \\ &= 1 - P(W = 0) \\ &\approx 1 - e^{-\lambda} \end{aligned}$$

which reduces, for λ small, the computation of

$$\lambda = \sum_{k=1}^{m} E(I_k / Y_k)$$

where

$$Y_k := \sum_{r \in A_k} I_r.$$

- Excursion plot suggests 2 interesting regions:

  Positions 1 to 1,000, and

  Positions 20,190 to 23,996.

- 3 interesting sites of palindrome clusters in SARS genome

  1. Segment 1: 5,100 to 5,200
  2. Segment 2: 22,700 to 22,800

     `TTATAA|TTATAA`

     This palindrome occurs twice at positions 22,696 and 22,780 in a stretch of 96 bases.

3. Segment 3: 25,800 to 26,100

   It has a very long perfect palindrome of length 22!

   TCTTTAACAAG|CTTGTTAAAGA

   located from 25,946 to 25,967.

   The next longest palindromes are of only lengths 14!

No.	start	end	s	lstem	rstem
72	5056	5067	6	CACTAC	GTAGTG
78	5191	5204	7	ATAACAA	TTGTTAT
87	5996	6007	6	TGATTT	AAATCA
97	6738	6749	6	AATTCT	AGAATT
134	10071	10082	6	ACAGTA	TACTGT
315	22696	22707	6	TTATAA	TTATAA
316	22777	22788	6	TAATTA	TAATTA
317	22780	22791	6	TTATAA	TTATAA
324	23220	23231	6	GTGTAA	TTACAC
337	23916	23927	6	GCTTCA	TGAAGC
391	28032	28045	7	ACCTTCA	TGAAGGT
405	29652	29665	7	TAAAATT	AATTTTA

What I have presented can be best summed up in the Chinese idiom:

抛砖引玉

Thank you very much for your attention!