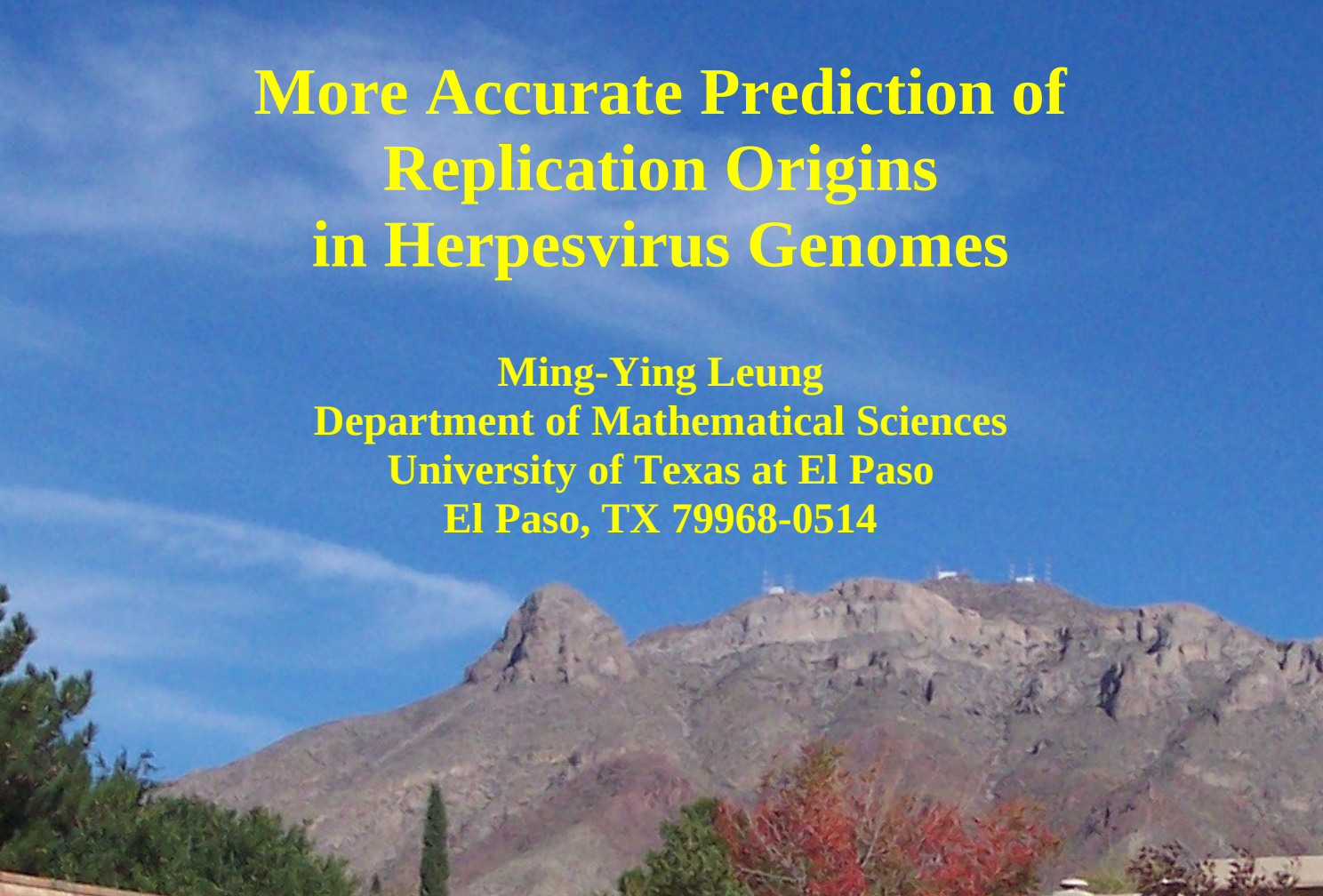
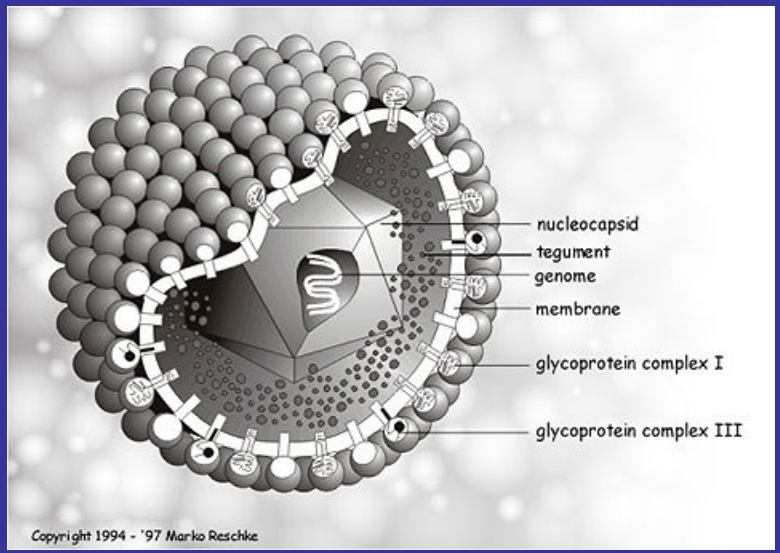
More Accurate Prediction of Replication Origins in Herpesvirus Genomes

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

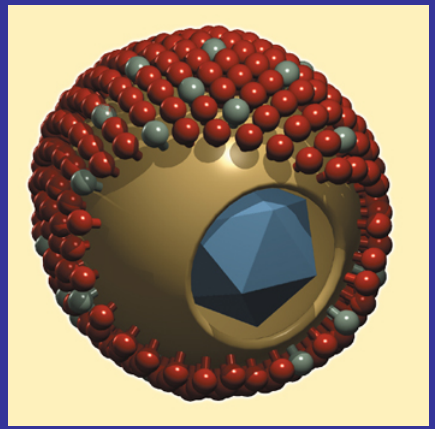


Outline:

- Herpesvirus genomes
- DNA palindromes
- Poisson process approximation of palindrome occurrences



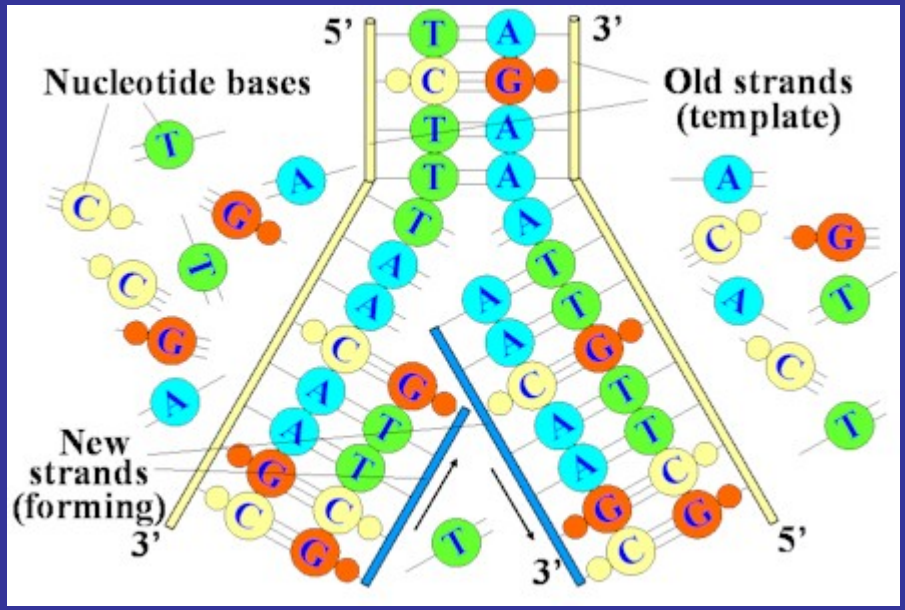
Cytomegalovirus (CMV) Particle



Genome sizes of ~100-250 kbp

Outline (cont'd):

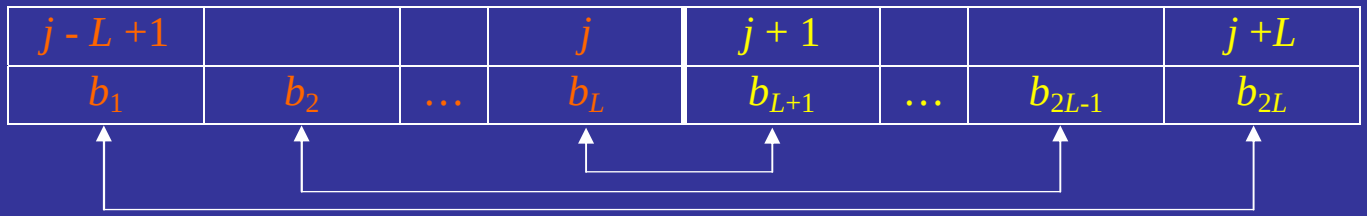
- Prediction of replication origins using scan statistics
- More accurate predictions using scoring schemes



DNA Replication at the Origin (Orilyt)

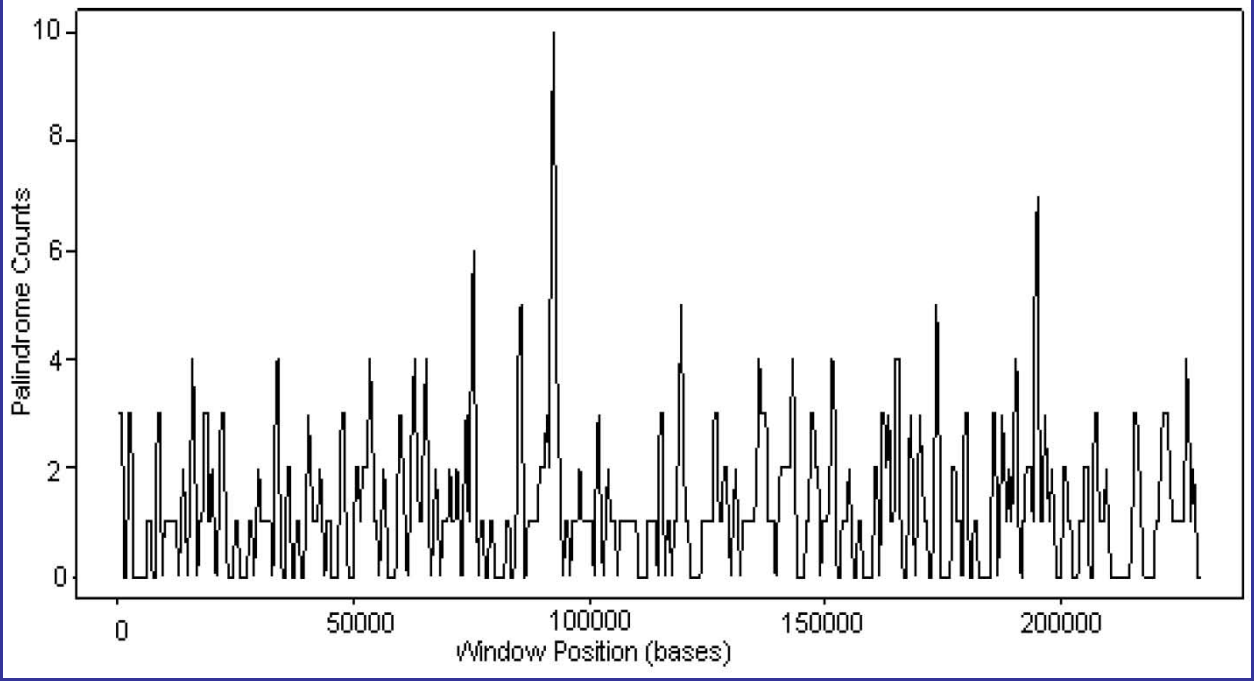
Palindrome: A string of nucleotide bases that reads the same as its reverse complement. A palindrome must be even in length, e.g. palindrome of length 10:

5’ ….. GCAATATTGC …..3’
3’ ….. CGTTATAACG …..5’



We say that a palindrome of length $2L$ occurs at position j when the $(j-i+1)$st and the $(j+i)$th bases are complementary to each other for $i=1,\ldots, L$. In an i.i.d. sequence model this occurs with probability $\left[2\left(p_A p_T + p_C p_G\right)\right]^L$.

Association of Palindromes Clusters with Replication Origins



Poisson process approximation

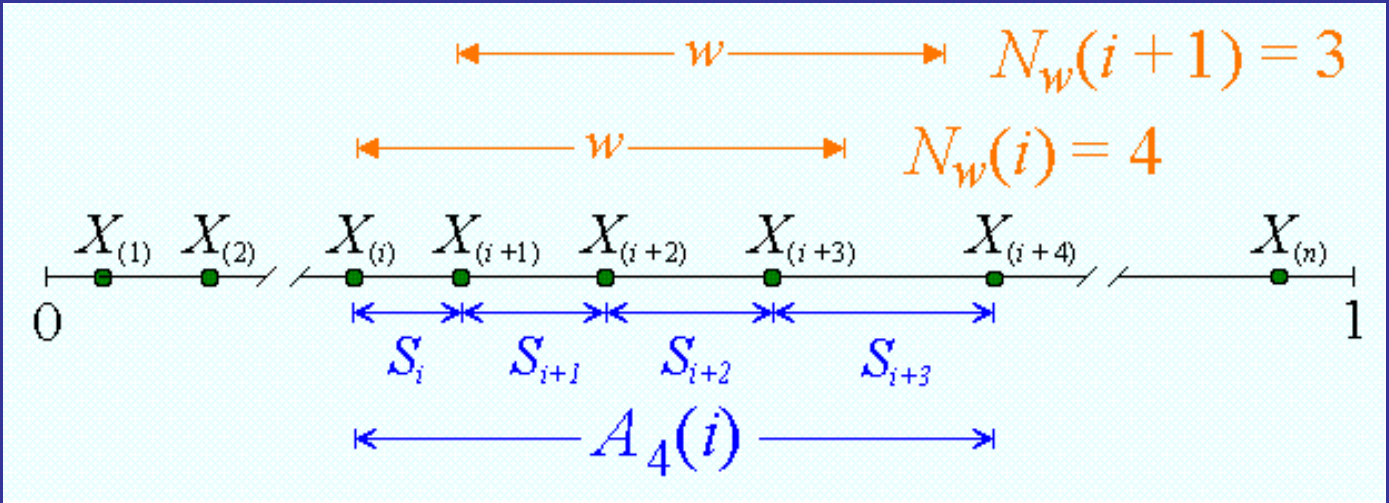
Let Ξ be the process representing the palindrome occurrences on a random nucleotide sequence generated by the i.i.d. model; and Z_λ be the Poisson process with rate λ.

Proposition (Leung *et al.* 2004 *J. Computat. Biol.*)
Assuming $p_A = p_T$, $p_C = p_G$ *and suppose that* $n, L \to \infty$ *in such a way that* $n\theta^L = \lambda$ where $\lambda \geq 1/32$ is a fixed positive constant, then

$$d_2(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) \leq cL\theta^{L/2} \to 0$$

Here d_2 *stands for the Wasserstein distance,* Ξ the palindrome process, and *c is an absolute constant no greater than* 131.

The Scan Statistic



$X_1, X_2, \ldots, X_n \sim$ i.i.d. Uniform (0,1)

$S_i = X_{(i+1)} - X_{(i)} = i^{\text{th}}$ spacing

$A_r(i) = S_i + \ldots + S_{i+r-1}$ = sum of r adjoining spacing

r-Scan Statistic $\quad A_r = \min_i A_r(i)$

Scan Statistics Prediction Results

Genome	Region	Palindrome	Feature
BHV1	77155 - 77168	3	
	102895 - 106948	22	
	113462 - 113636	5	1.75 mu[a] from Ori
	124582 - 124756	5	1.61 mu from Ori
	131268 - 135221	21	
EHV1	115125 - 119094	17	overlaps transcriptional regulator
	144064 - 148033	17	overlaps transcriptional regulator
EHV4	none		
HSV1	none		
HSV2	none		
VZV	none		
EBV	6772 - 11675	19	Contains OriP
	49460 - 54858	25	Contains OriLyt
HCMV	89585 - 94183	19	Contains OriLyt
	195029 - 195268	8	enhancer element

Scan Statistics Prediction Results (Cont'd)

Genome	Region	Palindrome	Feature
HCMV	89585 - 94183	19	Contains OriLyt
	195029 - 195268	8	enhancer element
HHV6	none		
HHV7	120758 -124422	16	
AHV1	113456 - 113759	5	
ATHV3	95350 - 100098	17	
MDV2	93143 - 93243	4	
	109331 - 110590	8	
MDV	none		
CCV1	none		
HVS2	none		

Scoring schemes

Palindrome count score (PCS): a palindrome is given a score 1 when its length is at or above $2L$.

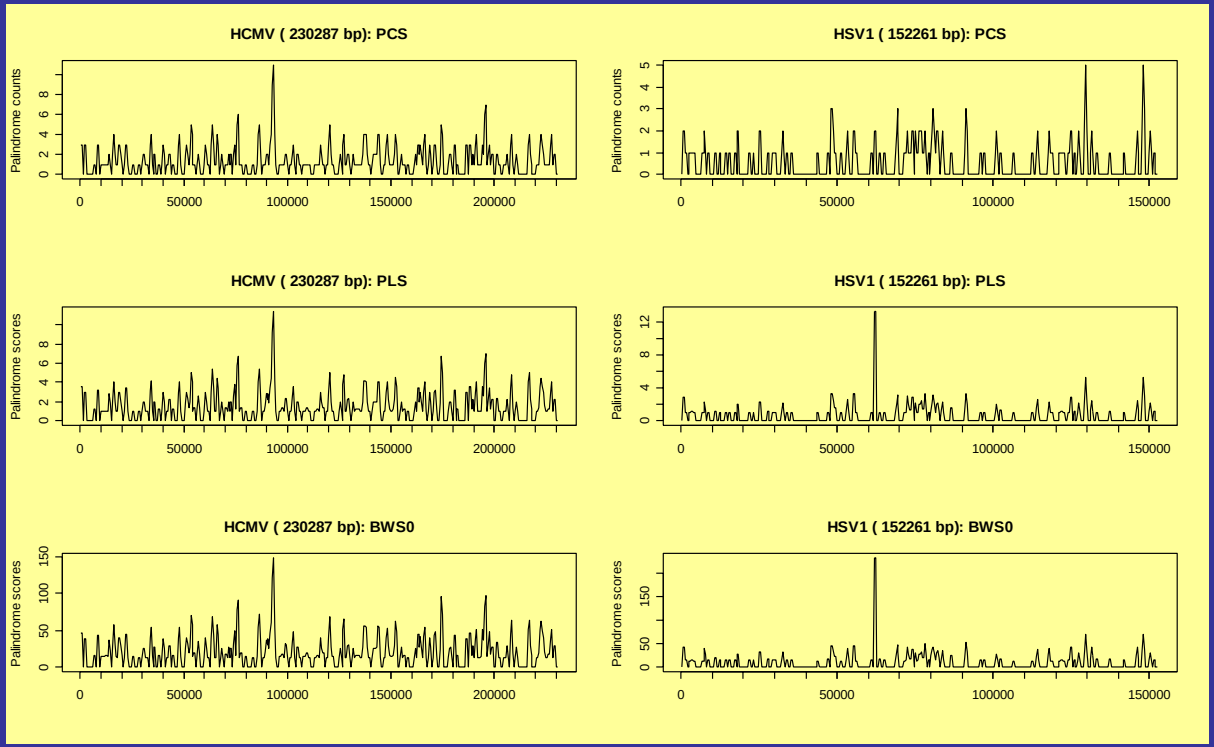
Palindrome length score (PLS): a palindrome of length at least $2L$ is given a score proportional to its length.
E.g., assign a score of s/L for a palindrome of length $2s$.

Base weighted score (BWS): a palindrome of length at least $2L$ is given a score equal to the negative log of the probability of its occurrence.
E.g., under the i.i.d. random sequence model, assign a score of

$$-(2\log p_A + 3\log p_C + 3\log p_G + 2\log p_T)$$

for the palindrome *CACGTACGTG*, where p_A, p_C, p_G, p_T are the percentages of the bases in the genome.

Sliding Window Plots for Various Scoring Schemes



Prediction results

Virus	Known ORIs/ Names			PCS	PLS	BWS
bohv1	111080-111300	(OriS)		1.75mu	1.6mu	1.6mu
	126918-127138	(OriS)		1.61mu	1.8mu	1.8mu
bohv4	97143-98850	(OriLyt)		-	-	-
cehv1	61592-61789	(OriL1)		-	0.1mu	0.1mu
	61795-61992	(OriL2)		-	0.2mu	0.2mu
	132795-132796	(OriS1)		-	0.1mu	0.1mu
	132998-132999	(OriS2)		-	0.002mu	0.002mu
	149425-149426	(OriS2)		-	0.02mu	0.02mu
	149628-149629	(OriS1)		-	0.1mu	0.1mu
cehv7	109627-109646			-	-	-
	118613-118632			-	-	-
ebv	7315-9312	(OriP)		contains ori	0.4mu	0.4mu
	52589-53581	(OriLyt)		contains ori	0.07mu	0.07mu
ehv1	126187-126338			-	-	-
ehv4	73900-73919	(OriL)		-	-	-
	119462-119481	(OriS)		-	-	-
	138568-138587	(OriS)		-	-	-

Prediction results (Cont'd)

Virus	Known ORIs/ Names			PCS	PLS	BWS
hcmv	93201-94646	(OriLyt)		contains ori	0.05mu	0.05mu
hhv6	67617-67993	(OriLyt)		-	-	-
hhv7	66685-67298			-	-	-
hsv1	62475	(OriL)		-	0.1mu	0.1mu
	131999	(OriS)		-	1.4mu	1.4mu
	146235	(OriS)		-	1.4mu	1.4mu
hsv2	62930	(OriL)		-	-	-
	132760	(OriS)		-	-	-
	148981	(OriS)		-	-	-
rcmv	75666-78970	(OriLyt)		overlaps ori	0.6mu	0.6mu
vzv	110087-110350			-	0.1mu	0.1mu
	119547-119810			-	0.2mu	0.2mu

Measures of Prediction Accuracy

$$\text{Sensitivity} = \frac{\text{no. of ORIs that are significant clusters}}{\text{no. of ORIs}}$$

$$\text{Specificity} = \frac{\text{no. of significant clusters that are ORIs}}{\text{no. of significant clusters}}$$

Improved prediction accuracy

	PCS	PLS					PWS				
		1	2	3	4	5	1	2	3	4	5
Sensitivity	0.17	0.28	0.48	0.59	0.66	0.69	0.28	0.48	0.59	0.62	0.66
Specificity	0.24	0.57	0.50	0.40	0.34	0.29	0.57	0.50	0.40	0.32	0.27

Ongoing work:

- Evaluation of statistical significance for the scoring schemes.
- Incorporate other sequence features such as close direct repeats and close inversions.

Acknowledgments

Collaborators

Louis H. Y. Chen (National University of Singapore)
David Chew (National University of Singapore)
Kwok Pui Choi (National University of Singapore)
Aihua Xia (University of Melbourne, Australia)

Funding Support

NIH Grants S06GM08194-23, S06GM08194-24, and 2G12RR008124
NSF DUE9981104
W.M. Keck Center of Computational & Struct. Biol. at Rice University
National Univ. of Singapore ARF Research Grant (R-146-000-013-112)
Singapore BMRC Grants 01/21/19/140 and 01/1/21/19/217