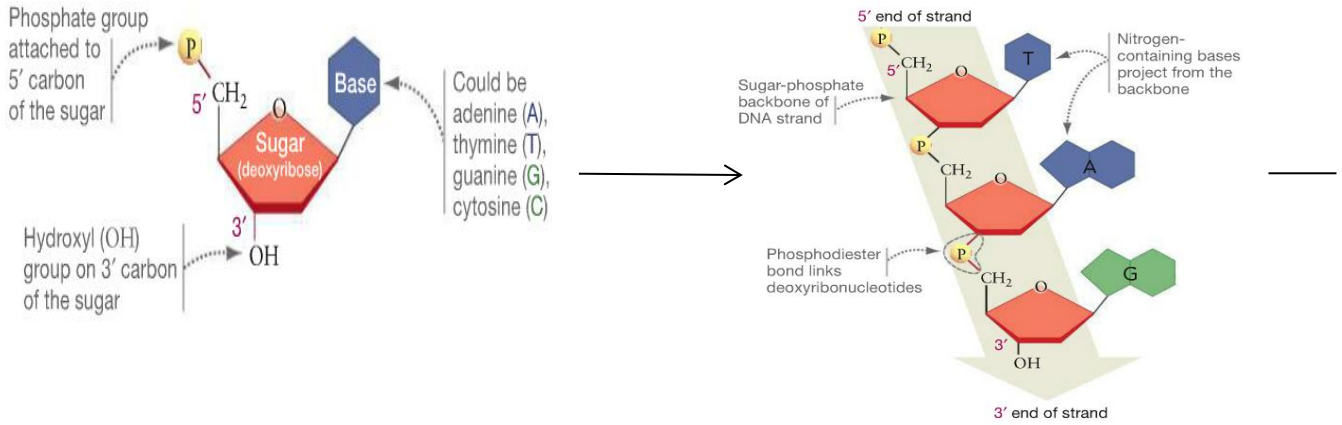
Application of R-Scan Statistic on Viral DNA Analysis

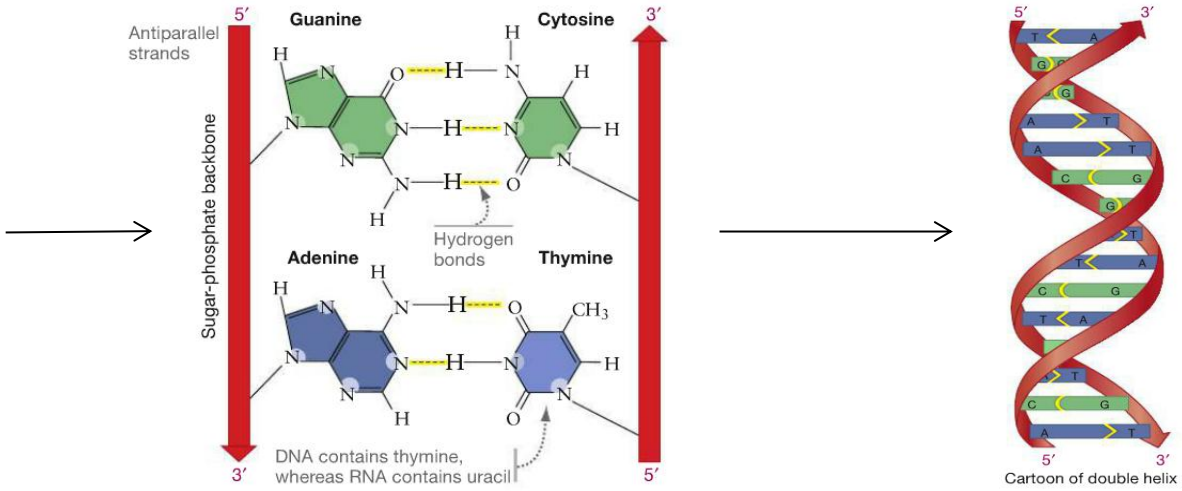
Si Yi Li

5/25/2014

Introduction

Deoxyribonucleic acid, DNA, is a very important character found in every organism. Generally, it is composed by two structures. In its primary structure, a strand called sugar-phosphate backbone is made by deoxyribonucleotides with one of four different nitrogenous bases attached which are adenine (A), thymine (T), guanine (G) and cytosine (C). The organic molecule, deoxyribonucleotide, has a pentagonal ring structure which is made of five carbons and one oxygen (one of the carbons is attached to the ring instead of being part of the ring). In order to specifies each carbon, biologists give each one an order based on the functional group it bonded. The 1' carbon is bonded to one of the four different bases, the 2' carbon is not bonded to any functional group, the 3' carbon is bonded to a hydroxyl group (OH), and the 5' carbon, the one that is not belong to the ring, is bonded to the 4' carbon and a phosphate group. In the secondary structure of DNA, two sugar-phosphate backbones are bonded antiparallel by their four different bases by a chemical bond called hydrogen bond. Because of their chemical properties, adenine and thymine are bonded together by two hydrogen bonds, and guanine and cytosine are bonded together by three hydrogen bonds. Since these two base pairs do not have the same number of hydrogen bonding, the two sugar-phosphate backbones are twisted in order to make the bonding to become possible. This twisting make the DNA to become double helix. Furthermore, during DNA synthesis, the phosphate group of 5' carbon is chemically added to the 3' carbon. Therefore, biologists always read the DNA strand from the base on the 5' end carbon to the base on the 3' end carbon. For example, 5' ATTGGCA 3'. Because DNA is made of two strands, the other strand would be 3' TAACCGT 5'. Even though it starts with a 3' carbon, it has to read from 5' to 3' which means from right to left.





The characteristic of DNA that makes it so important for organism is that it contains all the hereditary information of an organism even for a virus. Virus have a very simple structure compared to cells because they only have two parts which are the DNA of the virus and a protein called plasmid which is used to enclosed the virus DNA. They cannot be considered to be alive because they cannot replicate their own DNA based on their own structure. However, once they parasite to a cell, they can use the cell's materials to synthesize their DNA, and for most of the time, the viral DNA contains harmful information to the host. For example, Human Cytomegalovirus (HCMV) is a virus which infects from 30% to 80% of the human population. It can cause neurological disorders, gastrointestinal disease and pneumonia. Once HCMV infect a person, they enter a state called lytic cycle, or replicative growth. During this state, the virus use the person's cells and materials to replicate massively their own DNA. Because of this state, the infected person does not show any symptoms at the beginning. As the result it is so difficult for medicines to kill the virus at the beginning when there still are not many copy of the DNA of the virus in the host. However, one of the possibility to kill HCMV is to stop their DNA replication or DNA synthesis. Generally, the synthesis of the DNA of virus including HCMV starts at a DNA region called origin of replication. If biologists can expulse these region, it may stop the replication of the virus. It is a DNA region where has a lot of palindromes. A palindrome is a DNA segment which has the same reading from right to left and left to right direction. For example, 5' GAATTC 3' is a palindrome because it has the same reading as 3' CTTAAG 5'. An bacterial enzyme called restriction endonuclease can cut or "destroy" specific palindromes. As the result, if biologists can find the origin of replication of the virus and use specific restriction endonucleases to expulse it, they can stop the viral DNA replication in the host and cure the host. Unfortunately, a viral DNA sequence is usually about thousand hundred base pairs (bp) long. It is almost impossible to analyze a thousand

hundred base pairs long viral DNA sequence and accurately find out the origin of replication base pair by base pair. Therefore, we need some probability and statistic methods to estimate approximately the origin of replication in a DNA sequence.

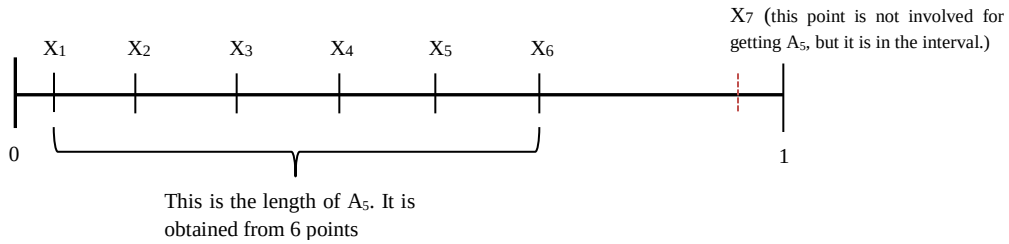
Statistical theory for DNA analysis

If we look at the whole virus DNA sequence and locate all the palindromes at once, palindromes are going to be looked like points on a chain because palindromes are just a small fragments compared to the whole DNA sequence. The origin of replication is a DNA area where has a lot of palindrome, and it is going to be looked like a region on the chain where points are more concentrated than the rest of the chain. However, we do not know what is the significance for a concentration of palindromes in a DNA sequence. Just by looking the whole virus DNA, can a 1000 base pair long DNA area where has 6 palindromes be considered as the origin of replication? Then, how about having 7 palindromes? Therefore the use of probability and statistic calculation can help us to get some stronger evidences for answering these confusing questions. The first part of the calculation is to describe the distribution of palindromes in a viral DNA sequence as the distribution of random points in a chain. Two statistical methods describe the distribution of random points in a chain, and they are called r-scan statistic and traditional scan statistic.

To begin with, let D_i denote the distance between two points X_i and X_{i+1}, so $D_i = X_{i+1} - X_i$, where i = 1, 2 …, N - 1. Now, let a set of points (X_1, X_2 …, X_N) distribute in a unit of interval (0, 1) independently and uniformly. From these information, the r-scan statistic indicates the cumulative lengths of r consecutive distance(s) between two points, D_i, from point X_i to X_{i+r-1} between point X_1 to X_{N-1}. In mathematic expression, if we want r consecutive distance started from point X_i, it is then $A_{r,i} = \sum_{j=i}^{i+r-1} D_{(j)}$ where i = 1, 2 …, N. In more detail, the minimal r-scam (A_r) indicates the minimum lengths for getting r consecutive D_i from points distributed in an interval, and it is $A_r = \min\{A_{r,i}, i = 1, \ldots, N - r\}$. The letter "i" is not included in symbol of the minimal r-scam because it can start at any point to get the minimum length.

In the other hand, the traditional scan statistic S_w indicates the maximum number of points, $Y_t(w)$, in the interval [t, t+w] where $0 < w < 1$ indicated the window length. The mathematical expression is then $S_w = \max_{0 < w < 1} Y_t(w)$. The traditional scan statistic and r-scan statistic are correlated. The event $\{A_{r,} \leq w\}$ shows that the minimum cumulative lengths of r consecutive D_i started at point X_i is at most to

the length of the interval, w. It is equivalent to the event $\{S_w = Y_{X_i}(w) \geq r+1\}$ which shows that the maximum number of points found in the window $[X_i, X_{i+w}]$ is at least r + 1. For example, if the minimum length for getting 5 consecutive D_i is 0.6 in an interval where has a length of 0.9 $\{A_5 = 0.6 < 0.9\}$, this interval has at least 6 maximum points $\{S_{0.9} \geq 5+1\}$because it needs at least 6 points to get 5 consecutive distances between two points.



This equivalent is true for all i where i = 1, 2 …, N, so the duality can be written mathematically as

$$\{S(w) \geq r+1\} = \{A_{r,} \leq w\}$$

This equivalent is very useful to analyze the DNA sequence. If we know the minimum length of r consecutive palindromes in a specific DNA segment (A_r) is less than the length of segment, we can also know the maximum number of palindromes, S(w), is at least r + 1. Due to its convenience, r-scan statistic is more practical to use in DNA sequence analyze.

The second part of statistical calculation is to obtain the probability distribution of random points (palindromes) distributed in a chain (viral DNA sequence). The probability distribution of A_r tells us if having the minimum length of a specific r consecutive D_i in a specific window length, w, is due to random chance or not. Unfortunately, the exact distribution of A_r is not known yet. But with the efforts of statisticians and a long period of work, better and better approximations have been developed. The current best approximation to find out A_r is called Compound Poisson approximation. The probabilities are computed as follow:

$$P\{Ar \leq w\} \approx 1 - \exp\{ -(N - r) \pi (1- p + pr(r + p - rp))\},$$

where $\pi = Q_1$ and $p = Q_1/ Q_2$ with $Q_1 = \sum_{j=r}^{n} b(j;\ N, w)$ and $Q_2 = \sum_{j=r}^{N} (-1)^{j+r} b(j; N, w)$.

Since $\{S(w) \geq r+1\} = \{A_{r,} \leq w\}$, for a particular r, the probability of $\{A_{r,} \leq w\}$ indicates also the probability of $\{S(w) \geq r+1\}$. With these statistical calculations, we can now start on analyzing the virus DNA and find out the origin of replication.

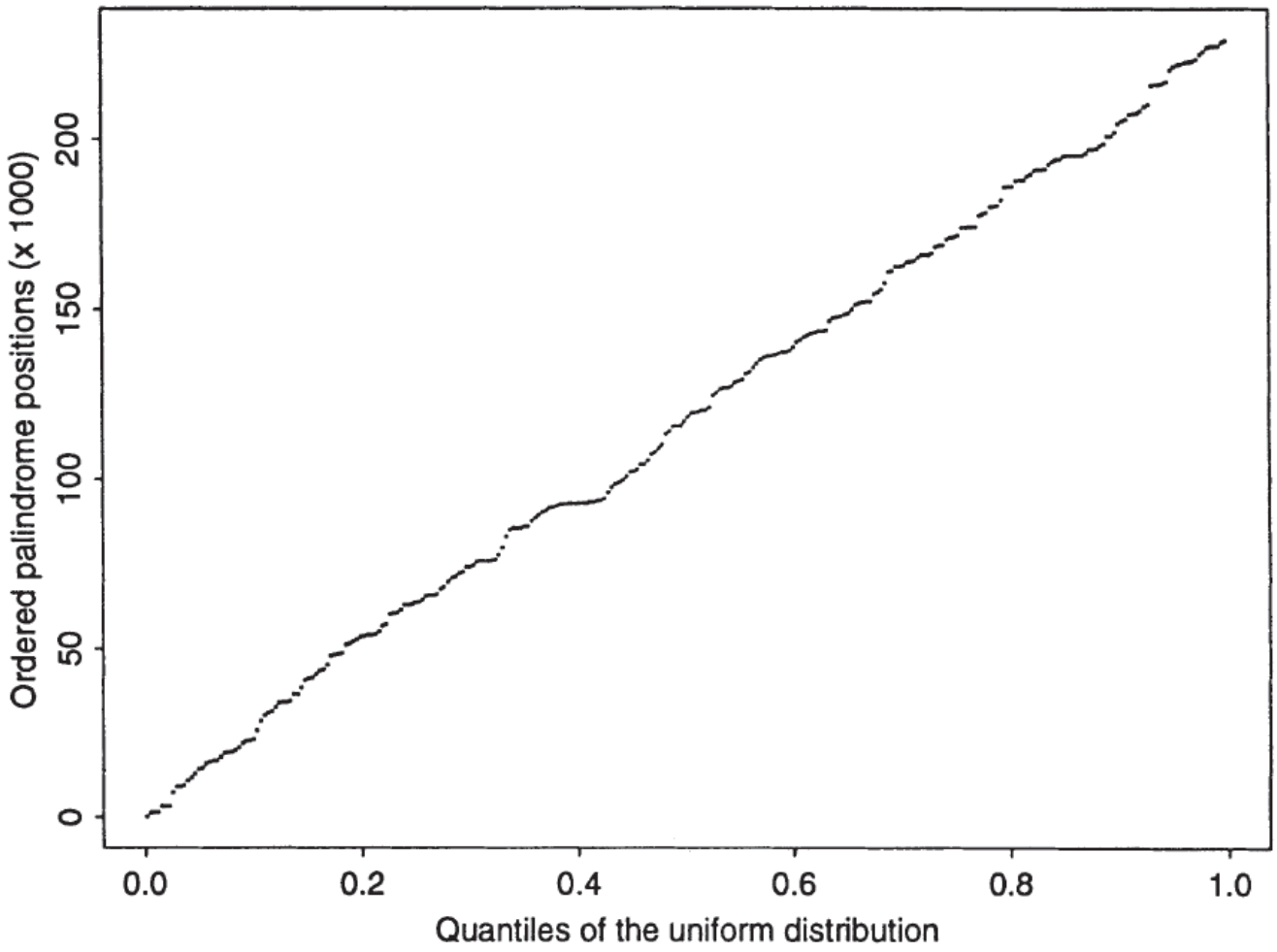
Analysis of the DNA sequence of HCMV

The DNA sequence of Human Cytomegalovirus has 229354 base pairs long and 296 palindromes. As we has mentioned before, comparing to the whole DNA sequence, the 296 palindromes seems like 296 points scattered in the whole viral DNA sequence. The table 1 below shows the location of 296 palindromes located in the HCMV DNA strand that are at least 10 base pairs long. Palindromic sequences that are less than 10 base long cannot be considered as a real palindrome because they may be occurred by random chance. In addition, the figure 1 displays the Q-Q plot of the location of the palindromes of HCMV versus the uniform quantiles.

Table 1. The location of 296 palindromes in HCMV DNA

177	19415	42376	61441	75812	92643	107414	128801	143549	162715	178574	195117	216190
1321	20030	43475	62946	75878	92701	108123	129057	143555	163745	180125	195151	216292
1433	20832	43696	63003	76043	92709	109185	129537	143738	163995	180374	195221	216539
1477	22027	45188	63023	76124	92747	110224	131200	146667	164072	180435	195262	217076
3248	22739	47905	63549	77642	92783	113378	131734	147612	165071	182195	195835	220549
3255	22910	48279	63769	79724	92859	114141	133040	147767	165883	186172	196992	221527
3286	23241	48370	64502	83033	93110	115627	134221	147878	165891	186203	197022	221949
7263	25949	48699	65555	85130	93250	115794	135361	148533	165931	186210	197191	222159
9023	28665	51170	65789	85513	93511	115818	136051	148821	166372	187981	198195	222573
9084	30378	51461	65802	85529	93601	117097	136405	150056	168261	188025	198709	222819
9333	30990	52243	66015	85640	94174	118555	136578	151314	168710	188137	201023	223001
10884	31503	52629	67605	86131	95975	119665	136870	151806	168815	189281	201056	223544
11754	32923	53439	68221	86137	97488	119757	137380	152045	170345	189810	202198	224994
12863	34103	53678	69733	87717	98493	119977	137593	152222	170988	190918	204548	225812
14263	34398	54012	70800	88803	98908	120411	137695	152331	170989	190985	205503	226936
14719	34403	54037	71257	89586	99709	120432	138111	154471	171607	190996	206000	227238
16013	34723	54142	72220	90251	100864	121370	139080	155073	173863	191298	207527	227249
16425	36596	55075	72553	90763	102139	124714	140579	155918	174049	192527	207788	227316
16752	36707	56695	74053	91490	102268	125546	141201	157617	174132	193447	207898	228424
16812	38626	57123	74059	91637	102711	126815	141994	161041	174185	193902	208572	228953
18009	40554	60068	74541	91953	104363	127024	142416	161316	174260	194111	209876	
19176	41100	60374	75622	92526	104502	127046	142991	162682	177727	195032	210469	
19325	41222	60552	75775	92570	105534	127587	143252	162703	177956	195112	215802	

Figure 1. Q-Q plot of palindromes vs. a uniform distribution (HCMV)



From the figure 1 above, the linear display shows that the 296 palindromes are distributed quite uniformly over the whole DNA sequence. From this characteristic, we can use the Compound Poisson approximation to estimate probability distribution of A_r in order to find out any nonrandom clusters in the DNA sequence. In other words, we are trying to find out an interval that does not fit significantly within the linear display in the figure 1, and this interval could be the origin of replication of HCMV. Just like other virus, experiments shows that many virus of the same family of HCMV have an unusual amount of palindrome at their origin of replication. Therefore, the use of Compound Poisson approximation can tell us if having a specific amount of palindromes in a specific region of the DNA of HCMV is due to random chance or not.

To start the analysis, statisticians divide the whole DNA sequence into windows of same length (w). Then, they calculate the probability for getting the minimum lengths of r consecutive palindromes

(A_r) within this length of windows. Because of the equivalent between the r-scan statistic and the traditional scan statistic, the probability for getting the minimum lengths of r consecutive palindromes in a window length is also the same probability for having at least r + 1 palindromes (S(w)) in the same window. For example in this experiment, we can divide the whole DNA sequence of HCMV into 2000 base pairs long windows, and if the probability for a minimum lengths of 5 consecutive palindromes within a 2000 bases pairs window length is 0.9, the probability for having at least 5 palindromes within the same 2000 base pairs window is also 0.9. However, because the r-scan works in an interval of (0, 1), we have to divide the 2000 bases pairs length to the whole viral DNA sequence which is 229354 bases pairs. Therefore, the equation of the example is going to be

$$P\{S(0.0087) \geq 5+1\} = P\{A_{5,} \leq 0.0087\} = 0.9$$

If the probability of a specific A_r is high, that means the distribution of r palindromes within the a specific window length is highly due to random chance. However, if the probability is significantly low, that means having a specific r palindromes in a specific interval is unusual.

Results

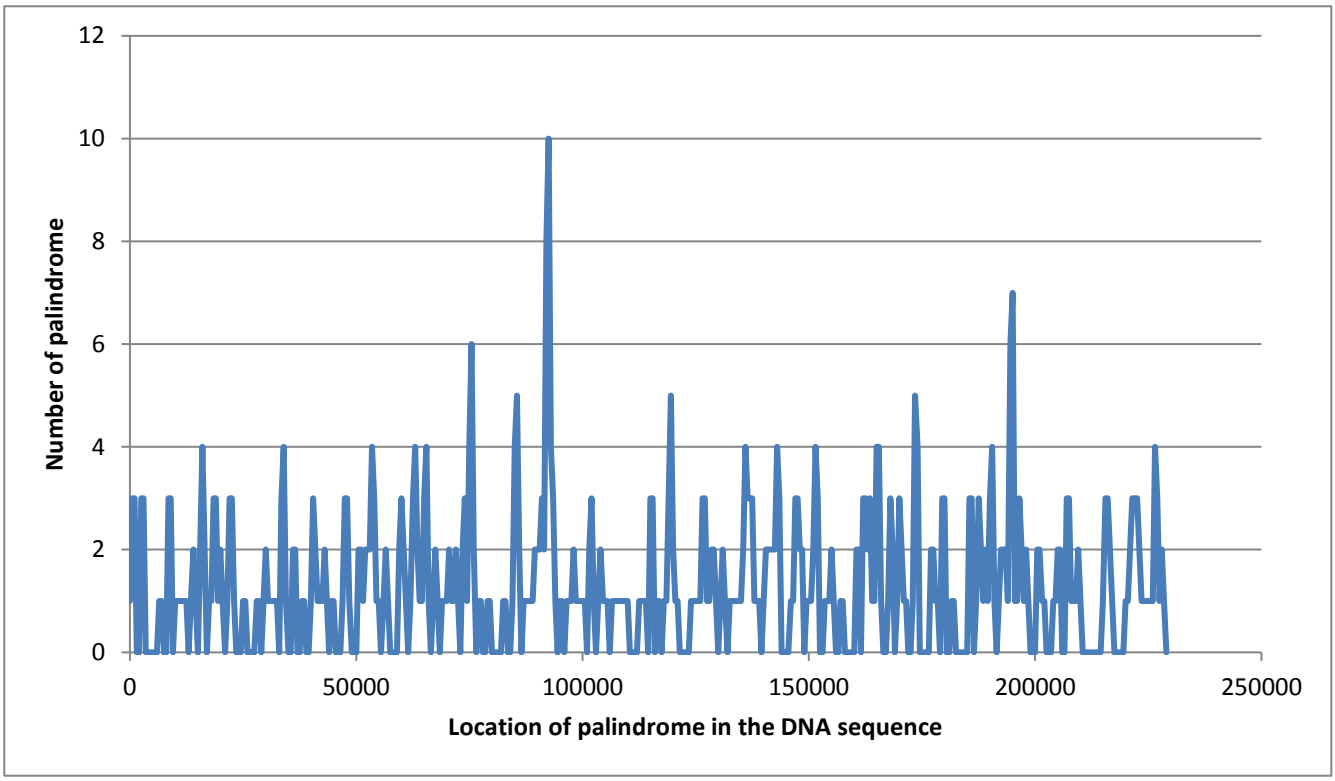
The null hypothesis of this experiment: having a A_r for a specific r palindromes within a specific DNA window length on the HCMV DNA sequence is due to random chance. Under the null hypothesis, the distribution of palindromes follows the random points distribution of r-scan.

The alternative hypothesis of this experiment: having a A_r for a specific r palindromes within a specific DNA window length is not due to random chance on the HCMV DNA sequence.

In order to test the hypothesis of palindromes distribution in HCMV DNA sequence, we are going to divide this viral DNA into three different length of DNA windows which 1000 bp, 1500 bp and 2000 bp. By analyzing three different length of windows, we can get a more accurate and precise estimation about the place of the origin of replication.

Case 1: for window length of 1000 bp

Figure 2. Location of Palindrome Scanning by 1000 bp



Note: The window forwards the successive windows by half which means 500 bp.

Figure 2 above shows that there is a unusual number of palindromes appearing at 75500 base pair where has 6 palindromes, at 92500 base pair where has 10 palindromes, and at 195000 base pair where has 7 palindromes.

Table 2. The probability distribution of A_r within 1000 bp window

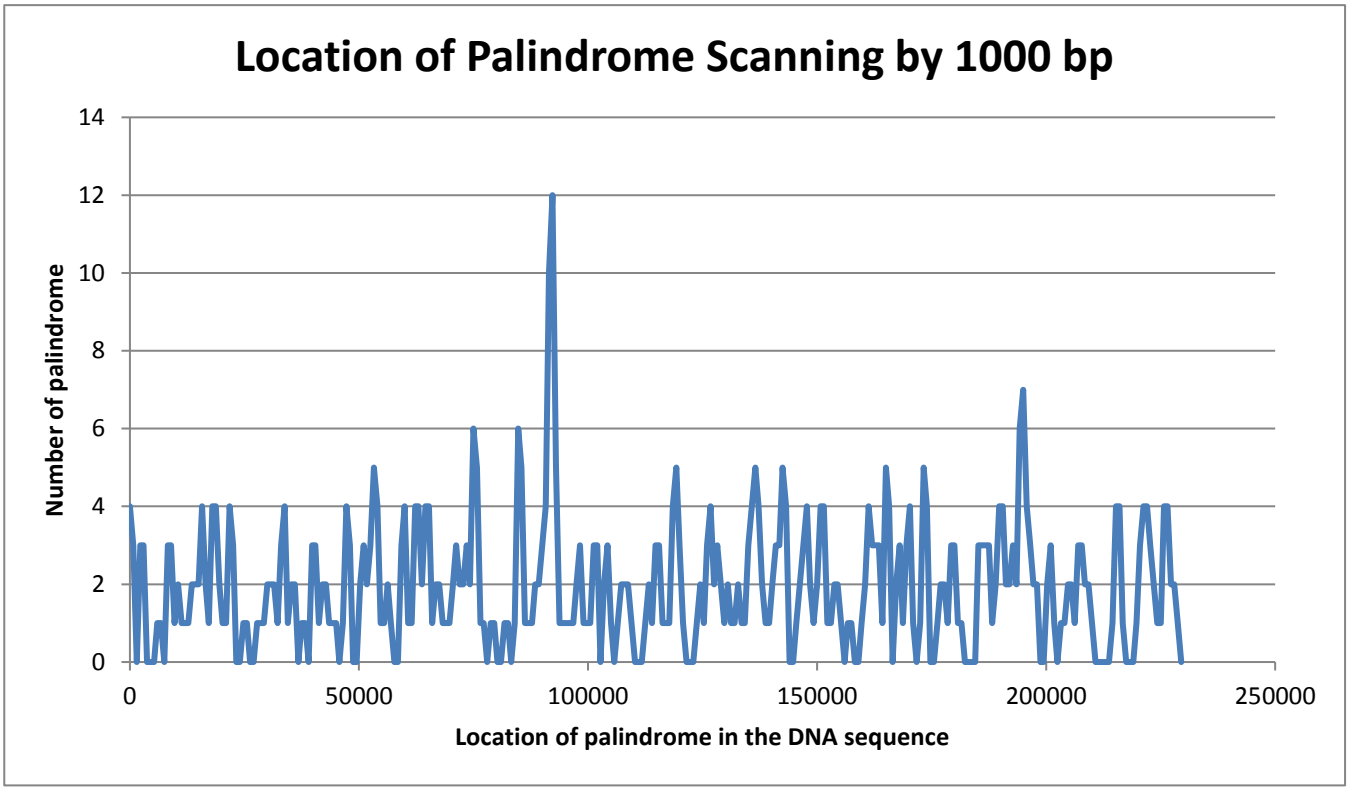
Window length (w)	r palindrome	Probability from Compound Poisson approximation
W = 1000 bp	1	1.0000
≈ 0.00436	2	1.0000
	3	1.0000
	4	0.9997
	5	0.86272

	6	0.34455
	7	0.07470
	8	0.01236
	9	0.00176
	10	0.00022
	≥11	0.00002

Table 2 shows that the p-value for having at least 6 and 7 palindromes within window length of 1000 bp is 0.344550 and 0.074703 respectively. However, the p-value for having at least 10 palindromes is 0.00022 which is significantly low because it is much smaller than the scientific statistical tolerant value which is 0.05. Therefore in this case, we have to reject the null hypothesis and accept the alternative hypothesis for having at least 10 palindromes at 92500 base pair.

Case 2: for window length of 1500 bp

Figure 3. Location of Palindrome Scanning by 1500 bp



Note: The window forwards the successive windows by half which means 750 bp.

Figure 3 above shows that there is a unusual number of palindromes appearing at 92250 bp where has 12 palindromes and at 195000 bp where has 7 palindromes.

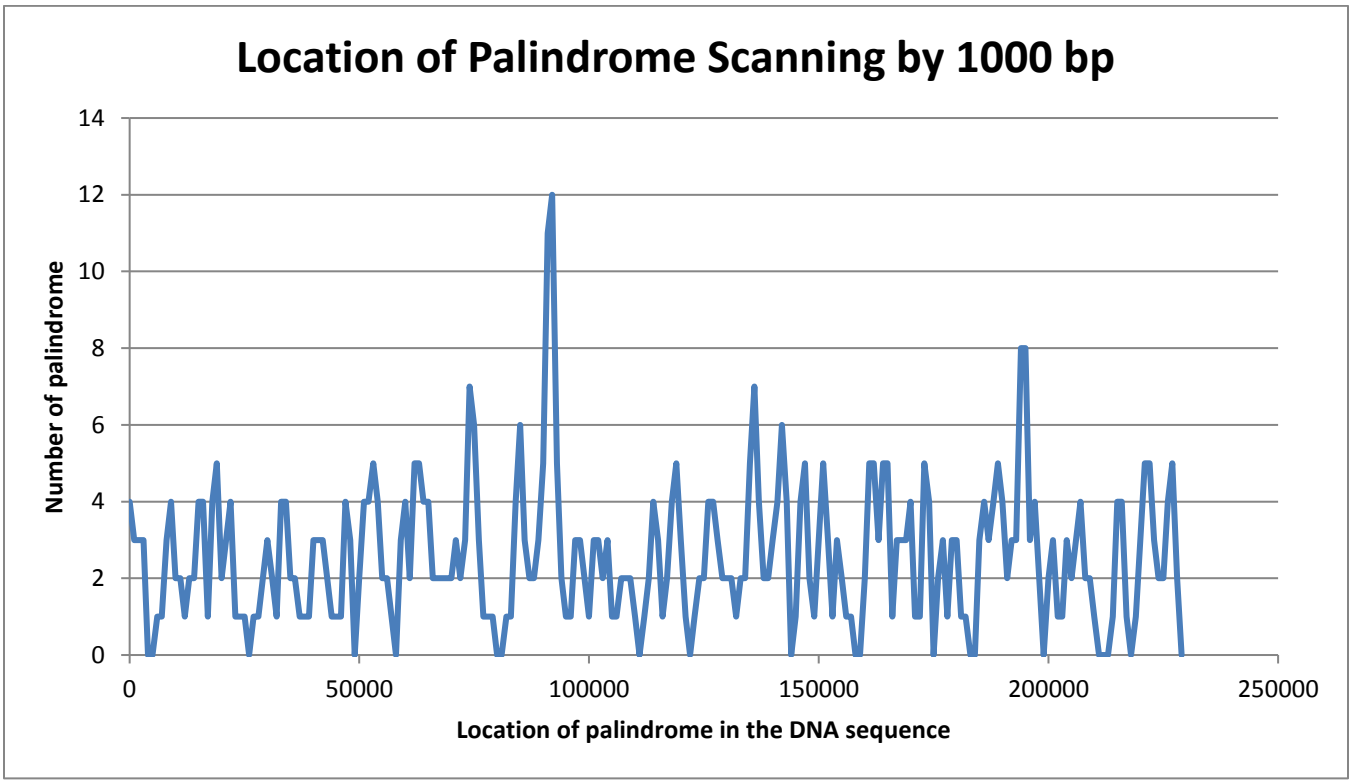
Table 3. The probability distribution of Ar within 1500 bp window

Window length (w)	r palindrome	Probability from Compound Poisson approximation
W = 1500 bp ≈ **0.0065**	1	1.0000
	2	1.0000
	3	1.0000
	4	1.0000
	5	0.9997
	6	0.91256
	7	0.48447
	8	0.14752
	9	0.03352
	10	0.00650
	11	0.00113
	12	0.00018
	≥13	0.00003

Table 3 shows that the p-value for having at least 7 palindromes is 0.48447 in 1500 bp windows. But for having at least 12 palindromes, the p-value is 0.00018 which is much smaller than the scientific tolerance, 0.05. Therefore in this case, we have to reject the null hypothesis and accept the alternative hypothesis for having at least 12 palindromes at 92250 base pair.

Case 3: for window length of 2000 bp

Figure 4. Location of Palindrome Scanning by 2000 bp



Note: The window forwards the successive windows by half which means 1000 bp .

Figure 3 above shows that there is a unusual number of palindromes appearing at 92000 bp where has 12 palindromes and at 194000 bp where has 8 palindromes.

Table 4. The probability distribution of A_r within 2000 bp window

Window length (w)	r palindrome	Probability from Compound Poisson approximation
W = 1000 bp	1	1.0000
≈ 0.0087	2	1.0000
	3	1.0000
	4	1.0000
	5	1.0000

	6	0.99941
	7	0.92182
	8	0.55331
	9	0.20517
	10	0.05713
	11	0.01354
	12	0.00287
	≥13	0.00056

Table 3 shows that the p-value for having at least 8 palindromes is 0.55331. But for having at least 12 palindromes, the p-value is 0.00287 which is significantly low. Thus in this case, we have to reject the null hypothesis and accept the alternative hypothesis for having at least 12 palindromes at 92000 base pair

Conclusion

From these three different observations, we see that there is unusual number of palindromes around 92000 bp and 92500 bp on the HCMV DNA sequence. Therefore, we conclude that the origin of replication is somewhere between these points. According to Proceedings of the National Academy of Science USA (1992), Masse et al. have done a very detailed experiment on the DNA sequence of HCMV around this part of the genome, and their results show that the segment between 92210 bp and 93715 bp is the lytic origin of replication for HCMV.

Bibliography

Masse, M. J., Karlin, S., Schachtel, G. A. and Mocarski, E. S. "Human cytomegalo virus origin of DNA replication (oriLyt) resides within Human cytomegalo virus origin of DNA replication (oriLyt) resides within." 1992.

Ming-Ying Leung & Trad E. Yamashita. "Applications of the Scan Statistic in DNA Sequence Analysis." Joseph GlazPozdnyakov, Sylvan WallensteinVladimir. Statistics for Industry and Technology. 2009. 269-286.