Introduction

Many important control signals of replication and gene expression are found in regions of the molecule with a high concentration of palindromes (e.g., see Masse et al. 1992). Statistical methods to identify significant clusters of palindromes would be helpful in finding putative locations of important functional sites on a long nucleotide sequence before experimentation. In this study, we investigate the scan statistic, which is a generalized likelihood ratio test for the presence of nonrandom clusters (Naus 1965) against uniformly distributed points on the unit interval. It is applied to locate significant palindrome clusters on one family of DNA viruses (the herpesviruses) and one family of RNA viruses (the paramyxoviruses).
Palindromes

A palindrome is a stretch of nucleotide bases which reads the same as its inverted complementary sequence. For example, ATGACCGGGTCAT is a palindrome of length 12.

Since short palindromes occur frequently even on randomly generated nucleotide sequences, we only take into consideration those palindromes of length $\geq 10$ for the DNA genomes of the herpesviruses (~$10^6$ bases), and $\geq 8$ for the RNA genomes of the paramyxoviruses (~$10^5$ bases).

An easy qualitative method to search for palindrome clusters is by examining a sliding window plot as shown in Figure 1. However, it is much more useful to have a criterion that helps to identify those clusters which are statistically significant, i.e., unlikely to occur by chance.
Figure 1: A sliding window plot is generated by choosing a window of fixed length and sliding it along the genome, beginning at the first base and continuing until the window reaches the last base of the genome. The window moves forward in steps of a pre-specified size. At each position of the window, the number of palindromes contained in it is counted and plotted against the window position. This is the sliding window plot for the human cytomegalovirus genome with a 1000 base window and step size of 500 bases. The peaks observed at window positions 92001 and 194501 suggest that there may be nonrandom palindrome clusters at these locations.
The Model

With few exceptions, the palindromes occurring in the genome sequences in this study have length no more than 20 bases. Compared to the entire genome lengths, these palindromes are extremely short. Hence it seems reasonable to idealize their occurrences on the genome as points on the unit interval.

As there are no established probabilistic results nor any generally accepted claims about the overall distribution of palindromes on nucleotide sequences, we start with a simple model assuming the palindromes to be identical and independently distributed random points on the interval (0,1). At least for some of the viral genomes, a Q-Q plot (Figure 2) can demonstrate that this assumption is a reasonable one.
The Q-Q plot

Figure 1: Q-Q plot for the palindrome positions of the human cytomegalovirus against quantiles of the uniform distribution. Here, we focus only on those palindromes with length $\geq 10$ bases in order to screen out the random noise generated by frequent fortuitous occurrences of very short palindromes (see Leung et al. 1994 for a full explanation). The overall straight line appearance of the Q-Q plot suggests that it would be reasonable to model the occurrences of palindromes above a prescribed length along the genome sequence as i.i.d. points uniformly distributed over $(0,1)$ and evaluate the significance of palindrome clusters with the scan statistic distribution.
Scan Statistics: Notations

Let $X_1, \ldots, X_n$ be the order statistics of $n$ i.i.d. uniformly distributed points on the unit interval $(0,1)$. Let

$$S_i = X_{i+1} - X_i, \ i = 1, \ldots, n-1$$

denote the spacing between adjacent points and

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

be the sum of the $r$ adjoining spacings starting at $X_i$.

$N_w(i)$ stands for the number of points contained in a window of length $w$ beginning at $X_i$.

Figure 3 Illustration of notations used in defining the scan statistics
Scan Statistics: Definitions and Duality Relation

For a fixed window length $0 < w < 1$, the traditional scan statistic $N_w = \max\{N_w(i) : i = 1, \ldots, n\}$ is the largest number of points contained in a window of size $w$ scanning over the unit interval.

For a fixed integer $r > 0$, the $r$-scan statistic $A_r = \min\{A_r(i) : i = 1, \ldots, n - r\}$ is the smallest of the aggregated $r$-spacings of the $n$ points.

$N_w$ and $A_r$ are related by a duality relation

$$\{ N_w \geq r + 1 \} = \{ A_r \leq w \}.$$ 

Hence, the traditional scan statistic and the $r$-scan can be used interchangeably. In molecular sequence analysis, usually the $r$-scan is preferred.
Poisson Approximation

Dembo and Karlin (1992) derive the limiting distribution
\[
\lim_{n \to \infty} P\left\{ A_r > \frac{x}{n^{1+1/r}} \right\} = e^{-x^r/r!}
\]

This follows from a Poisson limiting distribution for the counts \( C_r \) of those \( A_r(i) \)'s not exceeding \( x / n^{1+1/r} \).

If the above limiting distribution is used as an approximation for large \( n \), one can easily obtain a critical value
\[
c = -\frac{r!\ln(1-\alpha)}{n^{r+1}}
\]
for \( A_r \) below which a significant cluster is considered present.
When using an asymptotic result as an approximation, we need to be concerned about the rate of convergence to the limiting distribution. Unfortunately, the distribution of $A_r$ converges very slowly, at the rate of $O((\log n /n)^{1/2})$. Several alternative approximation distributions for $A_r$ have been proposed (see references in Glaz 1994).

The simulations of Leung and Yamashita (1999) compare these approximations and show that for moderately large $n$ (the range of $n$ in this study is 40 - 300), the best result obtained is based on a compound Poisson approximation of $C_r$ (Glaz 1994). Although this approximation formula is quite complicated and therefore not presented here, it is relatively straightforward to compute in S-Plus (or other statistical programming language).
Significant Palindrome clusters on the Herpesvirus and Paramyxovirus genomes

Using the compound Poisson approximation for the $r$-scan distribution, ten complete DNA genomes of herpesviruses and five complete RNA genomes of paramyxoviruses are analyzed. Table I gives a list of all statistically significant ($\alpha=0.05$) clusters of palindromes on these genomes.

Among the herpesviruses which have been extensively researched, most of the statistically significant clusters correspond to regions known to contain either an origin of replication or a transcriptional regulator (O’Brien 1993), suggesting that the regions of significant clusters in the other herpesviruses may be likely candidates for similar functional sites.
<table>
<thead>
<tr>
<th>Genome</th>
<th>Cluster Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human cytomegalovirus</td>
<td>91490-92643, 190532-195112</td>
</tr>
<tr>
<td>Epstein Barr virus</td>
<td>52787-53311, 85174</td>
</tr>
<tr>
<td>Herpes simplex 1</td>
<td>129511, 146228</td>
</tr>
<tr>
<td>Varicella zoster virus</td>
<td>1542</td>
</tr>
<tr>
<td>Equine herpes simplex</td>
<td>115125-115893, 144717-146485</td>
</tr>
<tr>
<td>Herpes saimiri</td>
<td>112418-112422, 109081-109238</td>
</tr>
<tr>
<td>Ateline herpesvirus 3</td>
<td>95350-96866</td>
</tr>
<tr>
<td>Bovine herpesvirus type 1.1</td>
<td>77554</td>
</tr>
<tr>
<td>Murine herpesvirus 68</td>
<td>26126, 46085-48154, 69174-69652, 70962-72987</td>
</tr>
<tr>
<td>Human parainfluenza 3</td>
<td>12679-12749</td>
</tr>
<tr>
<td>Tupaia paramyxovirus</td>
<td>16634</td>
</tr>
</tbody>
</table>

**Table 1:** Location of the significant clusters of palindromes found on the herpesviruses and paramyxoviruses. Apart from those genomes listed above, the set of genomes examined also include the Ictalurid herpesvirus, bovine parainfluenza virus 3, a mutant of the human parainfluenza virus 3, and the Sendai virus. They are not included in the table because no significant cluster is found.
References


