Bioinformatic Analysis of Palindromes in Bacterial Genome Sequences

Peter Thorpe, Leighton Pritchard, Peter Cock
Plant Pathology, Scottish Crop Research Institute, Dundee, DD2 5DA

scri living technology

UNIVERSITY OF LEEDS

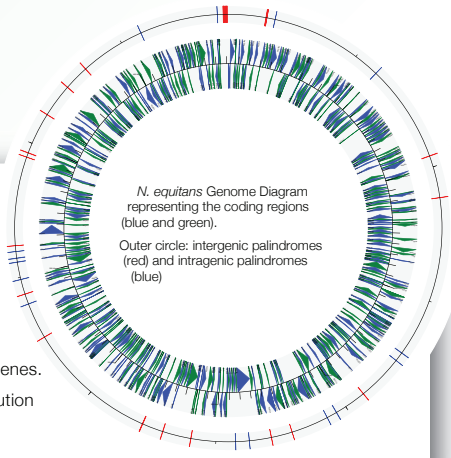
mylnefield research services

Aims of the Study

- To analyse bacterial genomes for the presence or absence of palindromic sequences.
- To investigate whether distribution of the palindromes is random or whether they are preferentially distributed within inter- or intragenic regions.
- To determine if a specific range of palindrome lengths are selected for in bacterial genomes.

Why Study Palindromes?

- Palindromes control gene expression through their interaction with transcription factors.
- They stabilise mRNA by inhibiting nuclease activity.
- They have been shown to be involved in mRNA localisation.
- Palindromes are present in both prokaryotes and eukaryotes.
- Organisms use palindromes as markers for self-DNA and non self-DNA, maybe as a way of preventing the expression of foreign genes.
- Palindromes can be used as markers of DNA to trace genome evolution and the acquisition of genes through horizontal gene transfer.
- Palindromes are involved in correct intron splicing.



N. equitans Genome Diagram representing the coding regions (blue and green).

Outer circle: intergenic palindromes (red) and intragenic palindromes (blue)

Materials and Methods

Python programs were developed to search for palindromes of length 15-150 with a probability less than 1/length of the genome. The probability function could be defined by:

$$P = \binom{L/2}{M} \bullet (Pmatch)^{(L/2)-M} \bullet (Pmismatch)^{M}$$

Where $\binom{L/2}{M}$ is the number of ways of getting M mismatches in the half length of palindrome (L/2).

The probability of getting a match (Pmatch) is: $\Sigma(A^*T, T^*A, C^*G, G^*C)^{(Number\ of\ matches)}$, the values for A, T, C and G are returned from the base composition function.

The probability of getting a mismatch (Pmismatch) is: $(1 - Pmatch)^{(Number\ of\ Mismatches)}$.

- Statistical support for the data was provided by running similar analysis on 30 randomly generated genomes of the same length and base composition as the parent genome.
- Four organisms were selected for this study: *Nanoarchaeum equitans, Escherichia coli* strain K-12 and plant pathogens *Pectobacterium atrosepticum* and *Pseudomonas syringae.*

Genome Sequence → Palindrome searcher ← 30 Random Sequences

↓

Repeat palindrome condenser

Inter – Intragenic Results

In each case the differences between real and random sequences were significant. This suggests there are a greater number of palindromes in intergenic regions in real bacterial genomes than would be expected to occur at random. This suggests palindromes have an intergenic function.

	Mean of Random Population (S.D.)	*N. equitans*	Chi-squared
Intergenic Score	2.4 (1.4)	31	P < 0.001 13.53 with 1 d.f.
Intragenic Score	14.1 (4.3)	15	

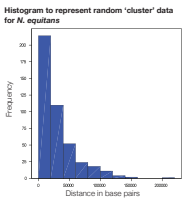
	Mean of Random Population (S.D.)	*E. coli*	Chi-squared
Intergenic Score	3.5 (1.7)	185	P < 0.001 48.93 with 1 d.f.
Intragenic Score	14.5 (4.1)	29	

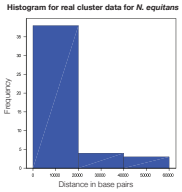
	Mean of Random Population (S.D.)	*P. atrosepticum*	Chi-squared
Intergenic Score	4.5	208	P < 0.001 82.15 with 1 d.f.
Intragenic Score	14.2	54	

	Mean of Random Population (S.D.)	*P. syringae*	Chi-squared
Intergenic Score	3.8	242	P < 0.001 30.07 with 1 d.f.
Intragenic Score	13.6	64	

Clustering Analysis

Using a one tailed Poisson test, in each case there was a significant difference between the random and real sequences: *N. equitans* P < 0.001, *E. coli* strain k-12 P < 0.001, *P. atrosepticum* P < 0.001 and *P. syringae* P < 0.001. This suggest palindrome clustering in intergenic regions may be involved in the function of palindromes.





Histograms using the same bin boundaries of 20 000 for the real and random data population for *N. equitans*. A clear difference can be seen.

Number Found Results

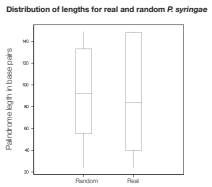
A significantly greater number of palindromes were present in the real genome sequence than in the random sequences generated from each parent sequence.

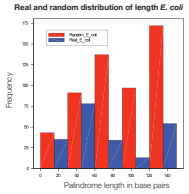
	Random mean (S.D)	'Real' number found	Probability of 'real' result occurring in the random population
N. equitans	16.5 (4.3)	46	P <0.001
E. coli strain k-12	18 (3.7)	214	P <0.001
P. syringae	17.4 (4.5)	306	P <0.001
P. atrosepticum	18.7 (4.5)	262	P <0.001

Distribution of Length Results

Using chi-squared analysis no significant difference in the distribution of lengths of palindromes in the real and randomly generated sequences was found for *N. equitans* (P = 0.39), *E. coli* strain k-12 (P = 0.41), *P. atrospeticum* (P = 0.14) or *P. syringae* (P = 0.681).

Histogram and Box-plot to represent the distribution of lengths in *E.coli* and *P. syringae.* Chi-squared analysis shows no statistical difference.





Conclusions

- Statistically greater number of palindromes were found in the real genome than the number expected to occur at random.
- Distribution was biased towards the intergenic regions.
- The results of this study support the hypothesis that palindromes are selected for within intergenic regions of bacterial genomes, where palindrome clustering may be occurring. This could be due to, and is consistent with published research, that palindromes are involved with the control of up and downstream gene regulators, from their intergenic locations.

Funded by BBSRC