Distributions of Palindromic Proportional Content in Bacteria

Oliver Bonham-Carter, Lotfollah Najjar, Ishwor Thapa and Dhundy R. Bastola

College of Information Science and Technology, University of Nebraska at Omaha,
Omaha, NE 68182

Research Problem: Motivation

DNA palindromes, the reversed and complemented genetic words, are read the same in the 3' to 5' as the 5' to 3' direction, and can form a unique restriction sites (RSs) where enzymes are able to cut DNA. Several studies have confirmed that short palindromes, behaving as active RSs, are few when compared to statistically expected values in bacterial genomes. These studies suggest that palindromes bring potential instability to intolerant coding regions of the genomes which appears to alter their concentrations. While this palindrome-avoidance phenomenon has been observed in bacteria, the exact location in the genome where palindromes are most rare has not been investigated.

Results:

In this paper, we provide evidence to suggest where the palindromic content is the least by comparing the content in *coding* and *non-coding* regions of bacterial DNA. We study the exhaustive lists of palindromes (lengths 4, 6, 8, and 10) to conclude that at least half of the motifs of each set (and sometimes, nearly all of the motifs of a set) show similar trends of reduced presence in the coding regions, when compared to the non-coding regions of bacteria.

What is a Palindrome?

A DNA palindrome is made up of {`A`,`C`,`G`,`T`} and is also a reversed complement of itself where `A` and `G` have complements of `T` and `C`.

<u>Definition of DNA Palindrome</u>

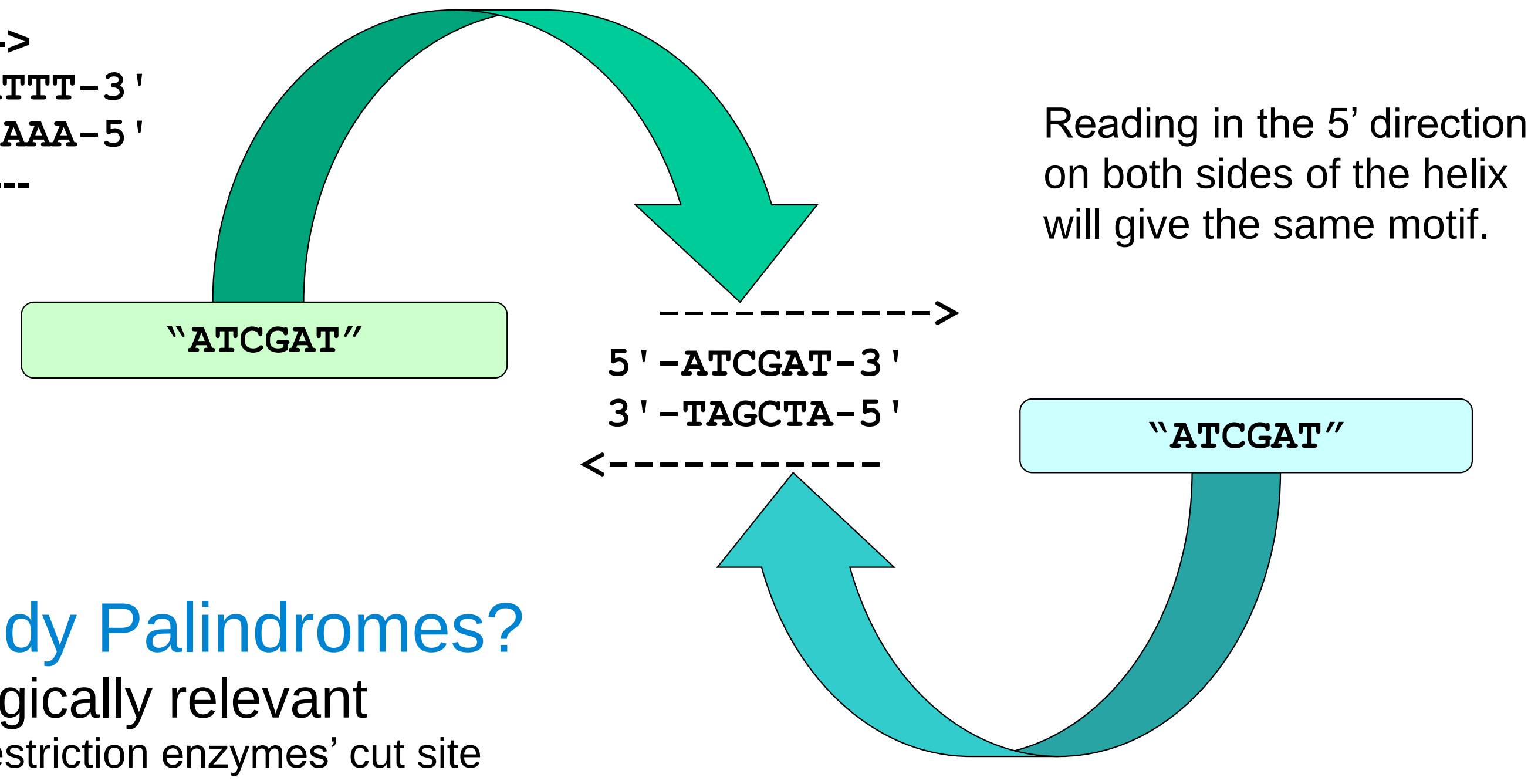
Our study defines a palindrome to be of an even length and to be the reversed compliment of the word.
For W = a DNA word, if W == reversed_compliment(W) then W is a palindrome.

<u>Example:</u>
W = **`AGCT`**
reversed(W) = **`TCGA`**
compliment(W) = **`AGCT`**

In a DNA helix, a palindromic sequence is read the same (in the 5' to 3' direction) from either strand due to the reverse complement property. For this reason, they form completely unique subsequences in DNA (on both strands) since all the other possible subsequences are represented by their matching complements on the other strand.

<u>Example</u>:
Palindromes are unique words in a 5' to 3' reading frame.
There are unique words which occur in the 5' to 3' reading frame at the same location on both strands;

```
   ------->
5'-AAATTT-3'
3'-TTTAAA-5'
   <-------
```



Reading in the 5' direction on both sides of the helix will give the same motif.

Why study Palindromes?

- Are biologically relevant
  - make restriction enzymes' cut site
  - can be used to track evolution
  - control gene expression
  - present in both prokaryotes and eukaryotes.
  - distinguish the host's DNA ("self") from that of the pathogen ("non-self")
  - stabilize mRNA by inhibiting nuclease activity
- Provides structure
  - They form hairpin loops and dimers which are necessary in protein folding.

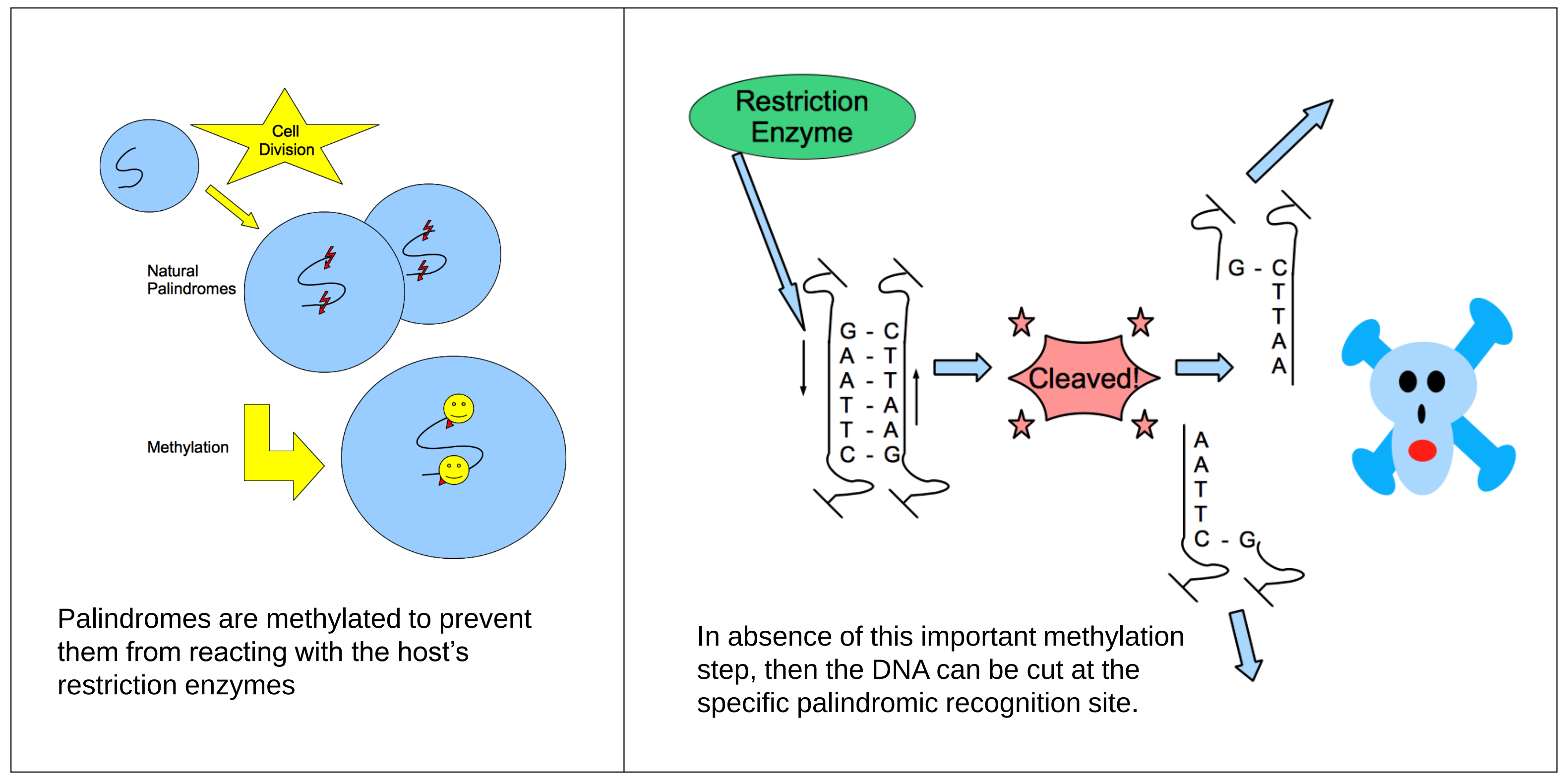
Palindromic Proportions

In this study, we compared the proportions of each palindrome from the coding and non-coding regions of bacterial DNA. There were no overlapping palindromes and we did not consider nested palindromes. This test was to determine whether a particular palindrome had more weight in one region over the other. The Mann-Whitney test, assuming no normal distribution, was appropriate for this biological data which was often non-normal in distribution, [1].

By using motif proportions, we calculate the amount of the sequence which is made up of the particular palindrome. The more common the palindrome in a region, then higher its proportion would be. For M_i, a motif, S_L the sequence space (coding or non-coding) and, $|M_i|$ and $|S_L|$ the lengths of the motifs and sequence spaces, respectively;
$Prop_i = (count(M_i) * |M_i|) / |S_L|$.

Significance of Methylation

Although palindromes are part of the host's immune defenses, they can also be dangerous to the host. The host undergoes methylation of its own palindromic content to prevent trouble.



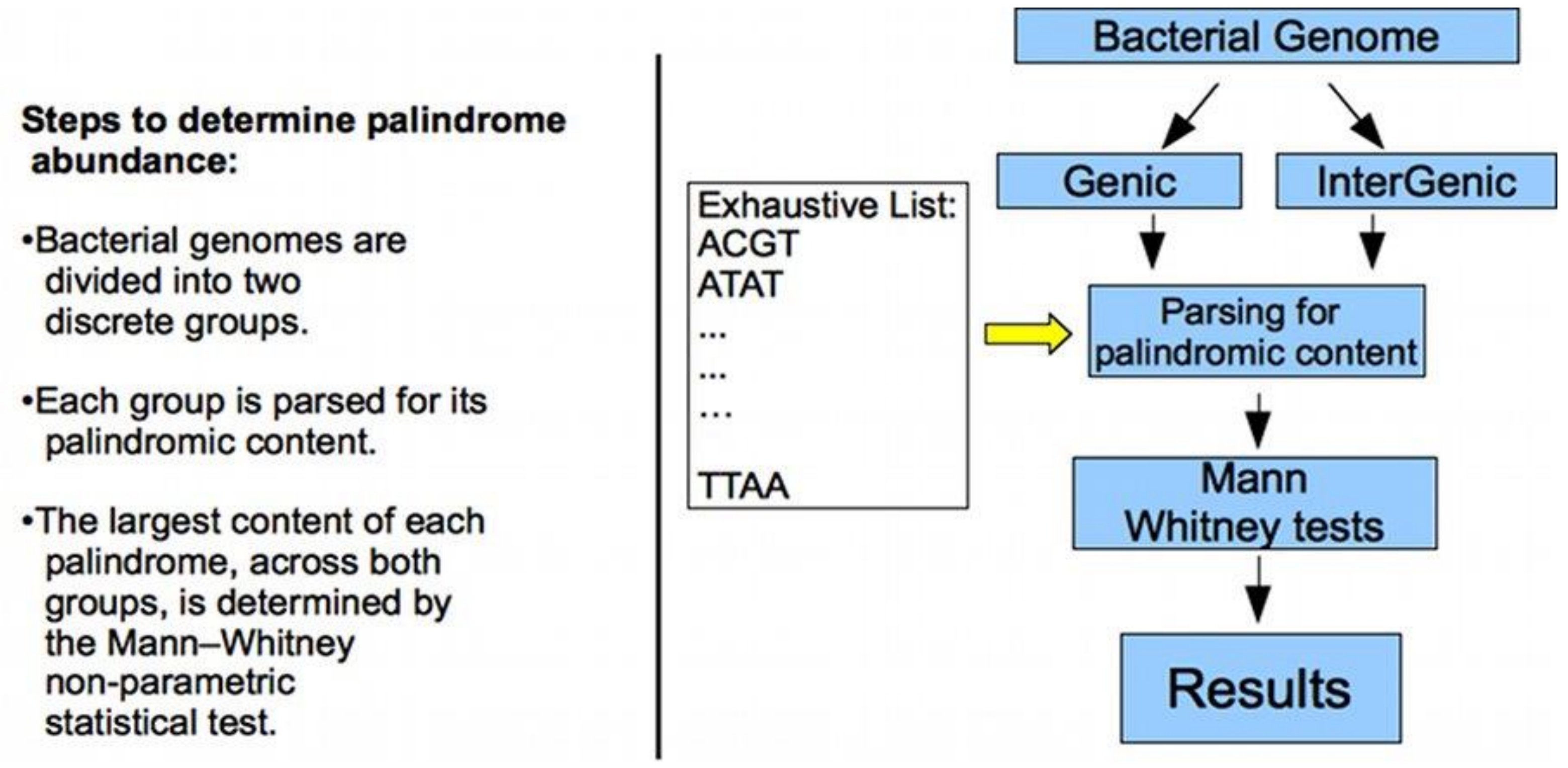
Palindromes are methylated to prevent them from reacting with the host's restriction enzymes

In absence of this important methylation step, then the DNA can be cut at the specific palindromic recognition site.

Alternative Hypothesis

A palindrome has higher proportions in the non-coding regions than in the coding regions of all evaluated genomes.

Materials and Methods



Steps to determine palindrome abundance:

- Bacterial genomes are divided into two discrete groups.
- Each group is parsed for its palindromic content.
- The largest content of each palindrome, across both groups, is determined by the Mann–Whitney non-parametric statistical test.

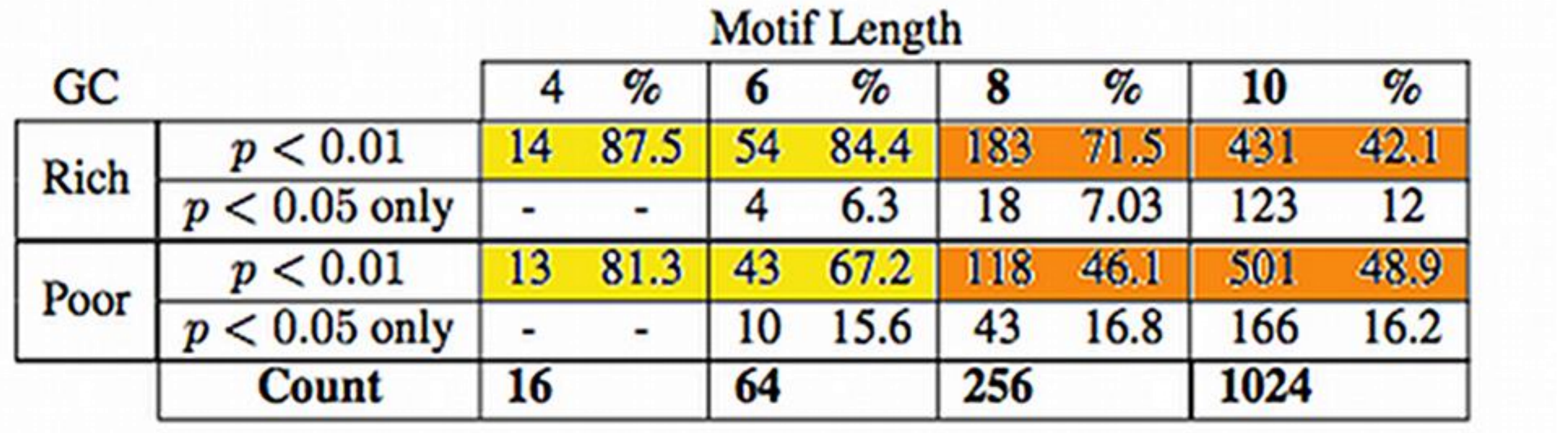
Organisms and GC Scoring

Because of the evidence that GC-groups have a tendency to mutate to AT groups [2, 3] and that similar GC composition implies similar genomic structure [4], our analysis was drawn from bacteria having both GC-rich and poor composition.

Table 1. List of bacterial genome used in this study. These organisms were classified based on GC content score. In the given GCrichness equation, |SL| = sequence length, G-, C-counts are the number of G's and C's (respectively) found in the current sequence.

GC-Rich	GC-Poor	Border Composition
Bifidobacterium Burkholderia Caulobacter Desulfovibrio Geobacter Xanthomonas	Agrobacterium Bifidobacterium Brucella Chloroflexus Corynebacterium Erwinia Geobacter Pantoea	Candidatus Mycobacterium Pseudomonas $GCrichnessScore = \frac{C_{count} + G_{count}}{\|S_L\|}$ $= \begin{cases} GCRich & \text{if } GCrichness \in (0.6, 1.0] \\ GCPoor & \text{if } GCrichness \in [0.0, 0.6] \end{cases}$

Results



GC		Motif Length 4	%	6	%	8	%	10	%
Rich	$p < 0.01$	14	87.5	54	84.4	183	71.5	431	42.1
	$p < 0.05$ only	-	-	4	6.3	18	7.03	123	12
Poor	$p < 0.01$	13	81.3	43	67.2	118	46.1	501	48.9
	$p < 0.05$ only	-	-	10	15.6	43	16.8	166	16.2
	Count	16		64		256		1024	

The "only" indicates that this number of palindromes is significant only at the alpha = 0.01 level. The percentages indicate the number of palindromes, from the exhaustive list, which are found in higher proportions in the non-coding regions than the coding regions. We note that most of the motifs of the exhaustive lists of palindromes length 4 and 6, tending to be restriction sites, are not commonly found in coding regions. We suspect that coding regions are generally intolerant of the instability brought by palindromes.

Conclusions

- Shorter palindromes are more abundant in the non-coding regions.
- Palindromic distribution is independent of GC-content of the genomes.
- Short palindromes do not have a uniform distribution and are under represented in coding regions of bacterial genomes.
- This mechanism behind the phenomenon will require biological validation (studies of DNA folding and structure, etc).

References

1. Karlin S., Burge C., Campbell A.M. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. 1992;20:1363-1370.
2. Hershberg R., Petrov D.A. (2010) Evidence that mutation is universally biased towards AT in bacteria.PLoS Genet. Sep 9;6(9).
3. Hildebrand F., Meyer A., and Eyre-Walker A. (2010): Evidence of selection upon genomic GC-content in bacteria. PLoS genetics, 6(9).
4. Lightfield J., Fram N.R., Ely B. (2011) Across Bacterial Phyla, Distantly Related Genomes with Similar Genomic GC Content Have Similar Patterns of Amino Acid Usage. PLoS ONE, 6(3).

Acknowledgements

We would like to thank the support staff in the UNO-Bioinformatics Core Facility, funded by the grants from the National Center for Research Resources (5P20RR016469) and the National Institute for General Medical Science (NIGMS) (8P20GM103427)