DNA Periodicity

Signals and Systems in Biology
Kushal Shah @ EE, IIT Delhi
Gene Periodicity

S. Tiwari et. al., Bioinformatics 1997
Genome Periodicity

Nucleotide Distribution

\[
\begin{pmatrix}
F_{A1} & F_{A2} & F_{A3} \\
F_{T1} & F_{T2} & F_{T3} \\
F_{C1} & F_{C2} & F_{C3} \\
F_{G1} & F_{G2} & F_{G3}
\end{pmatrix}
\times
\begin{pmatrix}
0.2696 & 0.3006 & 0.2259 \\
0.0676 & 0.2933 & 0.2605 \\
0.1676 & 0.2514 & 0.2987 \\
0.4973 & 0.1548 & 0.2149
\end{pmatrix}
\]

Organism

- Methanocaldococcus jannaschii: 0.1589 0.1598 0.1587 0.1583
- Saccharomyces cerevisiae CHR III: 0.1884 0.1909 0.1856 0.1886
- Escherichia coli: 0.2537 0.2553 0.2526 0.2532
Nucleotide Distribution

\[
\begin{pmatrix}
F_{A1} & F_{A2} & F_{A3} \\
F_{T1} & F_{T2} & F_{T3} \\
F_{C1} & F_{C2} & F_{C3} \\
F_{G1} & F_{G2} & F_{G3}
\end{pmatrix}
\begin{pmatrix}
0.2696 & 0.3006 & 0.2259 \\
0.0676 & 0.2933 & 0.2605 \\
0.1676 & 0.2514 & 0.2987 \\
0.4973 & 0.1548 & 0.2149
\end{pmatrix}
\]

Organism

<table>
<thead>
<tr>
<th>Methanocaldococcus jannaschii</th>
<th>( f_G )</th>
<th>( f_{G1} )</th>
<th>( f_{G2} )</th>
<th>( f_{G0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1589</td>
<td>0.1598</td>
<td>0.1587</td>
<td>0.1583</td>
</tr>
</tbody>
</table>
Nucleotide Distribution

\[
\begin{pmatrix}
F_{A1} & F_{A2} & F_{A3} \\
F_{T1} & F_{T2} & F_{T3} \\
F_{C1} & F_{C2} & F_{C3} \\
F_{G1} & F_{G2} & F_{G3}
\end{pmatrix}
\begin{pmatrix}
0.2696 \\
0.0676 \\
0.1676 \\
0.4973
\end{pmatrix}
= 
\begin{pmatrix}
0.3006 \\
0.2933 \\
0.2514 \\
0.1548
\end{pmatrix}
\begin{pmatrix}
0.2259 \\
0.2605 \\
0.2987 \\
0.2149
\end{pmatrix}
\]

**Organism**

<table>
<thead>
<tr>
<th>Organism</th>
<th>( f_G )</th>
<th>( f_{G1} )</th>
<th>( f_{G2} )</th>
<th>( f_{G0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanocaldococcus jannaschii</td>
<td>0.1589</td>
<td>0.1598</td>
<td>0.1587</td>
<td>0.1583</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae CHR III</td>
<td>0.1884</td>
<td>0.1909</td>
<td>0.1856</td>
<td>0.1886</td>
</tr>
</tbody>
</table>
Nucleotide Distribution

\[
\begin{pmatrix}
F_{A1} & F_{A2} & F_{A3} \\
F_{T1} & F_{T2} & F_{T3} \\
F_{C1} & F_{C2} & F_{C3} \\
F_{G1} & F_{G2} & F_{G3}
\end{pmatrix}
\]_{GENE}
= 
\begin{pmatrix}
0.2696 & 0.3006 & 0.2259 \\
0.0676 & 0.2933 & 0.2605 \\
0.1676 & 0.2514 & 0.2987 \\
0.4973 & 0.1548 & 0.2149
\end{pmatrix}

Organism

<table>
<thead>
<tr>
<th>Organism</th>
<th>$f_G$</th>
<th>$f_{G1}$</th>
<th>$f_{G2}$</th>
<th>$f_{G0}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanocaldococcus jannaschii</td>
<td>0.1589</td>
<td>0.1598</td>
<td>0.1587</td>
<td>0.1583</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae CHR III</td>
<td>0.1884</td>
<td>0.1909</td>
<td>0.1856</td>
<td>0.1886</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>0.2537</td>
<td>0.2553</td>
<td>0.2526</td>
<td>0.2532</td>
</tr>
</tbody>
</table>
Correlation Plot: Whole Genome of *E. coli*

![Figure 1](image)

\[ C(1) = 4(0.5 - p)^2 \]

Correlation Plot : GroEL gene of *E. coli*

Figure 2

3-base Periodicity: Evolutionary Perspective

- Sequences largely evolve by local repetition of triplets.
- These triplet repetitions then undergo mutations.
- Tandem repeats contain memory of recent expansion events.
Sequences largely evolve by local repetition of triplets
3-base Periodicity: Evolutionary Perspective

- Sequences largely evolve by local repetition of triplets.
- These triplet repetitions then undergo mutations.
Sequences largely evolve by local repetition of triplets

These triplet repetitions then undergo mutations

Tandem repeats contain memory of recent expansion events
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
- Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes
- Experimental value of helical repeat: $10.39 \pm 0.02$
- Difference of $0.16$ due to the super-helical winding in nucleosome

- Periodicities in protein sequences
  - 3.6 amino acids per alpha-helix turn
  - Leads to 10.8-base periodicity of nucleotides
  - 3.5 periodicity of hydrophilic and hydrophobic residues

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes
- Experimental value of helical repeat: $10.39 \pm 0.02$
  - Difference of 0.16 due to the super-helical winding in nucleosome

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes
- Experimental value of helical repeat: $10.39 \pm 0.02$
  - Difference of 0.16 due to the super-helical winding in nucleosome
- Periodicities in protein sequences

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes
- Experimental value of helical repeat: $10.39 \pm 0.02$
  - Difference of 0.16 due to the super-helical winding in nucleosome
- Periodicities in protein sequences
  - 3.6 amino acids per alpha-helix turn

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes
- Experimental value of helical repeat: $10.39 \pm 0.02$
  - Difference of 0.16 due to the super-helical winding in nucleosome
- Periodicities in protein sequences
  - 3.6 amino acids per alpha-helix turn
    Leads to 10.8-base periodicity of nucleotides

E. N. Trifonov, Physica A 1998
10.5 base periodicity

- Helical repeat: $10.55 \pm 0.01$ nucleotides per turn
- AA-TT and AG-CT cause DNA curvature
  - Non-optimal separation of 5 nt
- Sequence requirement for DNA folding in nucleosomes
- Experimental value of helical repeat: $10.39 \pm 0.02$
  - Difference of 0.16 due to the super-helical winding in nucleosome
- Periodicities in protein sequences
  - 3.6 amino acids per alpha-helix turn
  - Leads to 10.8-base periodicity of nucleotides
  - 3.5 periodicity of hydrophilic and hydrophobic residues

E. N. Trifonov, Physica A 1998
10.5 base periodicity: Correlation

\[ C_{AA}(k) = p_{AA}(k) - p_A^2 \]

If nucleotides at distance \( k \) are statistically independent, \( C_{AA}(k) = 0 \! \!

\[ \overline{C}(k) = \frac{C(k - 1) + C(k) + C(k + 1)}{3} \]

\( C_{WW} \): Correlation of weakly binding nucleotides \( A \) and \( T \)

H. Herzel et. al., Bioinformatics 1999
## Estimated periodicities of genomic DNA

<table>
<thead>
<tr>
<th>Organism</th>
<th>Length (Mbp)</th>
<th>AA</th>
<th>WW</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>4.6</td>
<td>11.0</td>
<td>11.0</td>
</tr>
<tr>
<td>B. subtilis</td>
<td>4.2</td>
<td>11.2</td>
<td>11.2</td>
</tr>
<tr>
<td>Synechocystis</td>
<td>3.6</td>
<td>11.5</td>
<td>11.6</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1.8</td>
<td>11.2</td>
<td>11.0</td>
</tr>
<tr>
<td>H. pylori</td>
<td>1.7</td>
<td>11.2</td>
<td>11.2</td>
</tr>
<tr>
<td>M. pnemoniae</td>
<td>0.8</td>
<td>11.3</td>
<td>11.4</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>0.6</td>
<td>11.5</td>
<td>11.5</td>
</tr>
<tr>
<td>A. fulgidus</td>
<td>2.2</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>M. thermo</td>
<td>1.8</td>
<td>10.1</td>
<td></td>
</tr>
<tr>
<td>M. jannaschii</td>
<td>1.7</td>
<td>10.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Periods < 10.5 : Left-handed superhelices or Positive supercoiling  
Periods > 10.5 : Right-handed superhelices or Negative supercoiling  
H. Herzel et. al., Bioinformatics 1999
10.5 base periodicity in Prokaryotes

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: Spacings between any of these tetranucleotides
    - AAAA, AAAT, AATT, ATTT, TTTT
  - A2T2: Spacing between AA or TT
- Measure of periodicity
  - \( N(s) \): Count of AT/AT4/A2T2 at distance \( s \) from each other
  - \( n_s \): \# times a pair of A/T/G/C is found at distance \( s \)
- \( p \): Probability of finding the particular pattern at any given location in the DNA sequence (Eg: \( p = f_A + f_T \) for AT)
- Odds ratio, \( R(s) = \frac{N(s)}{E(s)} \)
  
J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT : Spacing between A or T
- Measure of periodicity
  - $N(s)$: Count of AT/AT4/A2T2 at distance $s$ from each other
  - $n_s$: #times a pair of A/T/G/C is found at distance $s$
  - $p$: probability of finding the particular pattern at any given location in the DNA sequence (Eg: $p = f_A + f_T$ for AT)
  - Odds ratio, $R(s) = \frac{N(s)}{E(s)}$ where $E(s) = n_s p^2$

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: Spacings between any of these tetranucleotides AAAA, AAAT, AATT, ATTT, TTTT

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: Spacings between any of these tetranucleotides AAAA, AAAT, AATT, ATTT, TTTT
  - A2T2: Spacing between AA or TT

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

▶ 1025 prokaryotic chromosomes
▶ Nucleotide spacing
  ▶ AT : Spacing between A or T
  ▶ AT4 : Spacings between any of these tetranucleotides
    AAAA, AAAT, AATT, ATTT, TTTT
  ▶ A2T2 : Spacing between AA or TT
▶ Measure of periodicity

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: Spacings between any of these tetranucleotides AAAA, AAAT, AATT, ATTT, TTTT
  - A2T2: Spacing between AA or TT
- Measure of periodicity
  - \( N(s) \): Count of AT/AT4/A2T2 at distance \( s \) from each other
- Odds ratio, \( R(s) = \frac{N(s)}{E(s)} \) where \( E(s) = n_s p^2 \)

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: Spacings between any of these tetranucleotides AAAA, AAAT, AATT, ATTT, TTTT
  - A2T2: Spacing between AA or TT
- Measure of periodicity
  - $N(s)$: Count of AT/AT4/A2T2 at distance $s$ from each other
  - $n_s$: #times a pair of A/T/G/C is found at distance $s$

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: spacings between any of these tetranucleotides AAAAA, AAAT, AATT, ATTT, TTTT
  - A2T2: Spacing between AA or TT
- Measure of periodicity
  - $N(s)$: Count of AT/AT4/A2T2 at distance $s$ from each other
  - $n_s$: #times a pair of A/T/G/C is found at distance $s$
  - $p$: probability of finding the particular pattern at any given location in the DNA sequence (Eg: $p = f_{A+T}$ for AT)

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes

- 1025 prokaryotic chromosomes
- Nucleotide spacing
  - AT: Spacing between A or T
  - AT4: Spacings between any of these tetranucleotides
    AAAAA, AAAT, AATT, ATTT, TTTT
  - A2T2: Spacing between AA or TT
- Measure of periodicity
  - \( N(s) \): Count of AT/AT4/A2T2 at distance \( s \) from each other
  - \( n_s \): #times a pair of A/T/G/C is found at distance \( s \)
  - \( p \): probability of finding the particular pattern at any given location in the DNA sequence (Eg: \( p = f_{A+T} \) for AT)
  - Odds ratio, \( R(s) = N(s)/E(s) \) where \( E(s) = n_sp^2 \)

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes: Measures

- $N(s)$: Count of AT/AT4/A2T2 at distance $s$ from each other
- $n_s$: #times a pair of A/T/G/C is found at distance $s$
- $p$: probability of finding the particular pattern at any given location in the DNA sequence (Eg: $p = f_{A+T}$ for AT)
- Odds ratio, $R(s) = N(s)/E(s)$ where $E(s) = n_s p^2$

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes: Measures

- $N(s)$: Count of AT/AT4/A2T2 at distance $s$ from each other
- $n_s$: #times a pair of A/T/G/C is found at distance $s$
- $p$: probability of finding the particular pattern at any given location in the DNA sequence (Eg: $p = f_{A+T}$ for AT)
- Odds ratio, $R(s) = N(s)/E(s)$ where $E(s) = n_sp^2$
- $R'(s) = [R(s-1) + R(s) + R(s+1)]/3$
  For removing 3-base periodicity

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes : Measures

- \( N(s) \): Count of AT/AT4/A2T2 at distance \( s \) from each other
- \( n_s \): \#times a pair of A/T/G/C is found at distance \( s \)
- \( p \): probability of finding the particular pattern at any given location in the DNA sequence (Eg: \( p = f_{A+T} \) for AT)
- Odds ratio, \( R(s) = \frac{N(s)}{E(s)} \) where \( E(s) = n_s p^2 \)
- \( R'(s) = \frac{[R(s - 1) + R(s) + R(s + 1)]}{3} \)
  For removing 3-base periodicity
- \( R*(s) = R'(s) - (As^2 + Bs + C) \)
  A,B, C determined by least square fitting

J. Mrazek, Journal of Bacteriology 2010
10.5 base periodicity in Prokaryotes : Measures

- $N(s)$: Count of AT/AT4/A2T2 at distance $s$ from each other
- $n_s$: #times a pair of A/T/G/C is found at distance $s$
- $p$: probability of finding the particular pattern at any given location in the DNA sequence (Eg: $p = f_{A+T}$ for AT)
- Odds ratio, $R(s) = N(s)/E(s)$ where $E(s) = n_s p^2$
- $R'(s) = [R(s-1) + R(s) + R(s+1)]/3$
  For removing 3-base periodicity
- $R\ast(s) = R'(s) - (As^2 + Bs + C)$
  A, B, C determined by least square fitting
- $Q\ast(P)$ is the normalized Fourier Transform of $R\ast(s)$ taken from $s_{min} = 30$ up to $s_{max} = 100$

J. Mrazek, Journal of Bacteriology 2010
J. Mrazek, Journal of Bacteriology 2010
Nucleosome : Eukaryote
Supercoiling: Prokaryote

- Twist = -1, Writhe = 0.
- Twist = 0, Writhe = -1.
- Twist = 0, Writhe = 0.
- Twist = 0, Writhe = -2.
- Twist = 0, Writhe = 0.
- Twist = 0, Writhe = 2.
- Twist = 0, Writhe = 0.
- Twist = 0, Writhe = 4.

Plectonemic
Toroidal
200 base periodicity

- Nucleosome: 145 nt in core particle + 55 nt in linker DNA
- Revealed in size of DNA fragment obtained after digestion
  - Nucleases cleave at the more accessible linker DNA
- Nucleosome center-center distance: 200 ± 50

E. N. Trifonov, Physica A 1998