Sequence Logos and the Helix of DNA

Thomas D. Schneider, Ph.D.

Frederick National Laboratory for Cancer Research
Gene Regulation and Chromosome Biology Laboratory

Molecular Information Theory Group
<table>
<thead>
<tr>
<th>number of symbols</th>
<th>number of bits</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>H T</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>11 10 01 00</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\[ M = 2^B \quad B = \log_2 M \]
<table>
<thead>
<tr>
<th>number of symbols</th>
<th>number of bits</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>H T</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>11 10 01 00</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\[ M = 2^B \quad B = \log_2 M \]
## Information Theory: One-Minute Lesson

<table>
<thead>
<tr>
<th>number of symbols</th>
<th>number of bits</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>$B$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$H,T$</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>$11,10,01,00$</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

$M=2^B, B=\log_2M$
<table>
<thead>
<tr>
<th>number of symbols</th>
<th>number of bits</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>H T</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>11 10 01 00</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\[ M = 2^B \quad B = \log_2 M \]
## Information Theory: One-Minute Lesson

<table>
<thead>
<tr>
<th>number of symbols</th>
<th>number of bits</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>$B$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>H, T</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>11, 10, 01, 00</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

$M = 2^B$

$B = \log_2 M$
Gallery of DNA Binding Sites

12 Lambda cI and cro binding sites

8 Lambda O protein binding sites

12 434 cI and cro binding sites

34 ArgR binding sites

58 CRP binding sites

8 TrpR binding sites

14 FNR binding sites

38 LexA binding sites
El Duomo, Florence, Italy
Gallery of DNA Binding Sites

12 Lambda cI and cro binding sites

8 Lambda O protein binding sites

12 434 cI and cro binding sites

34 ArgR binding sites

58 CRP binding sites

8 TrpR binding sites

14 FNR binding sites

38 LexA binding sites
Gallery of DNA Binding Sites with Waves

12 Lambda cI and cro binding sites

12 434 cI and cro binding sites

8 Lambda O protein binding sites

34 ArgR binding sites

58 CRP binding sites

8 TrpR binding sites

14 FNR binding sites

38 LexA binding sites

Wavelength: 10.6 base pairs per turn
Gallery of DNA Binding Sites with Waves

Wave Peak corresponds to protein facing Major Groove

wavelength: 10.6 base pairs per turn
Wave Peak corresponds to protein facing Major Groove

Wave Trough corresponds to protein facing Minor Groove

wavelength: 10.6 base pairs per turn

12 Lambda cI and cro binding sites

8 Lambda O protein binding sites

12 434 cI and cro binding sites

34 ArgR binding sites

58 CRP binding sites

8 TrpR binding sites

14 FNR binding sites

38 LexA binding sites

Gallery of DNA Binding Sites with Waves
Gallery of DNA Binding Sites with Waves

Wave Peak corresponds to protein facing Major Groove

Wave Trough corresponds to protein facing Minor Groove

wavelength: 10.6 base pairs per turn

Why the match?
DNA Bases - Minor Groove Rule

Major Groove

Minor Groove

Major Groove

Minor Groove

Arnott & Hukins, Biochem. and Biophys. Res. Comm., 1972, 47: 1504-1509
The **MAJOR groove** has 4 distinct patterns so it can have up to **TWO BITS** of information.
DNA Bases - Minor Groove Rule

The **MAJOR groove** has 4 distinct patterns so it can have up to **TWO BITS** of information.

The **MINOR groove** has only 2 distinct patterns so it can only have up to **ONE BIT** of information.

Arnott & Hukins, Biochem. and Biophys. Res. Comm., 1972, 47: 1504-1509
Major and Minor groove contacts explain peak locations but . . .
Why do logos often smoothly follow the wave?

Major and Minor groove contacts explain peak locations but . . .
DNA Access

easiest

class difficult

most difficult
Skeleton Logos - just error bars
Skeleton Logos - just error bars

MAJOR groove accessibility curve
Skeleton Logos - just error bars

MAJOR groove accessibility curve

MINOR groove accessibility curve
Triangular area is empty ⇒ OR instead of Sum
Stuart Austin and Ann Abeles found binding sites... make a logo...

14 bacteriophage P1 RepA binding sites

Stuart Austin and Ann Abeles found binding sites...

14 bacteriophage P1 RepA binding sites

Violation of the 1 bit Minor Groove rule!

G: Methylation interference
●: hydroxyl radical footprint

P. P. Papp, D. K. Chattoraj and T. D. Schneider,
Information analysis of sequences that bind the replication, initiator RepA,
RepA orientation data

G: Methylation interference
●: hydroxyl radical footprint

P. P. Papp, D. K. Chattoraj and T. D. Schneider,
Information analysis of sequences that bind the replication, initiator RepA,

G: Methylation interference
●: hydroxyl radical footprint

P. P. Papp and D. K. Chattoraj,
Missing-base and ethylation interference footprinting of P1, plasmid replication initiator,
Orientation of proteins on DNA
Yeast Saccharomyces cerevisiae GAL4 Sequence Logo

22 Gal4 binding sites

Yeast Saccharomyces cerevisiae GAL4 Sequence Logo

22 Gal4 binding sites

This does not explain the RepA anomaly!

E. coli IHF Sequence Logo

27 IHF binding sites
△: IHF inserts a proline into the DNA to induce a bend

Minor groove binding - cracked open DNA

Rice et al. Cell, 1996, 87:1295-1306
E. coli IHF Sequence Logo

27 IHF binding sites

△: IHF inserts a proline into the DNA to induce a bend

Minor groove binding - cracked open DNA

This does not explain the RepA anomaly!

Rice et al. Cell, 1996, 87:1295-1306
HhaI methyltransferase Base Flipping

**DNA Replication Protein Rts1**

**PROKARYOTIC PLASMID ORIGIN:**

15 Rts1 RepA binding sites

<table>
<thead>
<tr>
<th>Bits</th>
<th>5'</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>3'</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
</tr>
</tbody>
</table>

**3'**

**5'**
A) PROKARYOTIC CHROMOSOMAL ORIGIN:
29 E. coli DnaA binding sites

B) PROKARYOTIC PLASMID ORIGIN:
11 P4 replicon binding sites

C) EUKARYOTIC PLASMID ORIGIN:
52 Epstein-Barr Virus EBNA1 sites

D) EUKARYOTIC CHROMOSOMAL ORIGIN:
18 S. cerevisiae ORC A sites
Sequence logos for $\sigma^{70}$ RNA transcription

Sequence logos for $\sigma^{70}$ RNA transcription

Our prediction of base 4 flipping was confirmed!
Feklistov & Darst, Cell, 2011, 147: 1257-1269

Experiment: change base ±7 on both strands.

Variant nucleotides at ±7 of the P1 RepA binding site

The four natural base pairs are boxed.
Gel Mobility Shift Assay of ±7 variant P1 RepA sites

Student’s t-test of ±7 variant P1 RepA site binding

The break separates the moieties into two classes

Variant nucleotides at ±7 of the P1 RepA binding site

A Thymine at position +7 is important

RepA interacts with the base at position +7 for two reasons: removing Thymine +7 leads to a decrease in the binding affinity, while removing Adenine +7′ does not.
The C5 methyl group does not interact with RepA because removing or changing it only has a slight effect on the binding affinity.
The O2 group is not important for RepA binding because 13.C/A (which has one) and 23.abasic/A (which does not) have similar binding energy.
The O4 group is not important because 9.N4-Me-C/A does not have an O4 group as does 1.T/A, but the ΔΔG is low.
A 7’ amino group (purple) inhibits RepA binding.
A +7 amino group (red) inhibits RepA binding more strongly than a +7′ amino group.
An important group that interacts with RepA is the N3 proton of Thymine or, in modified bases, a proton near the N3 atom. All bases that have a low Kd, below the distinct step, have a proton in this place.
• A. 1.T/A binds RepA stronger than 13.C/A.
• A. 1.T/A binds RepA stronger than 13.C/A.
• This low binding can be rescued by substituting the N4 proton with the methyl group in N4-Me-C.
A. 1.T/A binds RepA stronger than 13.C/A.

B. Two more structures that have a C-H proton near the N3 atom also bind RepA well.
• A. 1.T/A binds RepA stronger than 13.C/A.
• This low binding can be rescued by substituting the N4 proton with the methyl group in N4-Me-C.
• B. Two more structures that have a C-H proton near the N3 atom also bind RepA well.
• C. Outlines of the structures in A and B were aligned. Arrows point in the direction from which RepA would have to approach to form hydrogen bonds with the proton (green circles). Arrows with a red circle show contacts that are poor.

Conclusions

- Sequence logo sine waves indicate protein orientation on DNA
Conclusions

- Sequence logo sine waves indicate protein orientation on DNA
- Anomalies in the logo can reveal interesting binding modes
Conclusions

• Sequence logo sine waves indicate protein orientation on DNA

• Anomalies in the logo can reveal interesting binding modes

• Flipping a base out of DNA initiates DNA replication in bacteriophage P1 RepA and other DNA binding origin proteins
Acknowledgments

- Dhruba Chattoraj
- Peter Papp
- Rich Roberts
- Denise Rubens
- Paul Hengen
- Ilya Lyakhov

National Institutes of Health, National Cancer Institute

The Great Wave off Kanagawa by Katsushika Hokusai 1829-32
version = 1.24 of helixtalk.tex 2012 Nov 07