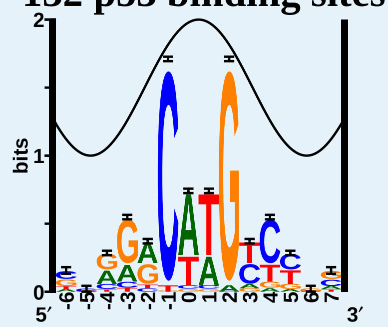
Sequence Logos and the Helix of DNA

Thomas D. Schneider, Ph.D.

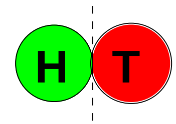
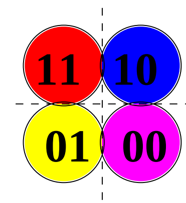
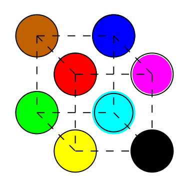
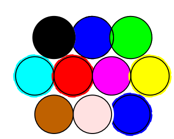
Frederick National Laboratory for Cancer Research
Gene Regulation and Chromosome Biology Laboratory

Molecular Information Theory Group

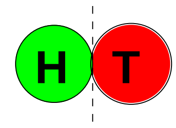
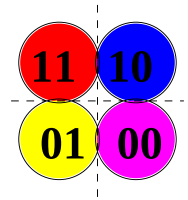
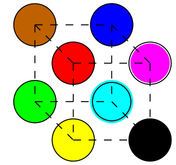
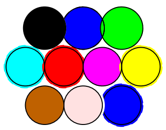
132 p53 binding sites

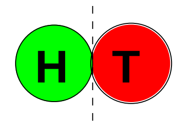
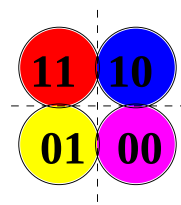
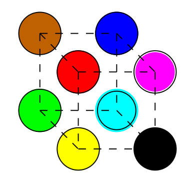
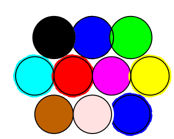


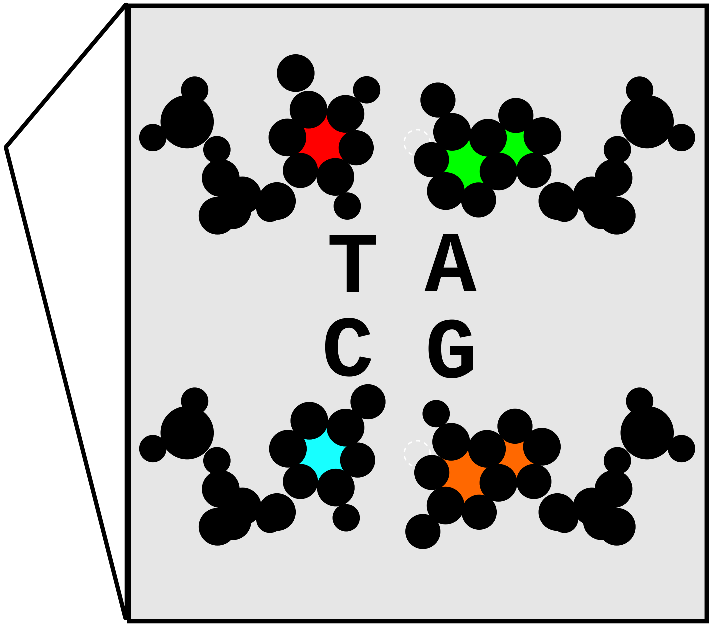
Information Theory: One-Minute Lesson

number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

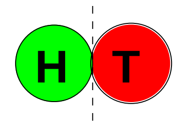
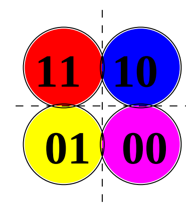
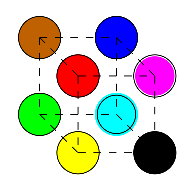
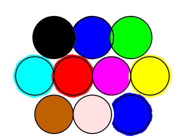
Information Theory: One-Minute Lesson

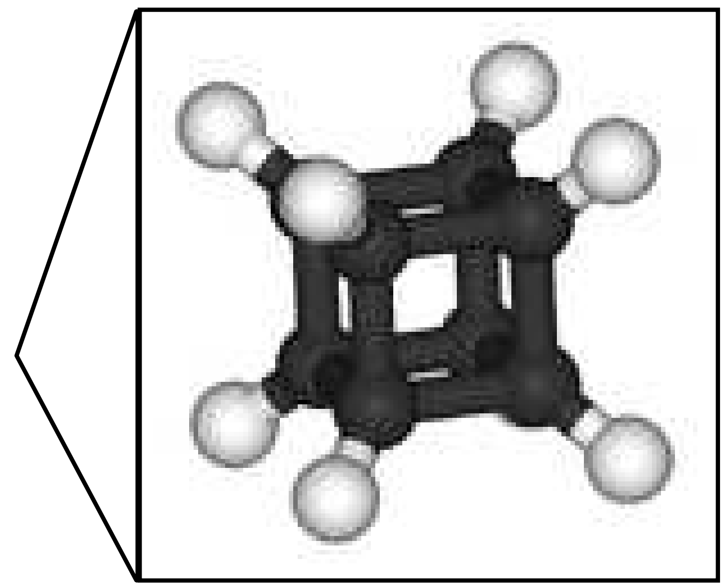
number of symbols	number of bits	example
M	B	
2	1	 
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

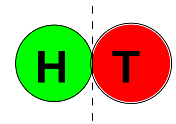
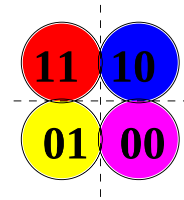
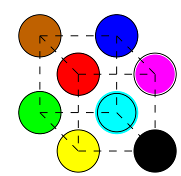
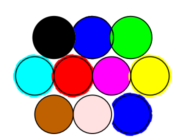


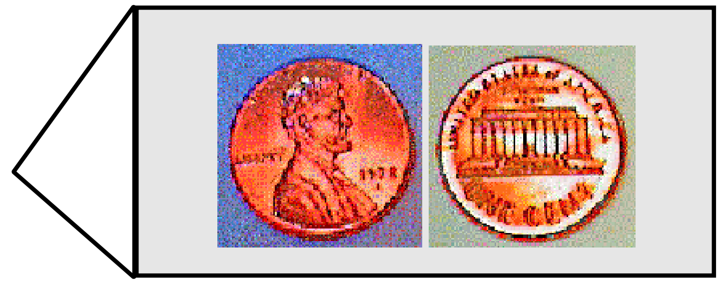
Information Theory: One-Minute Lesson

number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	



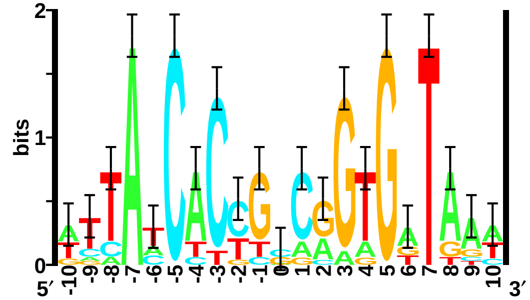
Information Theory: One-Minute Lesson

number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

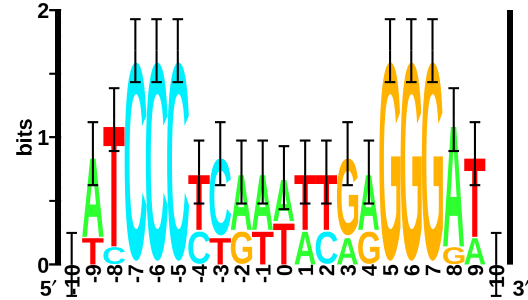


Gallery of DNA Binding Sites

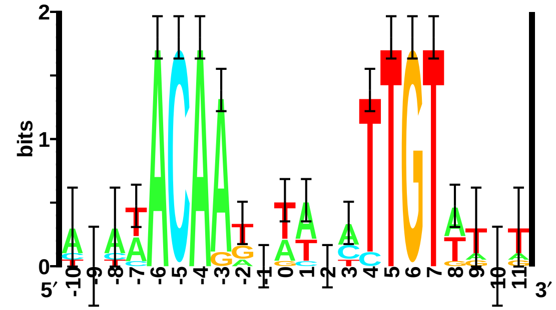
12 Lambda cI and cro binding sites



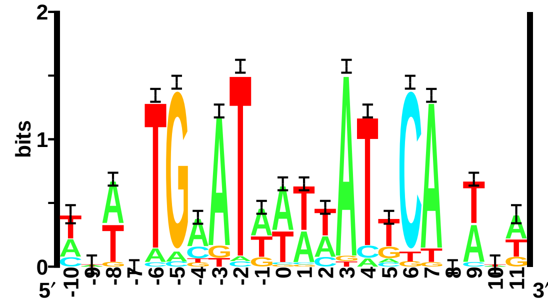
8 Lambda O protein binding sites



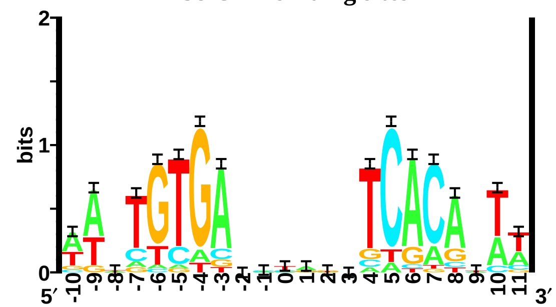
12 434 cI and cro binding sites



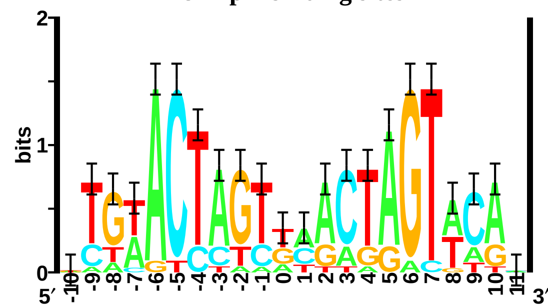
34 ArgR binding sites



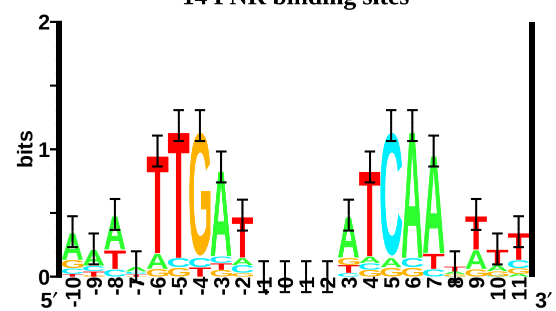
58 CRP binding sites



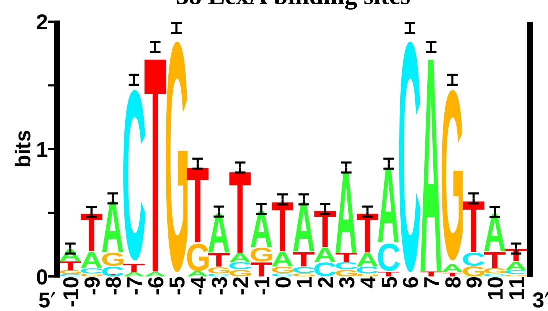
8 TrpR binding sites



14 FNR binding sites

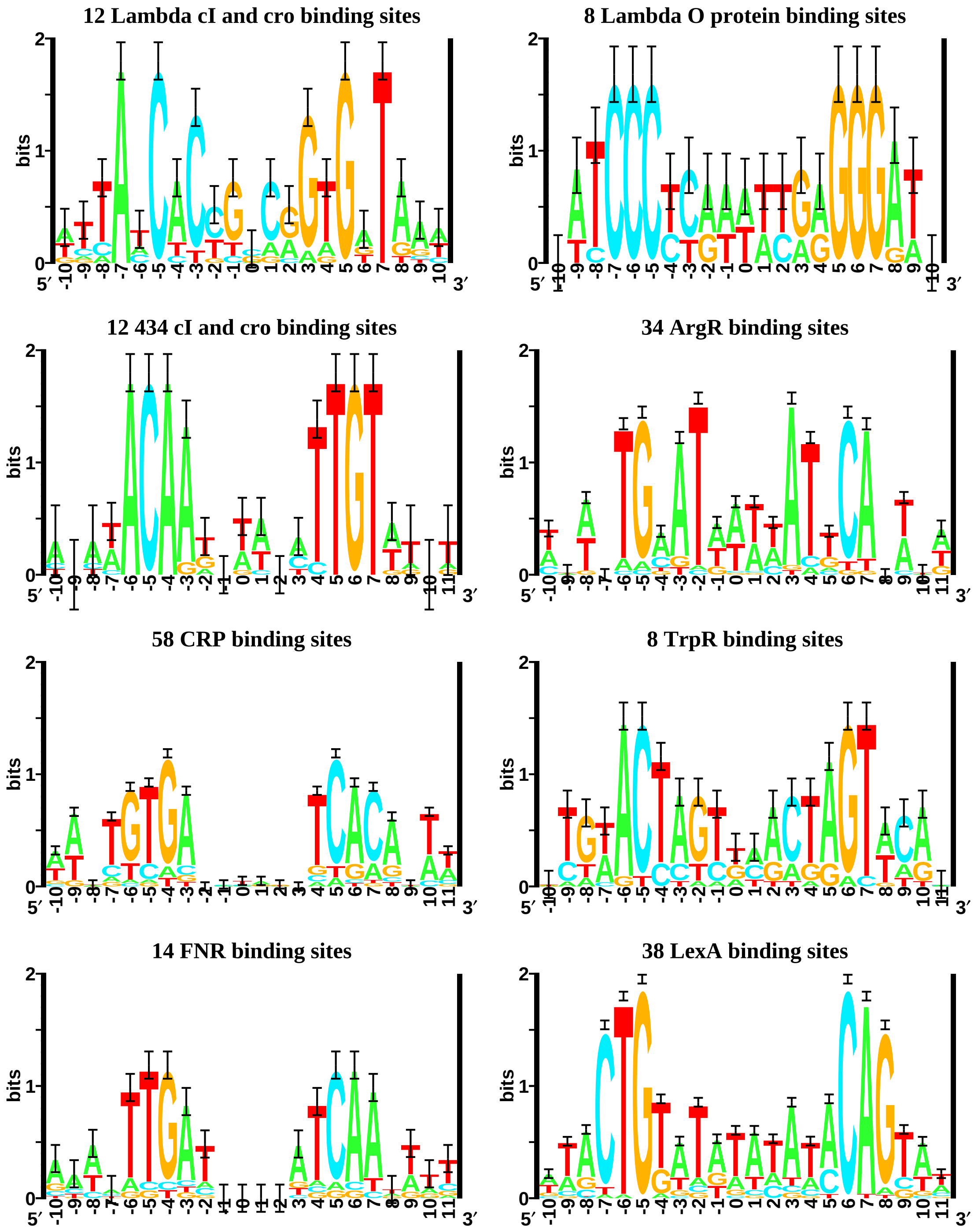


38 LexA binding sites

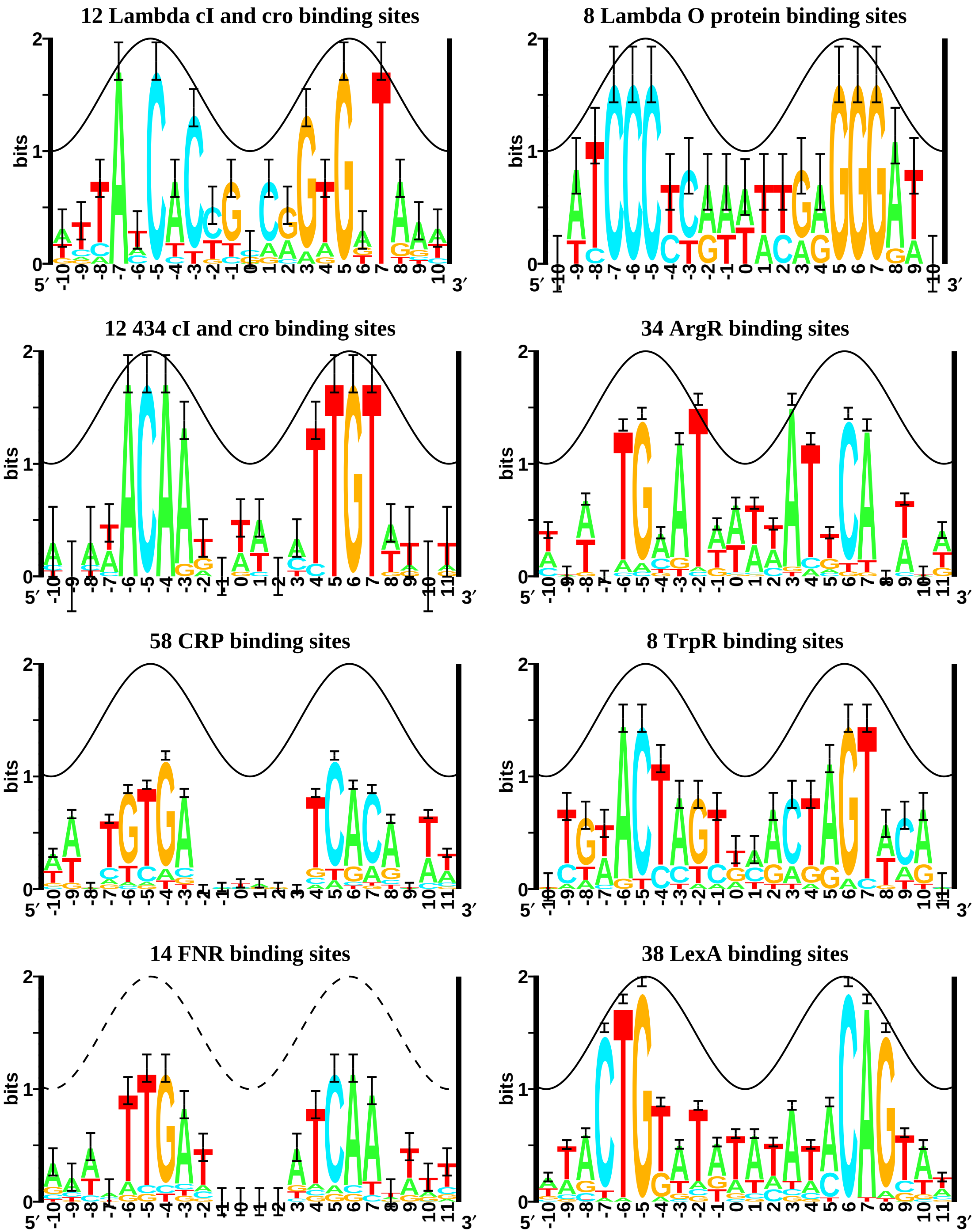


El Duomo, Florence, Italy

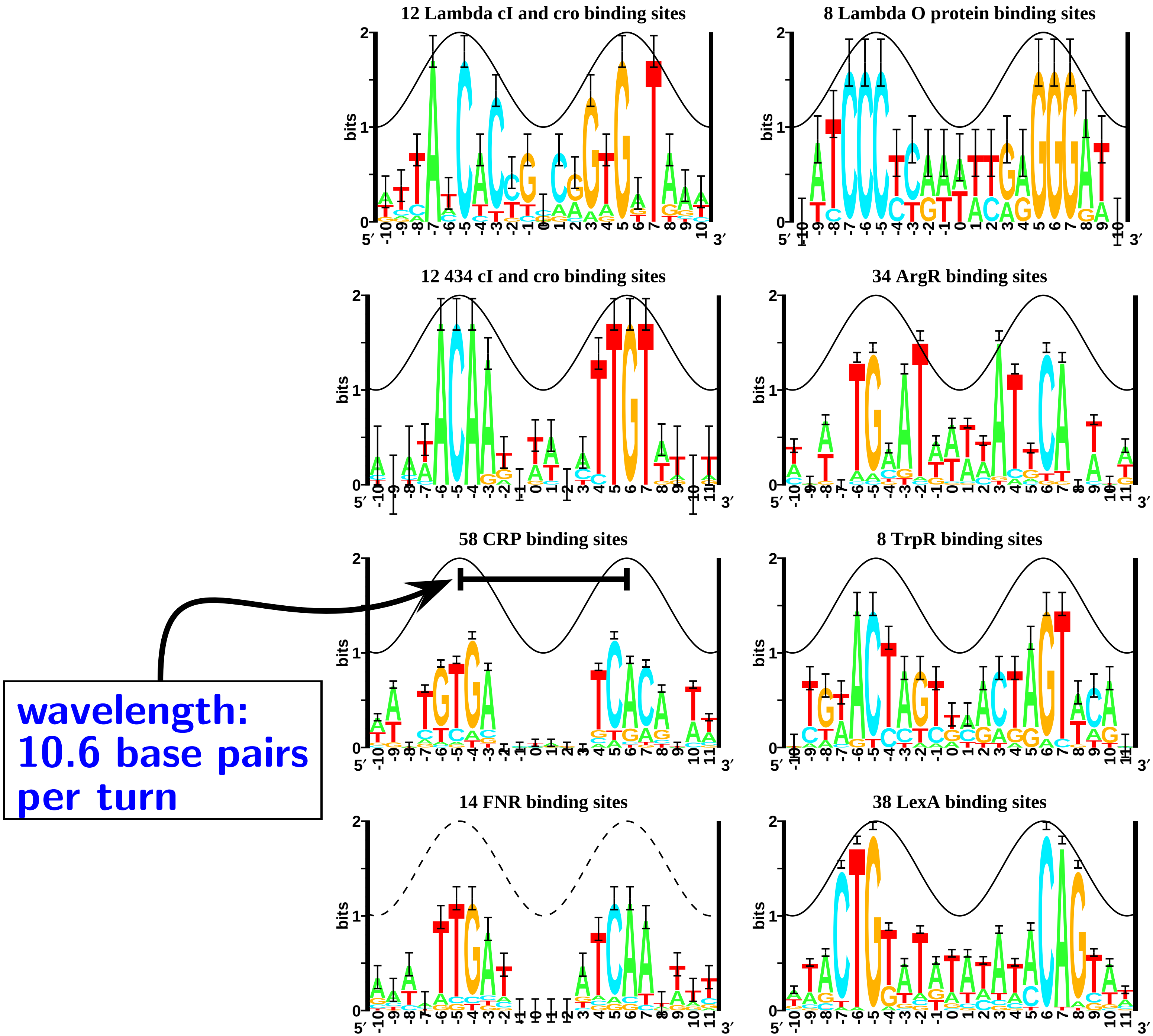
Gallery of DNA Binding Sites



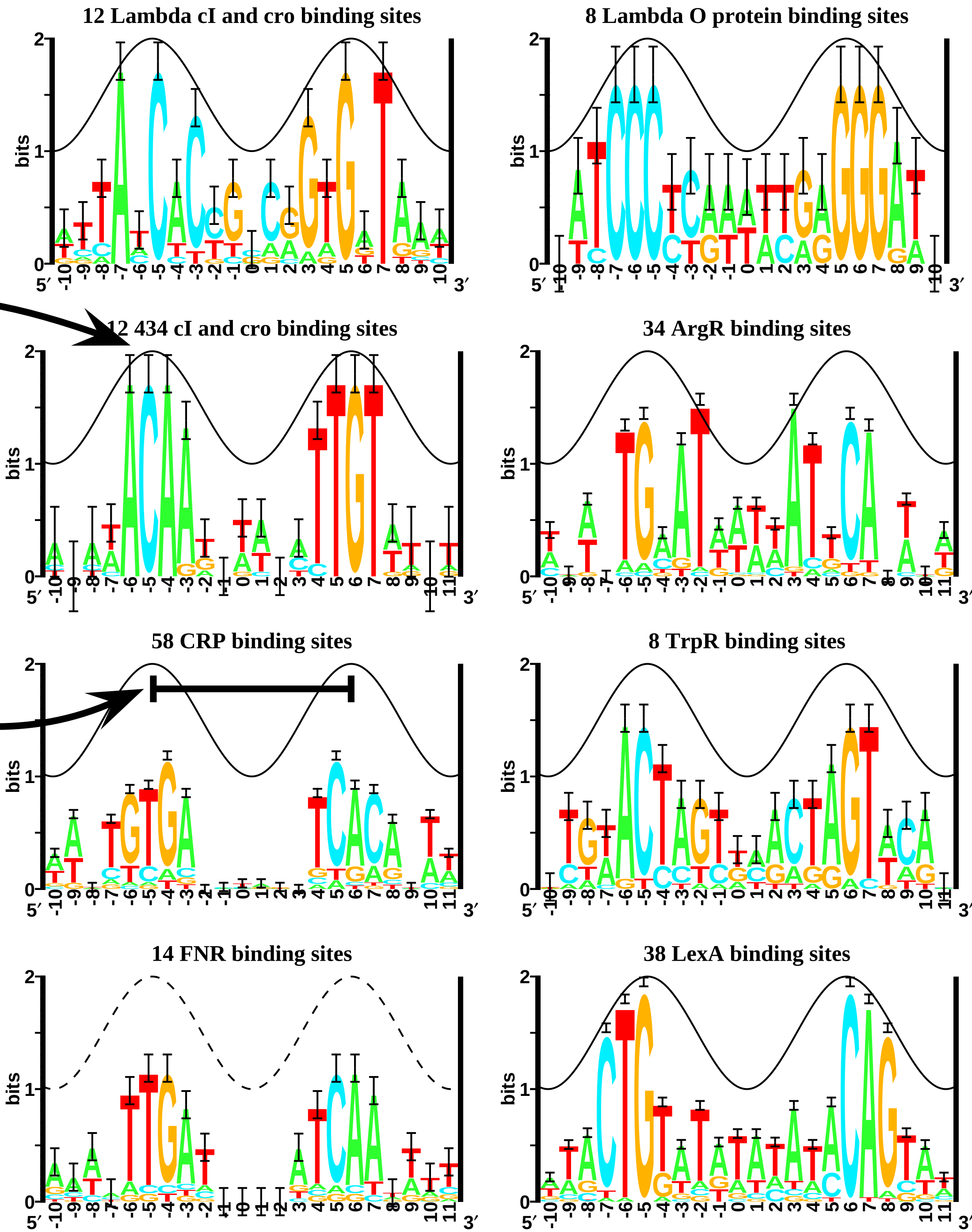
Gallery of DNA Binding Sites with Waves



Gallery of DNA Binding Sites with Waves



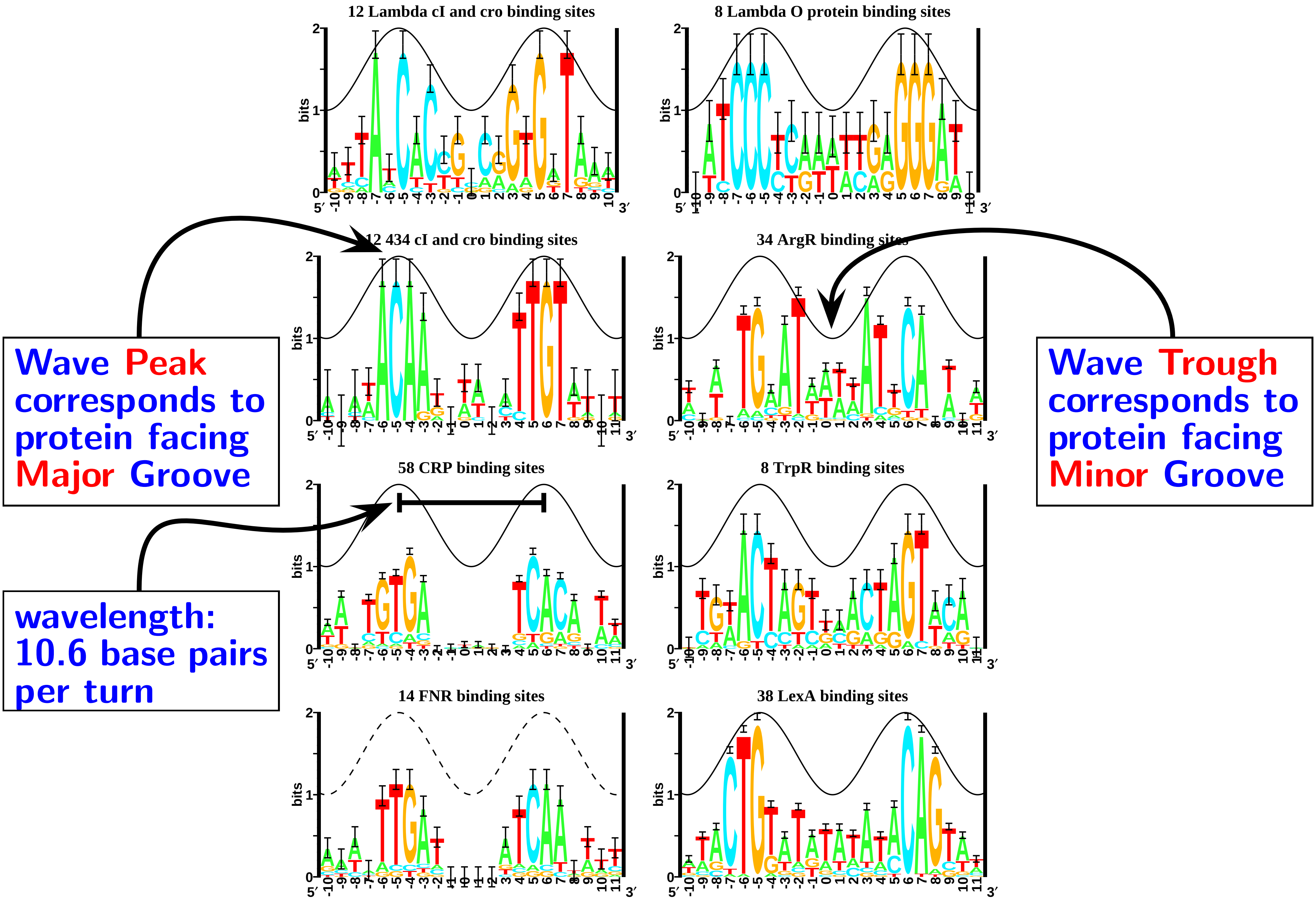
Gallery of DNA Binding Sites with Waves



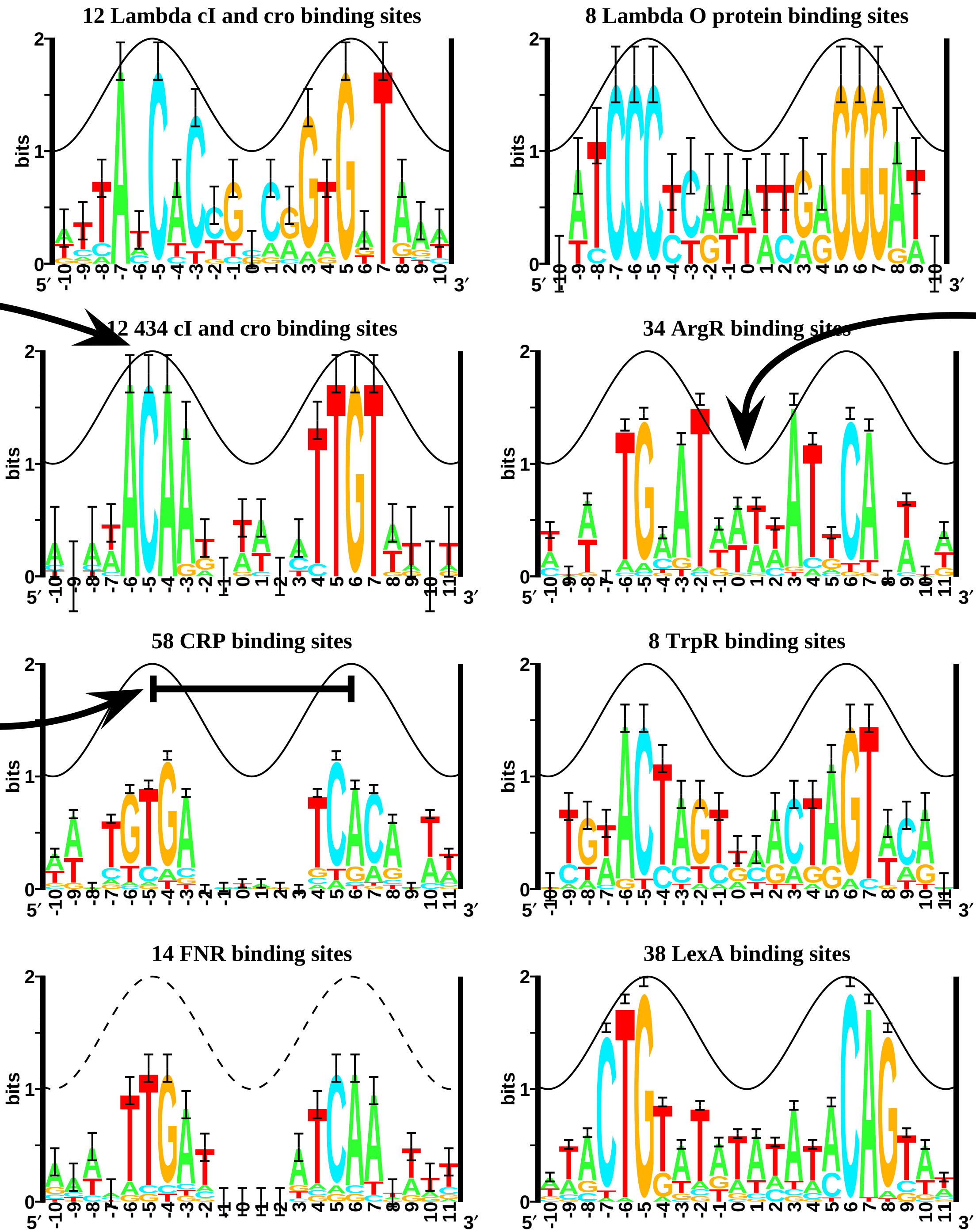
Wave Peak corresponds to protein facing Major Groove

wavelength: 10.6 base pairs per turn

Gallery of DNA Binding Sites with Waves



Gallery of DNA Binding Sites with Waves



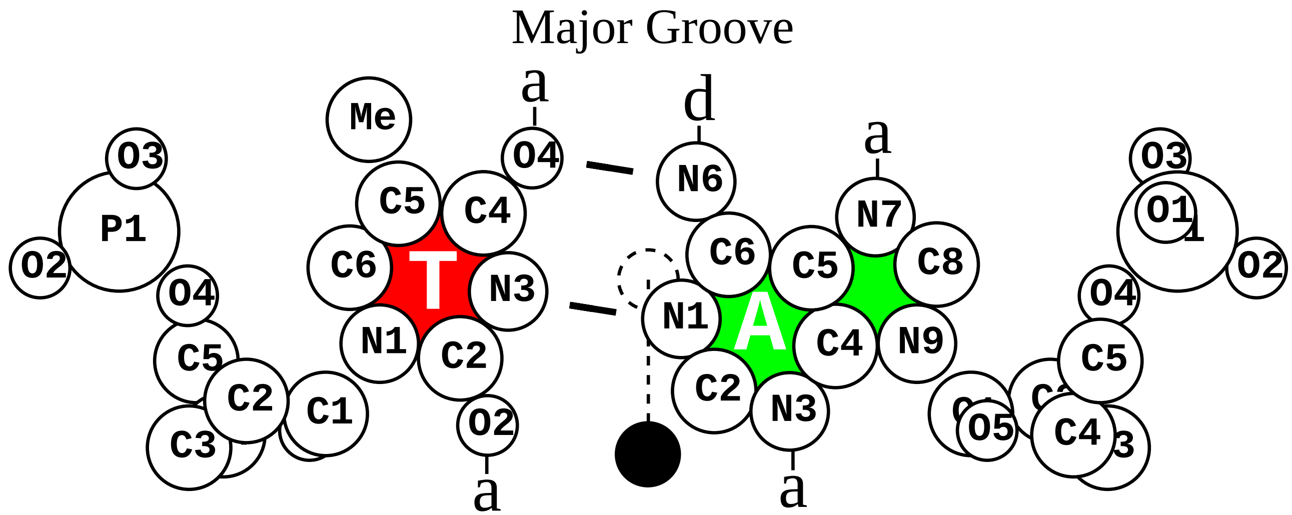
Wave Peak corresponds to protein facing Major Groove

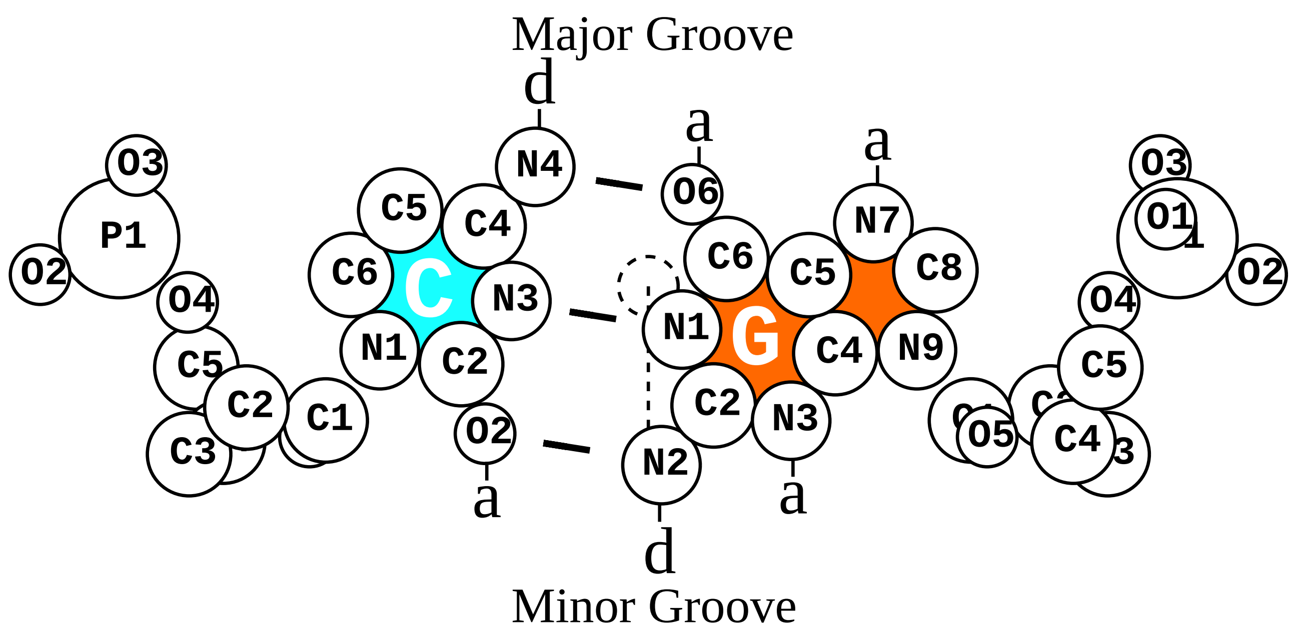
wavelength: 10.6 base pairs per turn

Wave Trough corresponds to protein facing Minor Groove

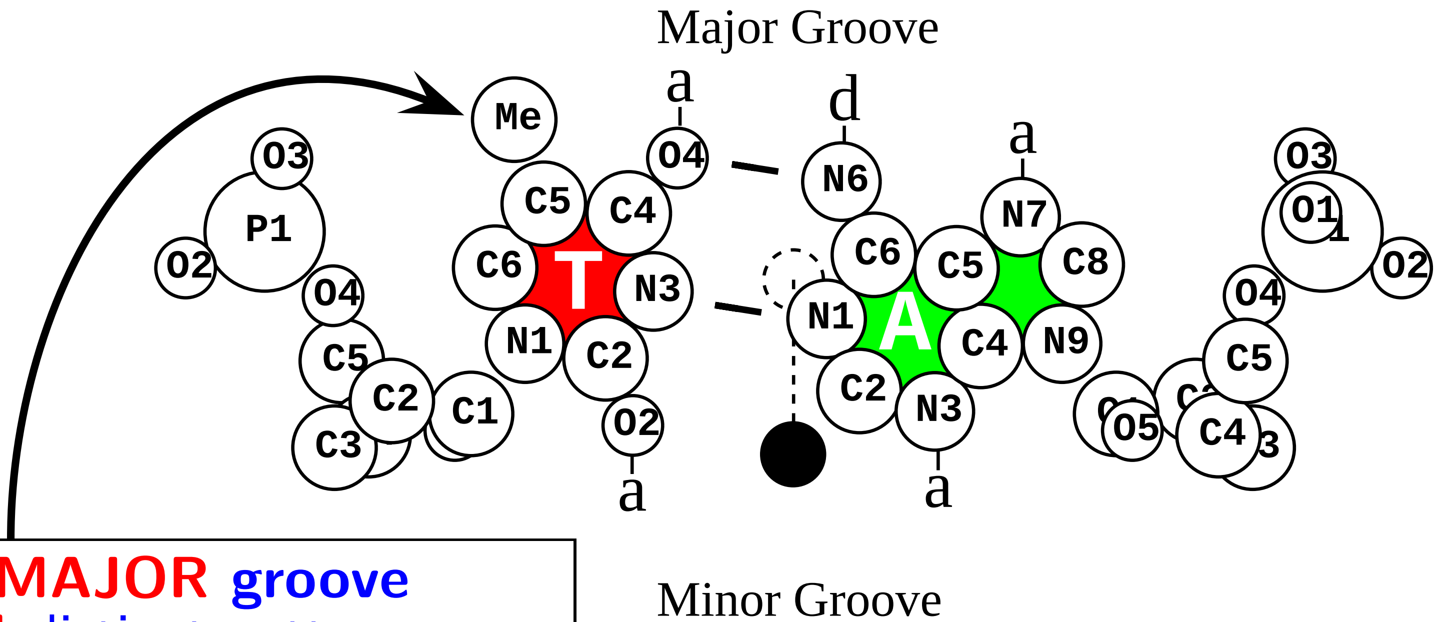
Why the match?

DNA Bases - Minor Groove Rule

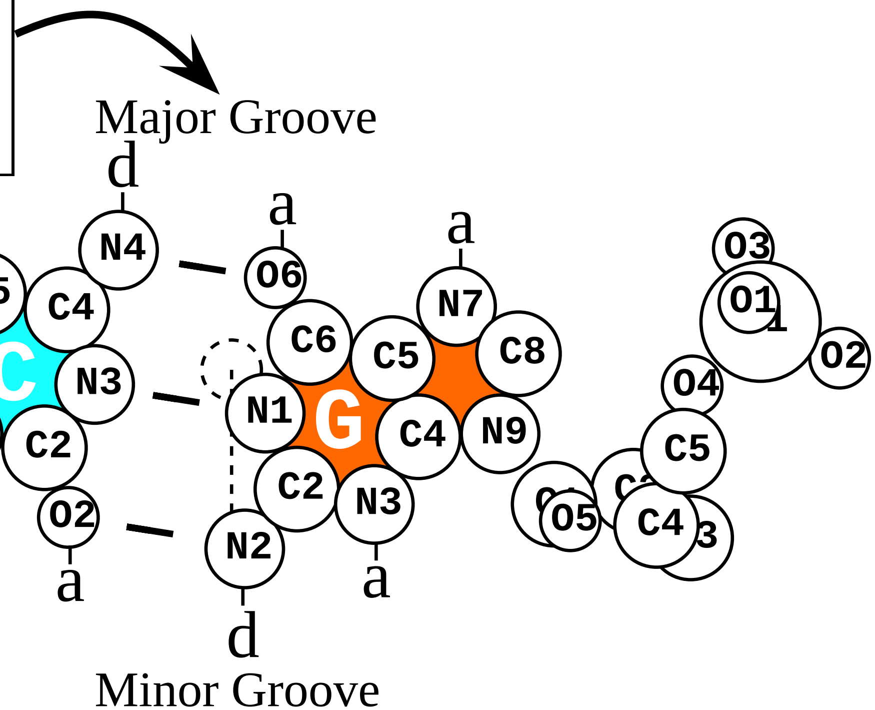




DNA Bases - Minor Groove Rule

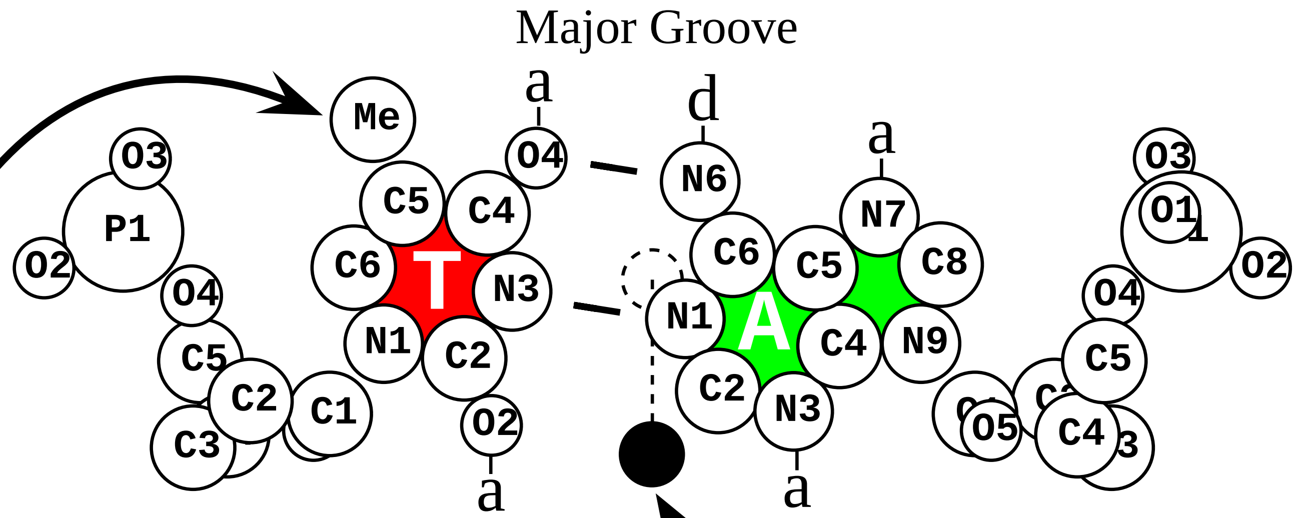


The **MAJOR groove**
has **4** distinct patterns
so it can have up to
TWO BITS of information



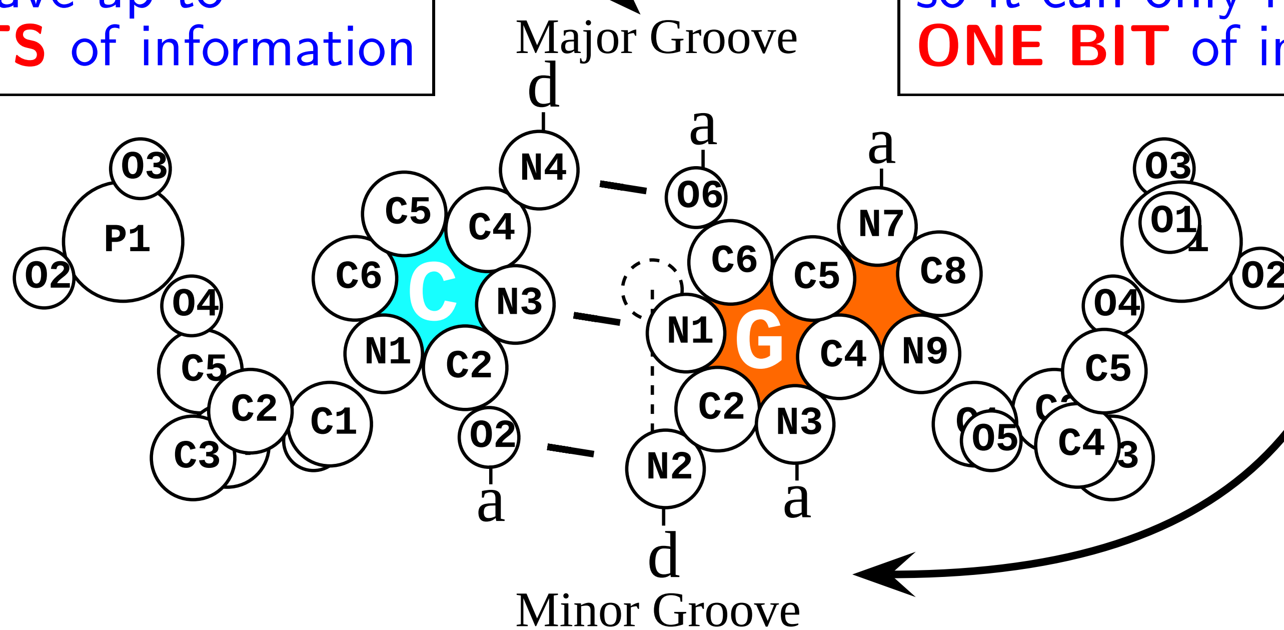
Arnott & Hukins, Biochem. and Biophys. Res. Comm., 1972, 47: 1504-1509

DNA Bases - Minor Groove Rule



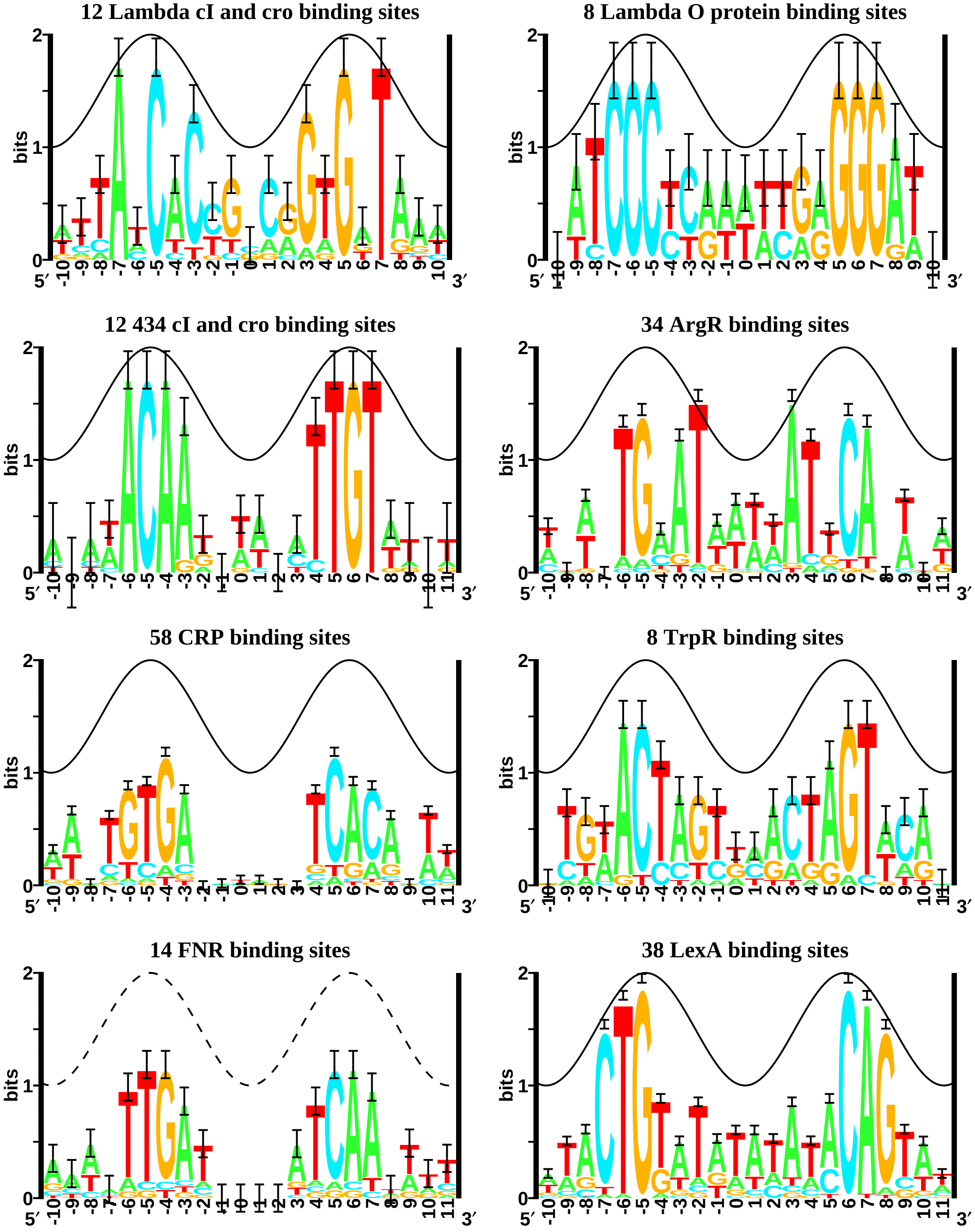
The **MAJOR groove**
has **4** distinct patterns
so it can have up to
TWO BITS of information

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



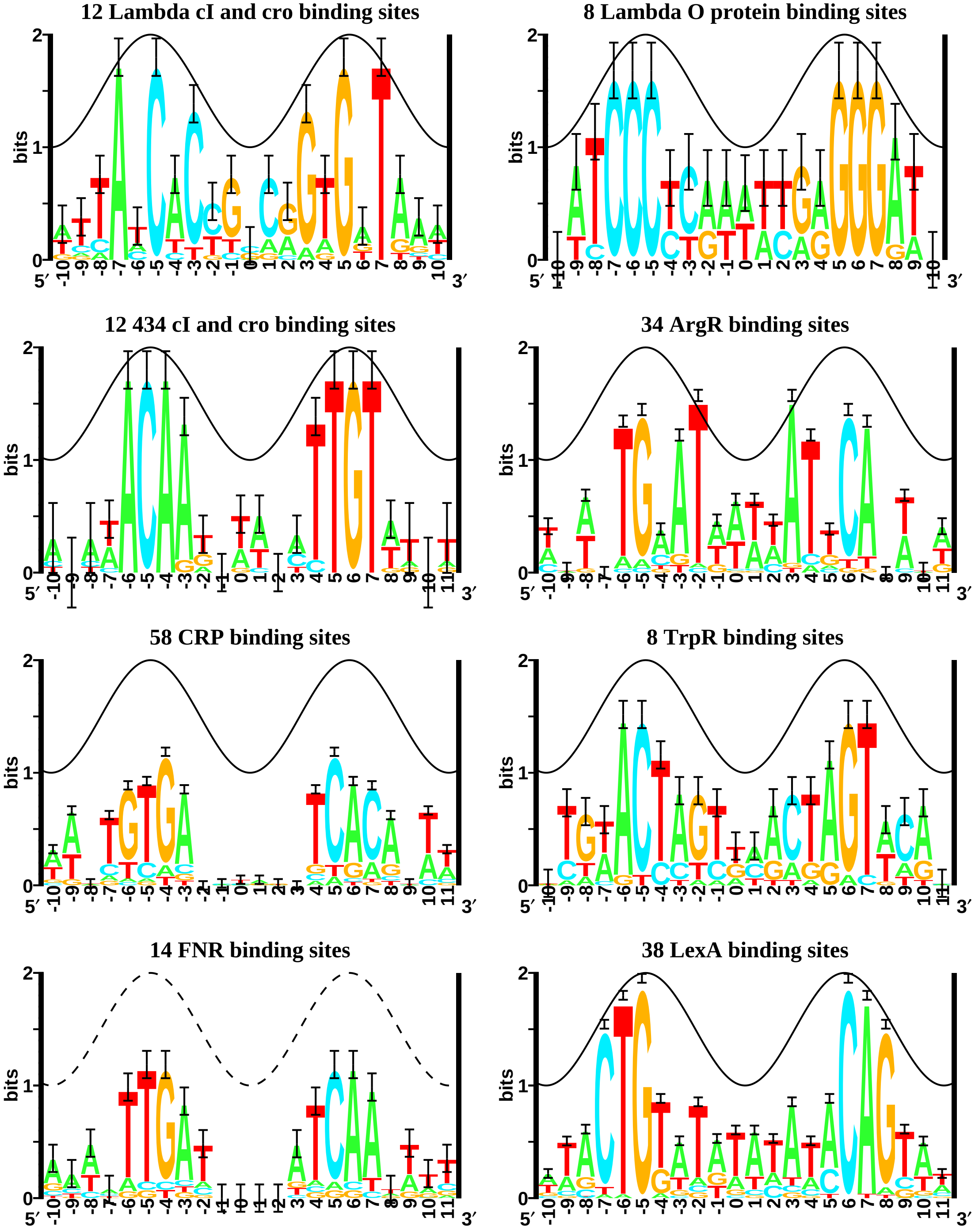
Gallery of DNA Binding Sites with Waves

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

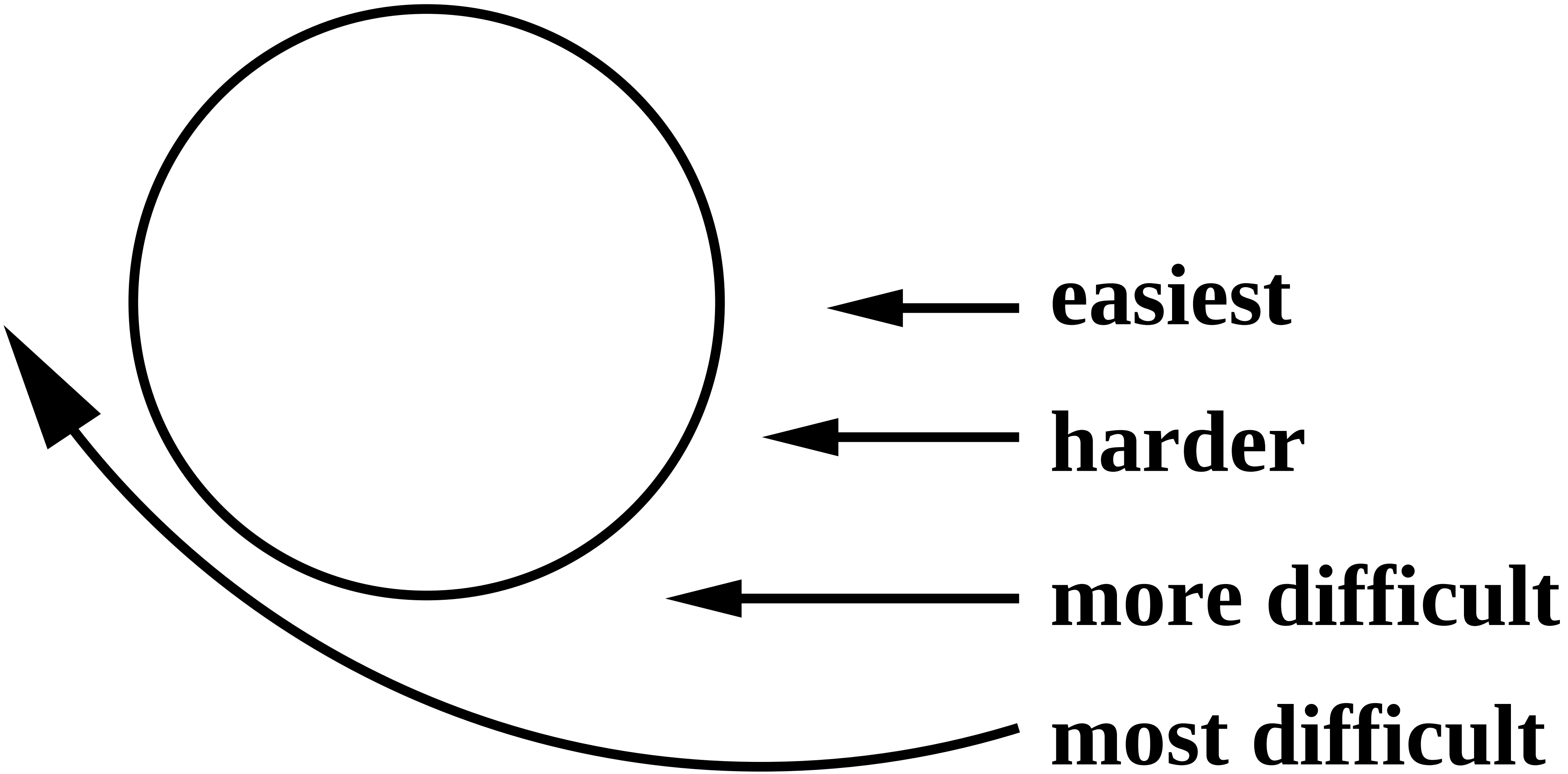


Gallery of DNA Binding Sites with Waves

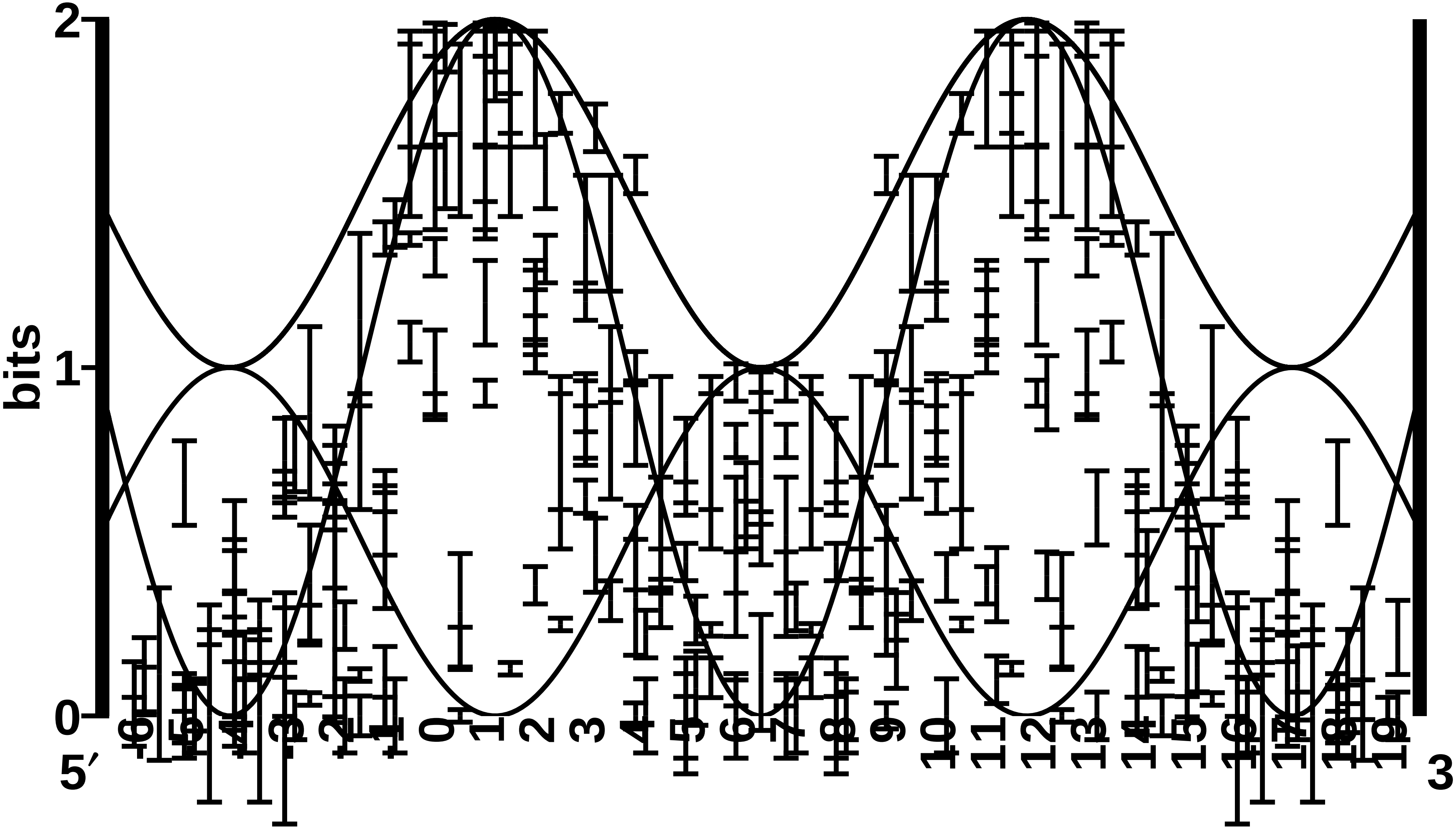
Major and Minor groove contacts explain peak locations but . . .



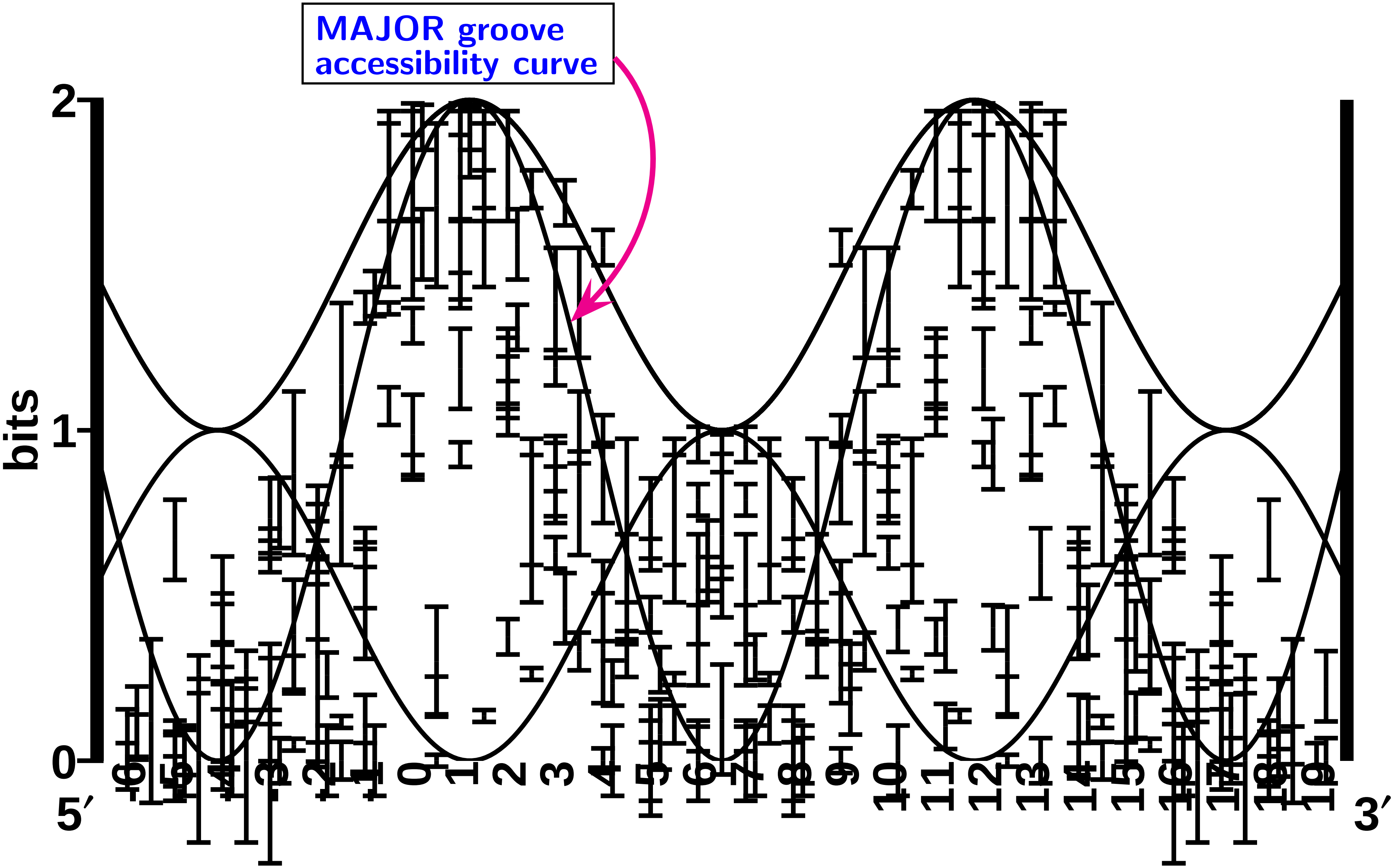
Why do logos often smoothly follow the wave?


easiest
harder
more difficult
most difficult

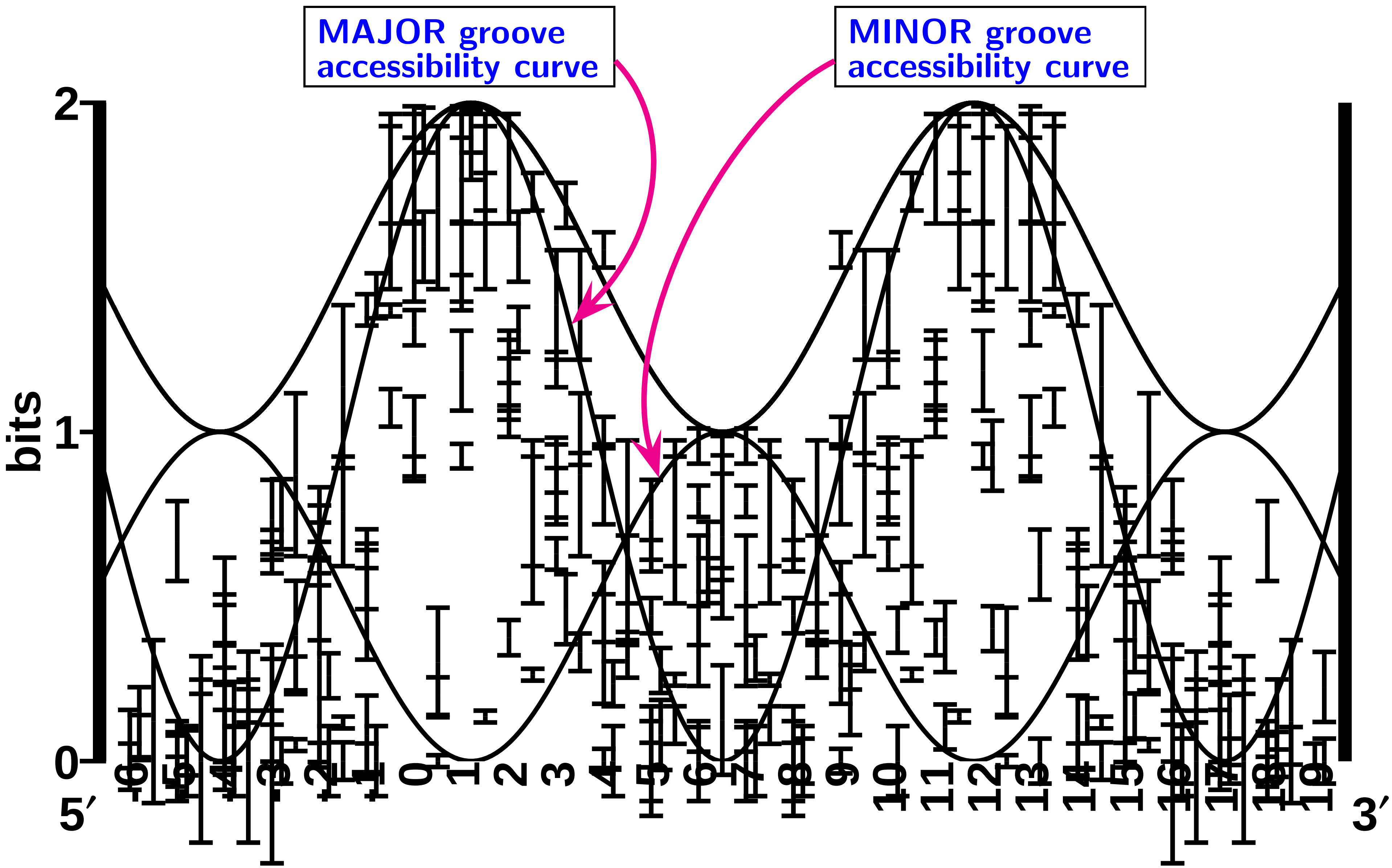
Skeleton Logos - just error bars



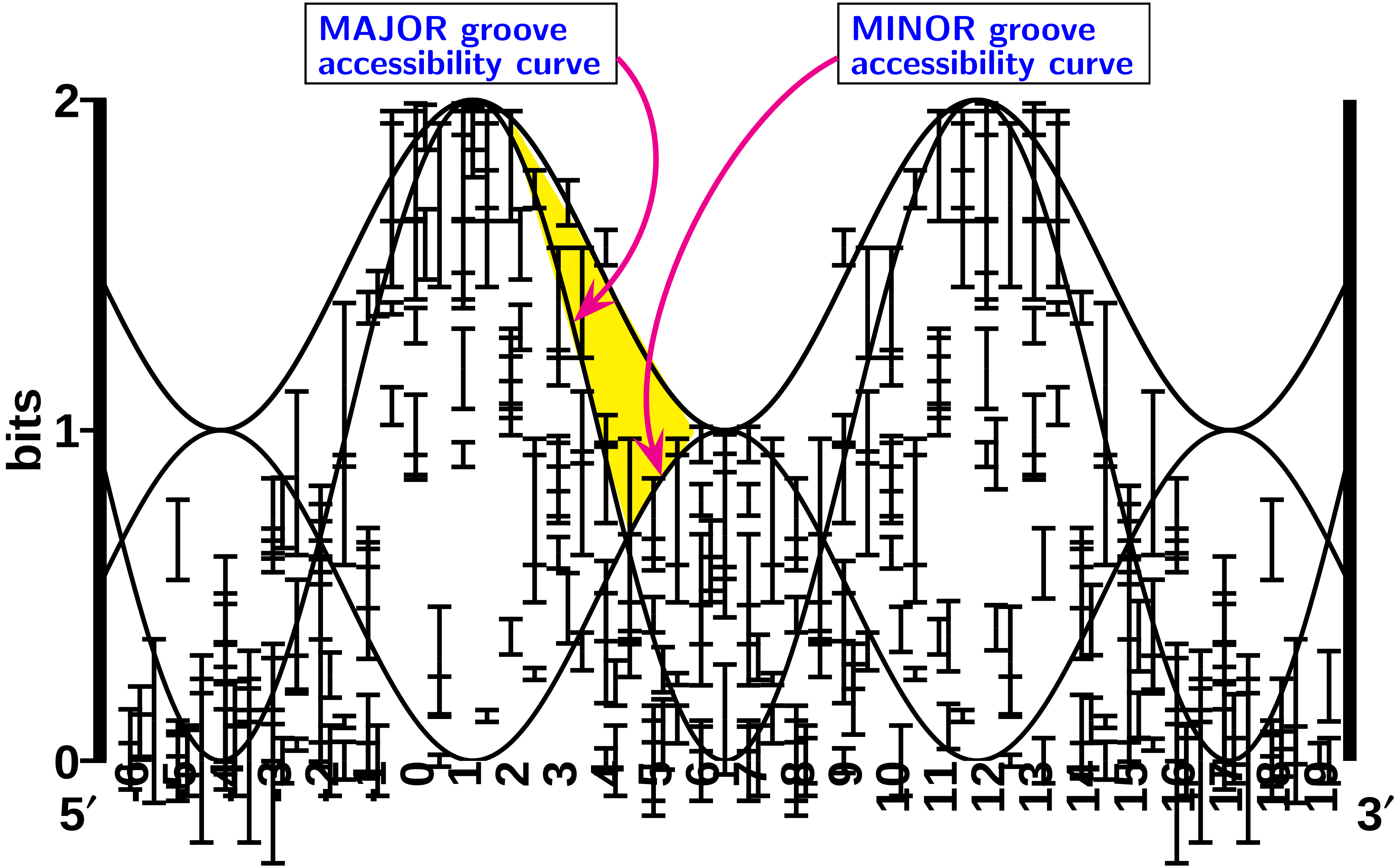
Skeleton Logos - just error bars



Skeleton Logos - just error bars



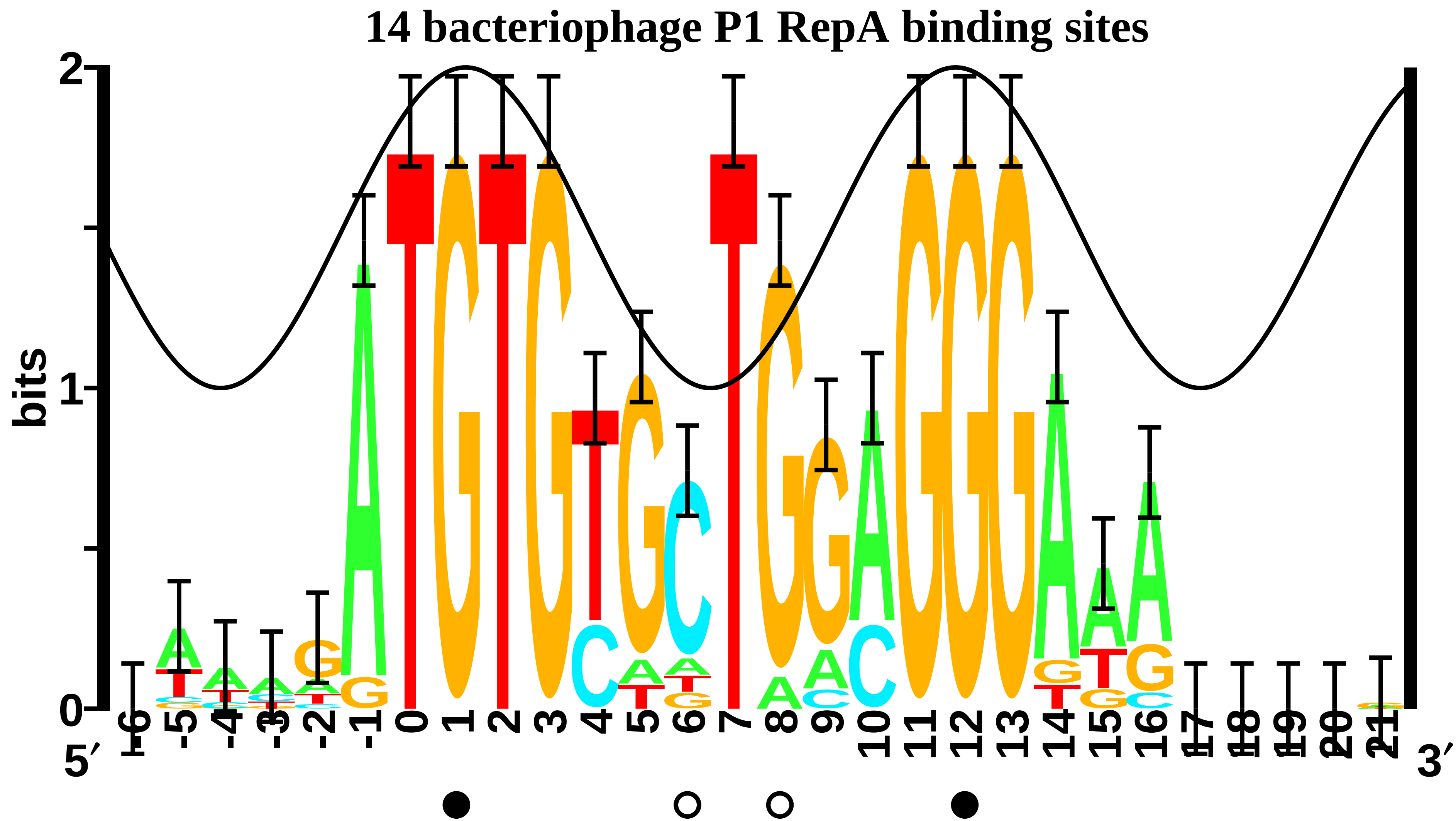
Skeleton Logos - just error bars



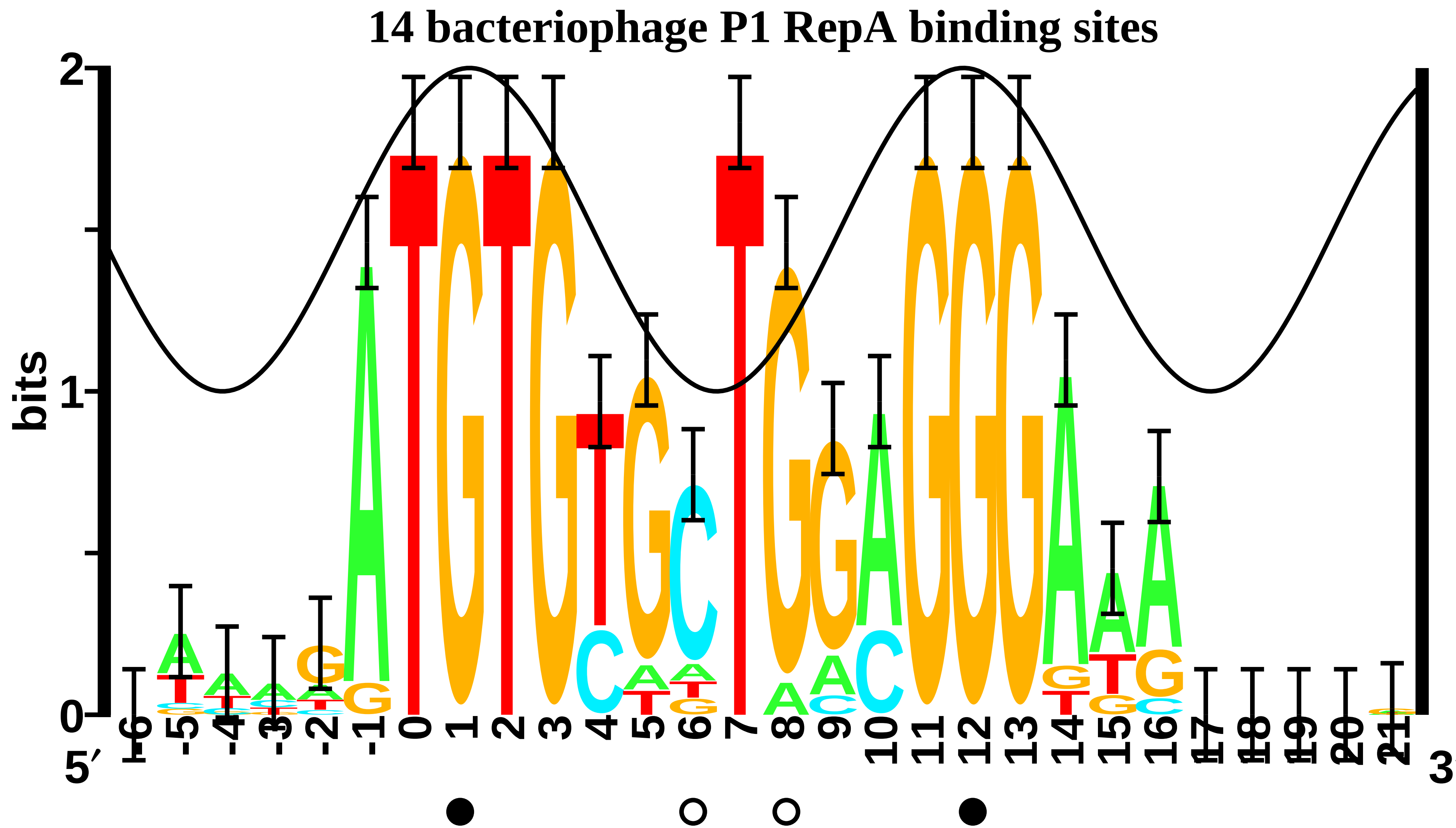
Triangular area is empty ⇒ OR instead of Sum

Bacteriophage P1 RepA Sequence Logo

Stuart Austin and Ann Abeles found binding sites . . . make a logo . . .



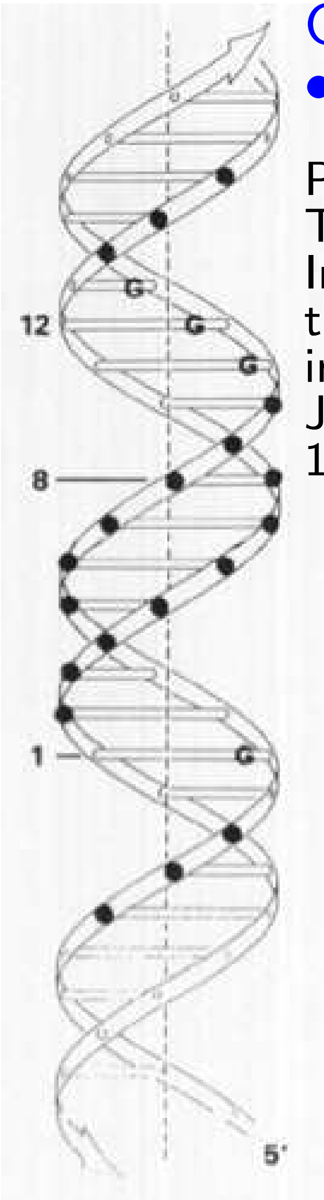
Stuart Austin and Ann Abeles found binding sites . . . make a logo . . .



Violation of the 1 bit Minor Groove rule!

Abeles, Reaves & Austin, J. Bacteriol. 1989 171:43-52

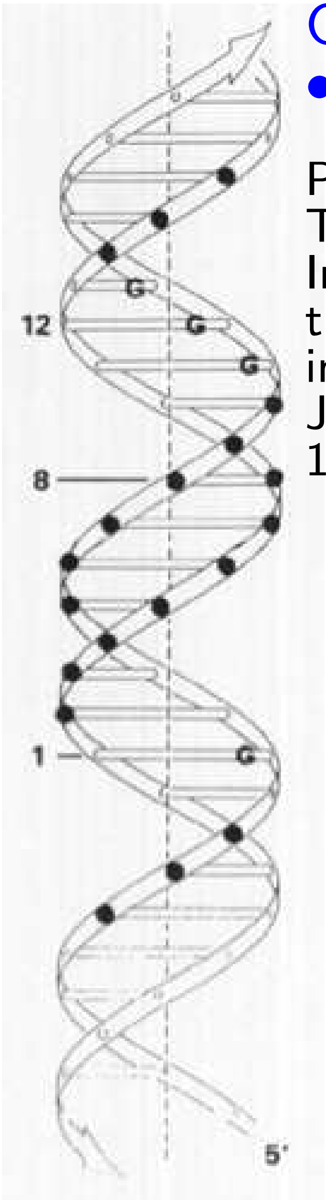
RepA orientation data



G: Methylation interference
•: hydroxyl radical footprint

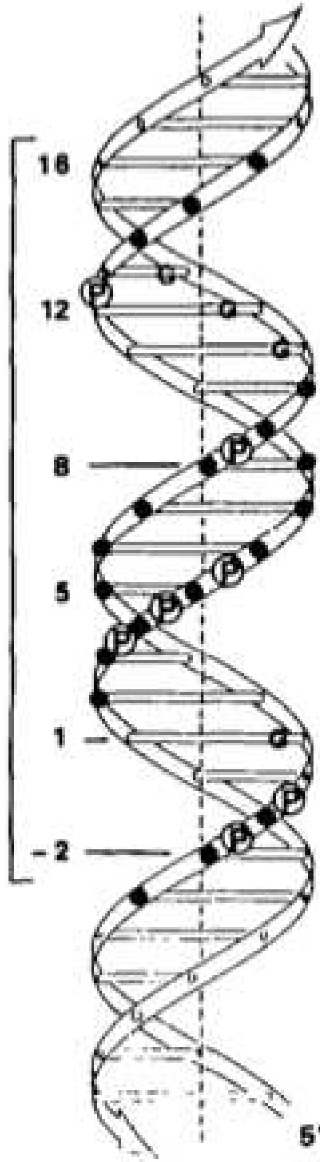
P. P. Papp, D. K. Chattoraj and T. D. Schneider,
Information analysis of sequences that bind the replication, initiator RepA,
J. Mol. Biol., 233, 219-230, 1993.

RepA orientation data



G: Methylation interference
•: hydroxyl radical footprint

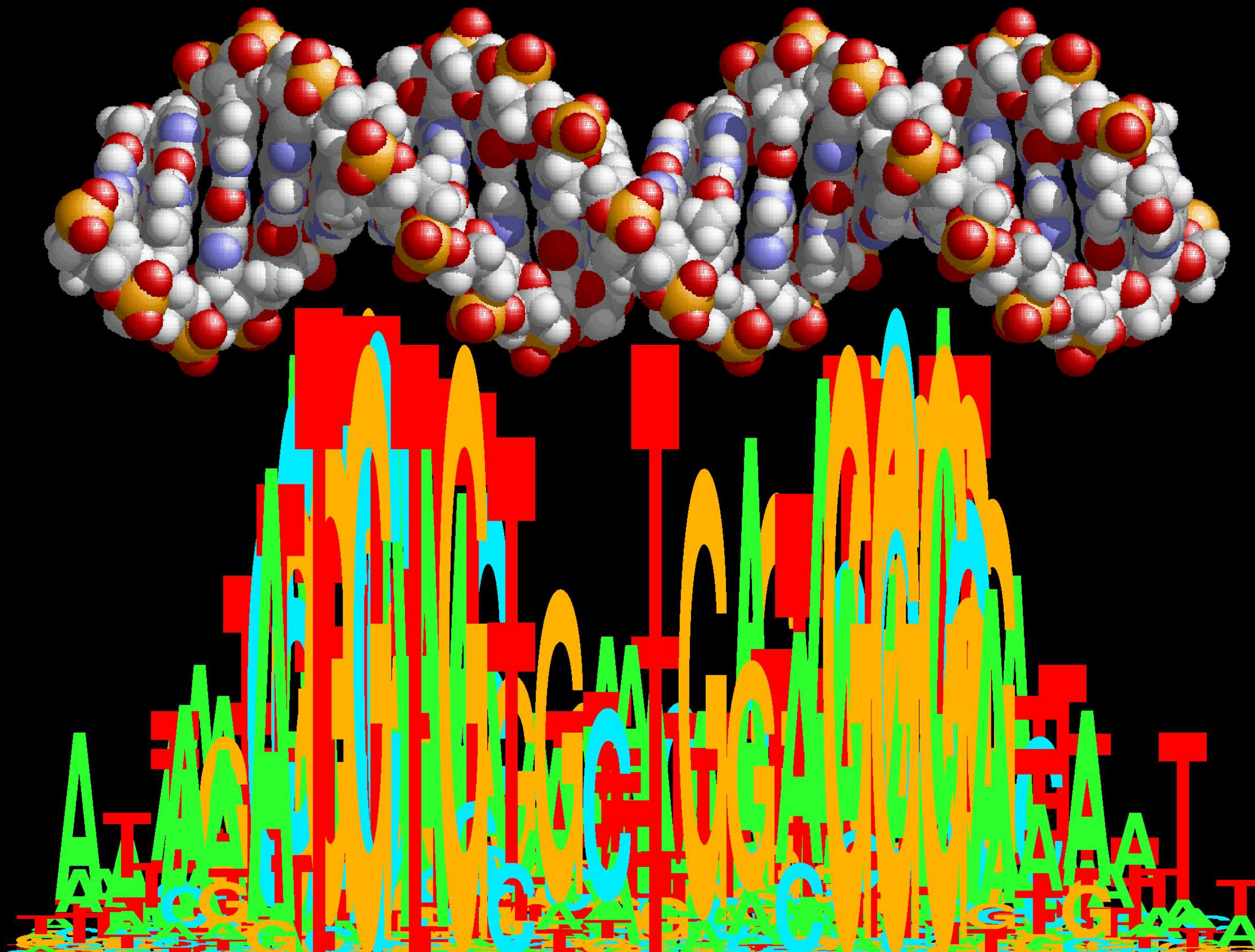
P. P. Papp, D. K. Chattoraj and T. D. Schneider,
Information analysis of sequences that bind the replication, initiator RepA,
J. Mol. Biol., 233, 219-230, 1993.



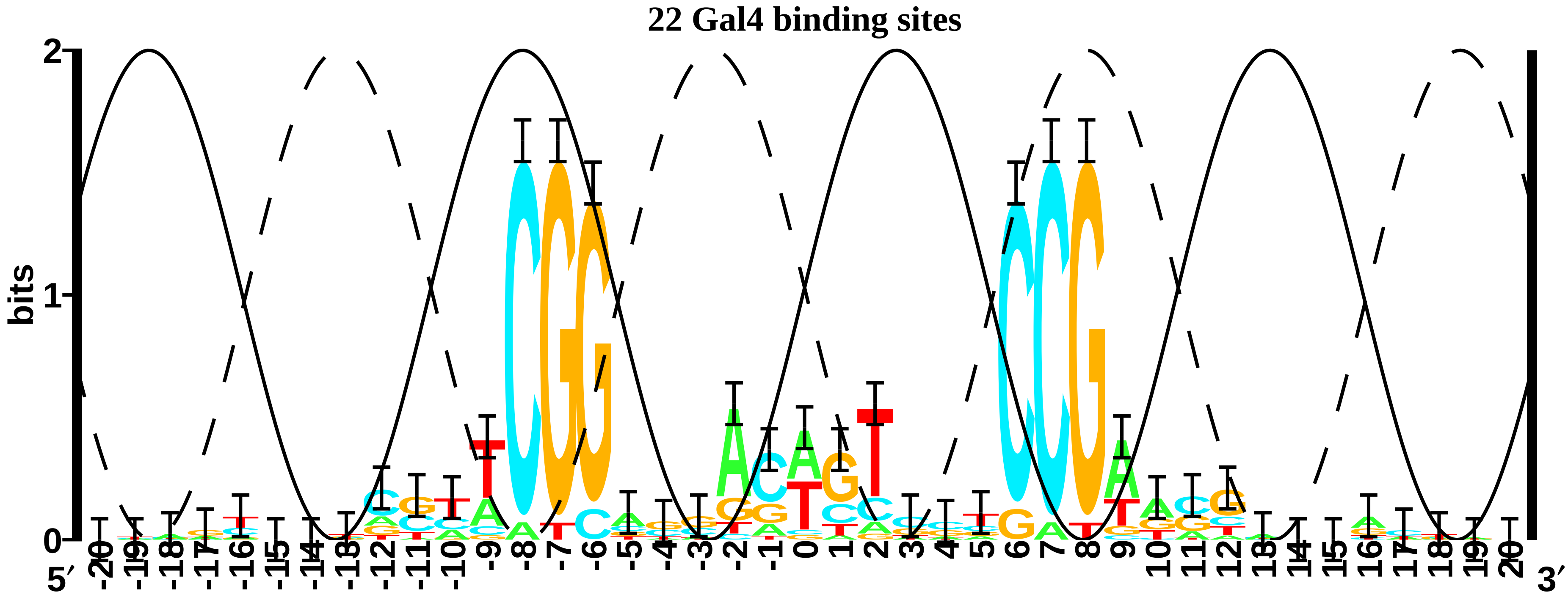
G: Methylation interference
Ⓟ: ethylation interference
•: hydroxyl radical footprint

P. P. Papp and D. K. Chattoraj,
Missing-base and ethylation interference footprinting of P1, plasmid replication initiator,
Nucleic Acids Res., 22, 152-157, 1994.

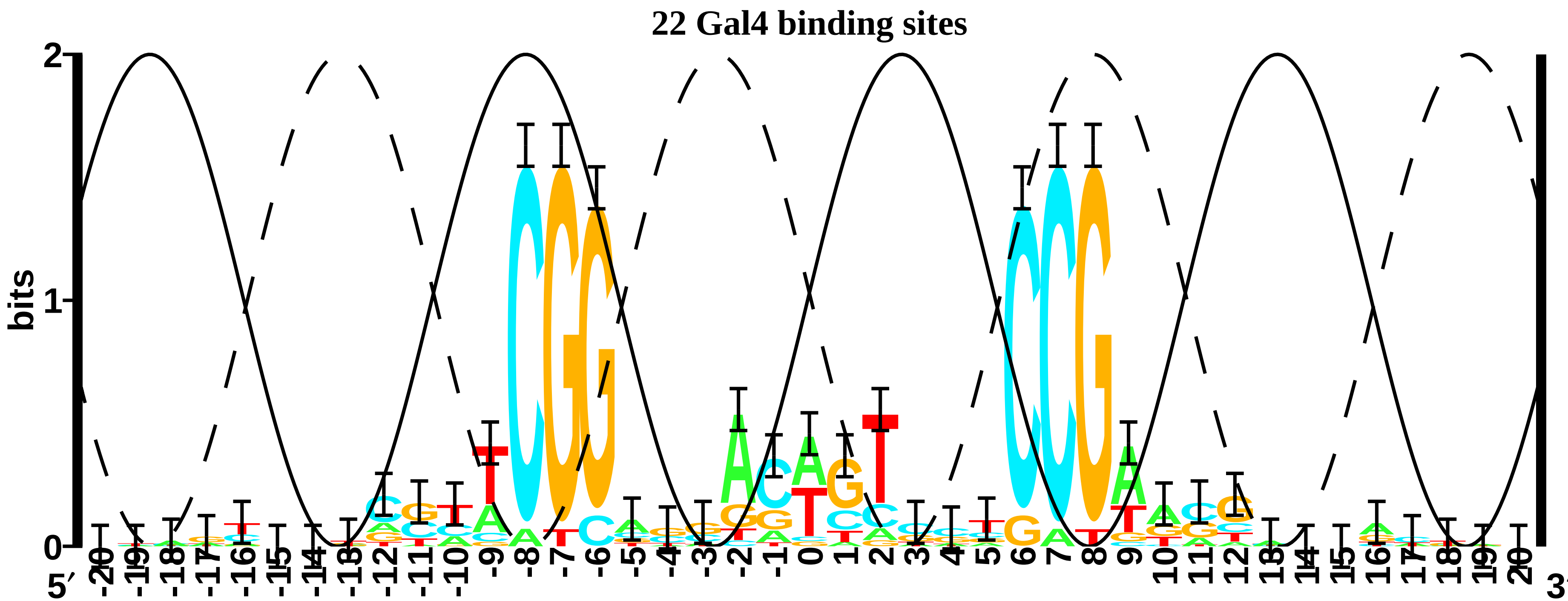
Orientation of proteins on DNA

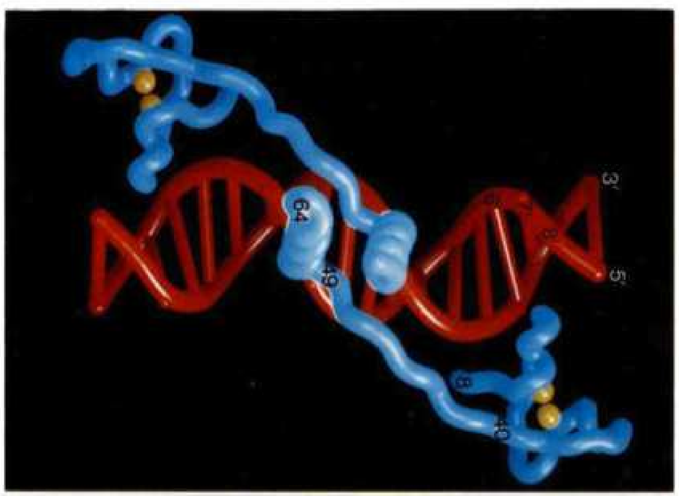


Yeast Saccharomyces cerevisiae GAL4 Sequence Logo

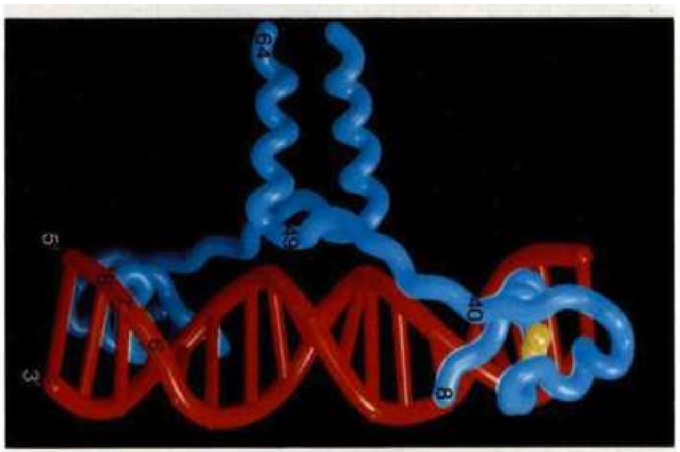


Yeast Saccharomyces cerevisiae GAL4 Sequence Logo





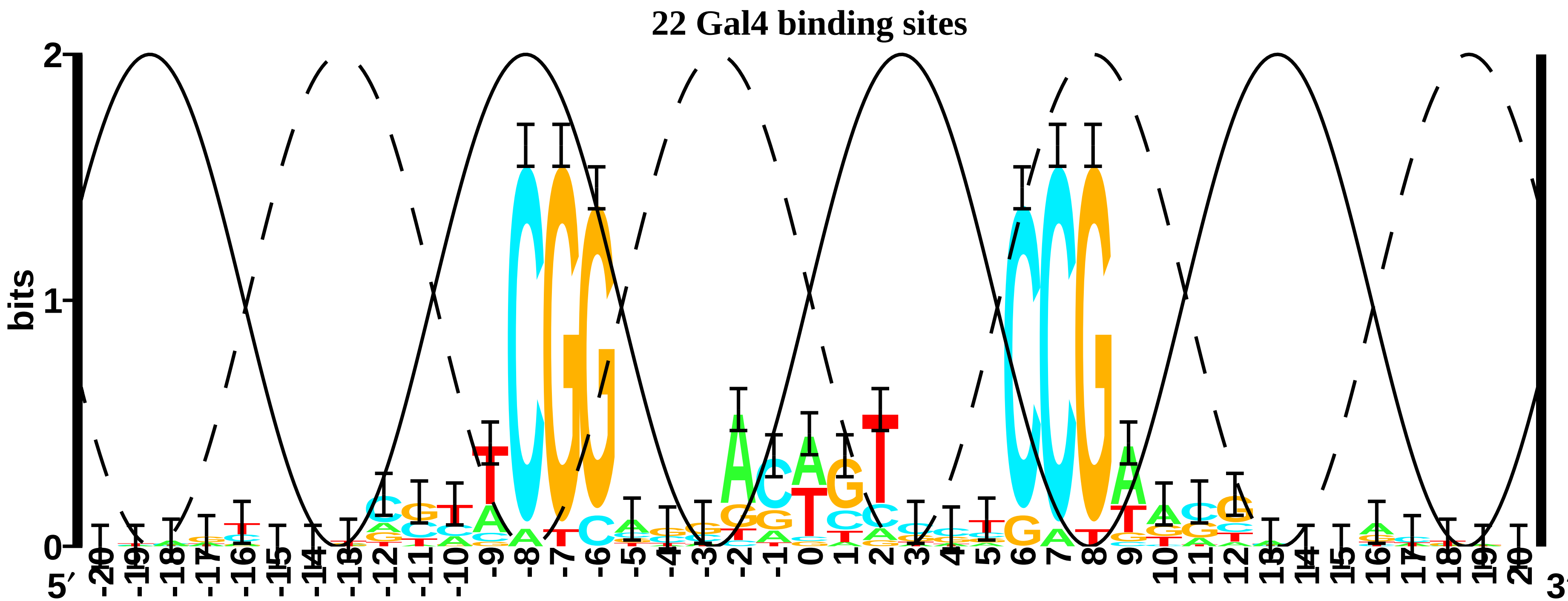
top view

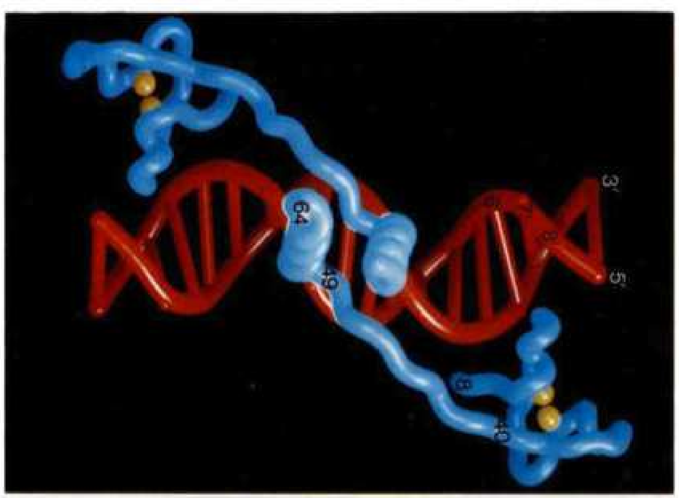


side view

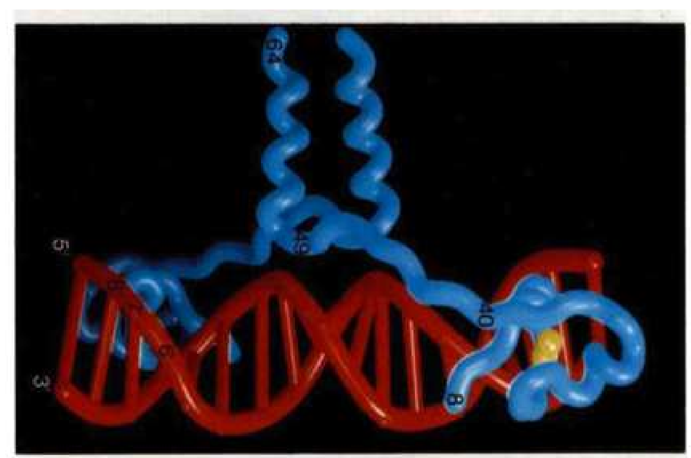
Marmorstein *et al.* 1992, Nature 356:408-414

Yeast Saccharomyces cerevisiae GAL4 Sequence Logo





top view

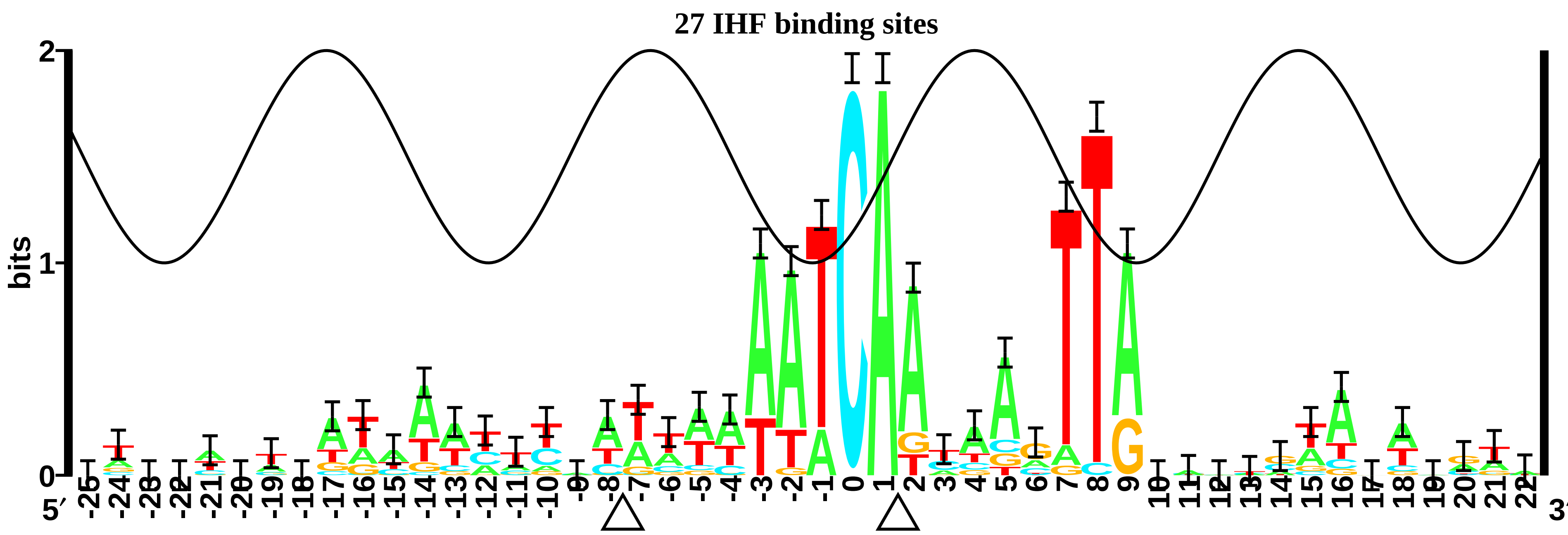


side view

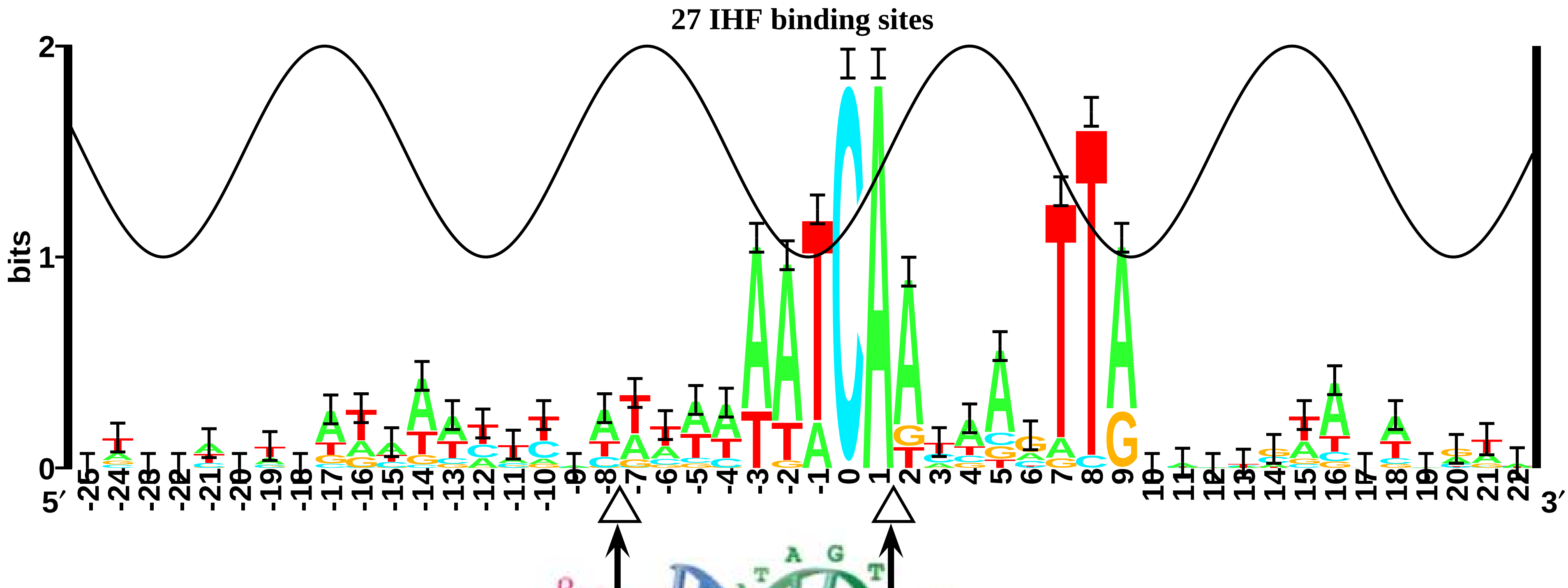
This does not explain the RepA anomaly!

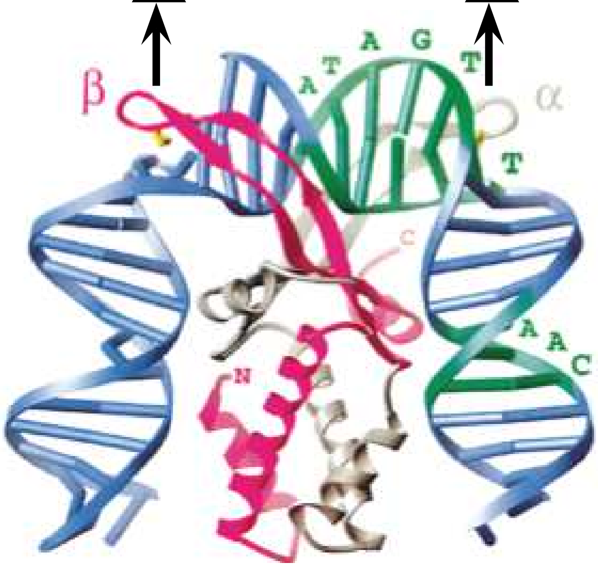
Marmorstein *et al.* 1992, Nature 356:408-414

E. coli IHF Sequence Logo



E. coli IHF Sequence Logo



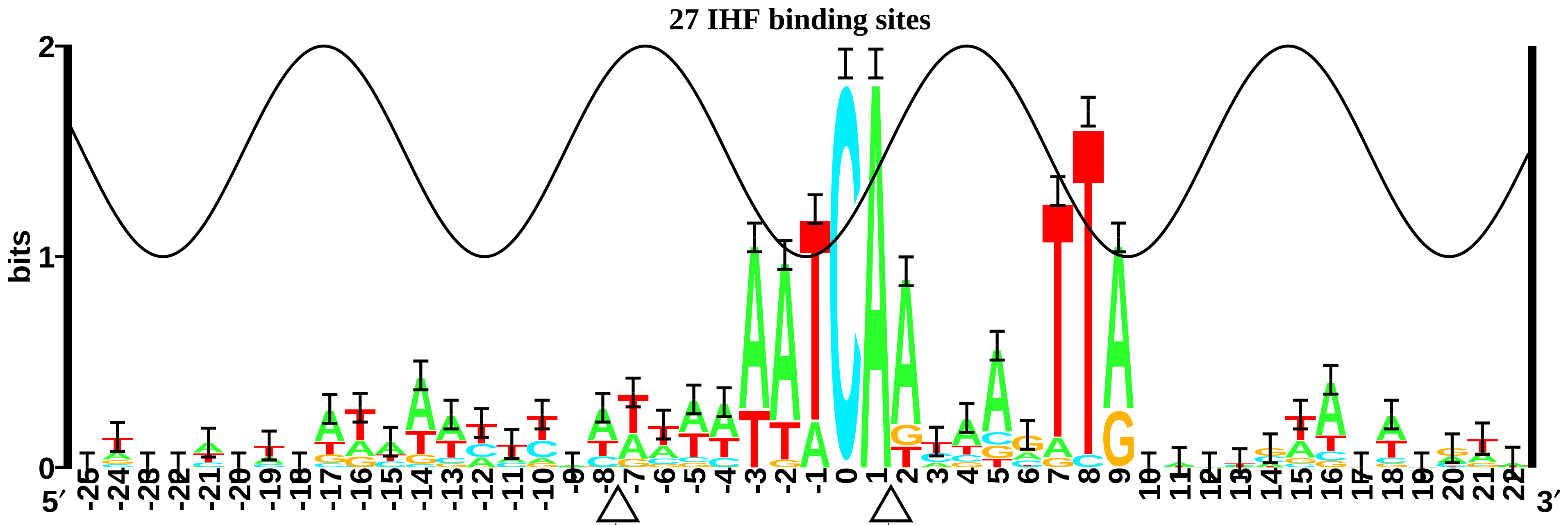


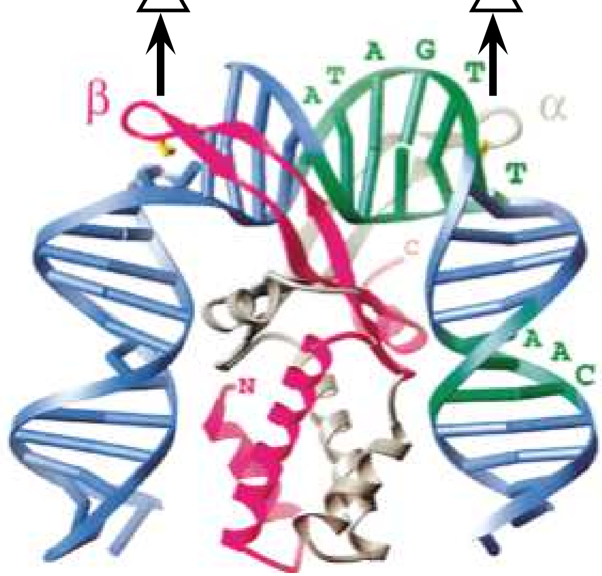
△: **IHF inserts a proline into the DNA to induce a bend**

Minor groove binding - cracked open DNA

Rice *et al.* Cell, 1996, 87:1295-1306

E. coli IHF Sequence Logo



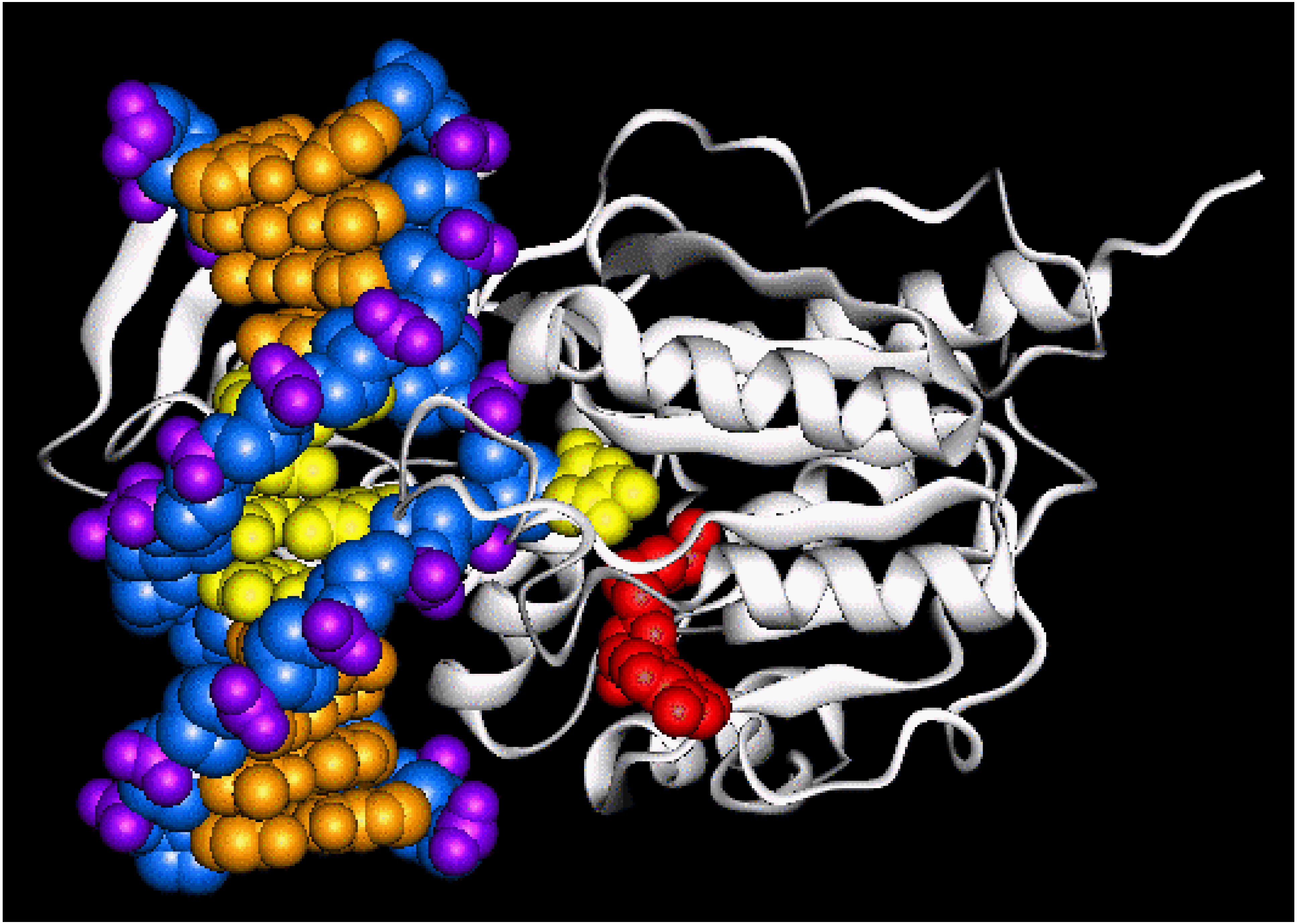


△: IHF inserts a proline into the DNA to induce a bend

Minor groove binding - cracked open DNA

This does not explain the RepA anomaly!

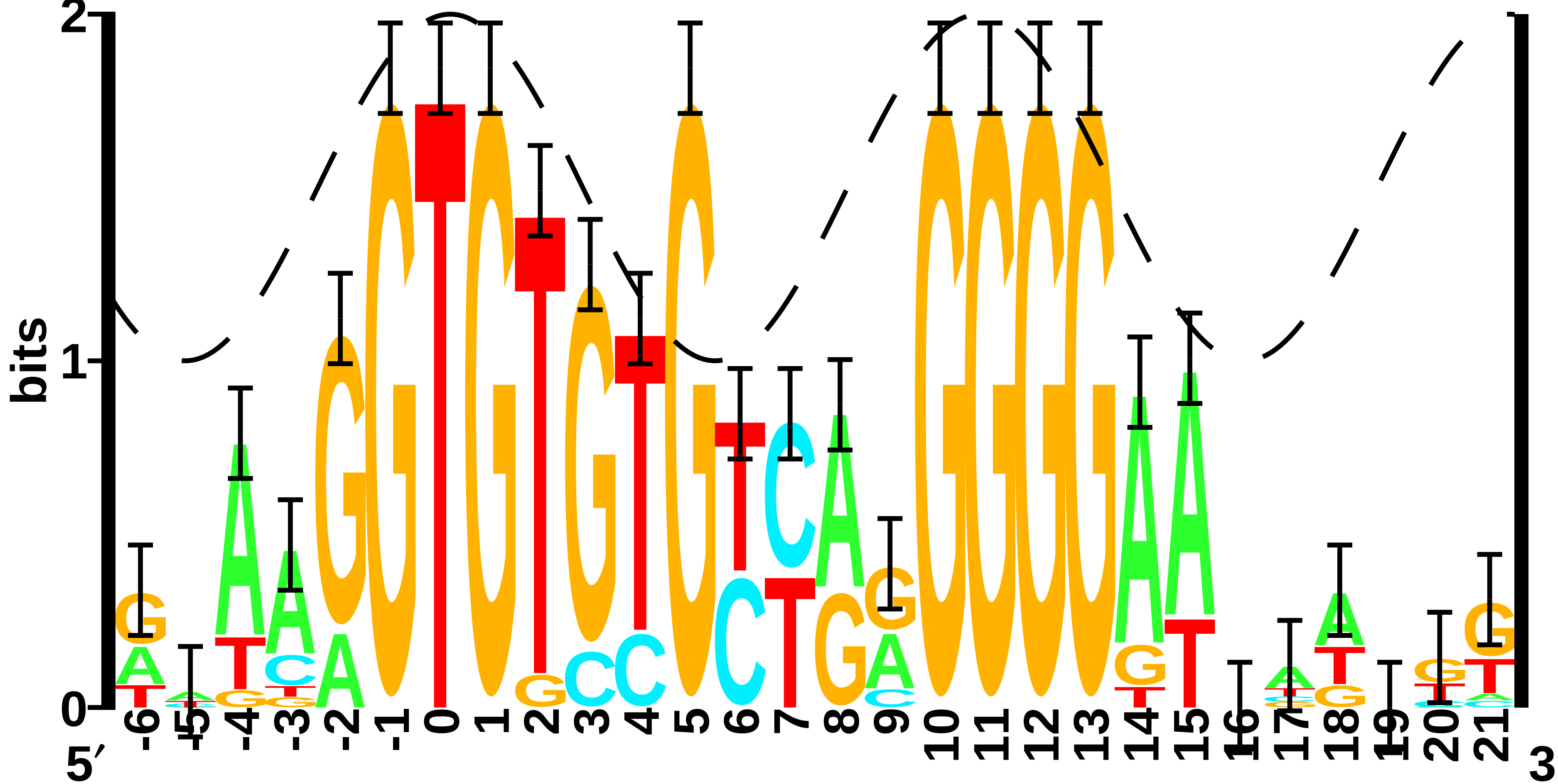
HhaI methyltransferase Base Flipping



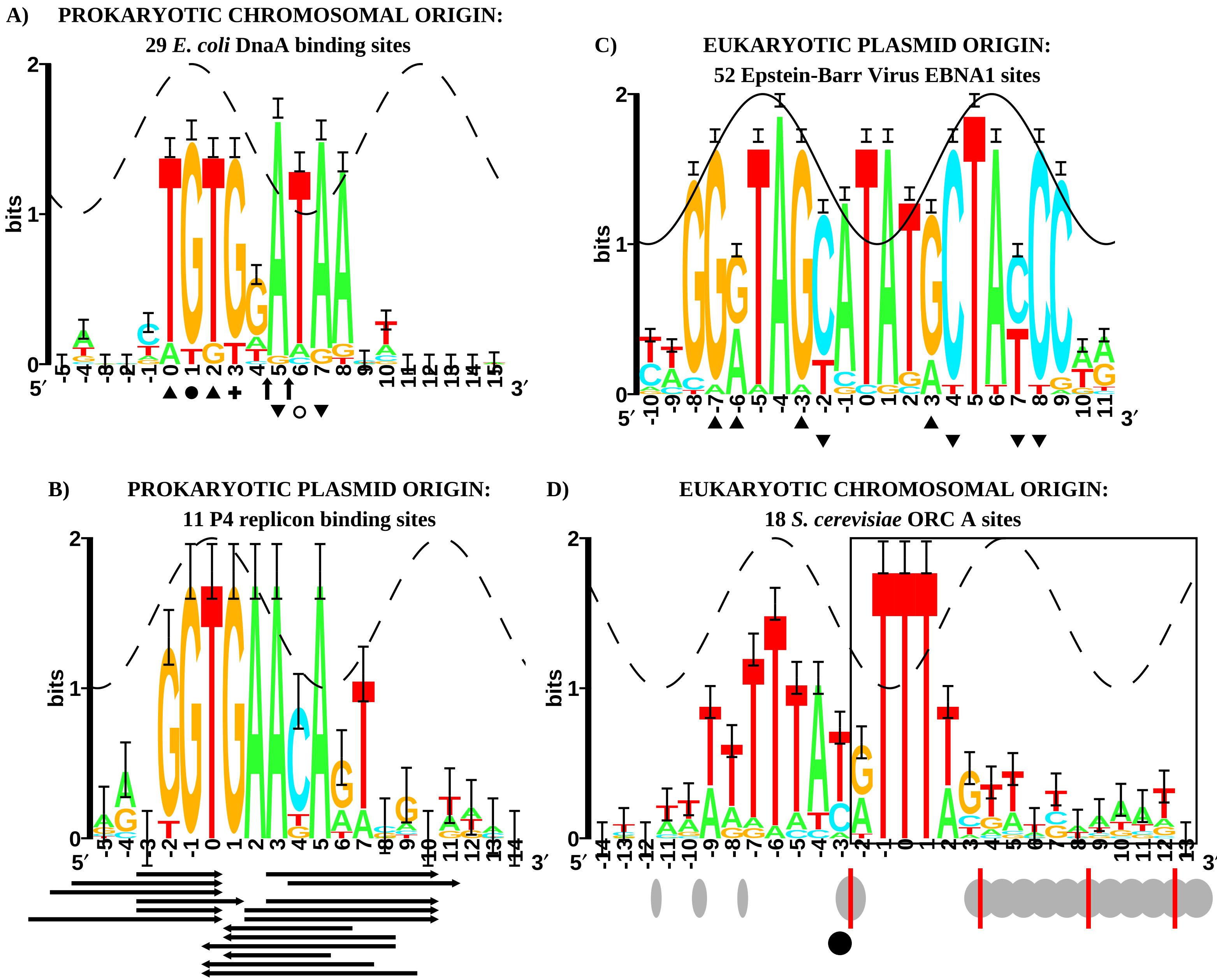
Richard J. Roberts and X. Cheng, Annu. Rev. Biochem. 1998, 67:181-198

PROKARYOTIC PLASMID ORIGIN:

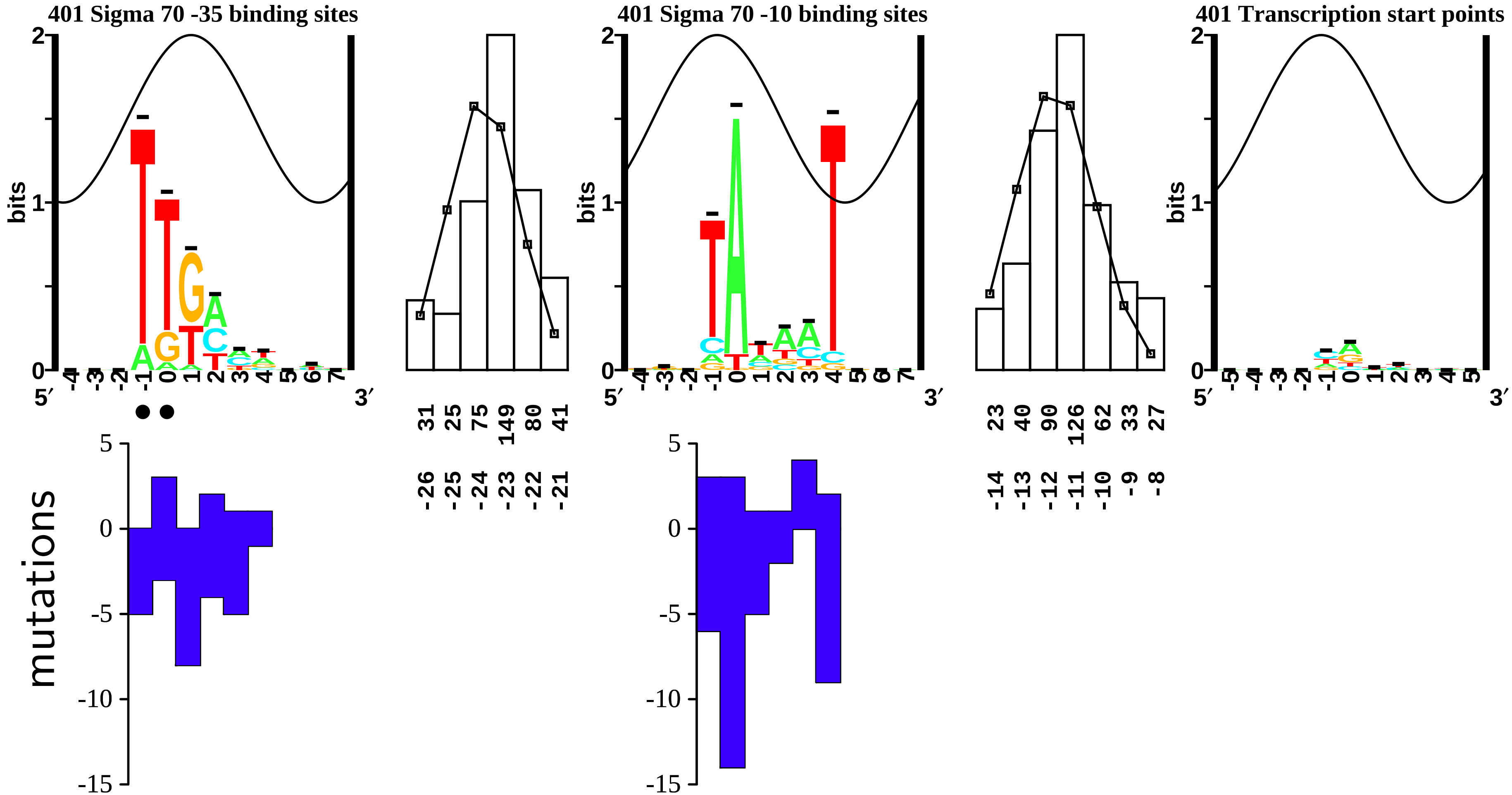
15 Rts1 RepA binding sites



DNA Replication Protein Sequence Logos

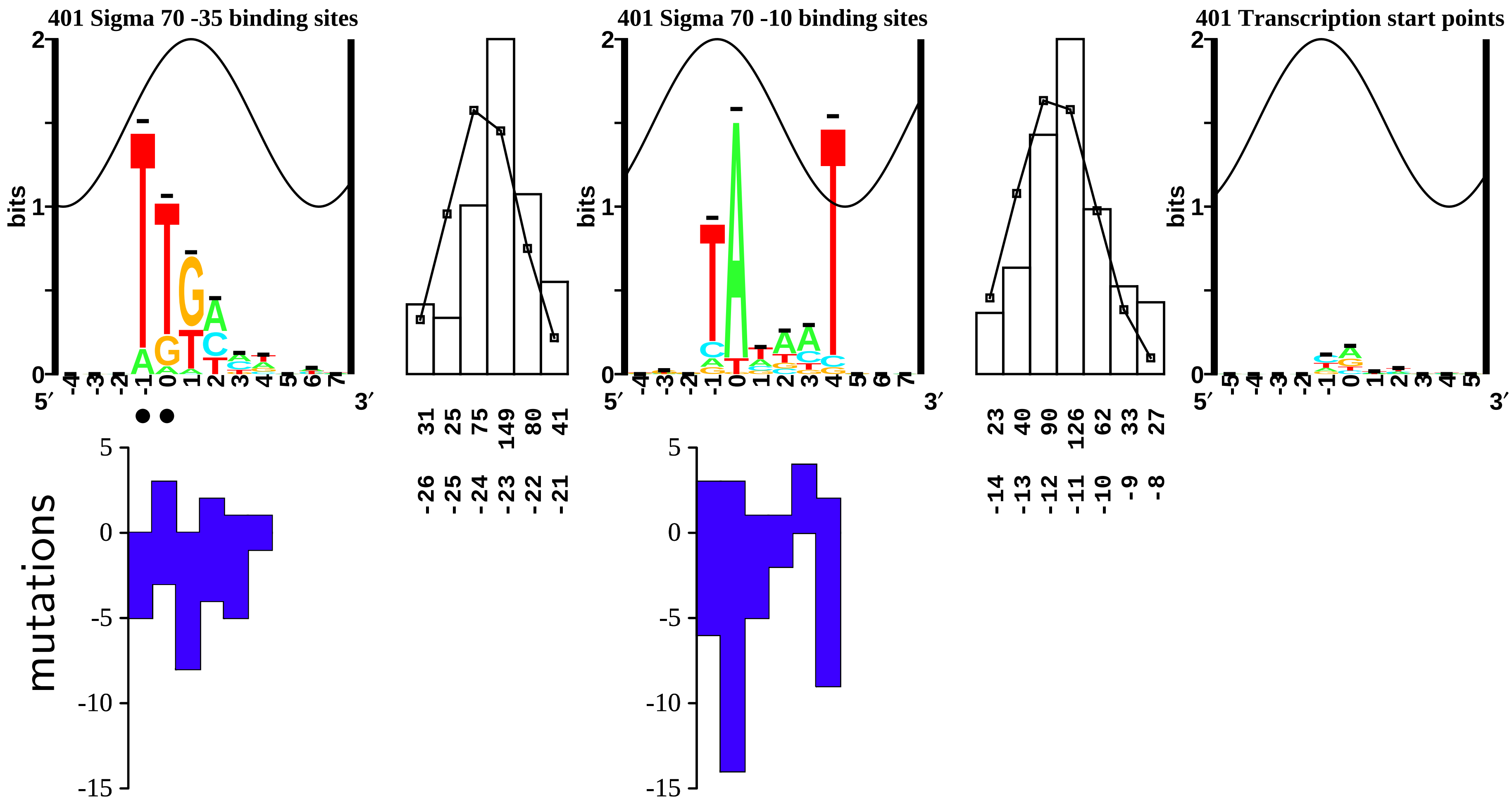


Sequence logos for σ^{70} RNA transcription



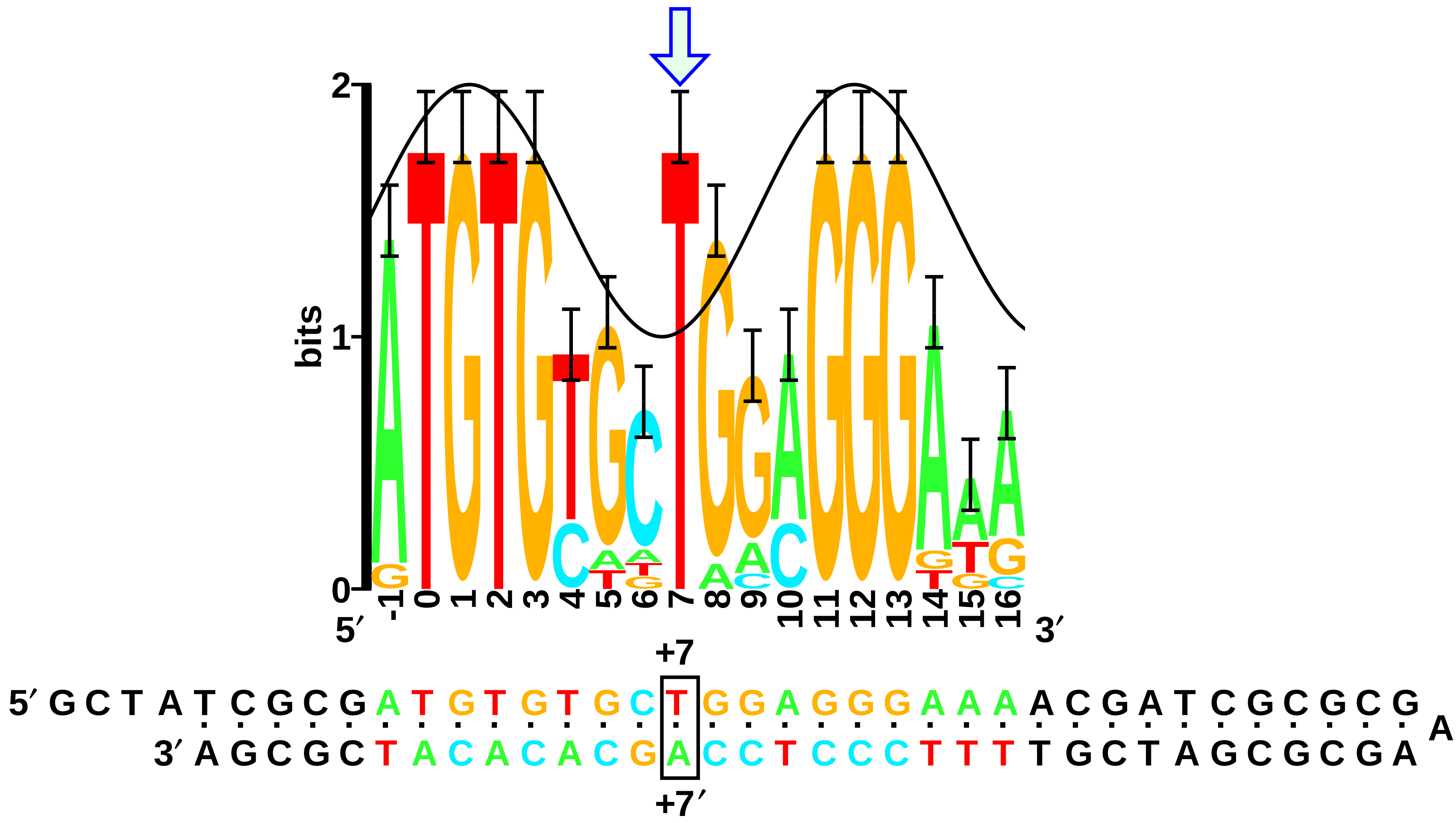
Shultzaberger *et al.* Nucleic Acids Res., 2007, 35:771-788

Sequence logos for σ^{70} RNA transcription



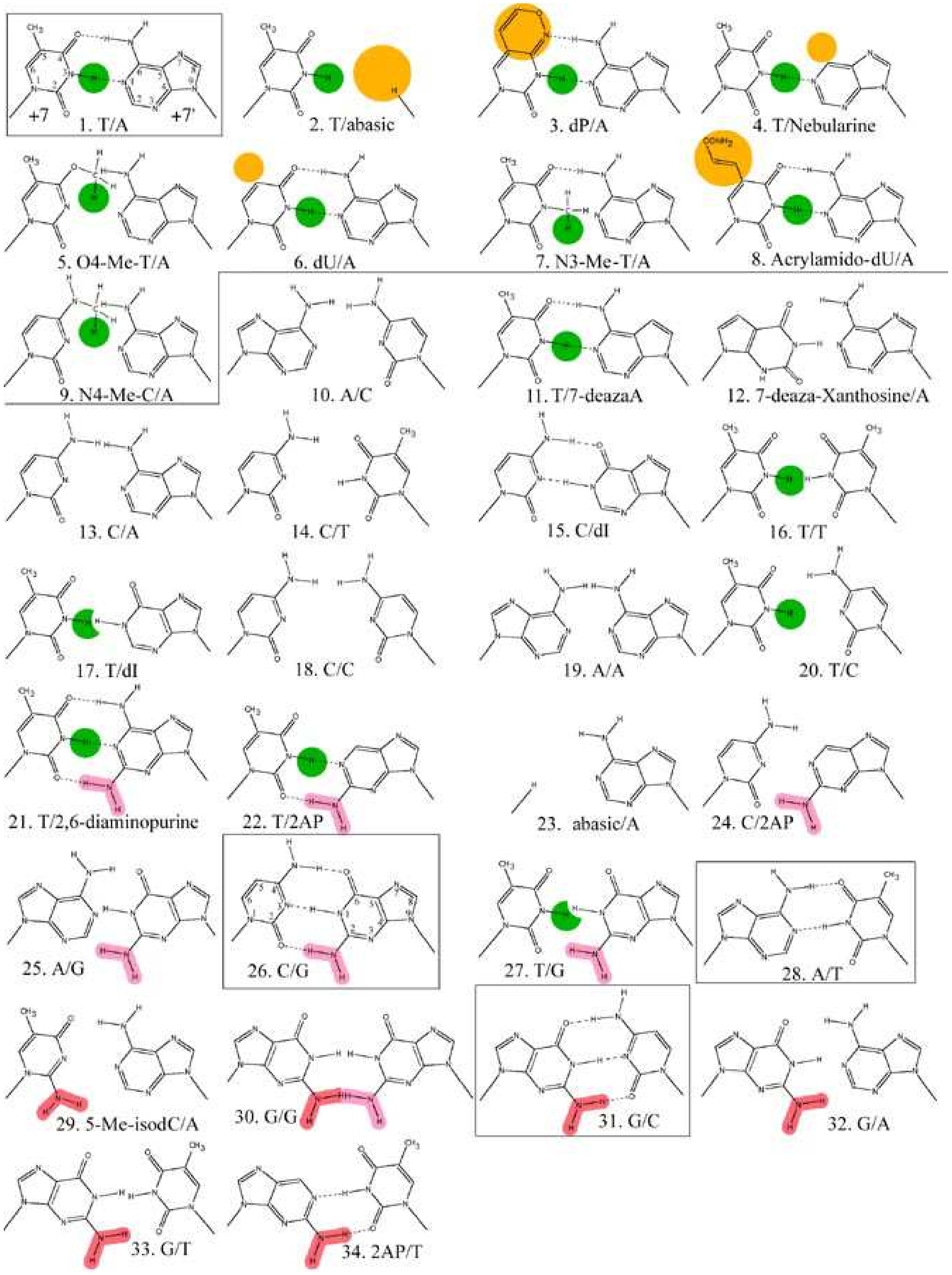
Our prediction of base 4 flipping was confirmed!
Feklistov & Darst, Cell, 2011, 147: 1257-1269

P1 RepA sequence logo and hairpin



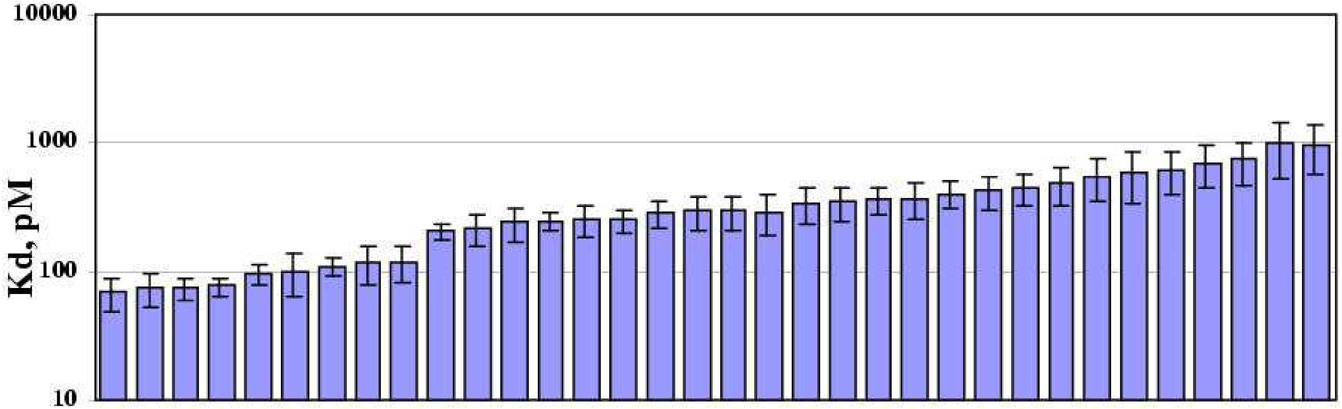
Experiment: change base ± 7 on both strands.

Variant nucleotides at ± 7 of the P1 RepA binding site

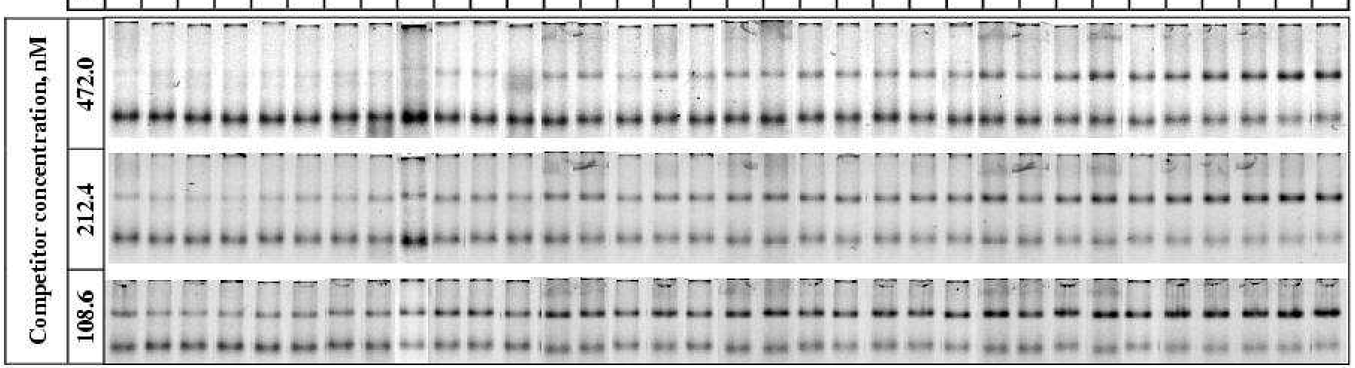


The four natural base pairs are boxed.

Gel Mobility Shift Assay of ±7 variant P1 RepA sites

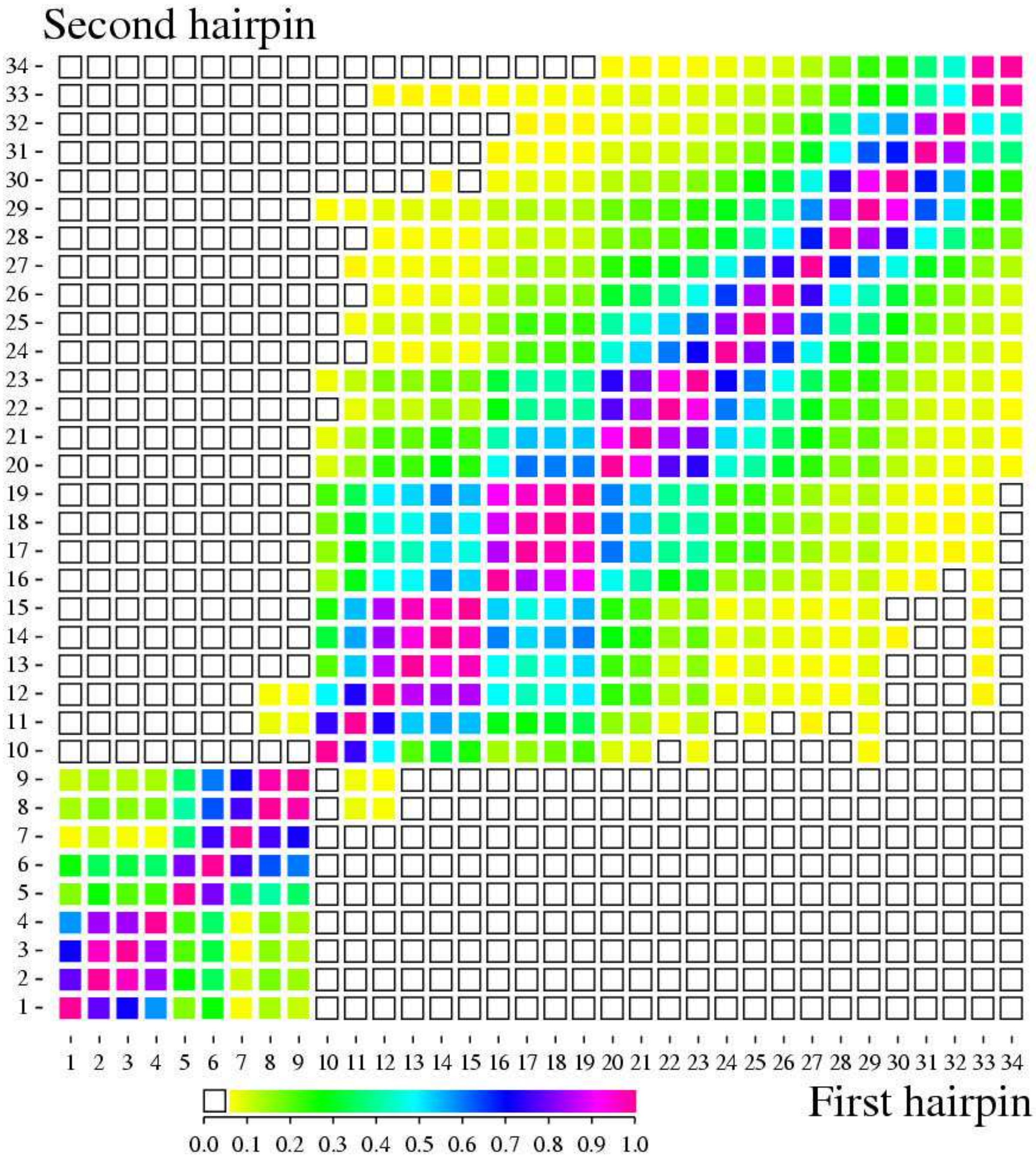


Name	Kd +/- SD, pM	
1. T/A	69	20
2. T/abasic	74	22
3. dP/A	75	15
4. T/Nebularine	77	12
5. O4-Me-T/A	95	17
6. dU/A	101	38
7. N3-MeT/A	109	18
8. Acrylamido-dU/A	118	40
9. N4-Me-C/A	118	36
10 A/C	205	26
11. T/7-deazaA	218	61
12. 7-deaza-Xanthosine/A	239	72
13. C/A	248	41
14. C/T	251	66
15. C/dI	249	50
16. T/T	283	68
17. T/dI	295	84
18. C/C	293	90
19. A/A	291	98
20. T/C	338	100
21. T/2,6-diaminopurine	346	102
22. T/2AP	361	85
23. abasic/A	369	112
24. C/2AP	403	93
25. A/G	425	131
26. C/G	447	128
27. T/G	488	160
28. A/T	553	199
29. 5-Me-isodC/A	591	254
30. G/G	614	221
31. G/C	701	262
32. G/A	743	271
33. G/T	986	466
34. 2AP/T	977	402



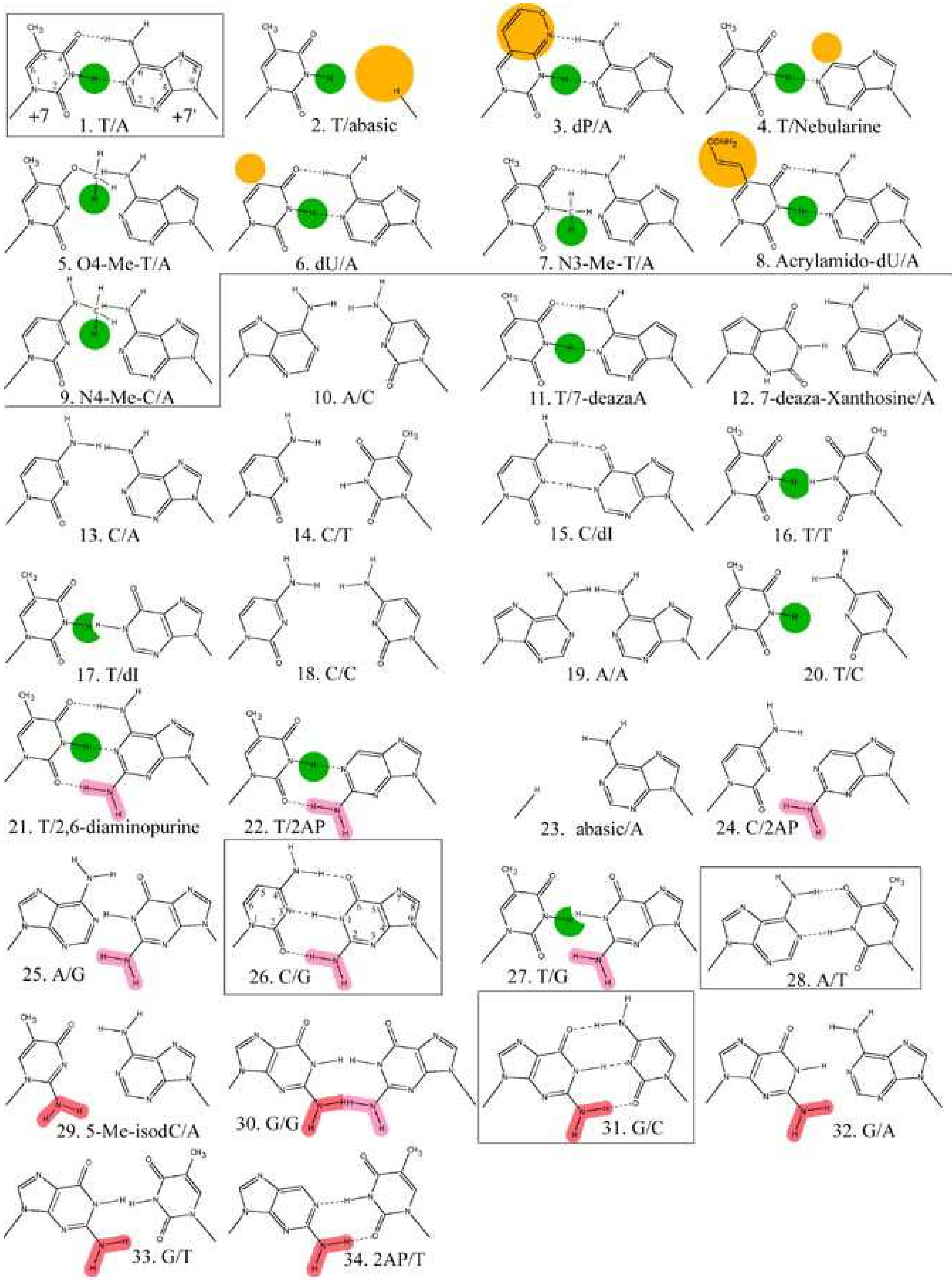
A break appears in the ordered moities

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



The break separates the moities into two classes

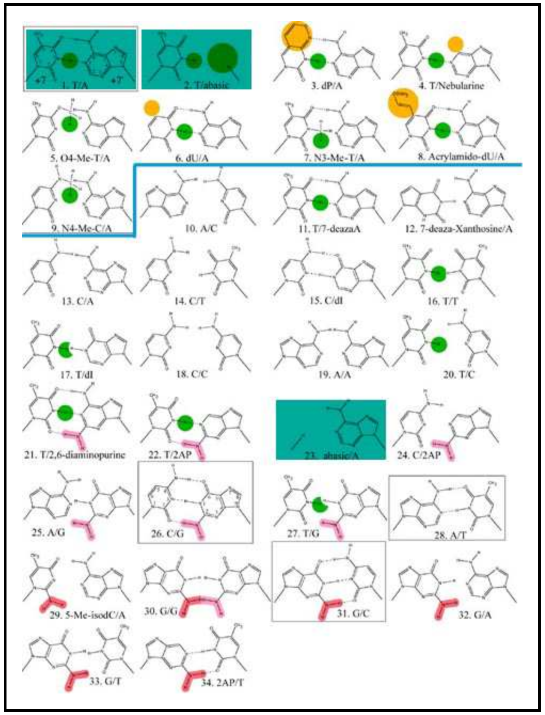
Variant nucleotides at ±7 of the P1 RepA binding site

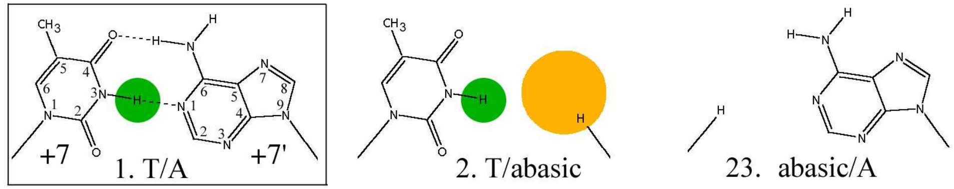


class separator line

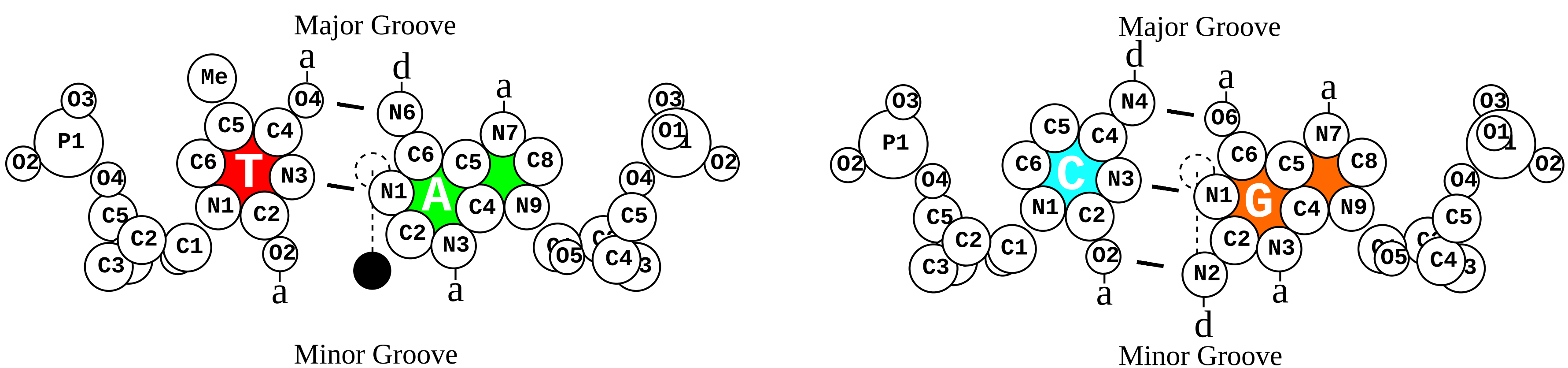
Lyakhov, Hengen, Rubens & Schneider, Nucleic Acids Res., 2001, 29:4892-4900

A Thymine at position +7 is important

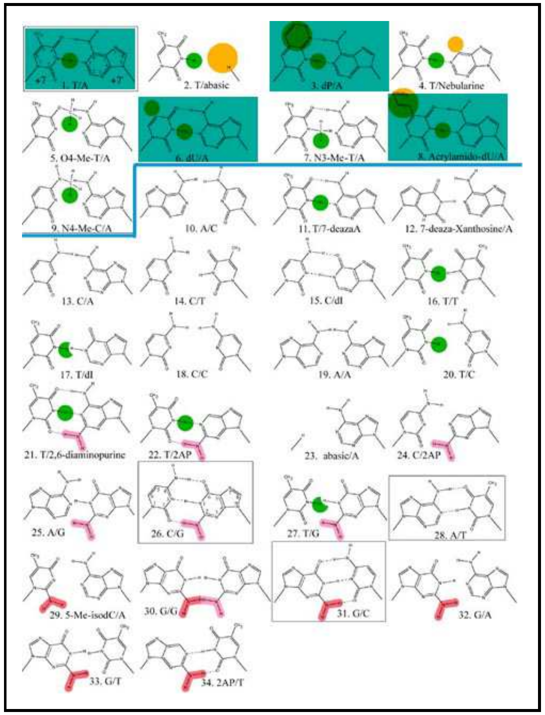


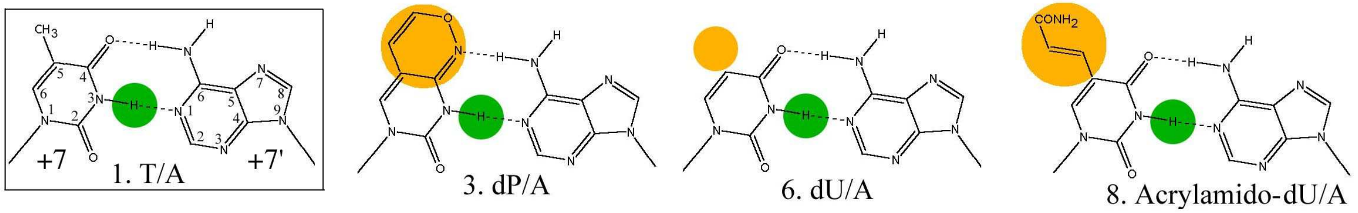


RepA interacts with the base at position +7 for two reasons: removing Thymine +7 leads to a decrease in the binding affinity, while removing Adenine +7$'$ does not.

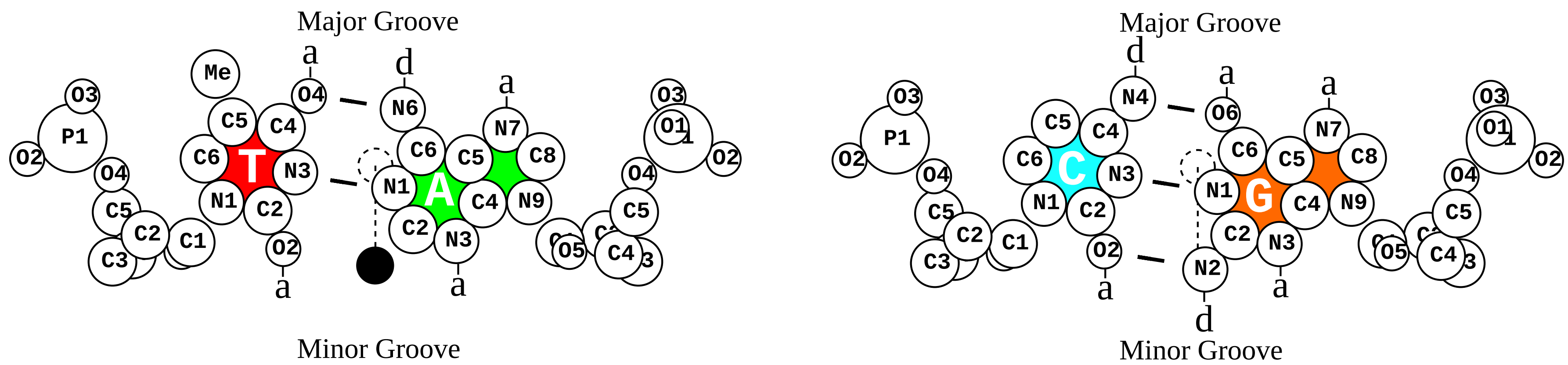


Lyakhov, Hengen, Rubens & Schneider, Nucleic Acids Res., 2001, 29:4892-4900

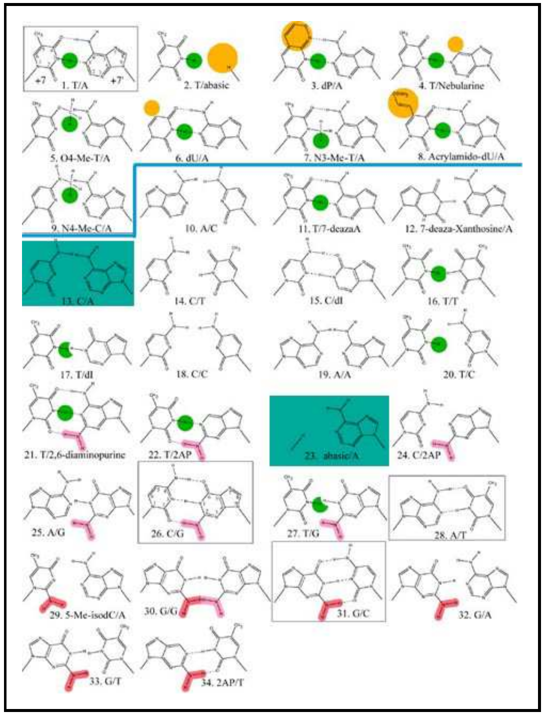


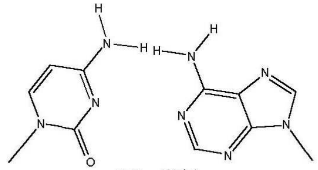


The C5 methyl group does not interact with RepA because removing or changing it only has a slight effect on the binding affinity.

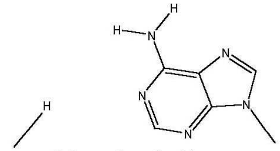


O2 is not important



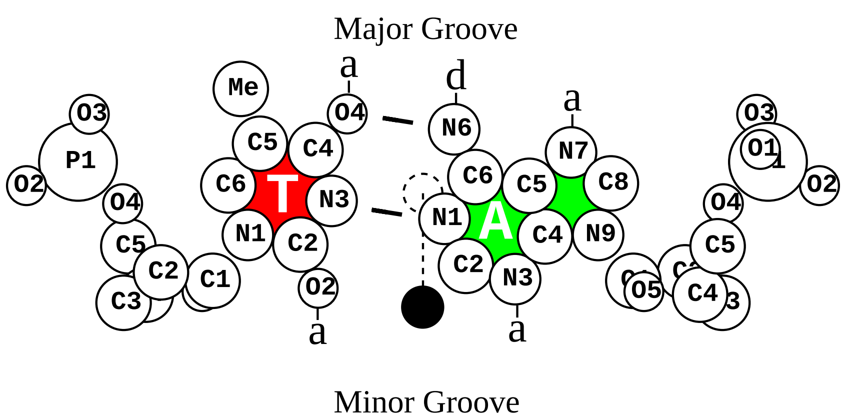


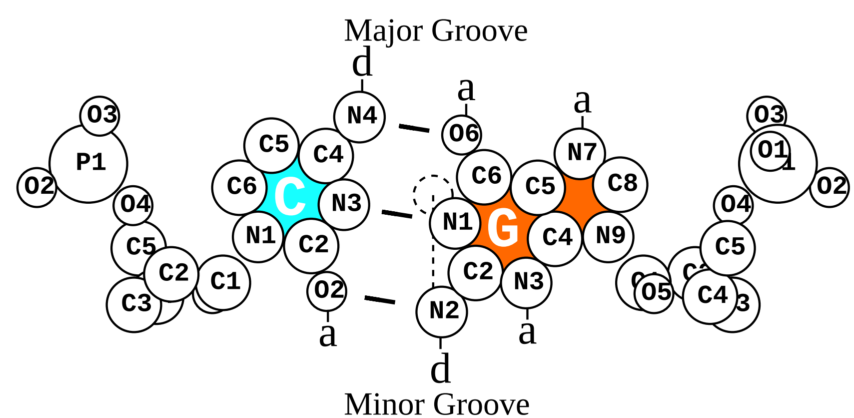
13. C/A



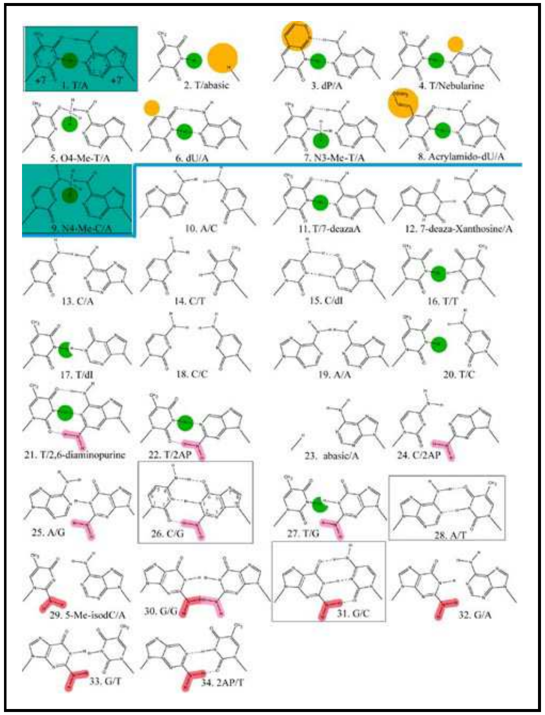
23. abasic/A

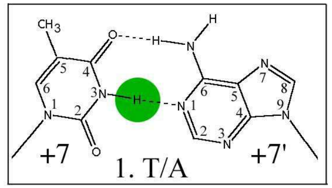
The O2 group is not important for RepA binding because 13.C/A (which has one) and 23.abasic/A (which does not) have similar binding energy.

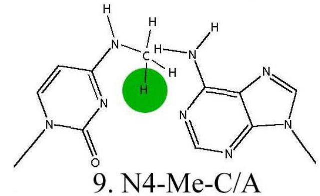




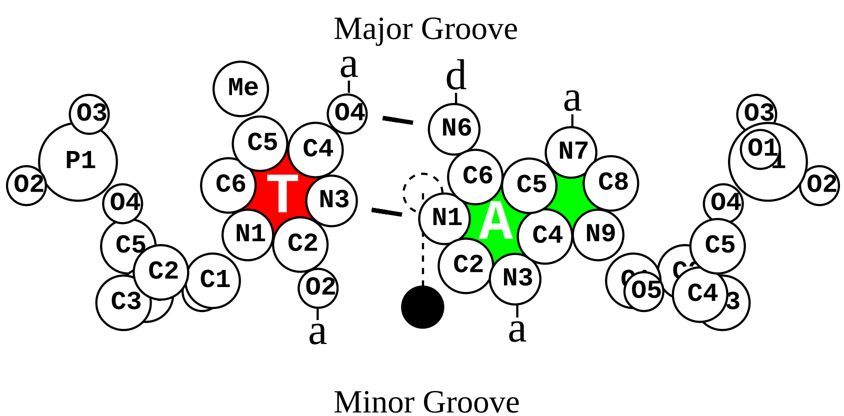
O4 is not important

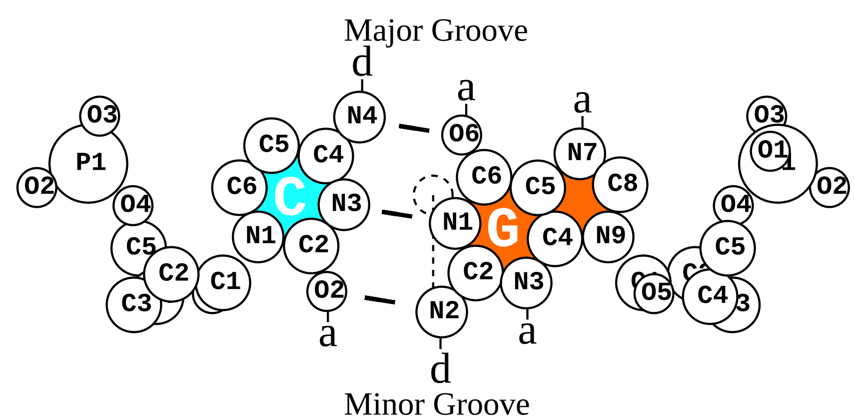




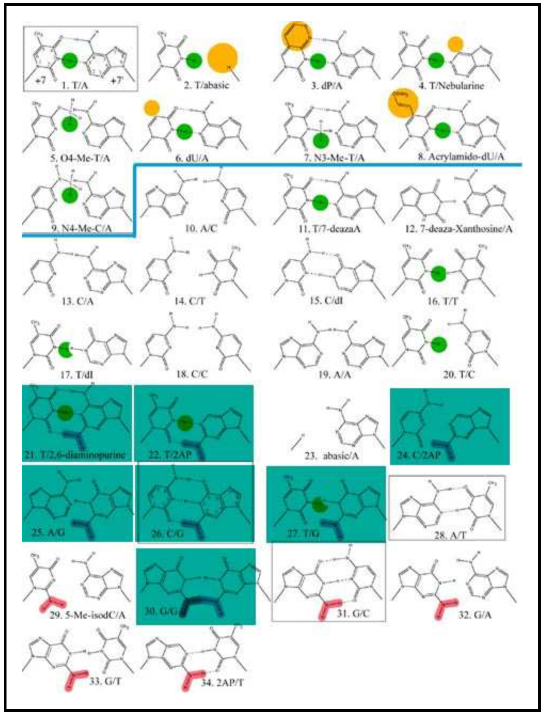


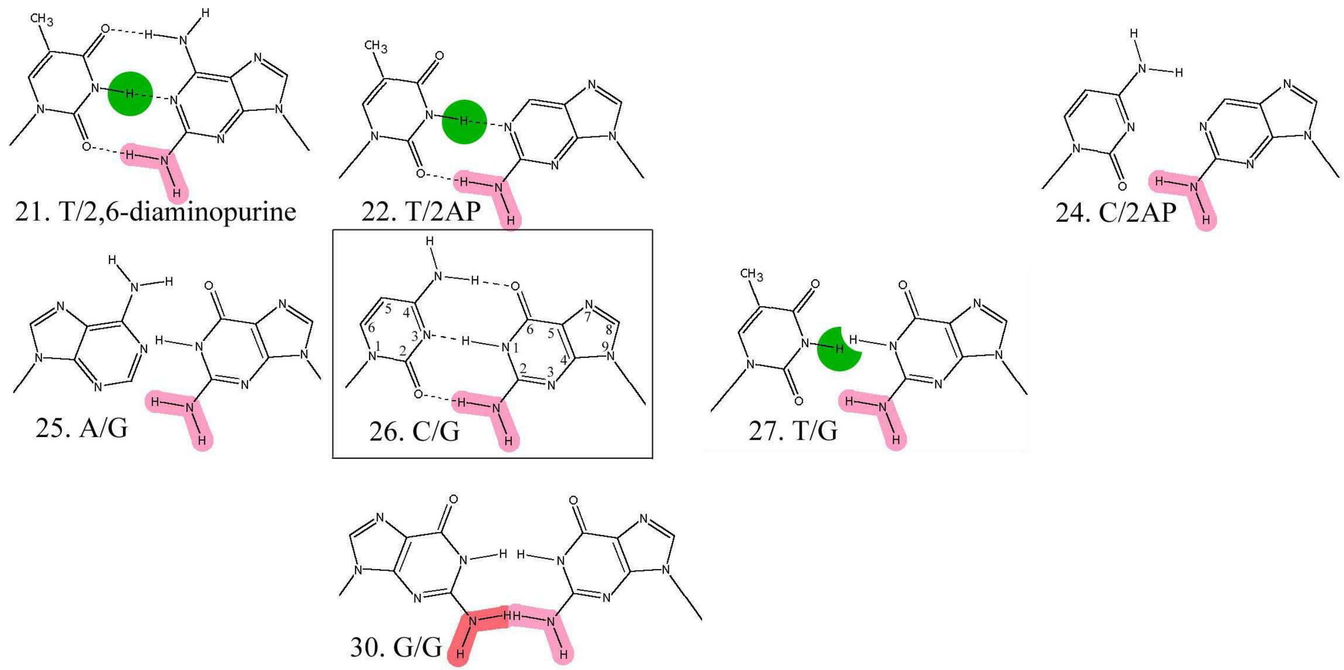
The O4 group is not important because 9.N4-Me-C/A does not have an O4 group as does 1.T/A, but the $\Delta\Delta G$ is low.



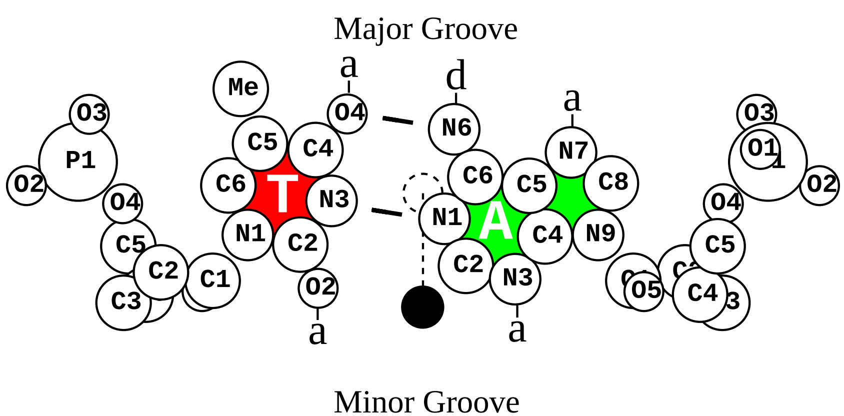


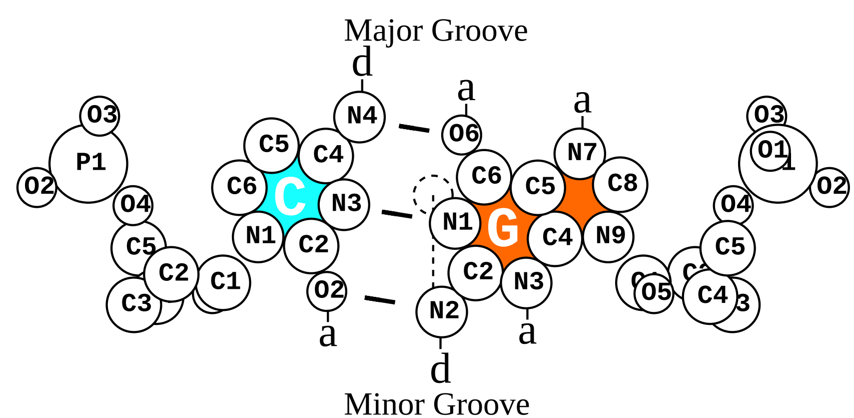
$+7'$ amino group in minor groove is bad



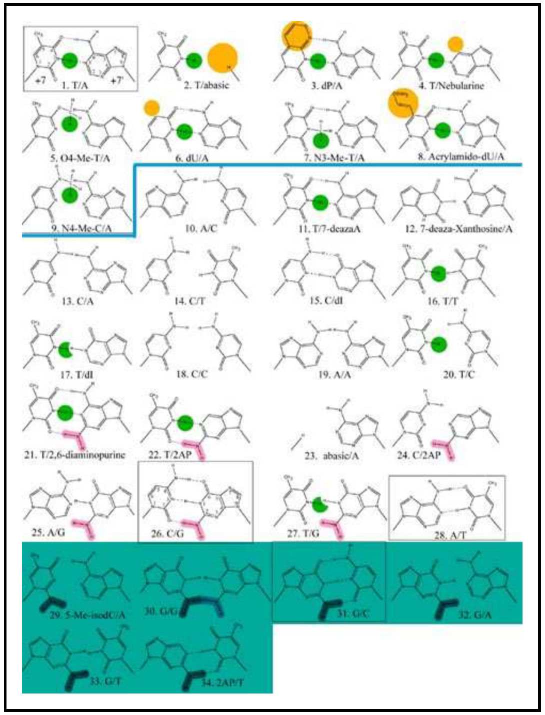


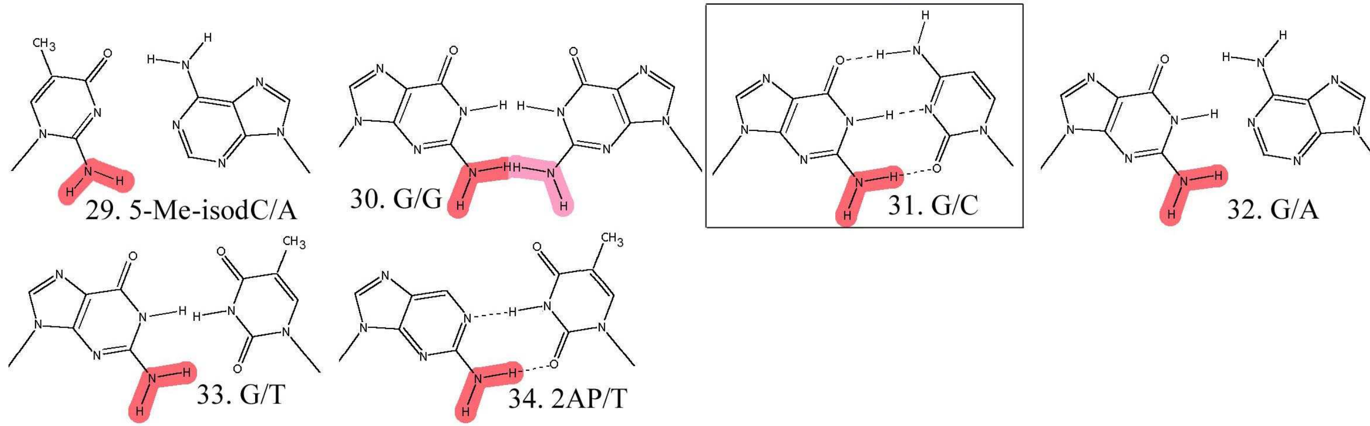
A $+7'$ amino group (purple) inhibits RepA binding.



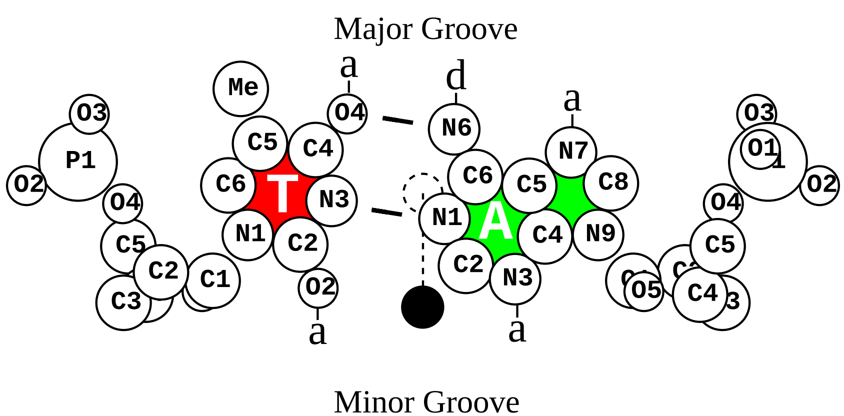


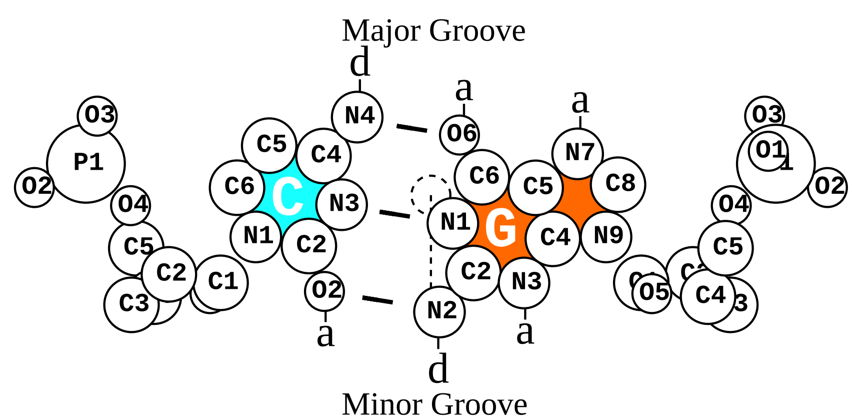
Lyakhov, Hengen, Rubens & Schneider, Nucleic Acids Res., 2001, 29:4892-4900



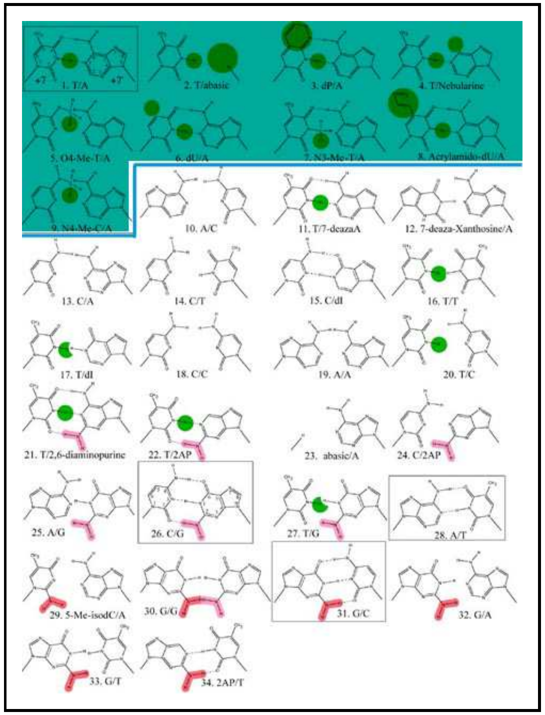


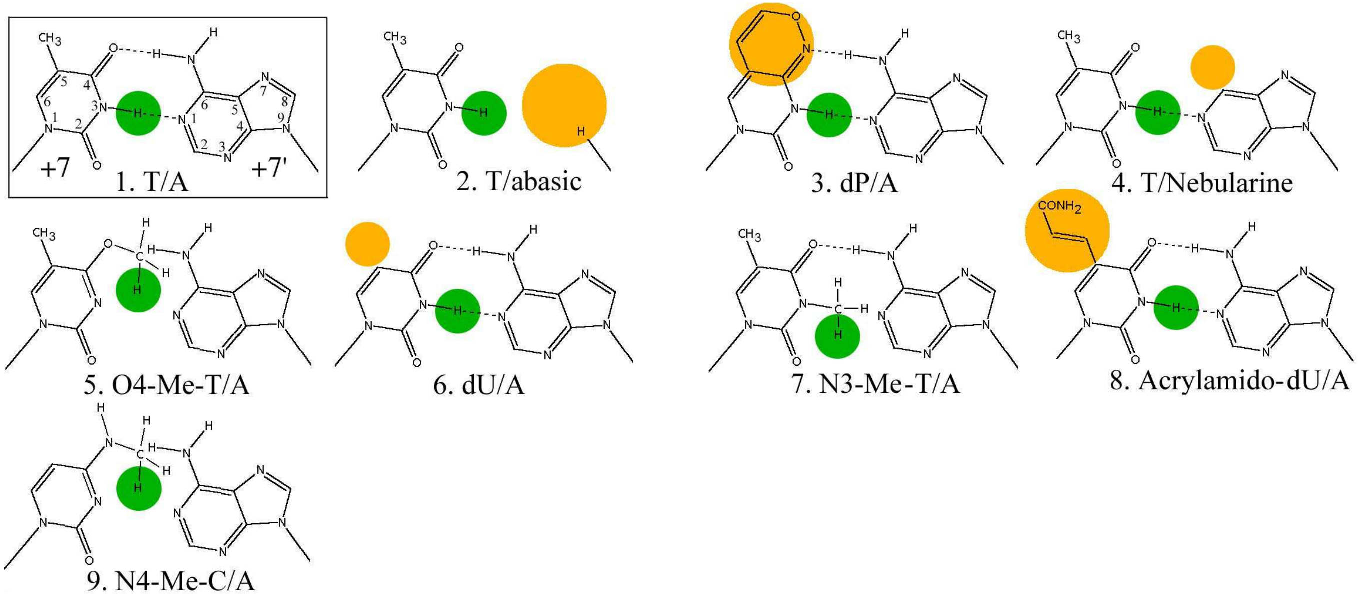
A +7 amino group (red) inhibits RepA binding more strongly than a +7′ amino group.



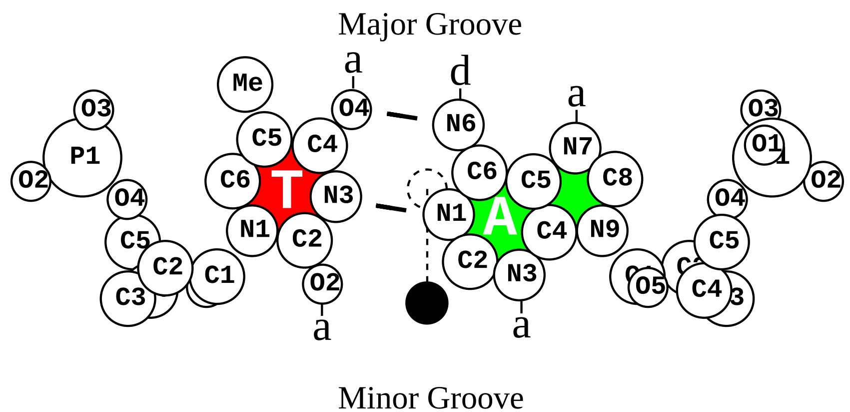


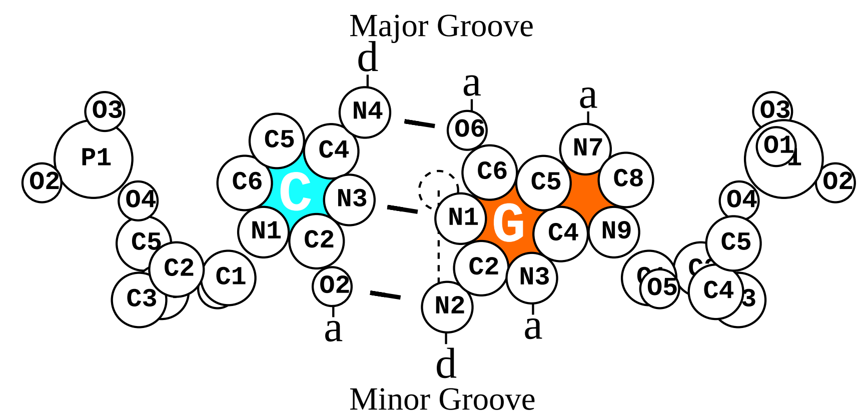
A proton near N3 interacts with RepA

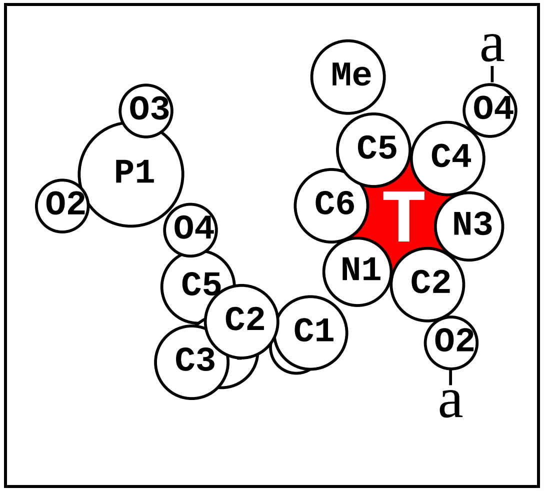


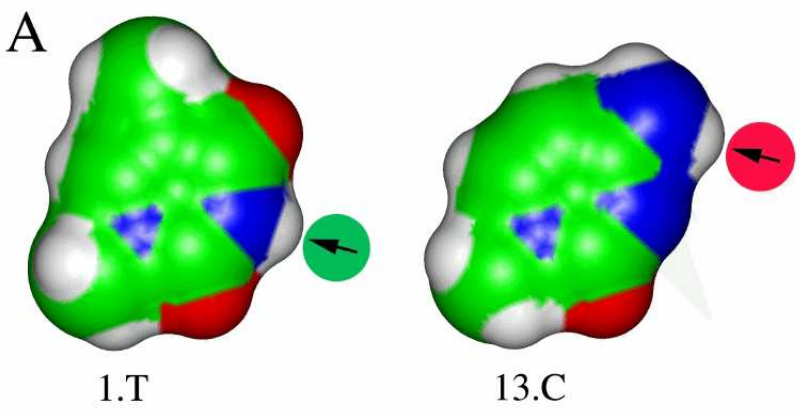


An important group that interacts with RepA is the N3 proton of Thymine or, in modified bases, a proton near the N3 atom. All bases that have a low Kd, below the distinct step, have a proton in this place.



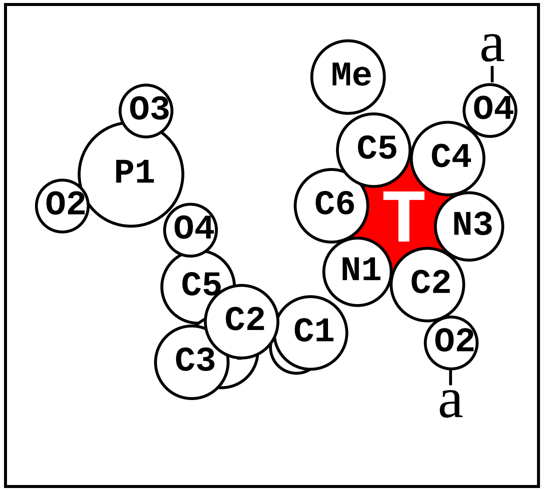


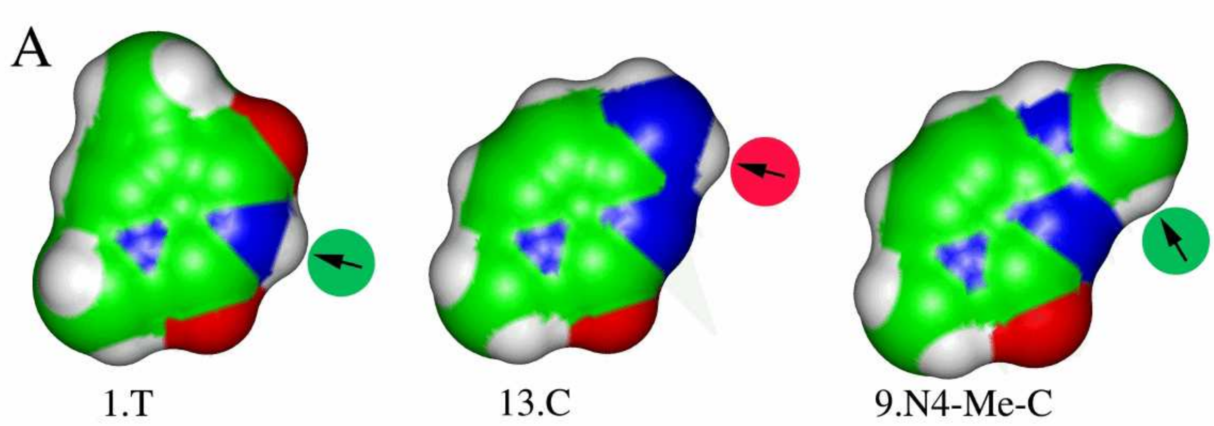




- **A. 1.T/A binds RepA stronger than 13.C/A.**

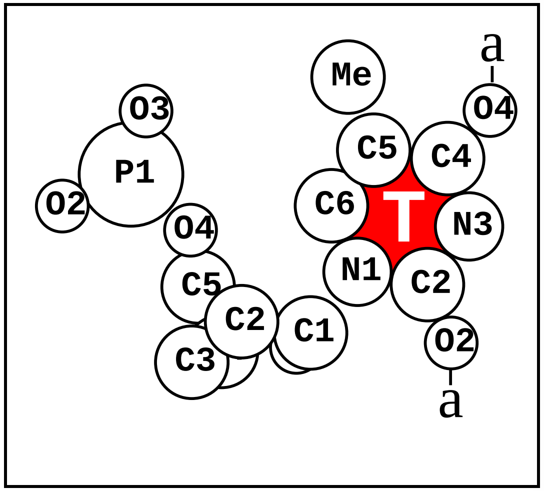
Base Structures at ± 7 P1 RepA sites

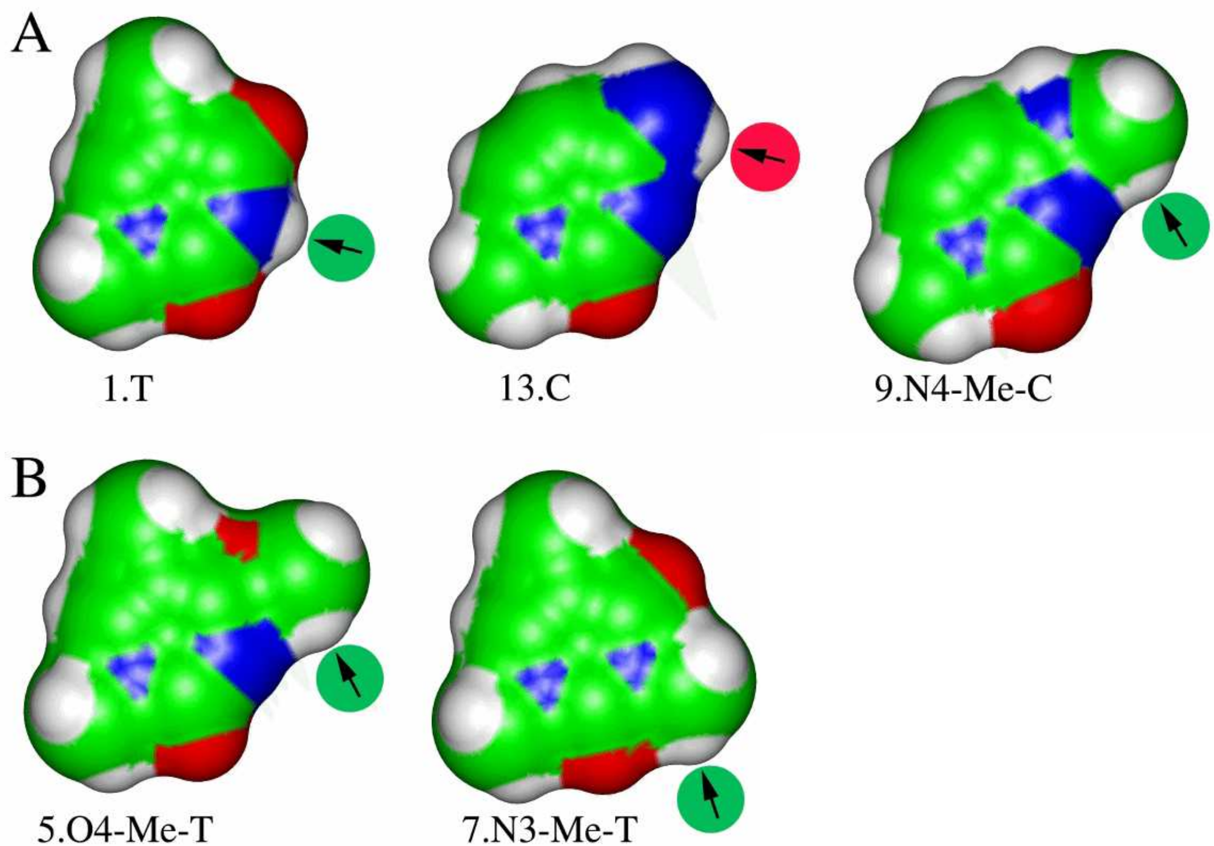




- **A. 1.T/A binds RepA stronger than 13.C/A.**
- **This low binding can be rescued by substituting the N4 proton with the methyl group in N4-Me-C.**

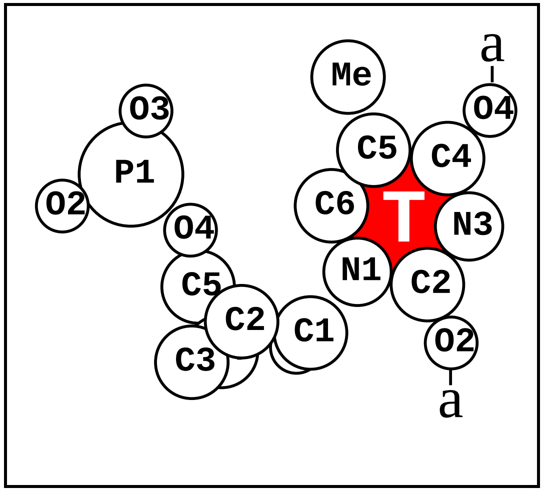
Base Structures at ± 7 P1 RepA sites

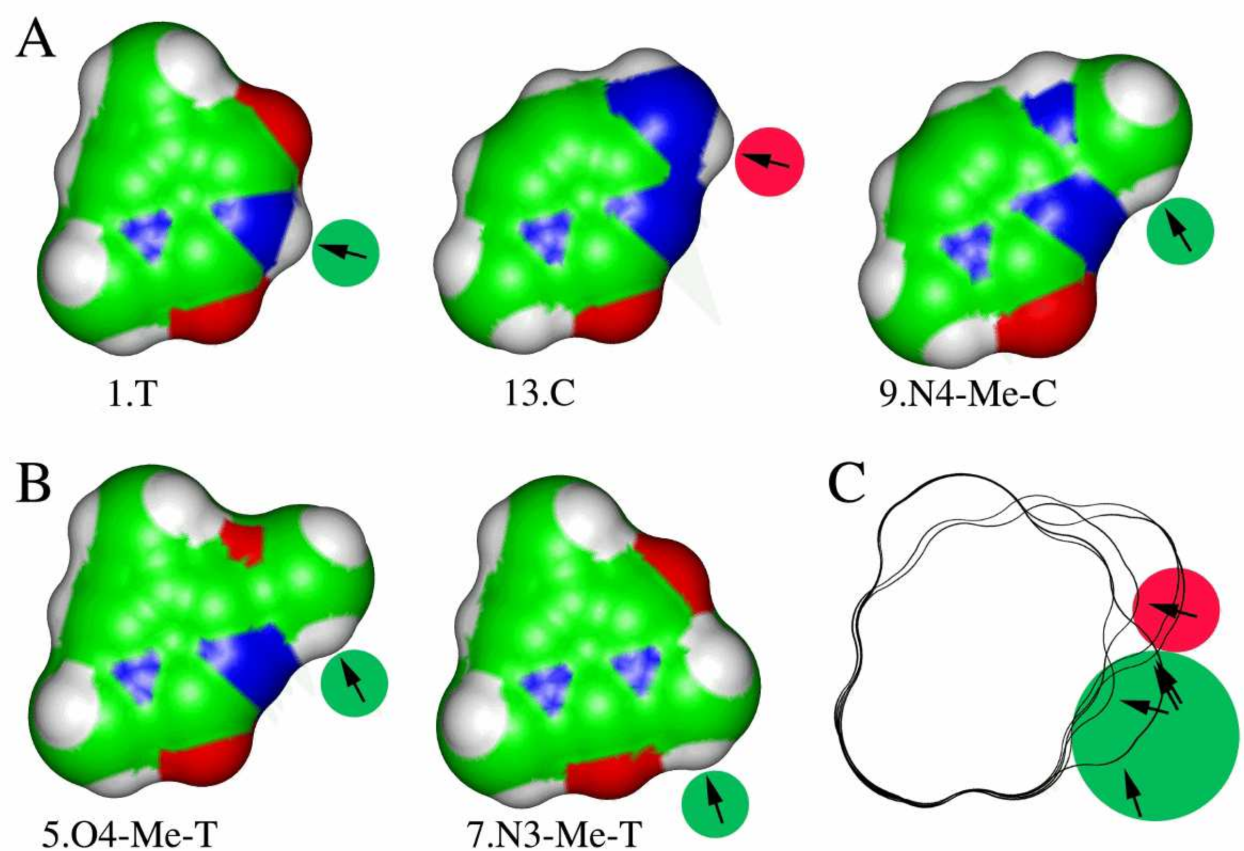




- **A. 1.T/A binds RepA stronger than 13.C/A.**
- **This low binding can be rescued by substituting the N4 proton with the methyl group in N4-Me-C.**
- **B. Two more structures that have a C-H proton near the N3 atom also bind RepA well.**

Base Structures at ±7 P1 RepA sites

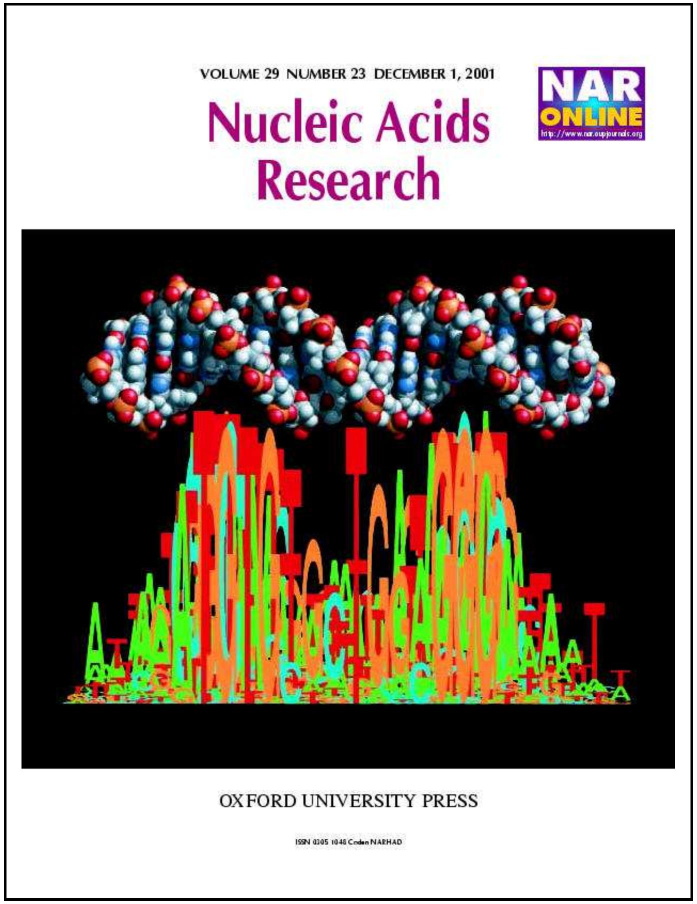




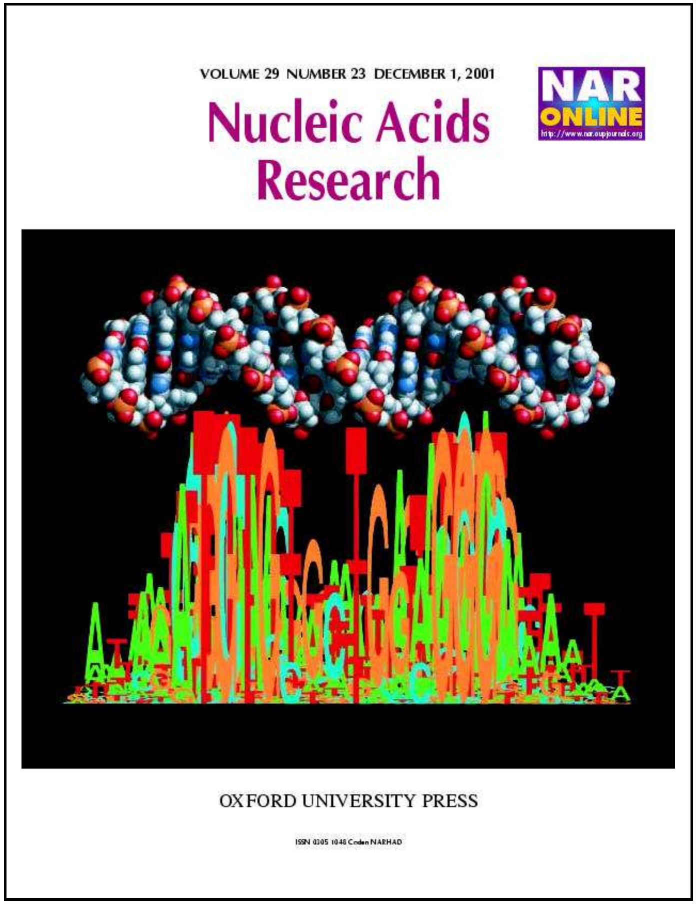
- **A. 1.T/A binds RepA stronger than 13.C/A.**
- **This low binding can be rescued by substituting the N4 proton with the methyl group in N4-Me-C.**
- **B. Two more structures that have a C-H proton near the N3 atom also bind RepA well.**
- **C. Outlines of the structures in A and B were aligned. Arrows 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.**

Lyakhov, Hengen, Rubens & Schneider, Nucleic Acids Res., 2001, 29:4892-4900

- **Sequence logo sine waves indicate protein orientation on DNA**

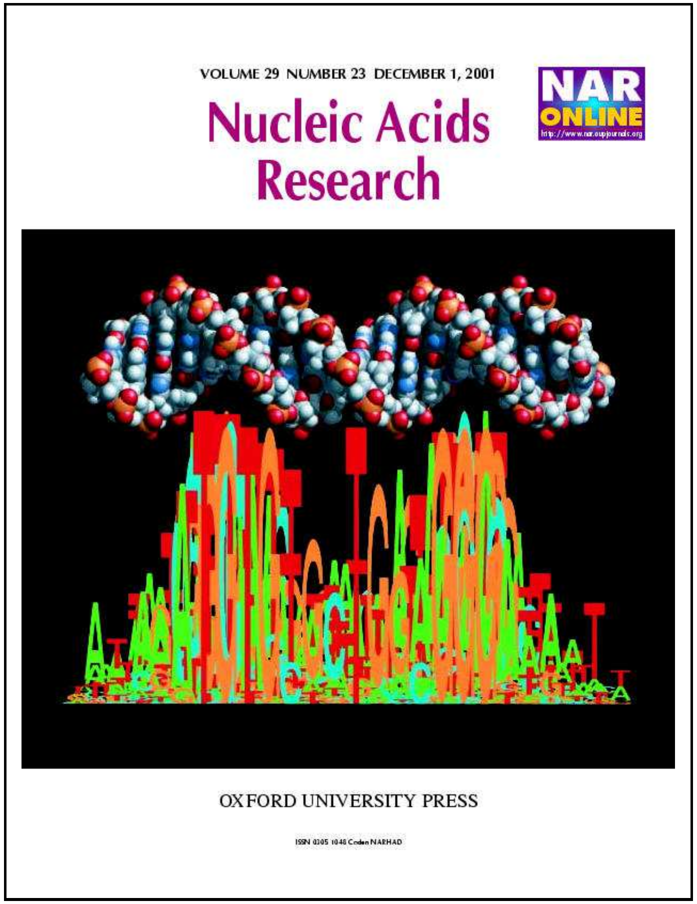


- **Sequence logo sine waves indicate protein orientation on DNA**
- **Anomalies in the logo can reveal interesting binding modes**



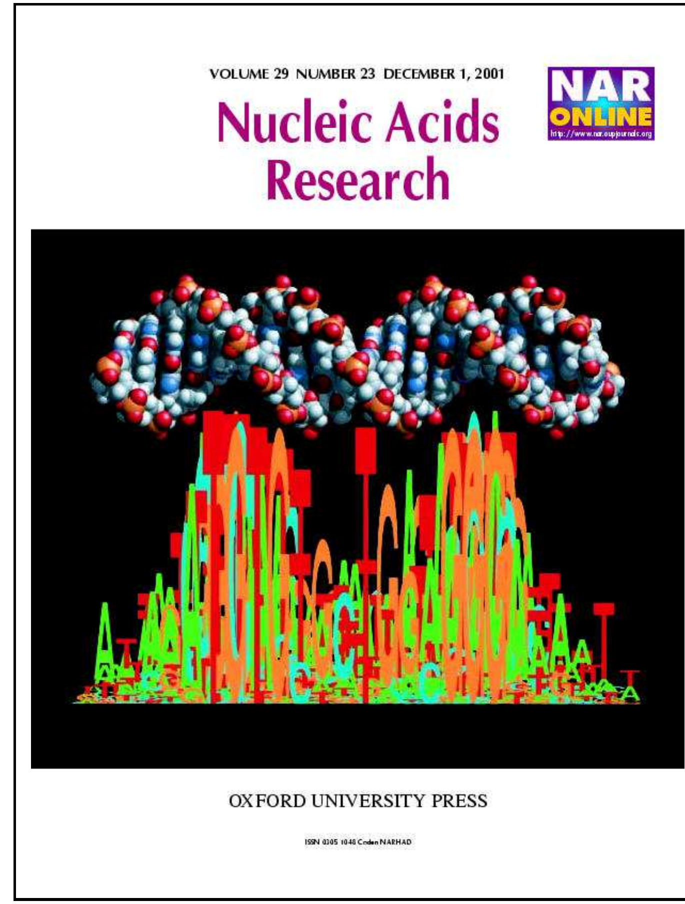
Conclusions

- **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

version = 1.24 of helixtalk.tex 2012 Nov 07