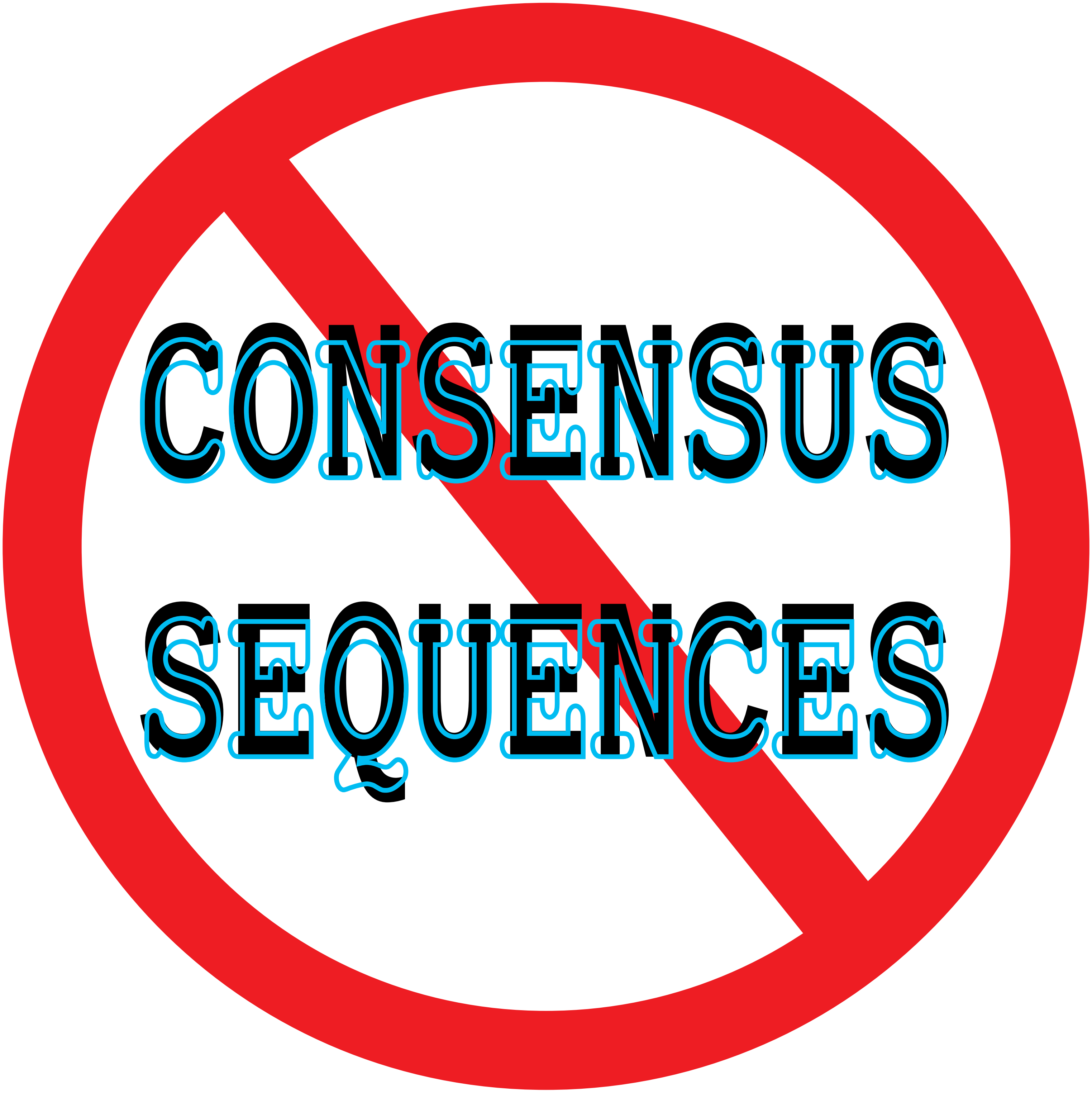
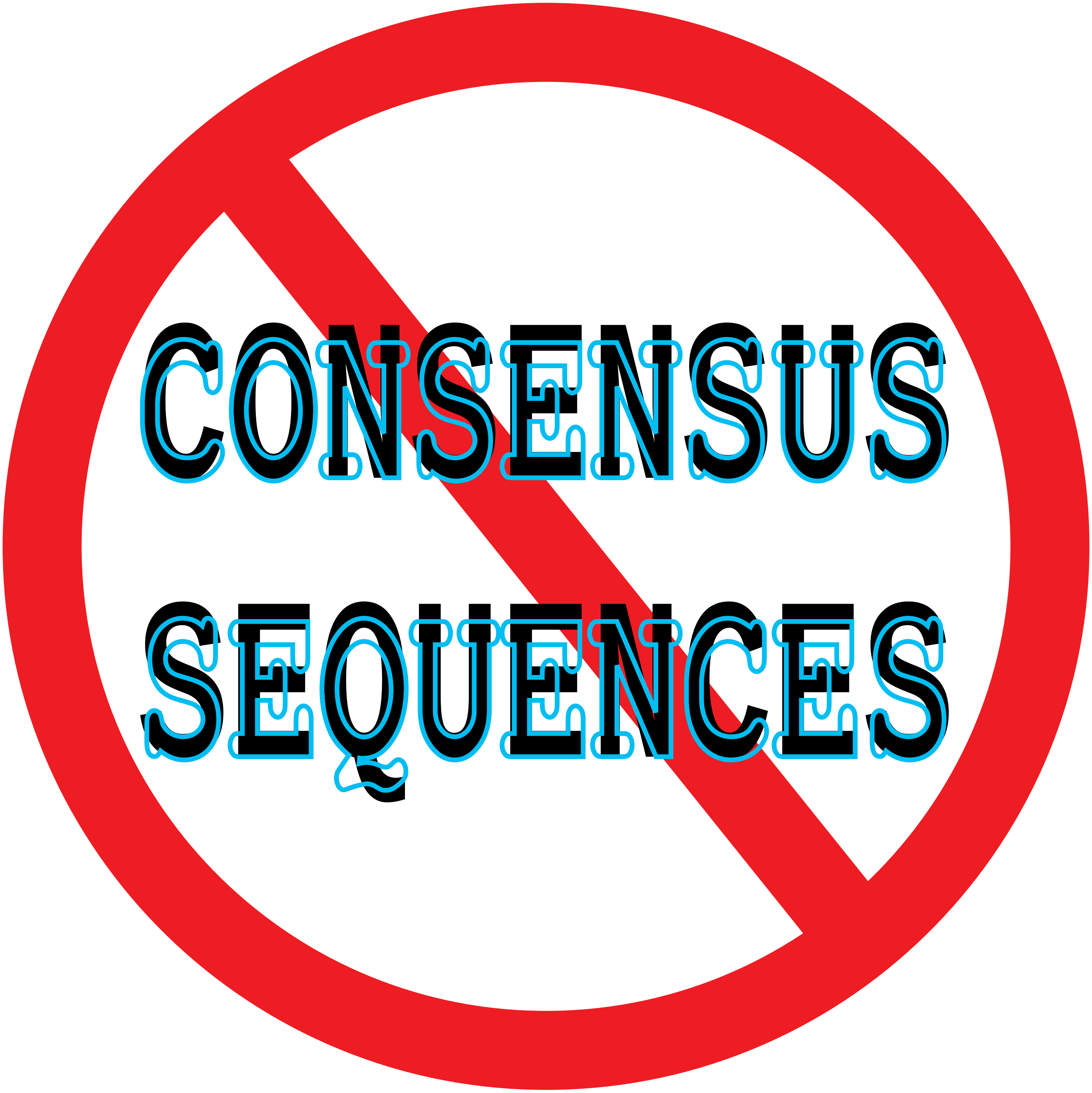
Consensus Sequences:
Just Say No!

Thomas D. Schneider
Laboratory of Mathematical Biology
National Cancer Institute
Frederick, MD 21702.

toms@ncifcrf.gov
http://www-lmmb.ncifcrf.gov/~toms/

1995 December 4


CONSENSUS
SEQUENCES



CONSENSUS SEQUENCES

Summary

Consensus sequences are being used to characterize the binding sites of macromolecules on DNA and RNA. After aligning a set of binding site sequences, the most frequent base is chosen. A position which contains 100% A's will be represented by an A, while a position that is only 75% A will also be represented by an A. The consensus is frequently used to search for binding sites, and the number of mismatches to the consensus is counted. A mismatch to a 100% A position is much more severe than one to a 75% A, but this is not accounted for so the researcher is mislead. We present mathematically robust graphical replacements for the consensus sequence called the sequence logo and the walker that do not discard your hard-earned data. Further information and examples may be found on the internet at http://www-lmmb.ncifcrf.gov/~toms/.

Consensus Sequences

Characterize what a binding site looks like

- ❄ Use Sequence Logos instead

Search for new sites

- ❄ Use Rindividual Matrix Scans instead

Investigate how well the bases of a sequence match to functional binding sites

- ❄ Use the Walker instead

Summary

Consensus sequences are being used to characterize the binding sites of macromolecules on DNA and RNA. After aligning a set of binding site sequences, the most frequent base is chosen. A position which contains 100% A's will be represented by an A, while a position that is only 75% A will also be represented by an A. The consensus is frequently used to search for binding sites, and the number of mismatches to the consensus is counted. A mismatch to a 100% A position is much more severe than one to a 75% A, but this is not accounted for so the researcher is mislead. We present mathematically robust graphical replacements for the consensus sequence called the sequence logo and the walker that do not discard your hard-earned data. Further information and examples may be found on the internet at http://www-lmmb.ncifcrf.gov/~toms/.

Consensus Sequences

Characterize what a binding site looks like

❄ Use Sequence Logos instead

Search for new sites

❄ Use Rindividual Matrix Scans instead

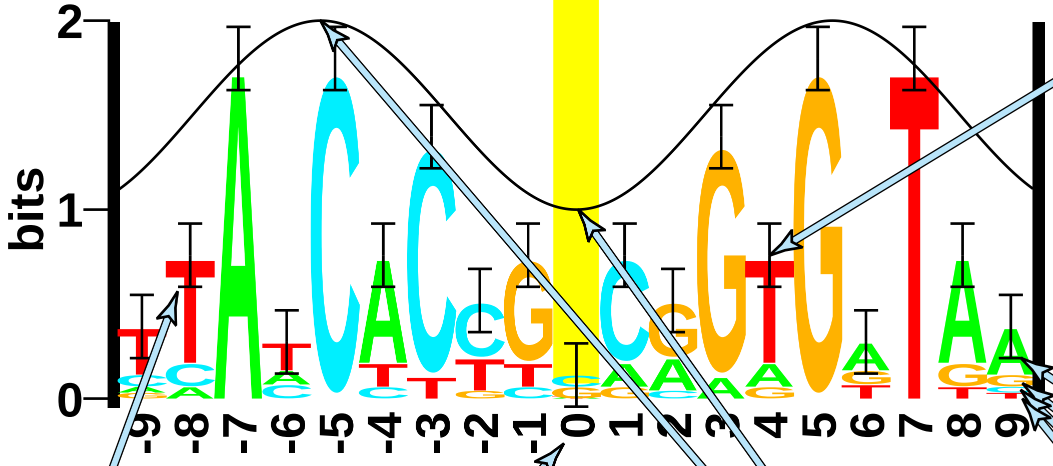
Investigate how well the bases of a sequence match to functional binding sites

❄ Use the Walker instead

SEQUENCE LOGO

```
    --------- +++++++++
    9876543210123456789
    ...................
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

raw sequences of binding sites



12 Lambda cI and cro binding sites

The height of each stack of letters is the information content (sequence conservation) in bits per base

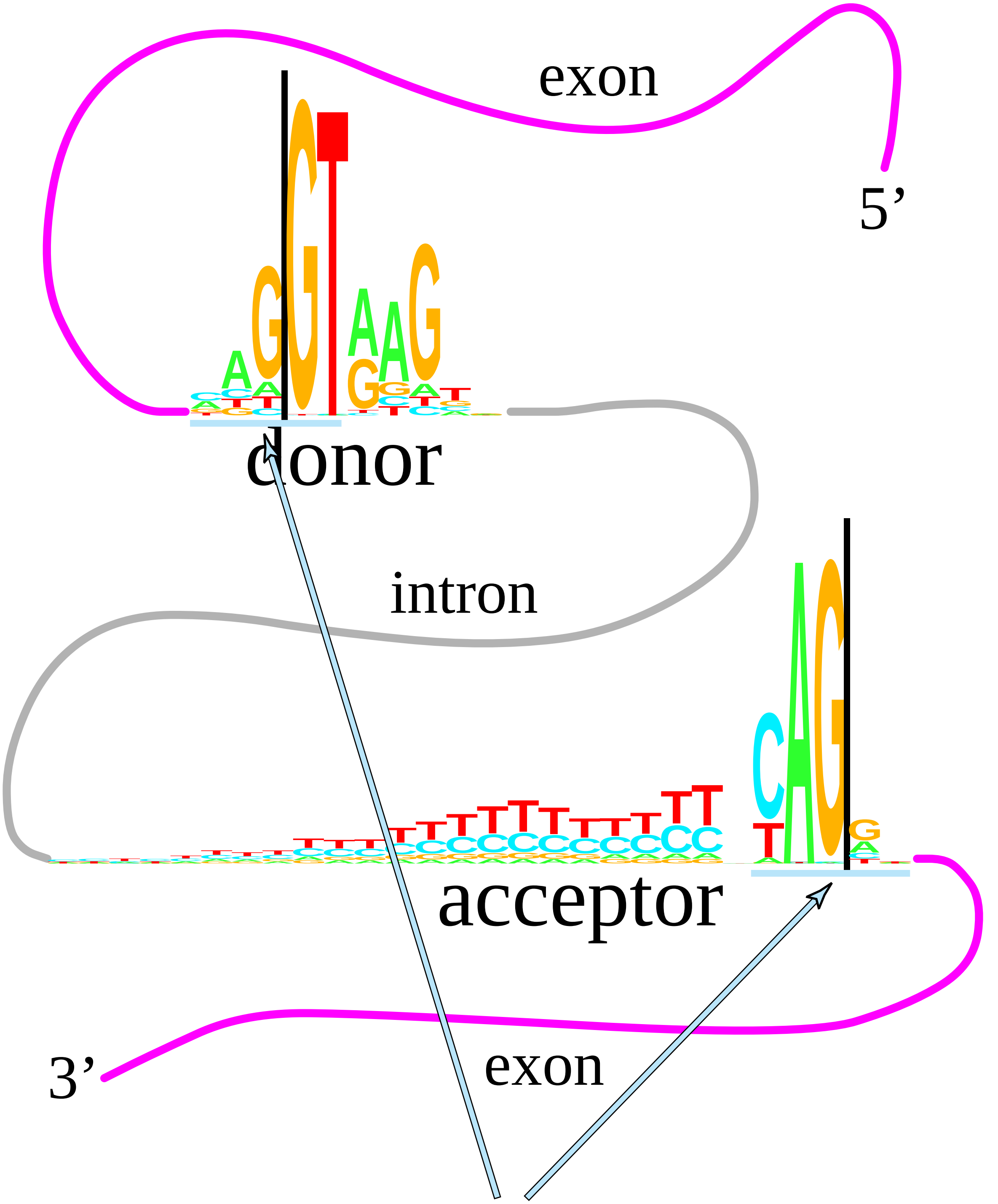
letter heights are proportional to frequency:
8/12 = 67% A
2/12 = 17% G
1/12 = 8% C
1/12 = 8% T

Error bar for entire stack of letters

Zero coordinate for alignment

Minor Groove Faces Protein
Major Groove Faces Protein

How can two binding sites be different but have the same consensus sequence?

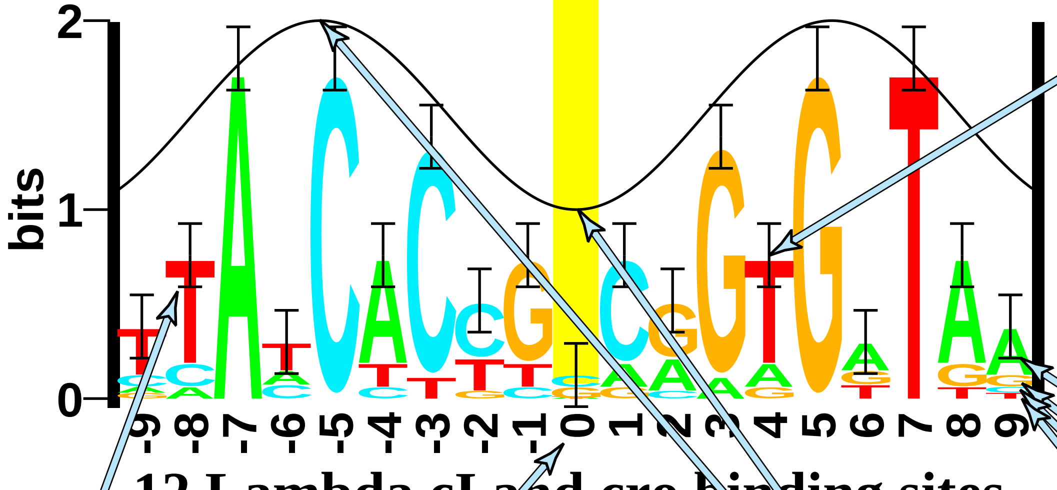


These two sequence logos have the same consensus sequence (CAGGT) but different emphasis

❄ SEQUENCE LOGO

```
    --------- +++++++++
    9876543210123456789
    ...................
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

raw sequences of binding sites



12 Lambda cI and cro binding sites

The height of each stack of letters is the information content (sequence conservation) in bits per base

letter heights are proportional to frequency:
8/12 = 67% A
2/12 = 17% G
1/12 = 8% C
1/12 = 8% T

Error bar for entire stack of letters

Zero coordinate for alignment

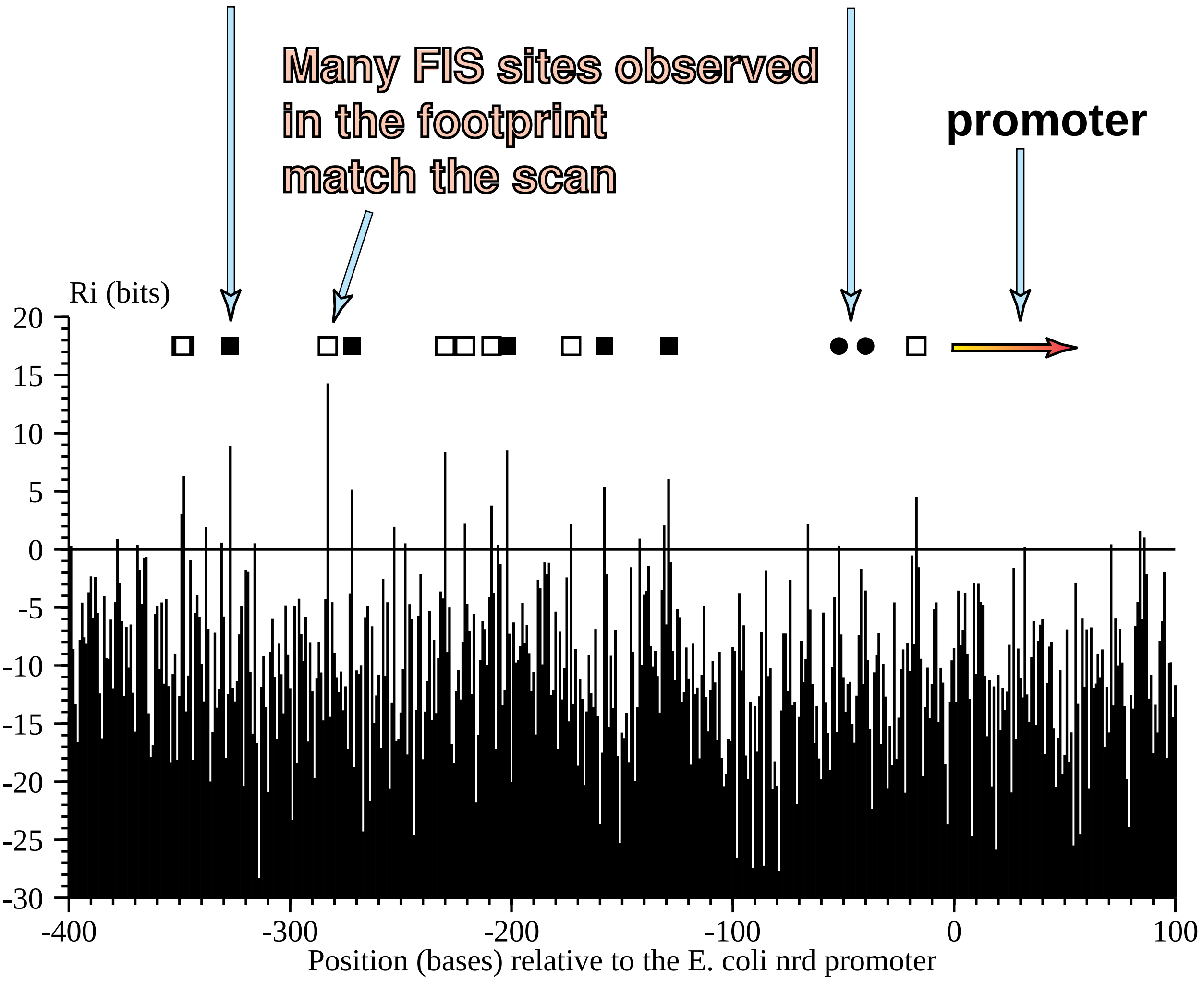
Minor Groove Faces Protein
Major Groove Faces Protein

❄ SCAN

The individual information weight matrix is put at every position of a sequence.

The weights are added together depending on the sequence.

This gives the total Rindividual (Ri) at every position in the sequence.



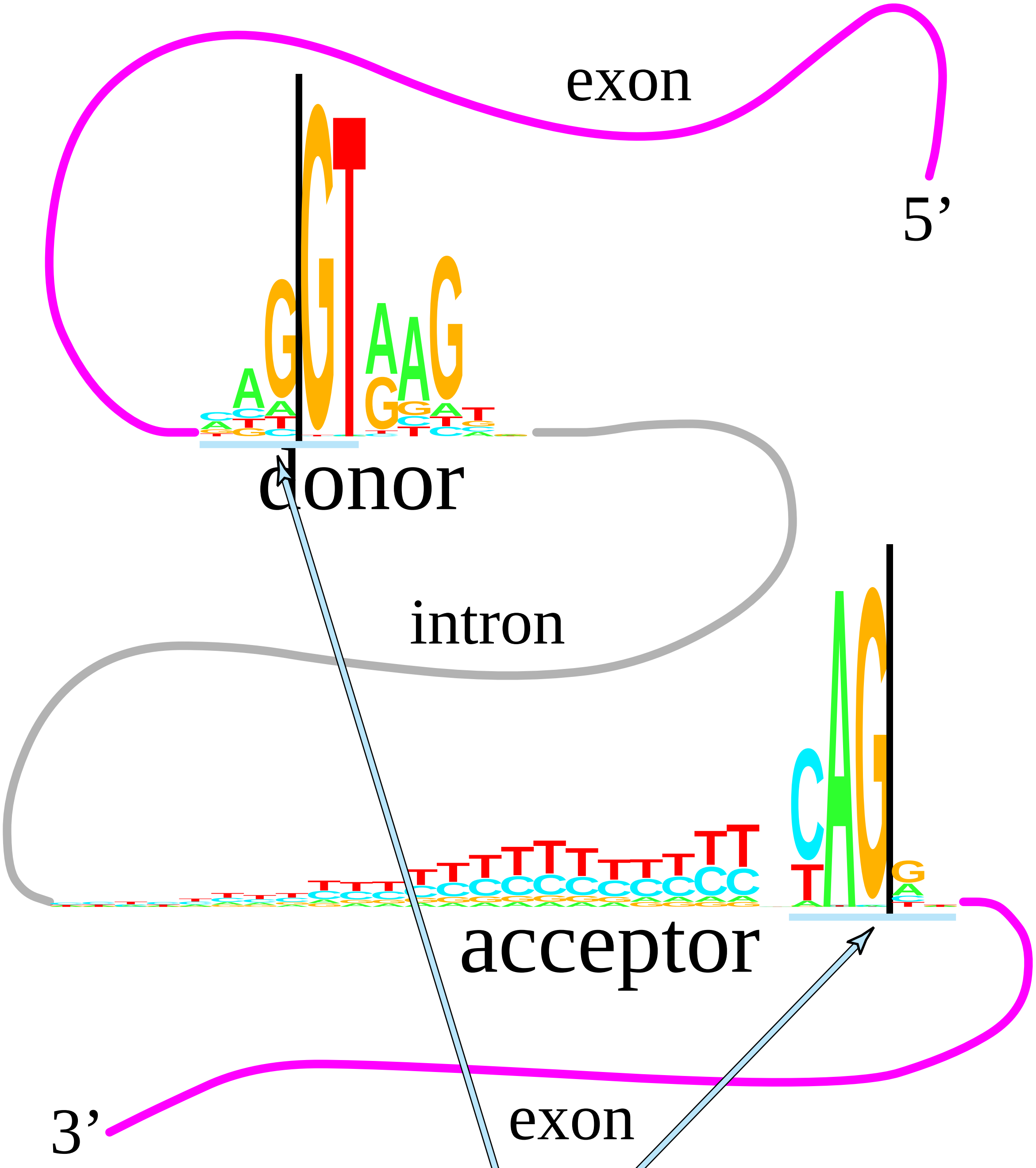
Fis sites predicted from consensus

DnaA sites

Many FIS sites observed in the footprint match the scan

promoter

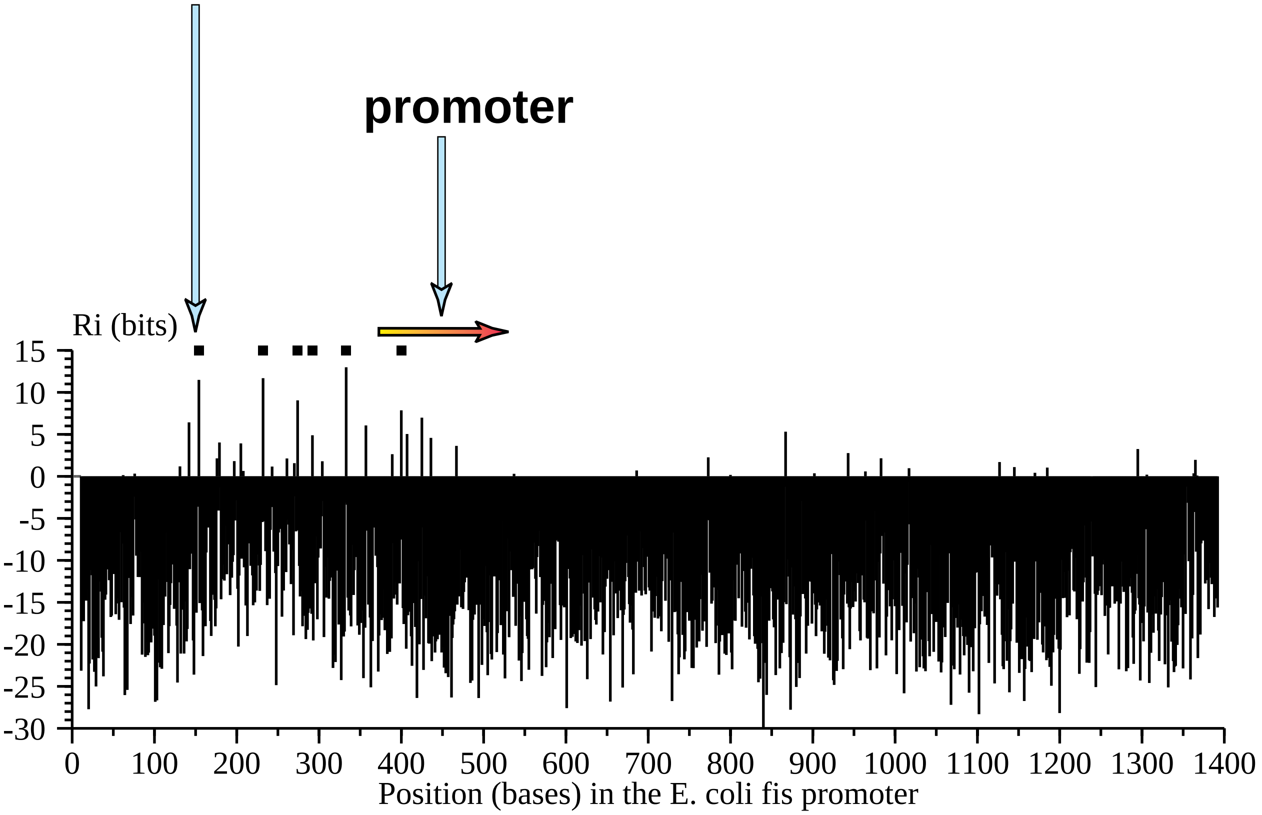
How can two binding sites be different but have the same consensus sequence?



These two sequence logos have the same consensus sequence (CAGGT) but different emphasis

Scan of Fis Promoter

6 Fis sites were predicted
on the *E. coli* Fis promoter
from footprints and consensus



promoter

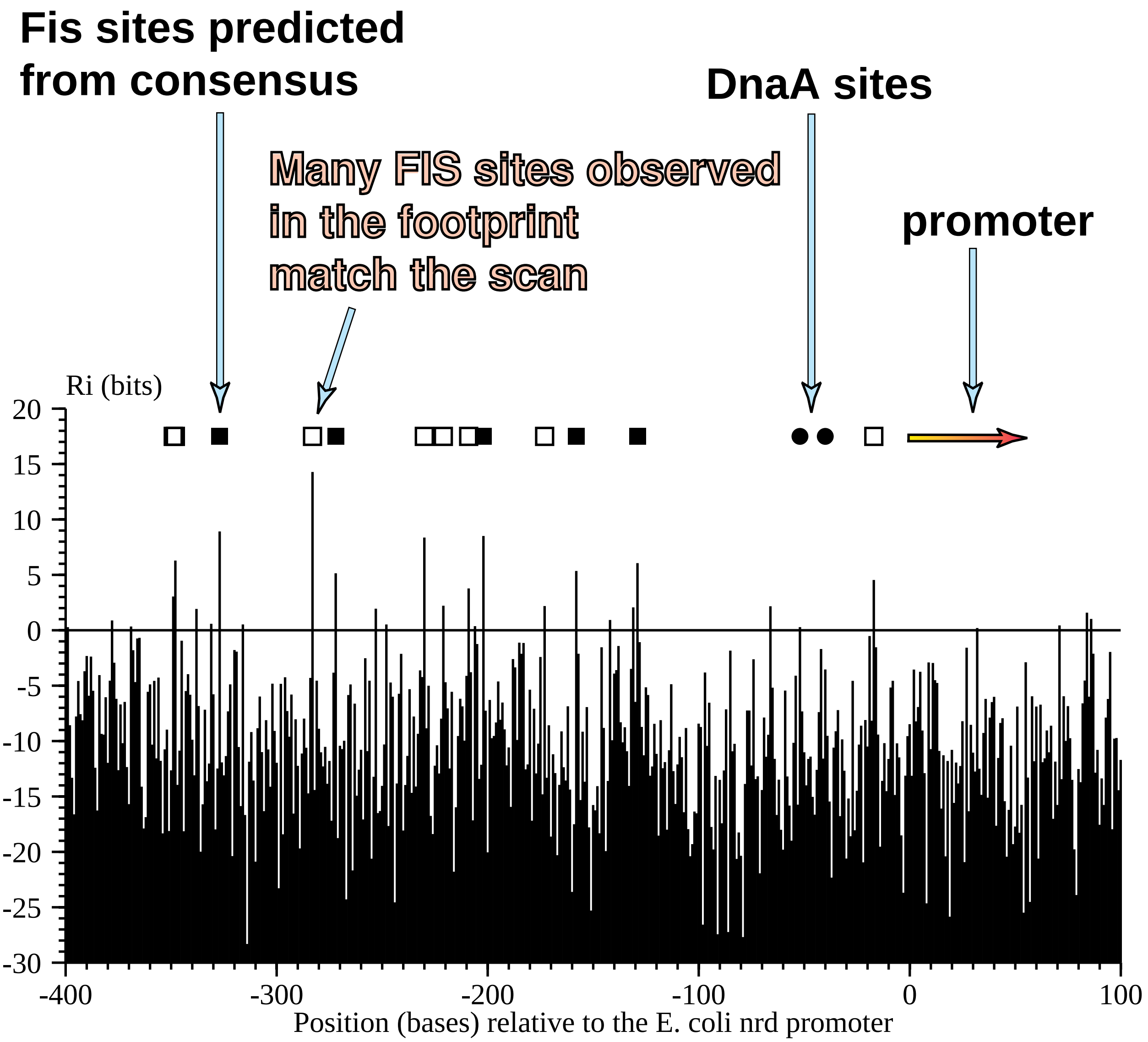
Many potential Fis sites besides those already identified on the Fis promoter are observed. The large number near the promoter indicates that Fis controls its own transcription.

❄ SCAN

The individual information weight matrix is put at every position of a sequence.

The weights are added together depending on the sequence.

This gives the total Rindividual (Ri) at every position in the sequence.



Fis sites predicted from consensus

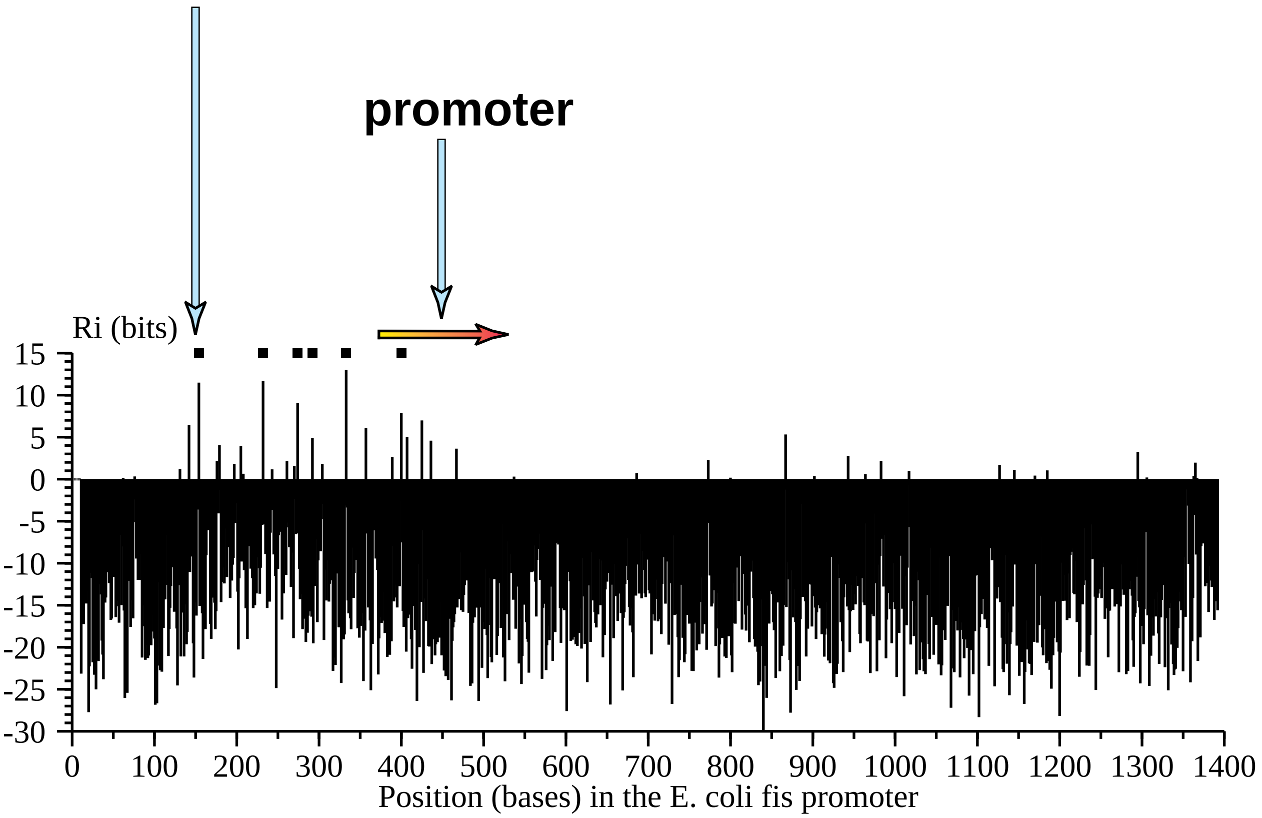
DnaA sites

Many FIS sites observed in the footprint match the scan

promoter

Scan of Fis Promoter

6 Fis sites were predicted
on the *E. coli* Fis promoter
from footprints and consensus

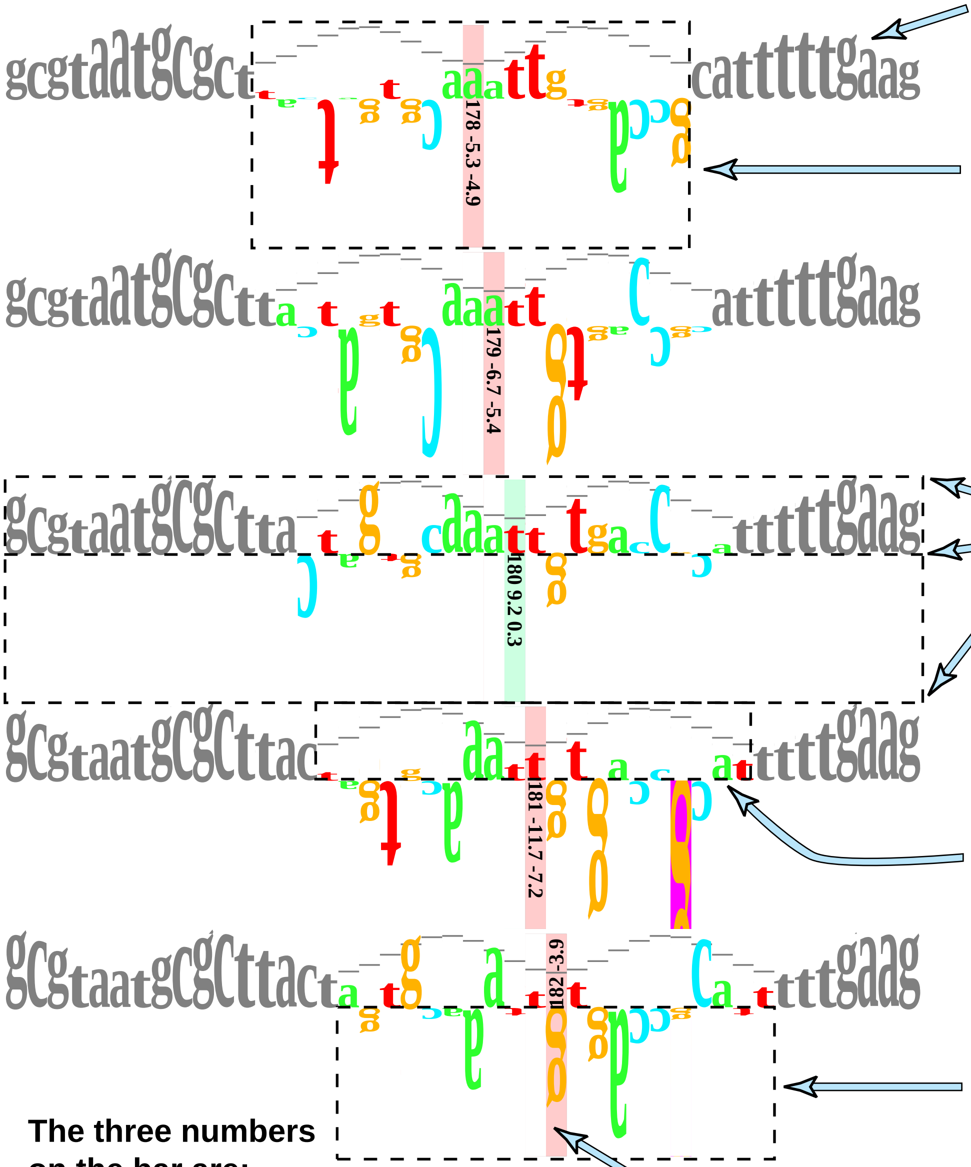


Many potential Fis sites besides
those already identified
on the Fis promoter are observed.
The large number near the promoter
indicates that Fis
controls its own transcription.

❄ WALKER

How well do the bases of a sequence match to functional binding sites?

5 copies of a single sequence are below:



The cosine wave represents the orientation of the DNA facing the protein.

The Walker is the colored letters.

The weight matrix is for the *E. coli* Fis protein.

The vertical bar is a scale:
the top is at 2 bits
the middle is at 0 bits
the bottom is at -4 bits

Letters that go up are preferred at that position.

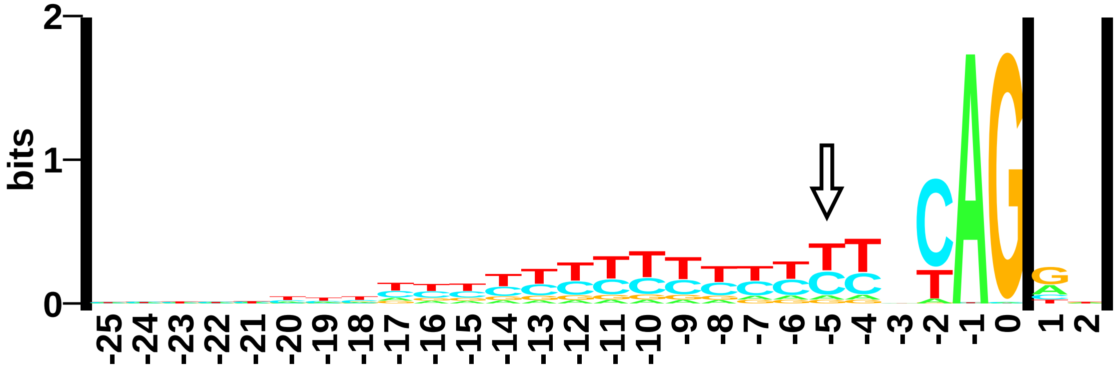
Letters that go down are detrimental at that position (purple goes below -4 bits).

The three numbers on the bar are:
❄ position
❄ Rindividual
❄ standard deviations from mean (Z)

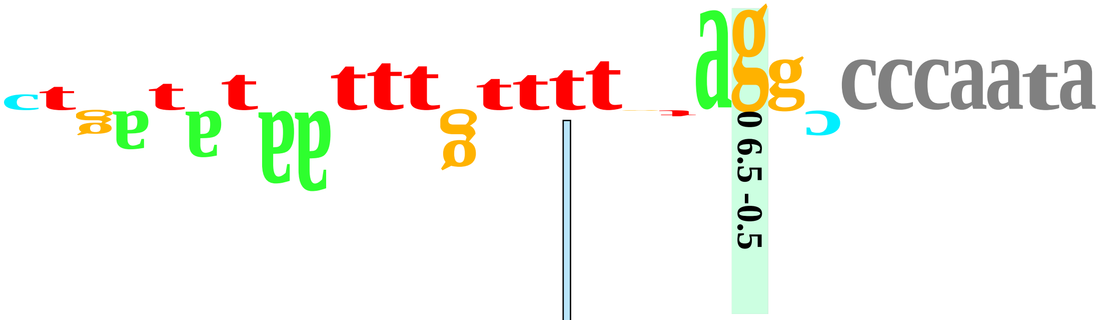
A green vertical bar indicates a binding site, red one means it is probably not a site.

Is a sequence change a Mutation or a Polymorphism?

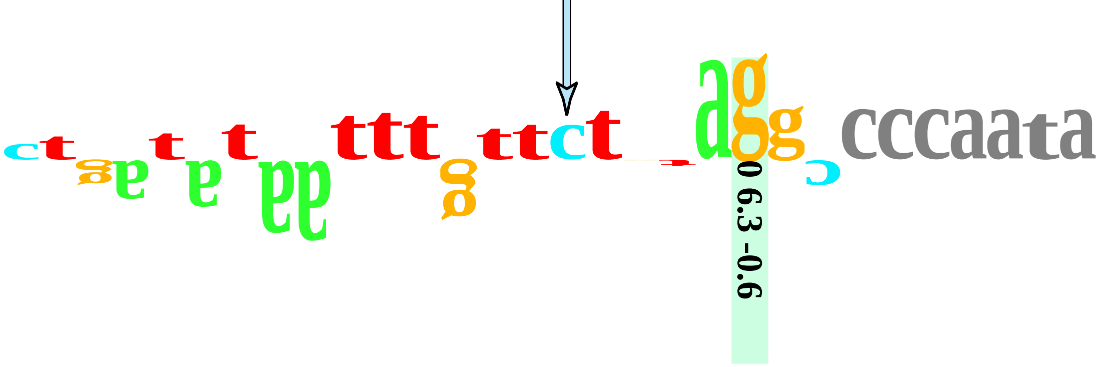
A T to C change seen in a splice acceptor of hMSH2 was interpreted to be the mutation which causes familial nonpolyposis colon cancer (Fishel *et al.*, Cell 75:1027-1038, 1993):



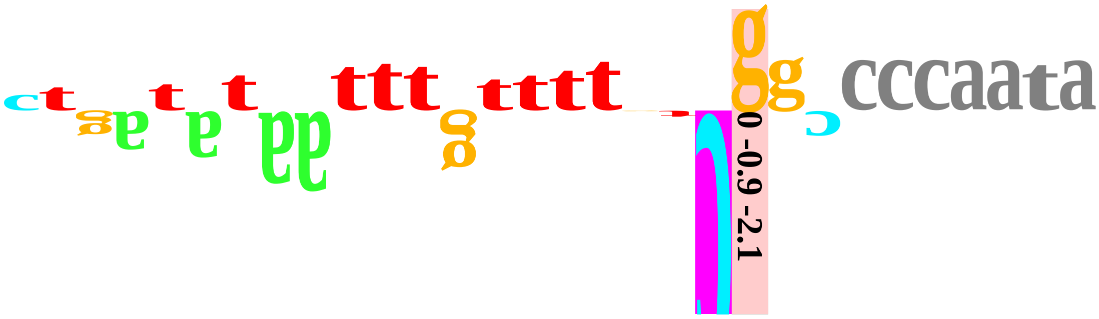
The sequence logo shows nearly equal frequencies of bases there.



wild-type as seen by a walker



The walker shows it is a polymorphism.

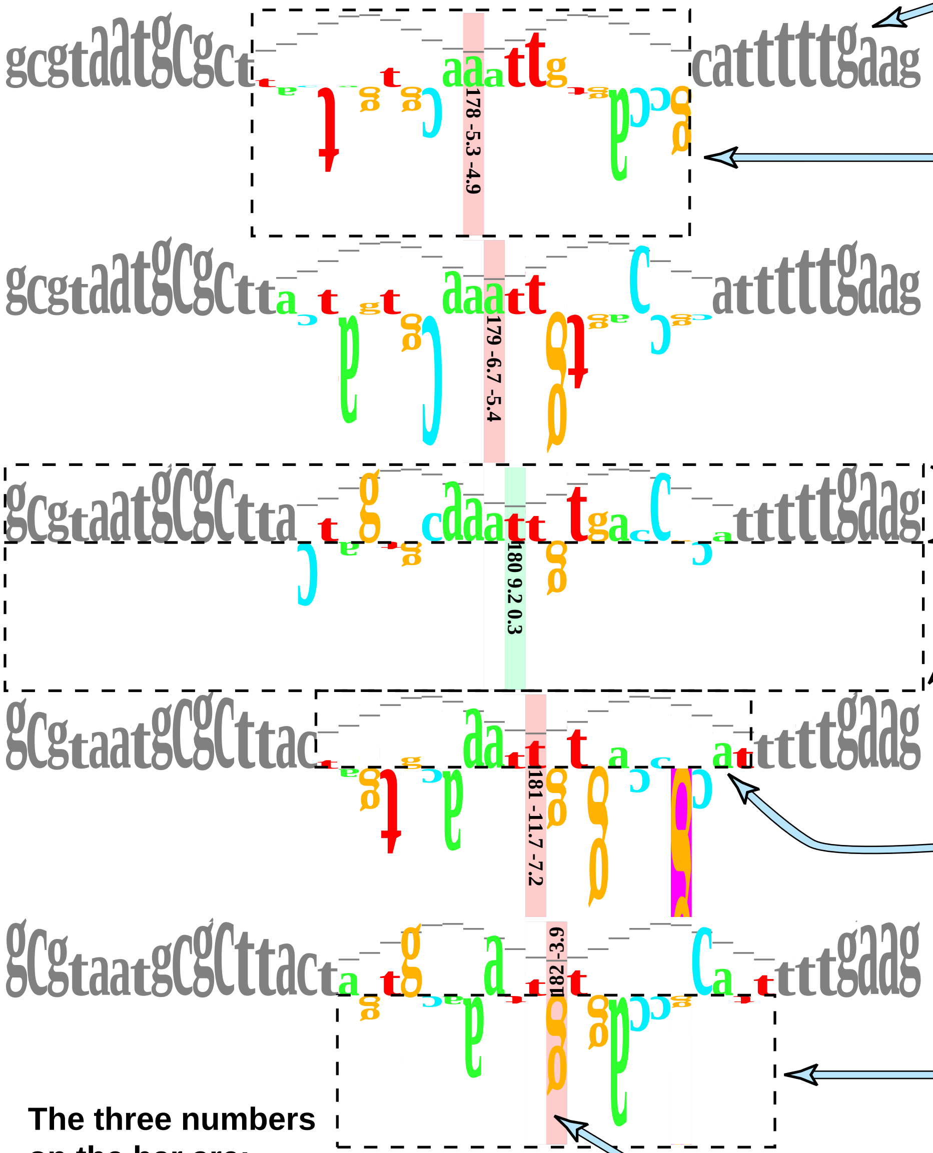


This is what a strong mutation would look like.

❄ WALKER

How well do the bases of a sequence match to functional binding sites?

5 copies of a single sequence are below:



The cosine wave represents the orientation of the DNA facing the protein.

The Walker is the colored letters.

The weight matrix is for the *E. coli* Fis protein.

The vertical bar is a scale:
the top is at 2 bits
the middle is at 0 bits
the bottom is at -4 bits

Letters that go up are preferred at that position.

Letters that go down are detrimental at that position (purple goes below -4 bits).

The three numbers on the bar are:

- ❄ position
- ❄ Rindividual
- ❄ standard deviations from mean (Z)

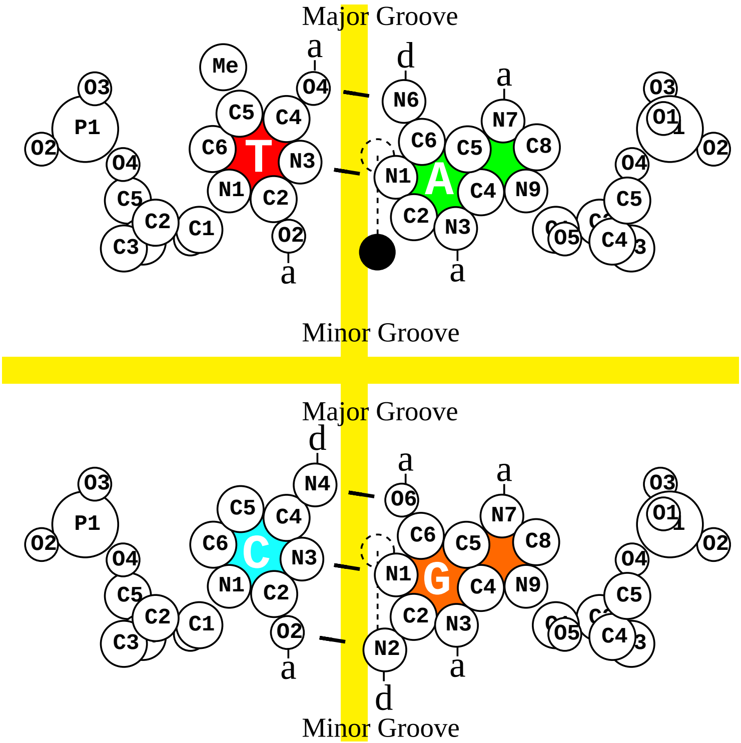
A green vertical bar indicates a binding site, red one means it is probably not a site.

INFORMATION THEORY

A *BIT* measures the choice between 2 equally likely possibilities:

one bit
is like a knife
slice

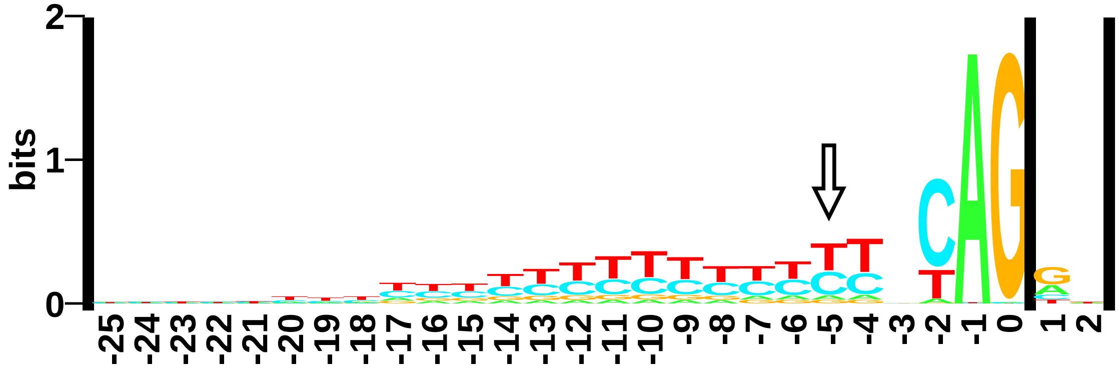
To choose one base in 4 requires two bits:



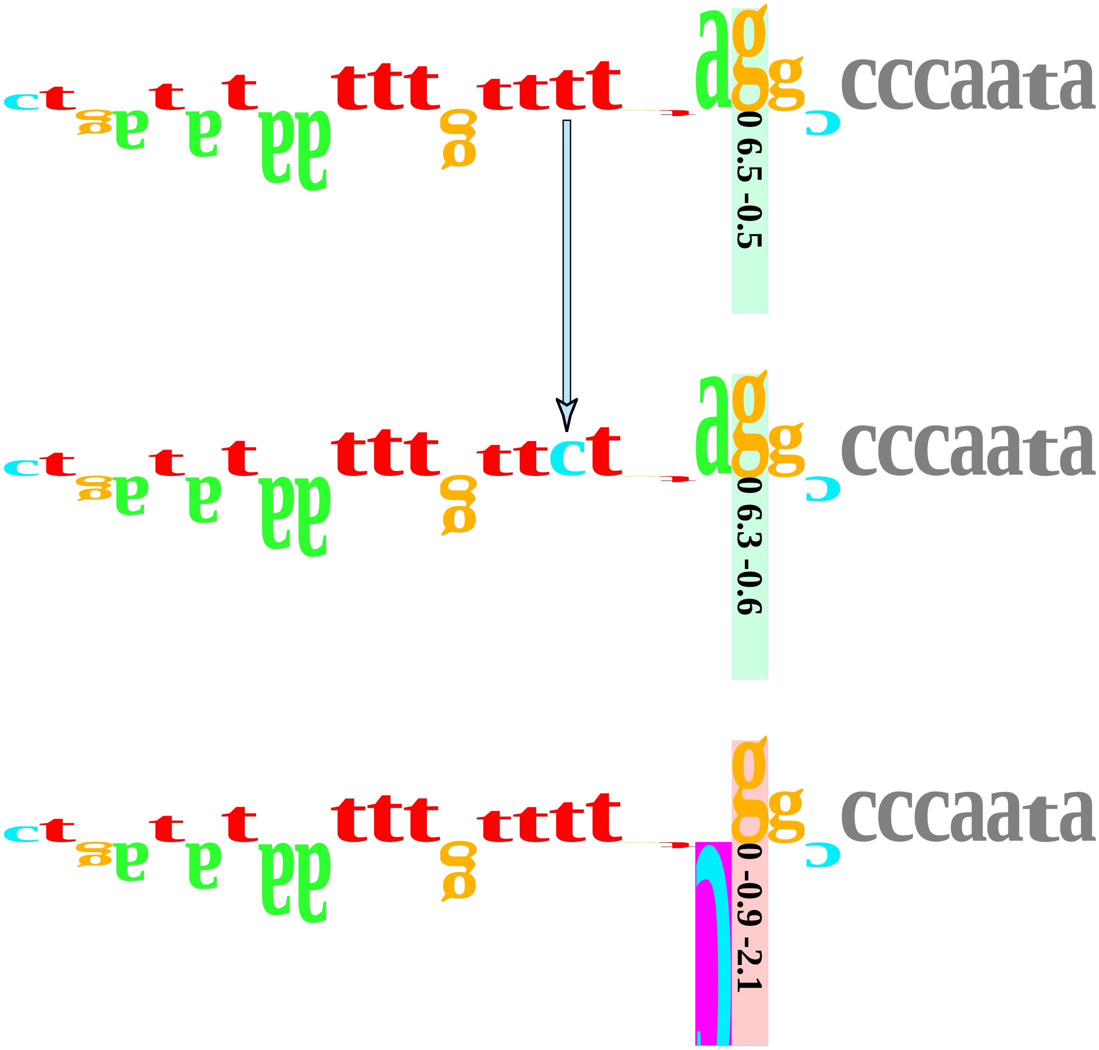
two bits
are like two
knife slices

Is a sequence change a Mutation or a Polymorphism?

A T to C change seen in a splice acceptor of hMSH2 was interpreted to be the mutation which causes familial nonpolyposis colon cancer (Fishel *et al.*, Cell 75:1027-1038, 1993):



The sequence logo shows nearly equal frequencies of bases there.



wild-type as seen by a walker

The walker shows it is a polymorphism.

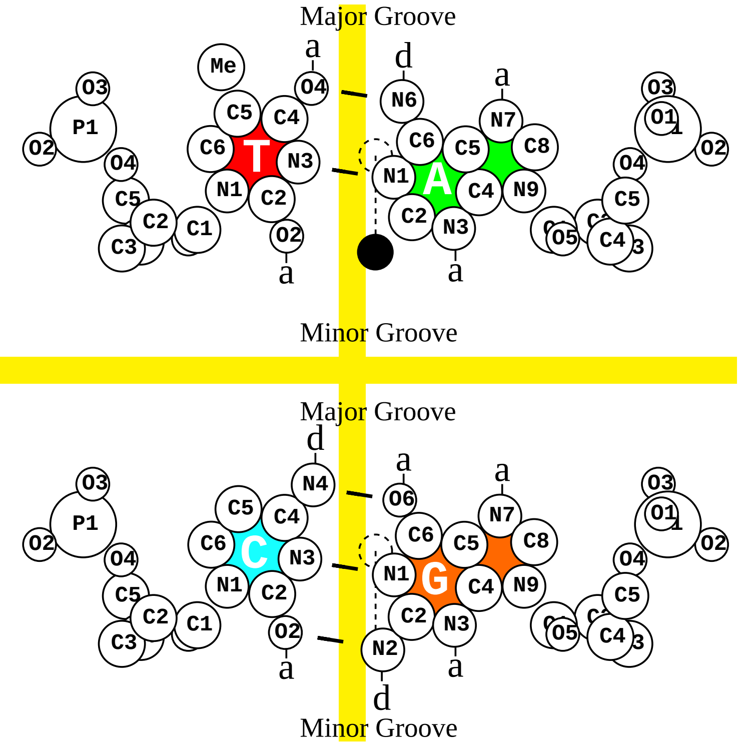
This is what a strong mutation would look like.

INFORMATION THEORY

A *BIT* measures the choice between 2 equally likely possibilities:

one bit
is like a knife
slice

To choose one base in 4 requires two bits:



two bits
are like two
knife slices

Information measured in bits

- is additive
- is well supported mathematically
- measures sequence conservation
- is related to the entropy
- is calculated as a decrease in the uncertainty H:

  $H = -\Sigma\, p_i \log_2 p_i$ (bits)

  as

  $R = -\Delta H$ (bits/symbol)

  R is the Rate of information transmission.

Every sequence has an individual information

```
    --------- +++++++++
    9876543210123456789
    ...................
 1  gtatcaccgccagtggtat 17.7 bits
 2  ataccactggcggtgatac 17.7 bits
 3  tcaacaccgccagagataa 19.3 bits
 4  ttatctctggcggtgttga 19.3 bits
 5  ttatcaccgcagatggtta 15.7 bits
 6  taaccatctgcggtgataa 15.7 bits
 7  ctatcaccgcaagggataa 17.3 bits
 8  ttatcccttgcggtgatag 17.3 bits
 9  ctaacaccgtgcgtgttga 11.0 bits
10  tcaacacgcacggtgttag 11.0 bits
11  ttacctctggcggtgataa 21.5 bits
12  ttatcaccgccagaggtaa 21.5 bits
```

$$\Sigma = 205 \text{ bits}$$

$$\Sigma /12 = 17.1 \text{ bits}$$

The average of the individual information values for the original sequence is the same as the sequence conservation and the area under the sequence logo.

Information measured in bits

❄ is additive

❄ is well supported mathematically

❄ measures sequence conservation

❄ is related to the entropy

❄ is calculated as a decrease in the uncertainty H:

$$H = -\Sigma\, p_i \log_2 p_i \text{ (bits)}$$

as

$$R = -\Delta H \qquad \text{(bits/symbol)}$$

R is the Rate of information transmission.

Every sequence has an individual information

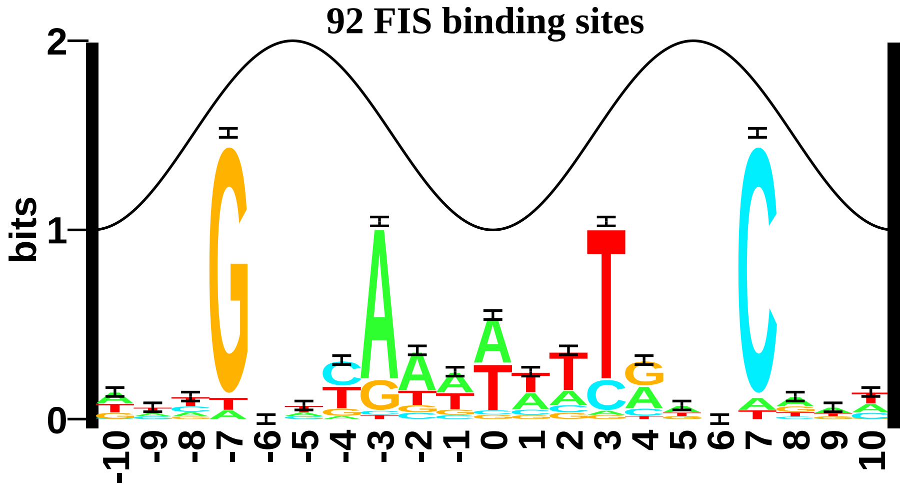
```
   --------- +++++++++
   9876543210123456789
   ...................
 1 gtatcaccgccagtggtat 17.7 bits
 2 ataccactggcggtgatac 17.7 bits
 3 tcaacaccgccagagataa 19.3 bits
 4 ttatctctggcggtgttga 19.3 bits
 5 ttatcaccgcagatggtta 15.7 bits
 6 taaccatctgcggtgataa 15.7 bits
 7 ctatcaccgcaagggataa 17.3 bits
 8 ttatcccttgcggtgatag 17.3 bits
 9 ctaacaccgtgcgtgttga 11.0 bits
10 tcaacacgcacggtgttag 11.0 bits
11 ttacctctggcggtgataa 21.5 bits
12 ttatcaccgccagaggtaa 21.5 bits
```

$$\Sigma = 205 \text{ bits}$$

$$\Sigma /12 = 17.1 \text{ bits}$$

The average of the individual information values for the original sequence is the same as the sequence conservation and the area under the sequence logo.

Individual Information



The area under a sequence logo is the total sequence conservation.

The mathematics makes it look like an average.

QUESTION: What is it the average of?

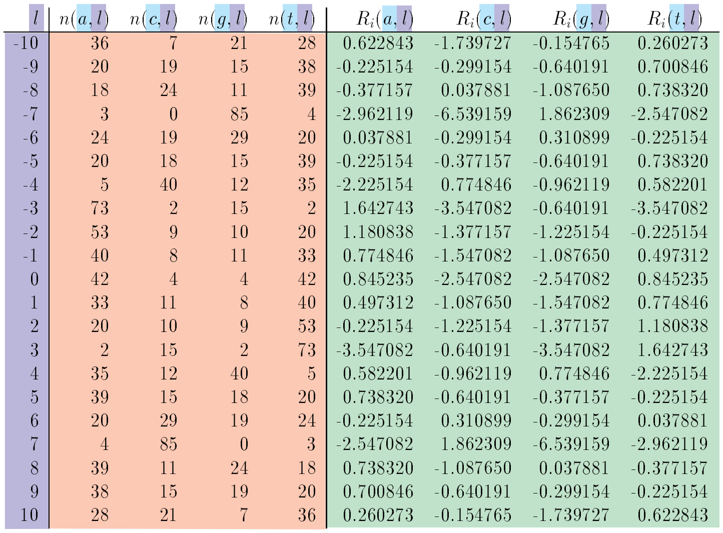
ANSWER: The information of each individual binding site.

This is found from the individual information weight matrix as:

$$Ri(b,l) = 2 + \log_2 f(b,l)$$

where f(b,l) is the frequency of each base b at every position l in the binding site.

number of bases b at each position l



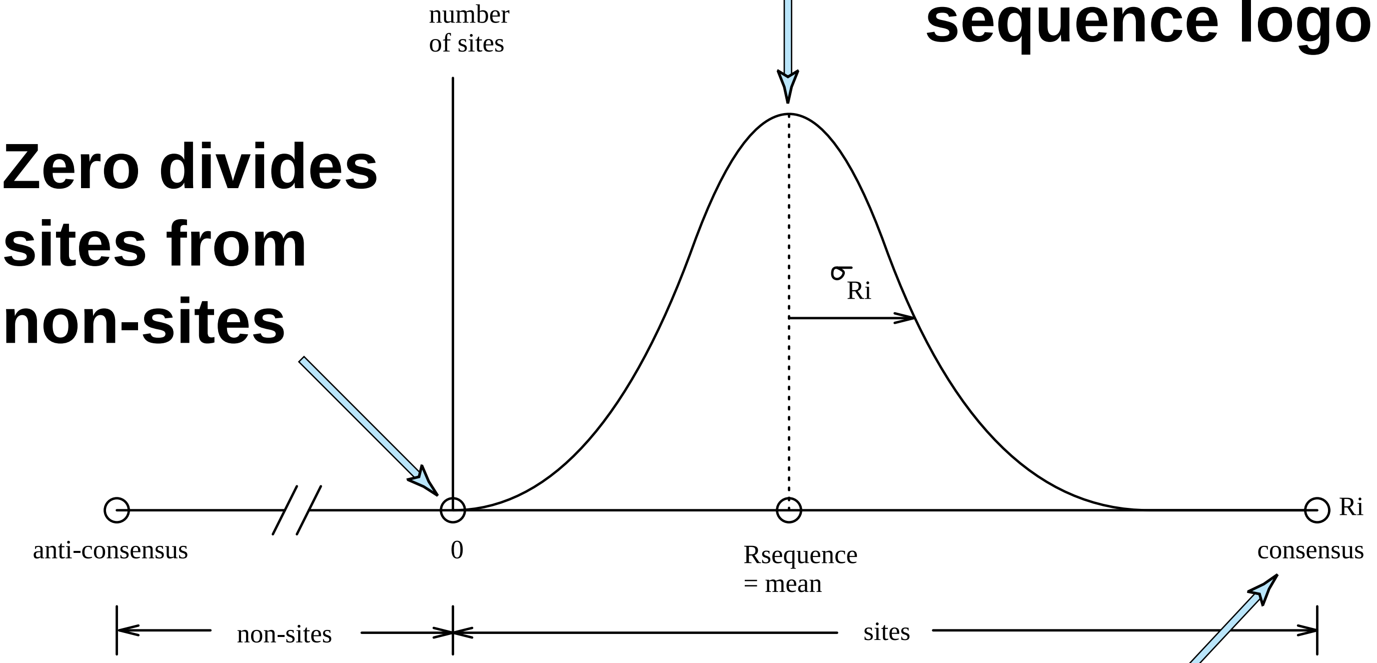
l	$n(a,l)$	$n(c,l)$	$n(g,l)$	$n(t,l)$	$R_i(a,l)$	$R_i(c,l)$	$R_i(g,l)$	$R_i(t,l)$
-10	36	7	21	28	0.622843	-1.739727	-0.154765	0.260273
-9	20	19	15	38	-0.225154	-0.299154	-0.640191	0.700846
-8	18	24	11	39	-0.377157	0.037881	-1.087650	0.738320
-7	3	0	85	4	-2.962119	-6.539159	1.862309	-2.547082
-6	24	19	29	20	0.037881	-0.299154	0.310899	-0.225154
-5	20	18	15	39	-0.225154	-0.377157	-0.640191	0.738320
-4	5	40	12	35	-2.225154	0.774846	-0.962119	0.582201
-3	73	2	15	2	1.642743	-3.547082	-0.640191	-3.547082
-2	53	9	10	20	1.180838	-1.377157	-1.225154	-0.225154
-1	40	8	11	33	0.774846	-1.547082	-1.087650	0.497312
0	42	4	4	42	0.845235	-2.547082	-2.547082	0.845235
1	33	11	8	40	0.497312	-1.087650	-1.547082	0.774846
2	20	10	9	53	-0.225154	-1.225154	-1.377157	1.180838
3	2	15	2	73	-3.547082	-0.640191	-3.547082	1.642743
4	35	12	40	5	0.582201	-0.962119	0.774846	-2.225154
5	39	15	18	20	0.738320	-0.640191	-0.377157	-0.225154
6	20	29	19	24	-0.225154	0.310899	-0.299154	0.037881
7	4	85	0	3	-2.547082	1.862309	-6.539159	-2.962119
8	39	11	24	18	0.738320	-1.087650	0.037881	-0.377157
9	38	15	19	20	0.700846	-0.640191	-0.299154	-0.225154
10	28	21	7	36	0.260273	-0.154765	-1.739727	0.622843

Ri(b,l) weight matrix

Individual Information Distributions

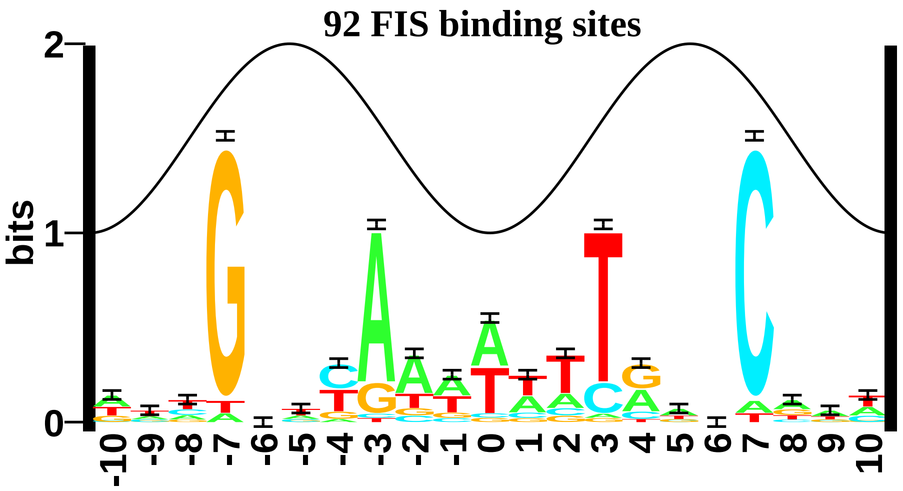
The average of the individual information values ... is the area under the sequence logo

Zero divides sites from non-sites



The consensus is an extremely abnormal, strong binding site with probability typically $< 10^{-5}$

Individual Information



The area under a sequence logo is the total sequence conservation.

The mathematics makes it look like an average.

QUESTION: What is it the average of?

ANSWER: The information of each individual binding site.

This is found from the individual information weight matrix as:

$$Ri(b,l) = 2 + \log_2 f(b,l)$$

where f(b,l) is the frequency of each base b
at every position l in the binding site.

number of
bases b
at each
position l

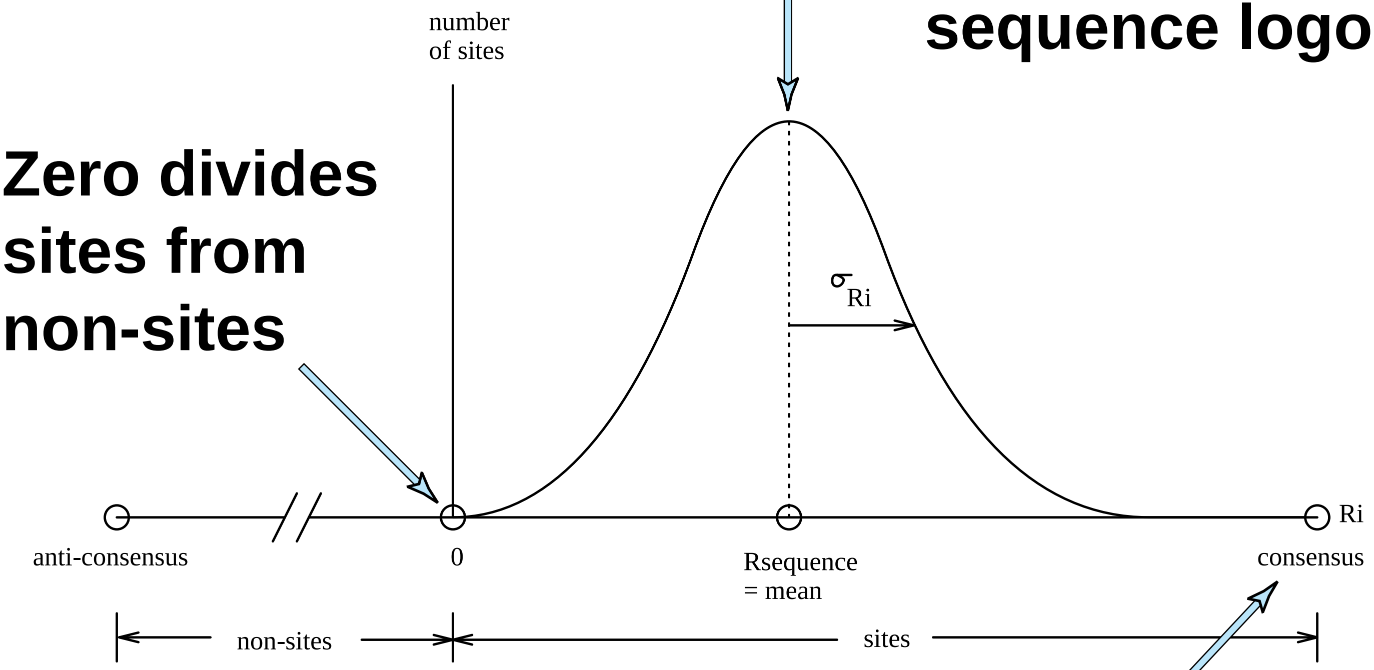
l	$n(a,l)$	$n(c,l)$	$n(g,l)$	$n(t,l)$	$R_i(a,l)$	$R_i(c,l)$	$R_i(g,l)$	$R_i(t,l)$
-10	36	7	21	28	0.622843	-1.739727	-0.154765	0.260273
-9	20	19	15	38	-0.225154	-0.299154	-0.640191	0.700846
-8	18	24	11	39	-0.377157	0.037881	-1.087650	0.738320
-7	3	0	85	4	-2.962119	-6.539159	1.862309	-2.547082
-6	24	19	29	20	0.037881	-0.299154	0.310899	-0.225154
-5	20	18	15	39	-0.225154	-0.377157	-0.640191	0.738320
-4	5	40	12	35	-2.225154	0.774846	-0.962119	0.582201
-3	73	2	15	2	1.642743	-3.547082	-0.640191	-3.547082
-2	53	9	10	20	1.180838	-1.377157	-1.225154	-0.225154
-1	40	8	11	33	0.774846	-1.547082	-1.087650	0.497312
0	42	4	4	42	0.845235	-2.547082	-2.547082	0.845235
1	33	11	8	40	0.497312	-1.087650	-1.547082	0.774846
2	20	10	9	53	-0.225154	-1.225154	-1.377157	1.180838
3	2	15	2	73	-3.547082	-0.640191	-3.547082	1.642743
4	35	12	40	5	0.582201	-0.962119	0.774846	-2.225154
5	39	15	18	20	0.738320	-0.640191	-0.377157	-0.225154
6	20	29	19	24	-0.225154	0.310899	-0.299154	0.037881
7	4	85	0	3	-2.547082	1.862309	-6.539159	-2.962119
8	39	11	24	18	0.738320	-1.087650	0.037881	-0.377157
9	38	15	19	20	0.700846	-0.640191	-0.299154	-0.225154
10	28	21	7	36	0.260273	-0.154765	-1.739727	0.622843

Ri(b,l) weight matrix

Individual Information Distributions

The average of the individual information values ... is the area under the sequence logo

Zero divides sites from non-sites



The consensus is an extremely abnormal, strong binding site with probability typically $\leq 10^{-5}$