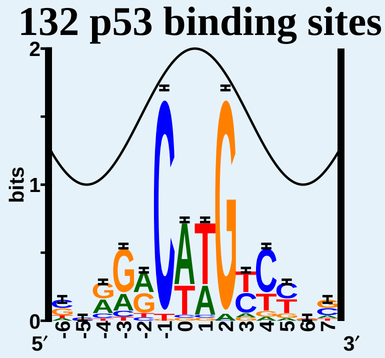
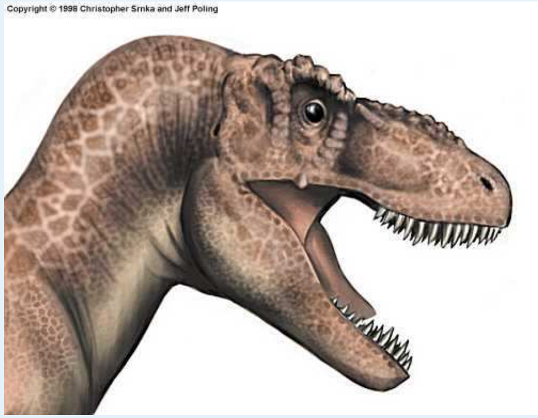
Evolution of Binding Sites

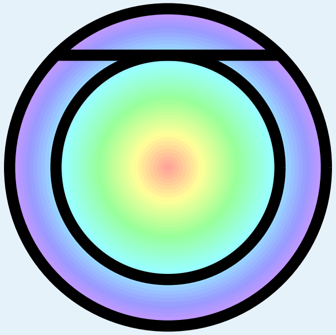
Thomas D. Schneider, Ph.D.

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

Molecular Information Theory Group

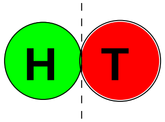
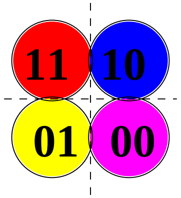
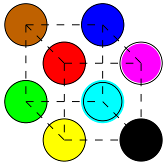
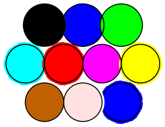



Copyright © 1998 Christopher Srnka and Jeff Poling

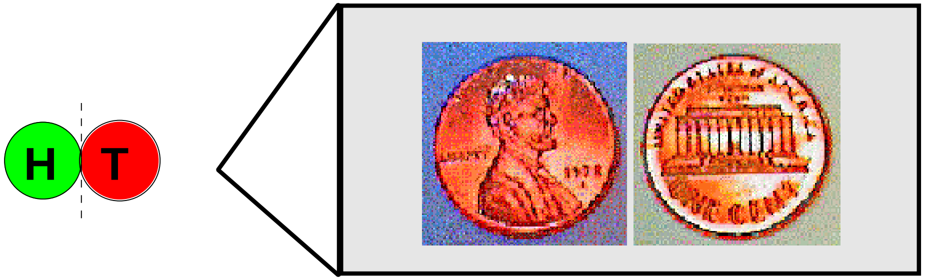
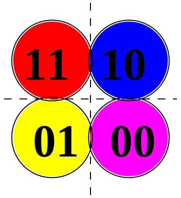
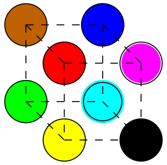
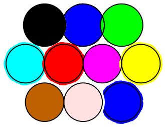


El Duomo, Florence, Italy

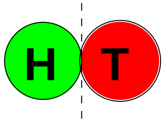
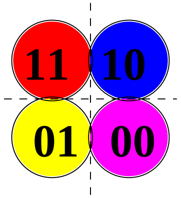
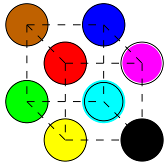
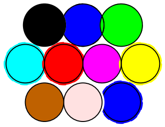
Information Theory: One-Minute Lesson

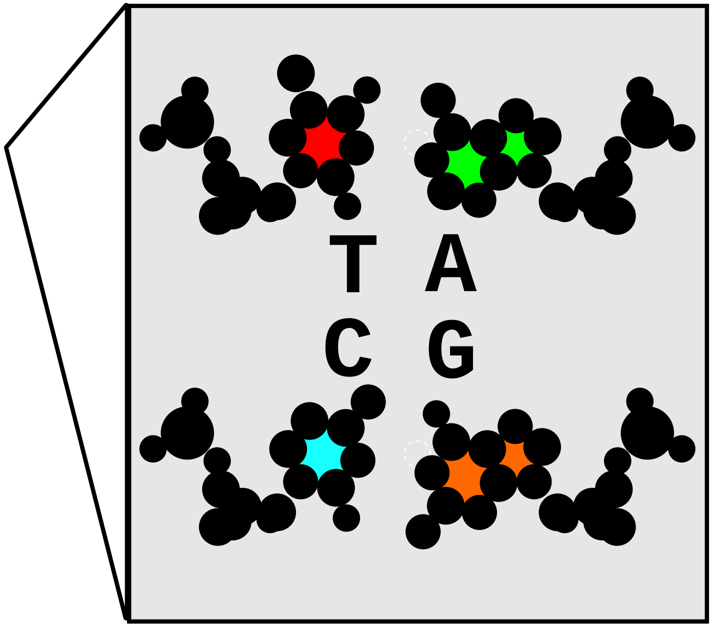
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

Information Theory: One-Minute Lesson

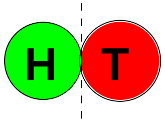
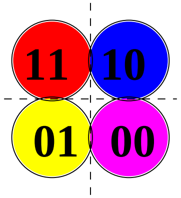
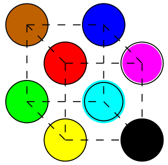
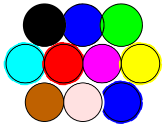
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

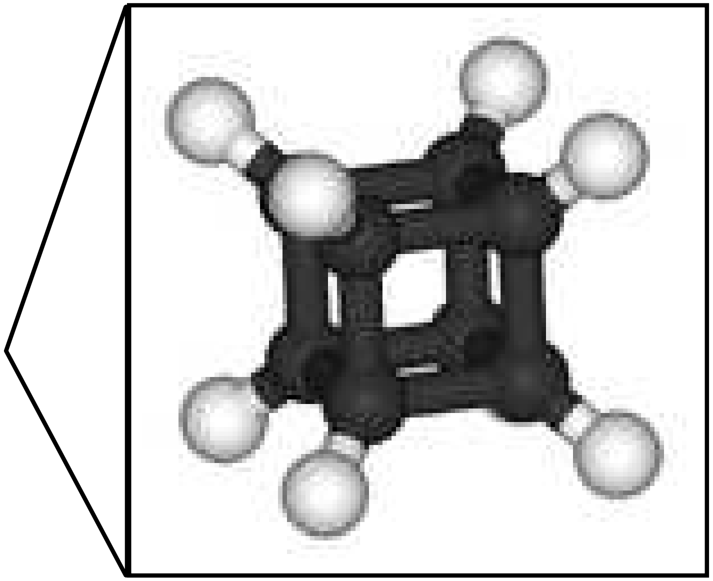
Information Theory: One-Minute Lesson

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

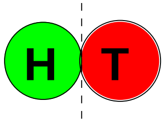
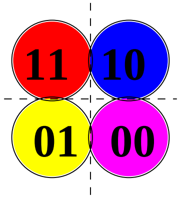
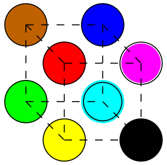
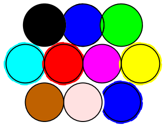


Information Theory: One-Minute Lesson

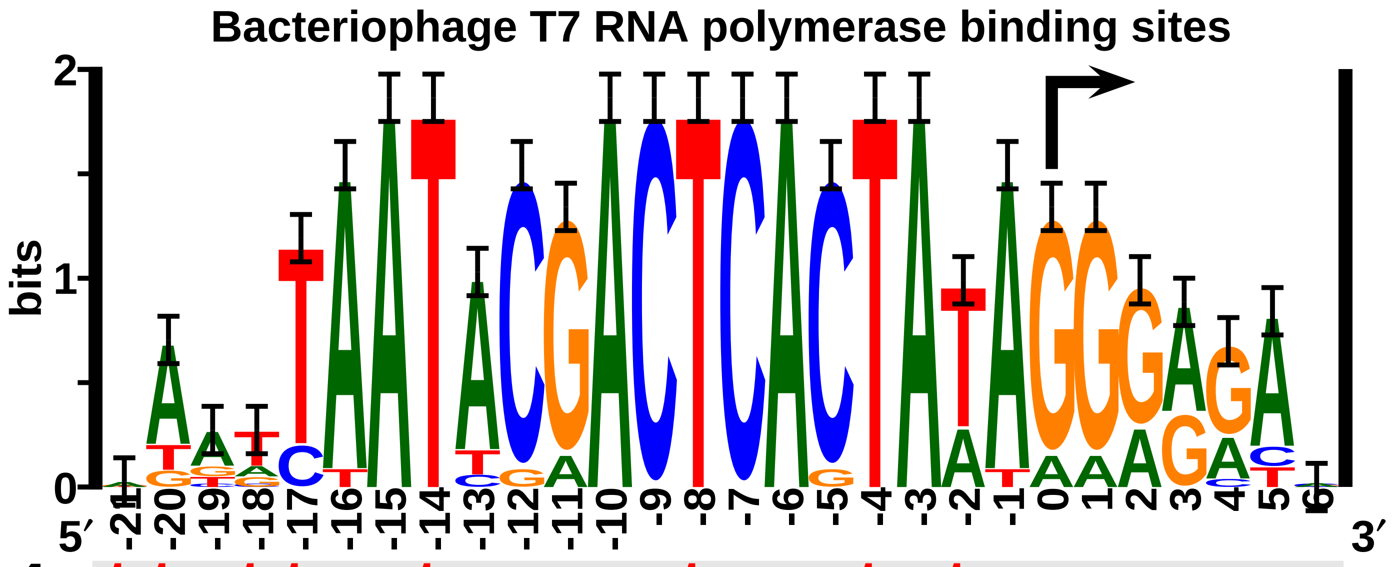
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	



Information Theory: One-Minute Lesson

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

Sequence Logo

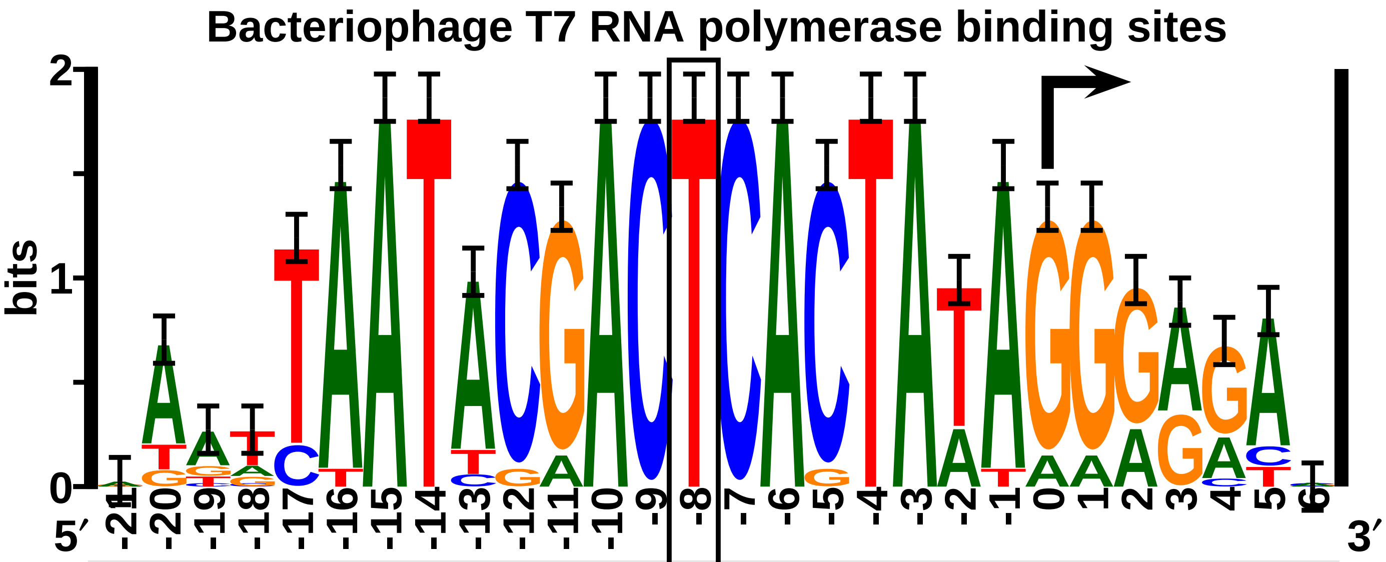


Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

```
1  ttattaatacaactcactataaggagag
2  aaatcaatacgactcactatagagggac
3  cggttaatacgactcactataggagaac
4  gaagtaatacgactcagtatagggacaa
5  taattaattgaactcactaaagggagac
6  cgcttaatacgactcactaaaggagaca
```

6 of 17 sites

Sequence Logo



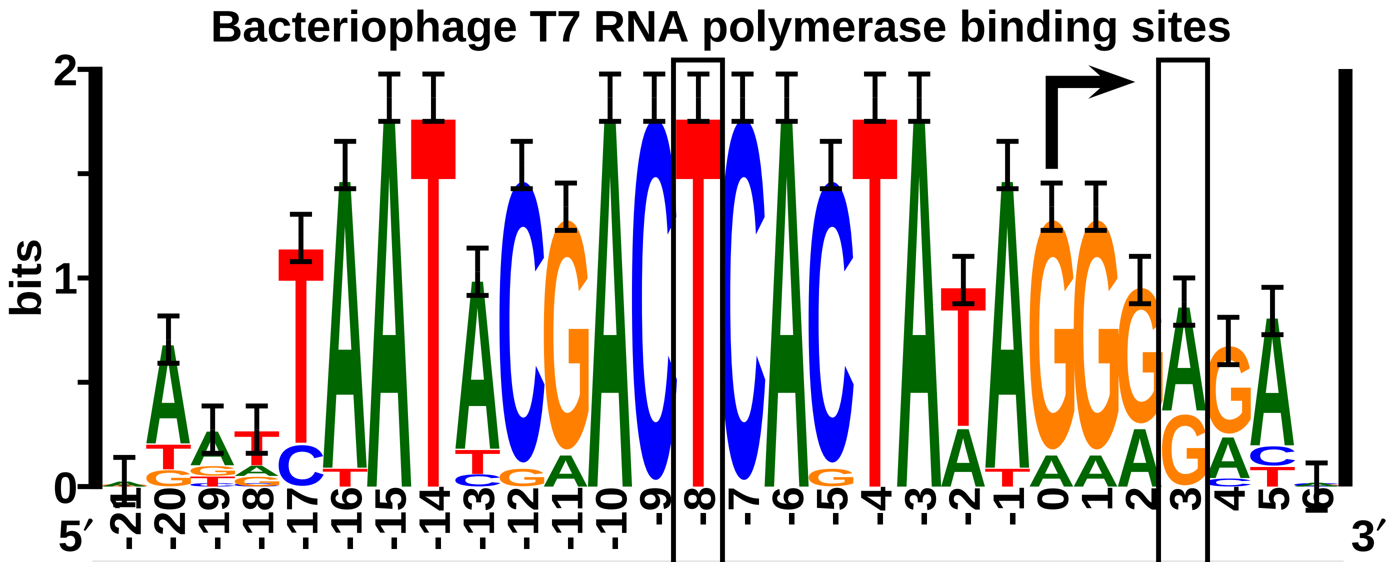
Bacteriophage T7 RNA polymerase binding sites

Schneider & Stephens *Nucl. Acids Res.* **18**: 6097-6100 1990

```
1  ttattaatacaactcactataaggagag
2  aaatcaatacgactcactatagagggac
3  cggttaatacgactcactataggagaac
4  gaagtaatacgactcagtatagggacaa
5  taattaattgaactcactaaagggagac
6  cgcttaatacgactcactaaaggagaca
```

6 of 17 sites

Sequence Logo



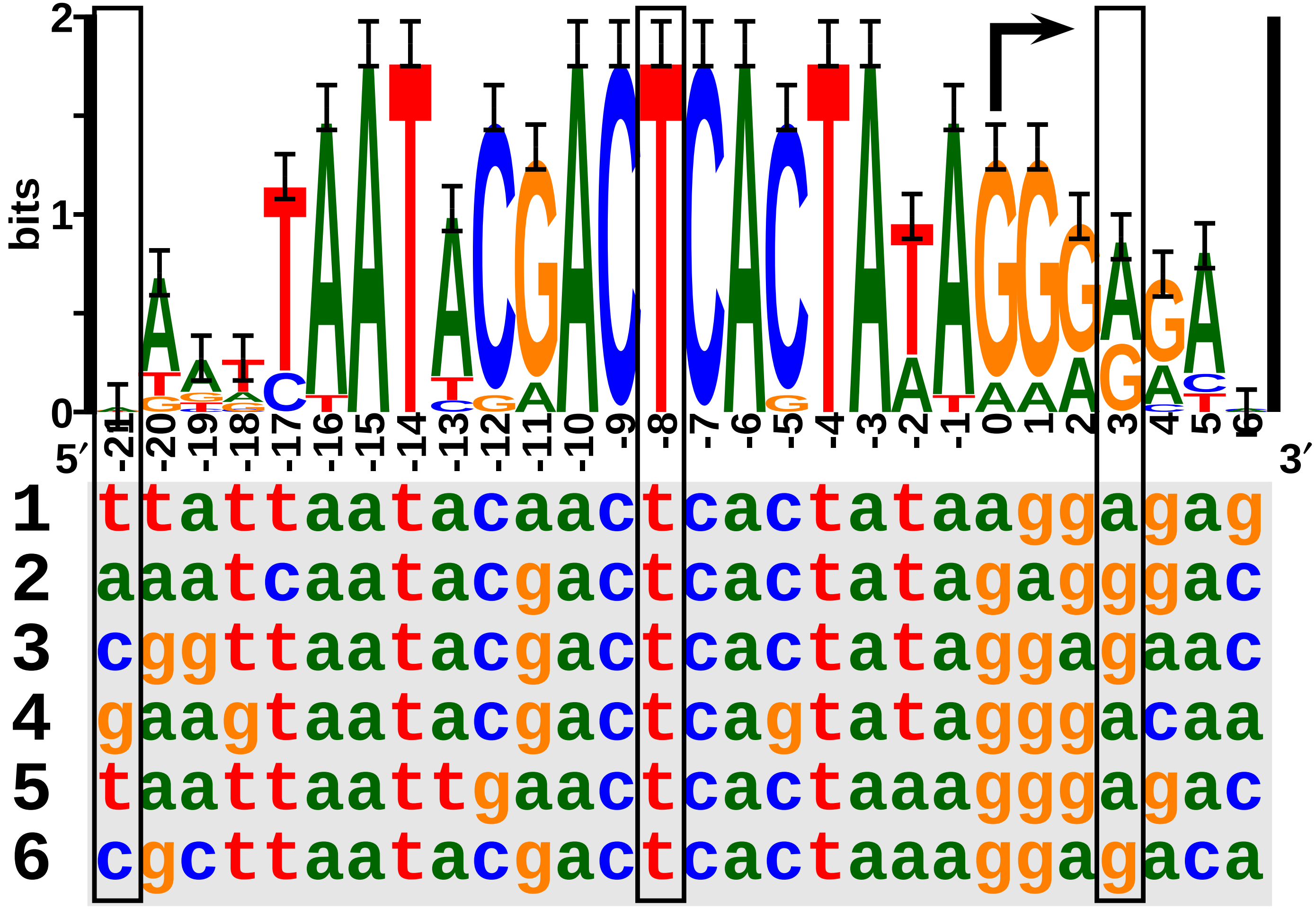
Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

```
1  ttattaatacaactcactataaggagag
2  aaatcaatacgactcactatagagggac
3  cggttaatacgactcactataggagaac
4  gaagtaatacgactcagtatagggacaa
5  taattaattgaactcactaaagggagac
6  cgcttaatacgactcactaaaggagaca
```

6 of 17 sites

Sequence Logo

Bacteriophage T7 RNA polymerase binding sites



1 ttattaatacaactcactataaggagag

2 aaatcaatacgactcactatagagggac

3 cggttaatacgactcactataggagaac

4 gaagtaatacgactcagtatagggacaa

5 taattaattgaactcactaaagggagac

6 cgcttaatacgactcactaaaggagaca

6 of 17 sites

Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

More Information Theory - 1

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \tag{1}$$

More Information Theory - 1

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\begin{aligned} \log_2 M & \\ &= -\log_2 1/M. \end{aligned} \tag{1}$$

$1/M$ is like the probability of a symbol.

More Information Theory - 1

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \tag{1}$$

$$= \quad -\log_2 1/M.$$

$1/M$ is like the probability of a symbol.

If the probabilities P_i of different symbols, i, are not equal, then the **surprisal** is:

$$u_i \equiv -\log_2 P_i. \tag{2}$$

how surprised one is to see a symbol

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$
\begin{aligned}
P_{\text{ring}} &= 1/1024 \qquad (3) \\
P_{\text{silent}} &= 1023/1024 \qquad (4)
\end{aligned}
$$

More Information Theory - 2

$\boxed{\text{EXAMPLE}}$

A phone rings once every 1024 seconds.

$$\begin{aligned} P_{\text{ring}} &= 1/1024 \qquad (3) \\ P_{\text{silent}} &= 1023/1024 \qquad (4) \end{aligned}$$

Surprisal:

$$\begin{aligned} \text{surprisal}_{\text{ring}} &= -\log_2(1/1024) = 10 \;\; \text{bits} \qquad (5) \\ \text{surprisal}_{\text{silent}} &= -\log_2(1023/1024) \approx 0 \;\; \text{bits} \qquad (6) \end{aligned}$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{3}$$

$$P_{\text{silent}} = 1023/1024 \tag{4}$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \text{ bits} \tag{5}$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \text{ bits} \tag{6}$$

The **average surprisal** is called the **uncertainty**, H:

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}}$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{3}$$

$$P_{\text{silent}} = 1023/1024 \tag{4}$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \text{ bits} \tag{5}$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \text{ bits} \tag{6}$$

The **average surprisal** is called the **uncertainty**, H:

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}} + P_{\text{silent}} \times \text{surprisal}_{\text{silent}} \tag{7}$$

More Information Theory - 2

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{3}$$

$$P_{\text{silent}} = 1023/1024 \tag{4}$$

Surprisal:

$$\text{surprisal}_{\text{ring}} = -\log_2(1/1024) = 10 \text{ bits} \tag{5}$$

$$\text{surprisal}_{\text{silent}} = -\log_2(1023/1024) \approx 0 \text{ bits} \tag{6}$$

The **average surprisal** is called the **uncertainty**, H:

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}} + P_{\text{silent}} \times \text{surprisal}_{\text{silent}} \tag{7}$$

$$H = P_{\text{ring}} \times \Big(-\log_2(P_{\text{ring}})\Big) + P_{\text{silent}} \times \Big(-\log_2(P_{\text{silent}})\Big) \tag{8}$$

More Information Theory - 3

For M symbols use the sum ($\sum$) notation:

$$H = \sum_{i=1}^{M} P_i \times (\text{surprisal for}P_i) \tag{9}$$

More Information Theory - 3

For M symbols use the sum ($\sum$) notation:

$$\begin{aligned} H &= \sum_{i=1}^{M} P_i \times (\text{surprisal for}P_i) \qquad (9) \\ &= \sum_{i=1}^{M} P_i \times (-\log_2 P_i) \qquad (10) \end{aligned}$$

More Information Theory - 3

For M symbols use the sum ($\sum$) notation:

$$
\begin{aligned}
H &= \sum_{i=1}^{M} P_i \times (\text{surprisal for} P_i) && (9)\\
&= \sum_{i=1}^{M} P_i \times (-\log_2 P_i) && (10)\\
&= -\sum_{i=1}^{M} P_i \log_2 P_i \quad \text{bits per symbol} && (11)
\end{aligned}
$$

More Information Theory - 4

$\boxed{\text{Information is a decrease in uncertainty}}$

$$R = H_{\text{before}} - H_{\text{after}} \tag{12}$$

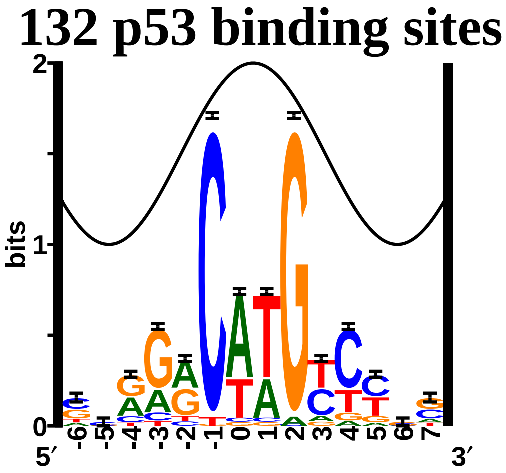
More Information Theory - 4

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \tag{12}$$

Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \tag{13}$$



More Information Theory - 4

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \tag{12}$$

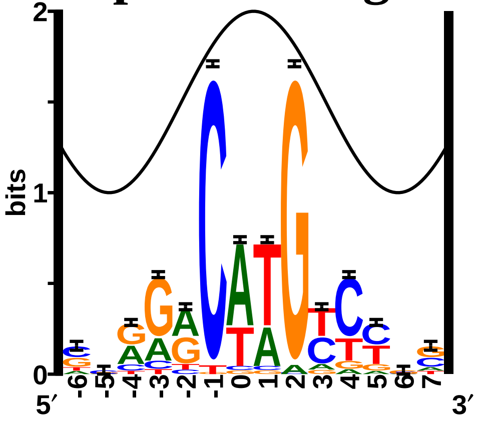
Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \tag{13}$$

and

$$\begin{aligned} H_{\text{after}} &= \text{uncertainty of bases} \\ &= -\sum_{base=A}^{T} P_{base} \log_2 P_{base} \end{aligned} \tag{14}$$

132 p53 binding sites



More Information Theory - 4

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \tag{12}$$

Example a sequence logo is computed from equiprobable bases before:

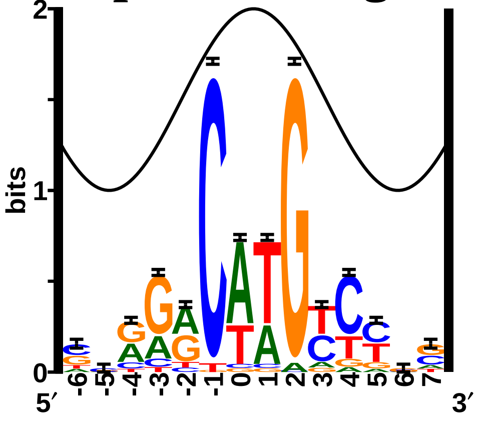
$$H_{\text{before}} = 2 \text{ bits/base} \tag{13}$$

and

$$\begin{aligned} H_{\text{after}} &= \text{uncertainty of bases} \\ &= -\sum_{base=A}^{T} P_{base} \log_2 P_{base} \end{aligned} \tag{14}$$

Note: with only one base, $H_{\text{after}} = 0$ so $R = 2$ bits/base.

132 p53 binding sites



Information required to find a set of binding sites

$G =$ # of potential binding sites

Information required to find a set of binding sites

$$\begin{aligned} G &= \#\ \text{of potential binding sites} \\ &= \text{genome size in some cases} \end{aligned}$$

Information required to find a set of binding sites

$$\begin{aligned} G &= \text{\# of potential binding sites} \\ &= \text{genome size in some cases} \end{aligned}$$

$\gamma =$ number of binding sites on genome

Information required to find a set of binding sites

$$\begin{aligned} G &= \text{ \# of potential binding sites} \\ &= \text{ genome size in some cases} \end{aligned}$$

$\gamma =$ number of binding sites on genome

$$R_{frequency} \ = \ H_{before} - H_{after}$$

Information required to find a set of binding sites

$$\begin{aligned} G = & \ \text{\# of potential binding sites} \\ = & \ \text{genome size in some cases} \end{aligned}$$

γ = number of binding sites on genome

$$\begin{aligned} R_{frequency} &= H_{before} - H_{after} \\ &= \log_2 G - \log_2 \gamma \end{aligned}$$

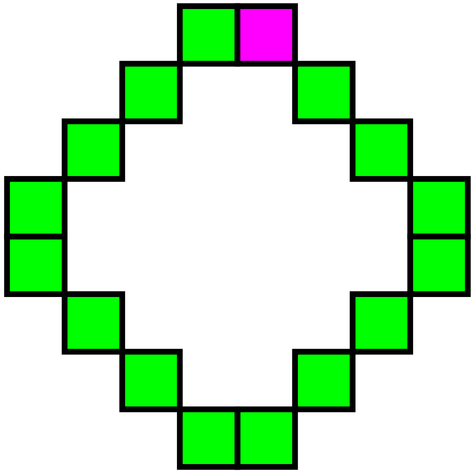
Information required to find a set of binding sites

$$\begin{aligned} G = &\ \text{\# of potential binding sites} \\ = &\ \text{genome size in some cases} \end{aligned}$$

$\gamma =$ number of binding sites on genome

$$\begin{aligned} R_{frequency} &= H_{before} - H_{after} \\ &= \log_2 G - \log_2 \gamma \\ &= -\log_2 \gamma / G \end{aligned}$$

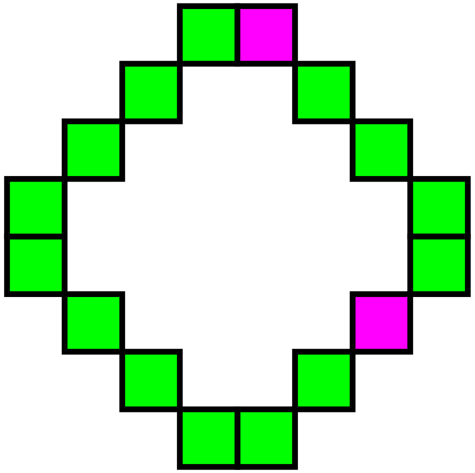
Information required to find a set of binding sites in a genome



16 positions

1 site

$\log_2 16/1 = 4$ bits

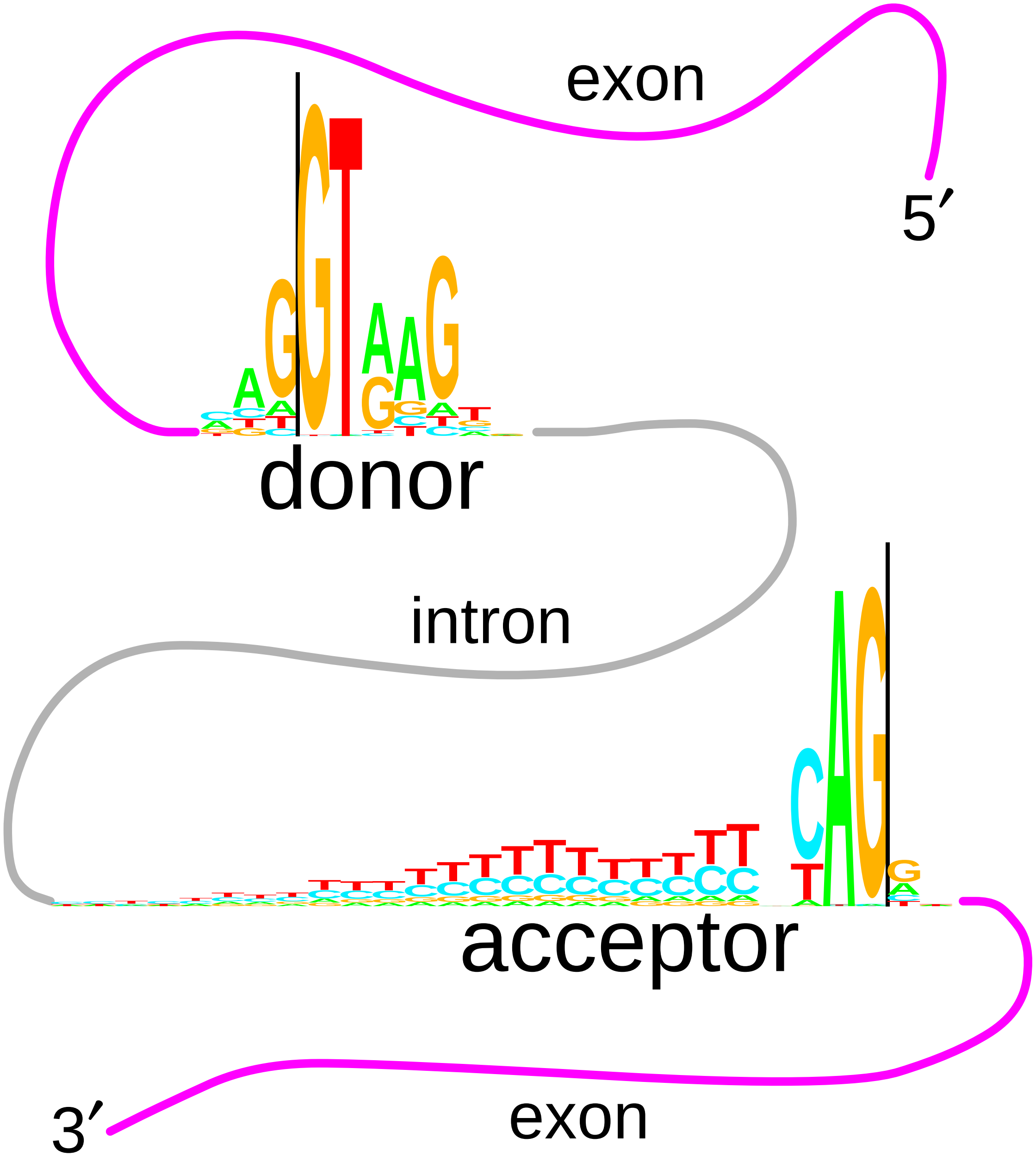


16 positions

2 sites

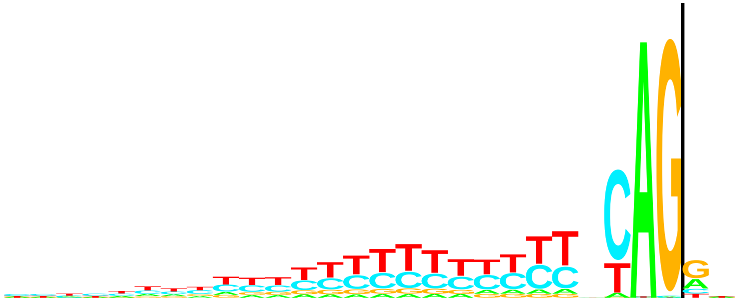
$\log_2 16/2 = 3$ bits

Donor and acceptor logos



Rsequence and Rfrequency for Splice Acceptors

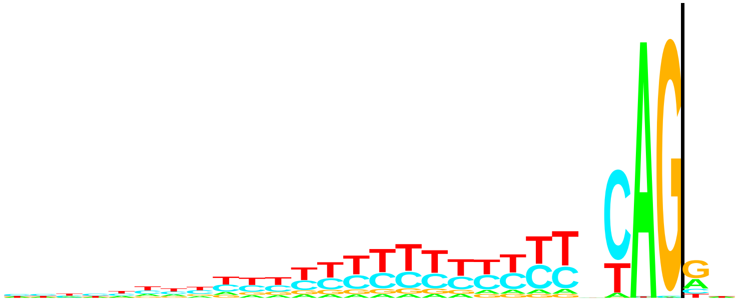
$\boxed{R_{sequence}}$



- **Information at binding site sequences (area under sequence logo)**
- **from: binding site sequences**
- **9.4 bits per site**

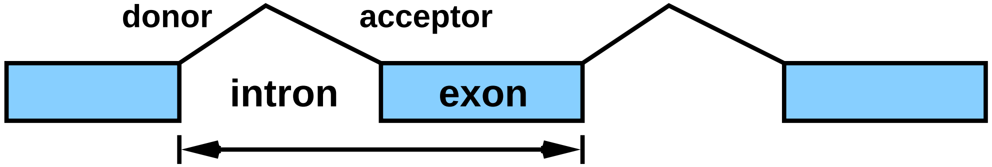
Rsequence and Rfrequency for Splice Acceptors

$\boxed{R_{sequence}}$



- Information at binding site sequences (area under sequence logo)
- from: binding site sequences
- 9.4 bits per site

$\boxed{R_{frequency}}$



- Information needed to locate the sites
- from: size of genome and number of sites (length of intron+exon)
- 9.7 bits per site

$$\boxed{R_{frequency}/R_{sequence} = 0.97}$$

Rsequence = Rfrequency Hypothesis

Hypothesis:

The information in
binding site patterns
is just sufficient
for the sites to be found
in the genome

Rsequence versus Rfrequency

Binding Site Recognizer[1]	Total Pattern Information $= \mathbf{R_{sequence}}$ (bits)	Information needed to Locate Site in Genome $= \mathbf{R_{frequency}}$ (bits)	$\frac{\textbf{Pattern Info}}{\textbf{Location Info}} = \frac{\mathbf{R_{sequence}}}{\mathbf{R_{frequency}}}$
Spliceosome acceptor[2]	$\mathbf{9.35 \pm 0.12}$	$\mathbf{9.66}$	$\mathbf{0.97 \pm 0.01}$
Spliceosome donor	$\mathbf{7.92 \pm 0.09}$	$\mathbf{9.66}$	$\mathbf{0.82 \pm 0.01}$
Ribosome	$\mathbf{11.0}$	$\mathbf{10.6}$	$\mathbf{1.0}$
λ cI/cro	$\mathbf{17.7 \pm 1.6}$	$\mathbf{19.3}$	$\mathbf{0.9 \pm 0.1}$
LexA	$\mathbf{21.5 \pm 1.7}$	$\mathbf{18.4}$	$\mathbf{1.2 \pm 0.1}$
TrpR	$\mathbf{23.4 \pm 1.9}$	$\mathbf{20.3}$	$\mathbf{1.2 \pm 0.1}$
LacI	$\mathbf{19.2 \pm 2.8}$	$\mathbf{21.9}$	$\mathbf{0.9 \pm 0.1}$
ArgR	$\mathbf{16.4}$	$\mathbf{18.4}$	$\mathbf{0.9}$
O (λ Origin)	$\mathbf{20.9}$	$\mathbf{19.9}$	$\mathbf{1.0}$
Ara C	$\mathbf{19.3}$	$\mathbf{19.3}$	$\mathbf{1.0}$
Transcription at TATA[3]	$\mathbf{3.3}$	$\sim \mathbf{3}$	$\sim \mathbf{1}$
T7 Promoter	$\mathbf{35.4}$	$\mathbf{16.5}$	$\mathbf{2.1}$

[1] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. J. Mol. Biol., 188:415-431, 1986.

[2] R. M. Stephens and T. D. Schneider. J. Mol. Biol., 228:1124-1136, 1992.

[3] F. E. Penotti. J Mol Biol, 213:37-52, 1990.

$R_{sequence}$ versus $R_{frequency}$ - meaning

The information in the binding site pattern ($R_{sequence}$)
is close to
The information needed to find the binding sites ($R_{frequency}$)

$R_{sequence}$ versus $R_{frequency}$ - meaning

The information in the binding site pattern ($R_{sequence}$)
is close to
The information needed to find the binding sites ($R_{frequency}$)

But for a species in a stable environment:

- size of genome (G) is fixed (*e. g. E. coli* has 4.7×10^6 bp)
- number of binding sites (γ) is fixed (*e. g.* there are ∼50 *E. coli* LexA sites)

so $R_{frequency} = \log_2 G/\gamma$ is fixed

$R_{sequence}$ **versus** $R_{frequency}$ **- meaning**

The information in the binding site pattern ($R_{sequence}$)
is close to
The information needed to find the binding sites ($R_{frequency}$)

But for a species in a stable environment:

- size of genome (G) is fixed (*e. g. E. coli* has 4.7×10^6 bp)
- number of binding sites (γ) is fixed (*e. g.* there are ~50 *E. coli* LexA sites)

so $R_{frequency} = \log_2 G/\gamma$ is fixed

Rsequence must evolve towards Rfrequency!

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Setup a Computer Model, ‘Ev’:
A population of “creatures” with

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Setup a Computer Model, ‘Ev’:
A population of “creatures” with

- genomes containing 4 bases (A, C, G, T)

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Setup a Computer Model, ‘Ev’:
A population of “creatures” with

- genomes containing 4 bases (A, C, G, T)
- a defined genome size (G)

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Setup a Computer Model, ‘Ev’:
A population of “creatures” with

- genomes containing 4 bases (A, C, G, T)
- a defined genome size (G)
- predetermined binding site locations (γ)
(to fix the frequency of sites)

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Setup a Computer Model, ‘Ev’:
A population of “creatures” with

- genomes containing 4 bases (A, C, G, T)
- a defined genome size (G)
- predetermined binding site locations (γ) (to fix the frequency of sites)

$\Big\}$ $R_{frequency}$ is fixed

Evolution of Binding Sites

- $R_{frequency}$ is fixed relative to $R_{sequence}$
- Does $R_{sequence}$ evolve toward $R_{frequency}$?

Setup a Computer Model, ‘Ev’:
A population of “creatures” with

- genomes containing 4 bases (A, C, G, T)
- a defined genome size (G)
- predetermined binding site locations (γ) (to fix the frequency of sites)

$\Big\}$ $R_{frequency}$ is fixed

- a recognizer gene encoded in the sequence: use a weight matrix

How A Weight Matrix Works

Sequence matrix, $s(b, l, j)$ for sequence j

base b	position l									
	C	A	G	G	T	C	T	G	C	A
	−3	−2	−1	0	1	2	3	4	5	6
A	0	1	0	0	0	0	0	0	0	1
C	1	0	0	0	0	1	0	0	1	0
G	0	0	1	1	0	0	0	1	0	0
T	0	0	0	0	1	0	1	0	0	0

Individual information weight matrix, $R_{iw}(b, l)$

base b	position l									
	−3	−2	−1	0	1	2	3	4	5	6
A	+0.4	$\boxed{+1.3}$	−1.4	−8.8	−5.8	+1.1	+1.5	−1.8	−0.7	$\boxed{+0.0}$
C	$\boxed{+0.6}$	−0.8	−2.4	−7.8	−5.5	$\boxed{-3.7}$	−1.6	−2.2	$\boxed{-0.5}$	−0.2
G	−0.6	−1.0	$\boxed{+1.6}$	$\boxed{+2.0}$	−6.2	+0.7	−1.1	$\boxed{+1.7}$	−0.3	+0.4
T	−1.0	−0.9	−1.7	−5.8	$\boxed{+2.0}$	−3.4	$\boxed{-1.6}$	−2.2	+0.9	−0.5

How A Weight Matrix Works

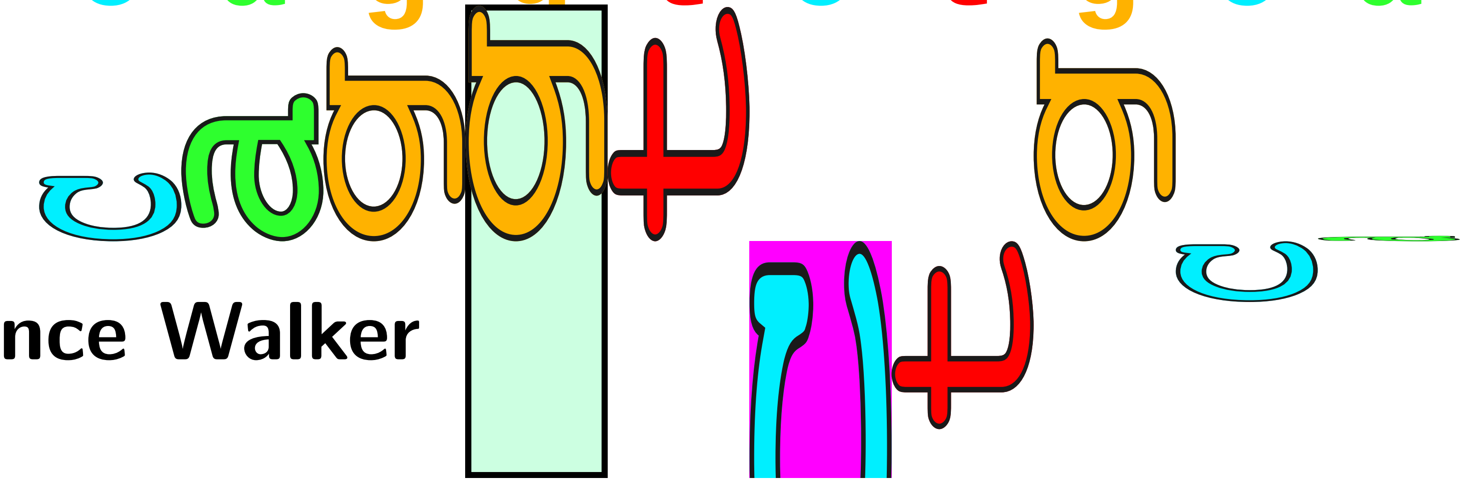
Sequence matrix, $s(b,l,j)$ for sequence j

base b	position l									
	C	A	G	G	T	C	T	G	C	A
	−3	−2	−1	0	1	2	3	4	5	6
A	0	1	0	0	0	0	0	0	0	1
C	1	0	0	0	0	1	0	0	1	0
G	0	0	1	1	0	0	0	1	0	0
T	0	0	0	0	1	0	1	0	0	0

Individual information weight matrix, $R_{iw}(b,l)$

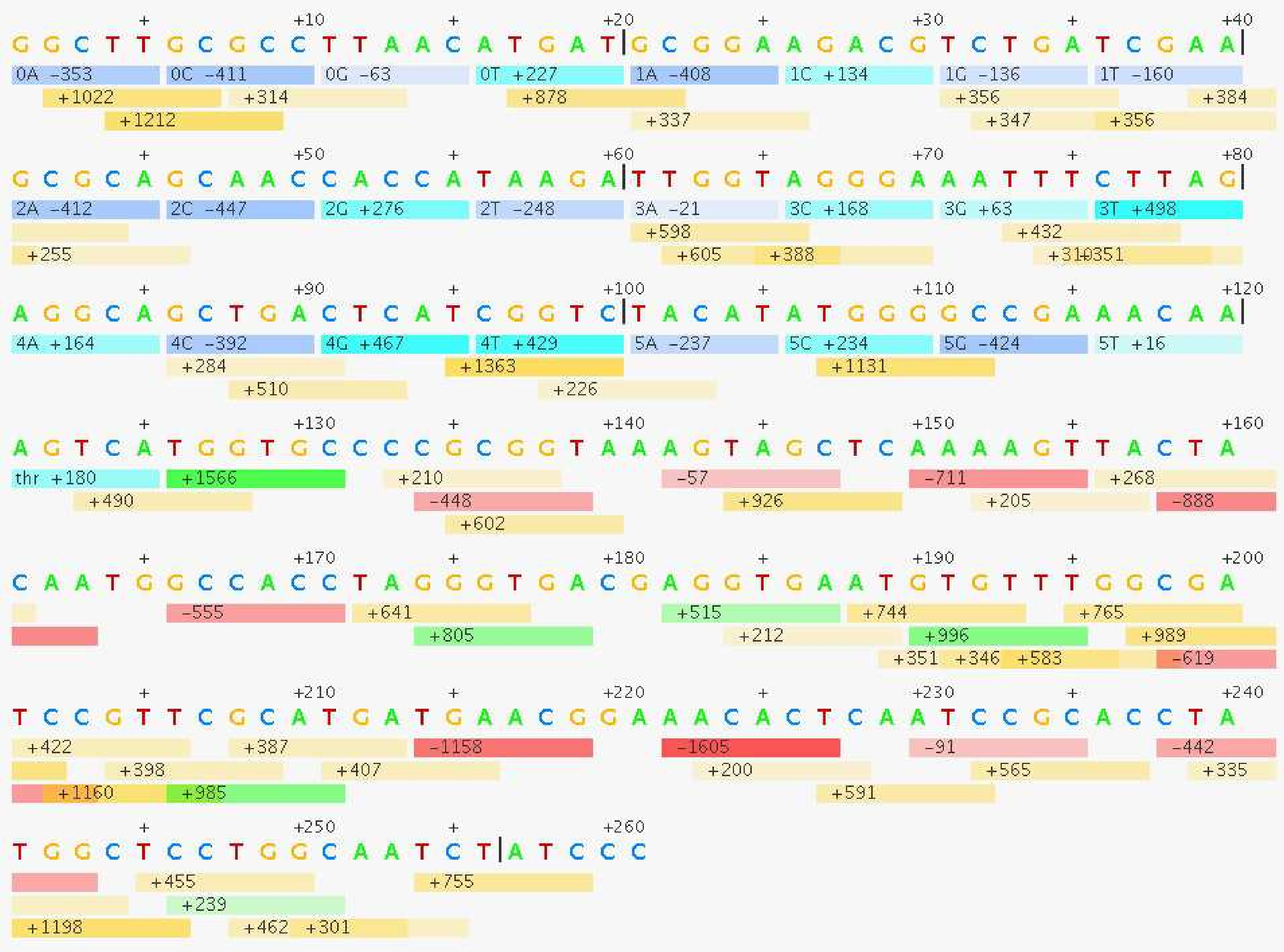
base b	position l									
	−3	−2	−1	0	1	2	3	4	5	6
A	+0.4	$\boxed{+1.3}$	−1.4	−8.8	−5.8	+1.1	+1.5	−1.8	−0.7	$\boxed{+0.0}$
C	$\boxed{+0.6}$	−0.8	−2.4	−7.8	−5.5	$\boxed{-3.7}$	−1.6	−2.2	$\boxed{-0.5}$	−0.2
G	−0.6	−1.0	$\boxed{+1.6}$	$\boxed{+2.0}$	−6.2	+0.7	−1.1	$\boxed{+1.7}$	−0.3	+0.4
T	−1.0	−0.9	−1.7	−5.8	$\boxed{+2.0}$	−3.4	$\boxed{-1.6}$	−2.2	+0.9	−0.5

5′ c a g g t c t g c a 3′

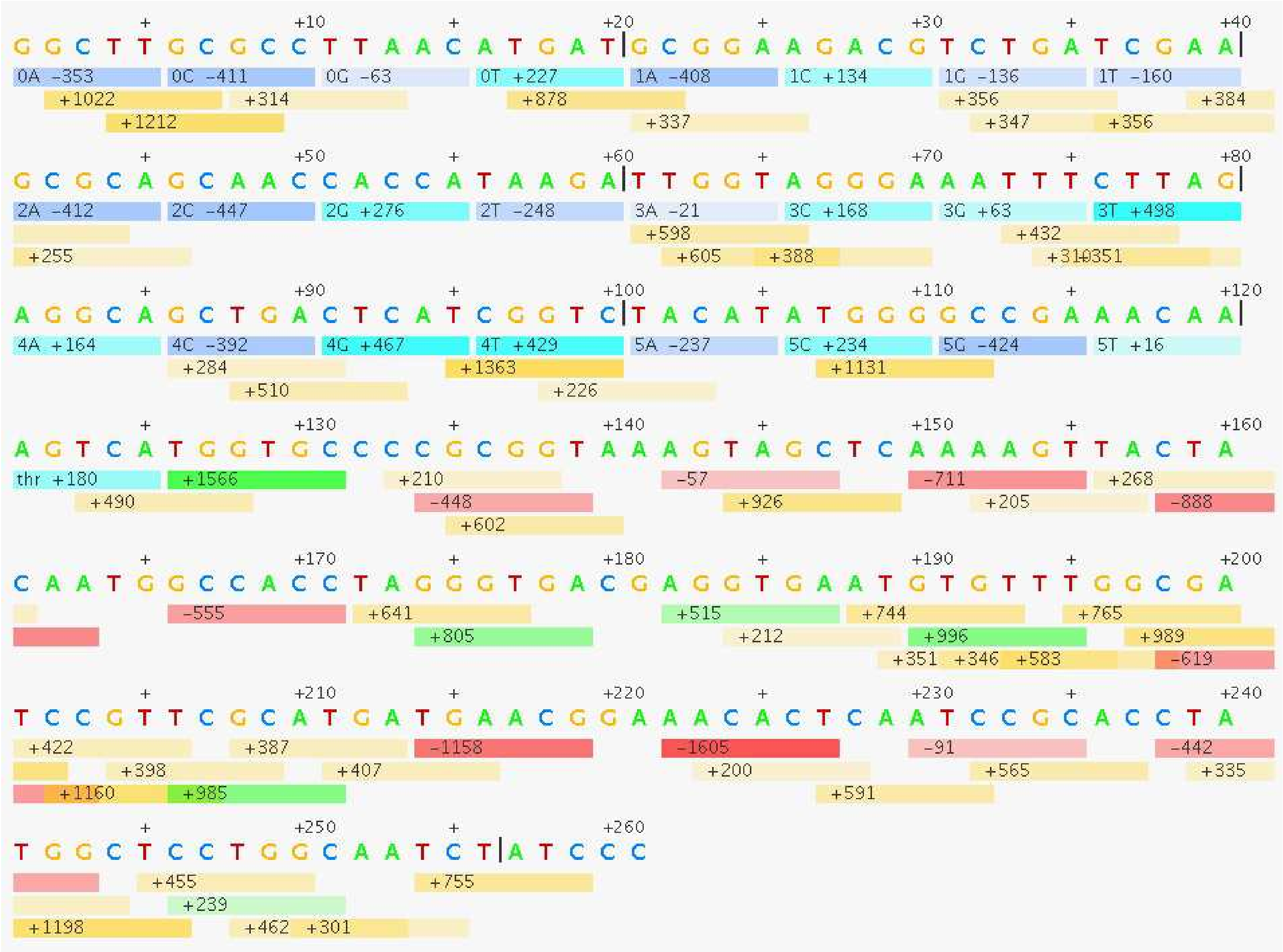


Sequence Walker

Unevolved Ev Creature

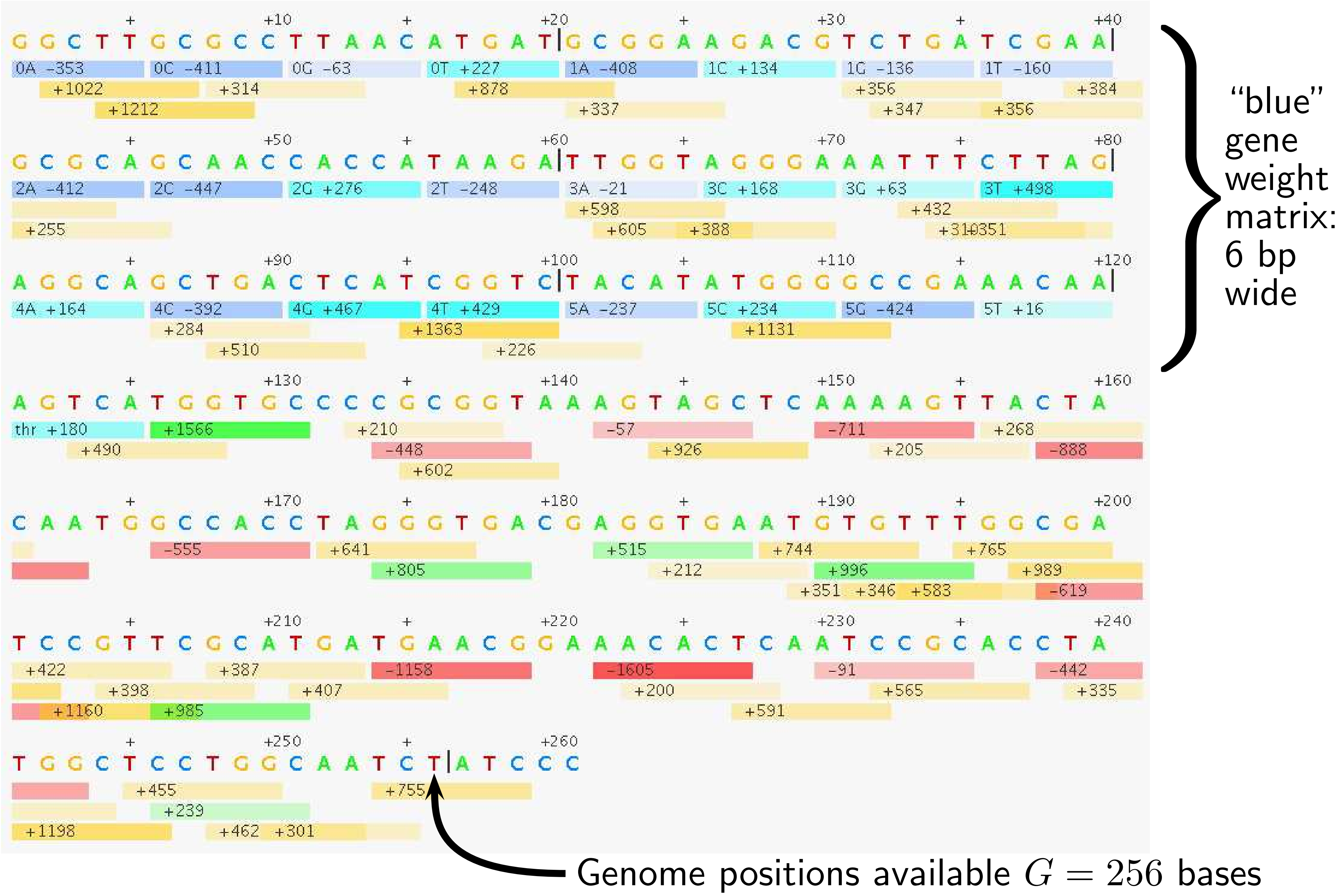

+ +10 + +20 + +30 + +40
G G C T T G C G C C T T A A C A T G A T G C G G A A G A C G T C T G A T C G A A
0A -353 0C -411 0G -63 0T +227 1A -408 1C +134 1G -136 1T -160
+1022 +314 +878 +356 +384
+1212 +337 +347 +356
+ +50 + +60 + +70 + +80
G C G C A G C A A C C A C C A T A A G A T T G G T A G G G A A A T T T C T T A G
2A -412 2C -447 2G +276 2T -248 3A -21 3C +168 3G +63 3T +498
+598 +432
+255 +605 +388 +310351
+ +90 + +100 + +110 + +120
A G G C A G C T G A C T C A T C G G T C T A C A T A T G G G G C C G A A A C A A
4A +164 4C -392 4G +467 4T +429 5A -237 5C +234 5G -424 5T +16
+284 +1363 +1131
+510 +226
+ +130 + +140 + +150 + +160
A G T C A T G G T G C C C C G C G G T A A A G T A G C T C A A A A G T T A C T A
thr +180 +1566 +210 -57 -711 +268
+490 -448 +926 +205 -888
+602
+ +170 + +180 + +190 + +200
C A A T G G C C A C C T A G G G T G A C G A G G T G A A T G T G T T T G G C G A
-555 +641 +515 +744 +765
+805 +212 +996 +989
+351 +346 +583 -619
+ +210 + +220 + +230 + +240
T C C G T T C G C A T G A T G A A C G G A A A C A C T C A A T C C G C A C C T A
+422 +387 -1158 -1605 -91 -442
+398 +407 +200 +565 +335
+1160 +985 +591
+ +250 + +260
T G G C T C C T G G C A A T C T A T C C C
+455 +755
+239
+1198 +462 +301

Unevolved Ev Creature



"blue" gene weight matrix: 6 bp wide

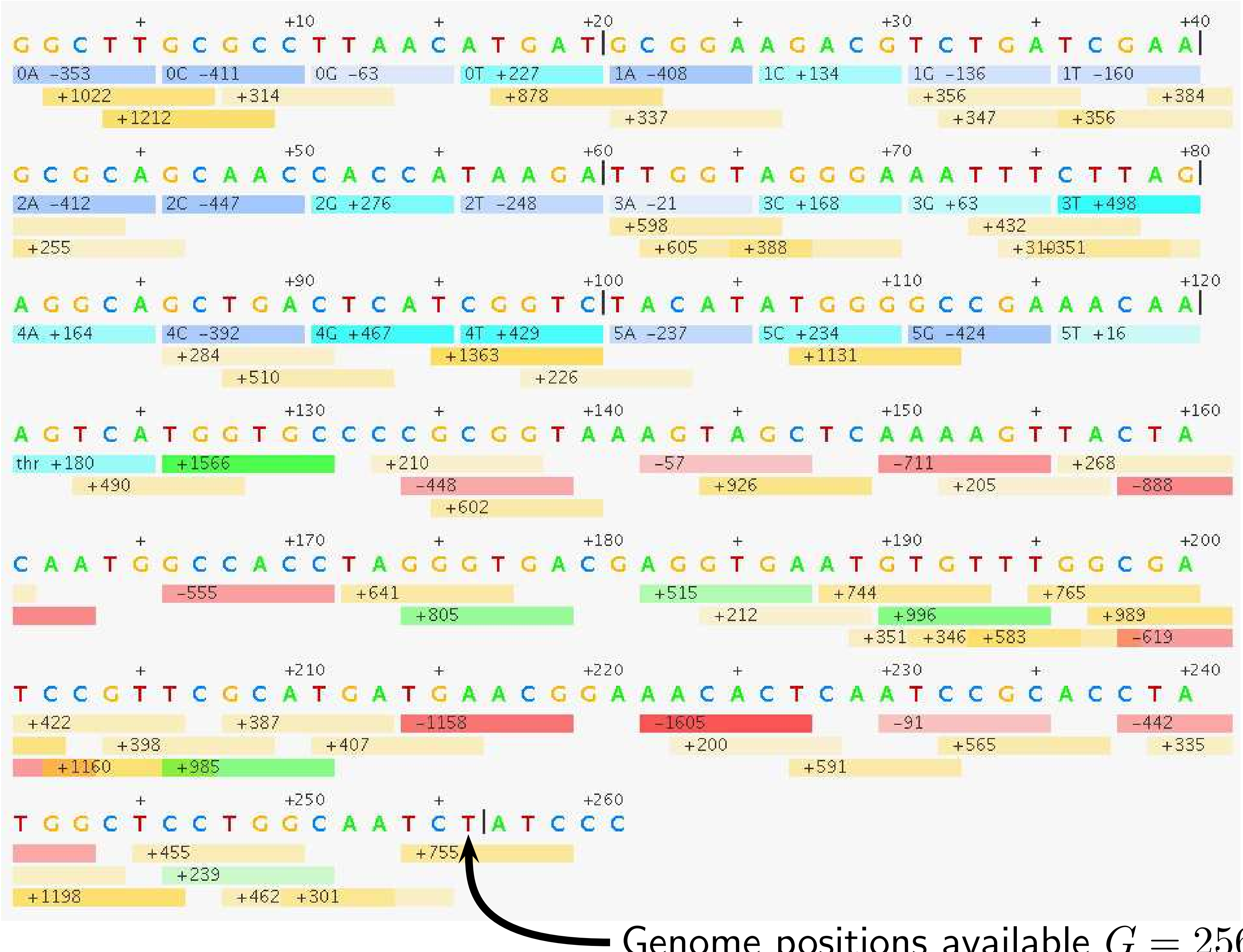
Unevolved Ev Creature


+ +10 + +20 + +30 + +40
G G C T T G C G C C T T A A C A T G A T G C G G A A G A C G T C T G A T C G A A
0A -353 0C -411 0G -63 0T +227 1A -408 1C +134 1G -136 1T -160
+1022 +314 +878 +356 +384
+1212 +337 +347 +356
+ +50 + +60 + +70 + +80
G C G C A G C A A C C A C C A T A A G A T T G G T A G G G A A A T T T C T T A G
2A -412 2C -447 2G +276 2T -248 3A -21 3C +168 3G +63 3T +498
+598 +432
+255 +605 +388 +310351
+ +90 + +100 + +110 + +120
A G G C A G C T G A C T C A T C G G T C T A C A T A T G G G G C C G A A A C A A
4A +164 4C -392 4G +467 4T +429 5A -237 5C +234 5G -424 5T +16
+284 +1363 +1131
+510 +226
+ +130 + +140 + +150 + +160
A G T C A T G G T G C C C C G C G G T A A A G T A G C T C A A A A G T T A C T A
thr +180 +1566 +210 -57 -711 +268
+490 -448 +926 +205 -888
+602
+ +170 + +180 + +190 + +200
C A A T G G C C A C C T A G G G T G A C G A G G T G A A T G T G T T T G G C G A
-555 +641 +515 +744 +765
+805 +212 +996 +989
+351 +346 +583 -619
+ +210 + +220 + +230 + +240
T C C G T T C G C A T G A T G A A C G G A A A C A C T C A A T C C G C A C C T A
+422 +387 -1158 -1605 -91 -442
+398 +407 +200 +565 +335
+1160 +985 +591
+ +250 + +260
T G G C T C C T G G C A A T C T A T C C C
+455 +755
+239
+1198 +462 +301

“blue” gene weight matrix: 6 bp wide

Genome positions available $G = 256$ bases

Unevolved Ev Creature


+ +10 + +20 + +30 + +40
G G C T T G C G C C T T A A C A T G A T | G C G G A A G A C G T C T G A T C G A A |
0A -353 0C -411 0G -63 0T +227 1A -408 1C +134 1G -136 1T -160
+1022 +314 +878 +356 +384
+1212 +337 +347 +356
+ +50 + +60 + +70 + +80
G C G C A G C A A C C A C C A T A A G A | T T G G T A G G G A A A T T T C T T A G |
2A -412 2C -447 2G +276 2T -248 3A -21 3C +168 3G +63 3T +498
+598 +432
+255 +605 +388 +310351
+ +90 + +100 + +110 + +120
A G G C A G C T G A C T C A T C G G T C | T A C A T A T G G G G C C G A A A C A A |
4A +164 4C -392 4G +467 4T +429 5A -237 5C +234 5G -424 5T +16
+284 +1363 +1131
+510 +226
+ +130 + +140 + +150 + +160
A G T C A T G G T G C C C C G C G G T A A A G T A G C T C A A A A G T T A C T A
thr +180 +1566 +210 -57 -711 +268
+490 -448 +926 +205 -888
+602
+ +170 + +180 + +190 + +200
C A A T G G C C A C C T A G G G T G A C G A G G T G A A T G T G T T T G G C G A
-555 +641 +515 +744 +765
+805 +212 +996 +989
+351 +346 +583 -619
+ +210 + +220 + +230 + +240
T C C G T T C G C A T G A T G A A C G G A A A C A C T C A A T C C G C A C C T A
+422 +387 -1158 -1605 -91 -442
+398 +407 +200 +565 +335
+1160 +985 +591
+ +250 + +260
T G G C T C C T G G C A A T C T | A T C C C
+455 +755
+239
+1198 +462 +301

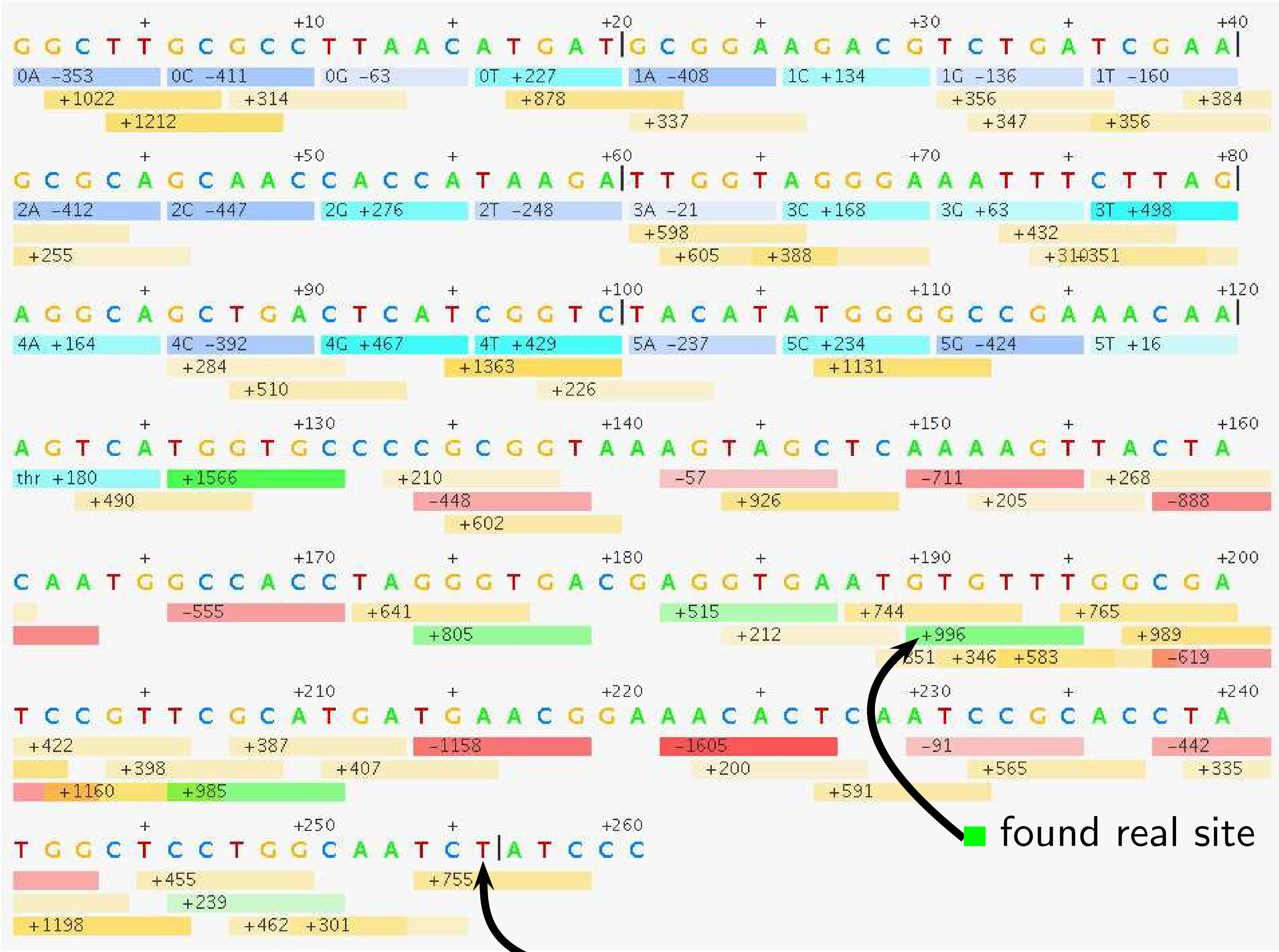
"blue" gene weight matrix: 6 bp wide

$\gamma = 16$ binding sites

Genome positions available $G = 256$ bases

$R_{frequency} = \log_2 256/16 = 4$ bits

Unevolved Ev Creature


+ +10 + +20 + +30 + +40
G G C T T G C G C C T T A A C A T G A T | G C G G A A G A C G T C T G A T C G A A |
0A -353 0C -411 0G -63 0T +227 1A -408 1C +134 1G -136 1T -160
+1022 +314 +878 +356 +384
+1212 +337 +347 +356
+ +50 + +60 + +70 + +80
G C G C A G C A A C C A C C A T A A G A | T T G G T A G G G A A A T T T C T T A G |
2A -412 2C -447 2G +276 2T -248 3A -21 3C +168 3G +63 3T +498
+598 +432
+255 +605 +388 +310351
+ +90 + +100 + +110 + +120
A G G C A G C T G A C T C A T C G G T C | T A C A T A T G G G G C C G A A A C A A |
4A +164 4C -392 4G +467 4T +429 5A -237 5C +234 5G -424 5T +16
+284 +1363 +1131
+510 +226
+ +130 + +140 + +150 + +160
A G T C A T G G T G C C C C G C G G T A A A G T A G C T C A A A A G T T A C T A
thr +180 +1566 +210 -57 -711 +268
+490 -448 +926 +205 -888
+602
+ +170 + +180 + +190 + +200
C A A T G G C C A C C T A G G G T G A C G A G G T G A A T G T G T T T G G C G A
-555 +641 +515 +744 +765
+805 +212 +996 +989
351 +346 +583 -619
+ +210 + +220 + +230 + +240
T C C G T T C G C A T G A T G A A C G G A A A C A C T C A A T C C G C A C C T A
+422 +387 -1158 -1605 -91 -442
+398 +407 +200 +565 +335
+1160 +985 +591
+ +250 + +260
T G G C T C C T G G C A A T C T | A T C C C
+455 +755
+239
+1198 +462 +301

"blue" gene weight matrix: 6 bp wide

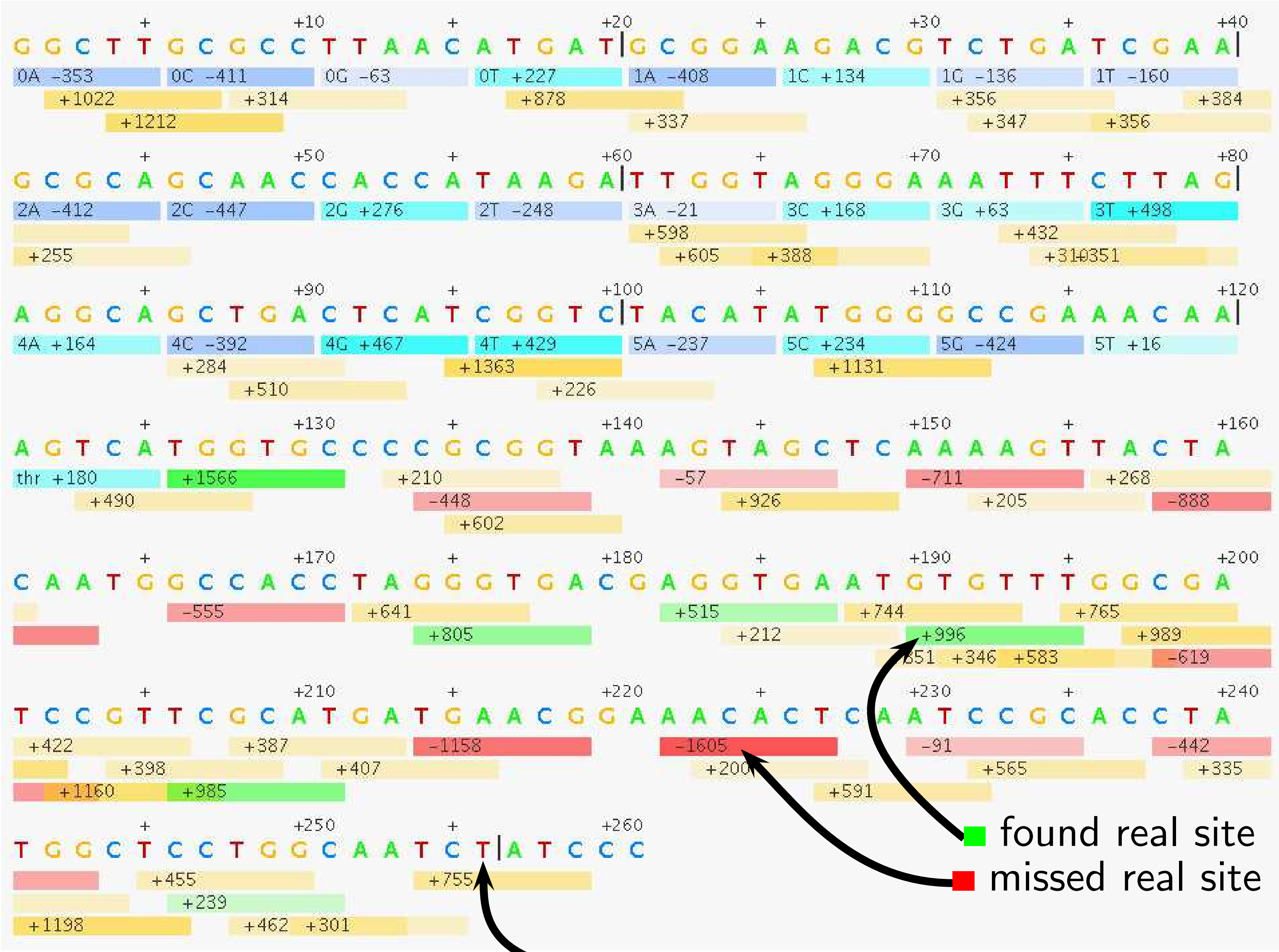
$\gamma = 16$ binding sites

🟩 found real site

Genome positions available $G = 256$ bases

$R_{frequency} = \log_2 256/16 = 4$ bits

Unevolved Ev Creature



```
         +         +10         +         +20         +         +30         +         +40
G G C T T G C G C C T T A A C A T G A T|G C G G A A G A C G T C T G A T C G A A|
0A -353   0C -411   0G -63   0T +227   1A -408   1C +134   1G -136   1T -160
  +1022     +314         +878                      +356                +384
     +1212                          +337              +347     +356

         +         +50         +         +60         +         +70         +         +80
G C G C A G C A A C C A C C A T A A G A|T T G G T A G G G A A A T T T C T T A G|
2A -412   2C -447   2G +276   2T -248   3A -21   3C +168   3G +63   3T +498
                                         +598                 +432
+255                                       +605  +388           +310351

         +         +90         +        +100         +        +110         +        +120
A G G C A G C T G A C T C A T C G G T C|T A C A T A T G G G G C C G A A A C A A|
4A +164   4C -392   4G +467   4T +429   5A -237   5C +234   5G -424   5T +16
          +284             +1363                 +1131
              +510               +226

         +        +130         +        +140         +        +150         +        +160
A G T C A T G G T G C C C C G C G G T A A A G T A G C T C A A A A G T T A C T A
thr +180  +1566        +210          -57             -711      +268
    +490             -448              +926             +205      -888
                       +602

         +        +170         +        +180         +        +190         +        +200
C A A T G G C C A C C T A G G G T G A C G A G G T G A A T G T G T T T G G C G A
          -555      +641             +515     +744       +765
                       +805              +212     +996      +989
                                             351  +346  +583  -619

         +        +210         +        +220         +        +230         +        +240
T C C G T T C G C A T G A T G A A C G G A A A C A C T C A A T C C G C A C C T A
+422          +387     -1158          -1605             -91       -442
     +398          +407                +200              +565      +335
  +1160   +985                                   +591

         +        +250         +        +260
T G G C T C C T G G C A A T C T|A T C C C
        +455                +755
          +239
+1198          +462  +301
```

“blue” gene weight matrix: 6 bp wide

$\gamma = 16$ binding sites

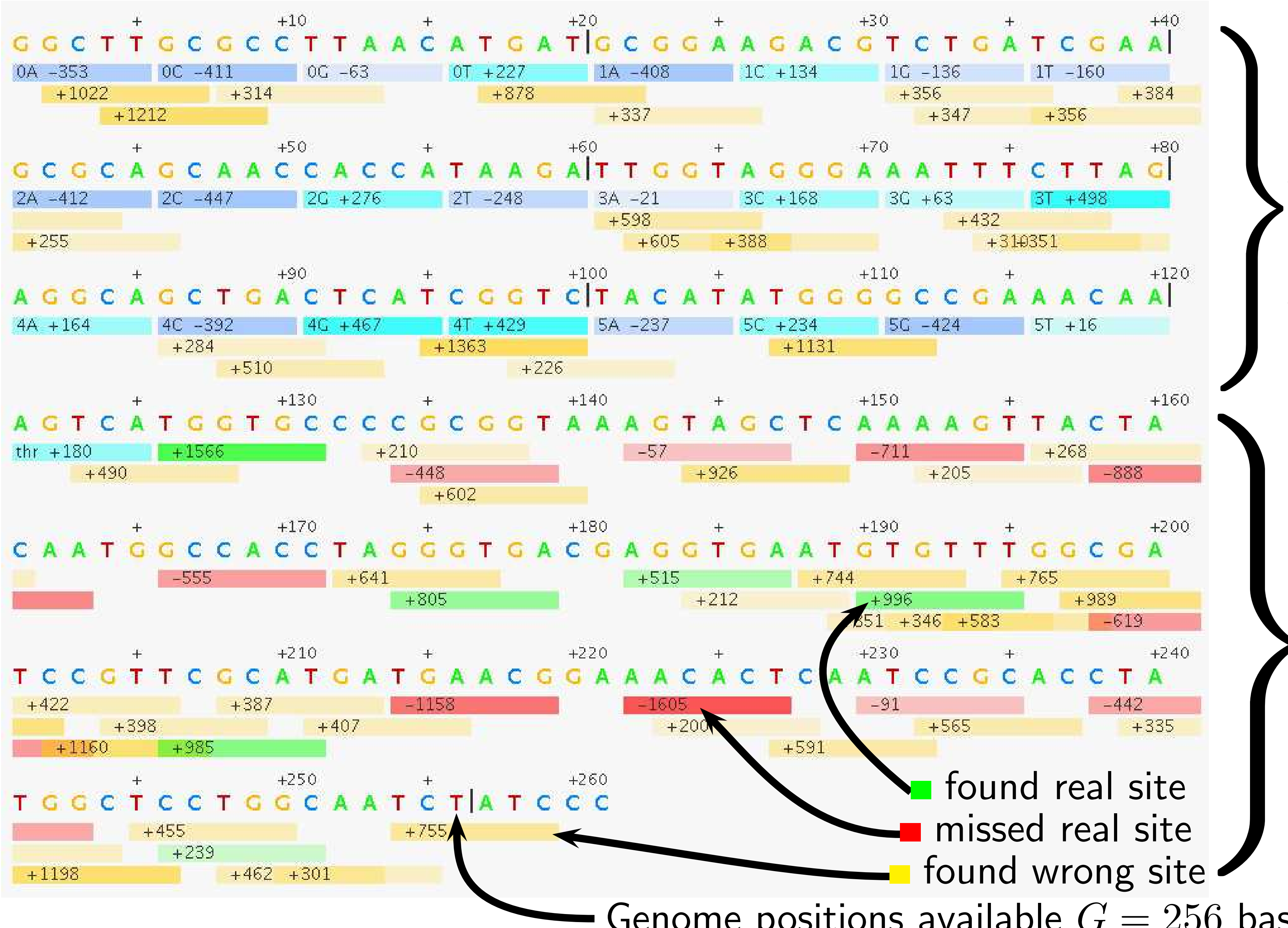
🟩 found real site

🟥 missed real site

Genome positions available $G = 256$ bases

$R_{frequency} = \log_2 256/16 = 4$ bits

Unevolved Ev Creature



+10 +20 +30 +40
G G C T T G C G C C T T A A C A T G A T G C G G A A G A C G T C T G A T C G A A
0A -353 0C -411 0G -63 0T +227 1A -408 1C +134 1G -136 1T -160
+1022 +314 +878 +356 +384
+1212 +337 +347 +356
+50 +60 +70 +80
G C G C A G C A A C C A C C A T A A G A T T G G T A G G G A A A T T T C T T A G
2A -412 2C -447 2G +276 2T -248 3A -21 3C +168 3G +63 3T +498
+598 +432
+255 +605 +388 +310351
+90 +100 +110 +120
A G G C A G C T G A C T C A T C G G T C T A C A T A T G G G G C C G A A A C A A
4A +164 4C -392 4G +467 4T +429 5A -237 5C +234 5G -424 5T +16
+284 +1363 +1131
+510 +226
+130 +140 +150 +160
A G T C A T G G T G C C C C G C G G T A A A G T A G C T C A A A A G T T A C T A
thr +180 +1566 +210 -57 -711 +268
+490 -448 +926 +205 -888
+602
+170 +180 +190 +200
C A A T G G C C A C C T A G G G T G A C G A G G T G A A T G T G T T T G G C G A
-555 +641 +515 +744 +765
+805 +212 +996 +989
351 +346 +583 -619
+210 +220 +230 +240
T C C G T T C G C A T G A T G A A C G G A A A C A C T C A A T C C G C A C C T A
+422 +387 -1158 -1605 -91 -442
+398 +407 +200 +565 +335
+1160 +985 +591
+250 +260
T G G C T C C T G G C A A T C T A T C C C
+455 +755
+239
+1198 +462 +301

"blue" gene weight matrix: 6 bp wide

$\gamma = 16$ binding sites

🟩 found real site

🟥 missed real site

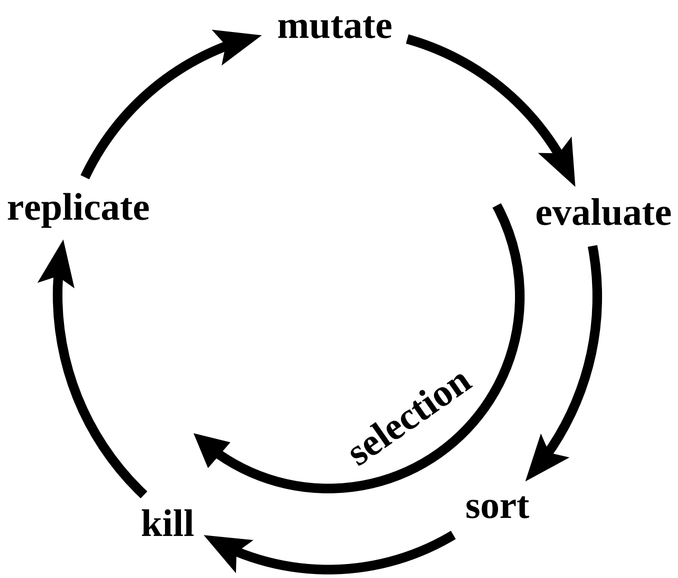
🟨 found wrong site

Genome positions available $G = 256$ bases

$R_{frequency} = \log_2 256/16 = 4$ bits

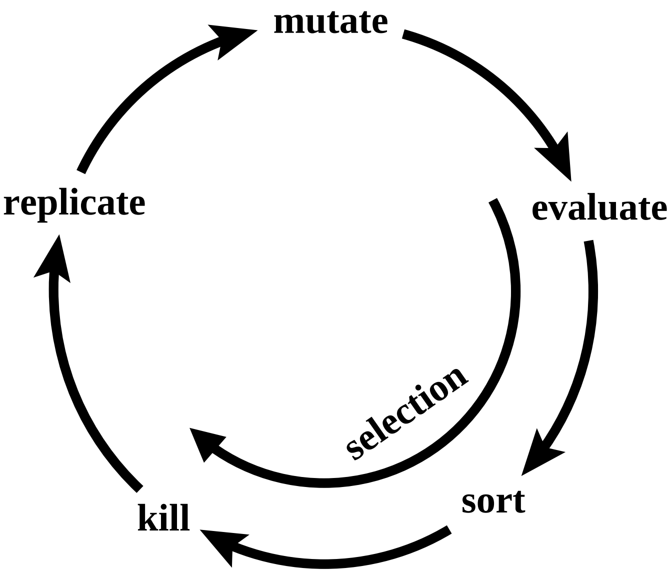
Evolution Cycle

- EVALUATE each creature



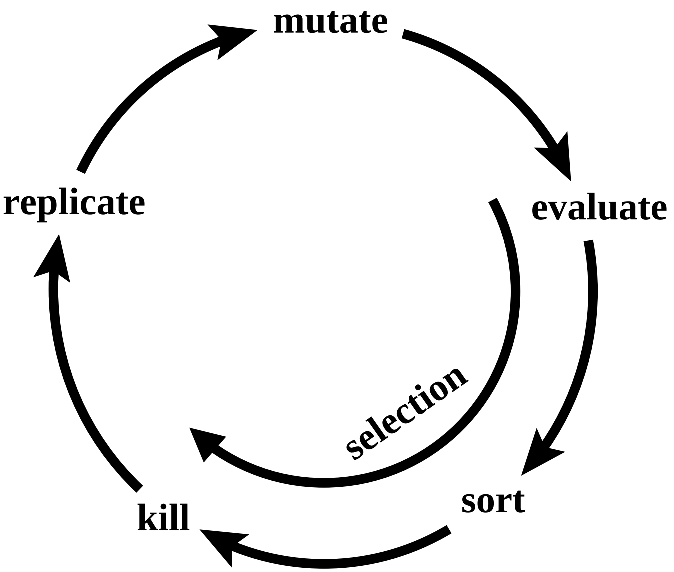
Evolution Cycle

- EVALUATE each creature
  - translate the recognizer gene into a weight matrix



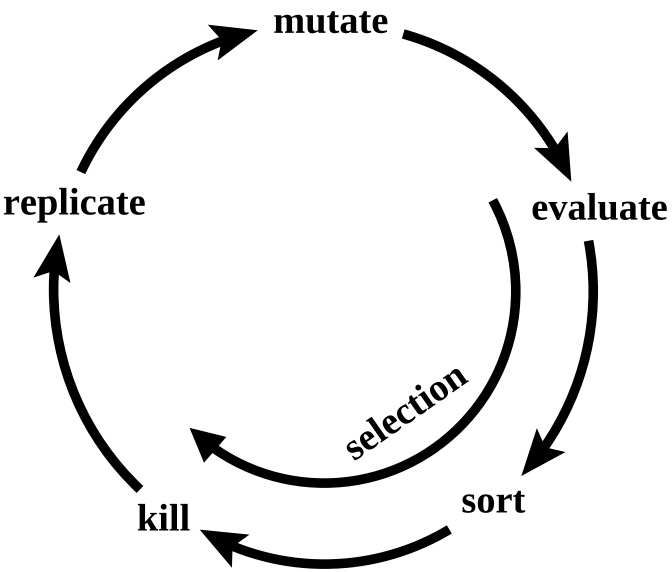
Evolution Cycle

- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome



Evolution Cycle

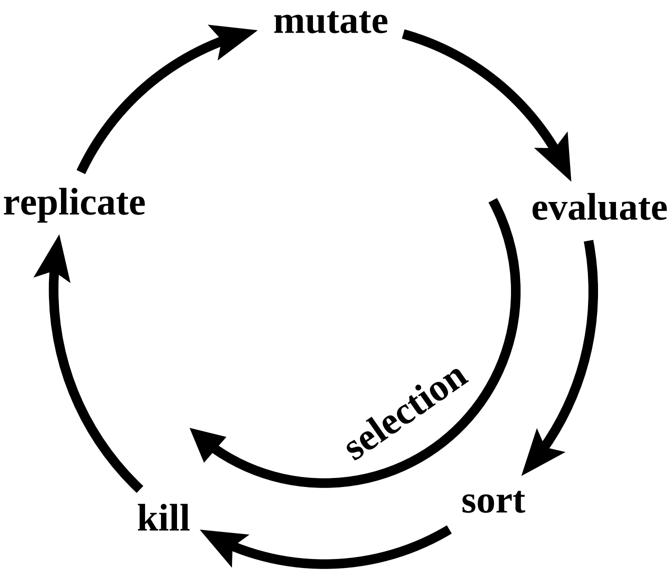
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:



Evolution Cycle

- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:

    🟥 missing a site at a right place

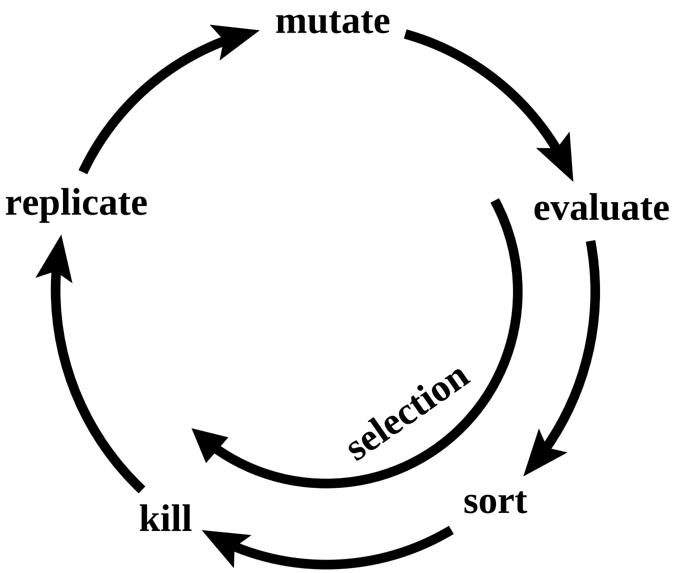


Evolution Cycle

- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:

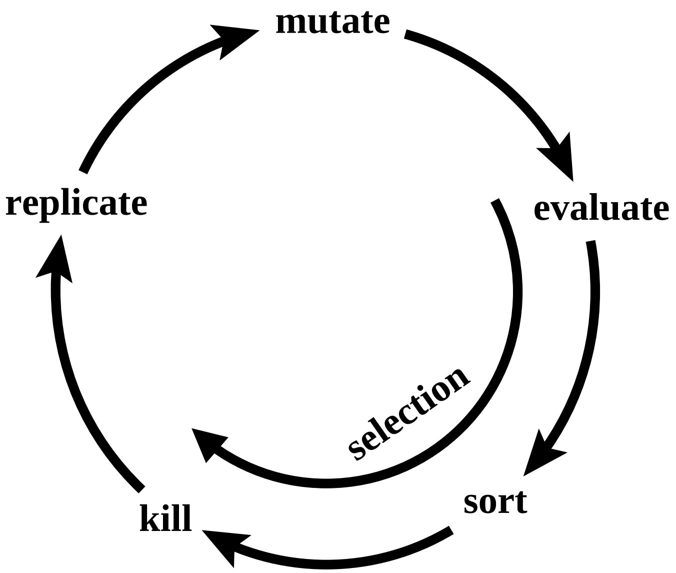
    🟥 missing a site at a right place

    🟨 finding a site at a wrong place



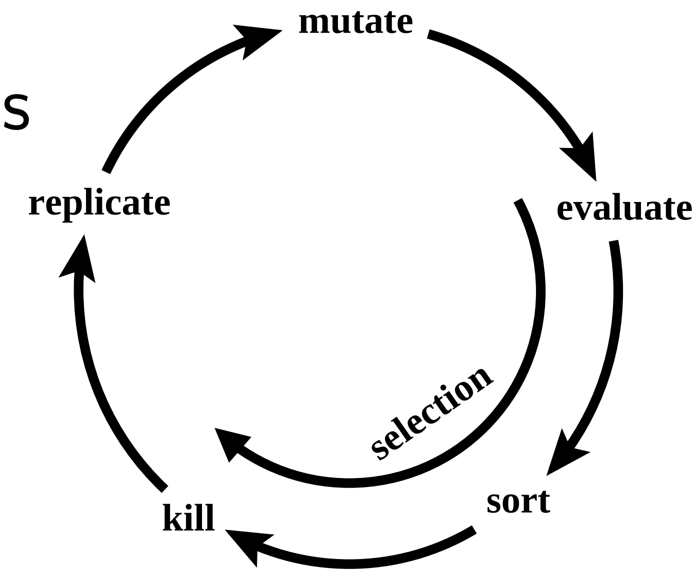
Evolution Cycle

- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    🟥 missing a site at a right place
    🟨 finding a site at a wrong place
  - Sort the creatures by their mistakes



Evolution Cycle

- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    🟥 missing a site at a right place
    🟨 finding a site at a wrong place
  - Sort the creatures by their mistakes
- REPLICATE: the best creatures are duplicated and replace the worst ones

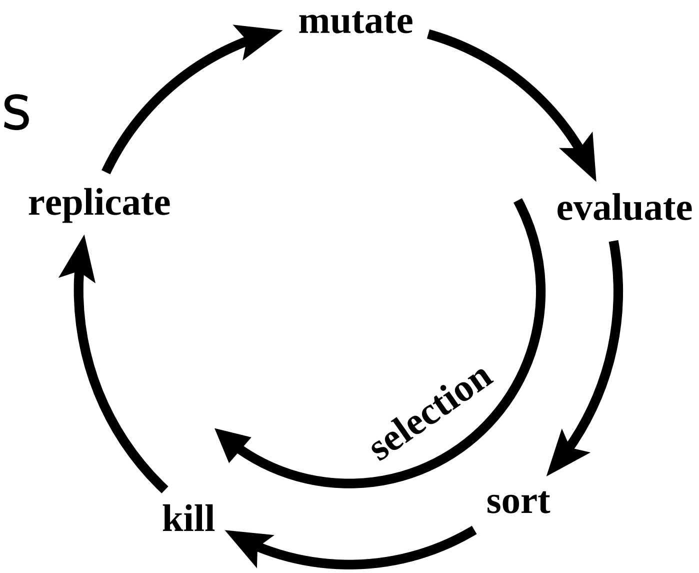


Evolution Cycle

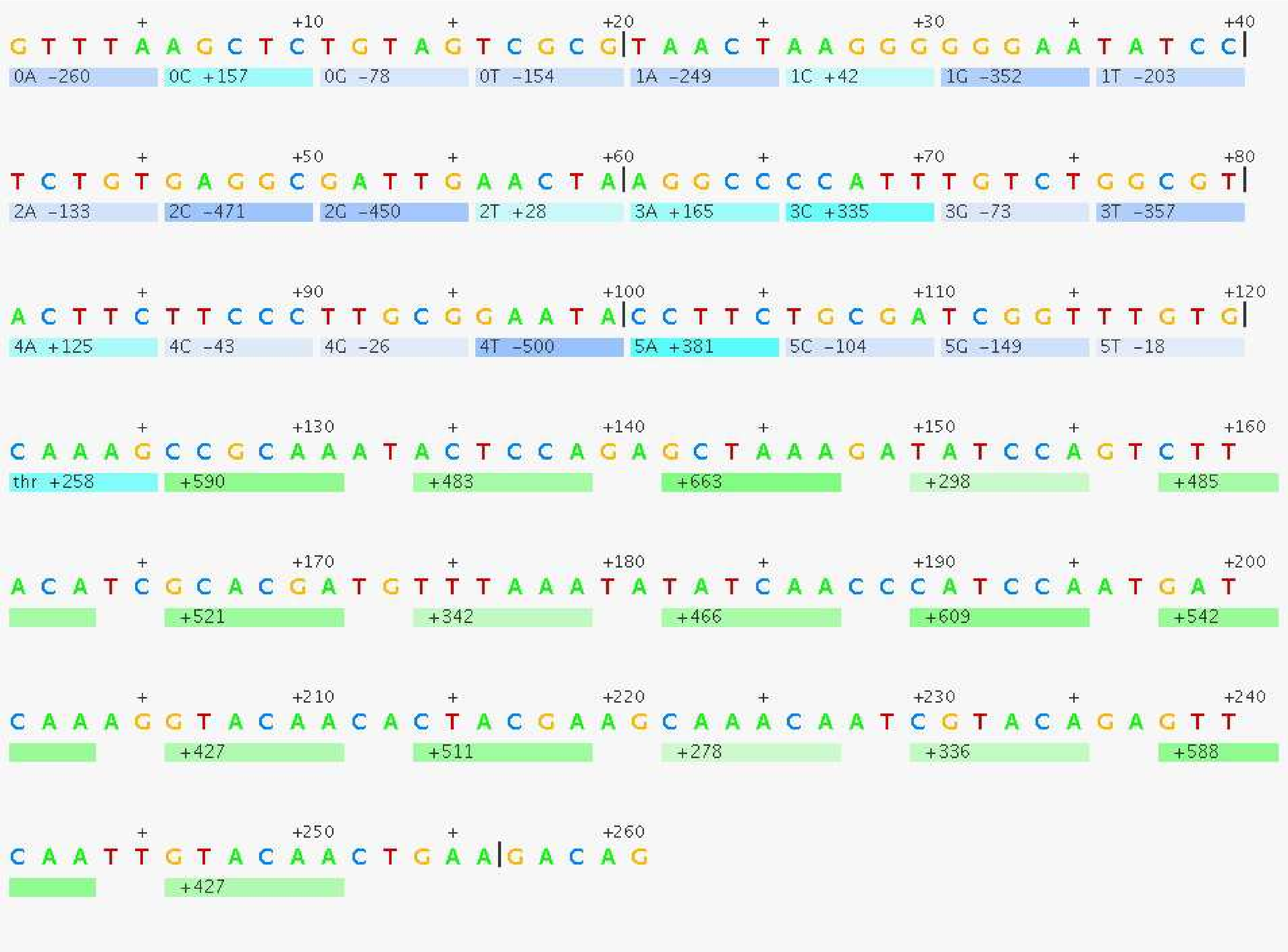
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:

    🟥 missing a site at a right place

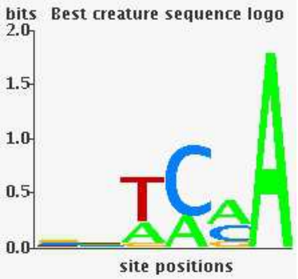
    🟨 finding a site at a wrong place
  - Sort the creatures by their mistakes
- REPLICATE: the best creatures are duplicated and replace the worst ones
- MUTATE all genomes randomly



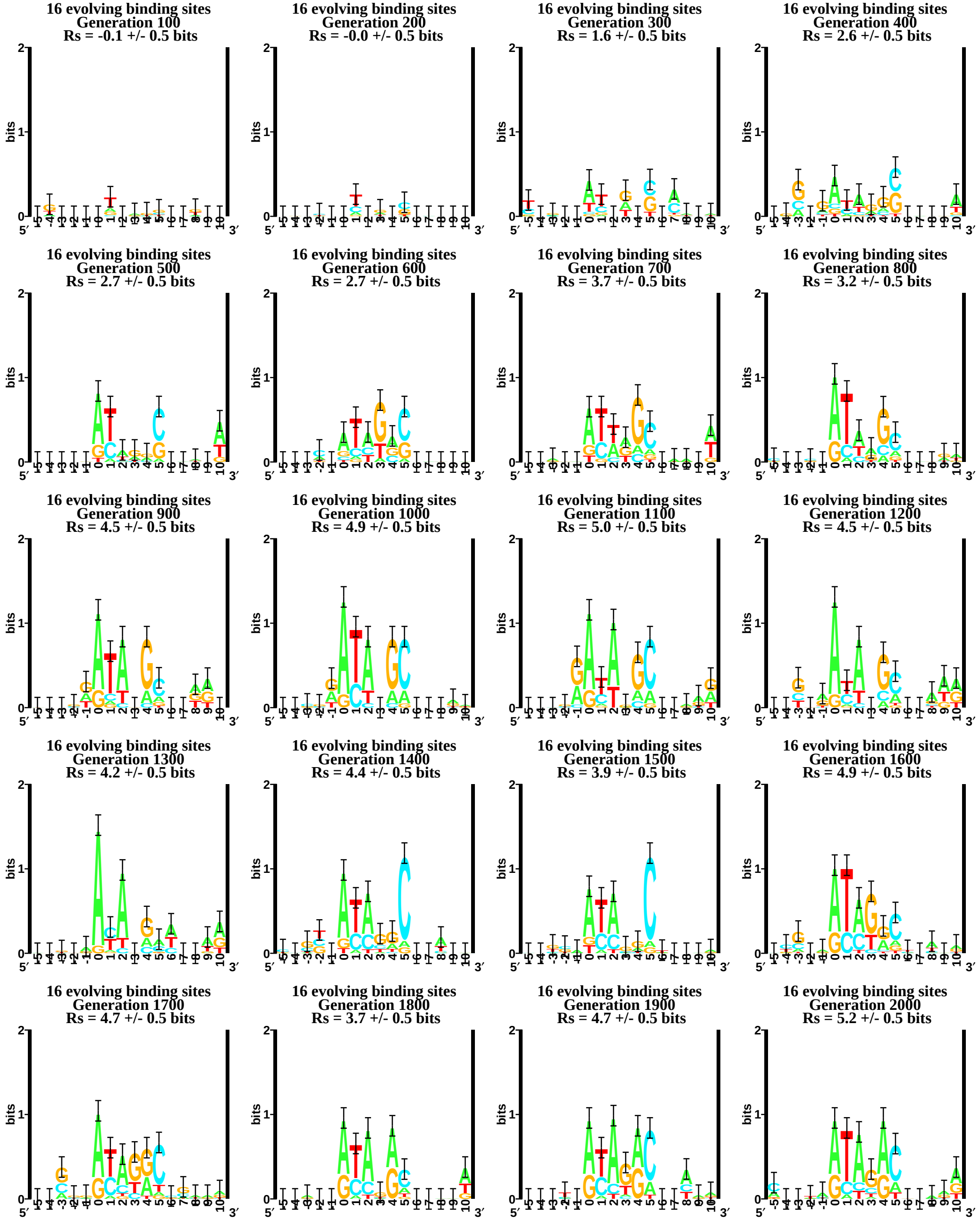
Evolved Ev Creature



```
        +         +10         +         +20         +         +30         +         +40
G T T T A A G C T C T G T A G T C G C G|T A A C T A A G G G G G G A A T A T C C|
0A -260   0C +157   0G -78    0T -154   1A -249   1C +42    1G -352   1T -203

        +         +50         +         +60         +         +70         +         +80
T C T G T G A G G C G A T T G A A C T A|A G G C C C C A T T T G T C T G G C G T|
2A -133   2C -471   2G -450   2T +28    3A +165   3C +335   3G -73    3T -357

        +         +90         +         +100        +         +110        +         +120
A C T T C T T C C C T T G C G G A A T A|C C T T C T G C G A T C G G T T T G T G|
4A +125   4C -43    4G -26    4T -500   5A +381   5C -104   5G -149   5T -18

        +         +130        +         +140        +         +150        +         +160
C A A A G C C G C A A A T A C T C C A G A G C T A A A G A T A T C C A G T C T T
thr +258   +590          +483            +663            +298            +485

        +         +170        +         +180        +         +190        +         +200
A C A T C G C A C G A T G T T T A A A T A T A T C A A C C C A T C C A A T G A T
           +521          +342            +466            +609            +542

        +         +210        +         +220        +         +230        +         +240
C A A A G G T A C A A C A C T A C G A A G C A A A C A A T C G T A C A G A G T T
           +427          +511            +278            +336            +588

        +         +250        +         +260
C A A T T G T A C A A C T G A A|G A C A G
           +427
```



Evolution of Binding Sites



16 evolving binding sites
Generation 100
Rs = -0.1 +/- 0.5 bits

16 evolving binding sites
Generation 200
Rs = -0.0 +/- 0.5 bits

16 evolving binding sites
Generation 300
Rs = 1.6 +/- 0.5 bits

16 evolving binding sites
Generation 400
Rs = 2.6 +/- 0.5 bits

16 evolving binding sites
Generation 500
Rs = 2.7 +/- 0.5 bits

16 evolving binding sites
Generation 600
Rs = 2.7 +/- 0.5 bits

16 evolving binding sites
Generation 700
Rs = 3.7 +/- 0.5 bits

16 evolving binding sites
Generation 800
Rs = 3.2 +/- 0.5 bits

16 evolving binding sites
Generation 900
Rs = 4.5 +/- 0.5 bits

16 evolving binding sites
Generation 1000
Rs = 4.9 +/- 0.5 bits

16 evolving binding sites
Generation 1100
Rs = 5.0 +/- 0.5 bits

16 evolving binding sites
Generation 1200
Rs = 4.5 +/- 0.5 bits

16 evolving binding sites
Generation 1300
Rs = 4.2 +/- 0.5 bits

16 evolving binding sites
Generation 1400
Rs = 4.4 +/- 0.5 bits

16 evolving binding sites
Generation 1500
Rs = 3.9 +/- 0.5 bits

16 evolving binding sites
Generation 1600
Rs = 4.9 +/- 0.5 bits

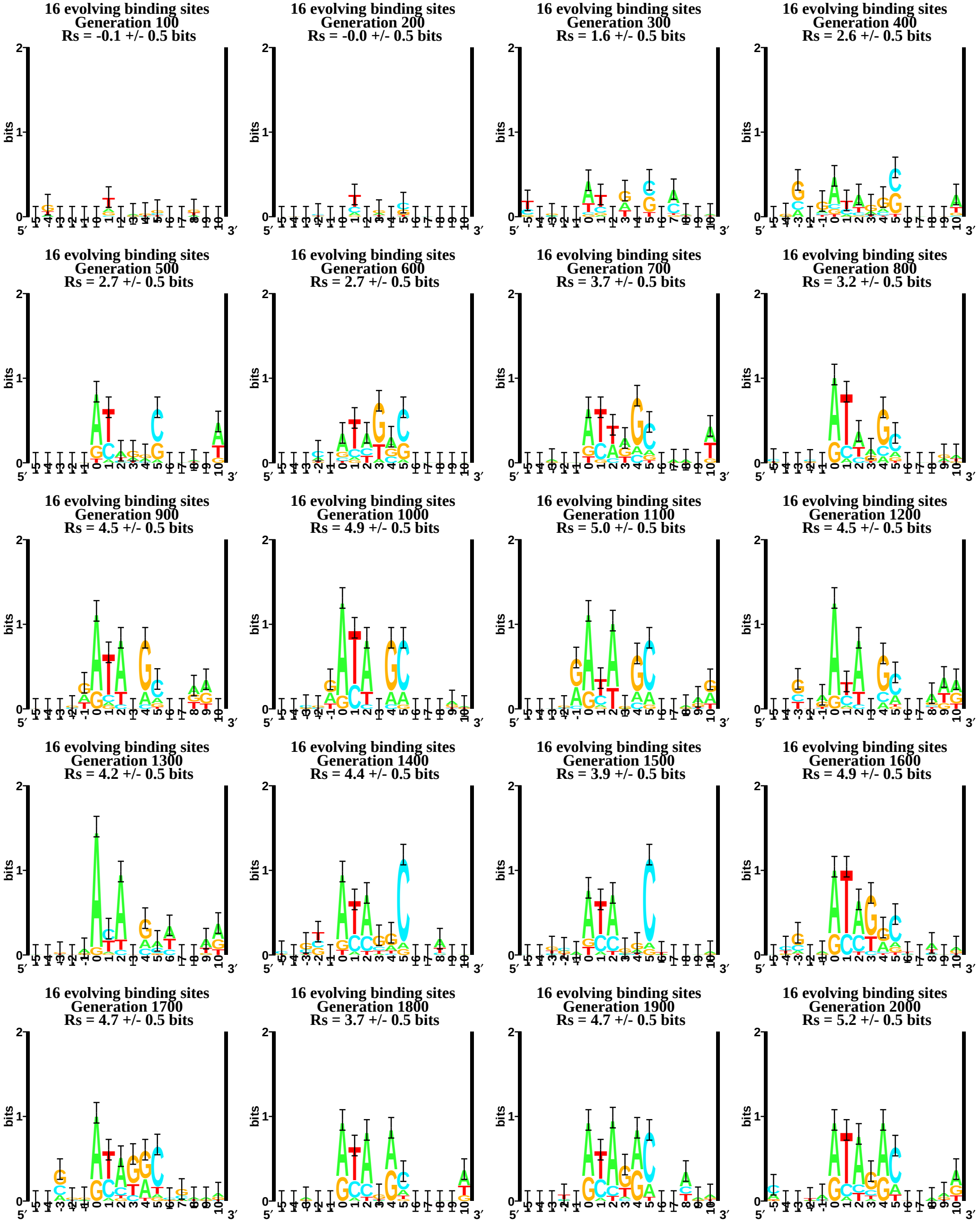
16 evolving binding sites
Generation 1700
Rs = 4.7 +/- 0.5 bits

16 evolving binding sites
Generation 1800
Rs = 3.7 +/- 0.5 bits

16 evolving binding sites
Generation 1900
Rs = 4.7 +/- 0.5 bits

16 evolving binding sites
Generation 2000
Rs = 5.2 +/- 0.5 bits

Evolution of Binding Sites



16 evolving binding sites Generation 100 Rs = -0.1 +/- 0.5 bits

16 evolving binding sites Generation 200 Rs = -0.0 +/- 0.5 bits

16 evolving binding sites Generation 300 Rs = 1.6 +/- 0.5 bits

16 evolving binding sites Generation 400 Rs = 2.6 +/- 0.5 bits

16 evolving binding sites Generation 500 Rs = 2.7 +/- 0.5 bits

16 evolving binding sites Generation 600 Rs = 2.7 +/- 0.5 bits

16 evolving binding sites Generation 700 Rs = 3.7 +/- 0.5 bits

16 evolving binding sites Generation 800 Rs = 3.2 +/- 0.5 bits

16 evolving binding sites Generation 900 Rs = 4.5 +/- 0.5 bits

16 evolving binding sites Generation 1000 Rs = 4.9 +/- 0.5 bits

16 evolving binding sites Generation 1100 Rs = 5.0 +/- 0.5 bits

16 evolving binding sites Generation 1200 Rs = 4.5 +/- 0.5 bits

16 evolving binding sites Generation 1300 Rs = 4.2 +/- 0.5 bits

16 evolving binding sites Generation 1400 Rs = 4.4 +/- 0.5 bits

16 evolving binding sites Generation 1500 Rs = 3.9 +/- 0.5 bits

16 evolving binding sites Generation 1600 Rs = 4.9 +/- 0.5 bits

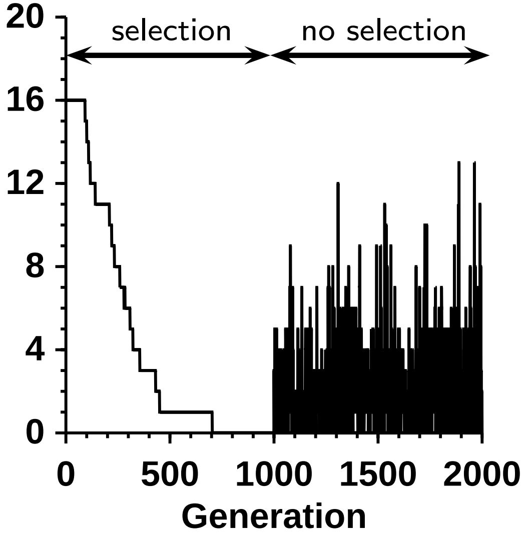
16 evolving binding sites Generation 1700 Rs = 4.7 +/- 0.5 bits

16 evolving binding sites Generation 1800 Rs = 3.7 +/- 0.5 bits

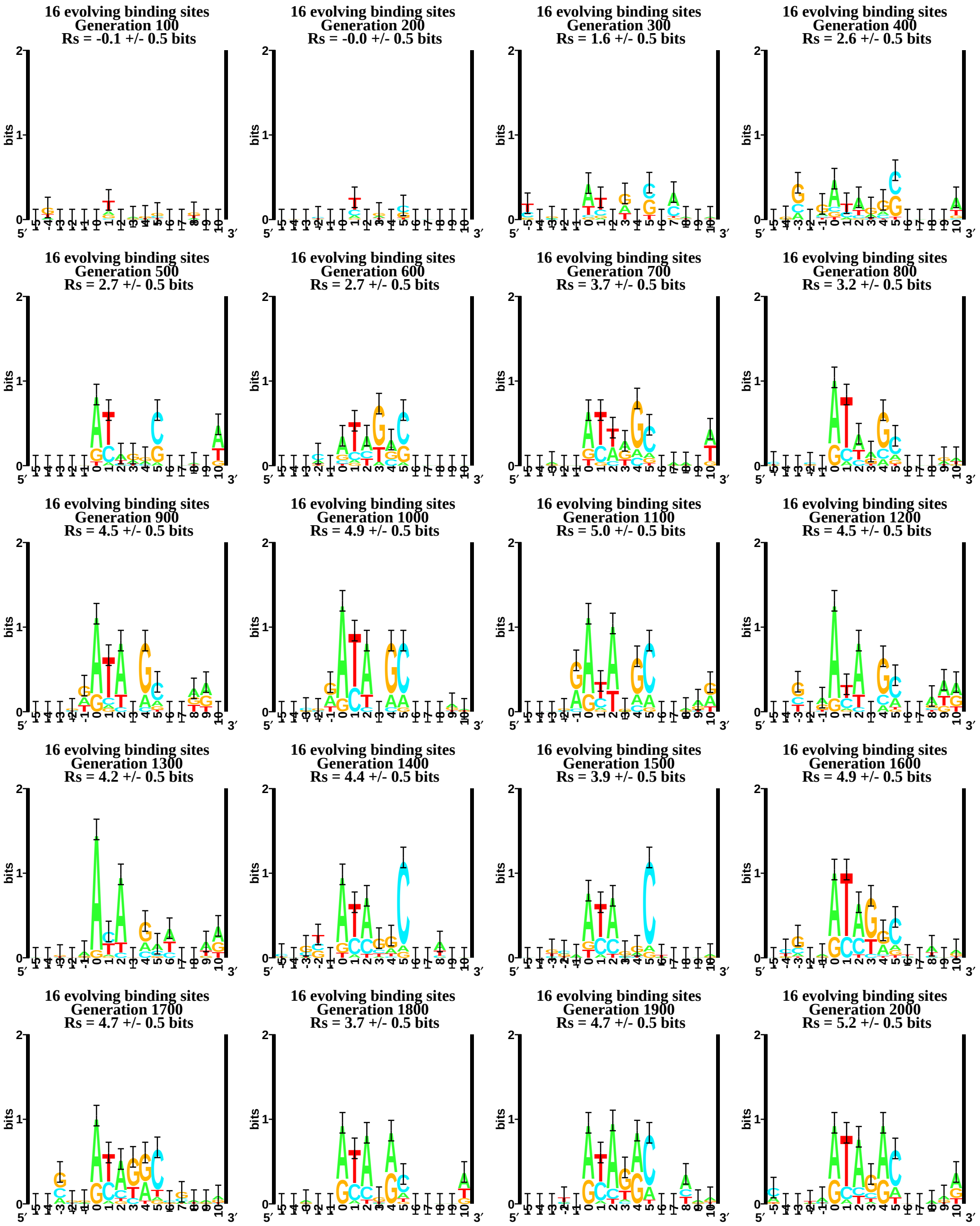
16 evolving binding sites Generation 1900 Rs = 4.7 +/- 0.5 bits

16 evolving binding sites Generation 2000 Rs = 5.2 +/- 0.5 bits

Mistakes of Best Organism



Evolution of Binding Sites



16 evolving binding sites
Generation 100
Rs = -0.1 +/- 0.5 bits

16 evolving binding sites
Generation 200
Rs = -0.0 +/- 0.5 bits

16 evolving binding sites
Generation 300
Rs = 1.6 +/- 0.5 bits

16 evolving binding sites
Generation 400
Rs = 2.6 +/- 0.5 bits

16 evolving binding sites
Generation 500
Rs = 2.7 +/- 0.5 bits

16 evolving binding sites
Generation 600
Rs = 2.7 +/- 0.5 bits

16 evolving binding sites
Generation 700
Rs = 3.7 +/- 0.5 bits

16 evolving binding sites
Generation 800
Rs = 3.2 +/- 0.5 bits

16 evolving binding sites
Generation 900
Rs = 4.5 +/- 0.5 bits

16 evolving binding sites
Generation 1000
Rs = 4.9 +/- 0.5 bits

16 evolving binding sites
Generation 1100
Rs = 5.0 +/- 0.5 bits

16 evolving binding sites
Generation 1200
Rs = 4.5 +/- 0.5 bits

16 evolving binding sites
Generation 1300
Rs = 4.2 +/- 0.5 bits

16 evolving binding sites
Generation 1400
Rs = 4.4 +/- 0.5 bits

16 evolving binding sites
Generation 1500
Rs = 3.9 +/- 0.5 bits

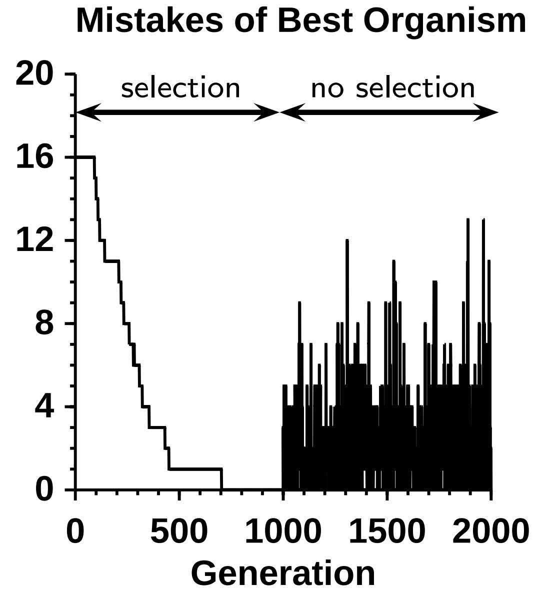
16 evolving binding sites
Generation 1600
Rs = 4.9 +/- 0.5 bits

16 evolving binding sites
Generation 1700
Rs = 4.7 +/- 0.5 bits

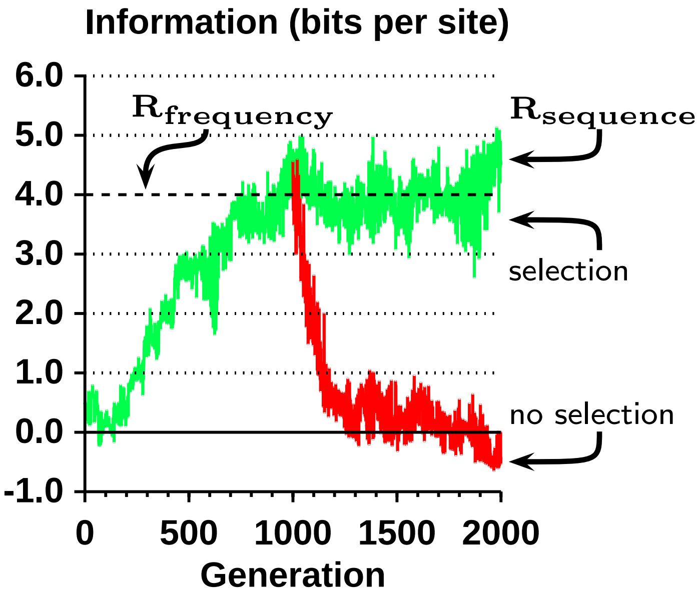
16 evolving binding sites
Generation 1800
Rs = 3.7 +/- 0.5 bits

16 evolving binding sites
Generation 1900
Rs = 4.7 +/- 0.5 bits

16 evolving binding sites
Generation 2000
Rs = 5.2 +/- 0.5 bits



Mistakes of Best Organism



Information (bits per site)

Shannon Information Measure of Binding Site Patterns

Information is measured as a **decrease in uncertainty**:

$$R = H_{before} - H_{after} \qquad \text{(bits per symbol)} \qquad (15)$$

Mathematics of Evolution 2

Before binding there are 4 possible bases at each position l, so the uncertainty is:

$$\begin{aligned} H_{before}(l) &= \log_2 4 \qquad \text{(bits per base)} \\ &\approx 2 \end{aligned} \tag{16}$$

Mathematics of Evolution 2

Before binding there are 4 possible bases at each position l, so the uncertainty is:

$$\begin{aligned} H_{before}(l) &= \log_2 4 \qquad \text{(bits per base)} \\ &\approx 2 \end{aligned} \tag{16}$$

After binding the uncertainty depends on the frequencies of bases b at positions l in a binding site, $f(b,l)$:

$$H_{after}(l) = -\sum_{b \in \{A,C,G,T\}} f(b,l) \log_2 f(b,l) \tag{17}$$

(bits per base)

Mathematics of Evolution 3

The **information at a position** l is:

$$\begin{aligned} R_{sequence}(l) &= H_{before}(l) - H_{after}(l) \qquad (18) \\ &\approx 2 - H_{after}(l) \qquad \text{(bits \textit{per base})} \end{aligned}$$

Mathematics of Evolution 3

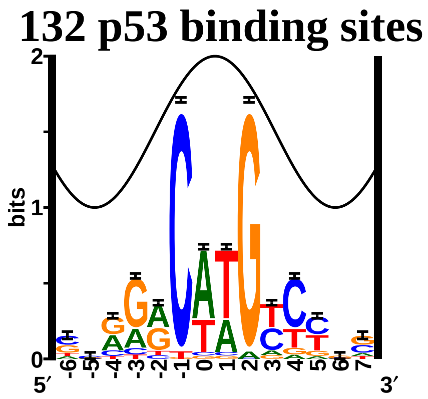
The **information at a position** l is:

$$\begin{aligned} R_{sequence}(l) &= H_{before}(l) - H_{after}(l) \\ &\approx 2 - H_{after}(l) \qquad \text{(bits \textit{per base})} \end{aligned} \tag{18}$$

The **total site information** is:

$$\begin{aligned} R_{sequence} &= \sum_{l} (H_{before}(l) - H_{after}(l)) \\ &\approx 2l - H_{after} \qquad \text{(bits \textit{per site})} \end{aligned} \tag{19}$$

During evolution,
as $H_{after} \downarrow$, $R_{sequence} \uparrow$



Acknowledgements

- Larry Gold
- Gary Stormo
- Andrzej Ehrenfeucht
- Paul Anagnostopoulos

Version

version = 1.20 of evtalk.tex 2012 Mar 15