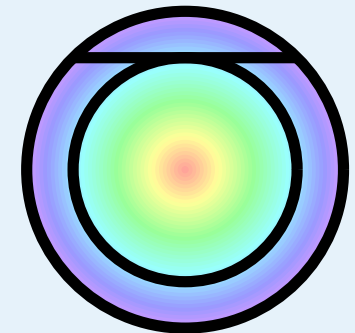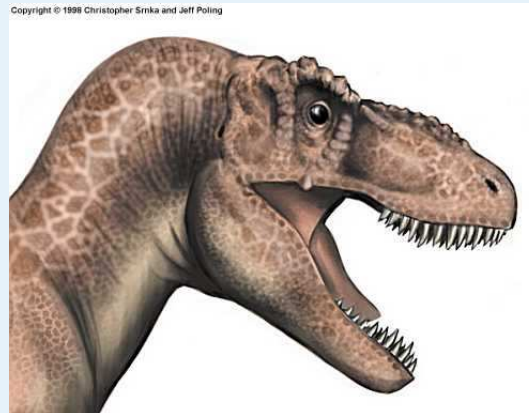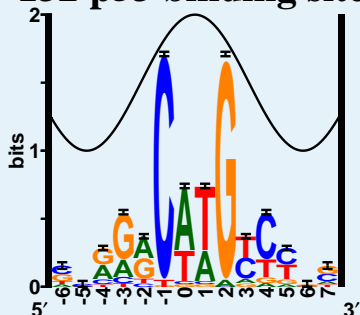# Evolution of Binding Sites

Thomas D. Schneider, Ph.D.

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

Molecular Information Theory Group

132 p53 binding sites

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	H T
4	2	11 10 01 00
8	3	
$M=2^B$	$B=\log_2 M$	

T A
C G

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	
$M=2^B$	$B=\log_2 M$	

Sequence Logo
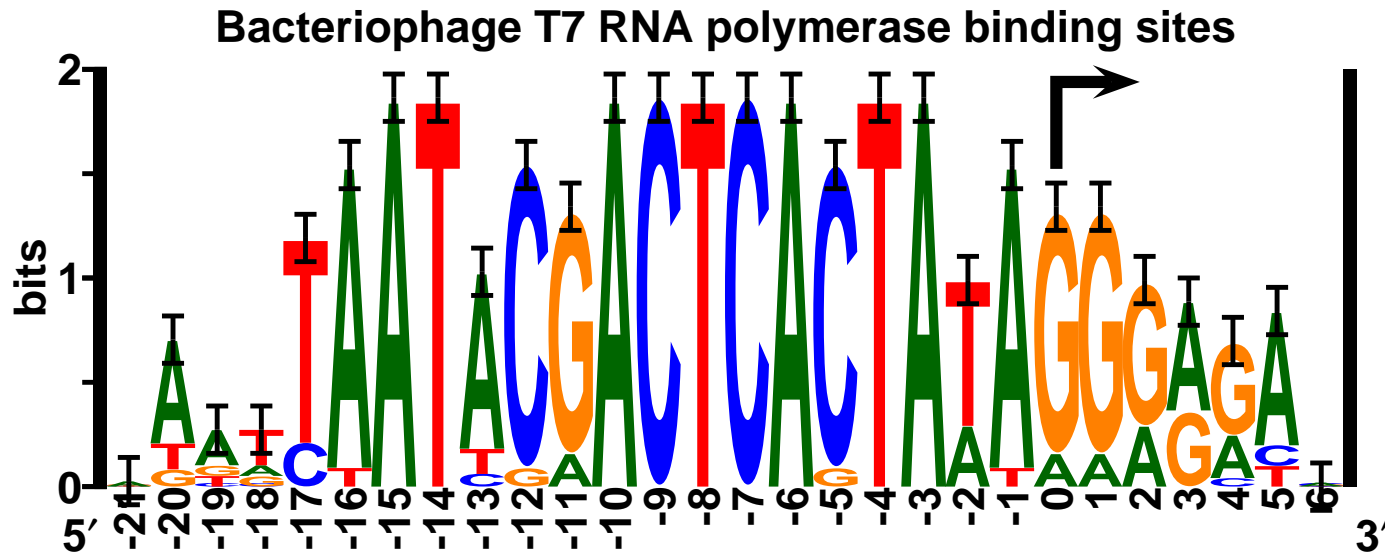
Bacteriophage T7 RNA polymerase binding sites

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

1	ttattaatacaactcactataaggagag
2	aaatcaatacgactcactatagaggggac
3	cggttaatacgactcactataggagaac
4	gaagtaatacgactcagtataggacaa
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	t t a t t a a t a c a a c t c a c t a t a a g g a g a g
2	a a a t c a a t a c g a c t c a c t a t a g a g g g a c
3	c g g t t a a t a c g a c t c a c t a t a g g a g a a c
4	g a a g t a a t a c g a c t c a g t a t a g g g a c a a
5	t a a t t a a t t g a a c t c a c t a a a g g g a g a c
6	c g c t t a a t a c g a c t c a c t a a a g g a g a c a

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	cggttaatacgactcactatagagaac
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	cggttaatacgactcactatagagaac
4	gaagtaatacgactcagtatagggacaa
5	taattaattgaactcactaaagggagac
6	cgcttaatacgactcactaaaggagaca
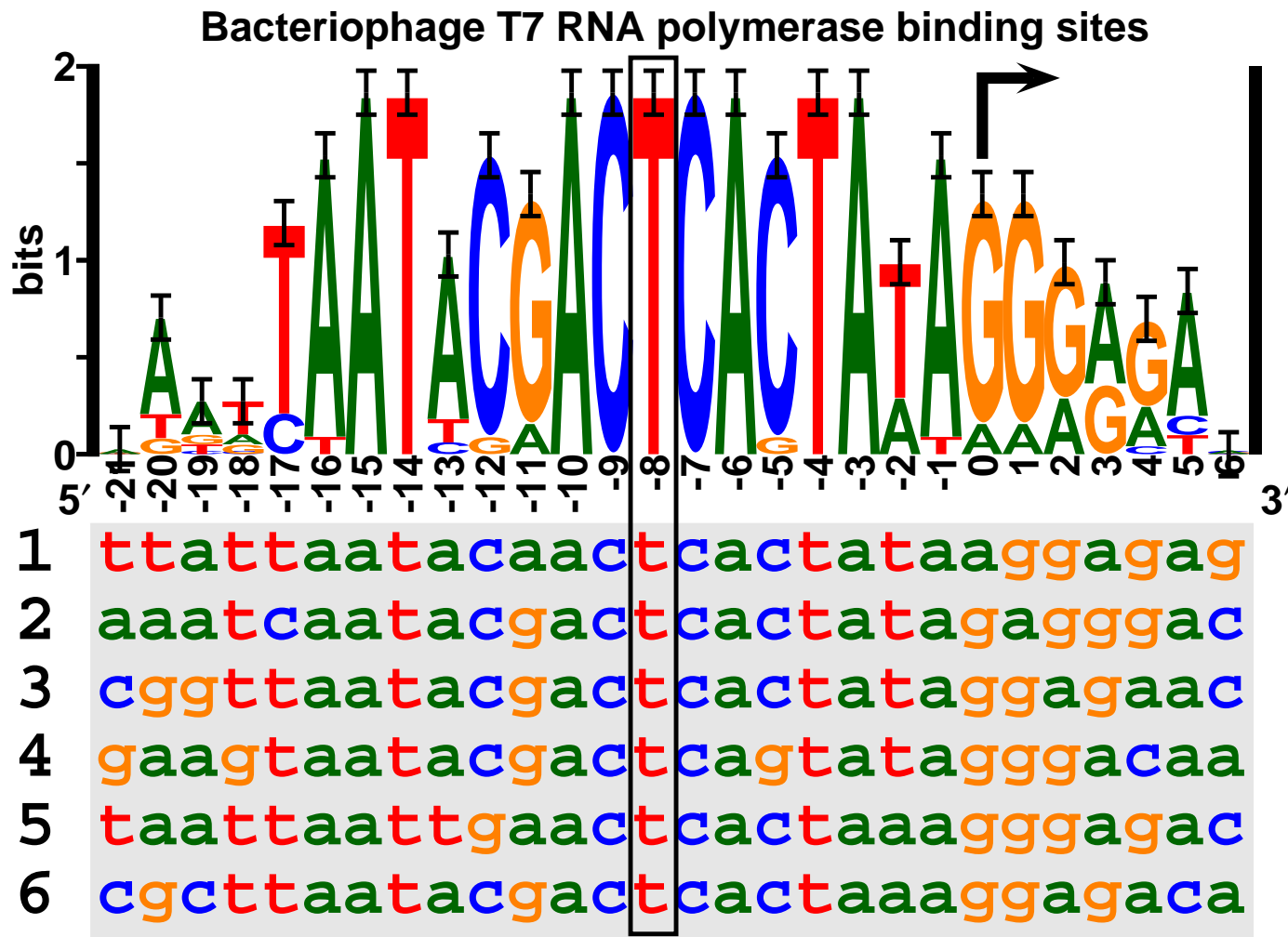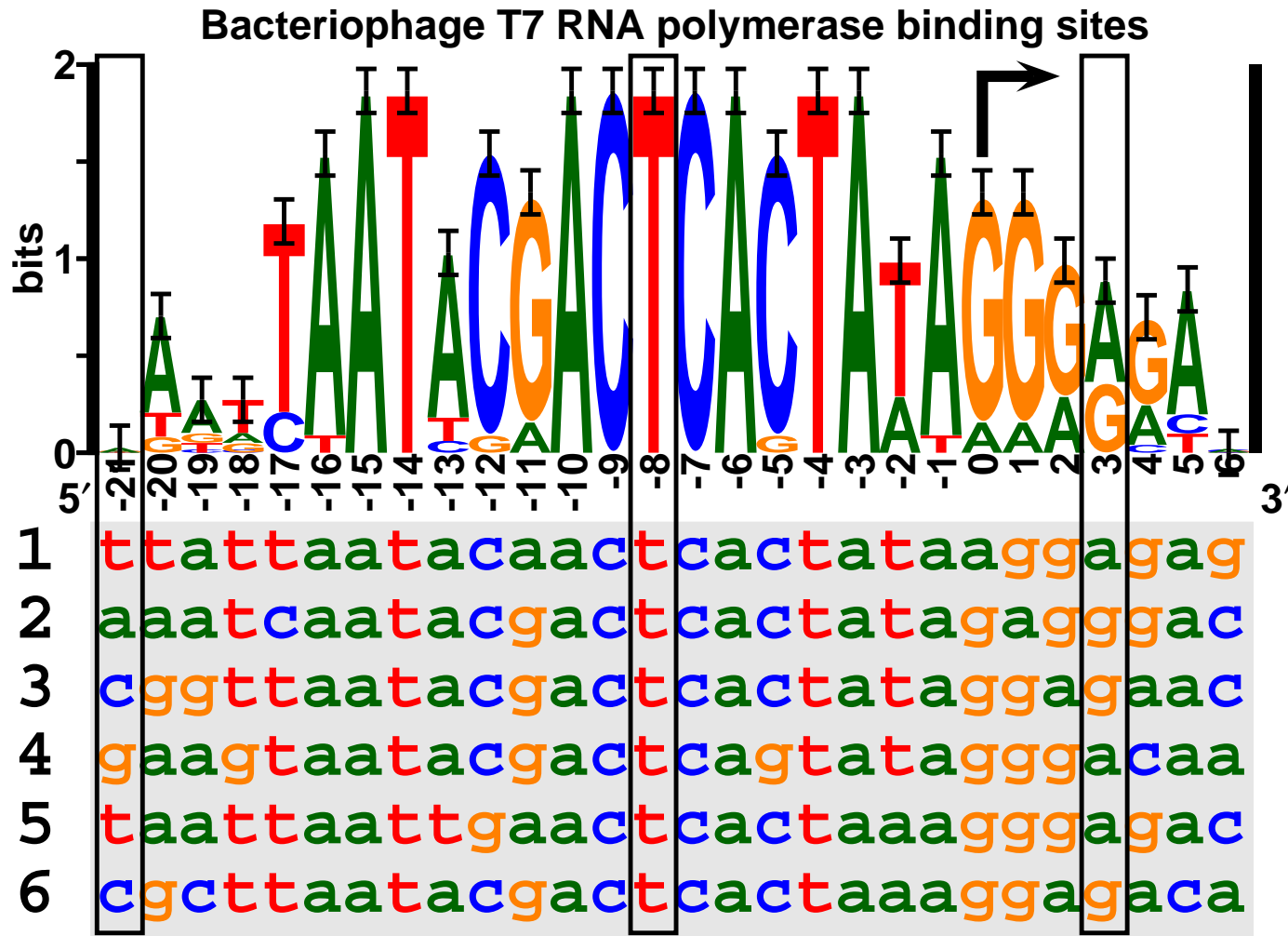
6 of 17 sites

An Intuitive Approach

Information to chose one symbol from M symbols:

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

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \qquad\qquad (1)$$
$$= \quad -\log_2 1/M.$$

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

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \tag{1}$$
$$= -\log_2 1/M.$$

$1/M$ 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

EXAMPLE

A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \tag{3}$$
$$P_{\text{silent}} = 1023/1024 \tag{4}$$

EXAMPLE

A phone rings once every 1024 seconds.

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

Surprisal:

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

EXAMPLE

A phone rings once every 1024 seconds.

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

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

Surprisal:

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

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

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

$$H = P_{ring} \times surprisal_{ring}$$

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:

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

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:

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

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

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

$$H \; = \; \sum_{i=1}^{M} P_i \times \left(\text{surprisal for} P_i\right) \qquad\qquad (9)$$

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

$$H \;=\; \sum_{i=1}^{M} P_i \times \left(\text{surprisal for} P_i\right) \qquad (9)$$

$$=\; \sum_{i=1}^{M} P_i \times \left(-\log_2 P_i\right) \qquad (10)$$

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

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

$$= \; \sum_{i=1}^{M} P_i \times \left(-\log_2 P_i\right) \tag{10}$$

$$= \; -\sum_{i=1}^{M} P_i \log_2 P_i \quad \text{bits per symbol} \tag{11}$$

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \qquad (12)$$

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \qquad (12)$$

Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \qquad (13)$$

132 p53 binding sites

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \qquad (12)$$

Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \qquad (13)$$

and

$$H_{\text{after}} = \text{uncertainty of bases}$$

$$= -\sum_{base=A}^{T} P_{base} \log_2 P_{base} \qquad (14)$$

132 p53 binding sites

Information is a decrease in uncertainty

$$R = H_{\text{before}} - H_{\text{after}} \qquad (12)$$

Example a sequence logo is computed from equiprobable bases before:

$$H_{\text{before}} = 2 \text{ bits/base} \qquad (13)$$

and

$$H_{\text{after}} = \text{uncertainty of bases}$$

$$= -\sum_{base=A}^{T} P_{base} \log_2 P_{base} \qquad (14)$$

132 p53 binding sites
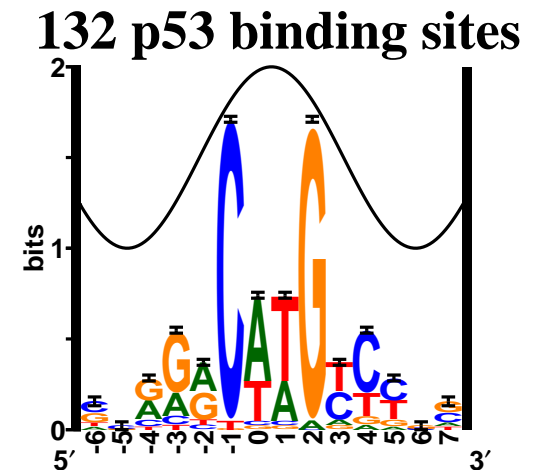
Note: with only one base, $H_{\text{after}} = 0$
so $R = 2$ bits/base.
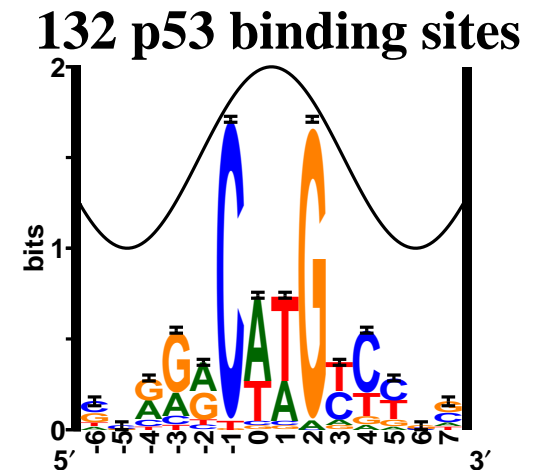
Information required
to find a set of binding sites

$G = \#$ of potential binding sites

Information required
to find a set of binding sites

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

Information required
to find a set of binding sites

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

$$\gamma = \text{number of binding sites on genome}$$

Information required
to find a set of binding sites

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

$$\gamma = \text{number of binding sites on genome}$$

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

Information required
to find a set of binding sites

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

$$\gamma = \text{number of binding sites on genome}$$

$$R_{frequency} = H_{before} - H_{after}$$
$$= \log_2 G - \log_2 \gamma$$

Information required
to find a set of binding sites

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

$$\gamma = \text{number of binding sites on genome}$$

$$R_{frequency} = H_{before} - H_{after}$$
$$= \log_2 G - \log_2 \gamma$$
$$= -\log_2 \gamma/G$$

**Information required
to find a set of binding sites
in a genome**
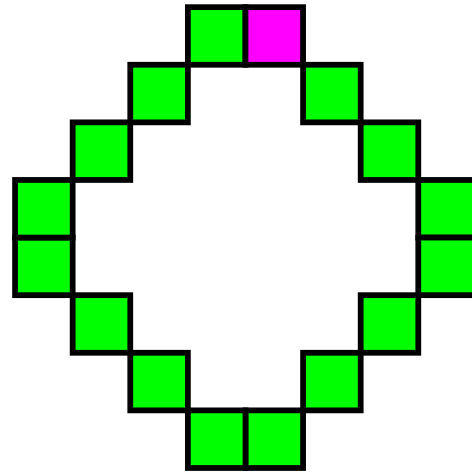
16 positions
  1 site
$\log_2 16/1 = 4$ bits
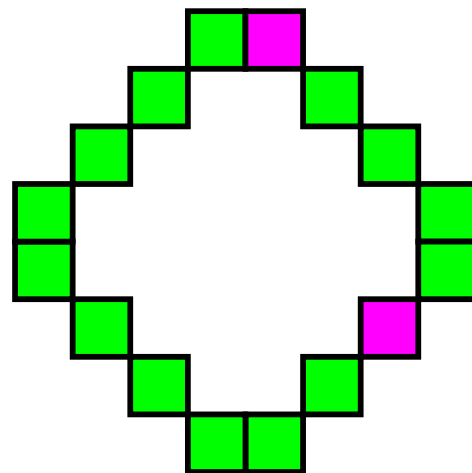
16 positions
  2 sites
$\log_2 16/2 = 3$ bits
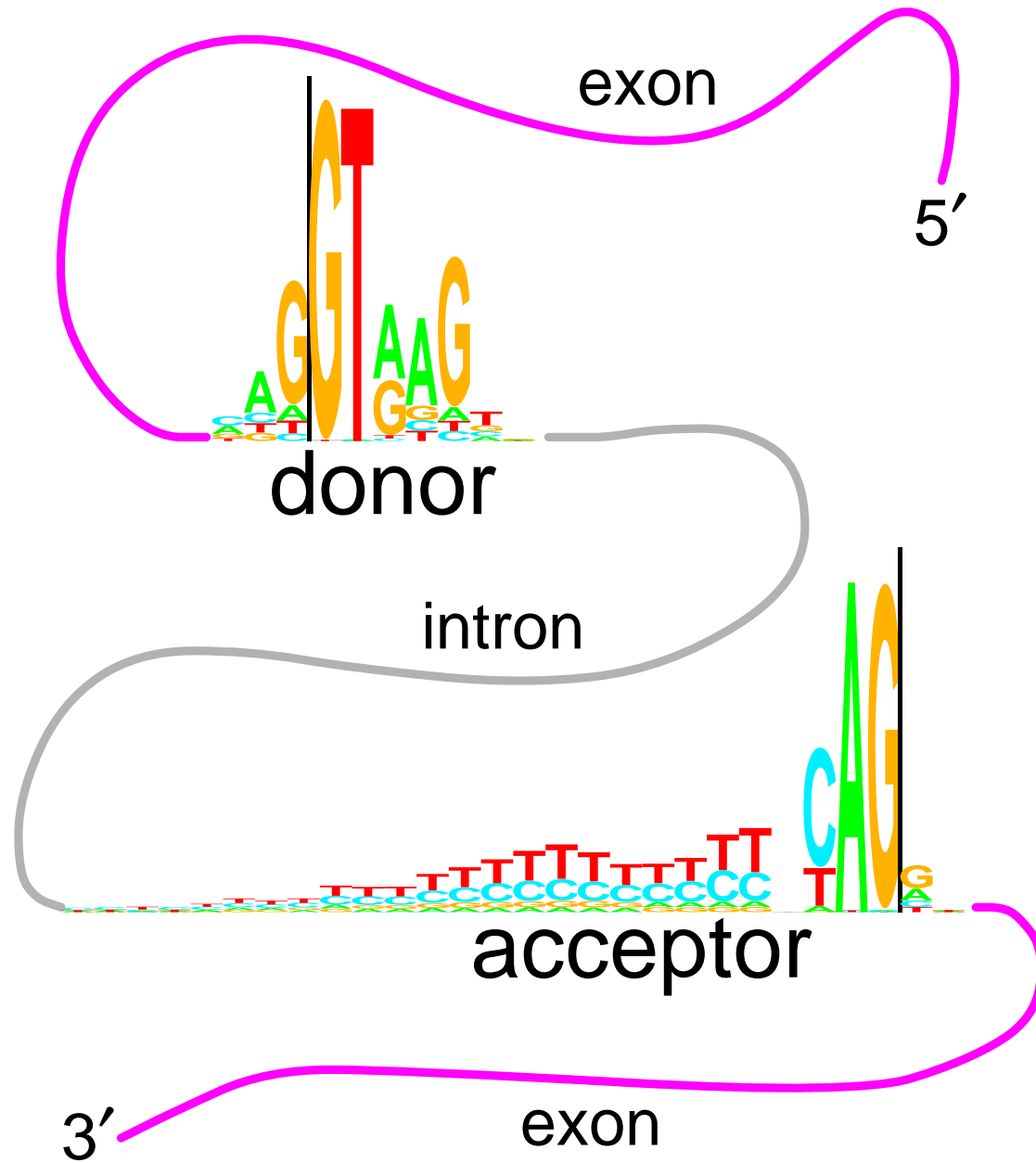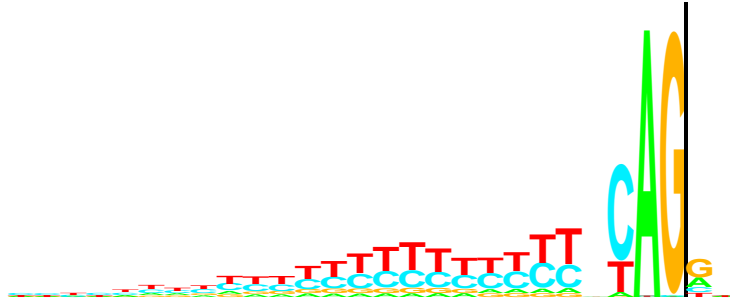
$R_{sequence}$

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

$R_{sequence}$
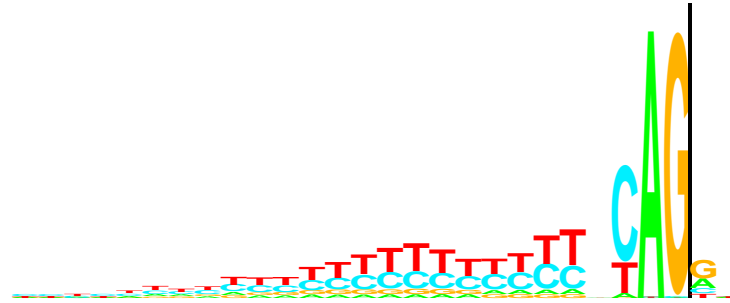
- **Information at binding site sequences (area under sequence logo)**
- **from: binding site sequences**
- **9.4 bits per site**
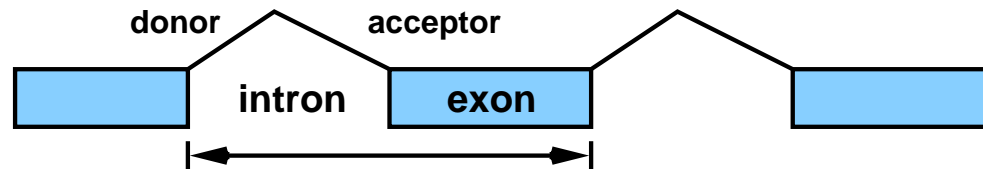
$R_{frequency}$

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

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

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)	$\dfrac{\textbf{Pattern Info}}{\textbf{Location Info}} = \dfrac{\mathbf{R_{sequence}}}{\mathbf{R_{frequency}}}$
Spliceosome acceptor[2]	9.35 ± 0.12	9.66	0.97 ± 0.01
Spliceosome donor	7.92 ± 0.09	9.66	0.82 ± 0.01
Ribosome	11.0	10.6	1.0
λ cI/cro	17.7 ± 1.6	19.3	0.9 ± 0.1
LexA	21.5 ± 1.7	18.4	1.2 ± 0.1
TrpR	23.4 ± 1.9	20.3	1.2 ± 0.1
LacI	19.2 ± 2.8	21.9	0.9 ± 0.1
ArgR	16.4	18.4	0.9
O (λ Origin)	20.9	19.9	1.0
Ara C	19.3	19.3	1.0
Transcription at TATA[3]	3.3	~ 3	~ 1
T7 Promoter	35.4	16.5	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.

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

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 $\sim$50 *E. coli* LexA sites)

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

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 $\sim$50 *E. coli* LexA sites)

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

Rsequence must evolve towards Rfrequency!

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

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

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

Setup a Computer Model, 'Ev':
A population of "creatures" with

- $R_{frequency}$ 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)

- $R_{frequency}$ 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)

- $R_{frequency}$ 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)

- $R_{frequency}$ 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)

$\left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} R_{frequency}$ is fixed

- $R_{frequency}$ 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) $\left.\right\}$ $R_{frequency}$ is fixed
- a recognizer gene encoded in the sequence:
  use a weight matrix

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	+1.3	−1.4	−8.8	−5.8	+1.1	+1.5	−1.8	−0.7	+0.0
C	+0.6	−0.8	−2.4	−7.8	−5.5	−3.7	−1.6	−2.2	−0.5	−0.2
G	−0.6	−1.0	+1.6	+2.0	−6.2	+0.7	−1.1	+1.7	−0.3	+0.4
T	−1.0	−0.9	−1.7	−5.8	+2.0	−3.4	−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	+1.3	−1.4	−8.8	−5.8	+1.1	+1.5	−1.8	−0.7	+0.0
C	+0.6	−0.8	−2.4	−7.8	−5.5	−3.7	−1.6	−2.2	−0.5	−0.2
G	−0.6	−1.0	+1.6	+2.0	−6.2	+0.7	−1.1	+1.7	−0.3	+0.4
T	−1.0	−0.9	−1.7	−5.8	+2.0	−3.4	−1.6	−2.2	+0.9	−0.5

5′ c a g g t c t g c a 3′
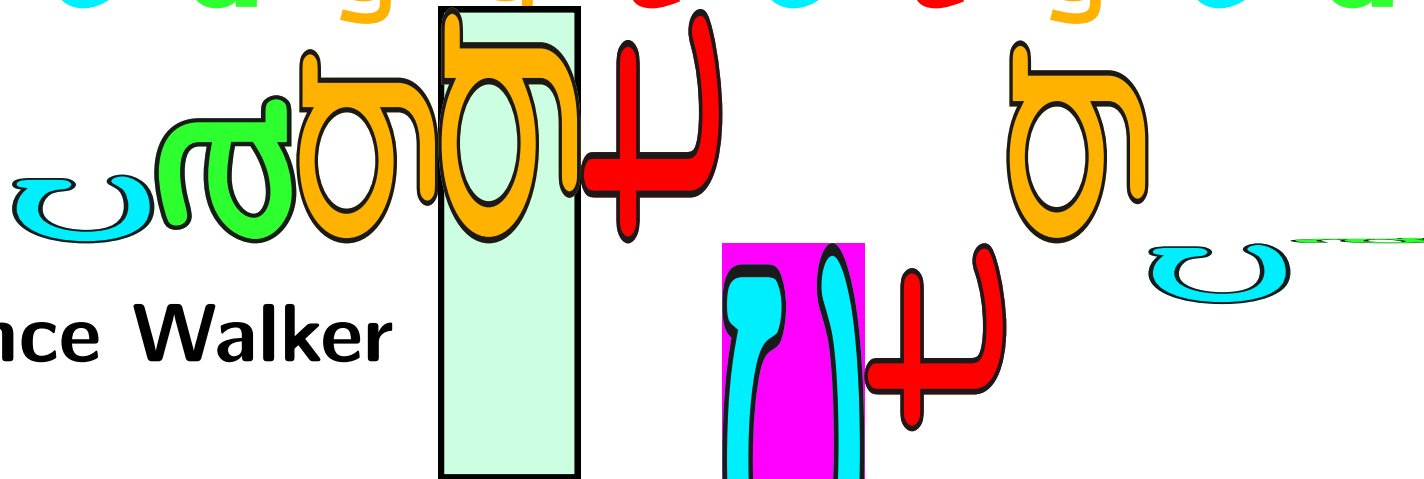
Sequence Walker

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
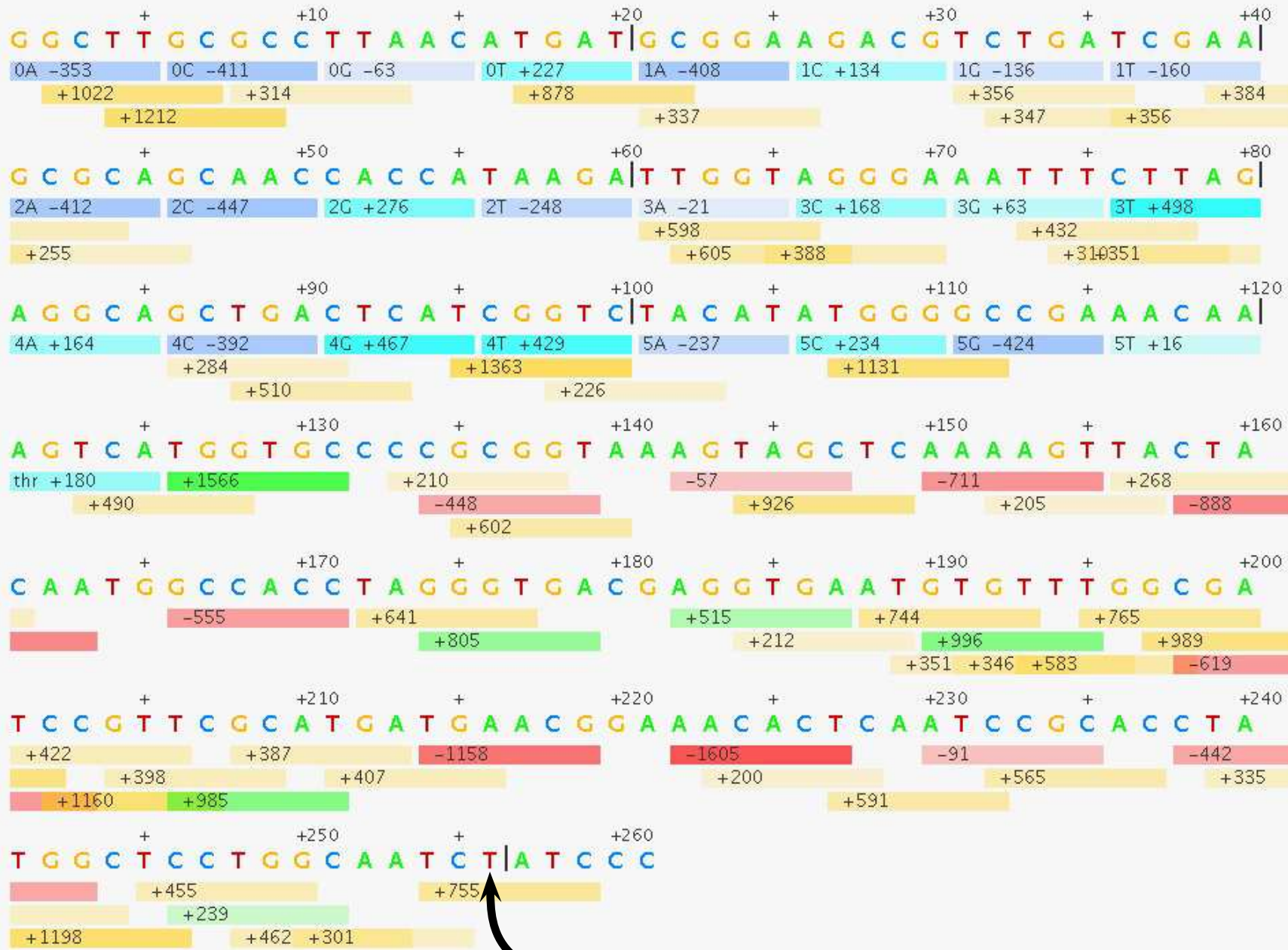   +239
+1198   +462 +301

Unevolved Ev Creature

"blue" gene weight matrix: 6 bp wide
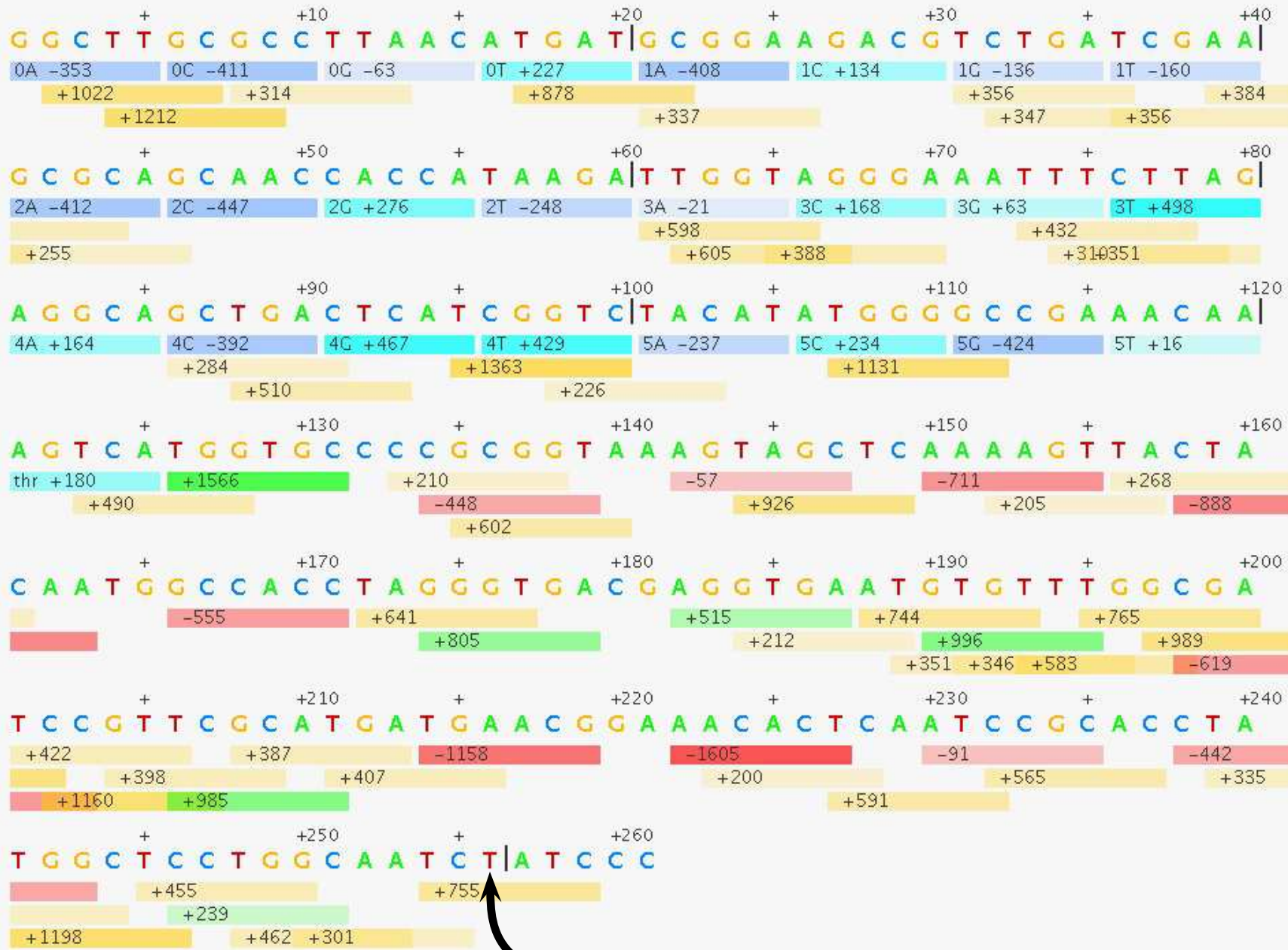
Unevolved Ev Creature

"blue" gene weight matrix: 6 bp wide

Genome positions available $G = 256$ bases

```
        +            +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                        +310 +351

        +            +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
```
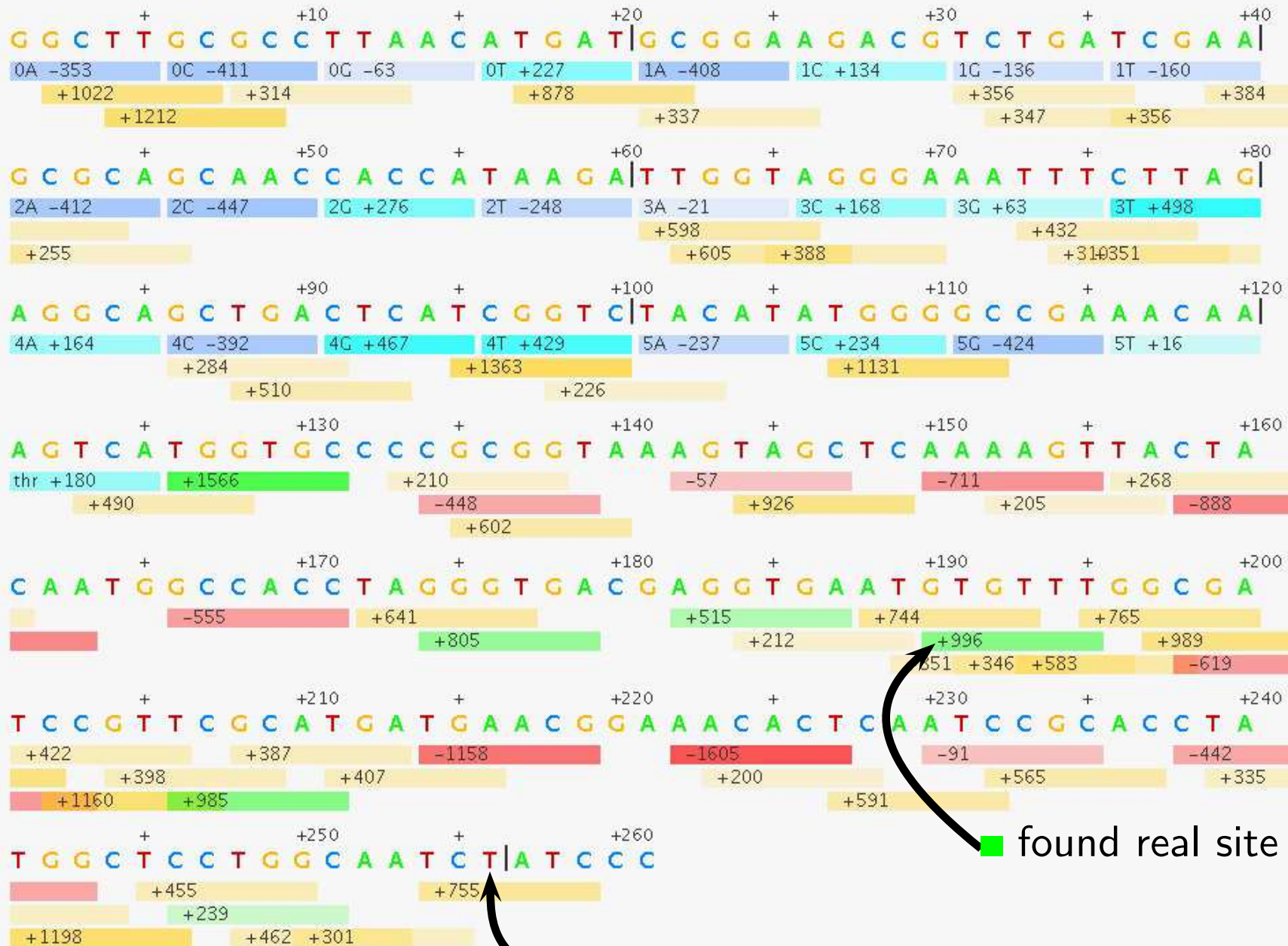
"blue" gene weight matrix: 6 bp wide

```
        +            +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
```

$\gamma = 16$ binding sites

Genome positions available $G = 256$ bases

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

Unevolved Ev Creature

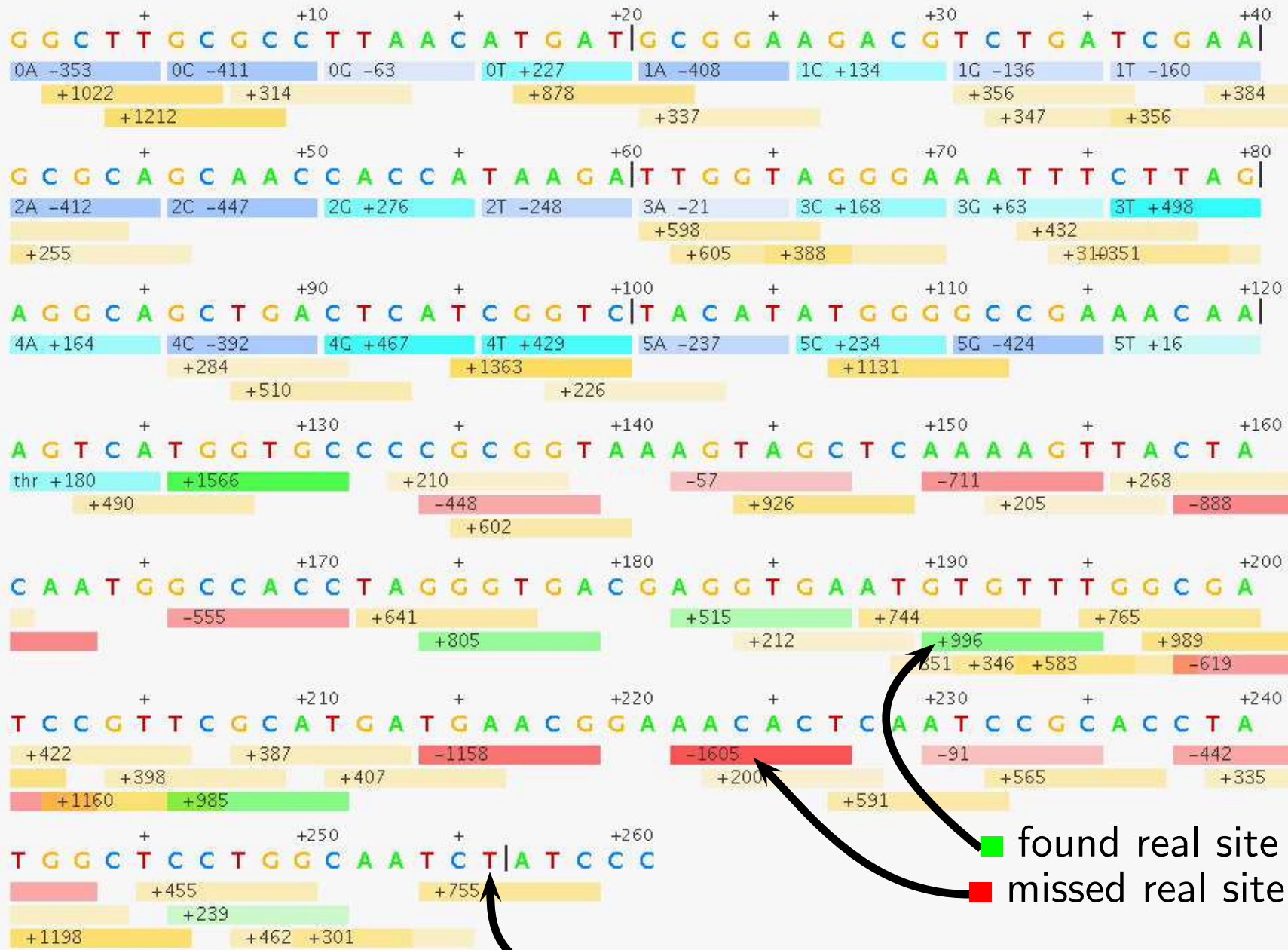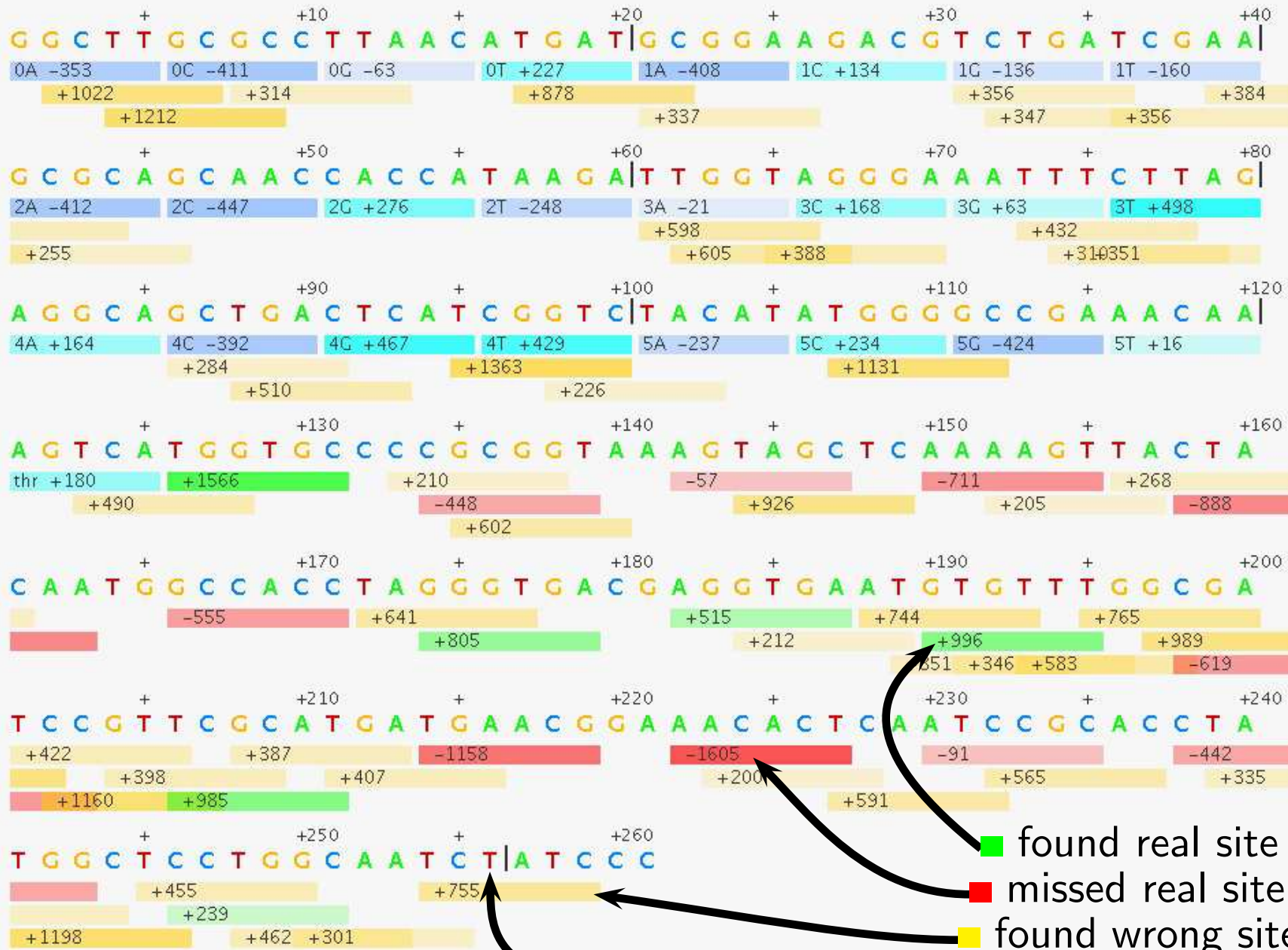"blue" gene weight matrix: 6 bp wide

$\gamma = 16$ binding sites

■ found real site

Genome positions available $G = 256$ bases
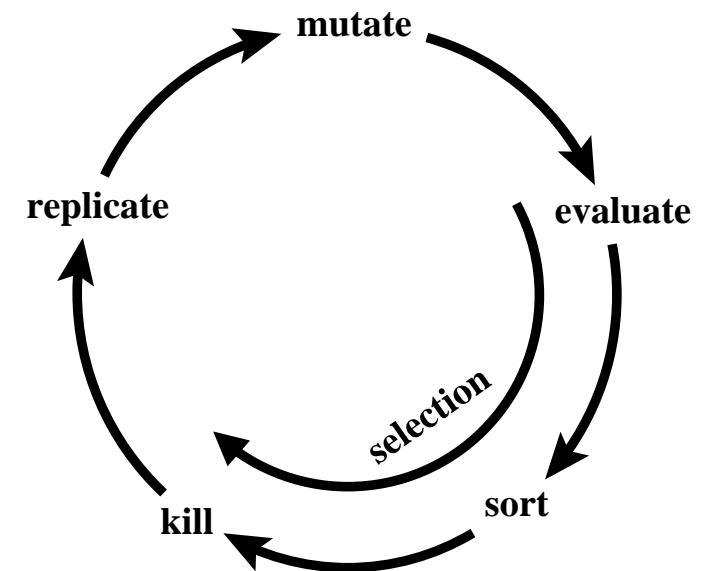
$R_{frequency} = \log_2 256/16 = 4$ bits

Unevolved Ev Creature

"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

"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
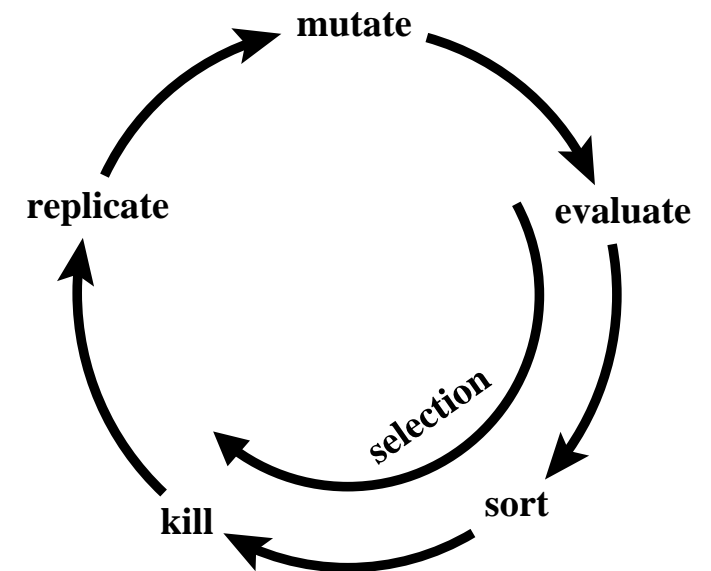
- EVALUATE each creature

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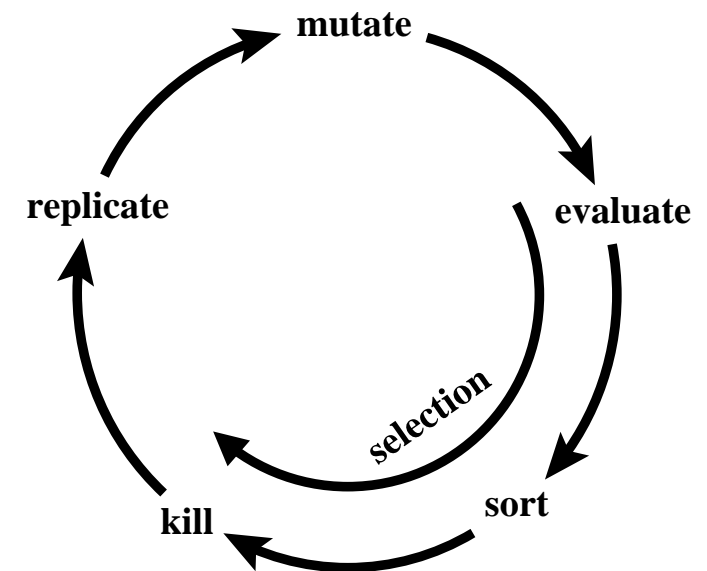
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
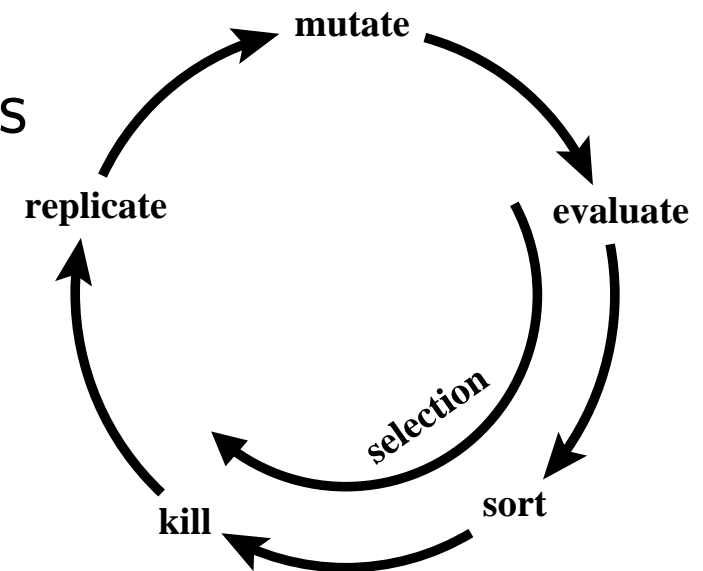
- EVALUATE each creature

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

- 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

- 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

mutate

replicate   evaluate

selection

kill   sort

- 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

mutate

replicate

evaluate

selection

kill

sort

- 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

- 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
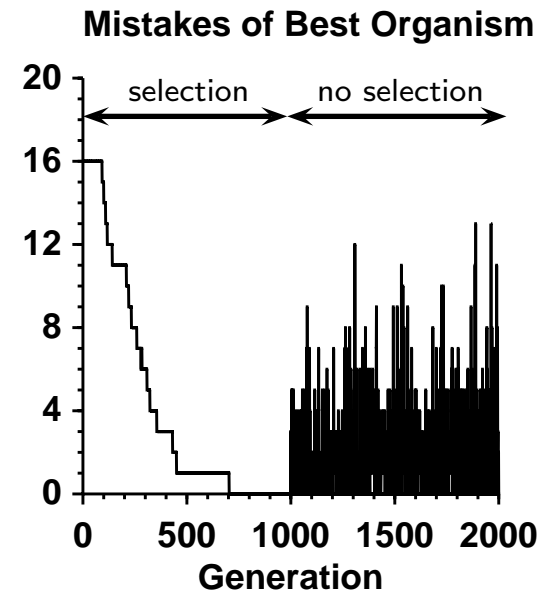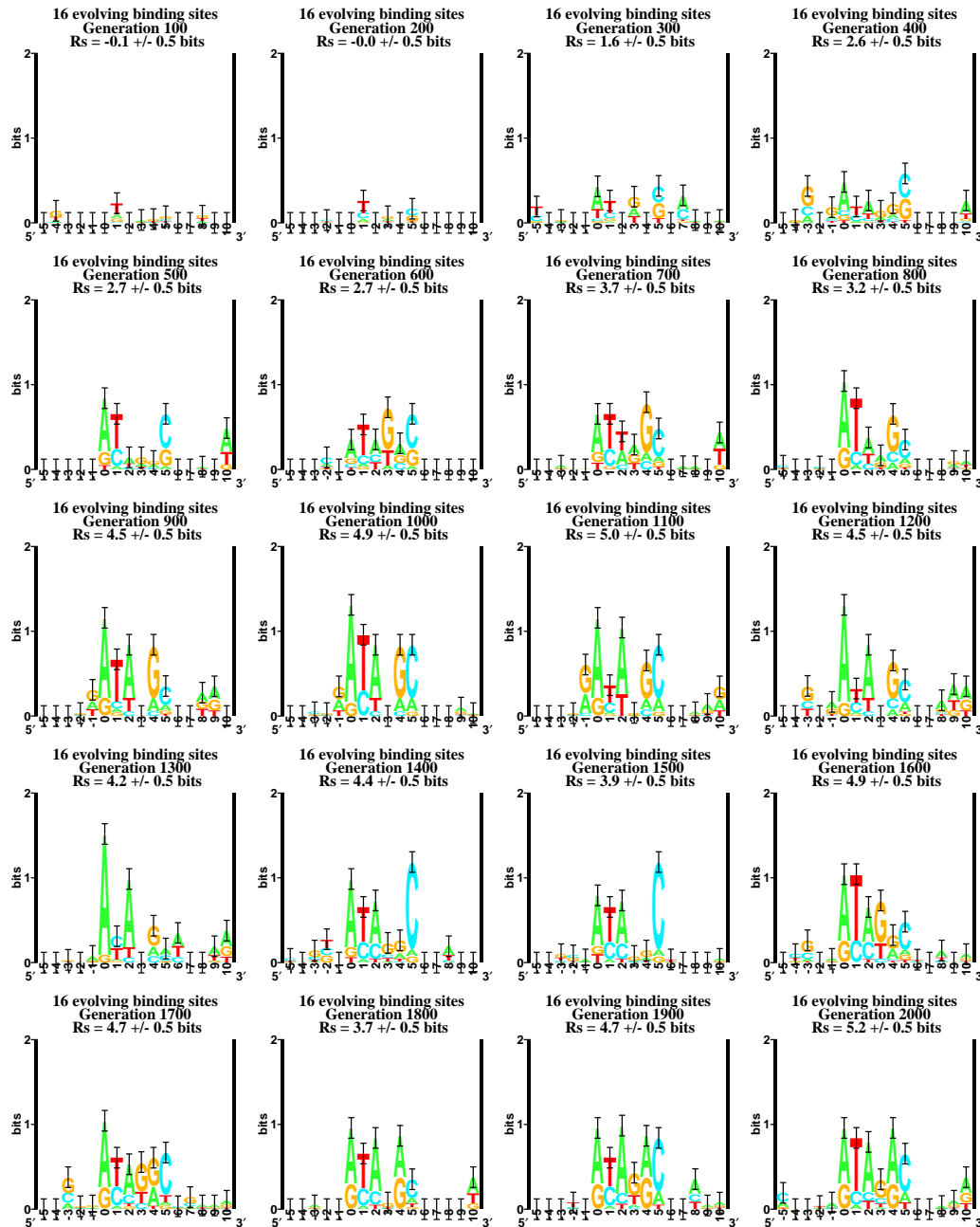- MUTATE all genomes randomly
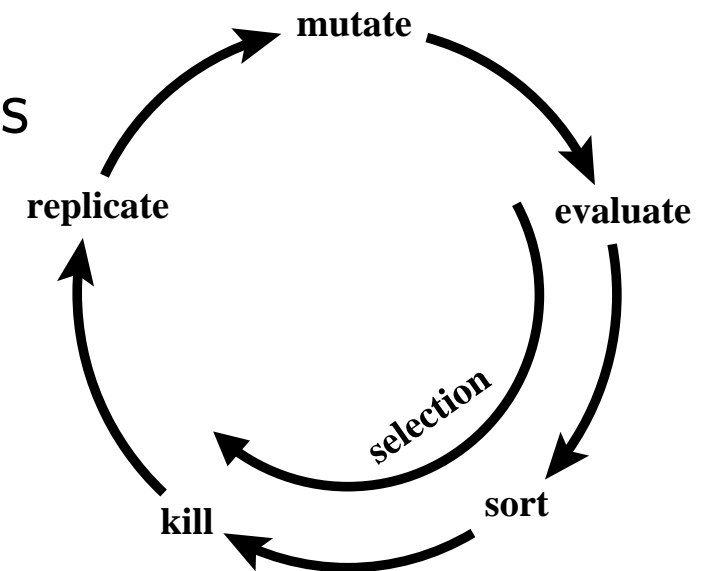
mutate

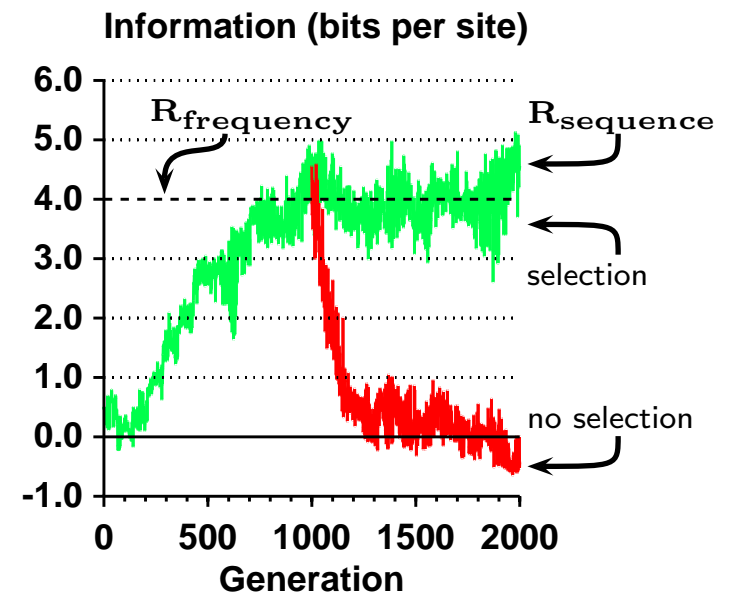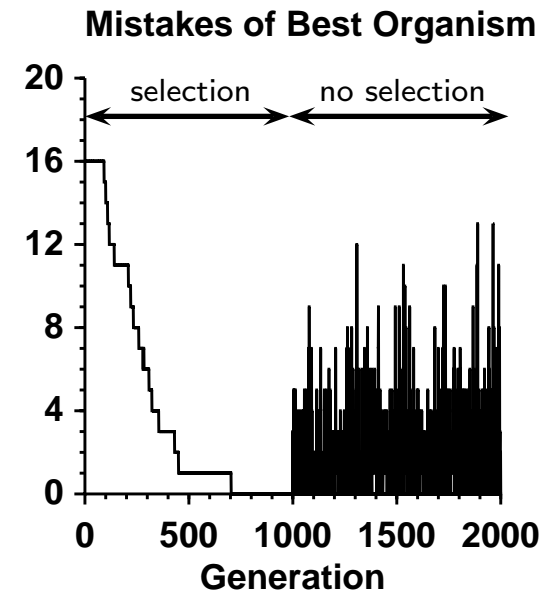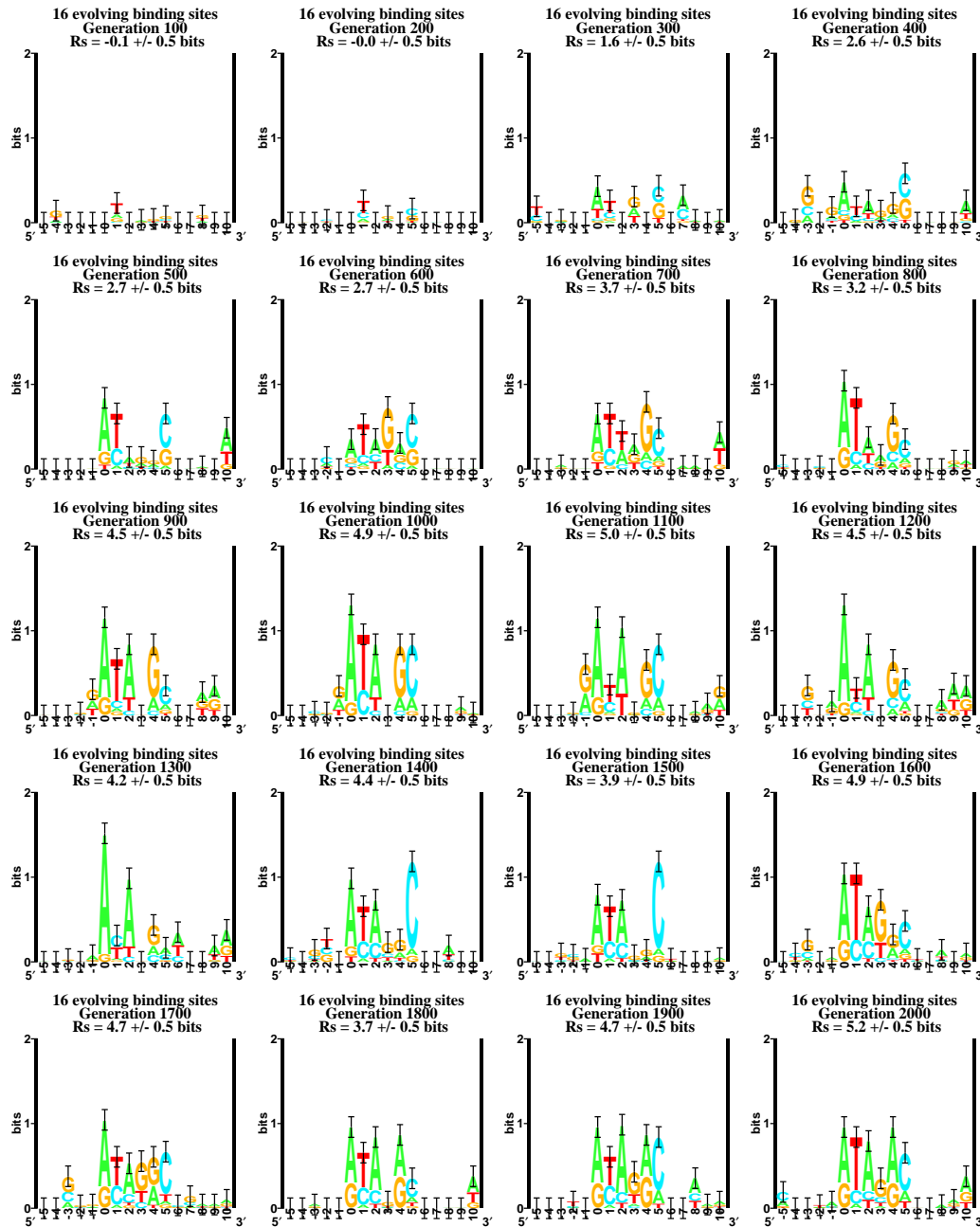replicate   evaluate

selection

kill   sort
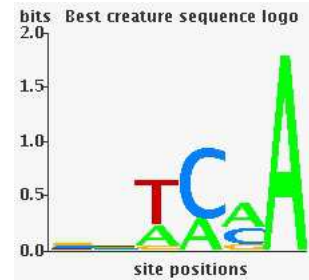
Evolution of Binding Sites
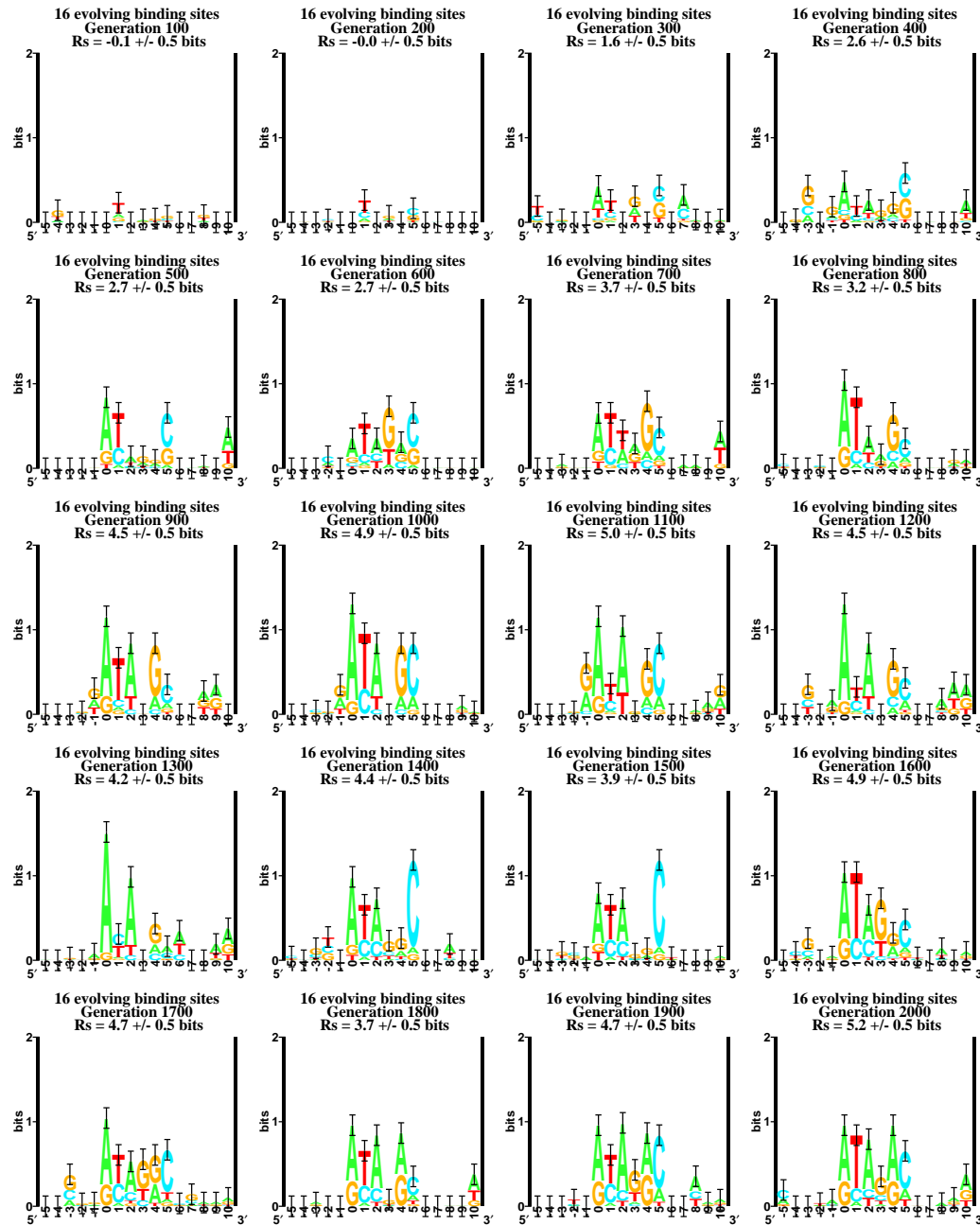
Evolution of Binding Sites

Evolution of Binding Sites

Shannon Information Measure of Binding Site Patterns

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

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

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

$$H_{before}(l) = \log_2 4 \quad \text{(bits per base)} \quad (16)$$
$$\approx 2$$

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

$$H_{before}(l) = \log_2 4 \quad \text{(bits per base)} \quad (16)$$
$$\approx 2$$

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

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

$$\text{(bits per base)}$$

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

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

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

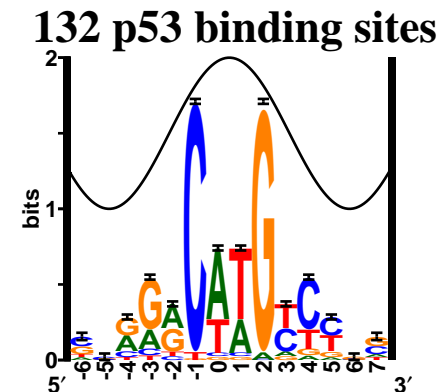$$R_{sequence}(l) = H_{before}(l) - H_{after}(l) \qquad (18)$$
$$\approx 2 - H_{after}(l) \qquad \text{(bits } per\ base)$$

The **total site information** is:

$$R_{sequence} = \sum_{l} (H_{before}(l) - H_{after}(l))$$
$$\approx 2l - H_{after} \qquad \text{(bits } per\ site) (19)$$

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

132 p53 binding sites

Acknowledgements

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

Version

version = 1.20 of evtalk.tex 2012 Mar 15