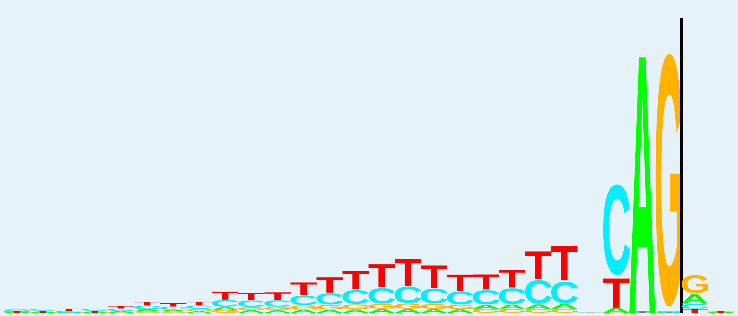
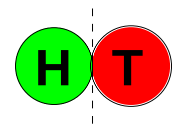
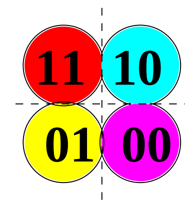
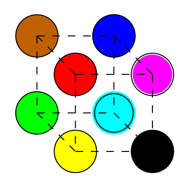
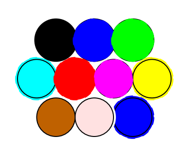
Information Theory in Biology

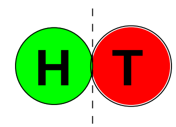
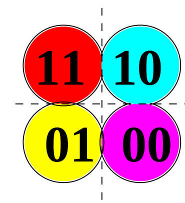
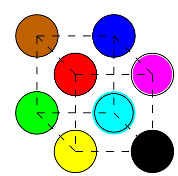
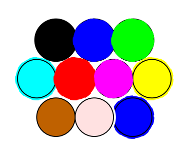
Thomas D. Schneider, Ph.D.

Molecular Information Theory Group
Center for Cancer Research
Gene Regulation and Chromosome Biology Laboratory
National Cancer Institute
Frederick, MD 21702-1201

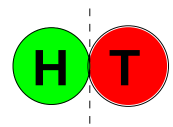
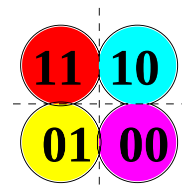
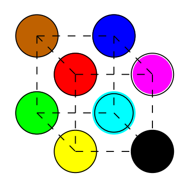
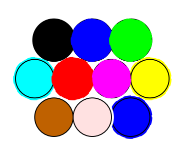


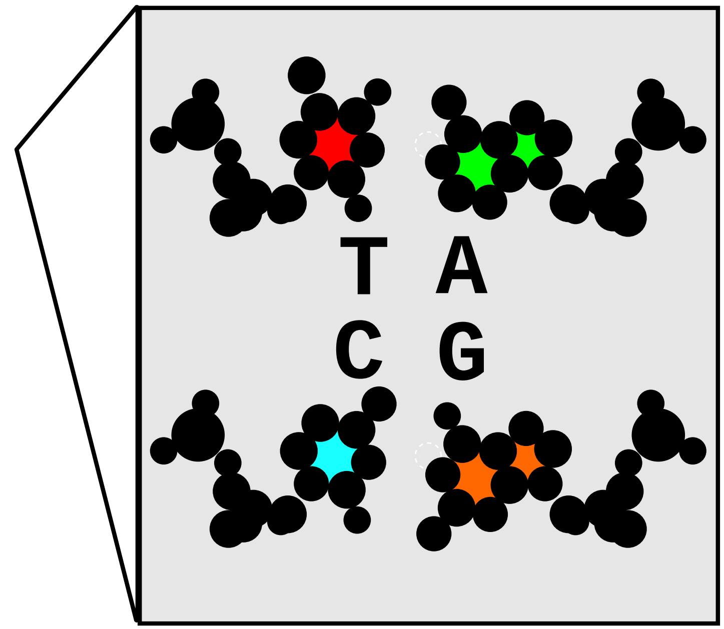
Information Theory: One-Minute Lesson

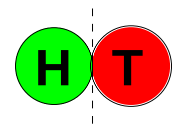
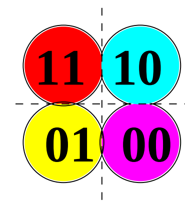
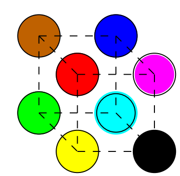
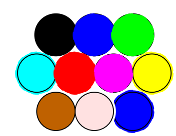
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

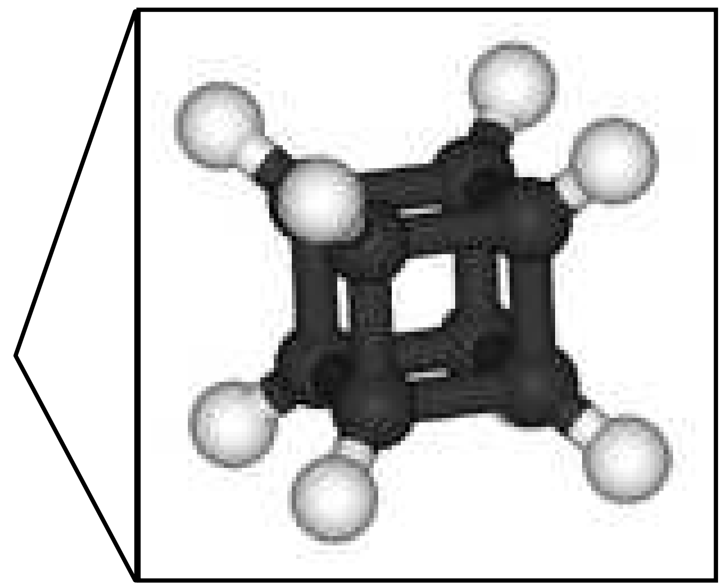
number of symbols	number of bits	example
M	B	
2	1	 
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	

Information Theory: One-Minute Lesson

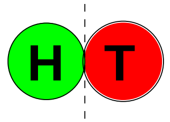
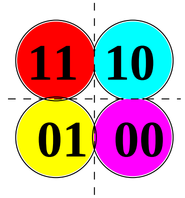
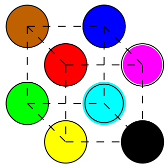
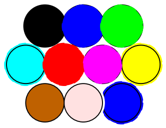
number of symbols	number of bits	example
M	B	
2	1	
4	2	
8	3	
$M=2^B$	$B=\log_2 M$	



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

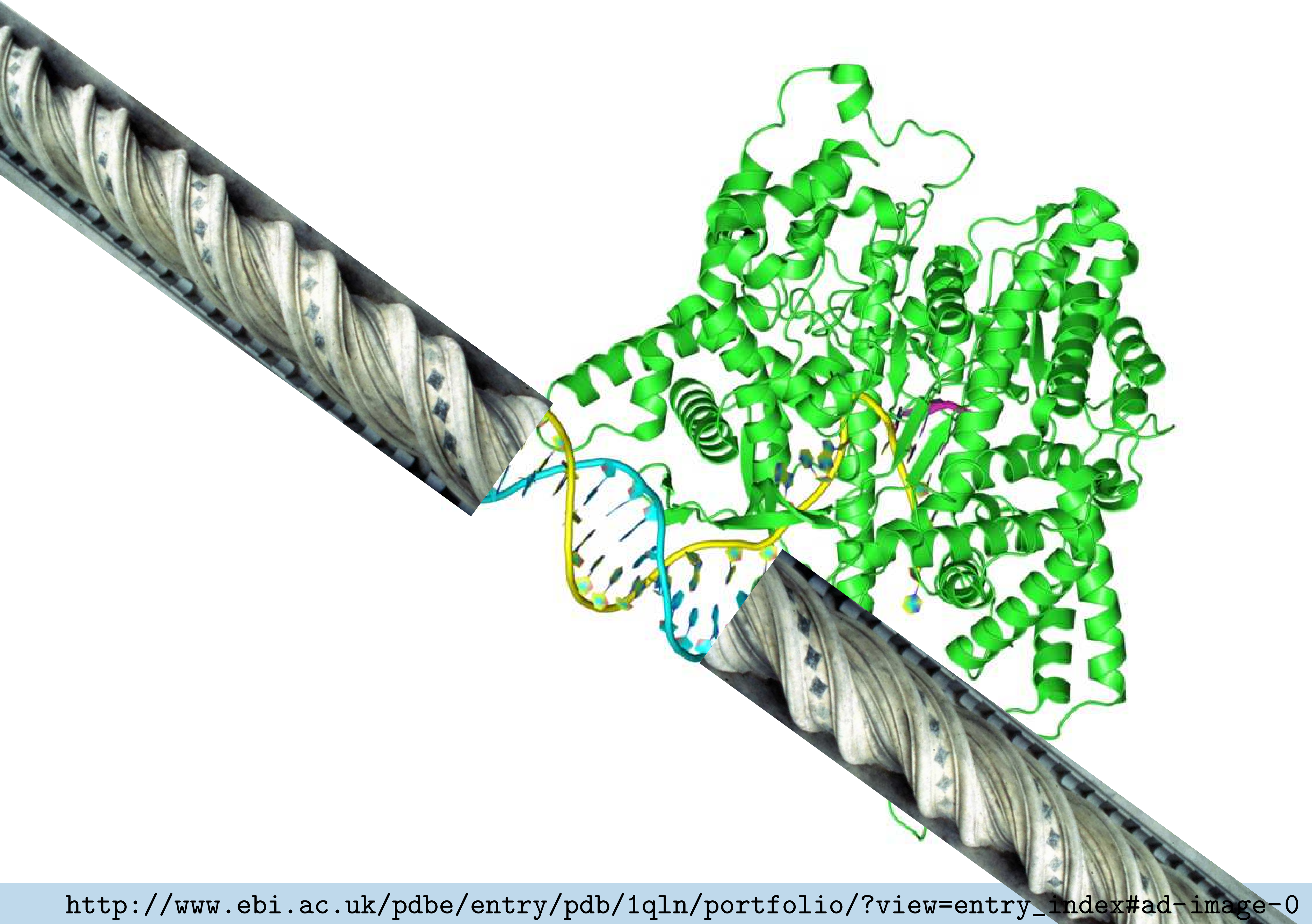


Information Theory: One-Minute Lesson

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

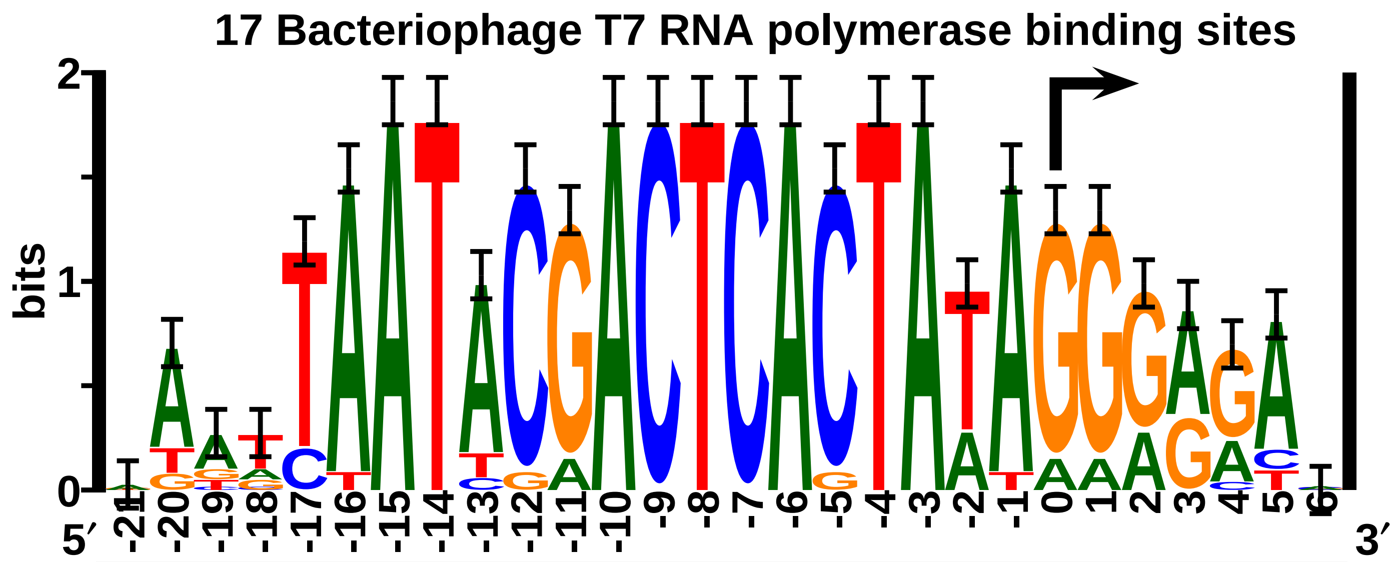
El Duomo, Florence, Italy

T7 RNA polymerase + DNA



http://www.ebi.ac.uk/pdbe/entry/pdb/1qln/portfolio/?view=entry_index#ad-image-0

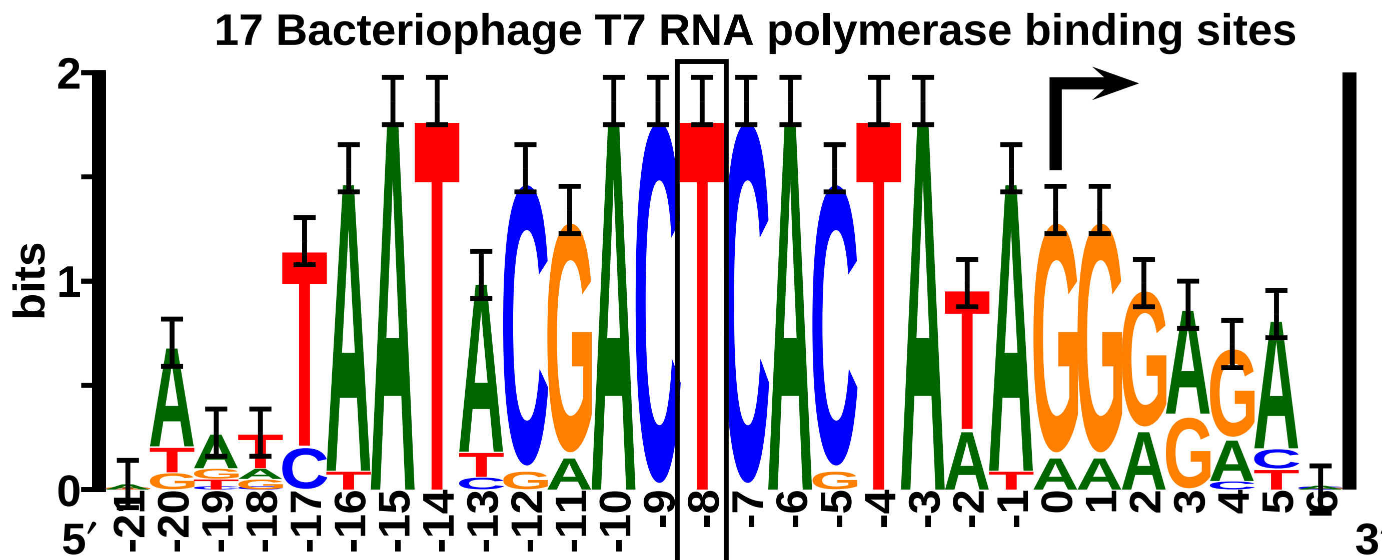
Sequence Logo



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



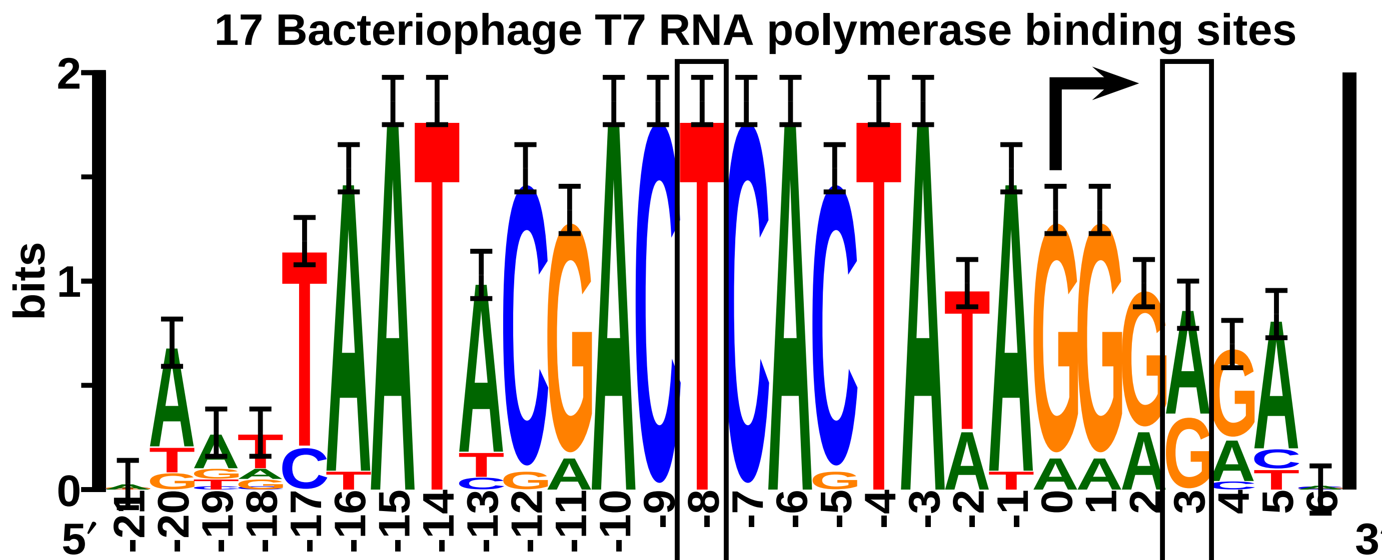
Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

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

6 of 17 sites

2 bits/base

Sequence Logo

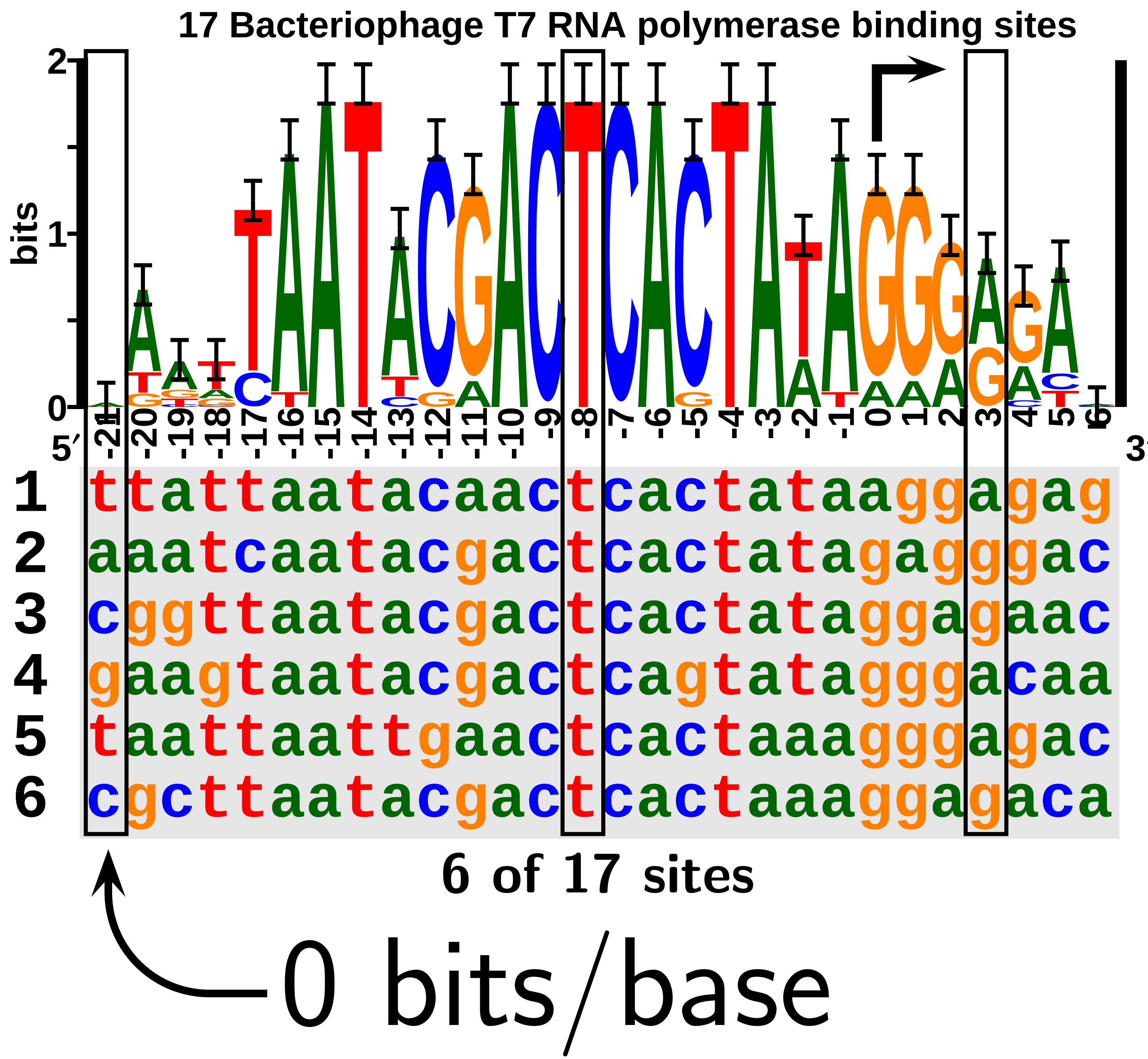


Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

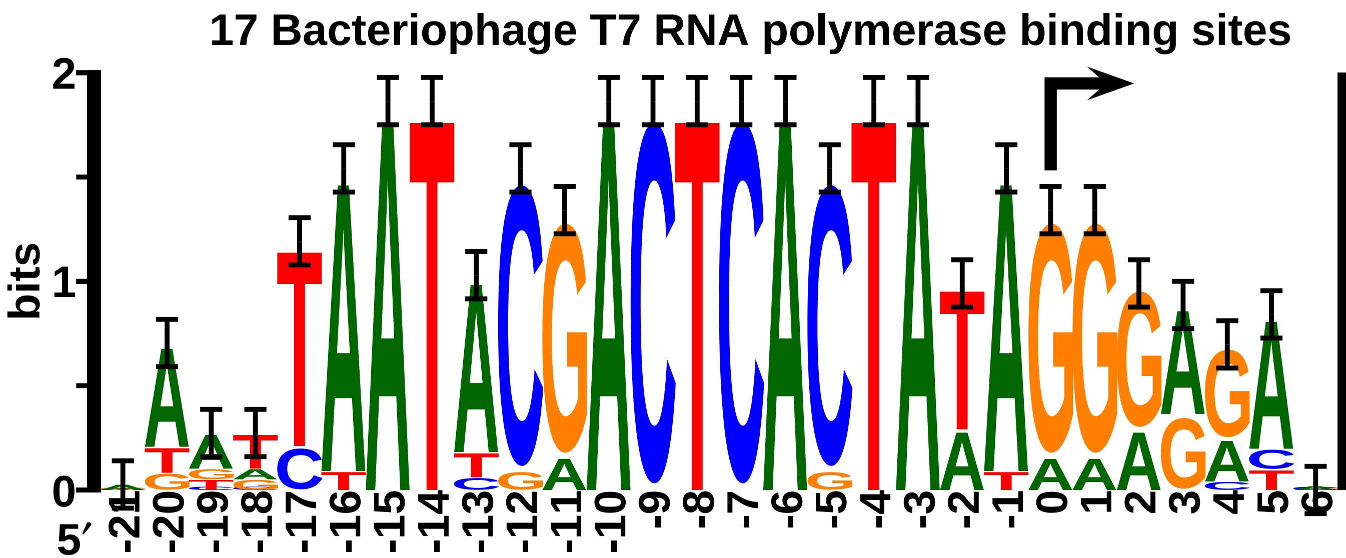
```
1 ttattaatacaactcactataaggagag
2 aaatcaatacgactcactatagaggac
3 cggttaatacgactcactataggagaac
4 gaagtaatacgactcagtatagggacaa
5 taattaattgaactcactaaagggagac
6 cgcttaatacgactcactaaaggagaca
```

6 of 17 sites

1 bit/base



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

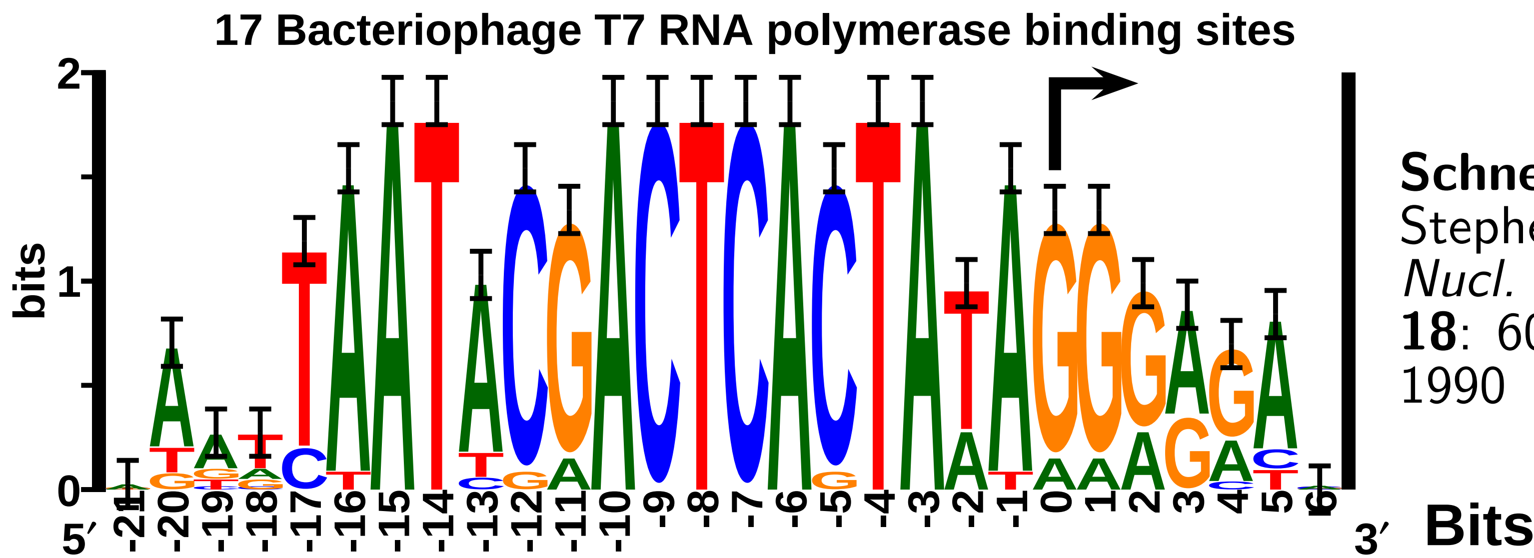


Schneider & Stephens
Nucl. Acids Res.
18: 6097-6100
1990

	5′ … 3′	Bits
1	ttattaatacaactcactataaggagag	33.3
2	aaatcaatacgactcactatagagggac	37.4
3	cggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaagggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1

Individual Information

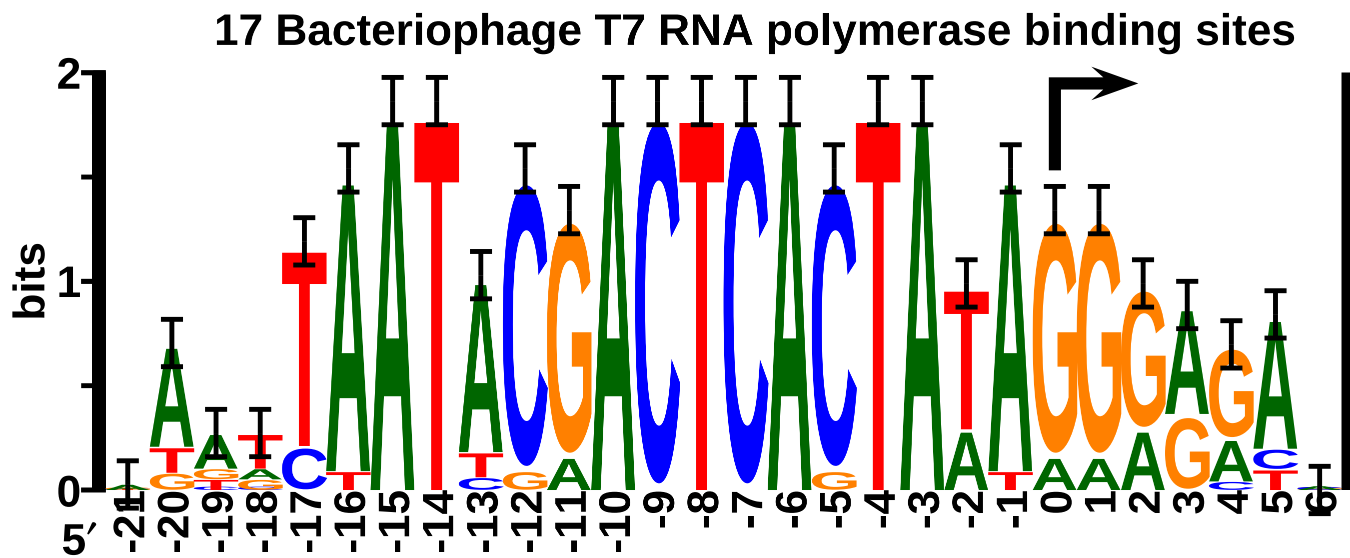
Sequence Logo and Sequence Walker



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

		Bits
1	ttattaatacaactcactataaggagag	33.3
2	aaatcaatacgactcactatagagggac	37.4
3	cggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaagggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1

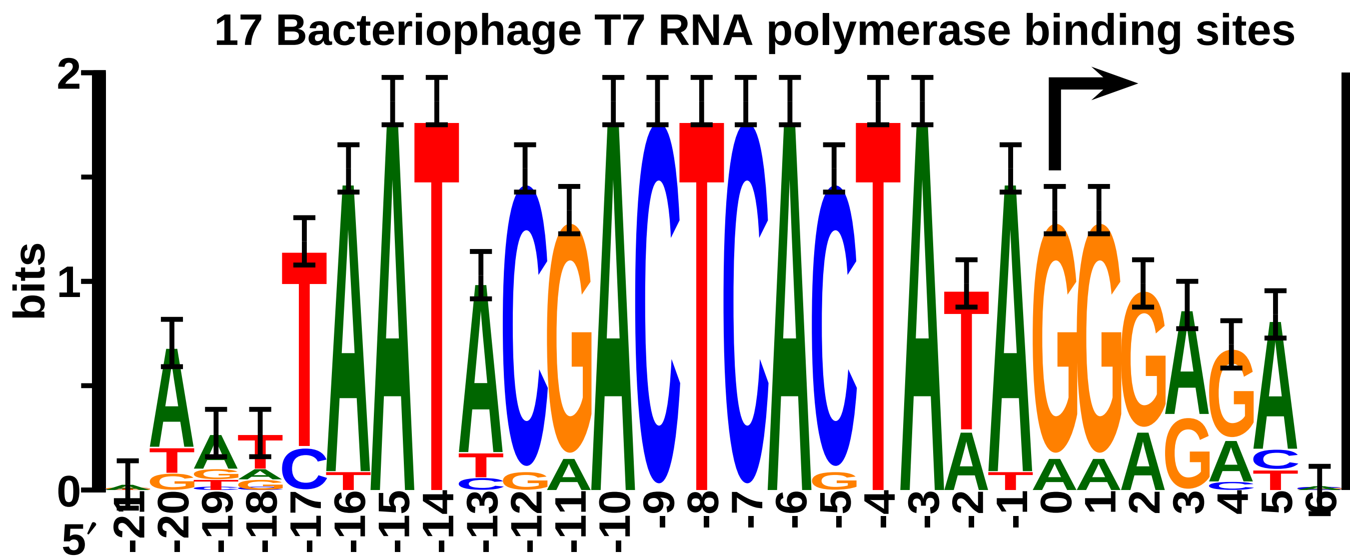
Sequence Walker Patent 5,867,402



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

		Bits
1	ttattaatacaactcactataaggagag	33.3
2	aaatcaatacgactcactatagagggac	37.4
3	cggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaagggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1

Rsequence is the average: 35.0 ± 0.6 bits



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

		Bits
1	ttattaatacaactcactataaggagag	33.3
2	aaatcaatacgactcactatagagggac	37.4
3	cggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaagggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1

Rsequence is the average: 35.0 ± 0.6 bits
$=$ "area under the logo"

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

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

γ = number of binding sites on genome

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

Information required to find a set of binding sites

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

$\gamma =$ number of binding sites on genome

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

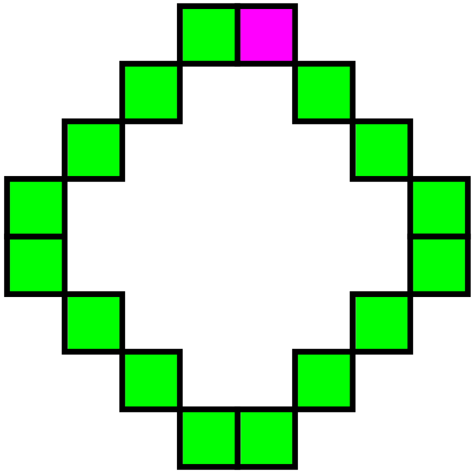
Information required to find a set of binding sites

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

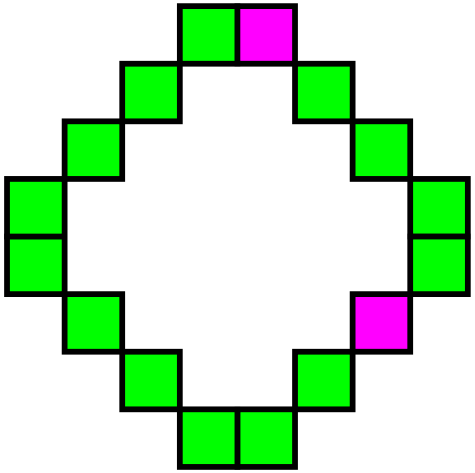
$\gamma =$ number of binding sites on genome

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

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

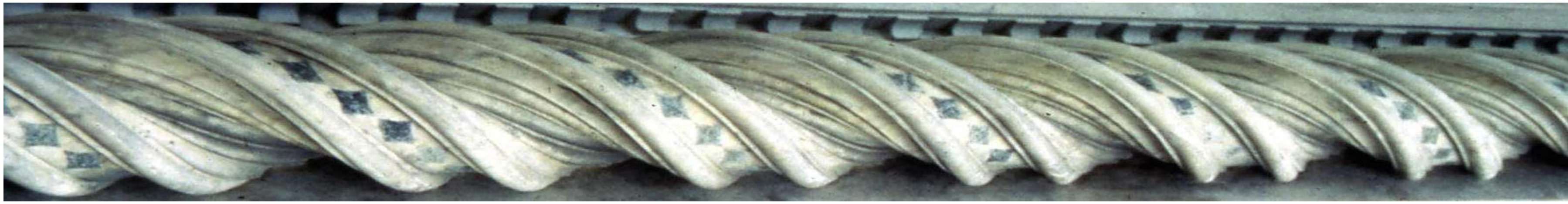


16 positions
1 site
$\log_2 16/1 = 4$ bits



16 positions
2 sites
$\log_2 16/2 = 3$ bits

DNA



Copy DNA (transcription)

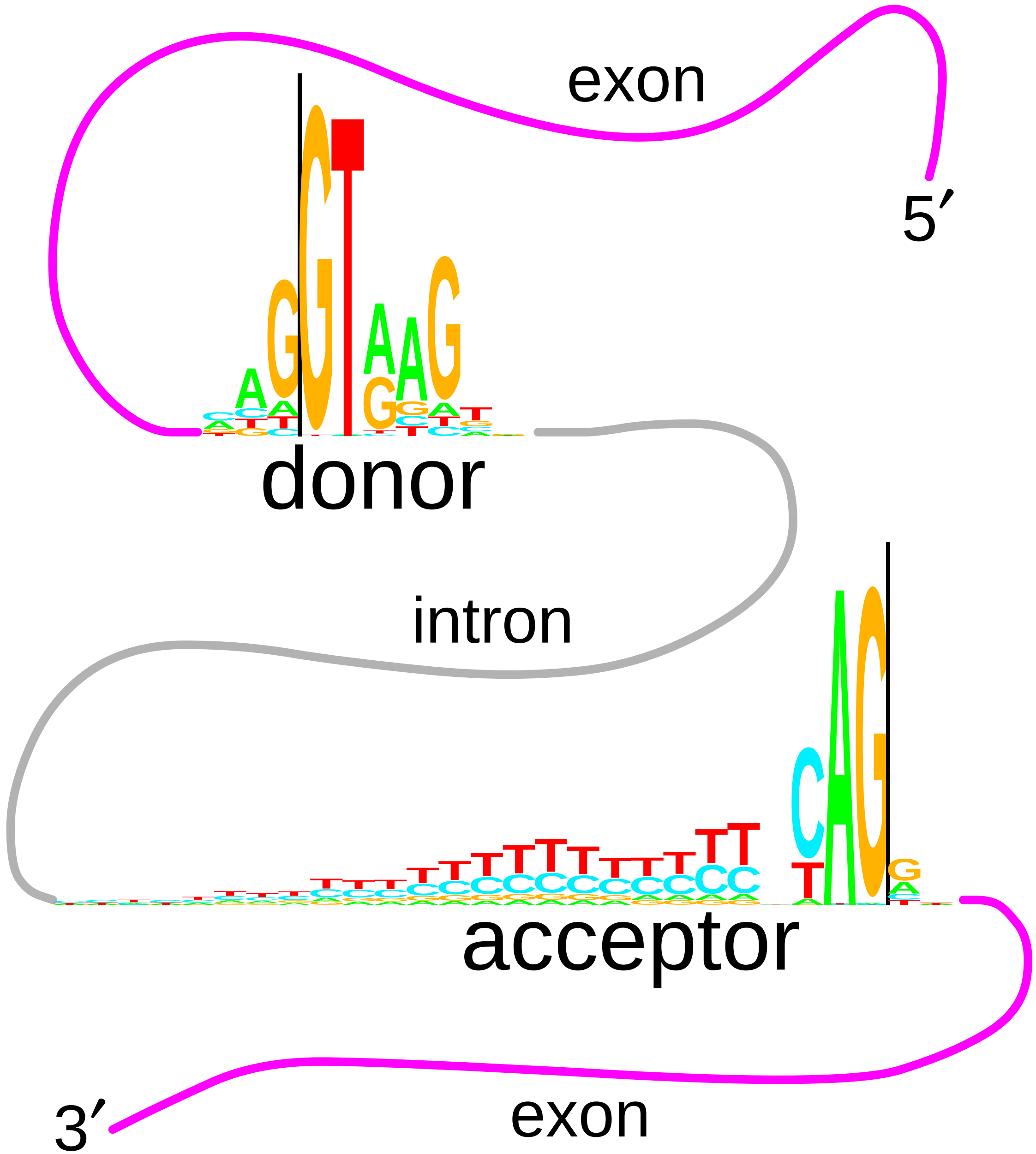
donor acceptor

RNA exon intron exon intron exon

RNA Splicing

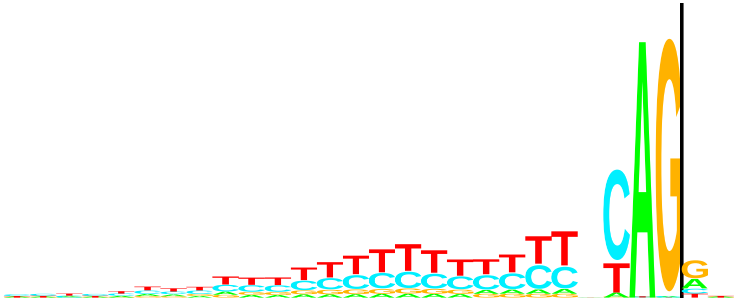
Spliced RNA exon exon exon

Donor and acceptor logos



Rsequence and Rfrequency for Splice Acceptors

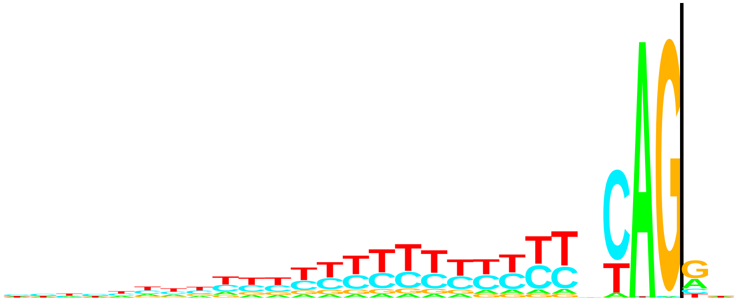
$\boxed{R_{sequence}}$



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

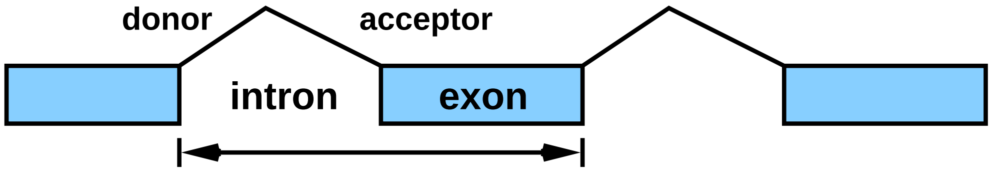
Rsequence and Rfrequency for Splice Acceptors

$\boxed{R_{sequence}}$



- **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**

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

$\boxed{\text{Hypothesis:}}$

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 $= R_{sequence}$ (bits)	Information needed to Locate Site in Genome $= R_{frequency}$ (bits)	$\frac{\textbf{Pattern Info}}{\textbf{Location Info}} = \frac{R_{sequence}}{R_{frequency}}$
Spliceosome acceptor[2]	$\mathbf{9.35 \pm 0.12}$	**9.66**	$\mathbf{0.97 \pm 0.01}$
Spliceosome donor	$\mathbf{7.92 \pm 0.09}$	**9.66**	$\mathbf{0.82 \pm 0.01}$
Ribosome	**11.0**	**10.6**	**1.0**
λ cI/cro	$\mathbf{17.7 \pm 1.6}$	**19.3**	$\mathbf{0.9 \pm 0.1}$
LexA	$\mathbf{21.5 \pm 1.7}$	**18.4**	$\mathbf{1.2 \pm 0.1}$
TrpR	$\mathbf{23.4 \pm 1.9}$	**20.3**	$\mathbf{1.2 \pm 0.1}$
LacI	$\mathbf{19.2 \pm 2.8}$	**21.9**	$\mathbf{0.9 \pm 0.1}$
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**	$\sim \mathbf{3}$	$\sim \mathbf{1}$
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.

$R_{sequence}$ 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 $R_{frequency}$?

- $R_{frequency}$ 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)

$\left.\right\}$ $R_{frequency}$ 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)

$R_{frequency}$ 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 -3	**A** -2	**G** -1	**G** 0	**T** 1	**C** 2	**T** 3	**G** 4	**C** 5	**A** 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

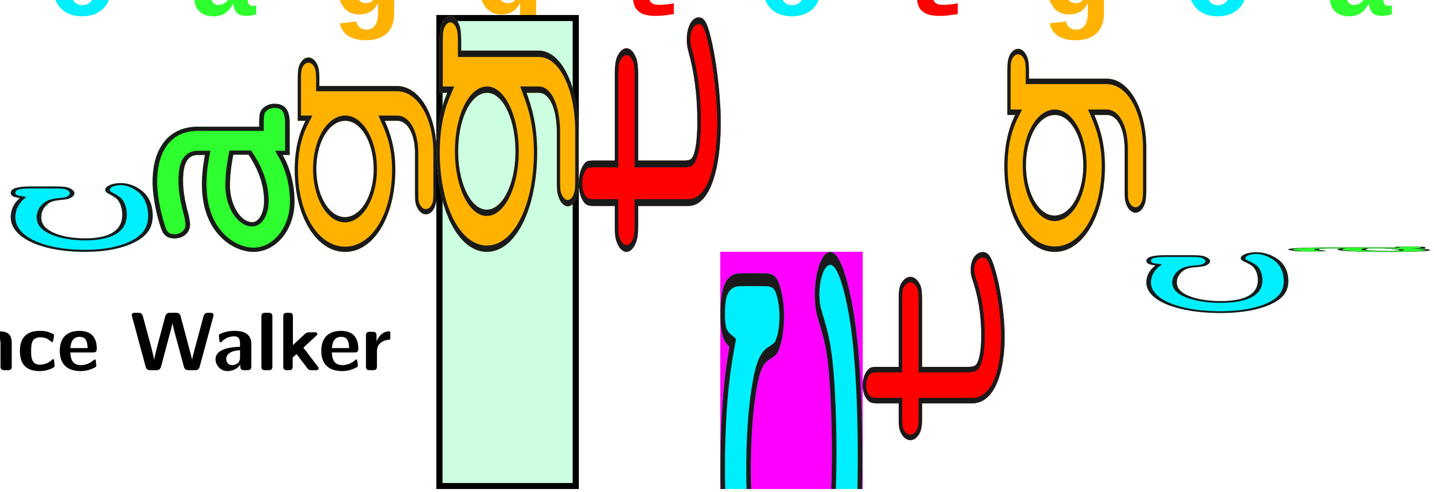
Sequence matrix, $s(b, l, j)$ **for sequence** j

base b	position l									
	C -3	A -2	G -1	G 0	T 1	C 2	T 3	G 4	C 5	A 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′



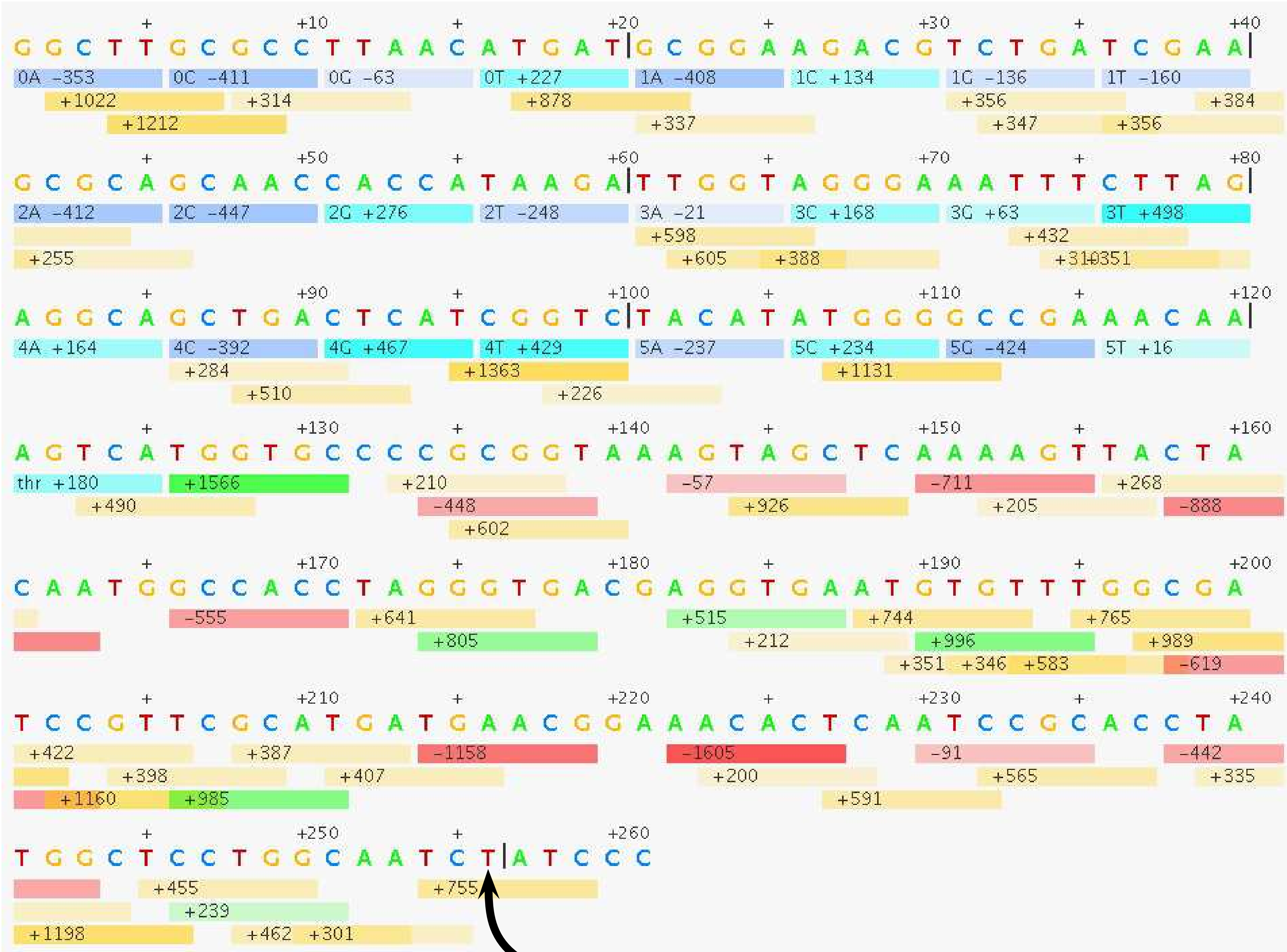
Sequence Walker

Unevolved Ev Creature

Unevolved Ev Creature

"blue" gene weight matrix: 6 bp wide

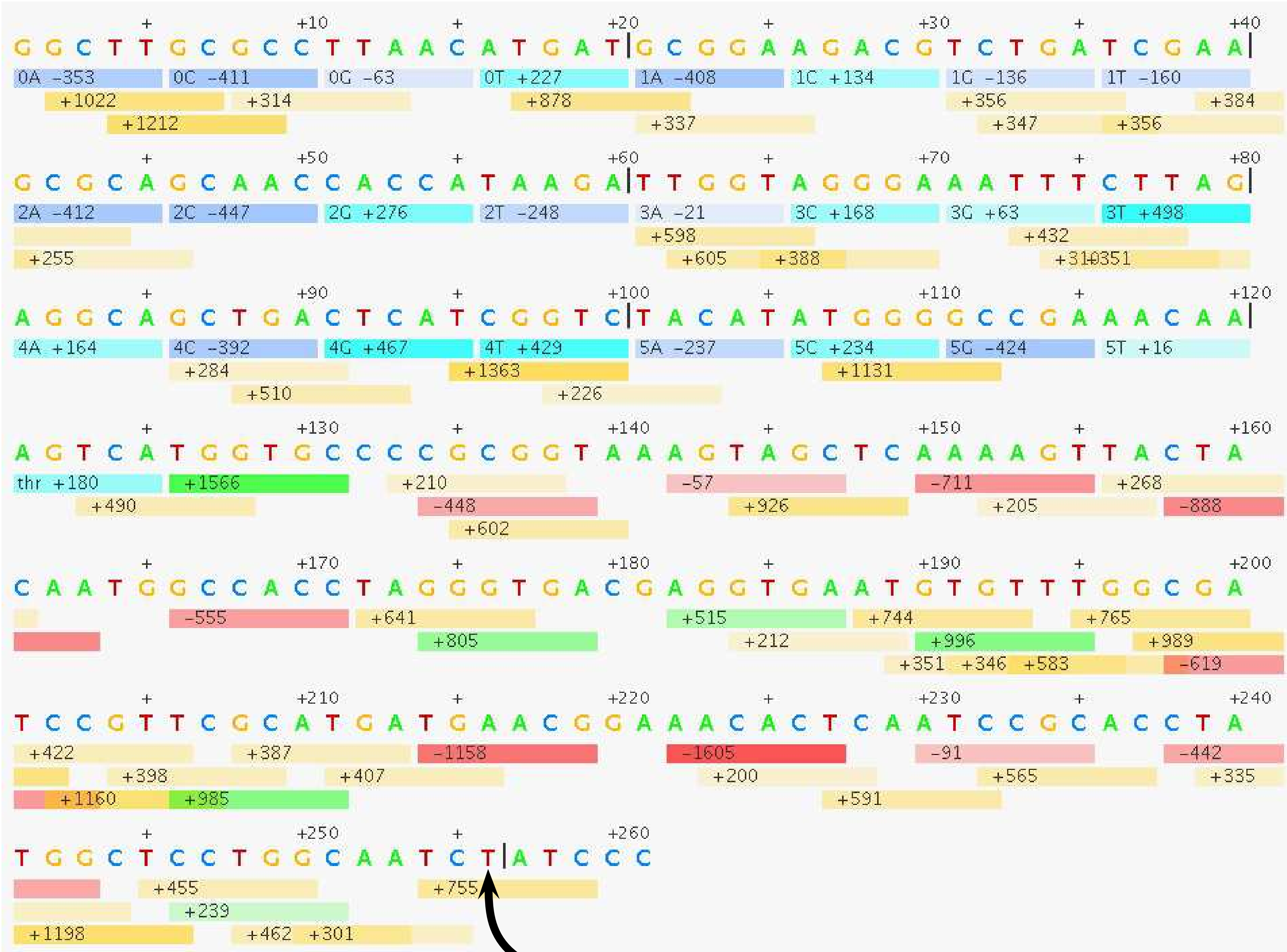
Unevolved Ev Creature



"blue" gene weight matrix: 6 bp wide

Genome positions available $G = 256$ bases

Unevolved Ev Creature



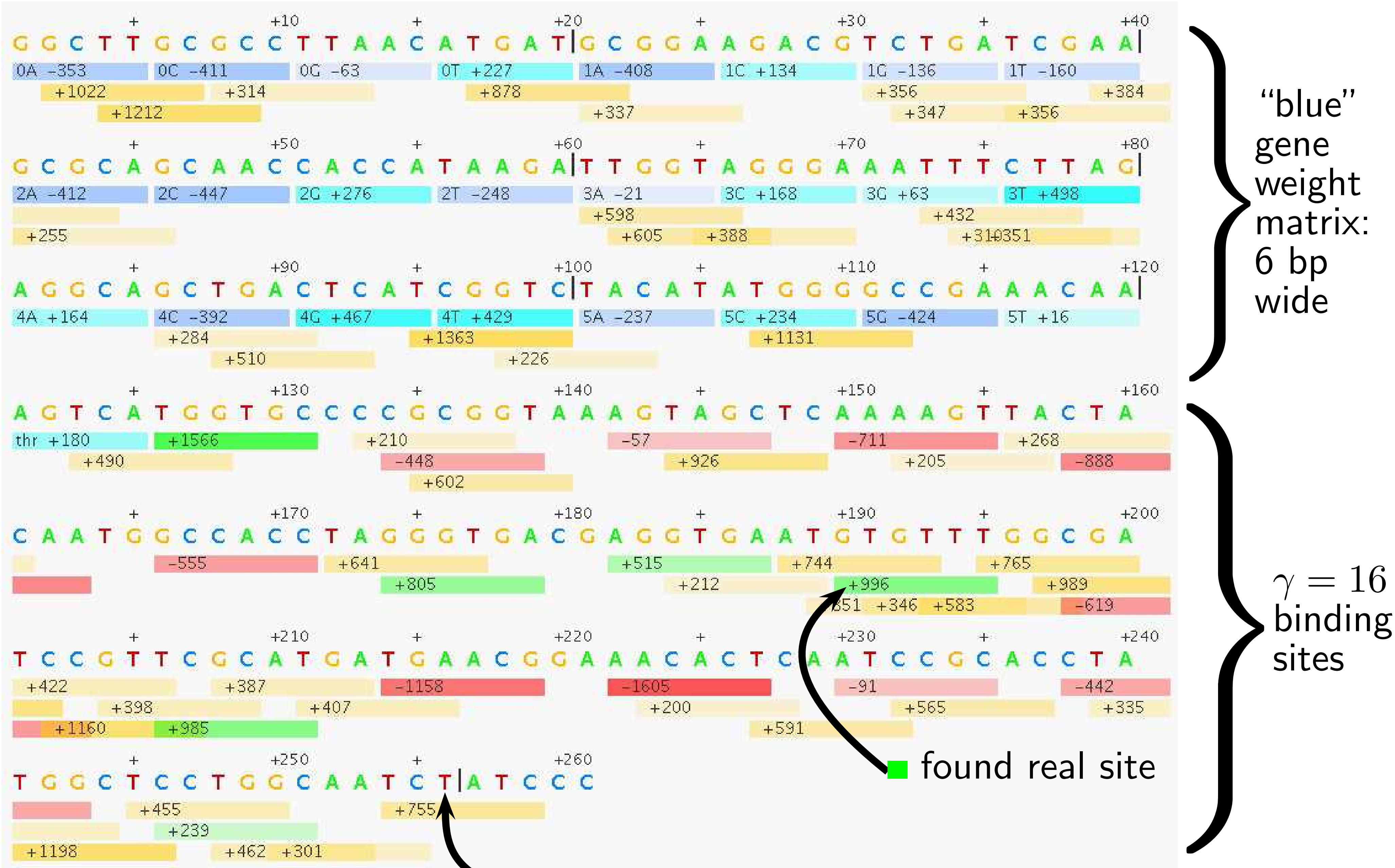
"blue" gene weight matrix: 6 bp wide

$\gamma = 16$ binding sites

Genome positions available $G = 256$ bases

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

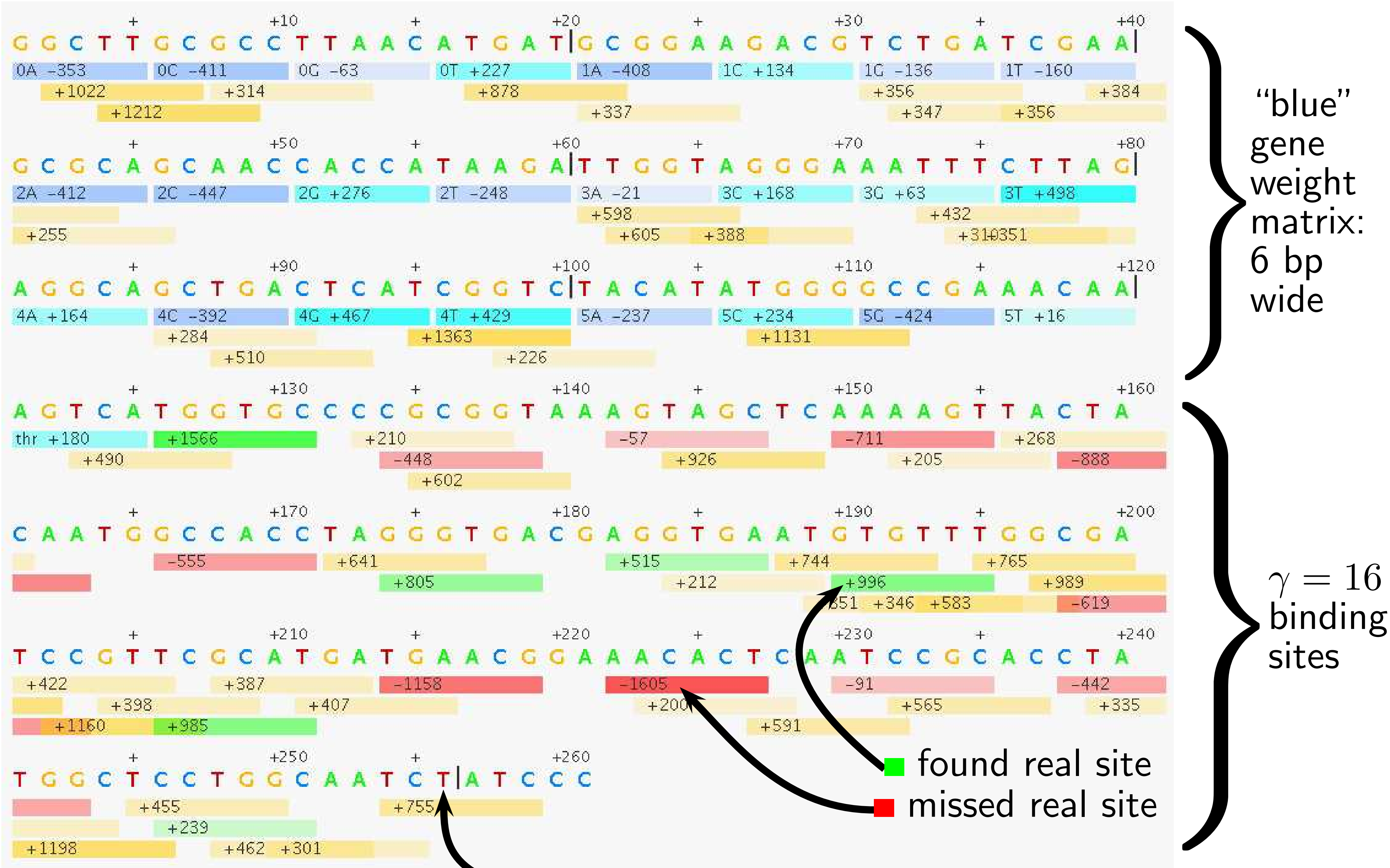
Unevolved Ev Creature



Genome positions available $G = 256$ bases

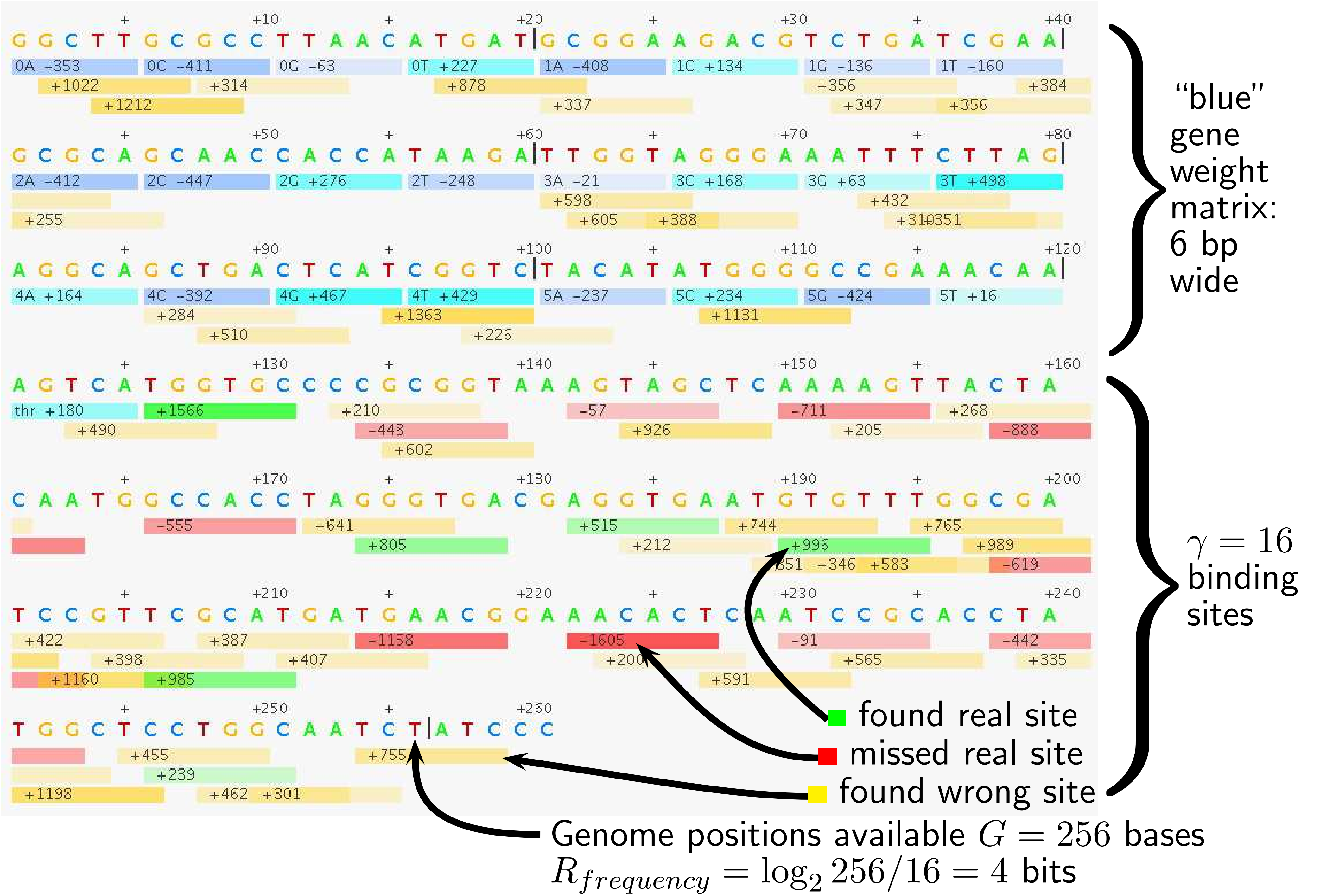
$R_{frequency} = \log_2 256/16 = 4$ bits

Unevolved Ev Creature

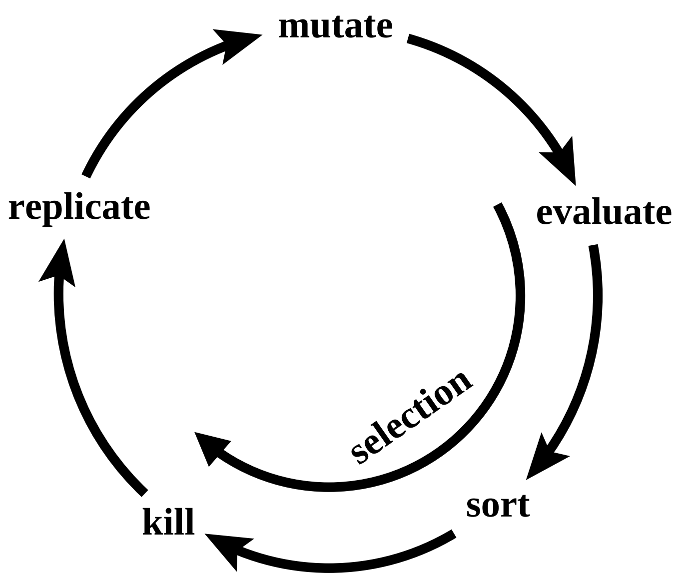


Genome positions available $G = 256$ bases
$R_{frequency} = \log_2 256/16 = 4$ bits

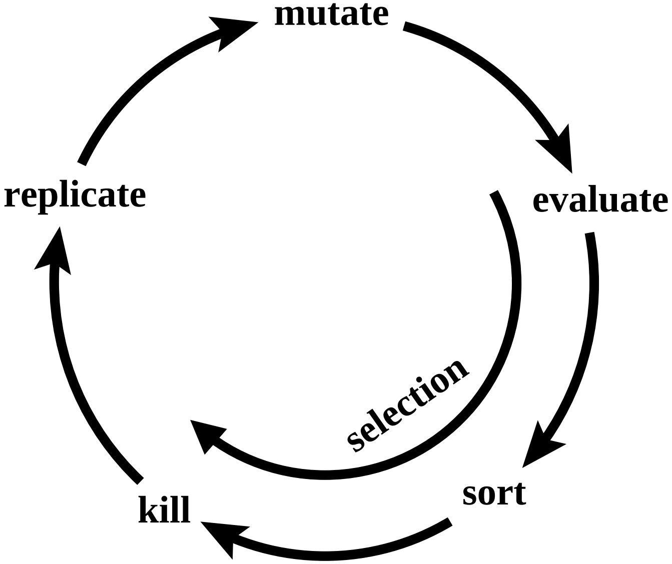
Unevolved Ev Creature



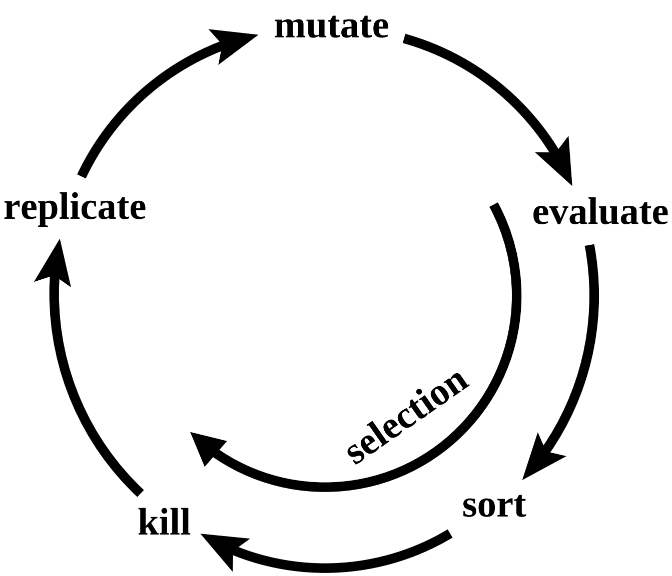
- EVALUATE each creature



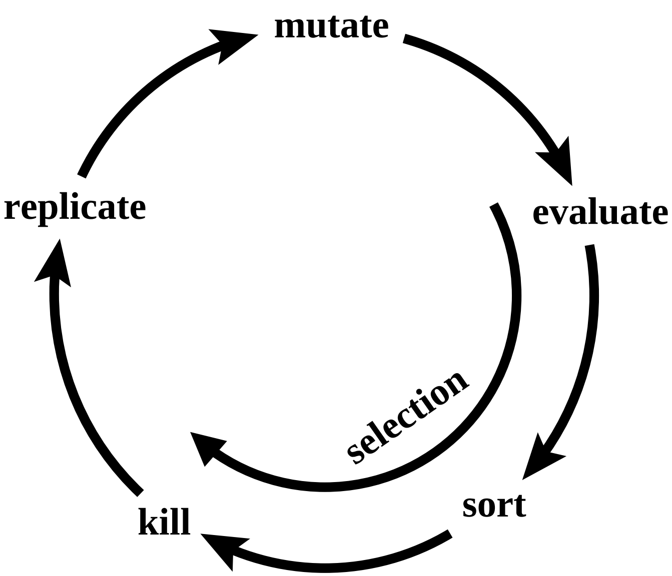
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix



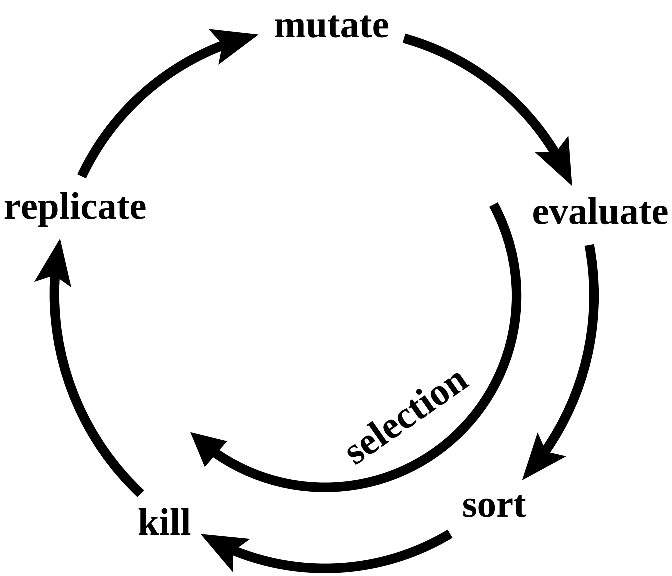
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome



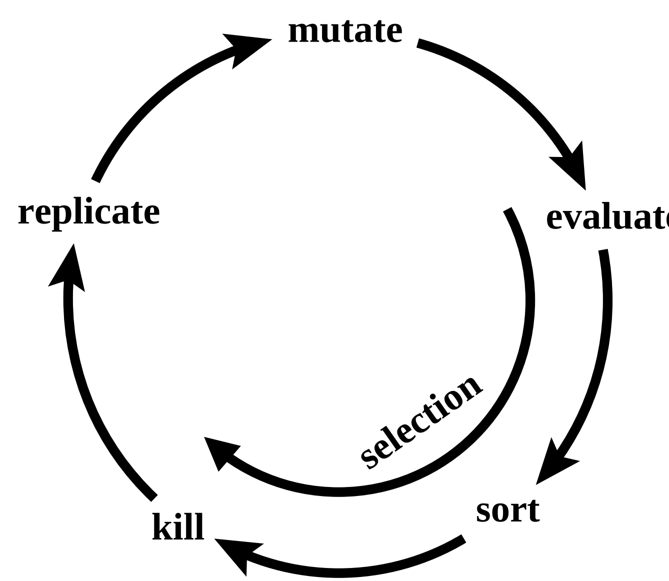
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:



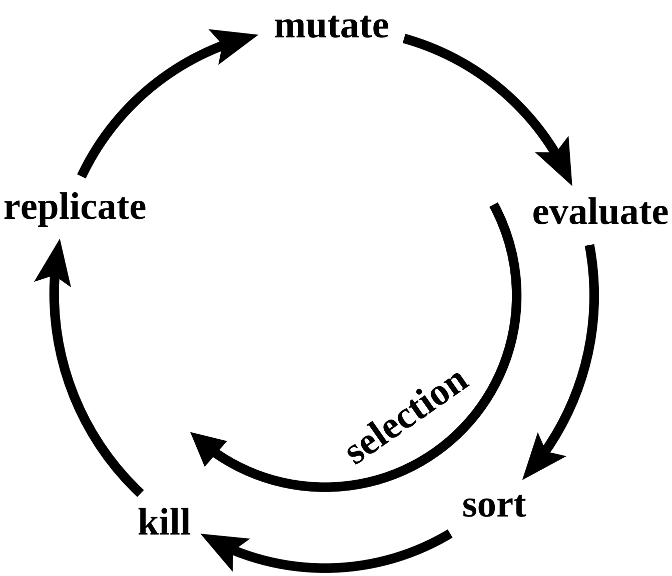
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    - missing a site at a right place



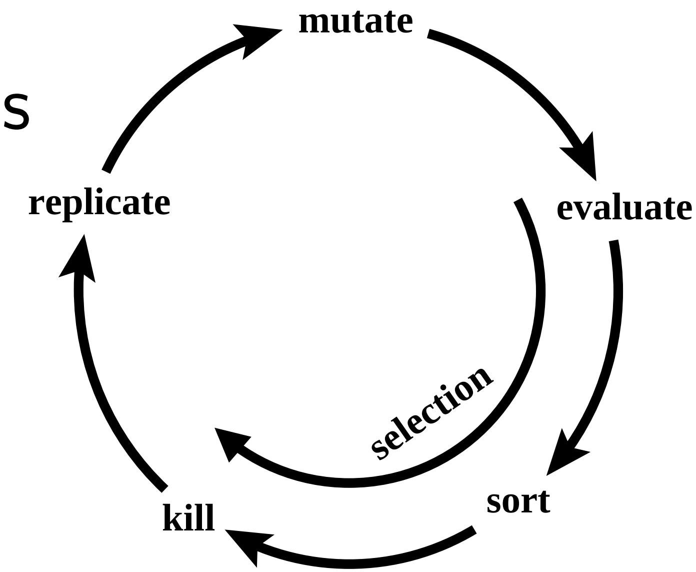
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    - ■ missing a site at a right place
    - □ finding a site at a wrong place



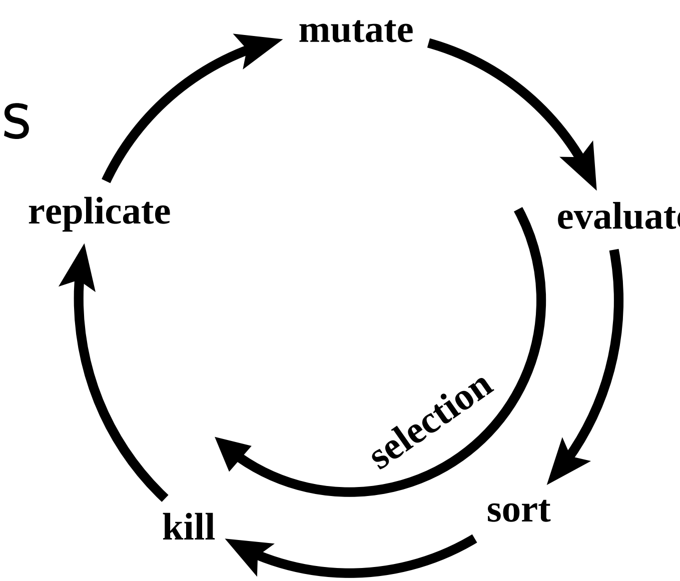
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    - missing a site at a right place
    - finding a site at a wrong place
  - Sort the creatures by their mistakes



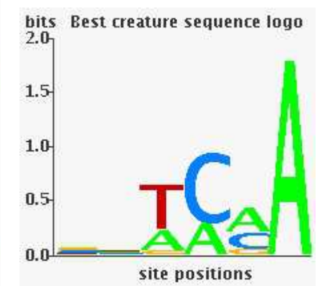
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    - missing a site at a right place
    - finding a site at a wrong place
  - Sort the creatures by their mistakes
- REPLICATE: the best creatures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



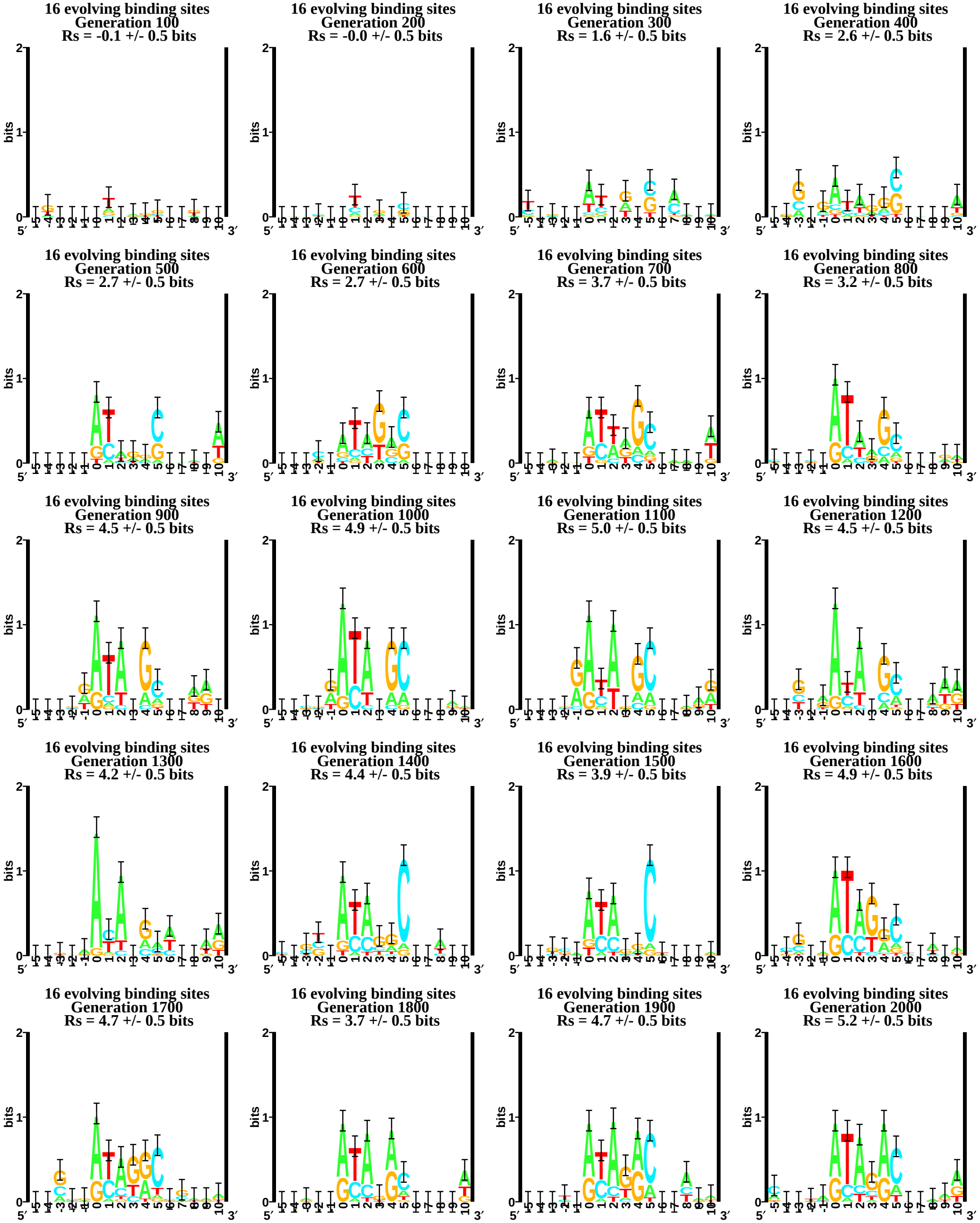
- EVALUATE each creature
  - translate the recognizer gene into a weight matrix
  - scan the weight matrix across the genome
  - count the number of mistakes:
    - missing a site at a right place
    - finding a site at a wrong place
  - Sort the creatures by their mistakes
- REPLICATE: the best creatures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



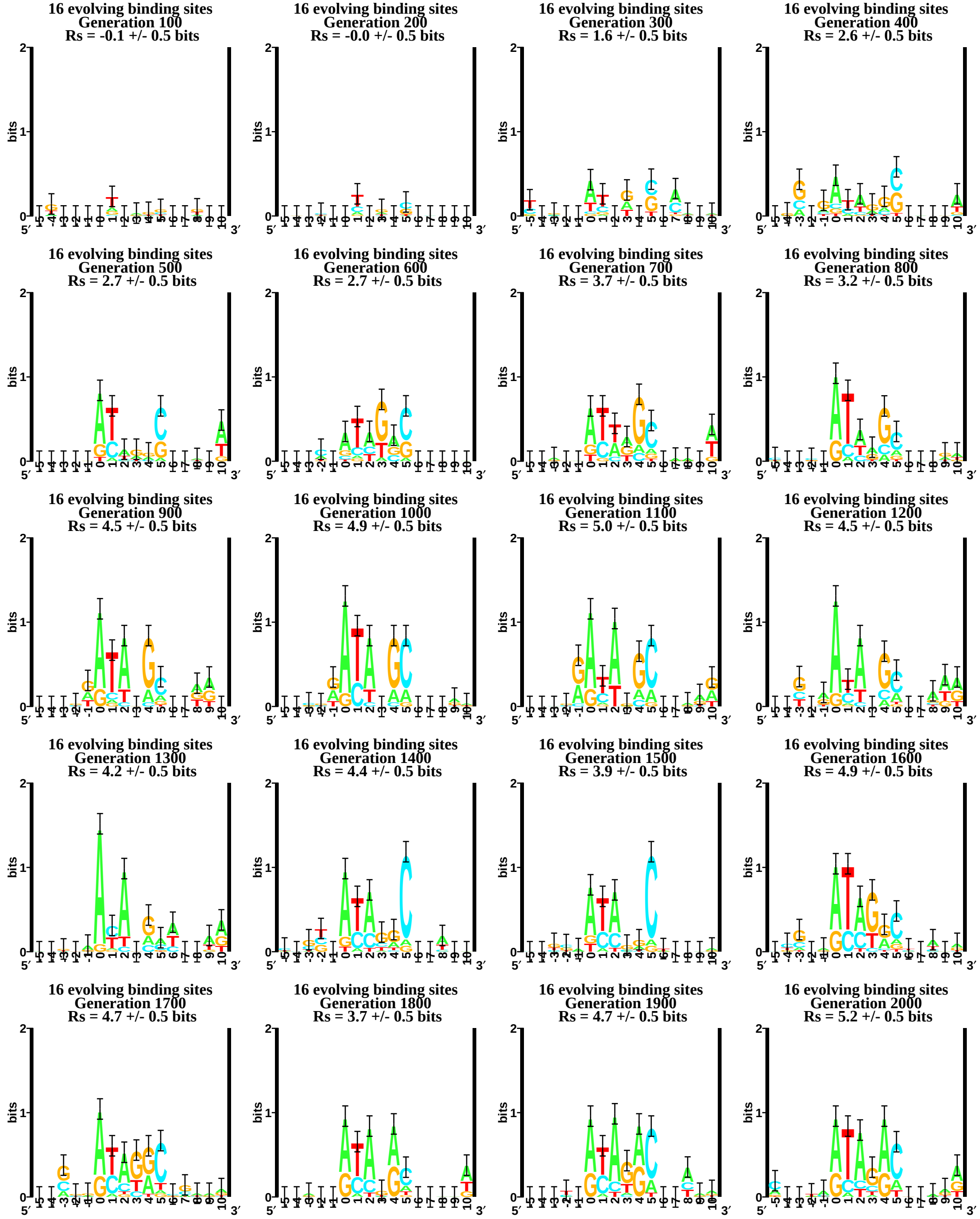
Evolved Ev Creature

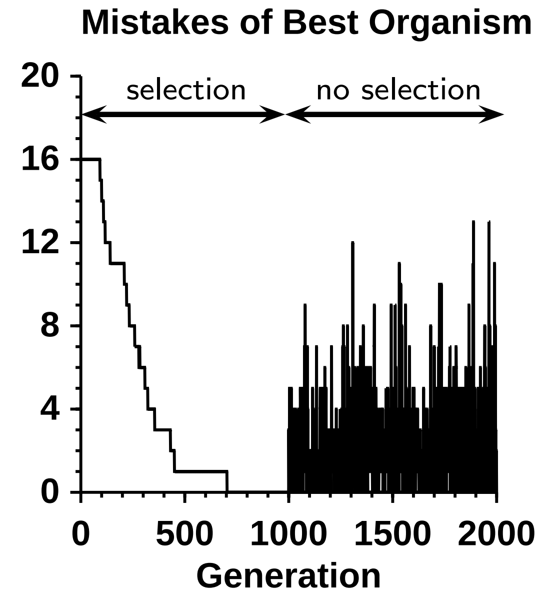


Evolution of Binding Sites

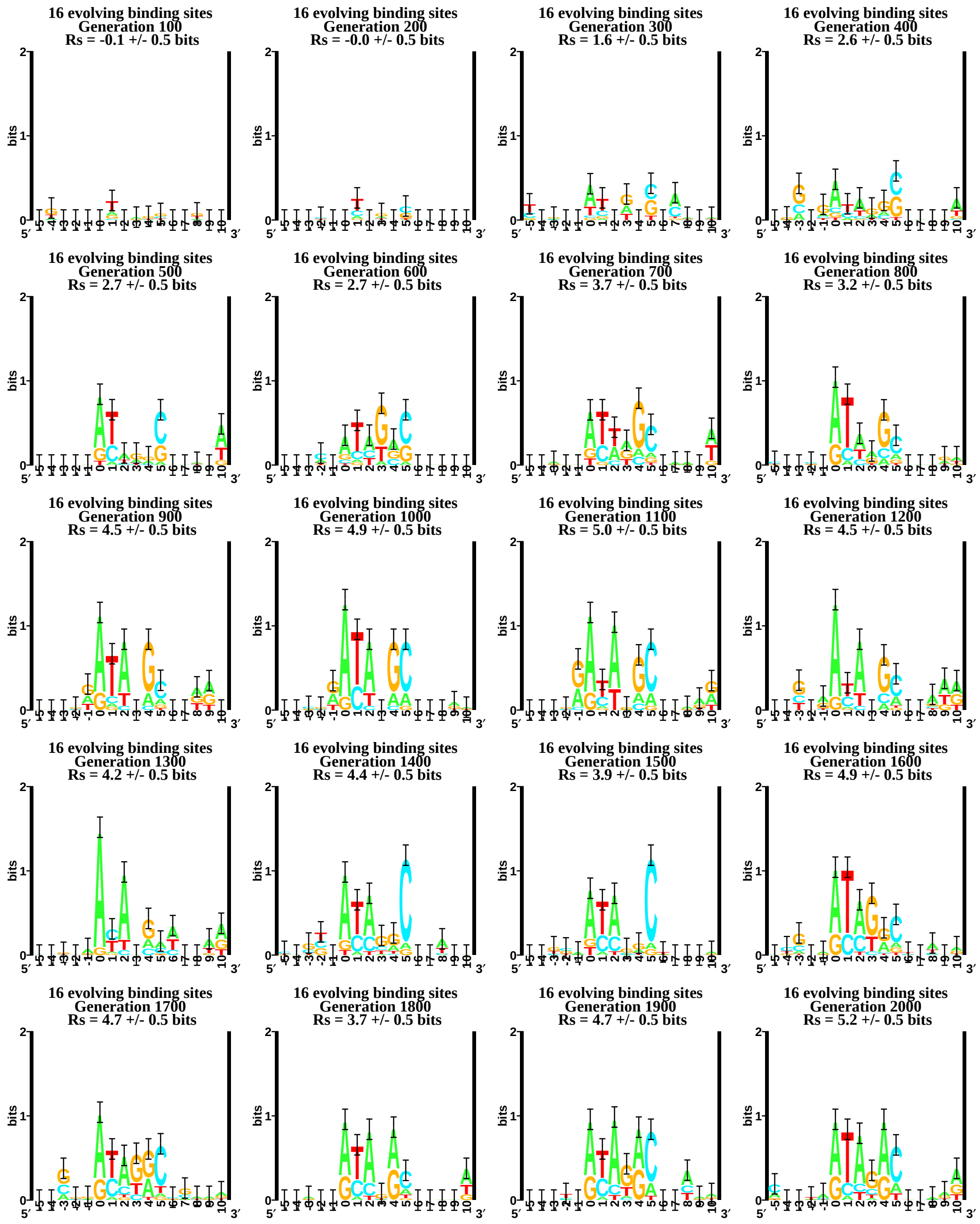


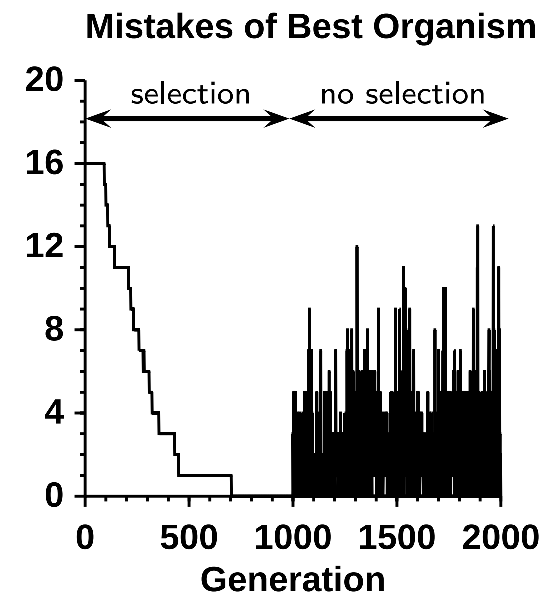
Evolution of Binding Sites

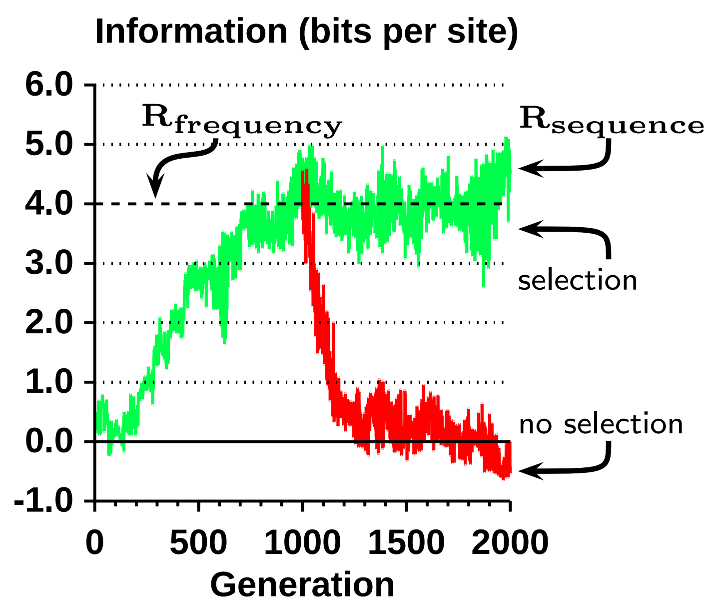




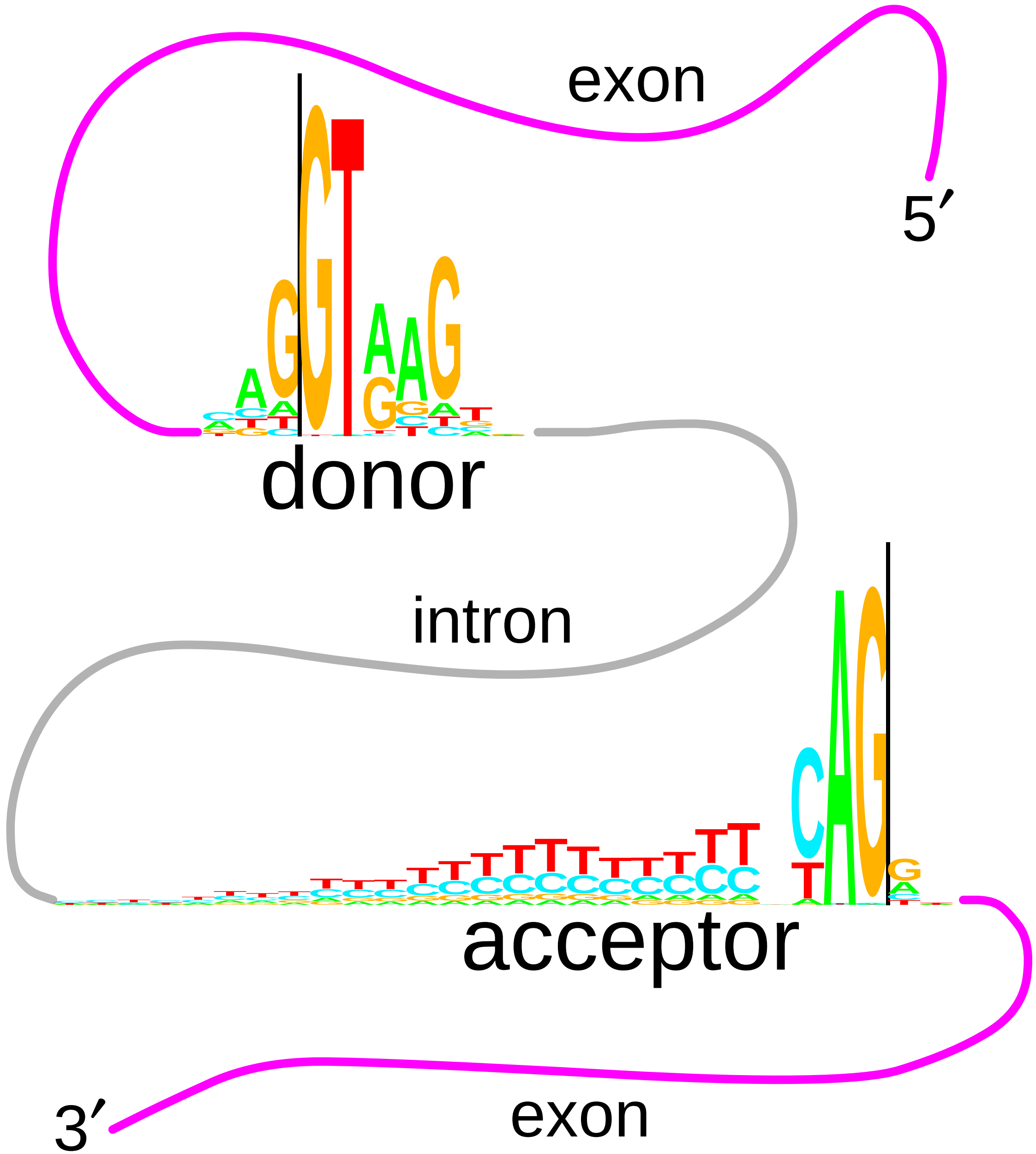
Evolution of Binding Sites







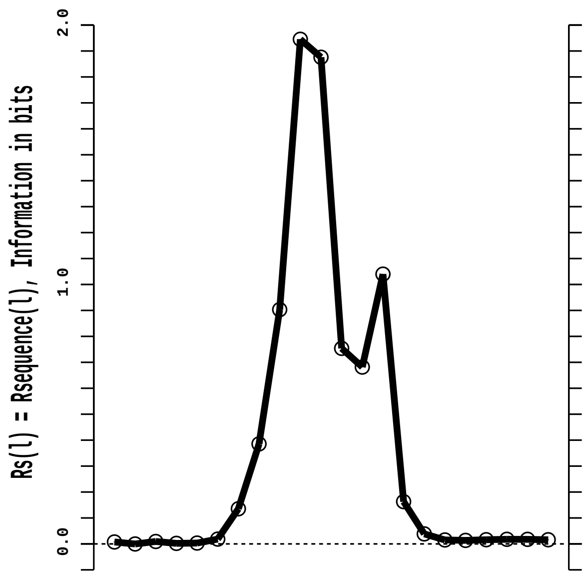
Donor and acceptor logos



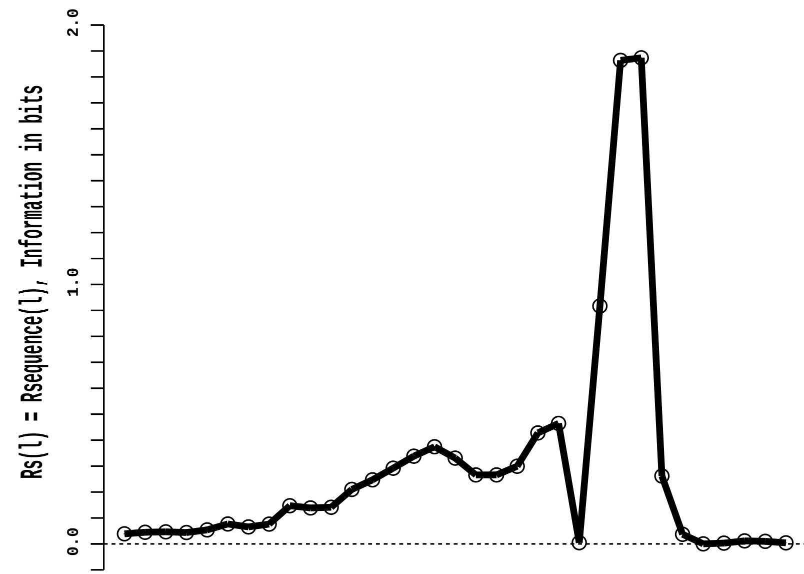
Human Splice Junction Information Curves

Sequence Conservation →
in bits per base

Donor



l	a	c	g	t	Rs(l)
-9	481	444	497	371	0.01
-8	461	461	453	422	-0.00
-7	434	457	530	378	0.01
-6	498	433	452	416	0.00
-5	474	489	398	438	0.00
-4	492	527	446	334	0.02
-3	603	673	301	222	0.14
-2	1071	263	219	246	0.39
-1	169	85	1404	141	0.90
0	0	2	1789	8	1.95
1	8	10	6	1775	1.88
2	974	35	738	43	0.75
3	1274	152	215	149	0.68
4	128	95	1468	97	1.04
5	270	304	348	812	0.16
6	428	369	564	296	0.04
7	319	492	408	424	0.02
8	332	494	413	384	0.01
9	302	437	463	408	0.02
10	281	419	424	342	0.02
11	274	382	436	369	0.02
12	273	414	400	355	0.02



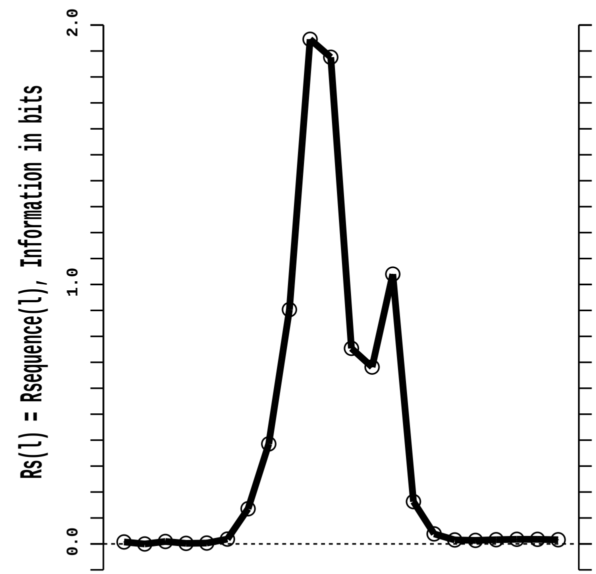
l	a	c	g	t	Rs(l)
-25	309	392	208	386	0.04
-24	282	420	216	388	0.04
-23	273	392	220	424	0.05
-22	290	411	214	397	0.04
-21	309	400	194	416	0.05
-20	248	412	203	461	0.08
-19	256	388	225	482	0.07
-18	213	442	246	455	0.08
-17	201	453	183	535	0.15
-16	193	471	199	517	0.14
-15	182	459	212	533	0.14
-14	158	476	193	619	0.21
-13	135	486	197	656	0.25
-12	122	511	174	673	0.29
-11	121	500	152	732	0.34
-10	110	483	149	775	0.37
-9	131	534	175	800	0.33
-8	156	583	192	727	0.27
-7	172	656	172	674	0.27
-6	165	653	158	726	0.30
-5	123	691	110	788	0.43
-4	127	590	103	914	0.46
-3	394	499	432	411	0.00
-2	67	1253	20	397	0.92
-1	1717	10	6	10	1.86
0	9	11	1720	4	1.87
1	449	231	871	187	0.26
2	426	308	413	590	0.04
3	424	426	419	468	0.00
4	400	487	440	410	0.00
5	424	502	347	464	0.01
6	394	527	424	392	0.01
7	441	431	485	380	0.00

Acceptor

- The consensus sequences match ...

Sequence Conservation → in bits per base

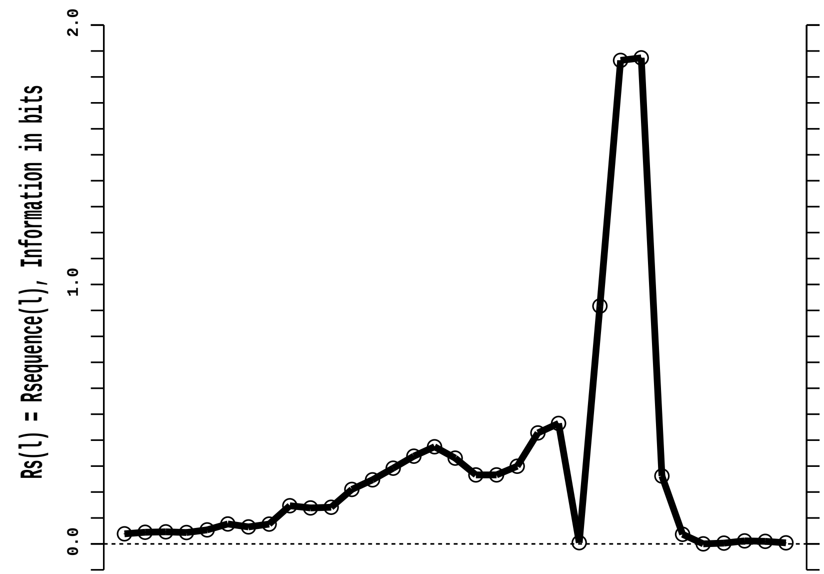
Donor



l	a	c	g	t	Rs(l)
-9	481	444	497	371	0.01
-8	461	461	453	422	-0.00
-7	434	457	530	378	0.01
-6	498	433	452	416	0.00
-5	474	489	398	438	0.00
-4	492	527	446	334	0.02
-3	603	673	301	222	0.14
-2	1071	263	219	246	0.39
-1	169	85	1404	141	0.90
0	0	2	1789	8	1.95
1	8	10	6	1775	1.88
2	974	35	738	43	0.75
3	1274	152	215	149	0.68
4	128	95	1468	97	1.04
5	270	304	348	812	0.16
6	428	369	564	296	0.04
7	319	492	408	424	0.02
8	332	494	413	384	0.01
9	302	437	463	408	0.02
10	281	419	424	342	0.02
11	274	382	436	369	0.02
12	273	414	400	355	0.02

Position L (in bases)

Acceptor



l	a	c	g	t	Rs(l)
-25	309	392	208	386	0.04
-24	282	420	216	388	0.04
-23	273	392	220	424	0.05
-22	290	411	214	397	0.04
-21	309	400	194	416	0.05
-20	248	412	203	461	0.08
-19	256	388	225	482	0.07
-18	213	442	246	455	0.08
-17	201	453	183	535	0.15
-16	193	471	199	517	0.14
-15	182	459	212	533	0.14
-14	158	476	193	619	0.21
-13	135	486	197	656	0.25
-12	122	511	174	673	0.29
-11	121	500	152	732	0.34
-10	110	483	149	775	0.37
-9	131	534	175	800	0.33
-8	156	583	192	727	0.27
-7	172	656	172	674	0.27
-6	165	653	158	726	0.30
-5	123	691	110	788	0.43
-4	127	590	103	914	0.46
-3	394	499	432	411	0.00
-2	67	1253	20	397	0.92
-1	1717	10	6	10	1.86
0	9	11	1720	4	1.87
1	449	231	871	187	0.26
2	426	308	413	590	0.04
3	424	426	419	468	0.00
4	400	487	440	410	0.00
5	424	502	347	464	0.01
6	394	527	424	392	0.01
7	441	431	485	380	0.00

Position L (in bases)

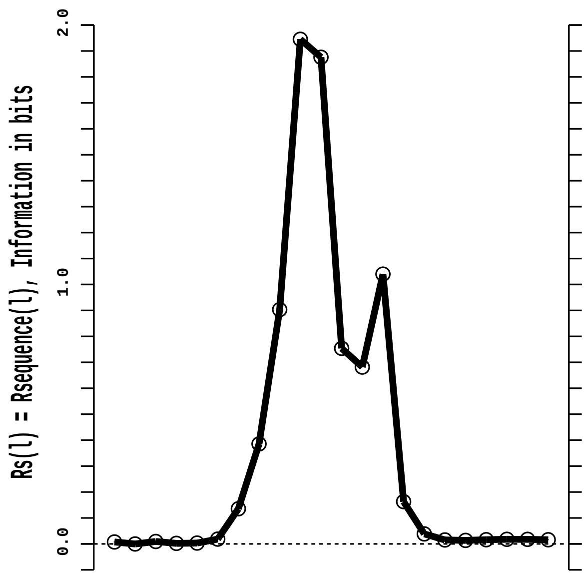
C

- The consensus sequences match ...

Human Splice Junction Information Curves

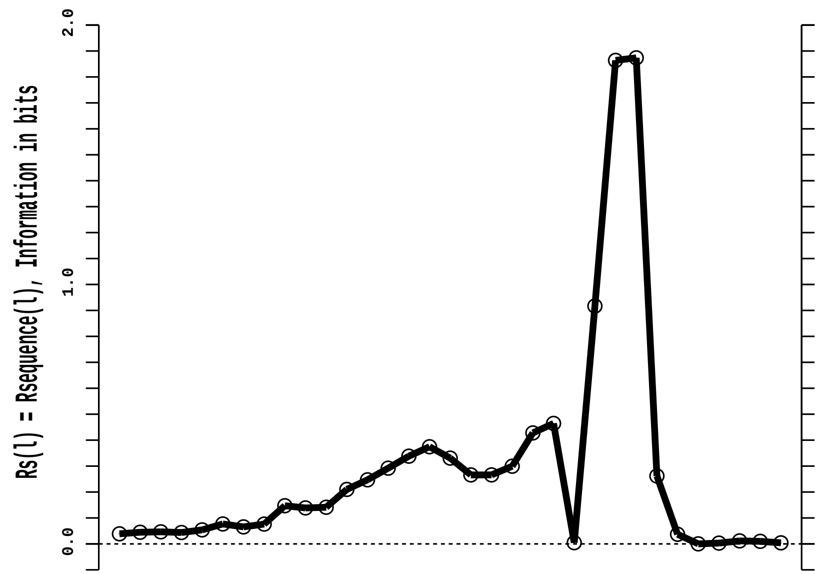
Sequence Conservation →
in bits per base

Donor



l	a	c	g	t	Rs(l)
-9	481	444	497	371	0.01
-8	461	461	453	422	-0.00
-7	434	457	530	378	0.01
-6	498	433	452	416	0.00
-5	474	489	398	438	0.00
-4	492	527	446	334	0.02
-3	603	673	301	222	0.14
-2	1071	263	219	246	0.39
-1	169	85	1404	141	0.90
0	0	2	1789	8	1.95
1	8	10	6	1775	1.88
2	974	35	738	43	0.75
3	1274	152	215	149	0.68
4	128	95	1468	97	1.04
5	270	304	348	812	0.16
6	428	369	564	296	0.04
7	319	492	408	424	0.02
8	332	494	413	384	0.01
9	302	437	463	408	0.02
10	281	419	424	342	0.02
11	274	382	436	369	0.02
12	273	414	400	355	0.02

Position L (in bases)



l	a	c	g	t	Rs(l)
-25	309	392	208	386	0.04
-24	282	420	216	388	0.04
-23	273	392	220	424	0.05
-22	290	411	214	397	0.04
-21	309	400	194	416	0.05
-20	248	412	203	461	0.08
-19	256	388	225	482	0.07
-18	213	442	246	455	0.08
-17	201	453	183	535	0.15
-16	193	471	199	517	0.14
-15	182	459	212	533	0.14
-14	158	476	193	619	0.21
-13	135	486	197	656	0.25
-12	122	511	174	673	0.29
-11	121	500	152	732	0.34
-10	110	483	149	775	0.37
-9	131	534	175	800	0.33
-8	156	583	192	727	0.27
-7	172	656	172	674	0.27
-6	165	653	158	726	0.30
-5	123	691	110	788	0.43
-4	127	590	103	914	0.46
-3	394	499	432	411	0.00
-2	67	1253	20	397	0.92
-1	1717	10	6	10	1.86
0	9	11	1720	4	1.87
1	449	231	871	187	0.26
2	426	308	413	590	0.04
3	424	426	419	468	0.00
4	400	487	440	410	0.00
5	424	502	347	464	0.01
6	394	527	424	392	0.01
7	441	431	485	380	0.00

Position L (in bases)

Acceptor

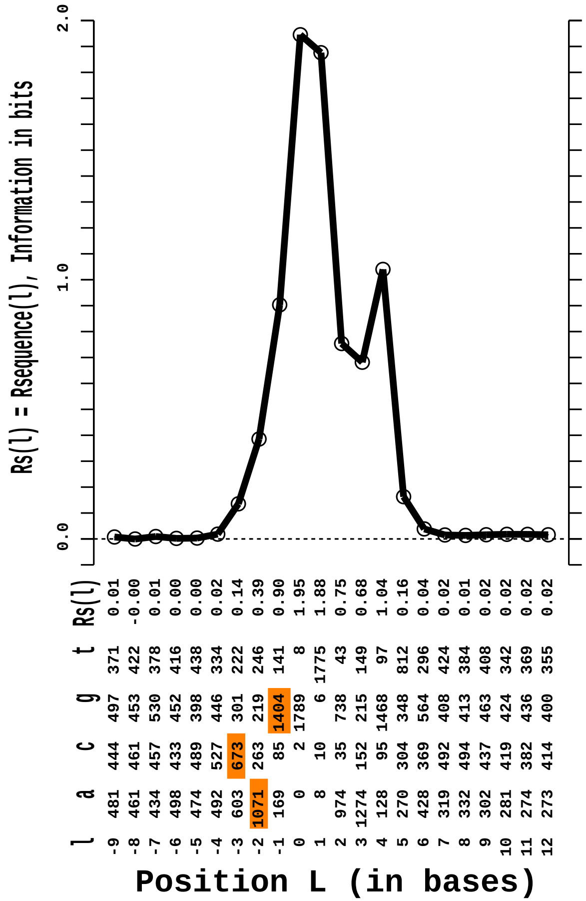
C A

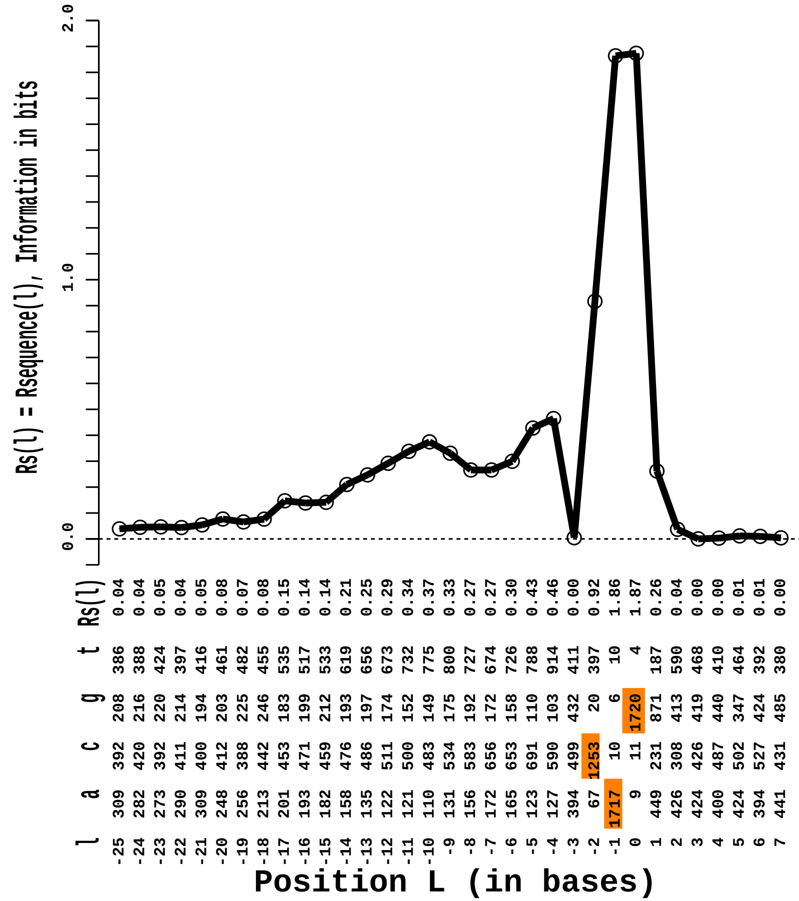
- The consensus sequences match …

Human Splice Junction Information Curves

Sequence Conservation →
in bits per base

Donor





Acceptor

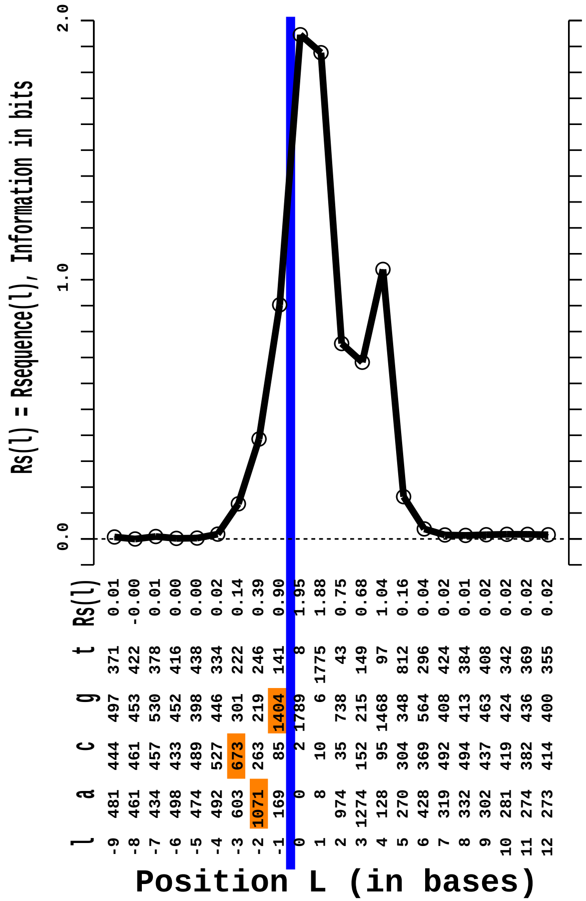
C A G

- The consensus sequences match . . .

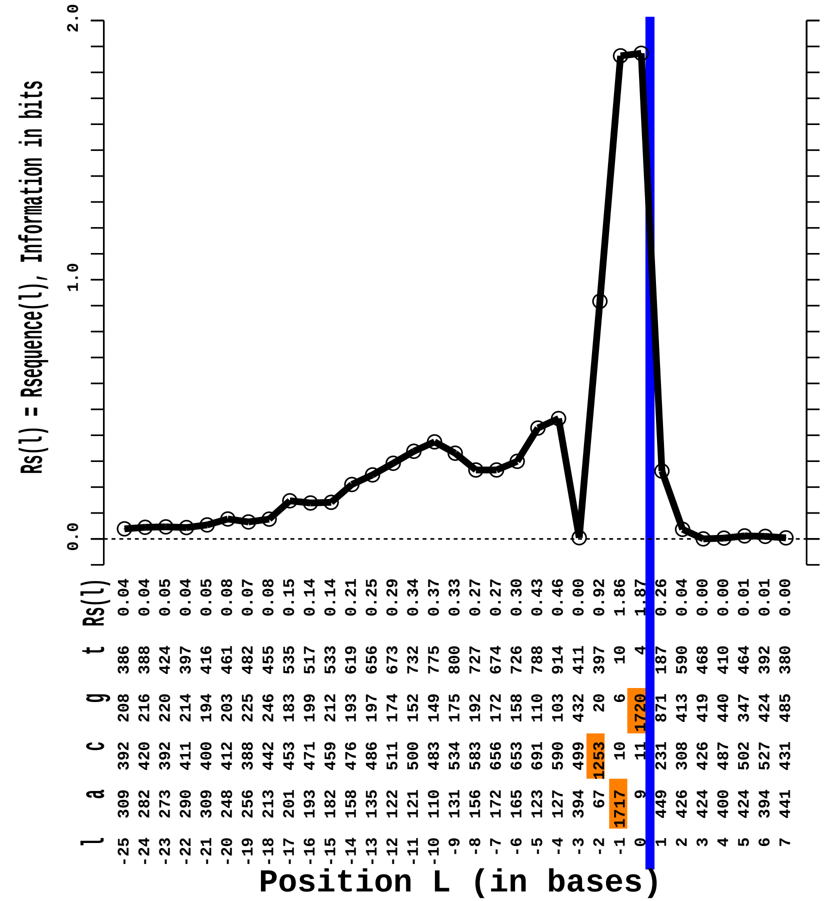
https://alum.mit.edu/www/toms/papers/splice/

Sequence Conservation → in bits per base

Donor



Acceptor



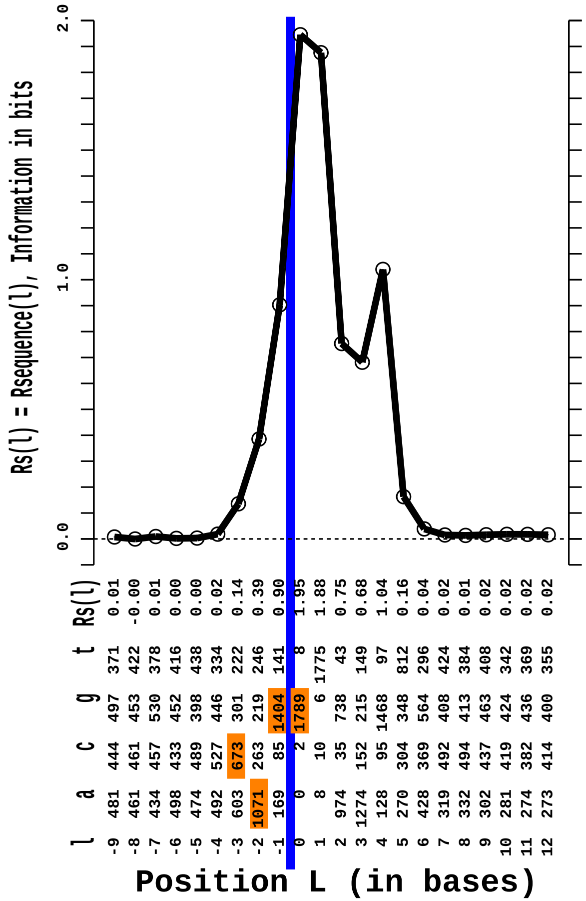
C A G —

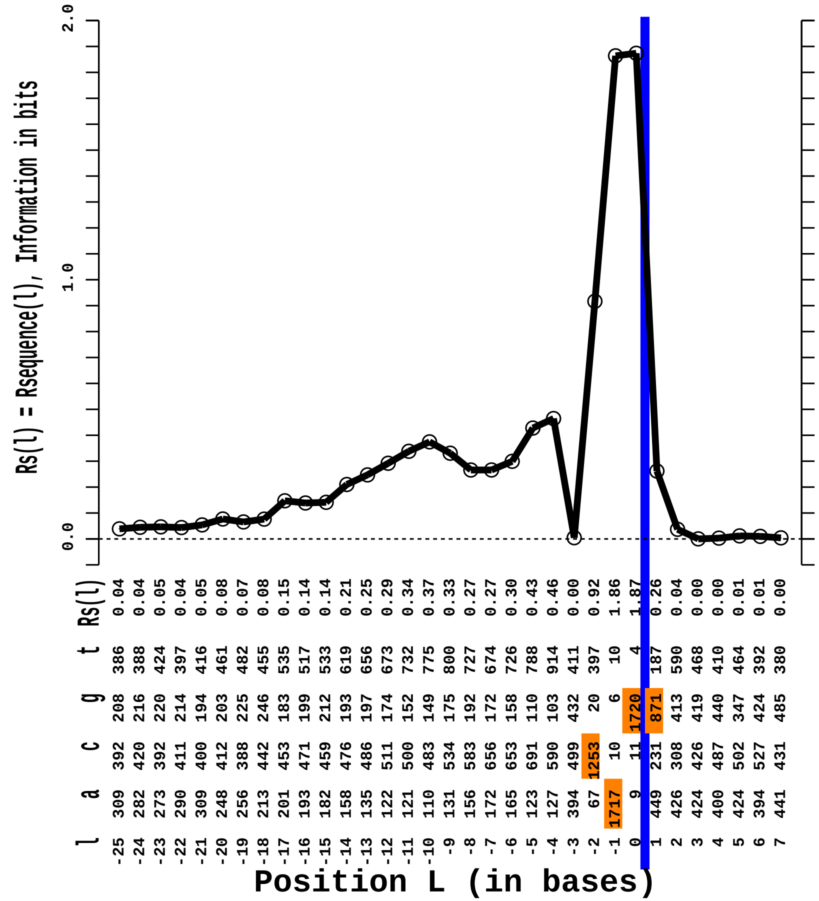
- The consensus sequences match . . .

Human Splice Junction Information Curves

Sequence Conservation → in bits per base

Donor





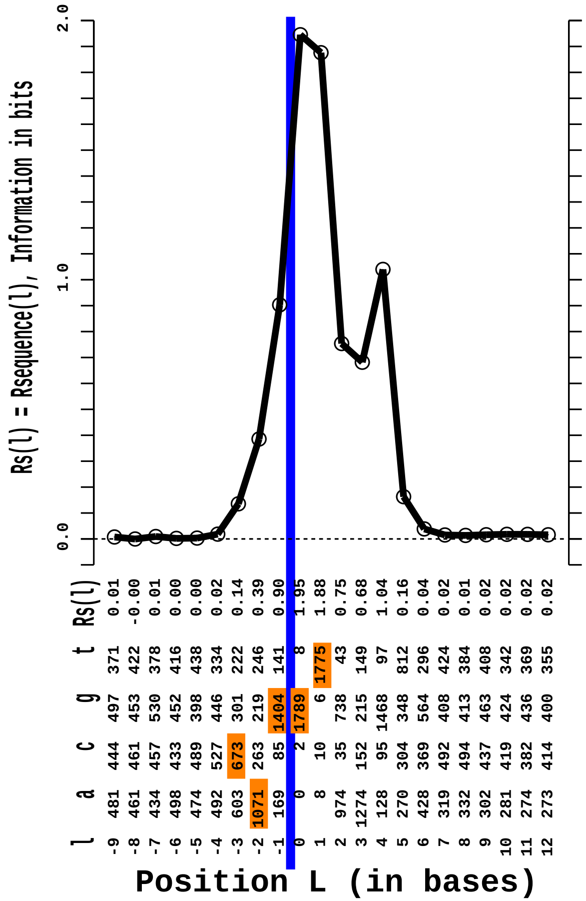
Acceptor

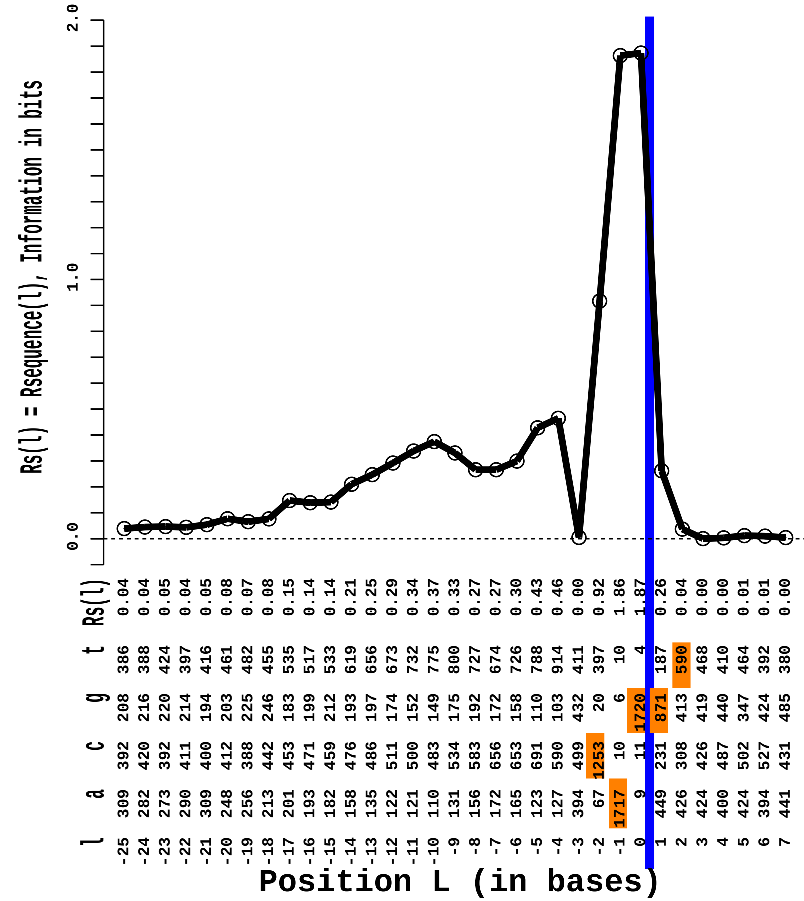
C A G — G

- The consensus sequences match . . .

Sequence Conservation →
in bits per base

Donor





Acceptor

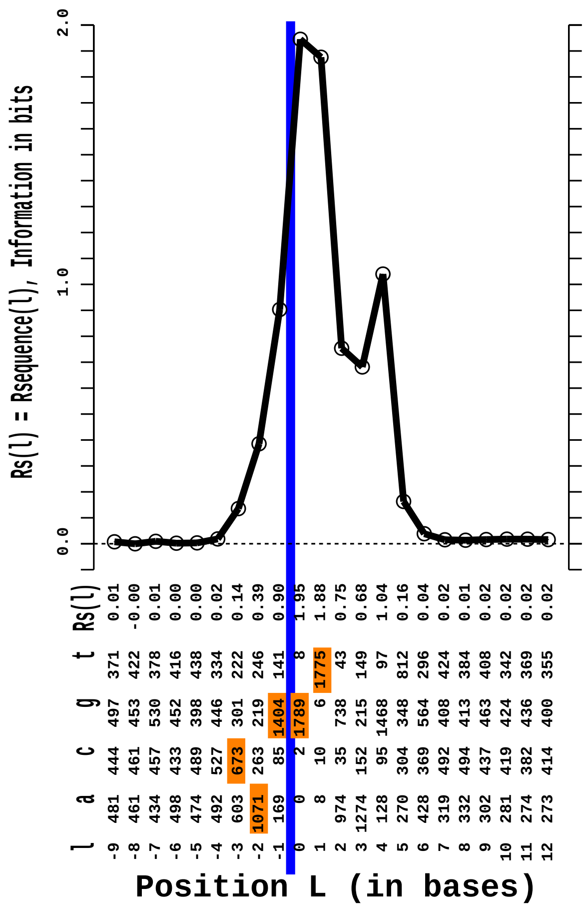
C A G — G T

- The consensus sequences match ...

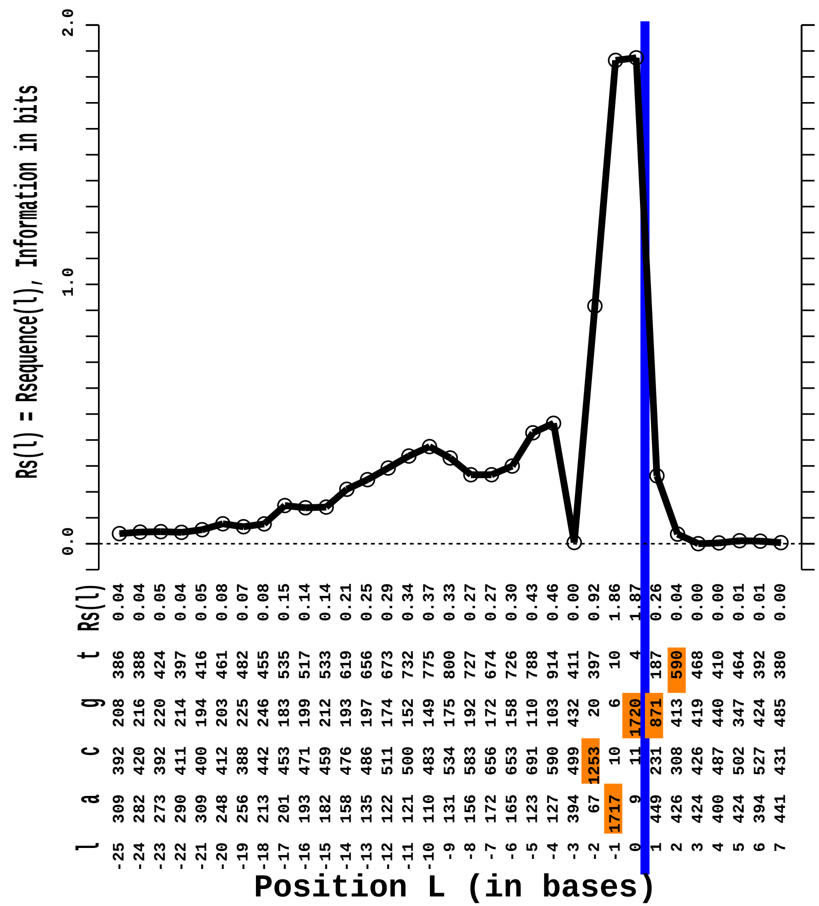
Human Splice Junction Information Curves

Sequence Conservation →
in bits per base

Donor



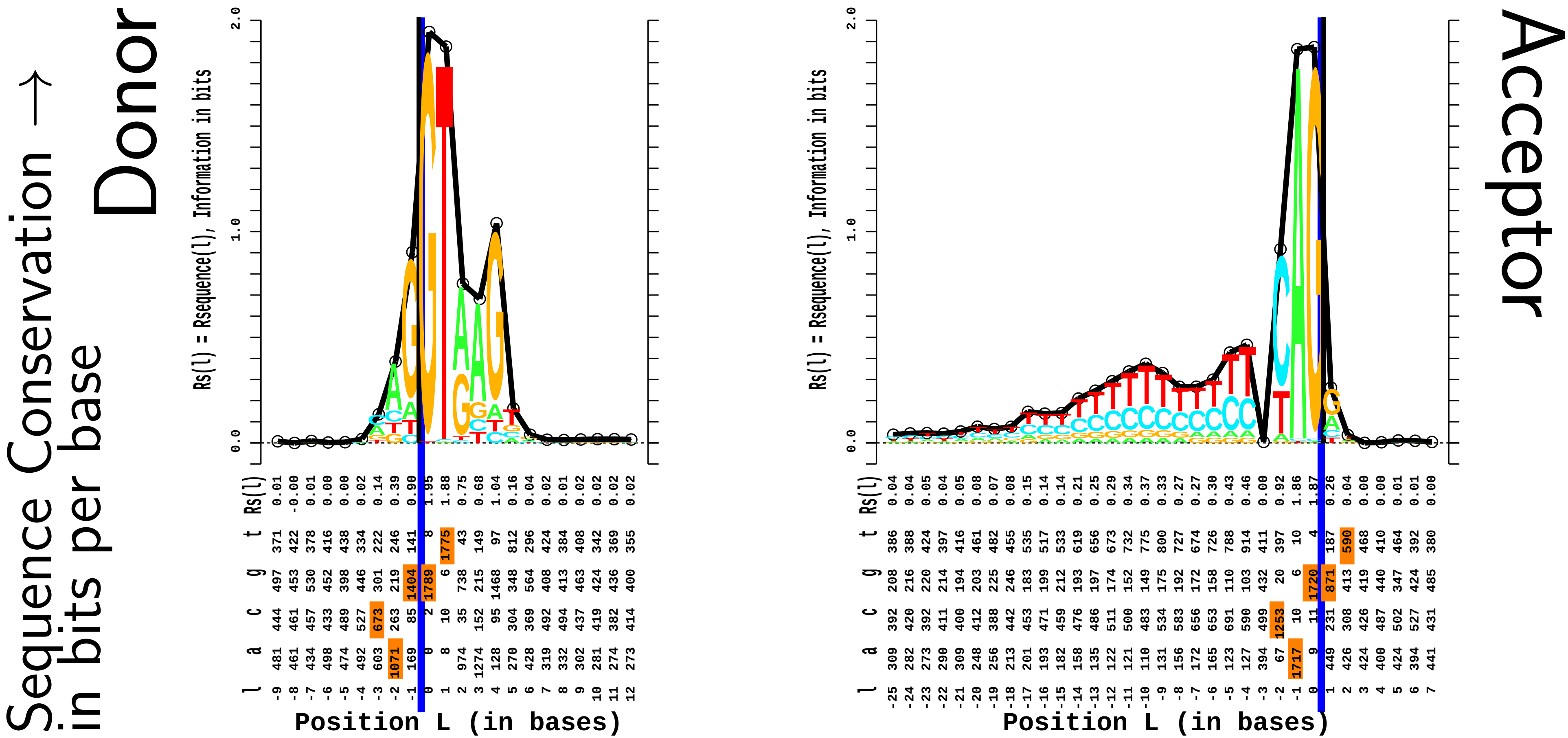
Acceptor



C A G — G T

- The consensus sequences match . . .
- BUT the information curves (sequence conservation) differ!

Human Splice Junction Information Curves



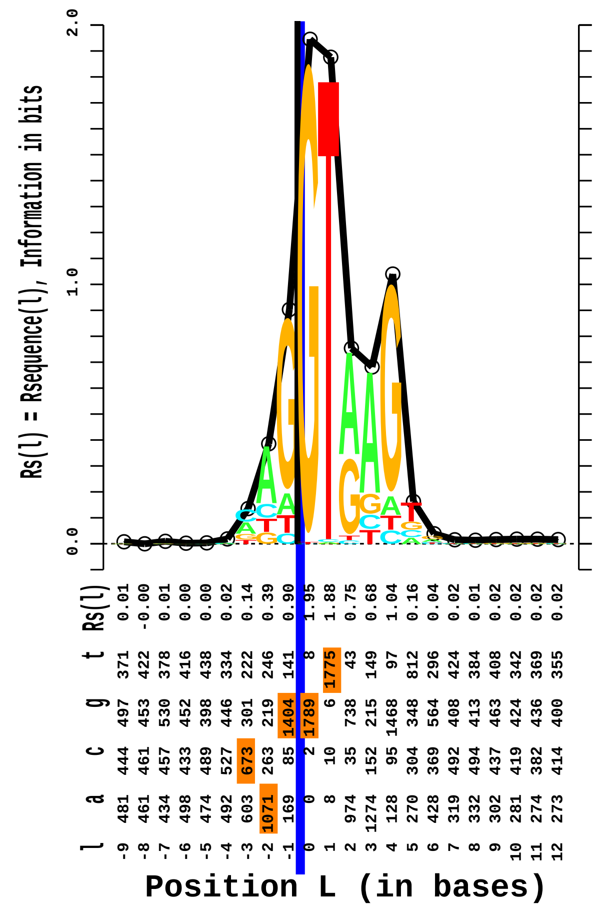
C A G — G T

- The consensus sequences match . . .
- BUT the information curves (sequence conservation) differ!
- Put letters into the graph proportional to their frequency!

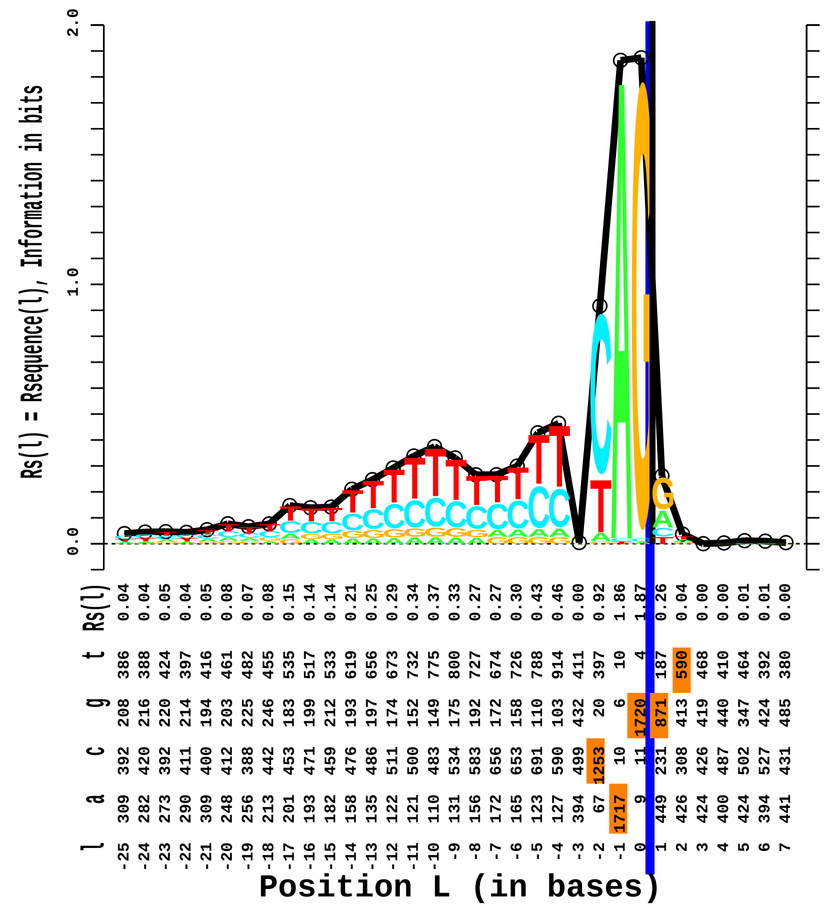
Human Splice Junction Information Curves

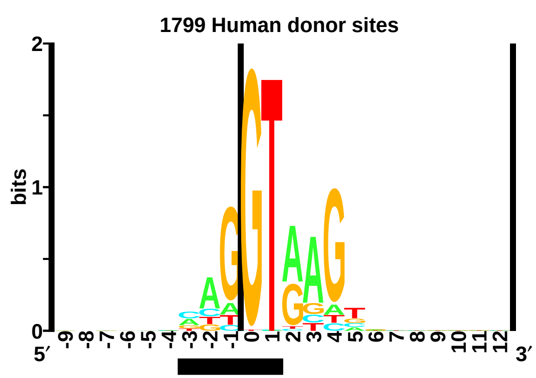
Sequence Conservation →
in bits per base

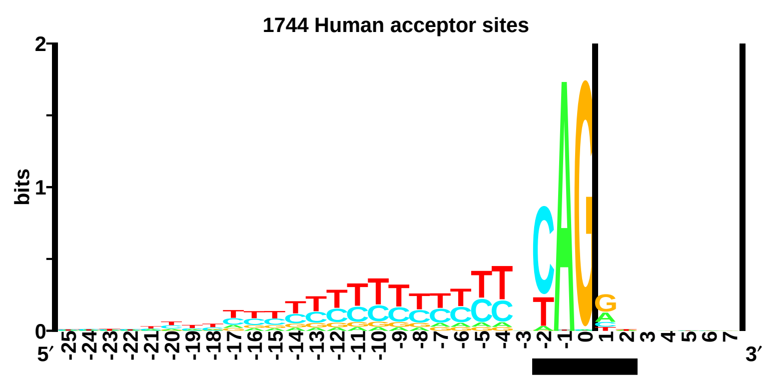
Donor



Acceptor

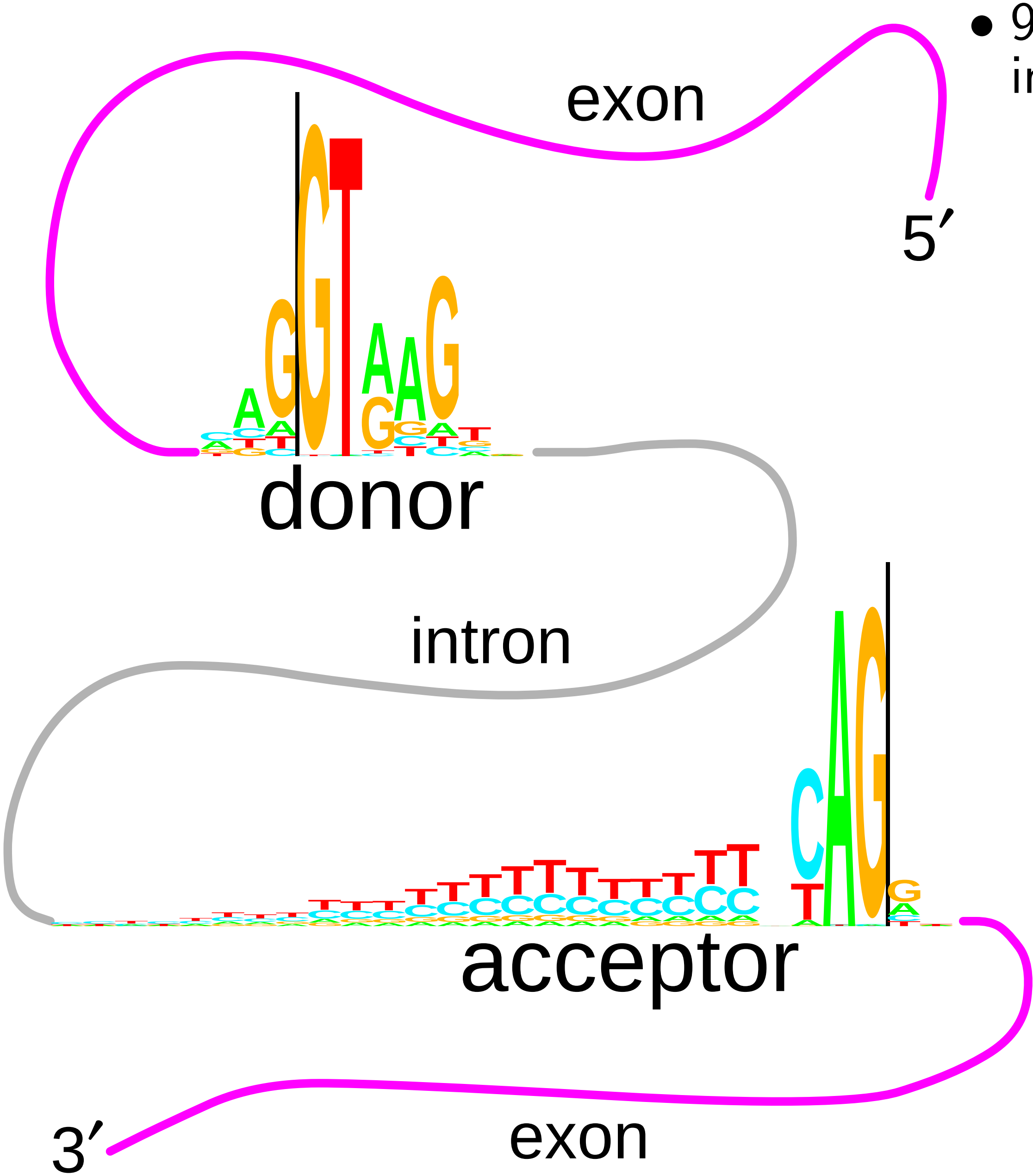






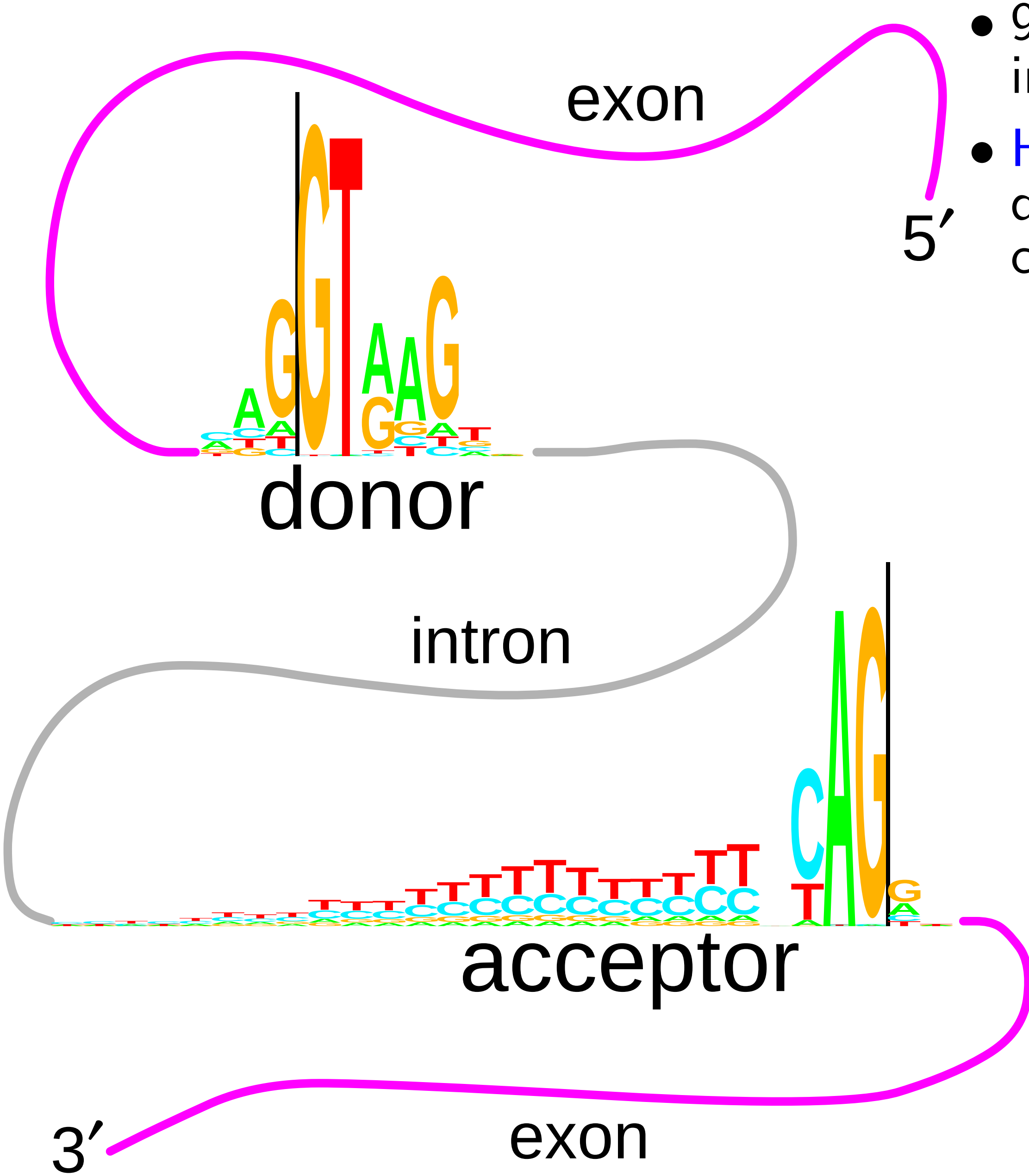
- That's how and why we invented sequence logos!

Splice Junction Sequence Logos

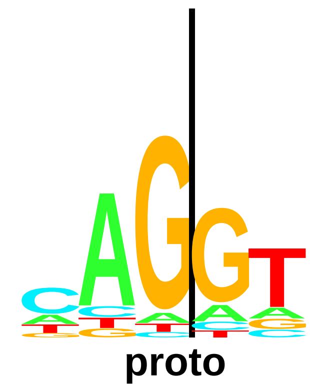


- 90% of the splice junction information is on the intron side

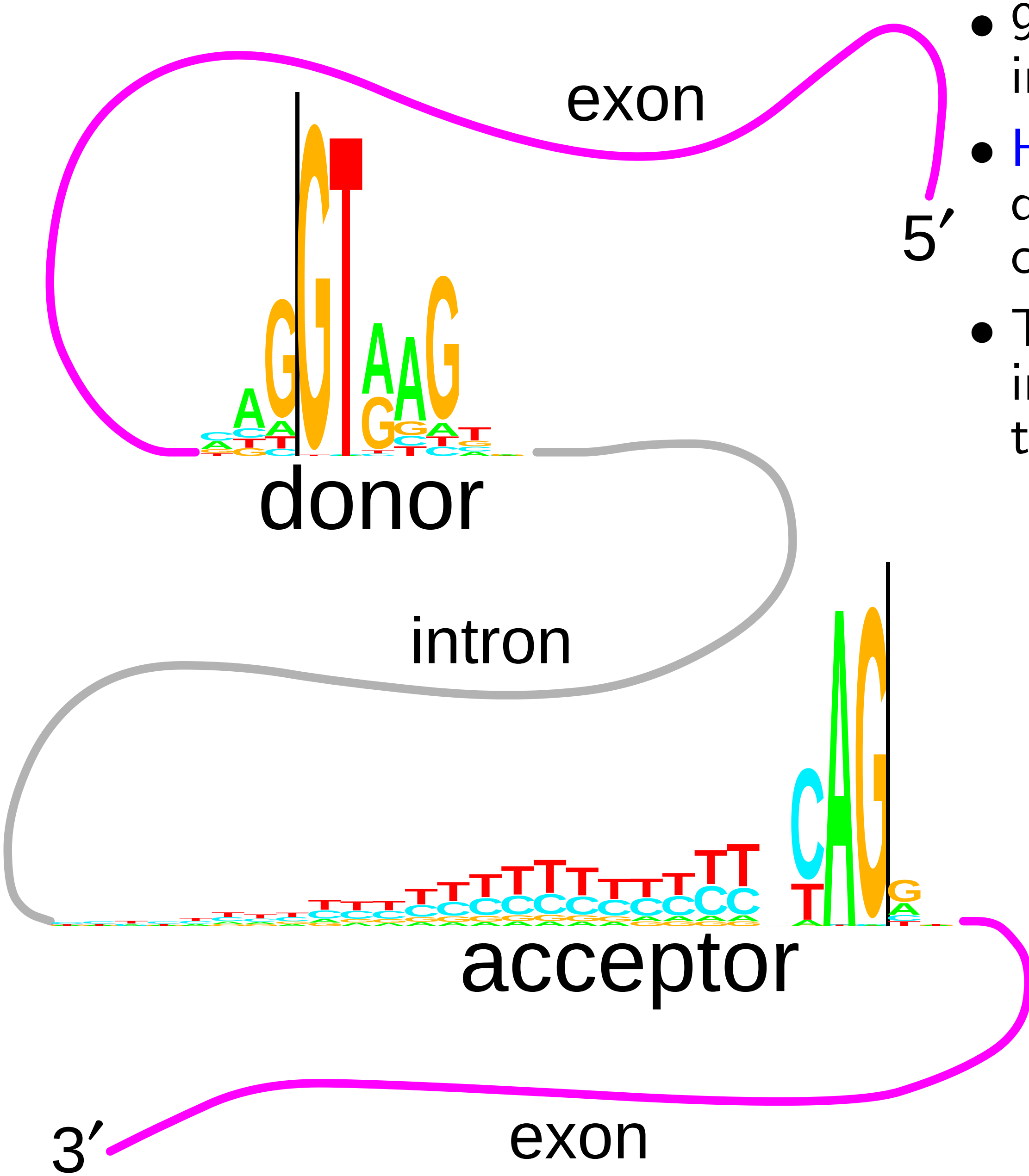
Splice Junction Sequence Logos



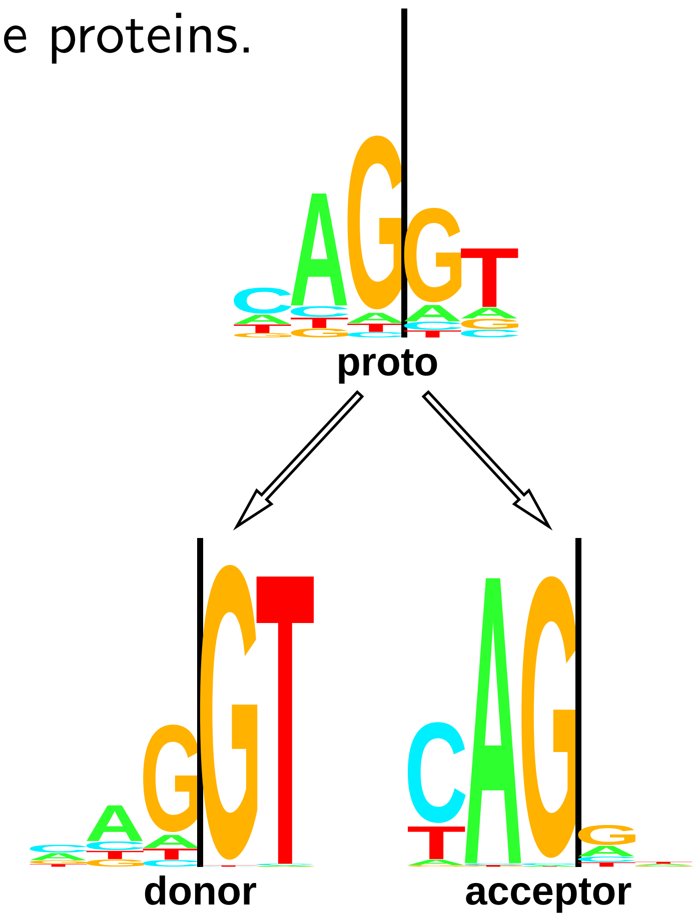
- 90% of the splice junction information is on the intron side
- Hypothesis:
  donor and acceptor sites had a common ancestor that duplicated



Splice Junction Sequence Logos

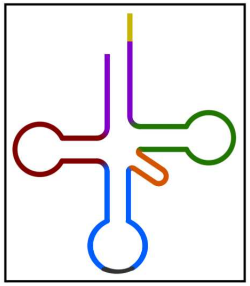


- 90% of the splice junction information is on the intron side
- Hypothesis:
  donor and acceptor sites had a common ancestor that duplicated
- They evolved to put the information into the intron. This avoids affecting the proteins.



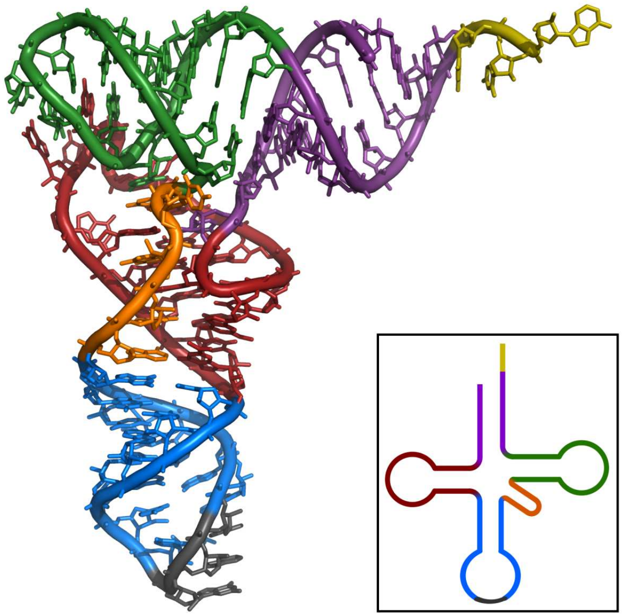
https://alum.mit.edu/www/toms/papers/splice/

3D Sequence Logos for tRNA Correlations



- tRNA reads RNA to make protein

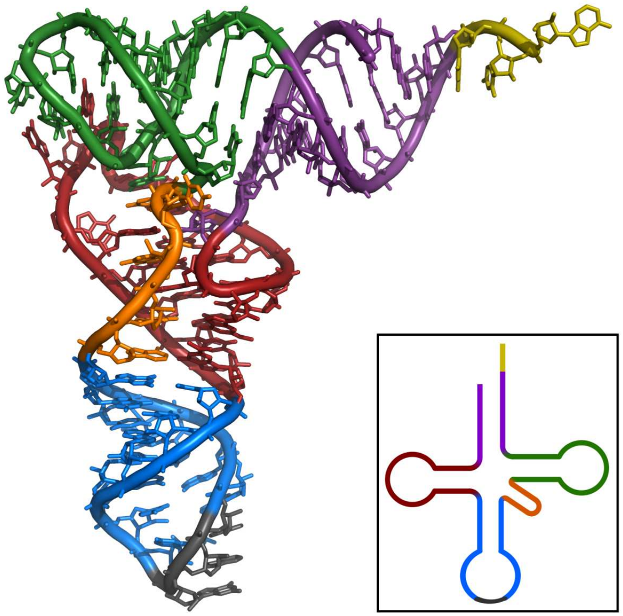
3D Sequence Logos for tRNA Correlations

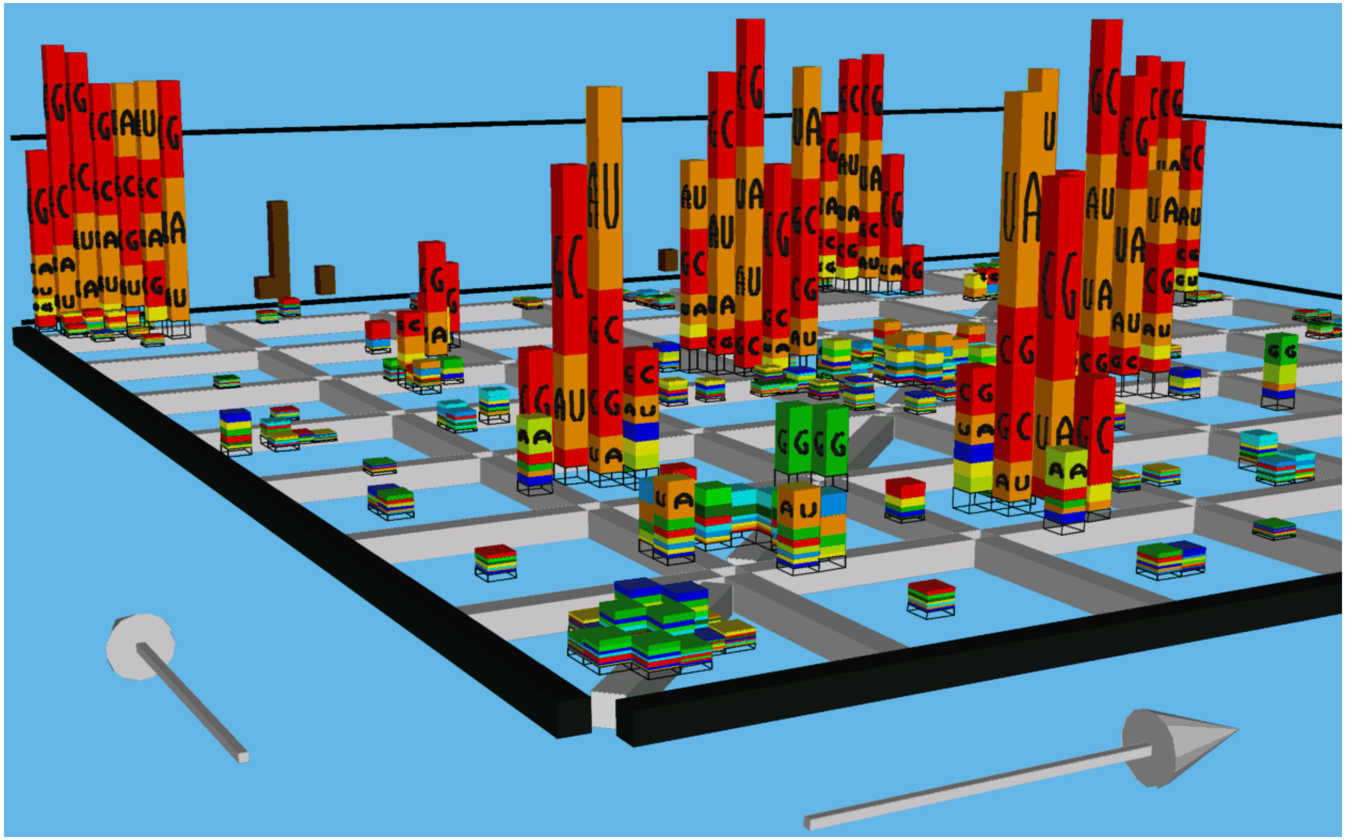


- tRNA reads RNA
  to make protein
- Correlations can be measured in bits!

https://alum.mit.edu/www/toms/papers/correlogo/ Nucleic Acids Res. 2006 34:W405-11

3D Sequence Logos for tRNA Correlations

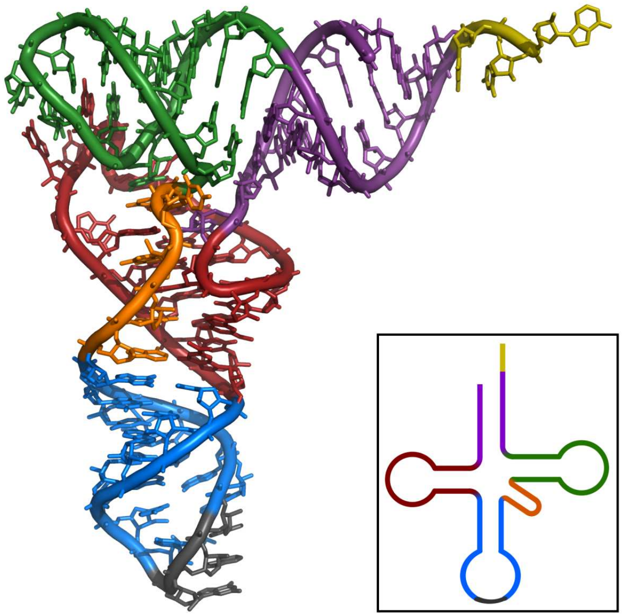


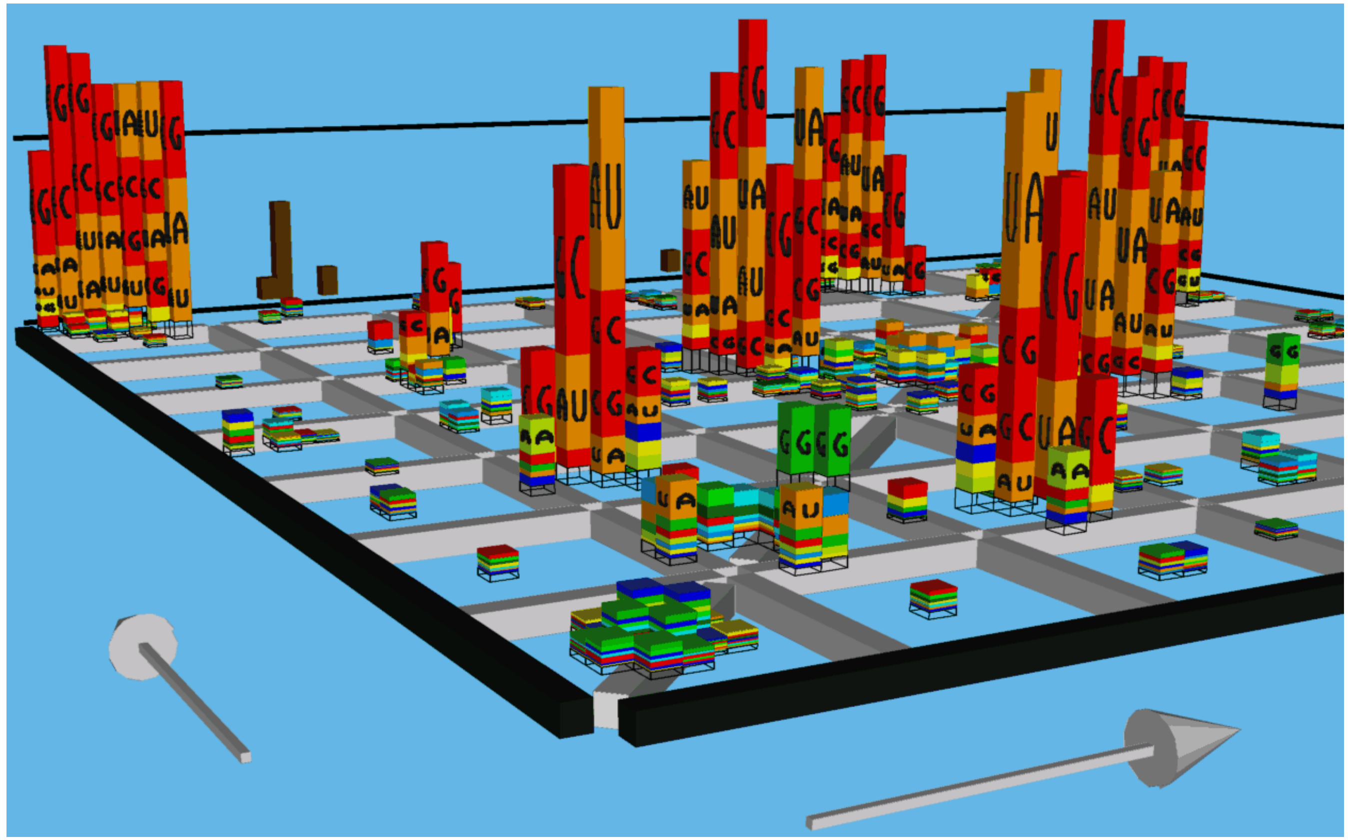


- tRNA reads RNA to make protein
- Correlations can be measured in bits!
- 3D Sequence logo

https://alum.mit.edu/www/toms/papers/correlogo/ Nucleic Acids Res. 2006 34:W405-11

3D Sequence Logos for tRNA Correlations

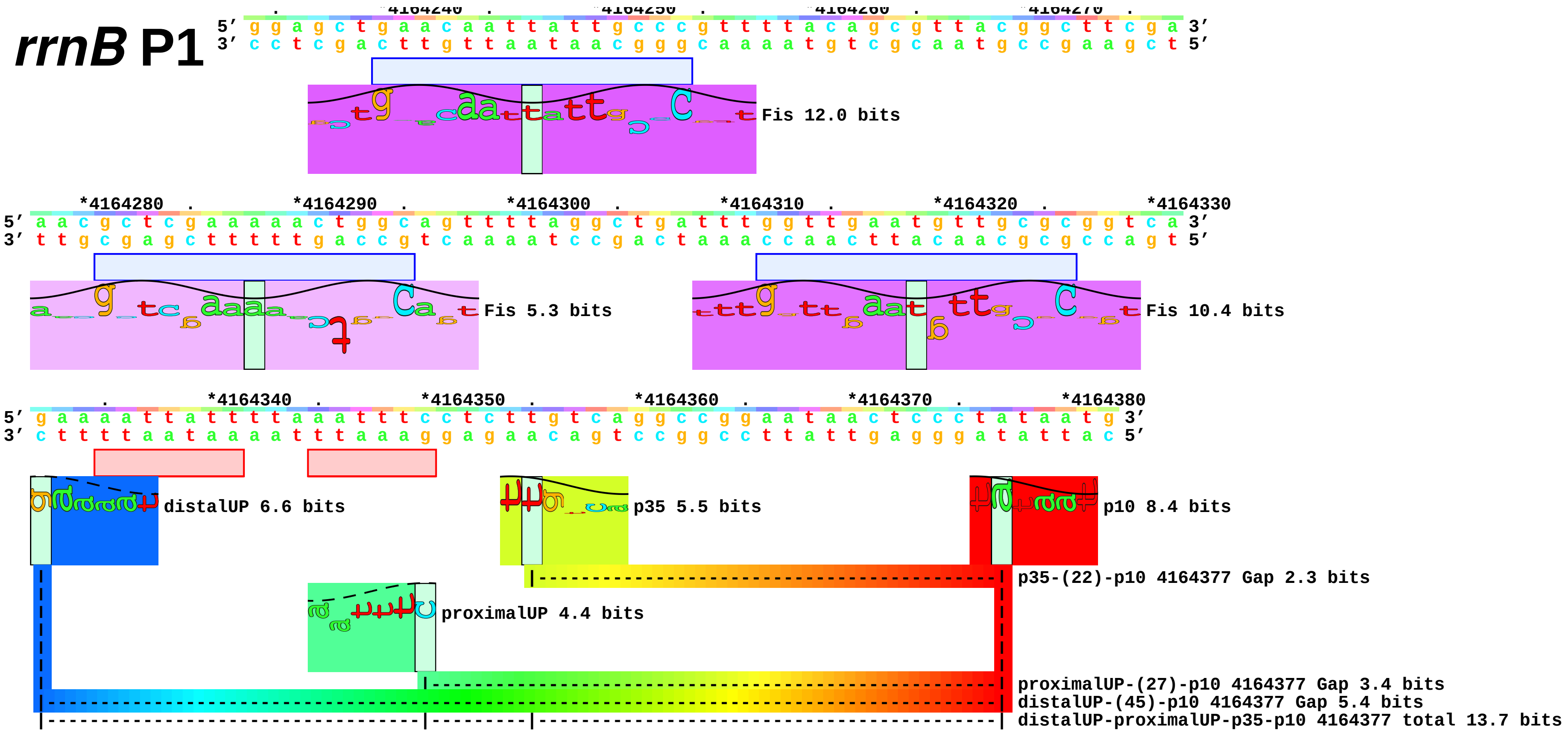




- tRNA reads RNA to make protein
- Correlations can be measured in bits!
- 3D Sequence logo
- OBSERVED: tRNA stems

https://alum.mit.edu/www/toms/papers/correlogo/ Nucleic Acids Res. 2006 34:W405-11

Sequence Walker example: rrnB P1



Complex Sequence Walker Example

- σ^{70} promoters have a -35 and a -10

Complex Sequence Walker Example

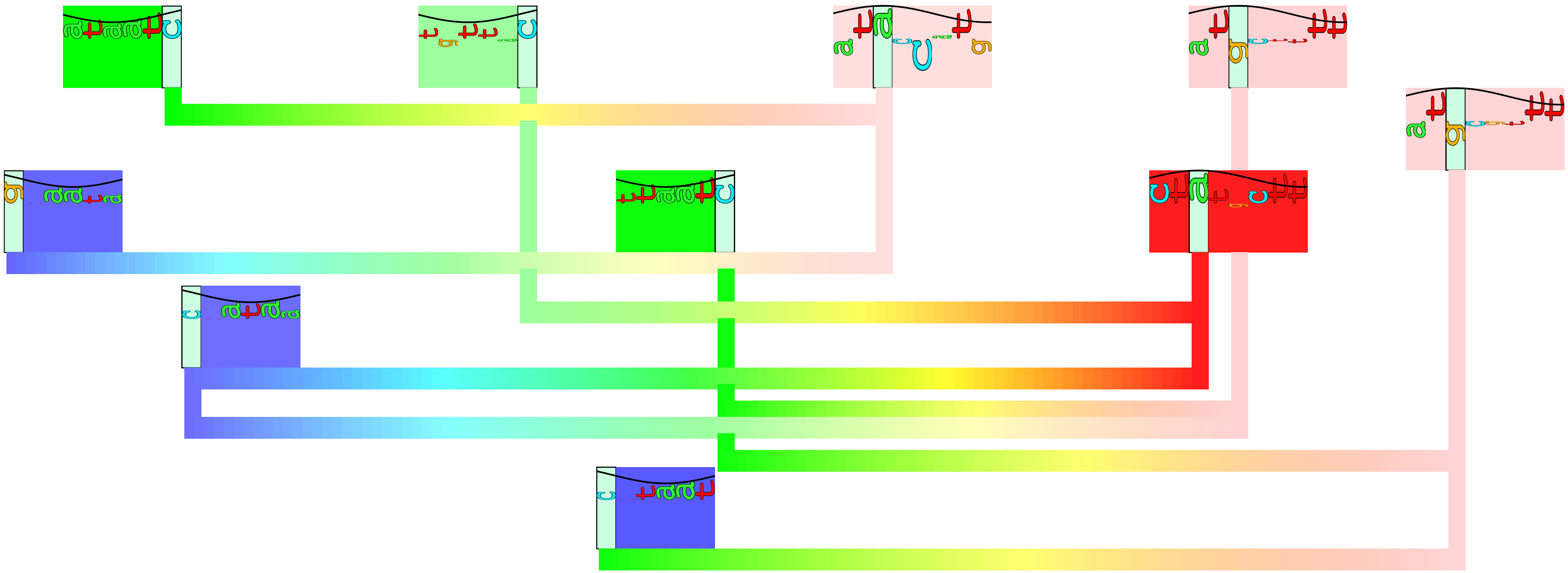
- σ^{70} promoters have a -35 and a -10
- Using information theory we discovered that stress-response σ^{38} promoters do not have a -35

Complex Sequence Walker Example

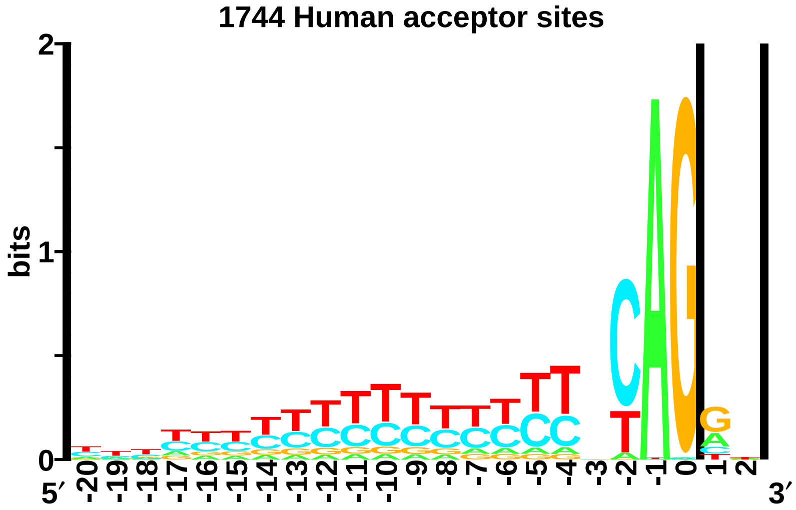
- σ^{70} promoters have a -35 and a -10
- Using information theory we discovered that stress-response σ^{38} promoters do not have a -35
- Instead, they have a -10 and two UP elements

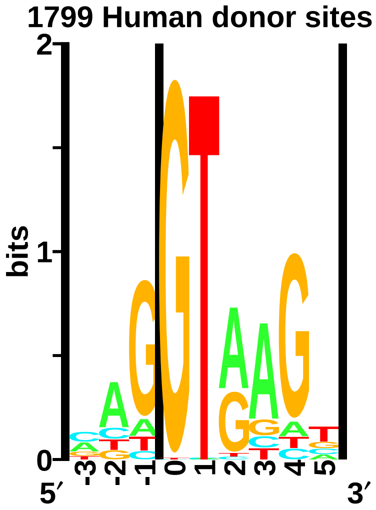
Complex Sequence Walker Example

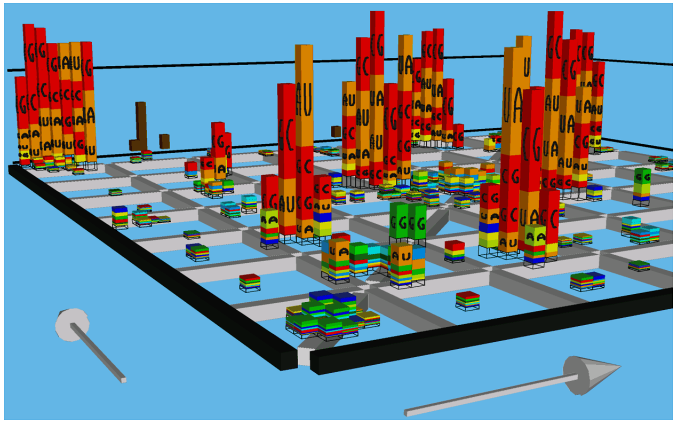
- σ^{70} promoters have a -35 and a -10
- Using information theory we discovered that stress-response σ^{38} promoters do not have a -35
- Instead, they have a -10 and two UP elements
- σ^{38} promoter *talA* P1 is complex!



Important Discovery

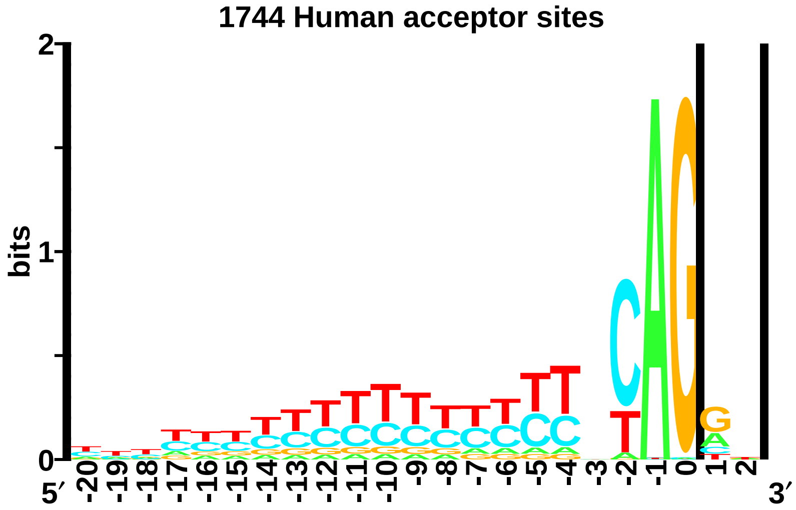


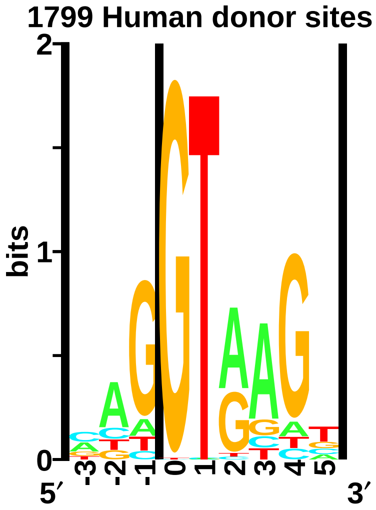


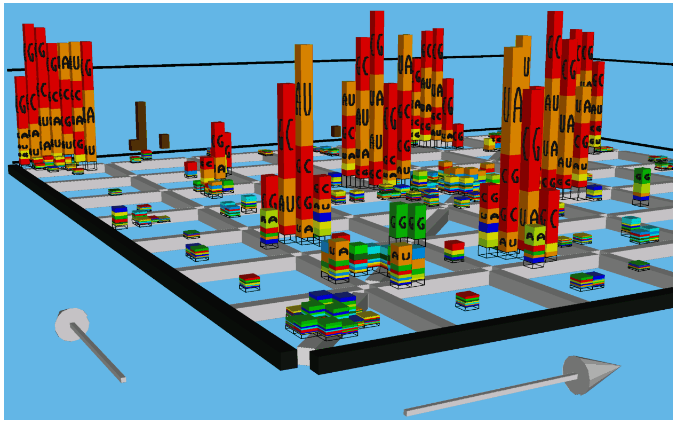


- Area under a sequence logo is the total information.

Important Discovery

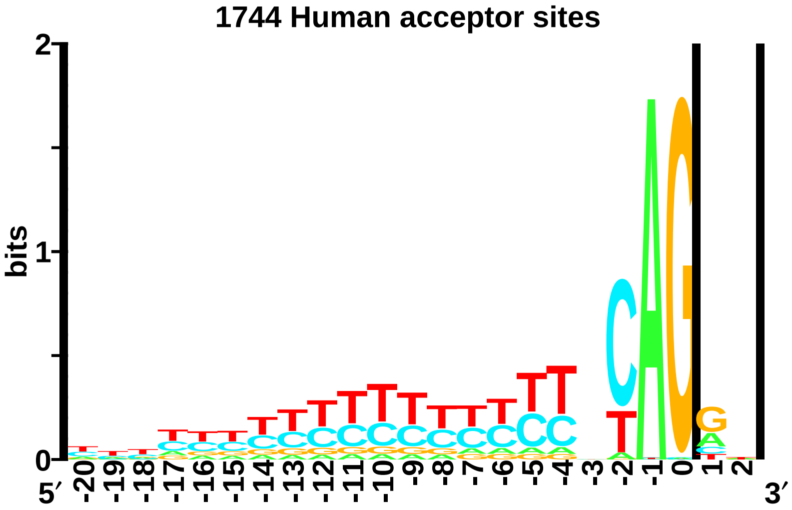


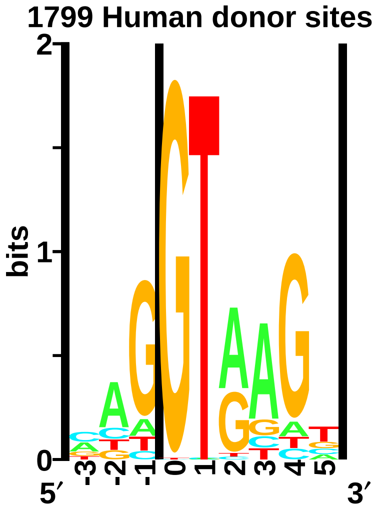


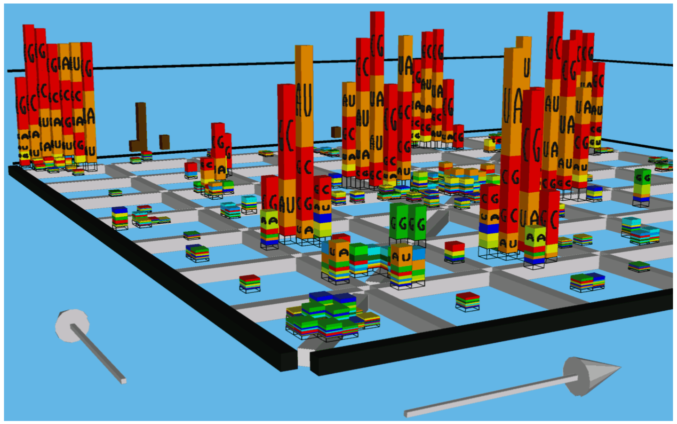


- Area under a sequence logo is the total information.
- How is that related to the binding energy?

Important Discovery

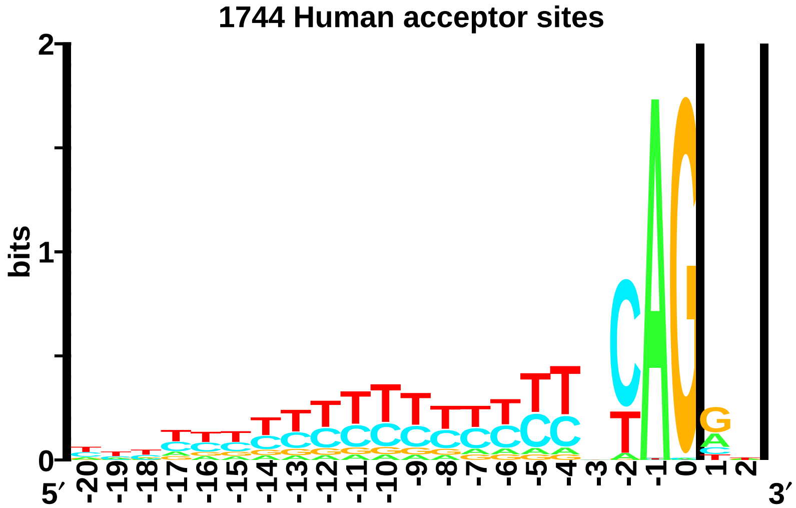


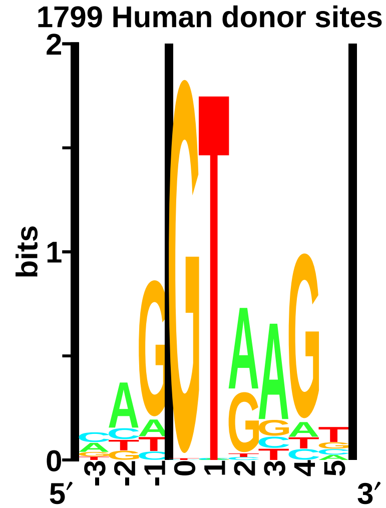


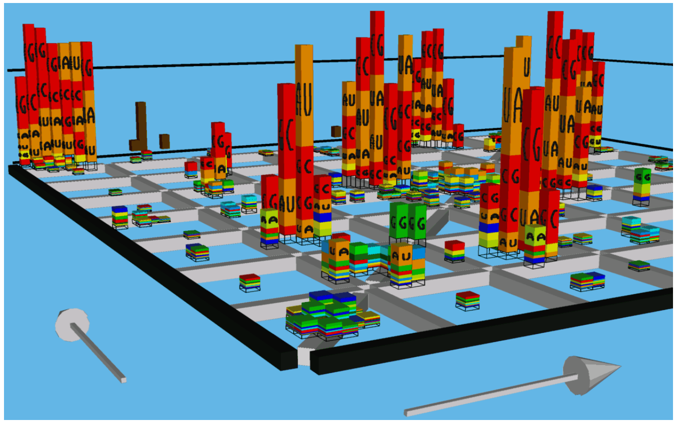


- Area under a sequence logo is the total information.
- How is that related to the binding energy?
- Information gained for energy dissipated

Important Discovery

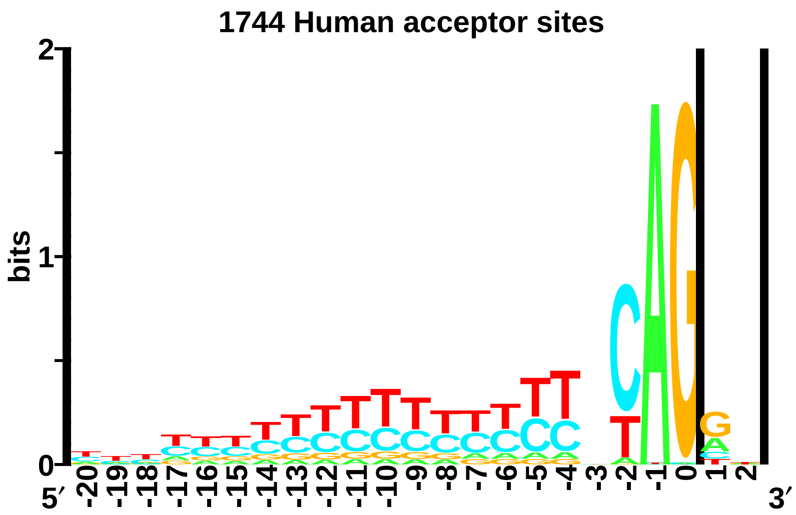


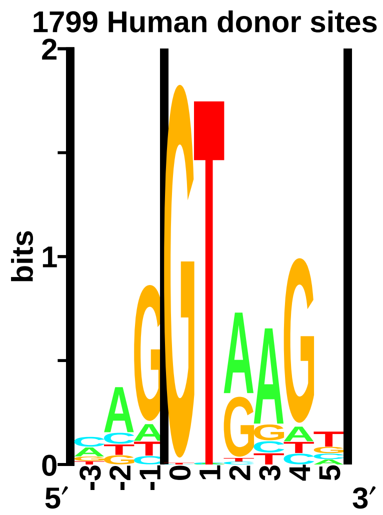


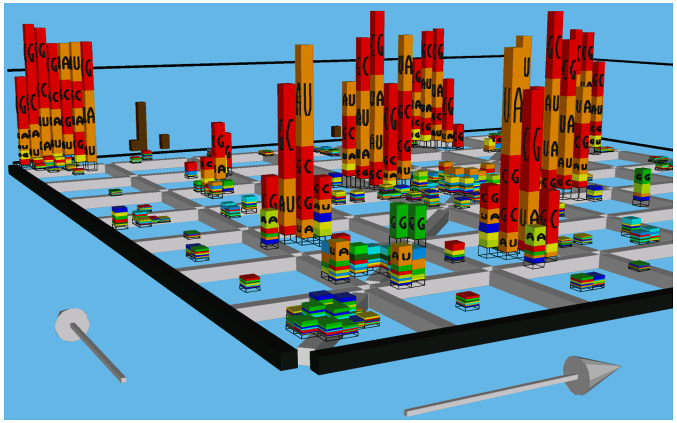


- Area under a sequence logo is the total information.
- How is that related to the binding energy?
- Information gained for energy dissipated
- Isothermal efficiency

Important Discovery



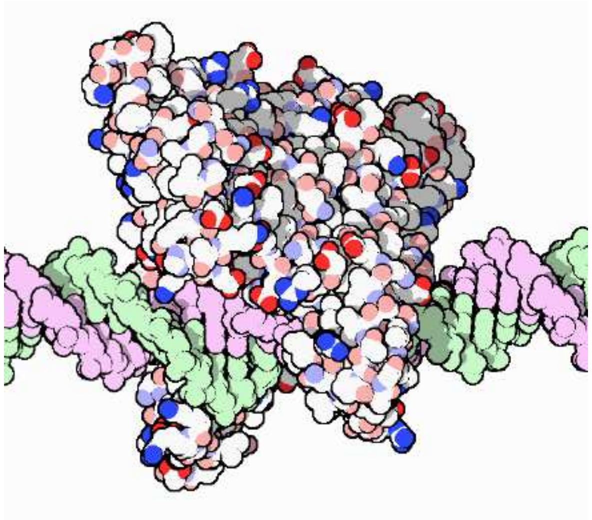




- Area under a sequence logo is the total information.
- How is that related to the binding energy?
- Information gained for energy dissipated
- Isothermal efficiency
- My most important discovery:

$\boxed{\text{Molecules are often 70\% efficient}}$

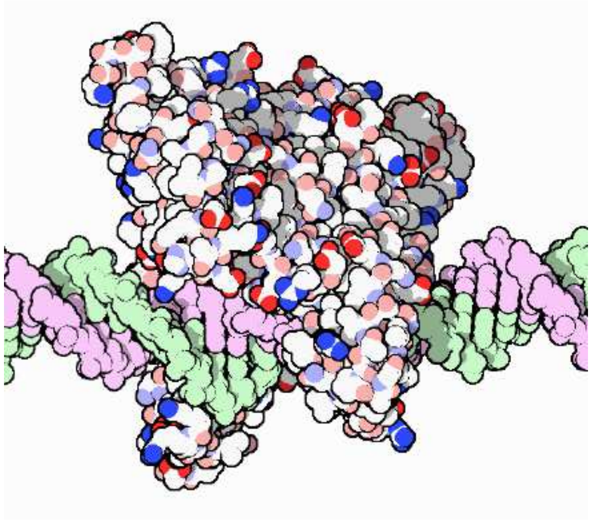
Information of EcoRI DNA Binding

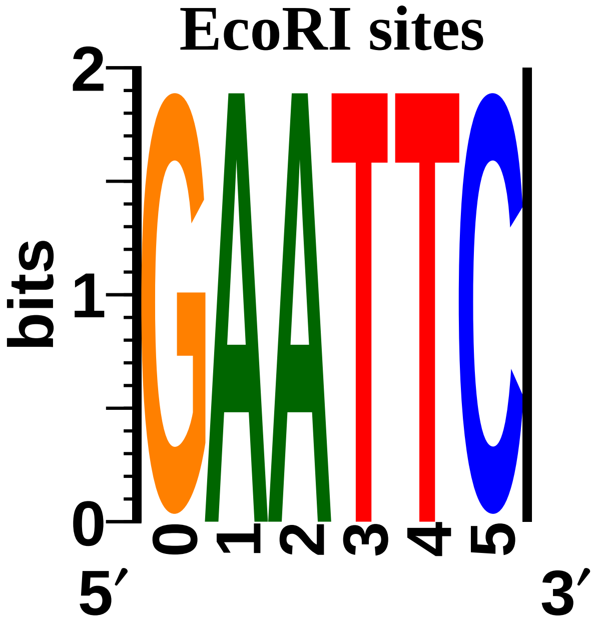


- EcoRI - restriction enzyme

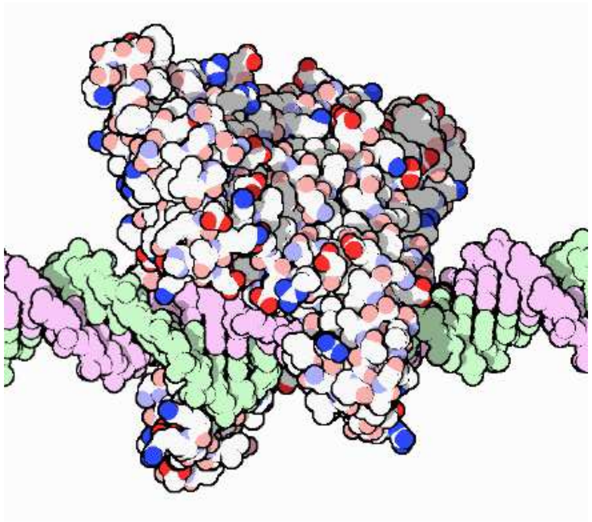
Information of EcoRI DNA Binding

- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$

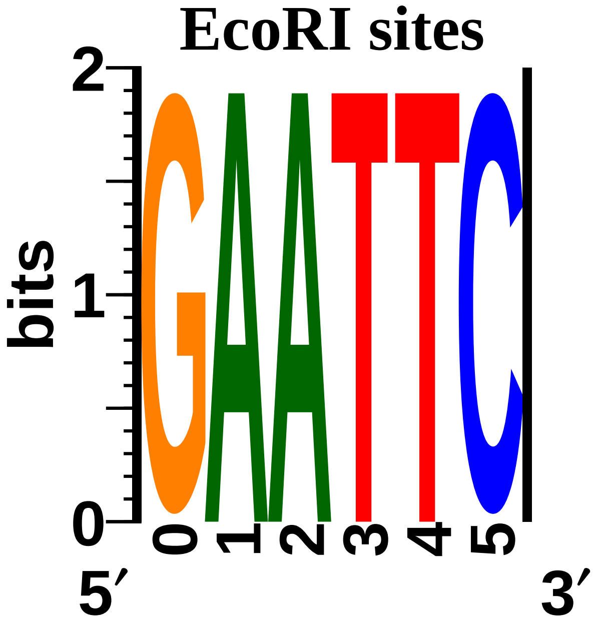




Information of EcoRI DNA Binding



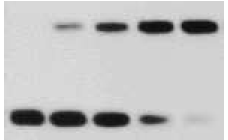
- EcoRI - restriction enzyme
- EcoRI binds DNA at $5'$ GAATTC $3'$
- information required:
  6 bases $\times$ 2 bits per base $=$ $\boxed{\text{12 bits}}$



Energy Dissipation by EcoRI

- Measured specific binding constant:

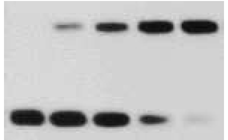
$$K_{spec} = 1.6 \times 10^5$$



Energy Dissipation by EcoRI

- Measured specific binding constant:

$$K_{spec} = 1.6 \times 10^5$$



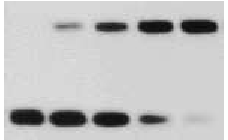
- Average energy dissipated by one molecule as it binds:

$$\Delta G^\circ_{spec} = -k_\mathsf{B} T \ln K_{spec} \qquad \text{(joules per binding)}$$

Energy Dissipation by EcoRI

- Measured specific binding constant:

$$K_{spec} = 1.6 \times 10^5$$



- Average energy dissipated by one molecule as it binds:

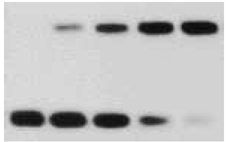
$$\Delta G^{\circ}_{spec} = -k_{\mathsf{B}} T \ln K_{spec} \qquad \text{(joules per binding)}$$

- The Second Law of Thermodynamics as a conversion factor:

$$\mathcal{E}_{min} = k_{\mathsf{B}} T \ln 2 \qquad \text{(joules per bit)}$$

Energy Dissipation by EcoRI

- Measured specific binding constant:



$$K_{spec} = 1.6 \times 10^5$$

- Average energy dissipated by one molecule as it binds:

$$\Delta G^{\circ}_{spec} = -k_{\mathsf{B}} T \ln K_{spec} \qquad \text{(joules per binding)}$$

- The Second Law of Thermodynamics as a conversion factor:

$$\mathcal{E}_{min} = k_{\mathsf{B}} T \ln 2 \qquad \text{(joules per bit)}$$

- Number of bits that could have been selected:

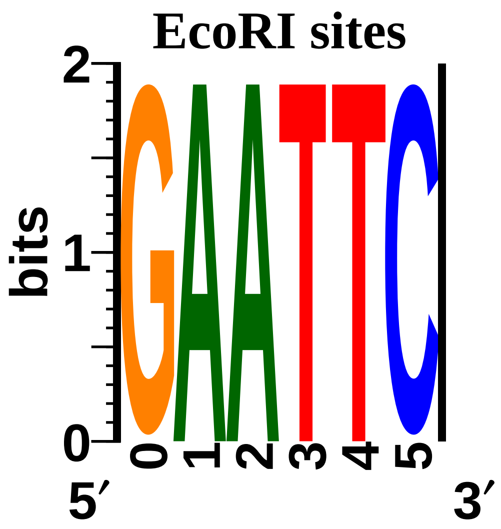
$$\begin{aligned} R_{energy} &= -\Delta G^{\circ} / \mathcal{E}_{min} \\ &= k_{\mathsf{B}} T \ln K_{spec} / k_{\mathsf{B}} T \ln 2 \\ &= \log_2 K_{spec} \qquad \Leftarrow \text{SO SIMPLE!} \\ &= \boxed{\text{17.3 bits per binding}} \end{aligned}$$

Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices

Information/Energy = Efficiency of EcoRI

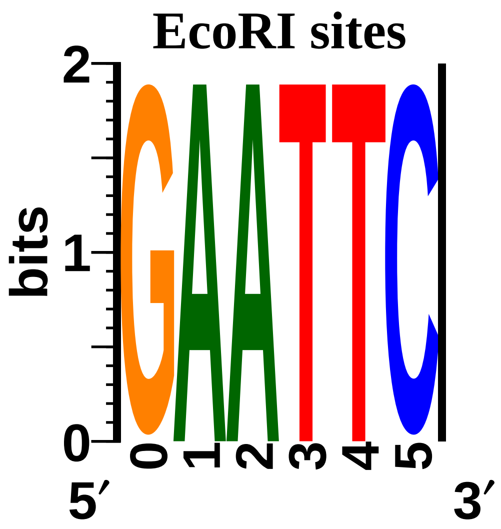
EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.



Information/Energy = Efficiency of EcoRI

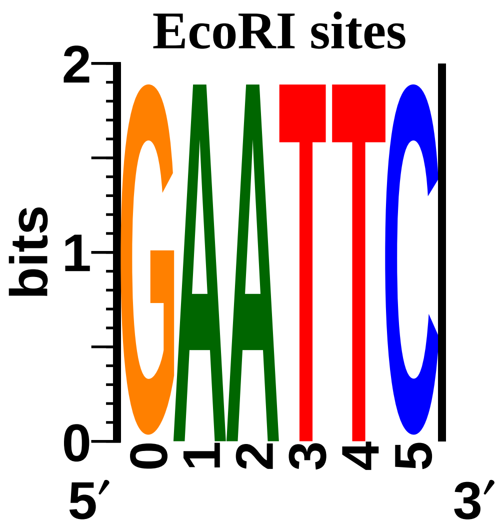
EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.

Efficiency is
'WORK' DONE / ENERGY DISSIPATED



Information/Energy = Efficiency of EcoRI

EcoRI could have made 17.3 binary choices
... but it only made 12 choices.



Efficiency is
'WORK' DONE / ENERGY DISSIPATED

$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

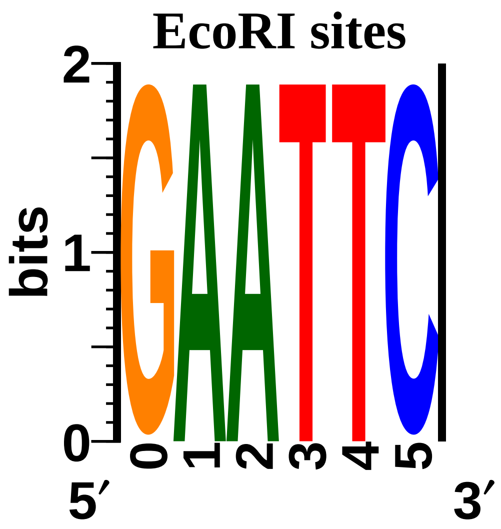
Information/Energy = Efficiency of EcoRI = 70%

EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.

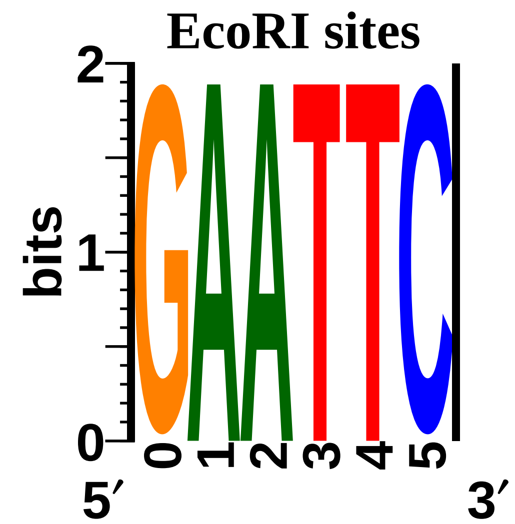
Efficiency is
'WORK' DONE / ENERGY DISSIPATED

$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

The efficiency is 70%.



EcoRI could have made 17.3 binary choices
. . . but it only made 12 choices.



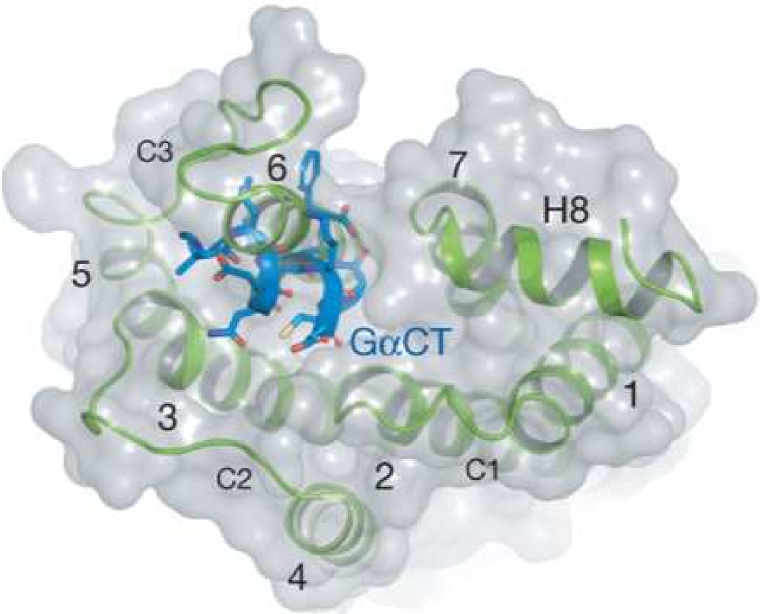
Efficiency is
'WORK' DONE / ENERGY DISSIPATED

$$\frac{12 \text{ bits per binding}}{17.3 \text{ bits per binding}} = 0.7$$

The efficiency is 70%.

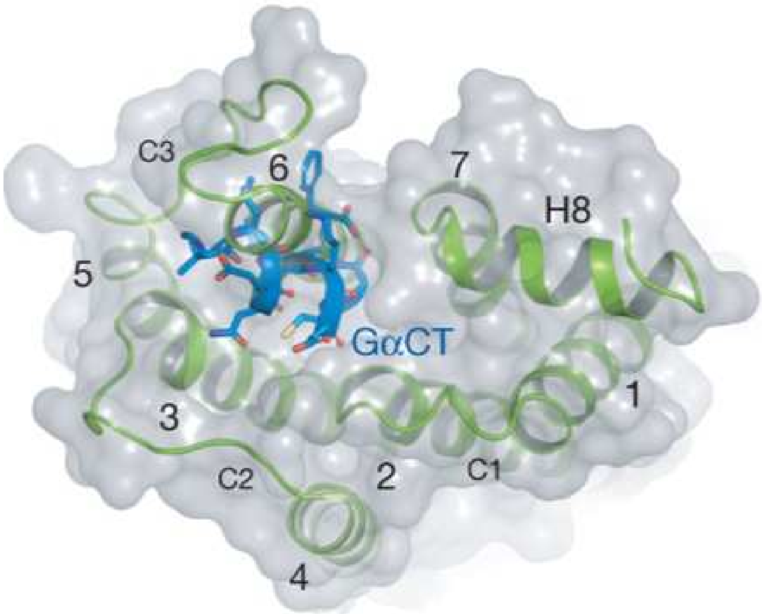
18 out of 19 DNA binding proteins give ∼70% efficiency.

Dark State



Rhodopsin Shape Change

Dark State



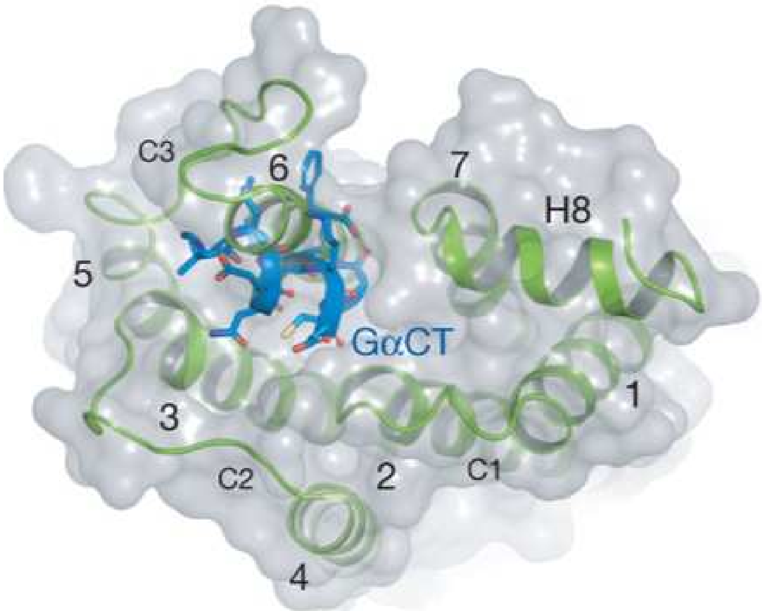
$$\mathbf{h}\nu$$

P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

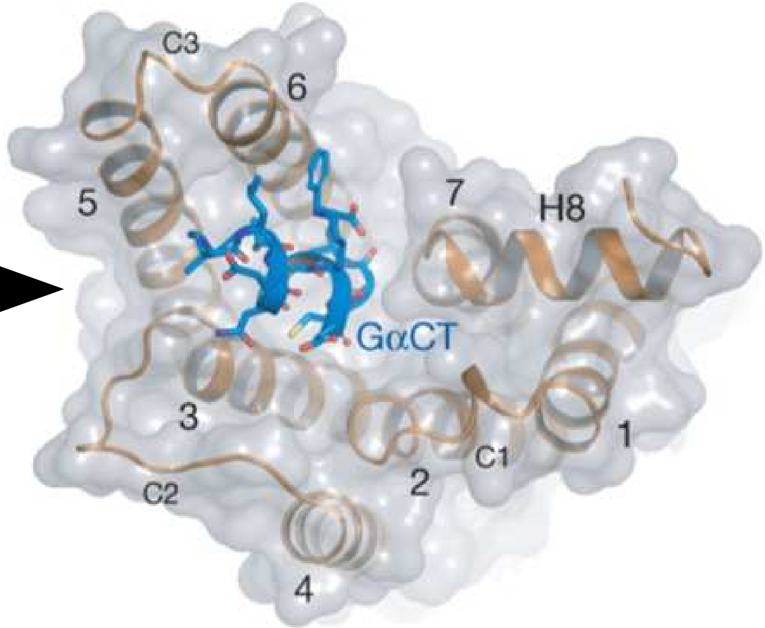
Rhodopsin Shape Change

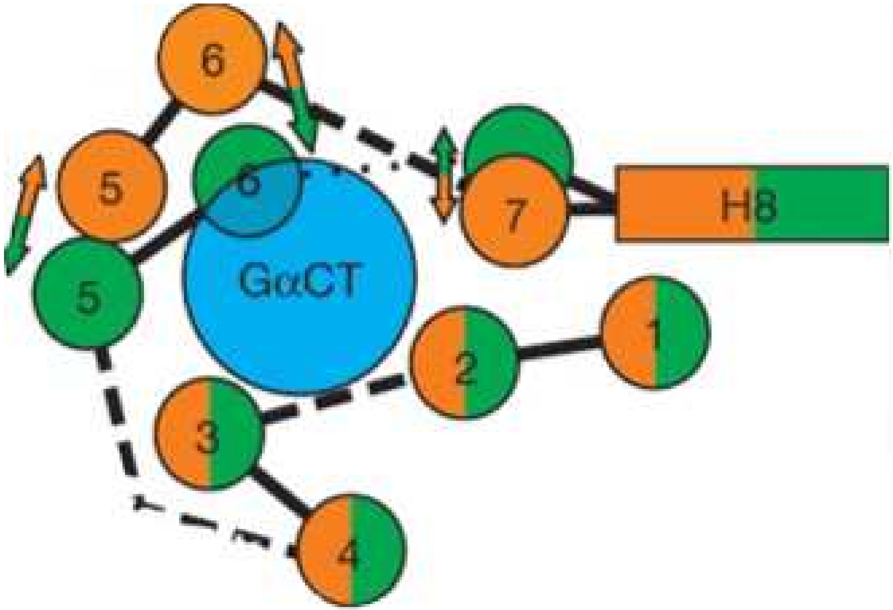
Dark State

After Photon - Light State



$\mathbf{h}\nu$



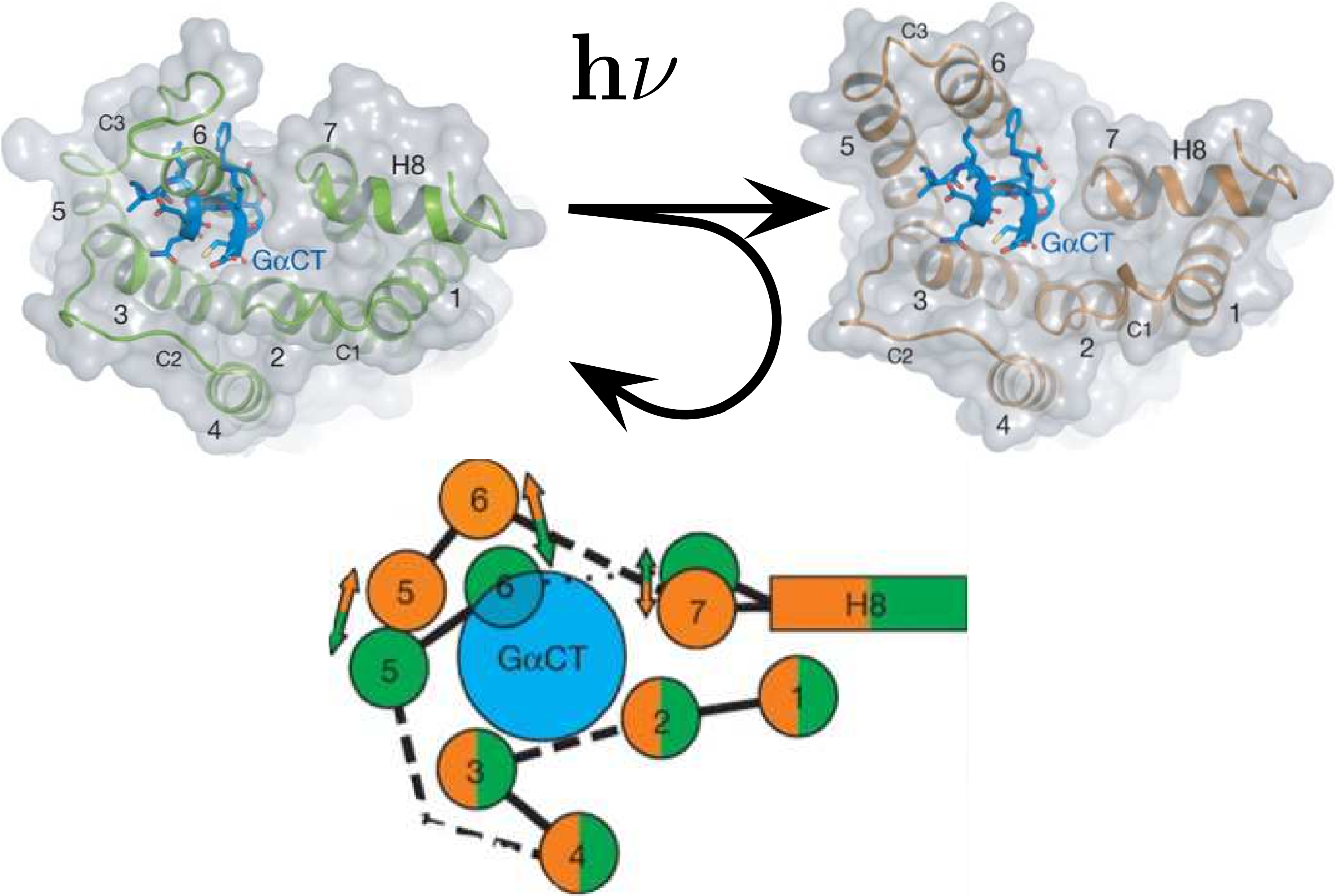


P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Rhodopsin Shape Change

Dark State

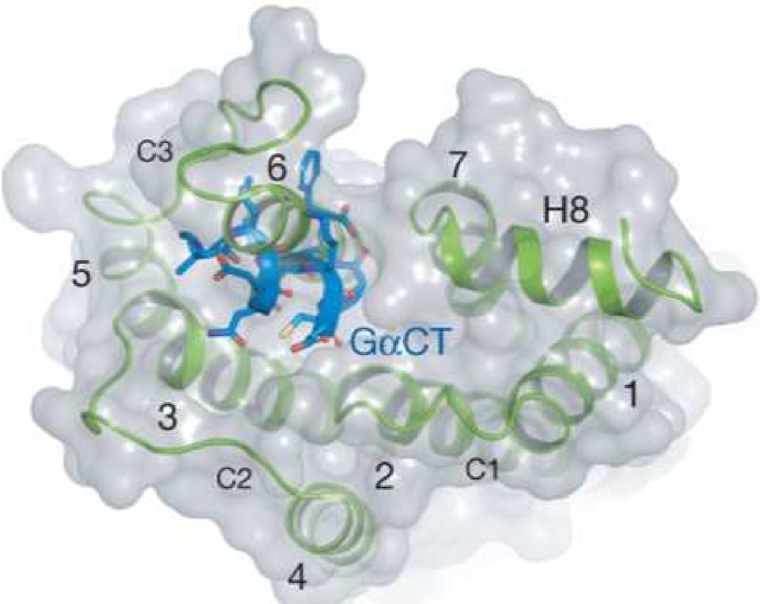
After Photon - Light State



Rhodopsin Shape Change

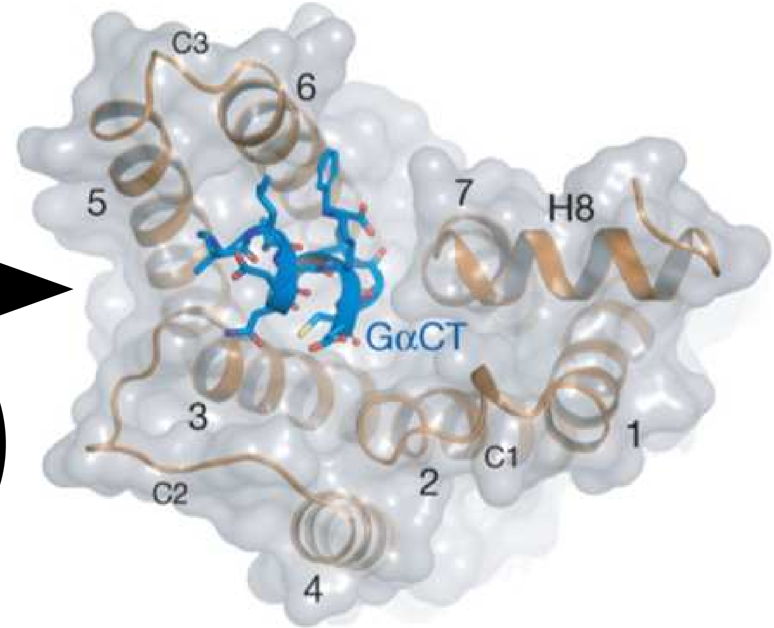
Dark State

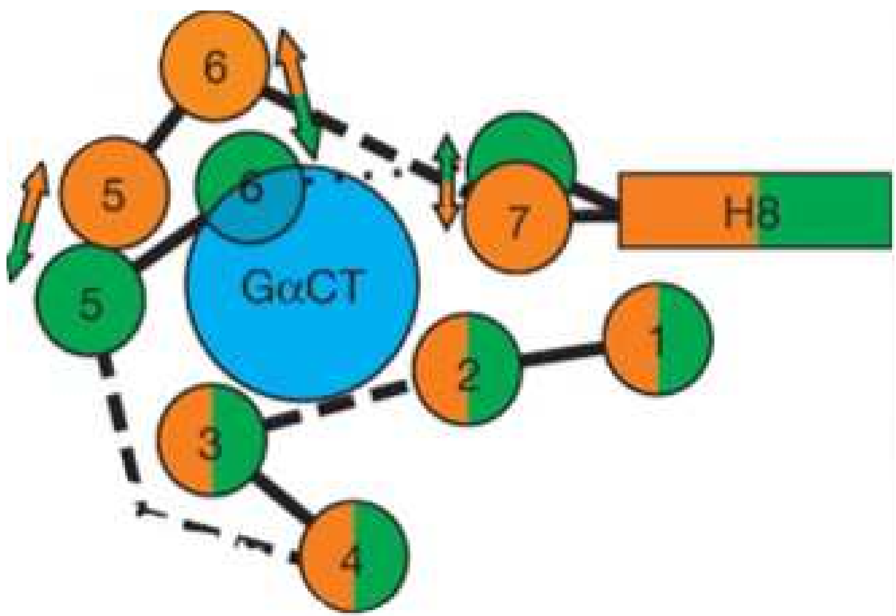
After Photon - Light State



$\mathbf{h}\nu$

70%

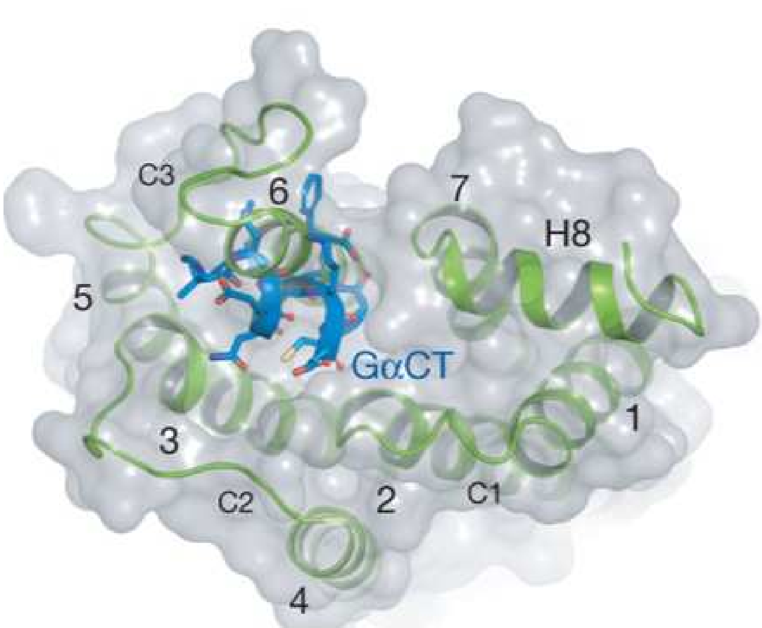




P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Rhodopsin Shape Change

Dark State

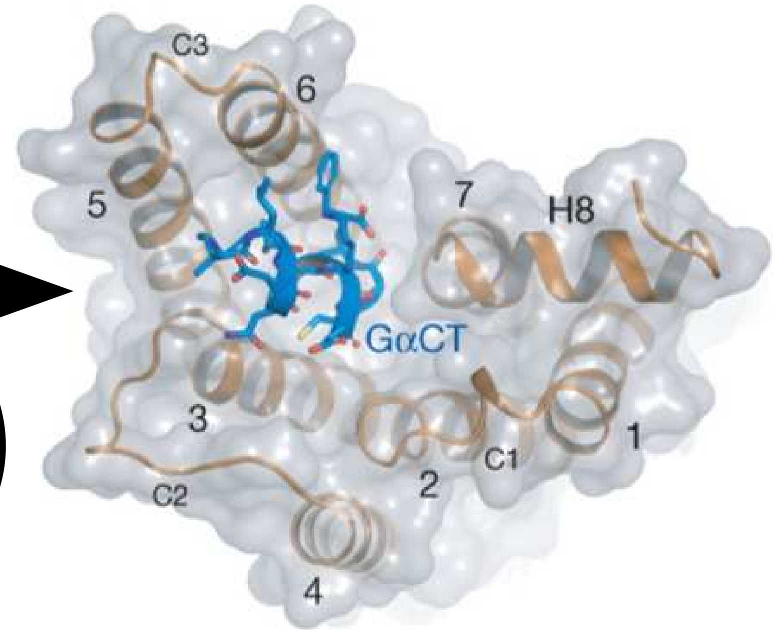


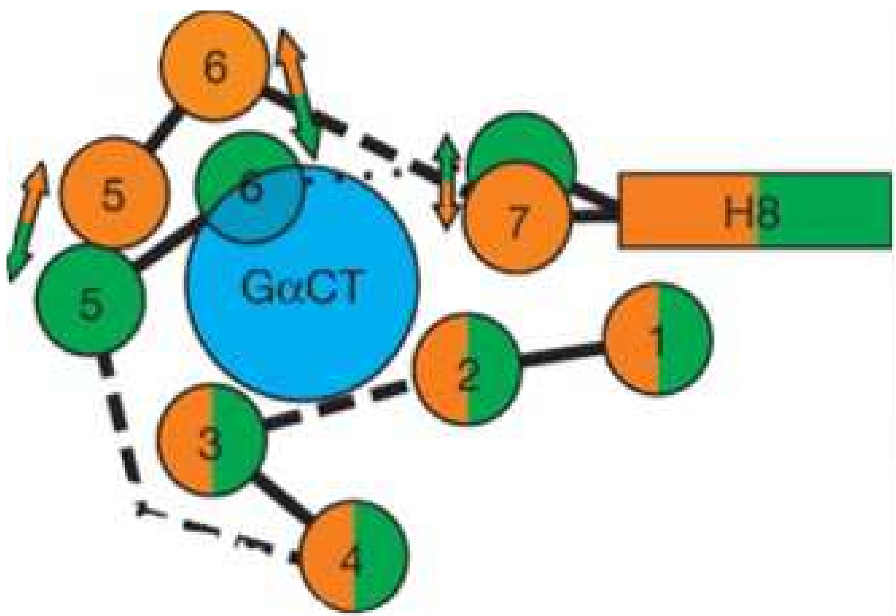
hν

70%

30%

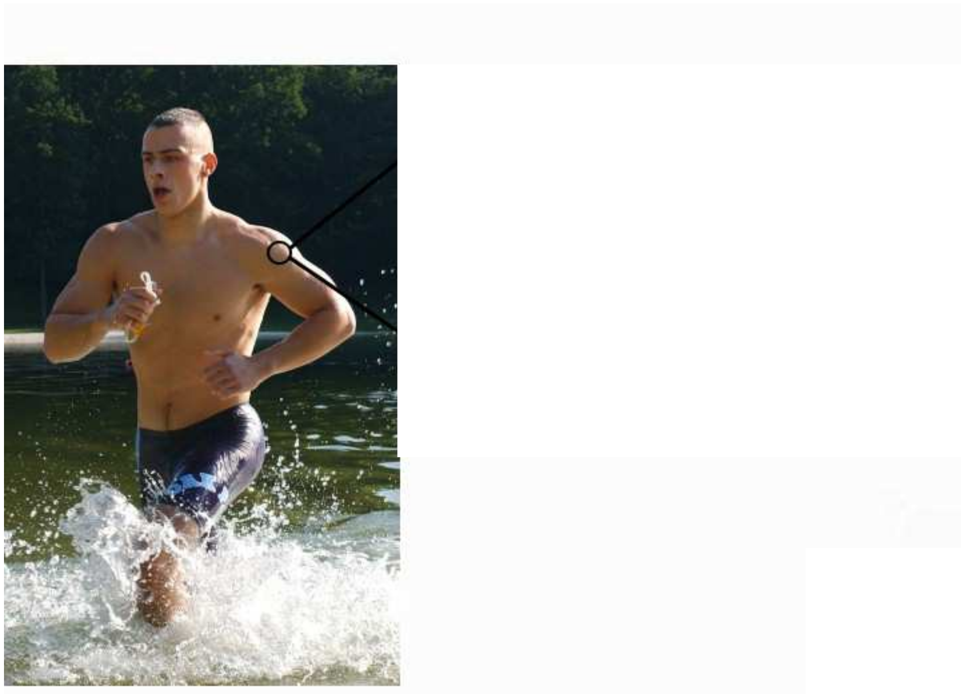
After Photon - Light State





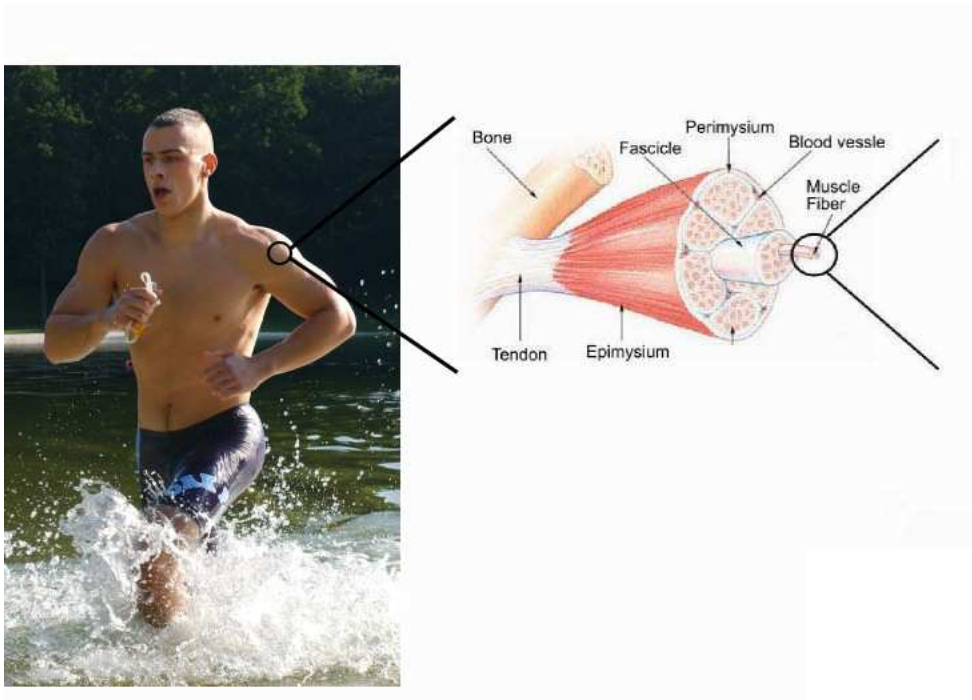
P. Scheerer et al. Nature 455, 497-502(2008) doi:10.1038/nature07330

Muscle Structure

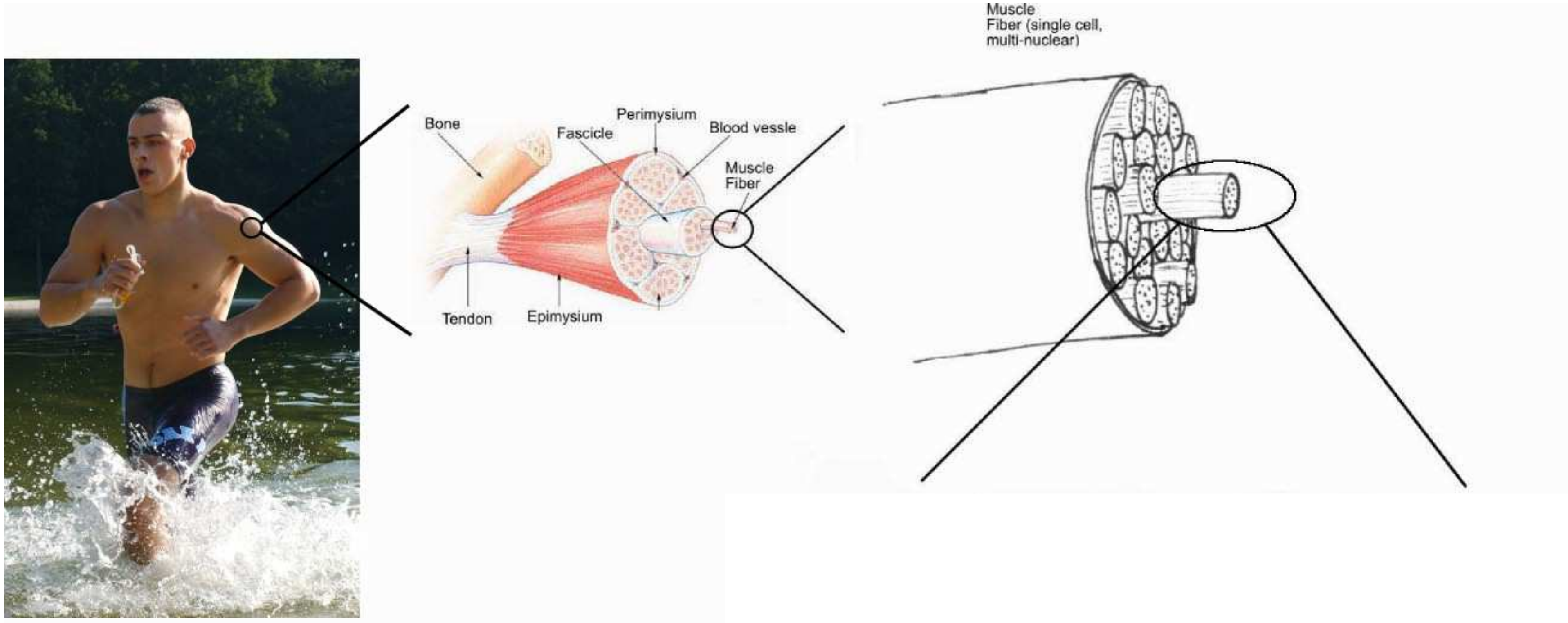


https://en.wikipedia.org/wiki/Muscle

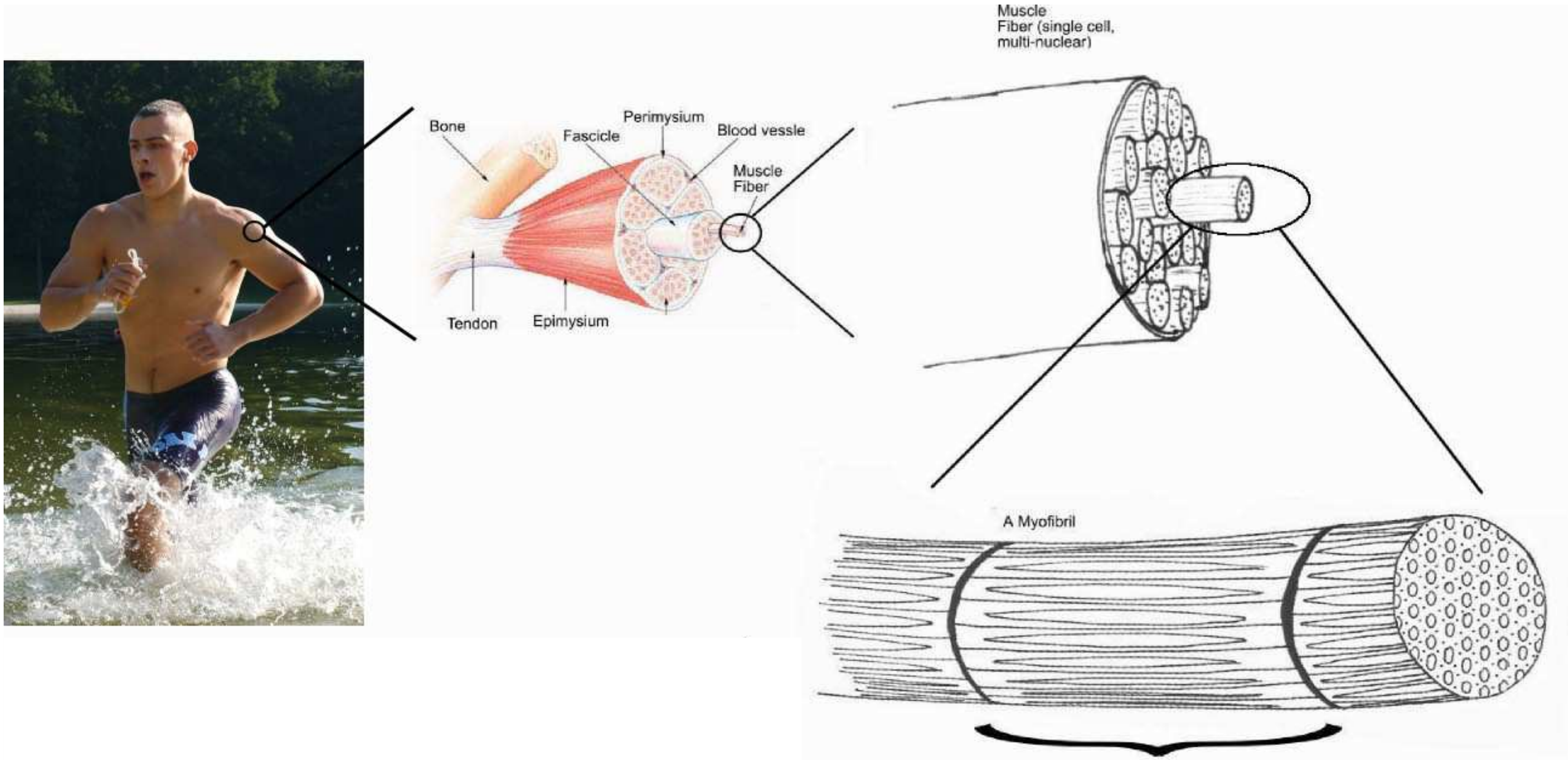
Muscle Structure



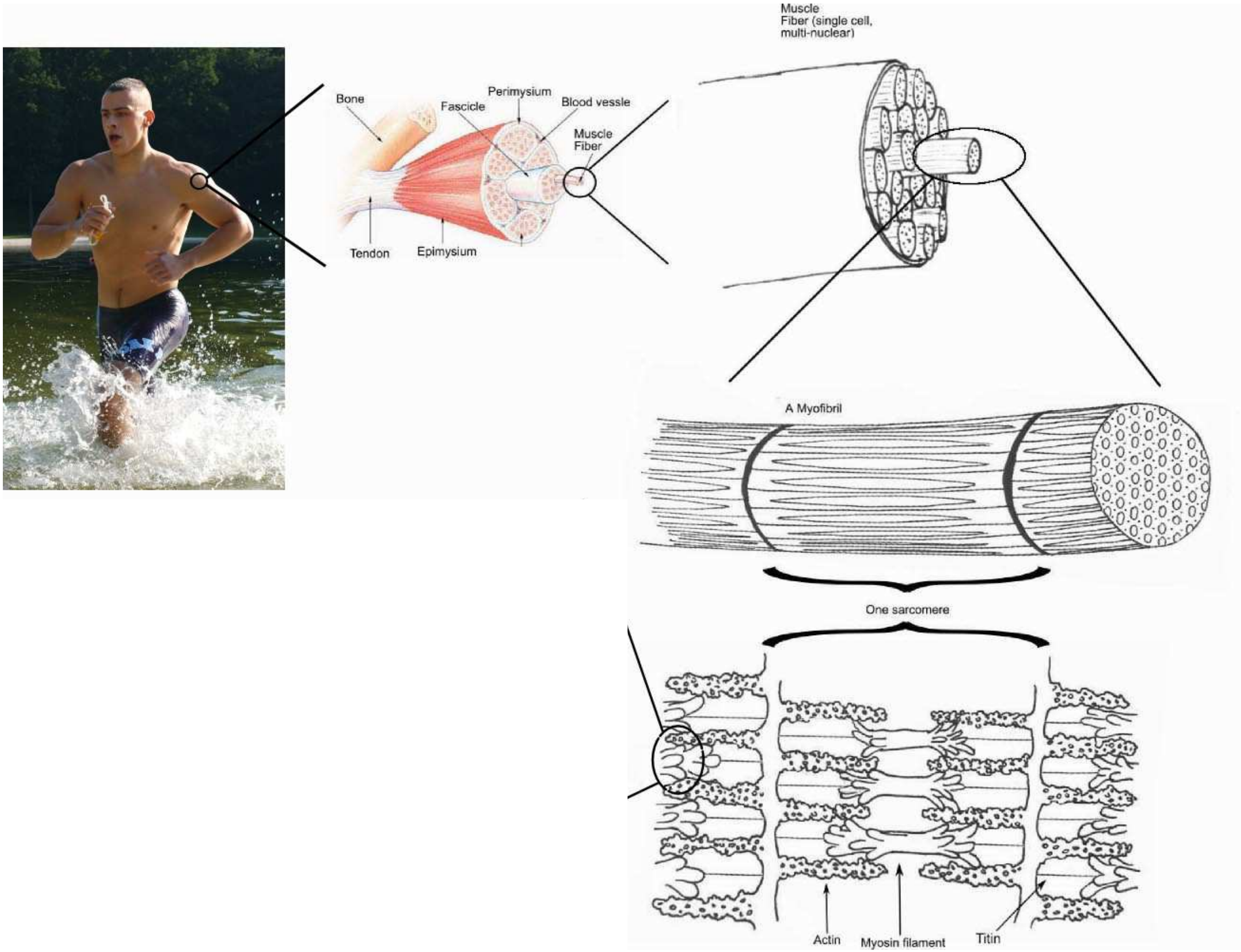
Muscle Structure



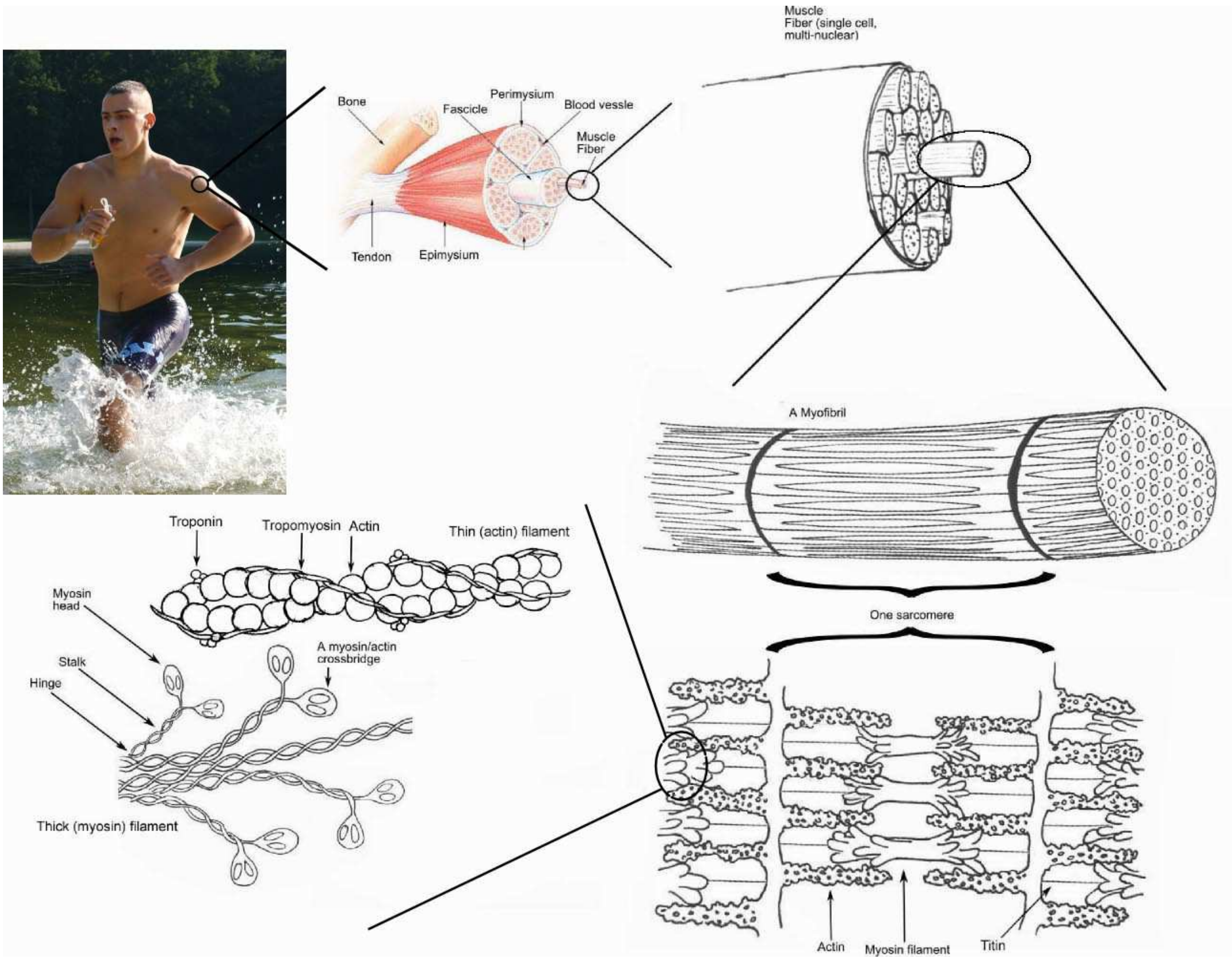
Muscle Structure



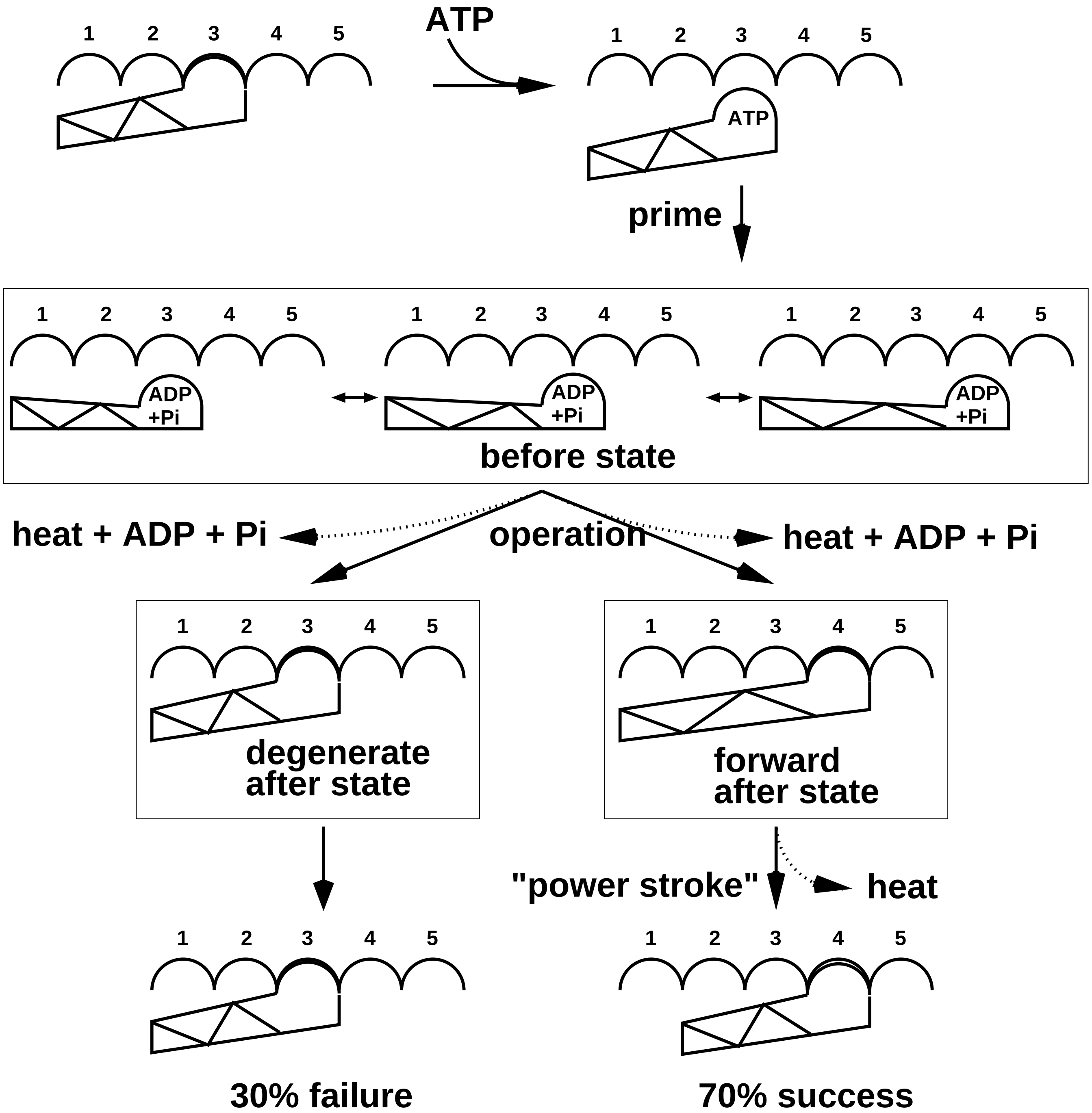
https://en.wikipedia.org/wiki/Muscle


Muscle
Fiber (single cell,
multi-nuclear)
Bone
Fascicle
Perimysium
Blood vessle
Muscle
Fiber
Tendon
Epimysium
A Myofibril
One sarcomere
Actin
Myosin filament
Titin

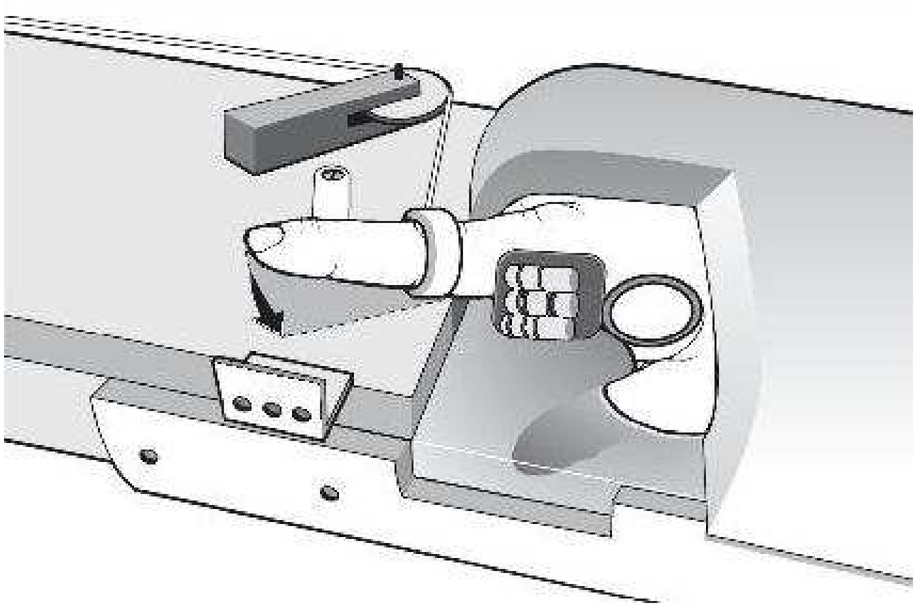
Muscle Structure

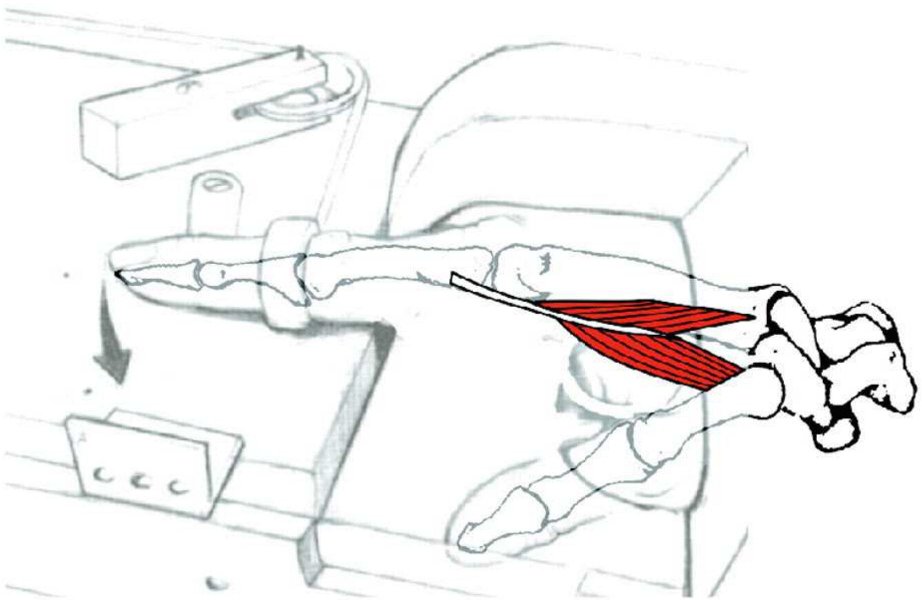


Tom's Model of Muscle Mechanism



Efficiency of Muscle

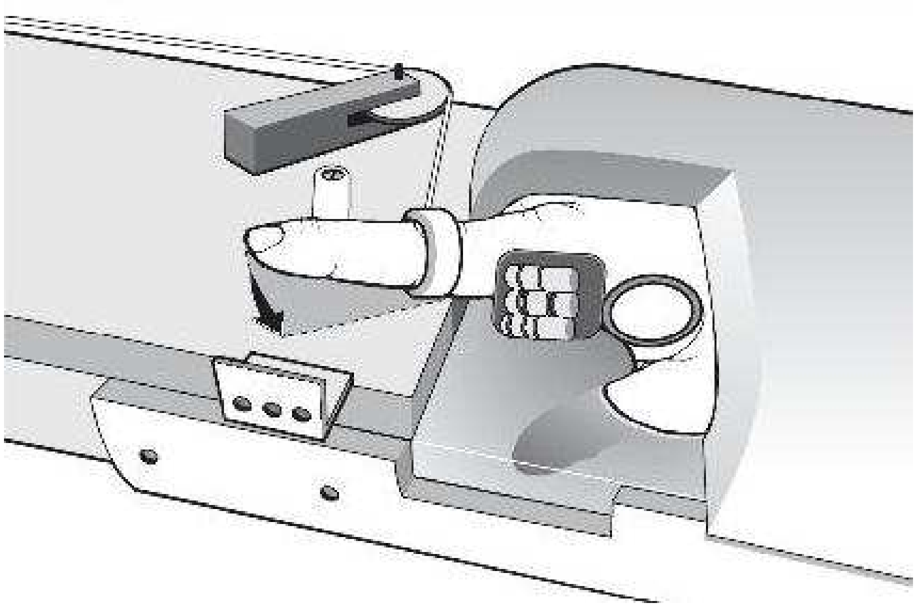


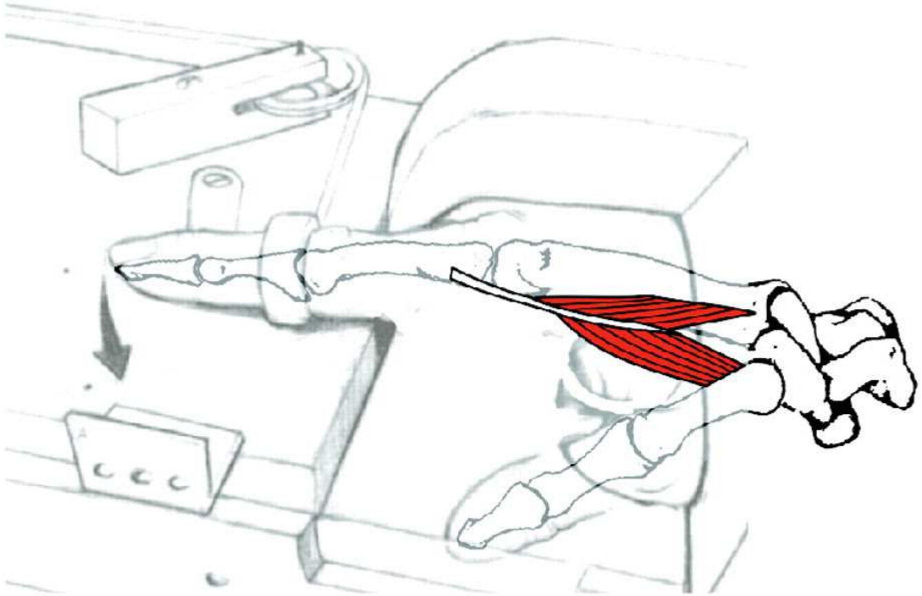


- Experiments by Kushmerick's lab since (at least) 1969

http://dx.doi.org/10.1113/jphysiol.2007.146829 http://dx.doi.org/10.1242/jeb.052985

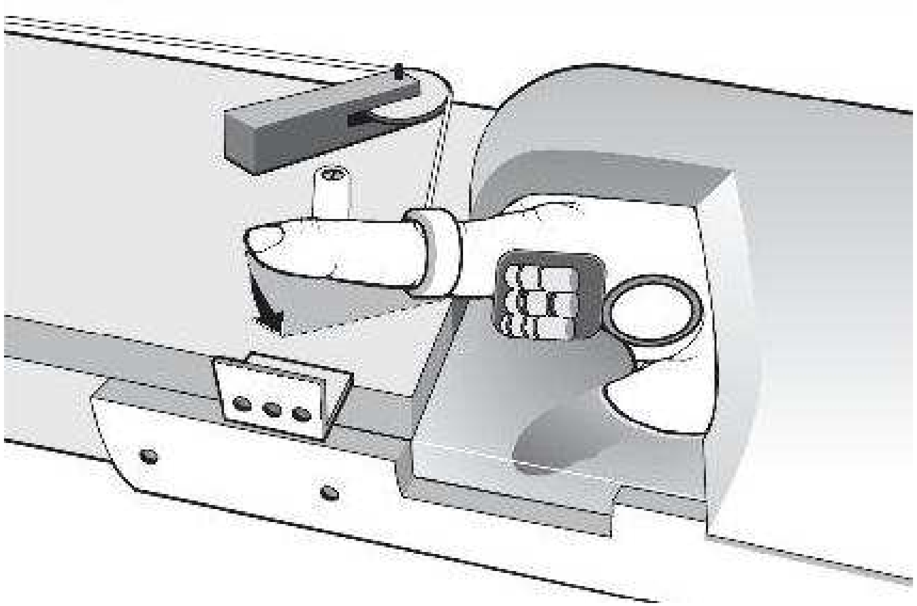
Efficiency of Muscle

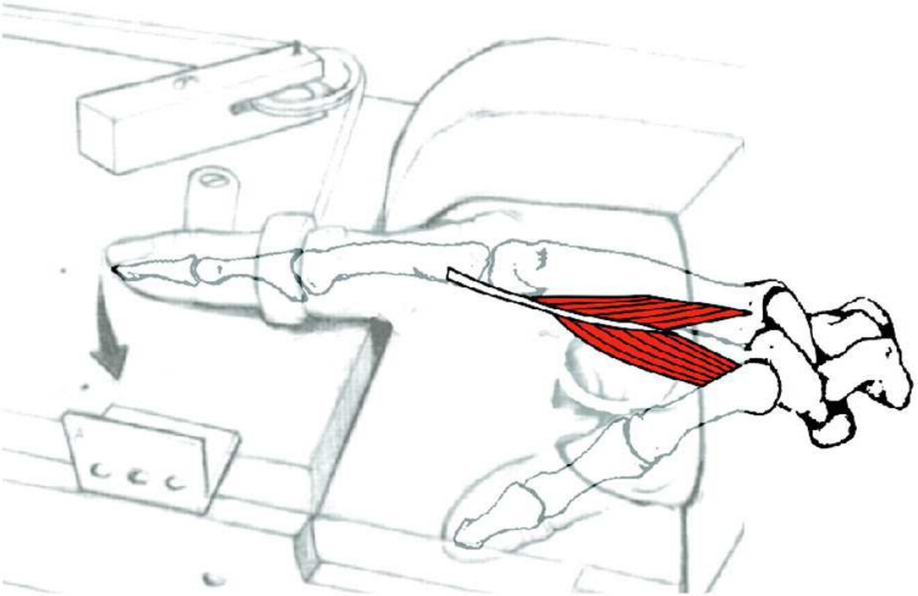




- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011

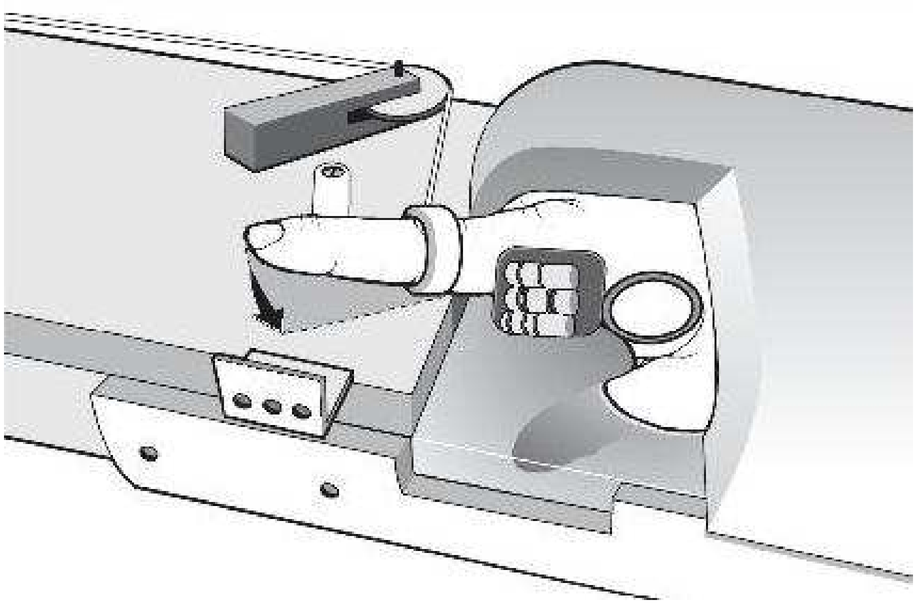
Efficiency of Muscle

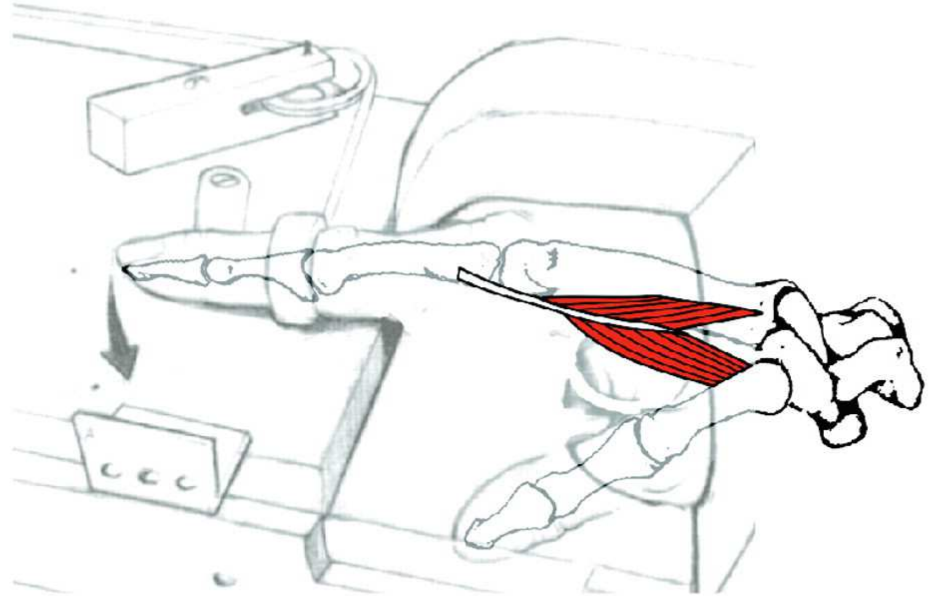




- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011
- Weight lifting gives work done

Efficiency of Muscle

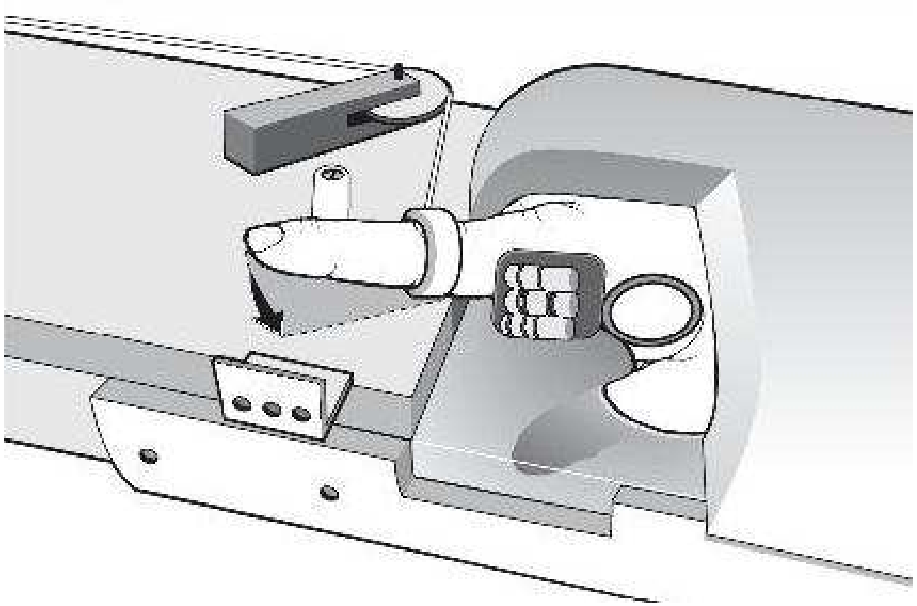


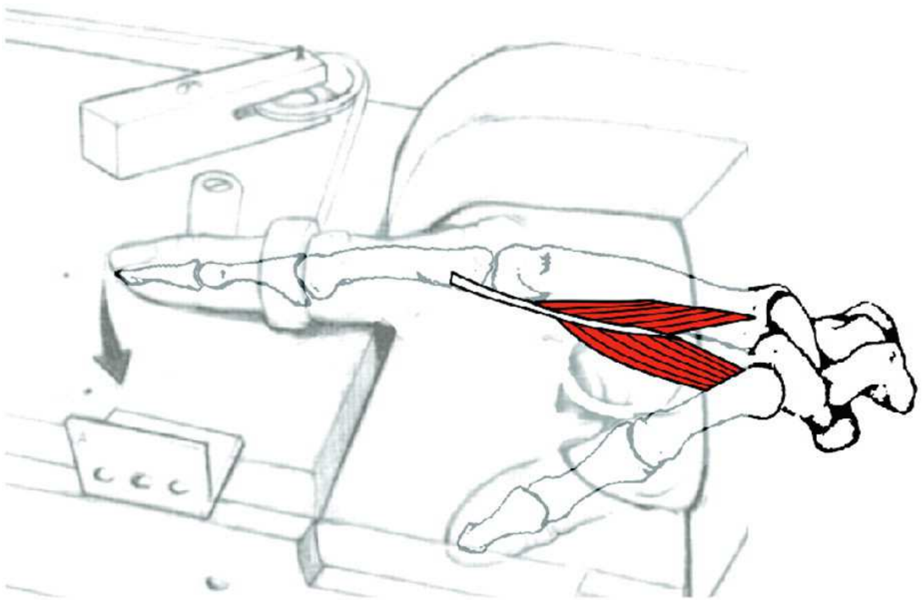


- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011
- Weight lifting gives work done
- NMR coil gives ATP = energy used

http://dx.doi.org/10.1113/jphysiol.2007.146829 http://dx.doi.org/10.1242/jeb.052985

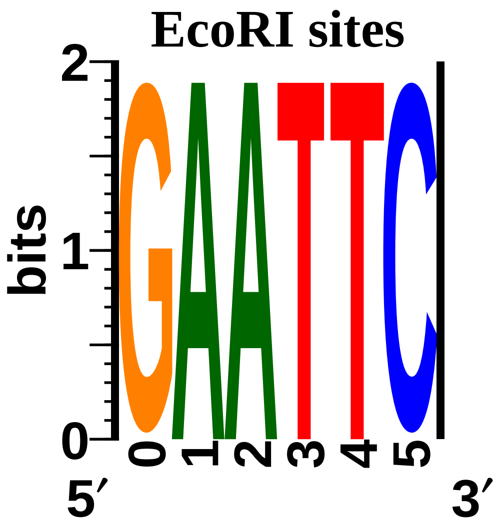
Efficiency of Muscle

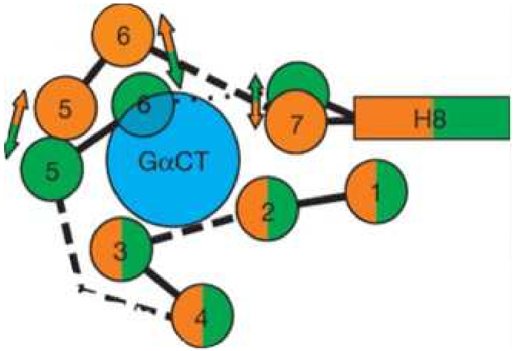




- Experiments by Kushmerick's lab since (at least) 1969
- new work: 2008, 2011
- Weight lifting gives work done
- NMR coil gives ATP = energy used
- $\boxed{\text{Efficiency: } 0.68 \pm 0.09}$

70% efficiency appears widely in biology:

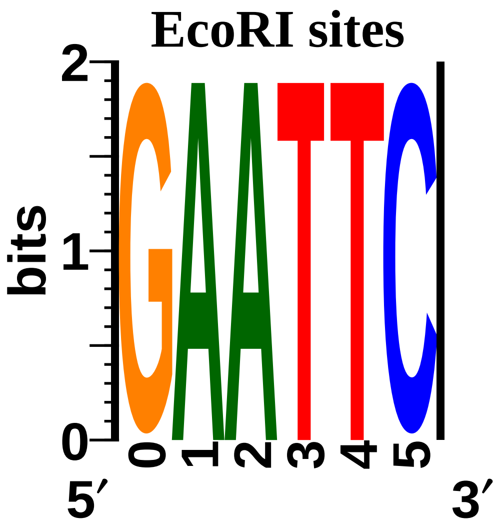


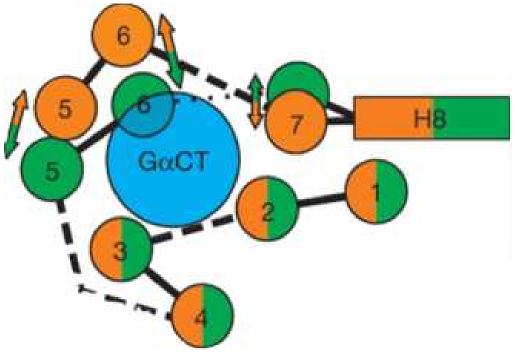


Why are molecular machines 70% efficient?

70% efficiency appears widely in biology:

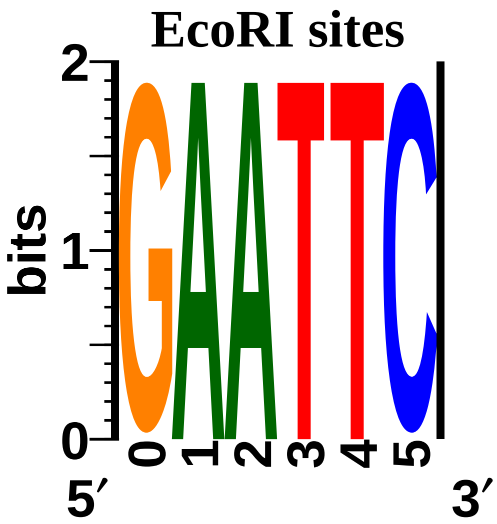
- DNA - protein binding

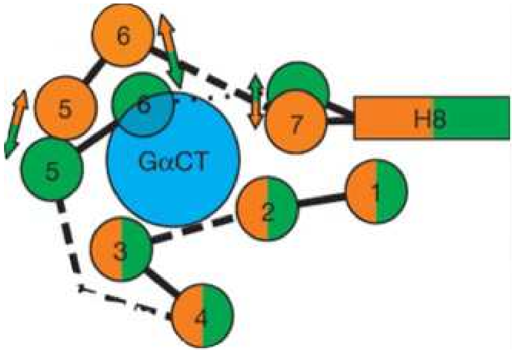




70% efficiency appears widely in biology:

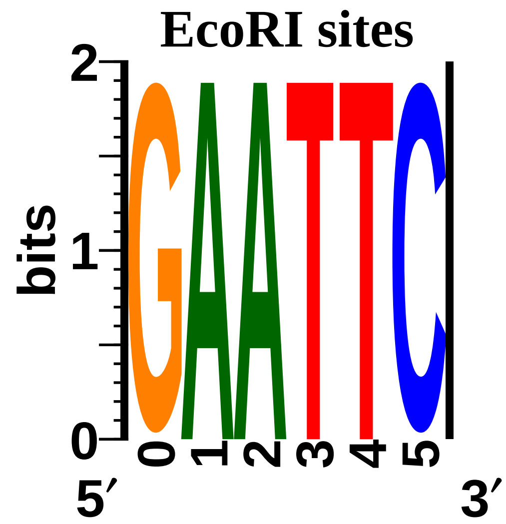
- DNA - protein binding
- rhodopsin

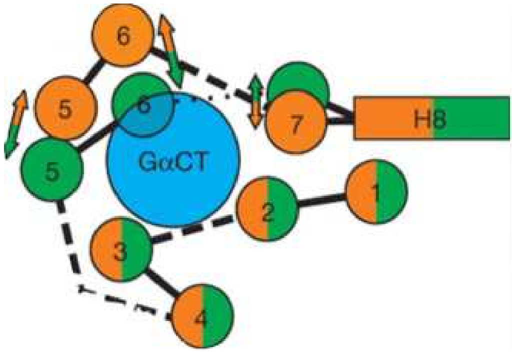




70% efficiency appears widely in biology:

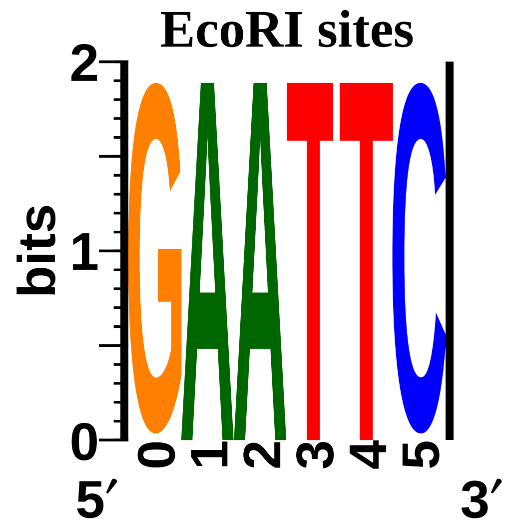
- DNA - protein binding
- rhodopsin
- muscle

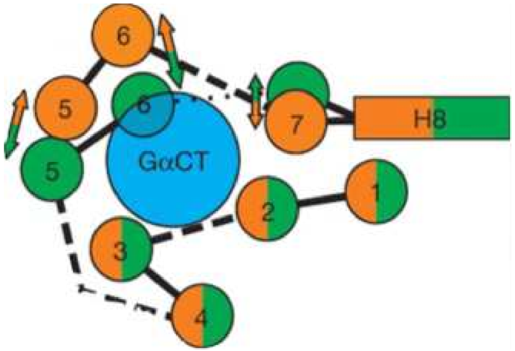




70% efficiency appears widely in biology:

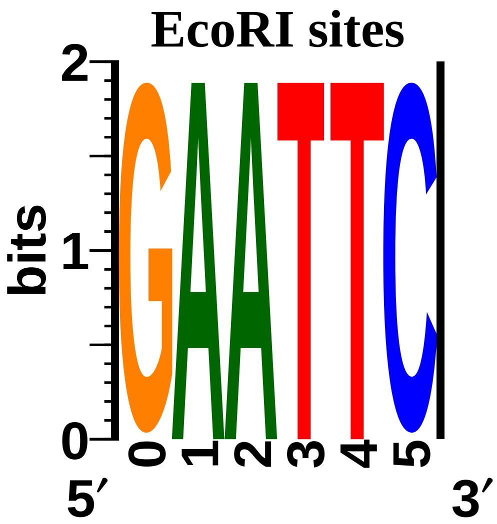
- DNA - protein binding
- rhodopsin
- muscle
- other systems



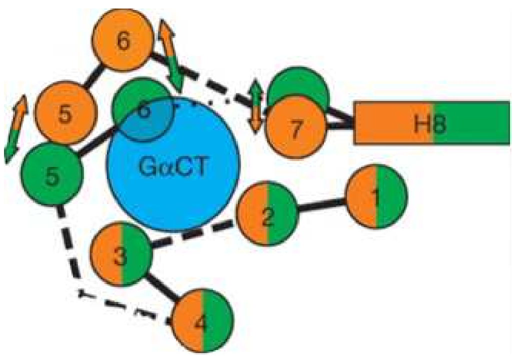


70% efficiency appears widely in biology:

- DNA - protein binding
- rhodopsin
- muscle
- other systems

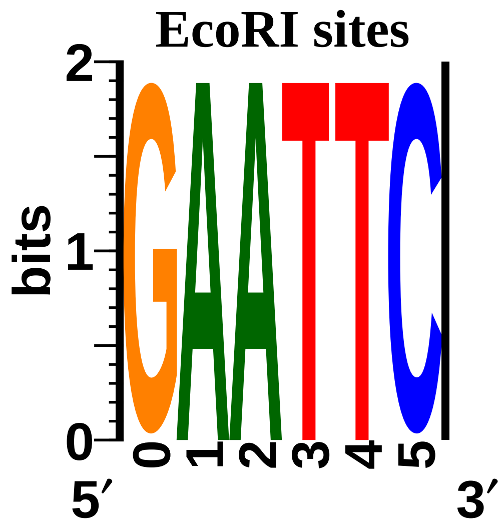


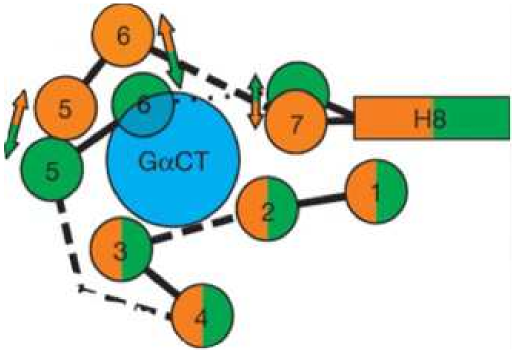
Why 70% efficiency?



70% efficiency appears widely in biology:

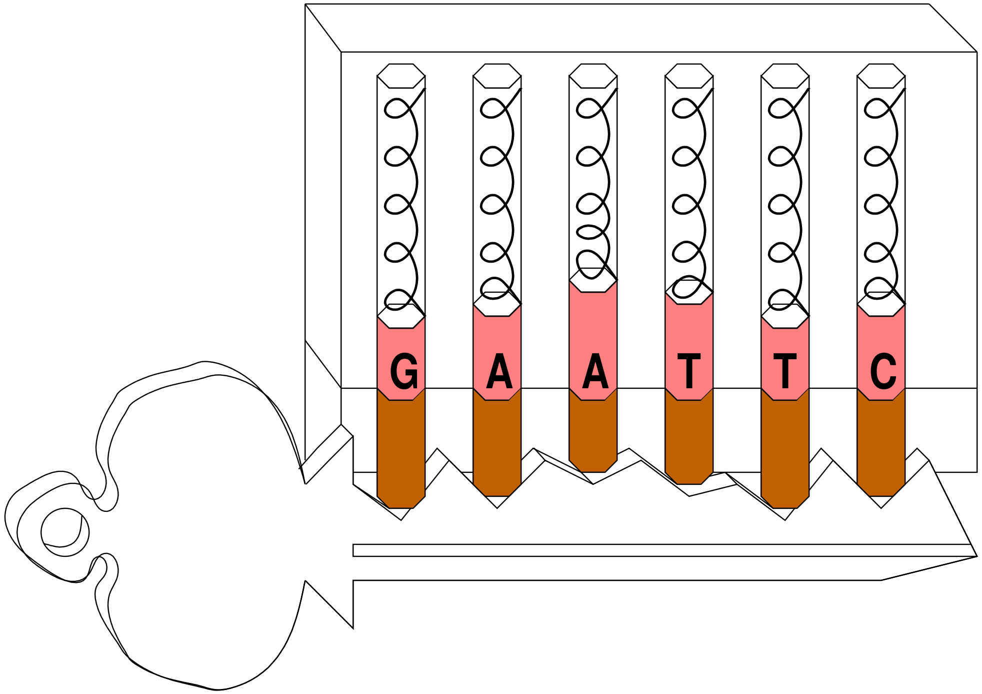
- DNA - protein binding
- rhodopsin
- muscle
- other systems





Why 70% efficiency?

Information theory explanation

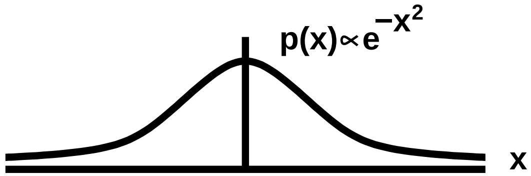


**Like a key in a lock
which has many independent pins,
it takes many numbers
to describe the vibrational state
of a molecular machine**

Gaussians

- Pin motion x has a Gaussian distribution:

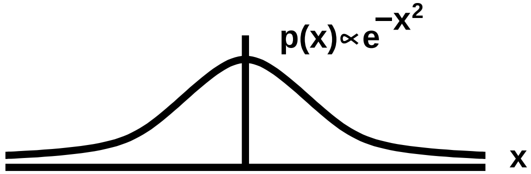
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ = mean, σ = standard deviation

- Pin motion x has a Gaussian distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

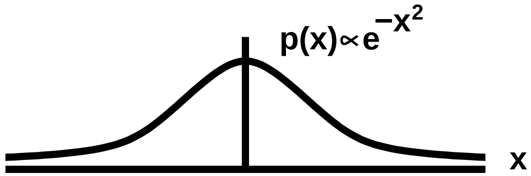


μ = mean, σ = standard deviation

- Gaussian distributions are generated by the sum of many small random variables

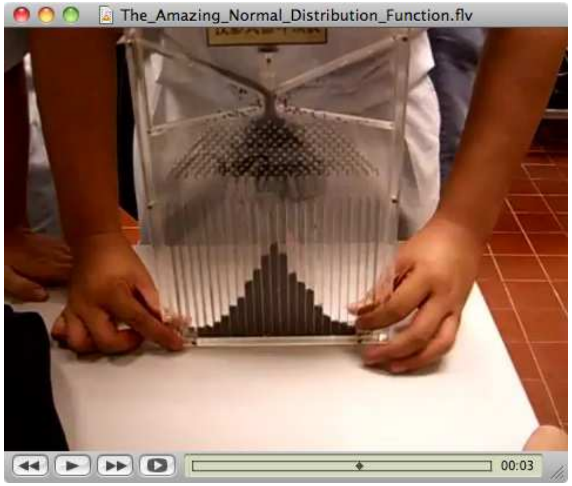
- Pin motion x has a Gaussian distribution:

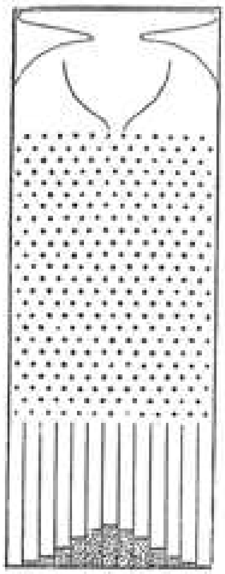
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



μ = mean, σ = standard deviation

- Gaussian distributions are generated by the sum of many small random variables
- Drunkard's walk: Galton's quincunx device!





Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{2}$$

Two Gaussians

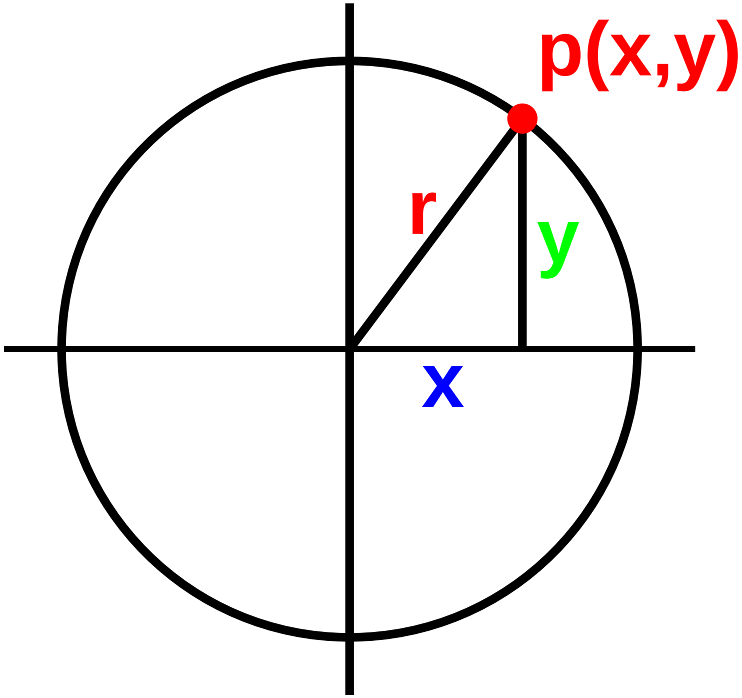
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

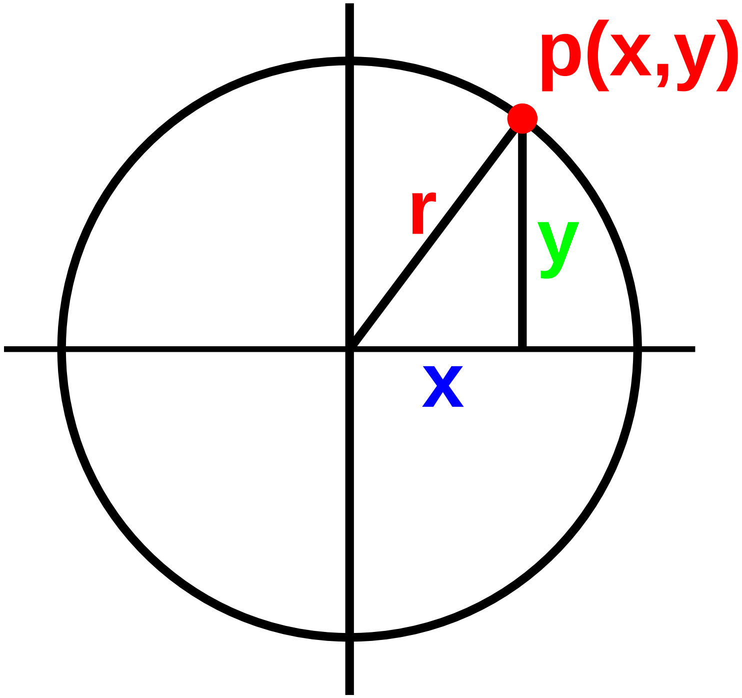
$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$



Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$

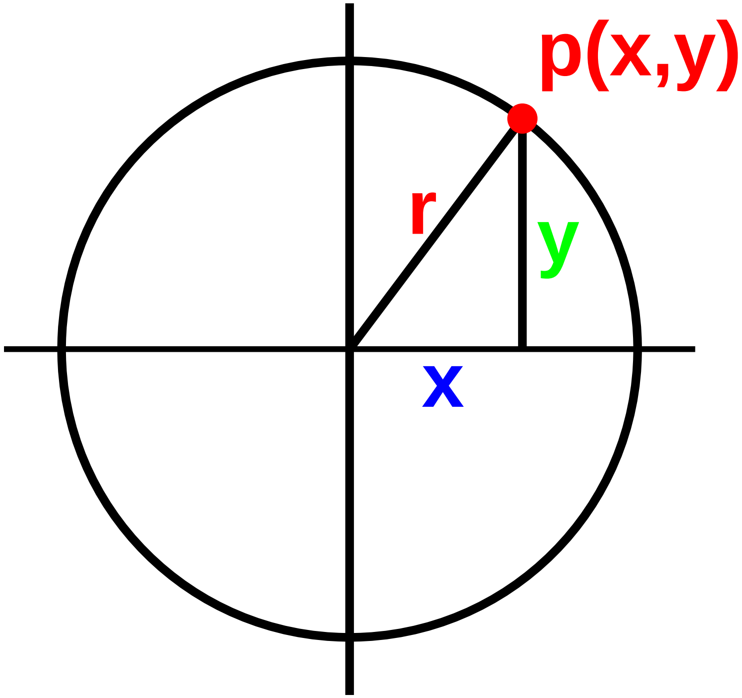


$$p(x,y) = \; p(x) \times p(y) \tag{3}$$

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$



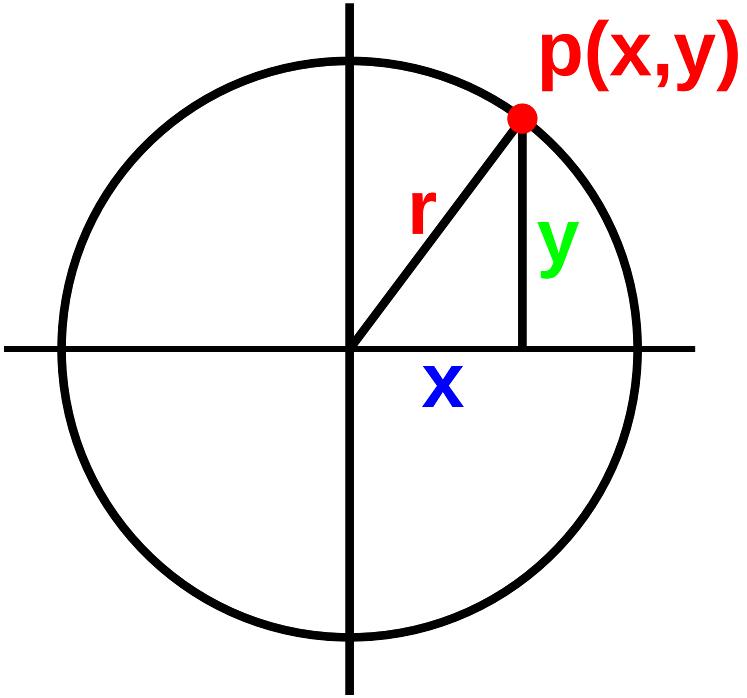
$$p(x,y) = p(x) \times p(y) \quad (3)$$

$$\propto e^{-x^2} \times e^{-y^2} \quad (4)$$

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$

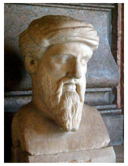


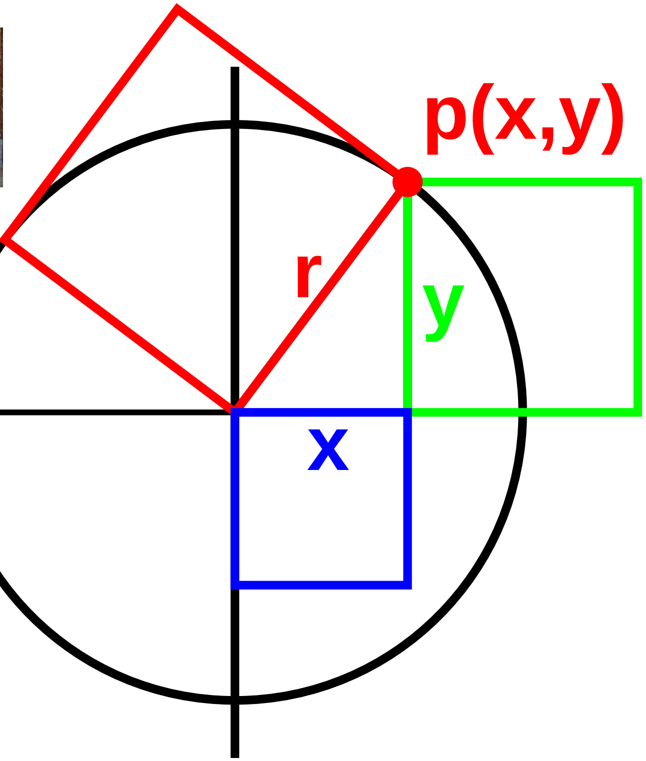
$$\begin{aligned} p(x,y) &= \; p(x) \times p(y) && (3)\\ &\propto \; e^{-x^2} \times e^{-y^2} && (4)\\ &\propto \; e^{-(x^2+y^2)} && (5) \end{aligned}$$

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \quad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \quad (2)$$



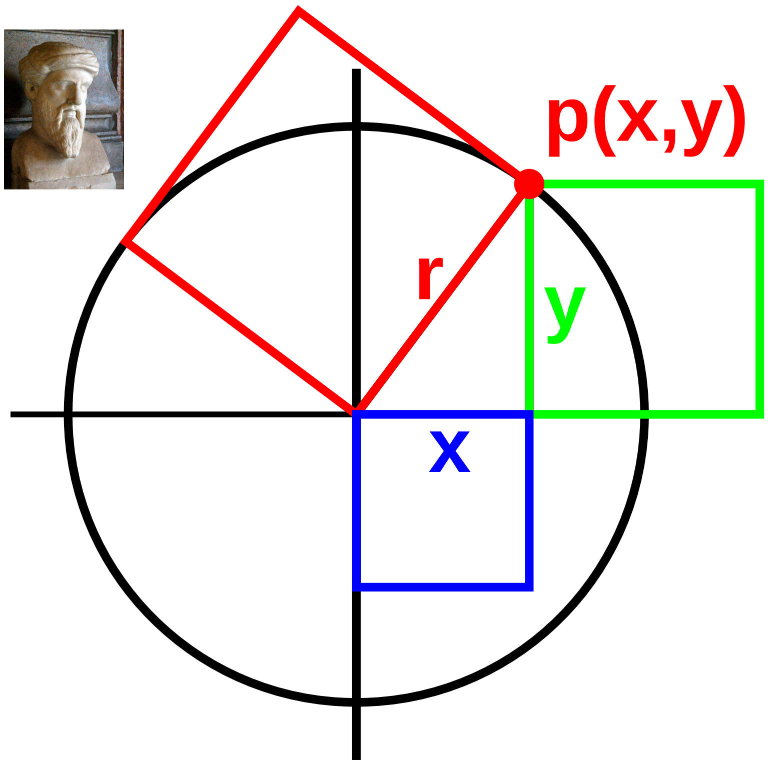


$$\begin{aligned} p(x,y) &= p(x) \times p(y) && (3)\\ &\propto e^{-x^2} \times e^{-y^2} && (4)\\ &\propto e^{-(x^2+y^2)} && (5)\\ &\propto e^{-r^2} && (6) \end{aligned}$$

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \qquad (1)$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \qquad (2)$$



$$\begin{aligned} p(x,y) &= p(x) \times p(y) && (3) \\ &\propto e^{-x^2} \times e^{-y^2} && (4) \\ &\propto e^{-(x^2+y^2)} && (5) \\ &\propto e^{-r^2} && (6) \end{aligned}$$

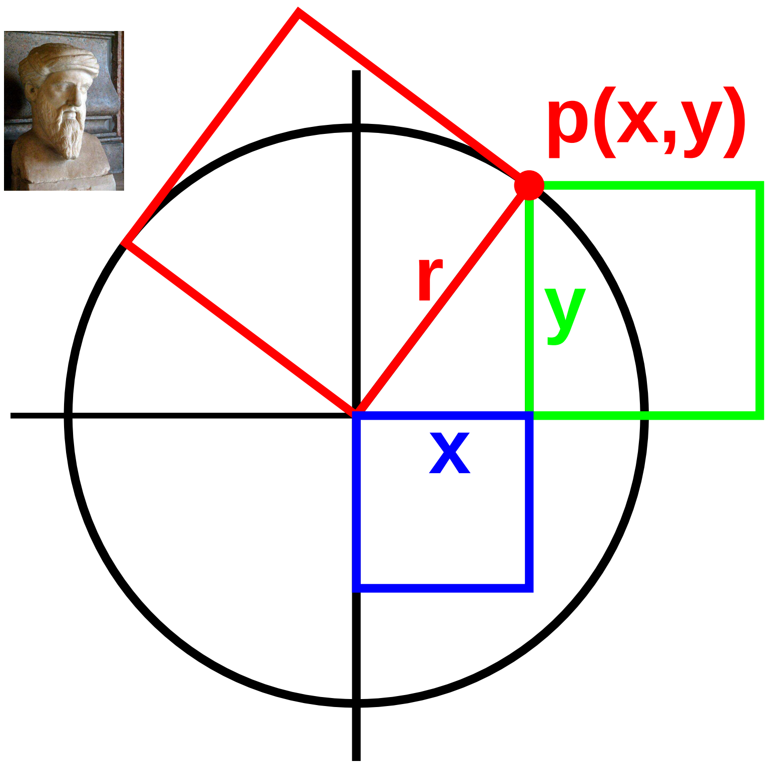
If $p(x,y)$ is a constant, then r is a constant.

credit: http://en.wikipedia.org/wiki/Pythagoras

Two Gaussians

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \propto e^{-x^2} \tag{1}$$

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \propto e^{-y^2} \tag{2}$$

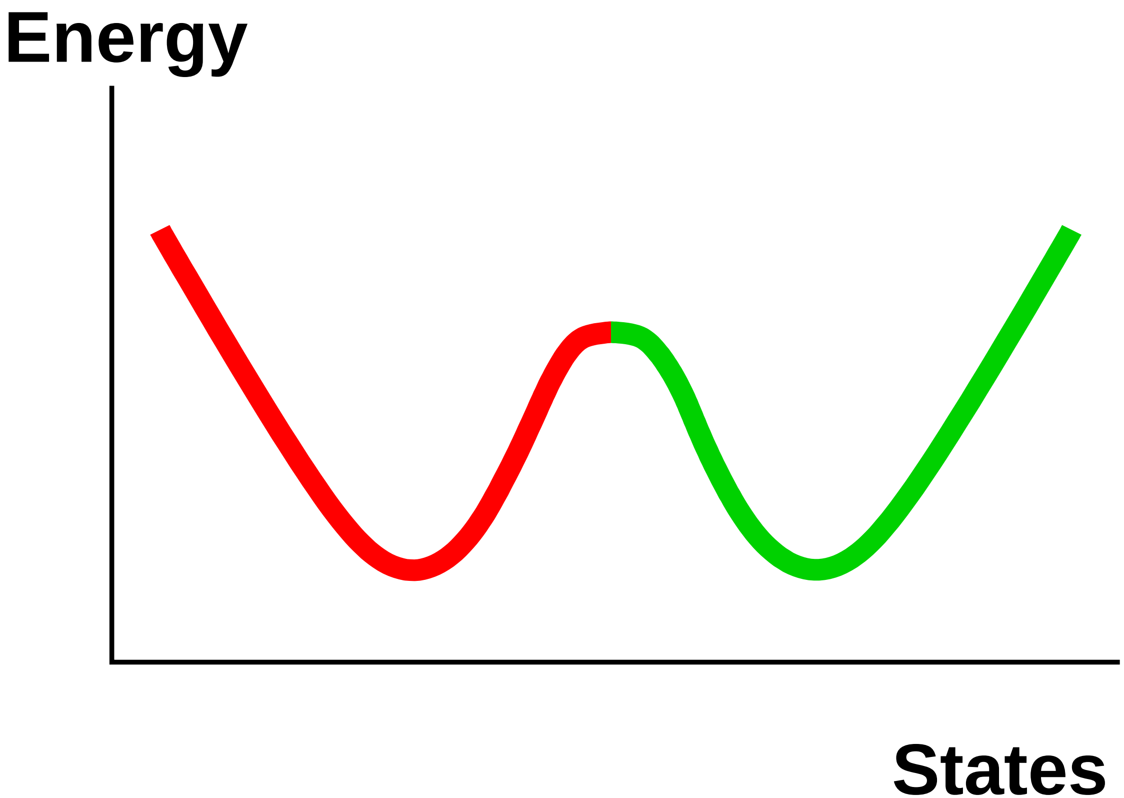


$$\begin{aligned} p(x,y) &= p(x) \times p(y) && (3)\\ &\propto e^{-x^2} \times e^{-y^2} && (4)\\ &\propto e^{-(x^2+y^2)} && (5)\\ &\propto e^{-r^2} && (6) \end{aligned}$$

If $p(x,y)$ is a constant,
then r is a constant.

$\boxed{\text{Circular distribution!}}$

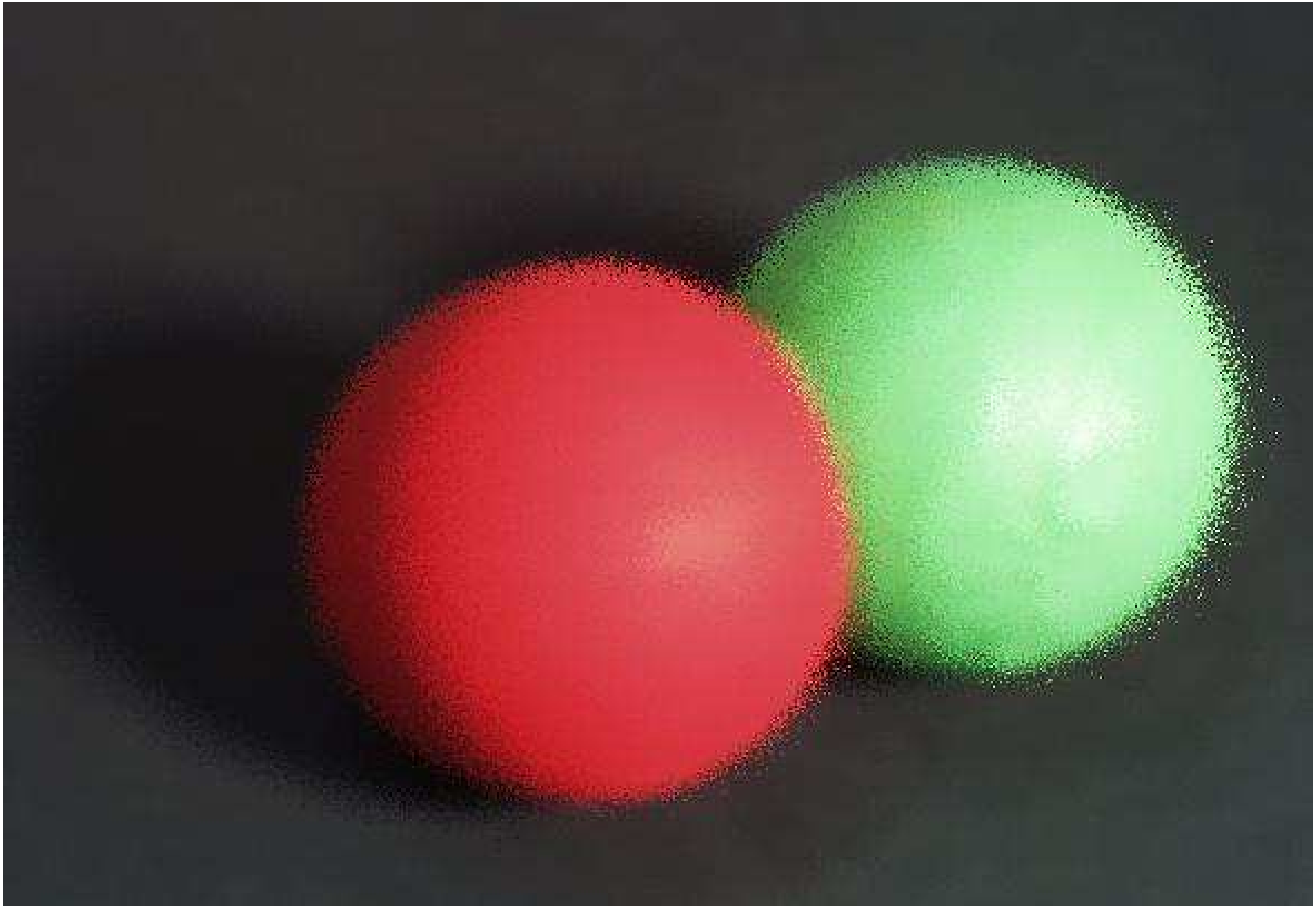
credit: http://en.wikipedia.org/wiki/Pythagoras



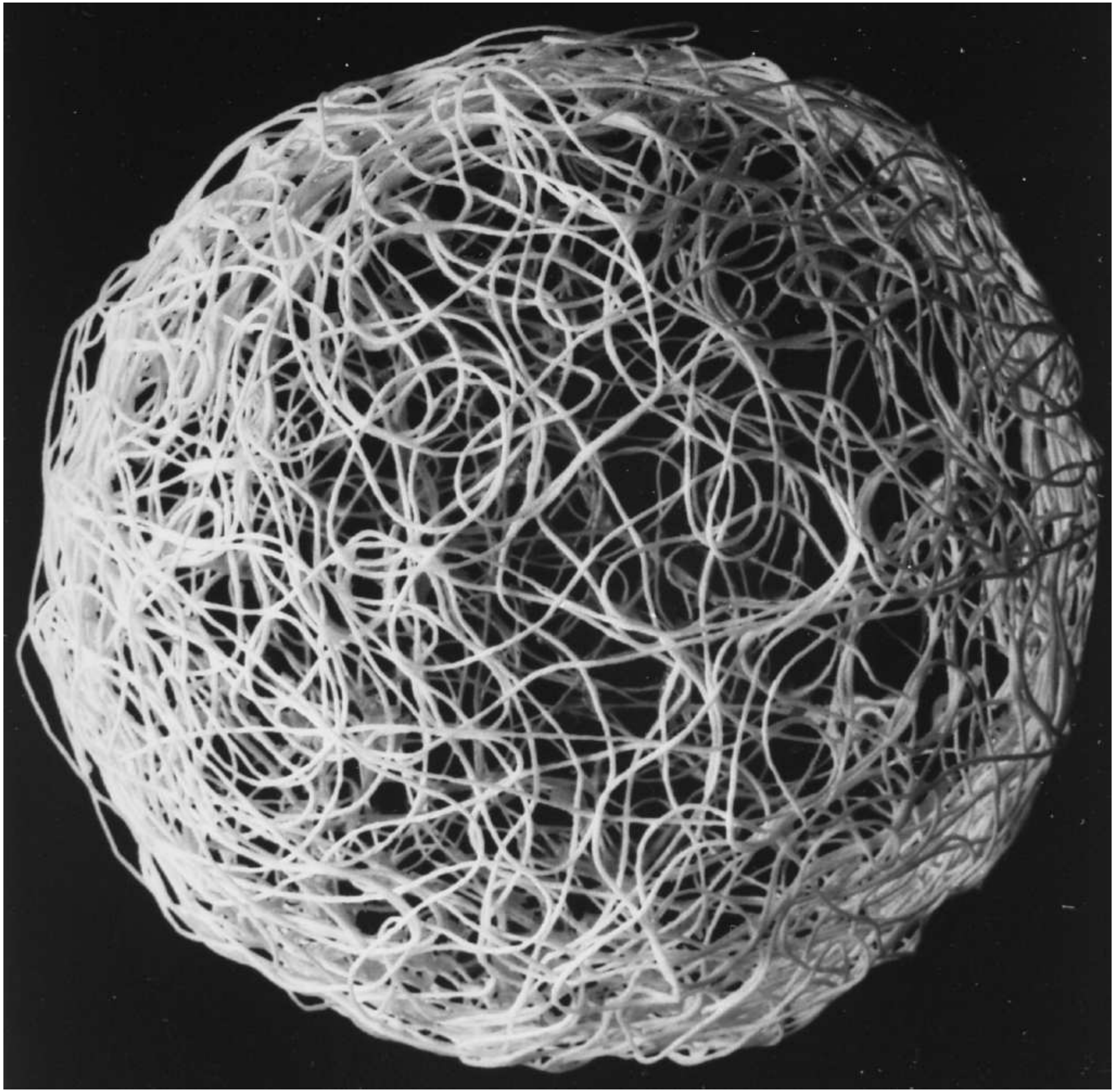
1 dimension is too simple!

Bowls in 2 Dimensions

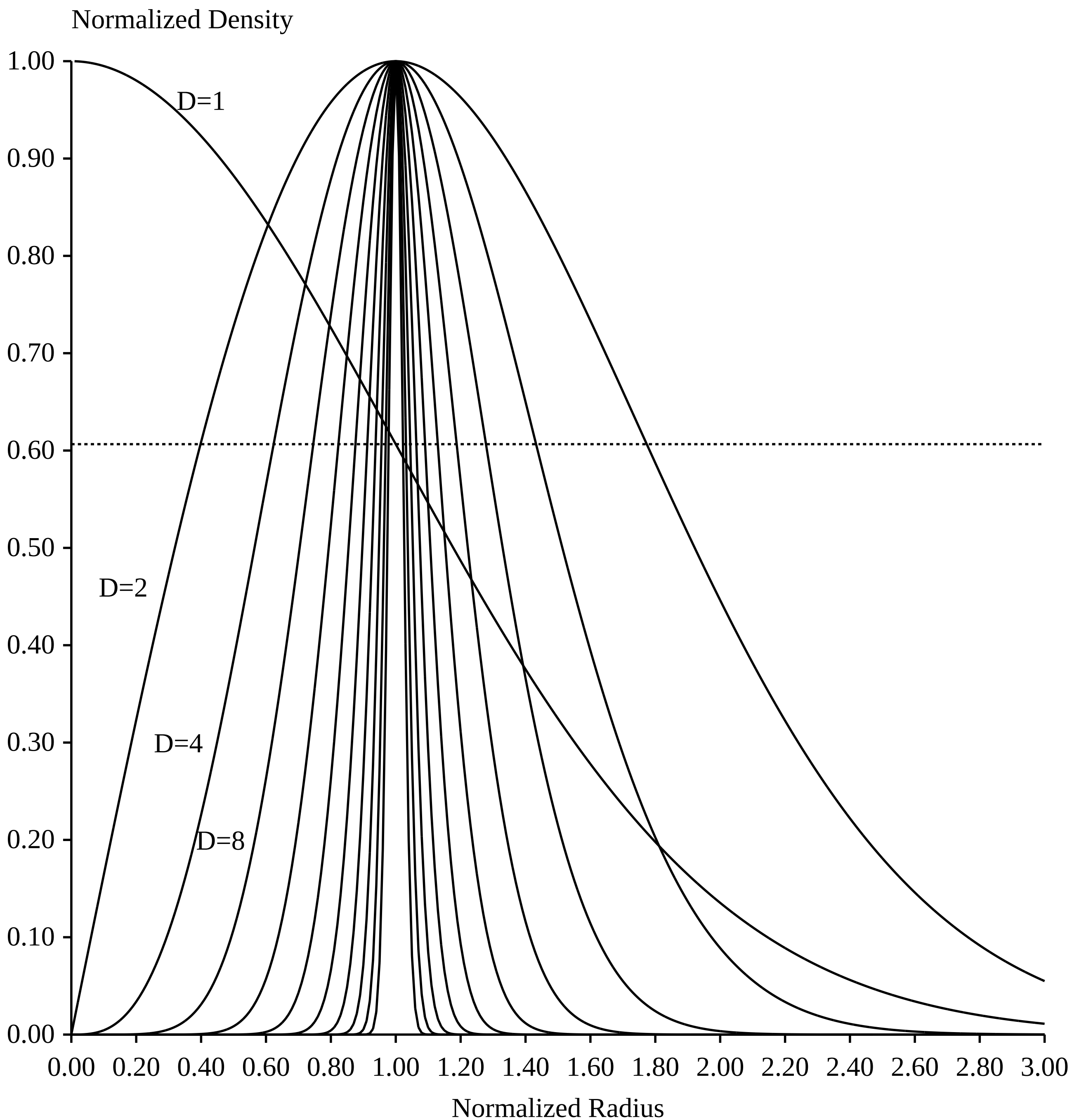
Spheres in 3 Dimensions



N Dimensional Sphere



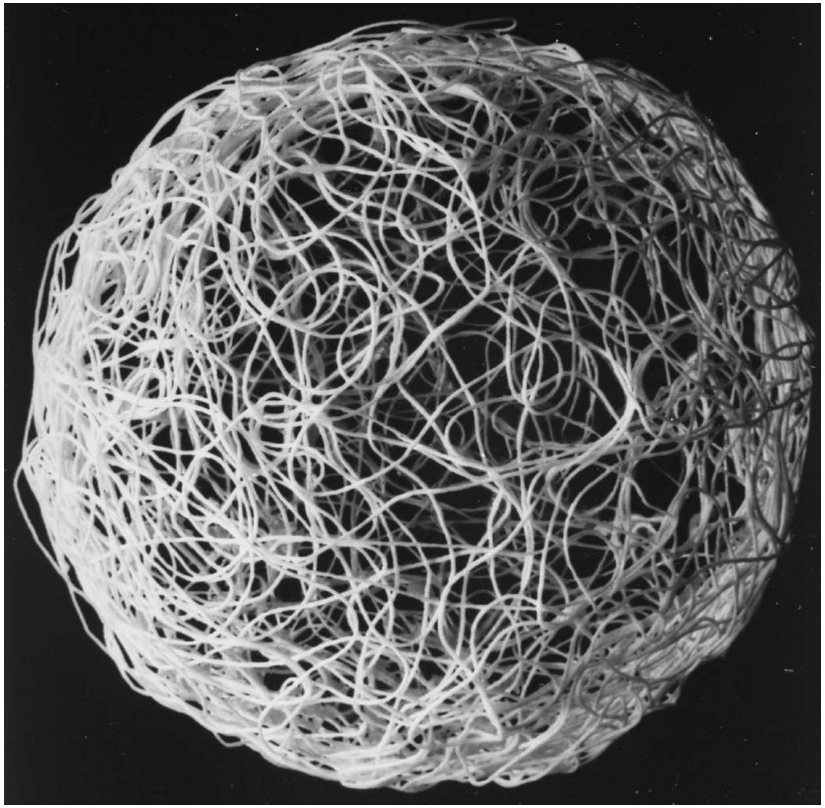
Spheres tighten in high dimensions



Good packing of spheres
gives a molecule
the capacity
to make selections efficiently

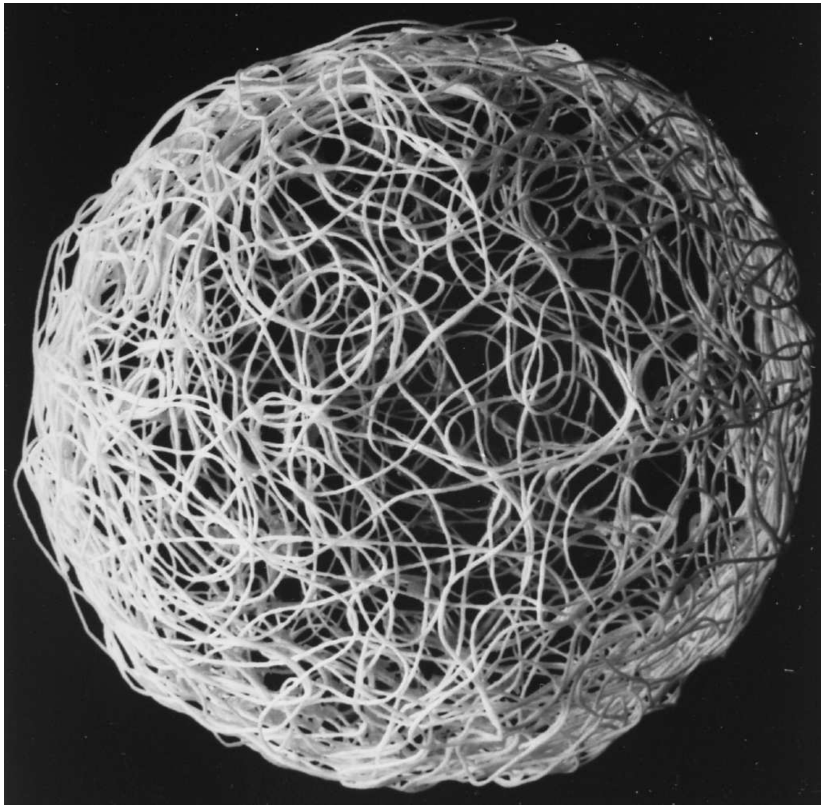
N Dimensional Sphere Separation

Degenerate Sphere

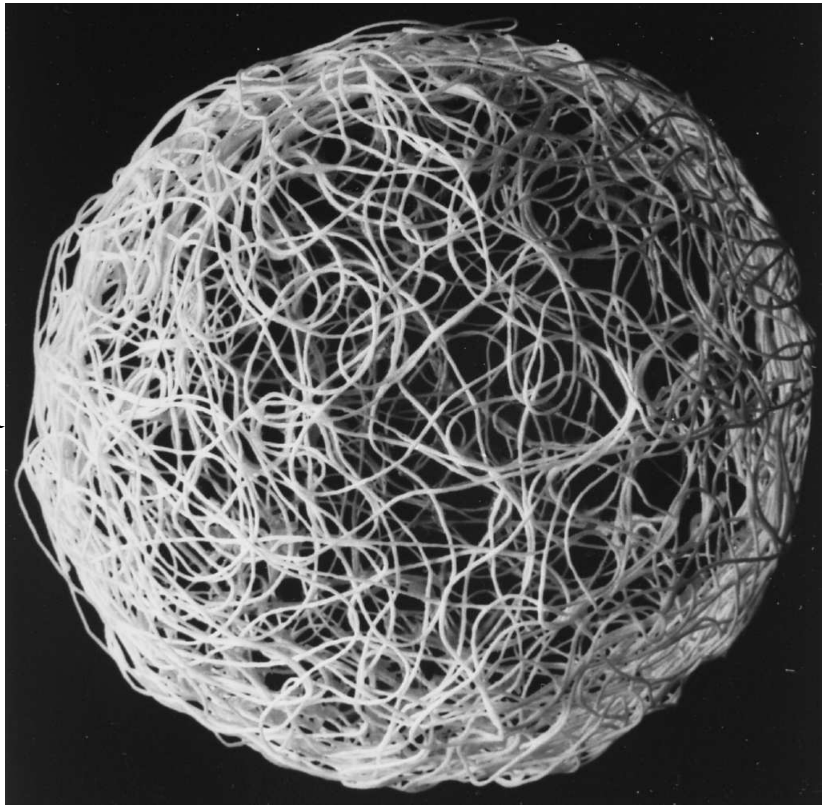


N Dimensional Sphere Separation

Degenerate Sphere

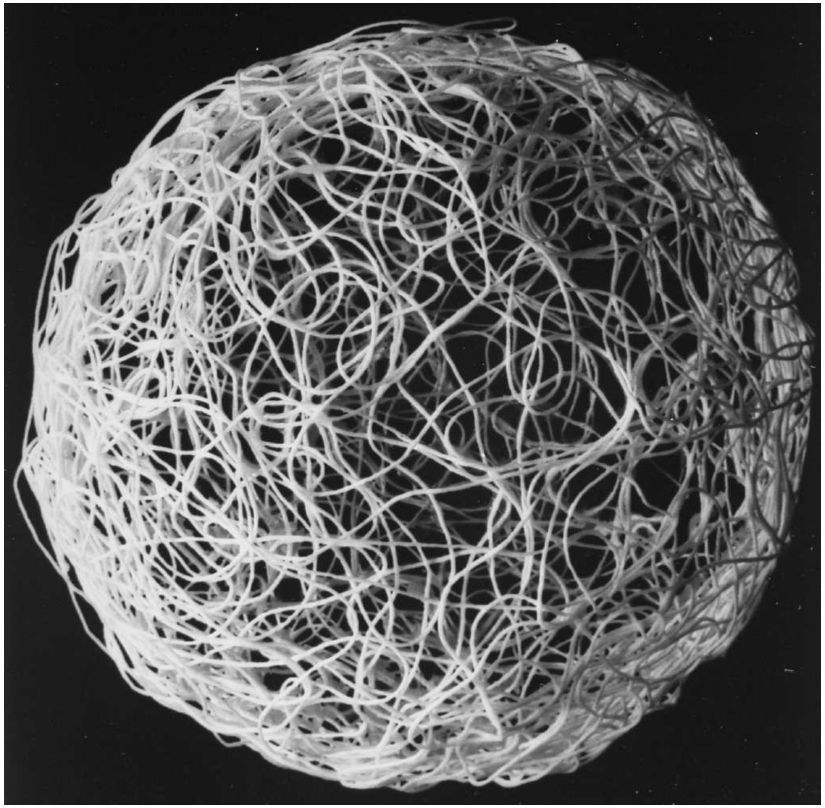


Forward Sphere

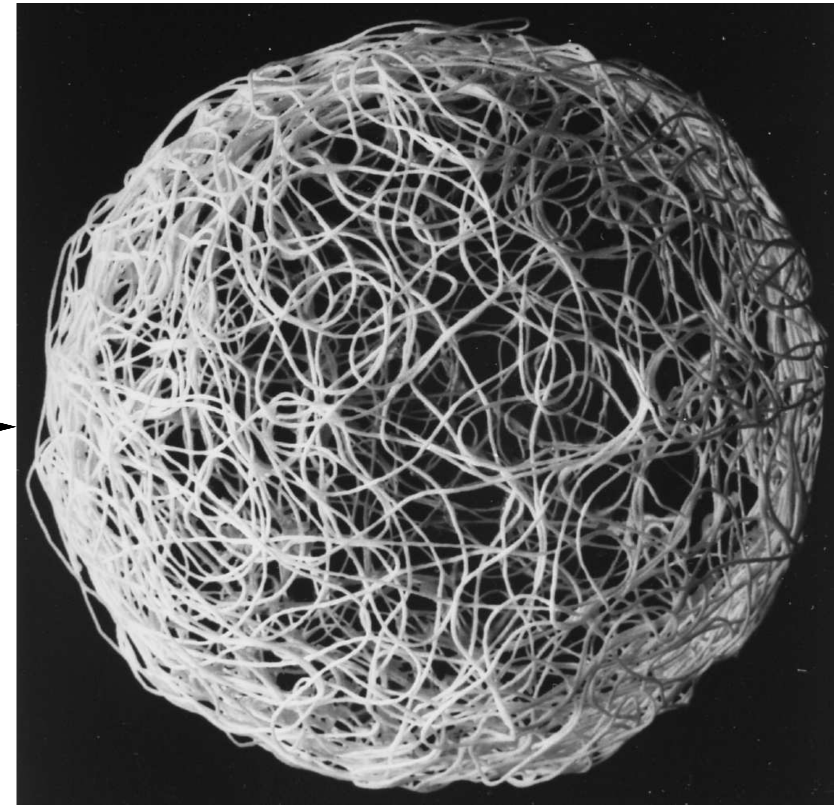


N Dimensional Sphere Separation

Degenerate Sphere



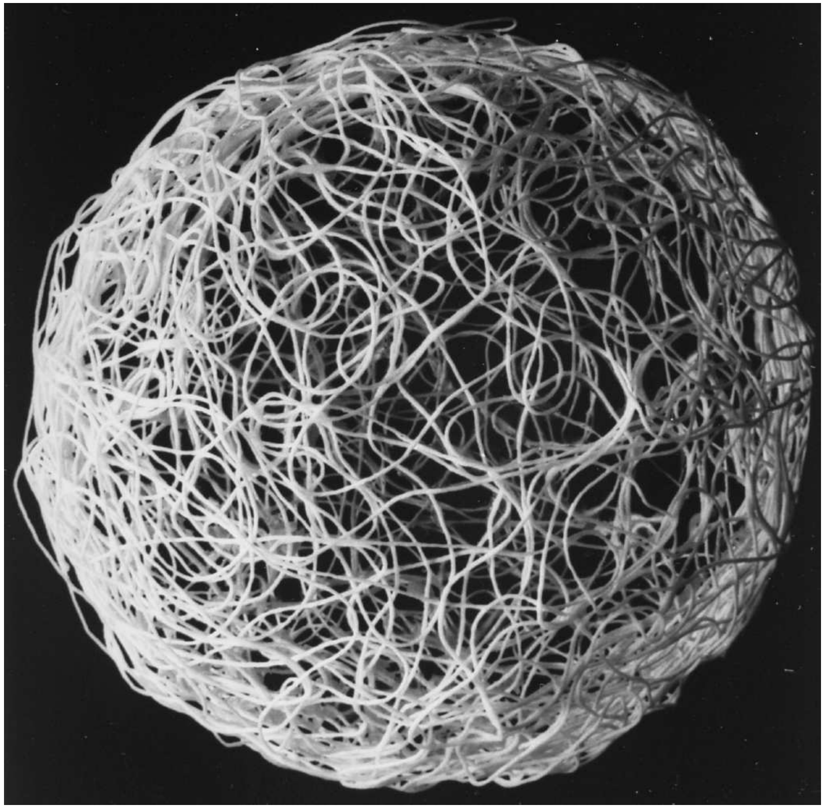
Forward Sphere



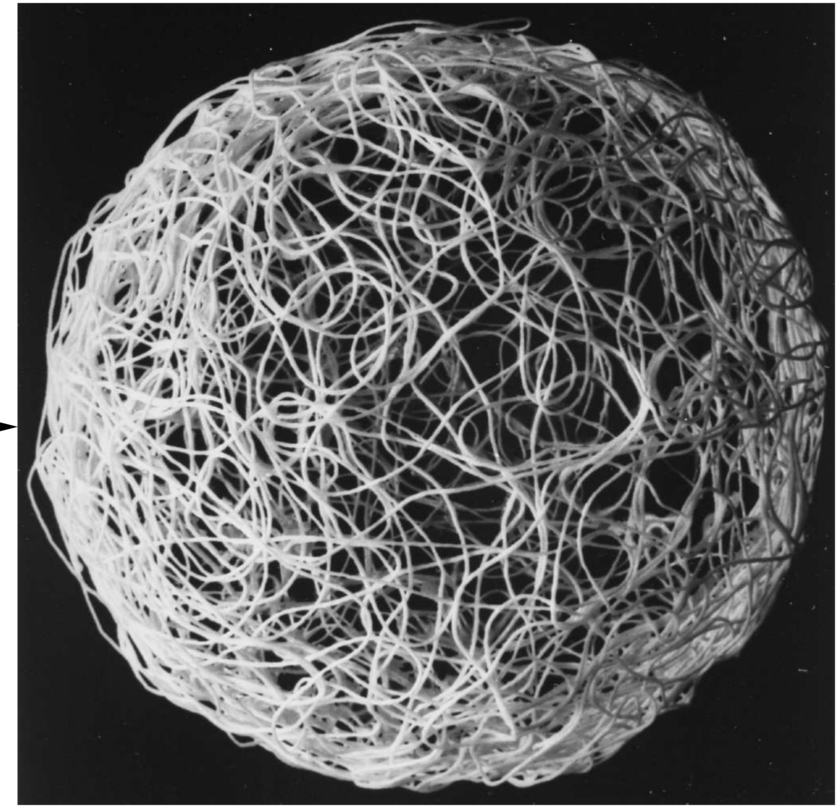
$\sqrt{\text{Noise}}$

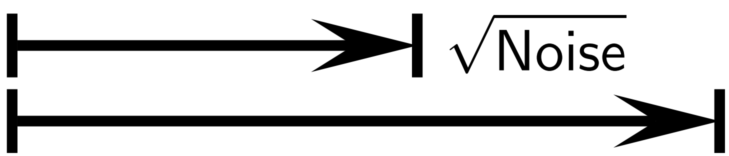
N Dimensional Sphere Separation

Degenerate Sphere



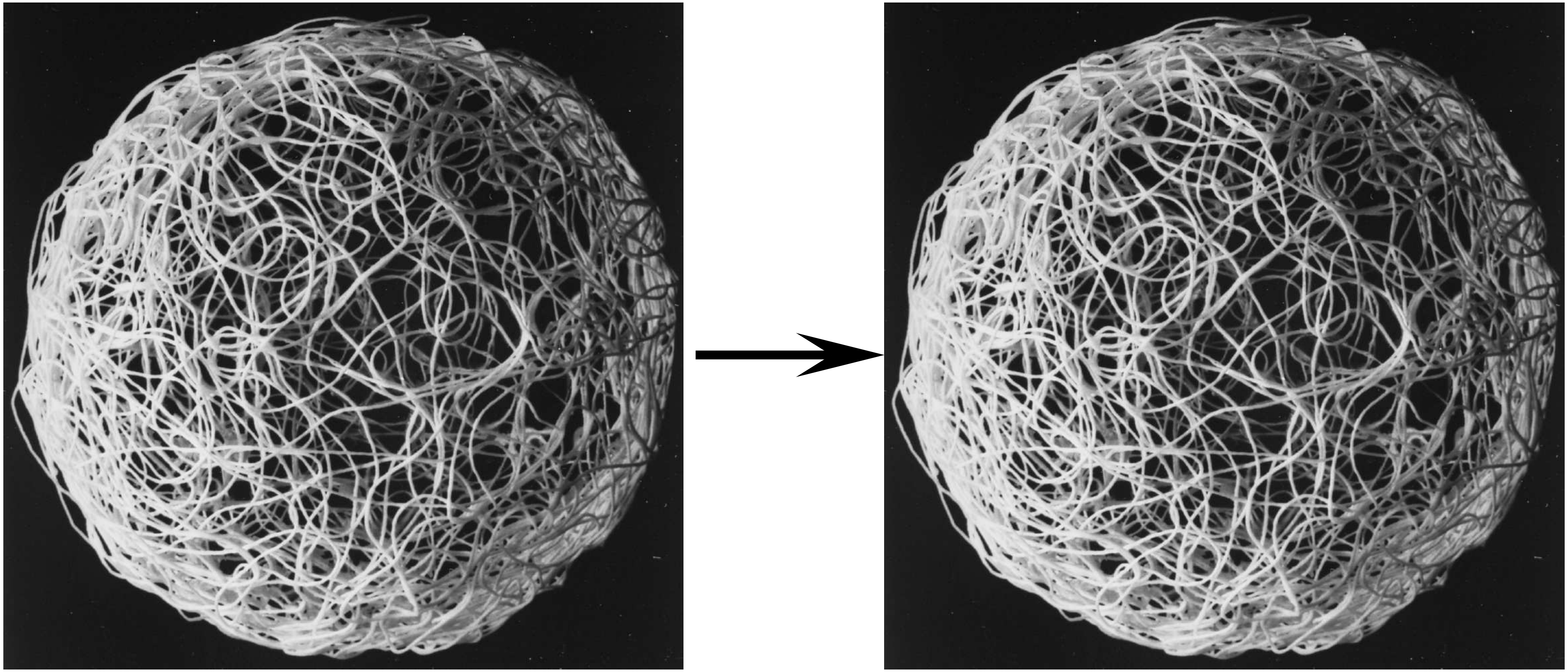
Forward Sphere

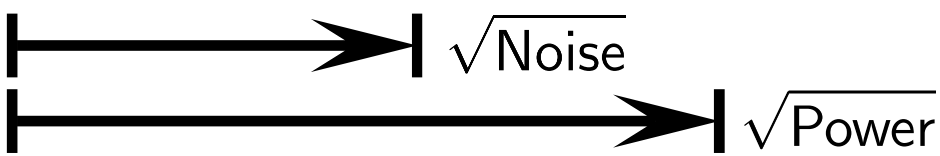




N Dimensional Sphere Separation

Degenerate Sphere　　　　Forward Sphere

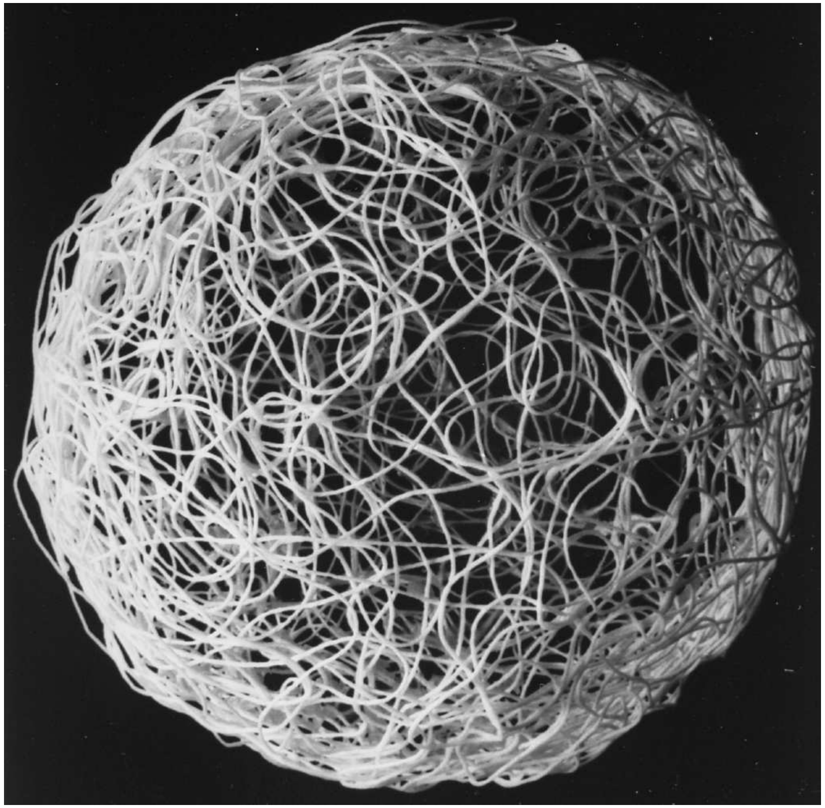




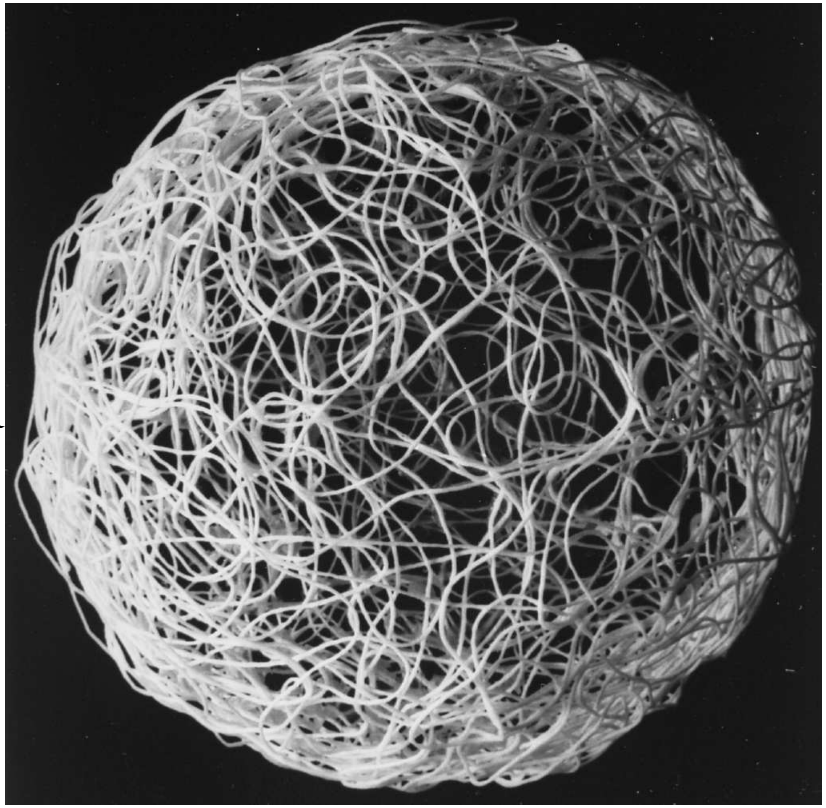
Energy dissipated to escape the Degenerate Sphere must exceed the Noise

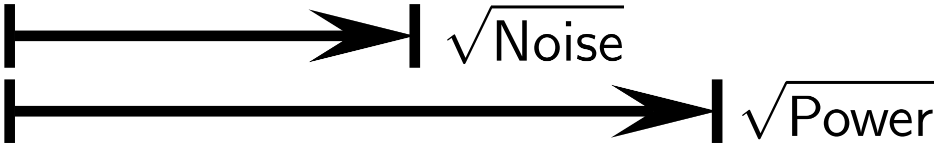
N Dimensional Sphere Separation

Degenerate Sphere



Forward Sphere

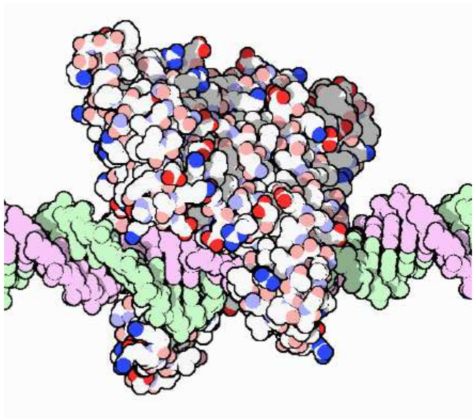




Energy dissipated to escape the Degenerate Sphere must exceed the Noise

$$\sqrt{\text{Power}} > \sqrt{\text{Noise}}$$

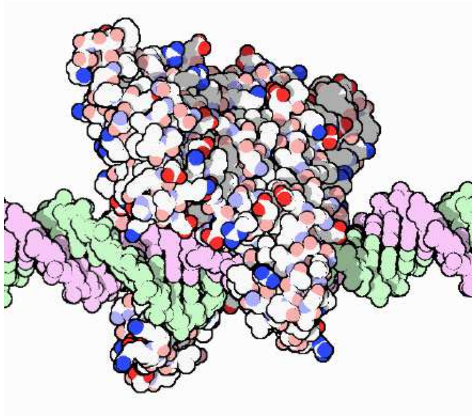
Theoretical Isothermal Efficiency



- For molecular states of molecules with d_{space} 'parts' P_y energy is dissipated for noise N_y and

  $C_y = d_{space} \log_2(P_y/N_y + 1) \leftarrow$ machine capacity

Theoretical Isothermal Efficiency

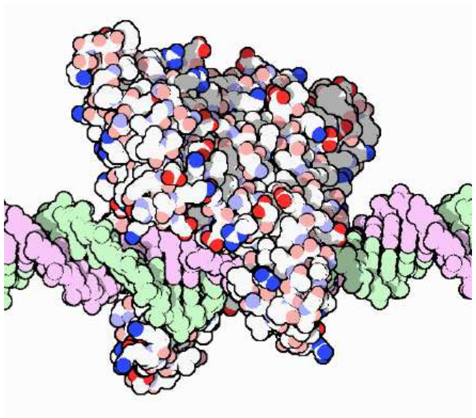


- For molecular states of molecules with d_{space} 'parts' P_y energy is dissipated for noise N_y and

  $C_y = d_{space} \log_2(P_y/N_y + 1) \leftarrow$ machine capacity

  $\epsilon_t = \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow$ isothermal efficiency

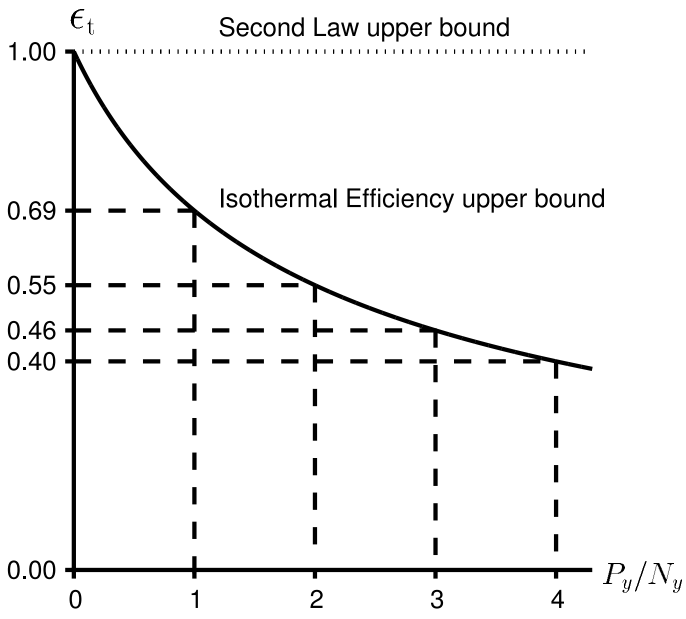
Theoretical Isothermal Efficiency



- For molecular states of molecules with d_{space} 'parts' P_y energy is dissipated for noise N_y and

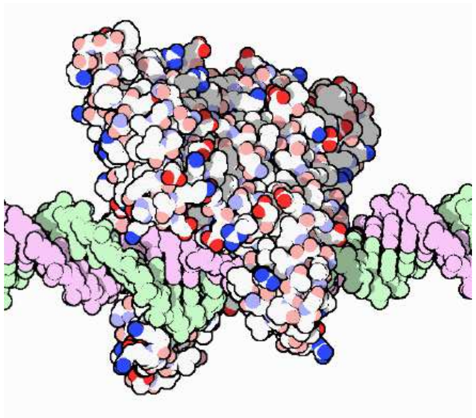
$$C_y = d_{space} \log_2(P_y/N_y + 1) \leftarrow \text{machine capacity}$$

$$\epsilon_t = \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow \text{isothermal efficiency}$$



The curve is an upper bound

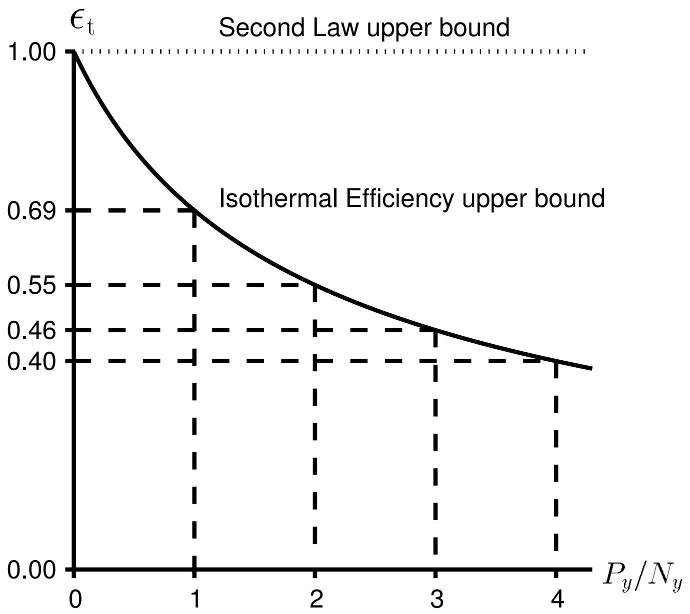
Theoretical Isothermal Efficiency



- For molecular states of molecules with d_{space} 'parts' P_y energy is dissipated for noise N_y and

  $$C_y = d_{space} \log_2(P_y/N_y + 1) \leftarrow \text{machine capacity}$$

  $$\epsilon_t = \frac{\ln\left(\frac{P_y}{N_y}+1\right)}{\frac{P_y}{N_y}} \leftarrow \text{isothermal efficiency}$$

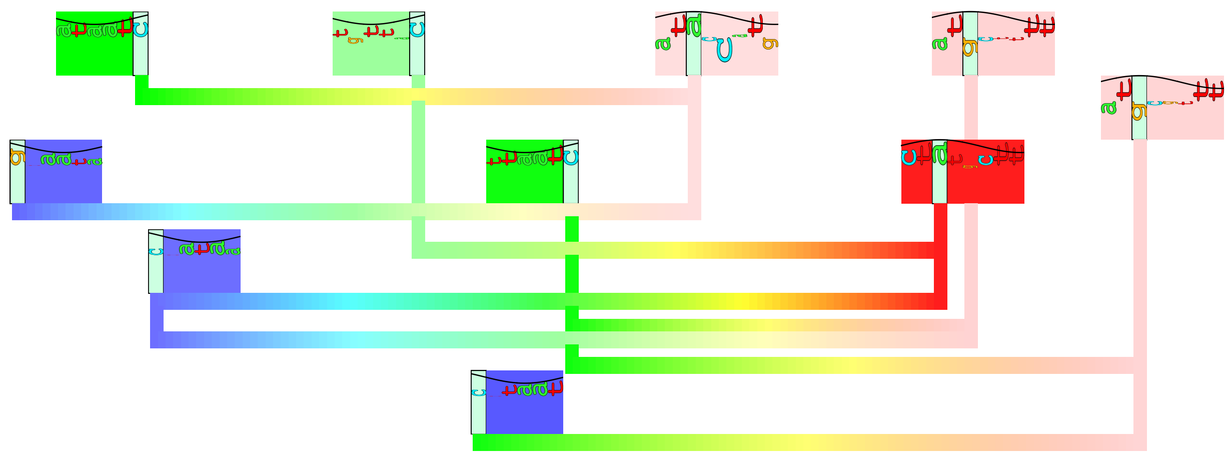


The curve is an upper bound

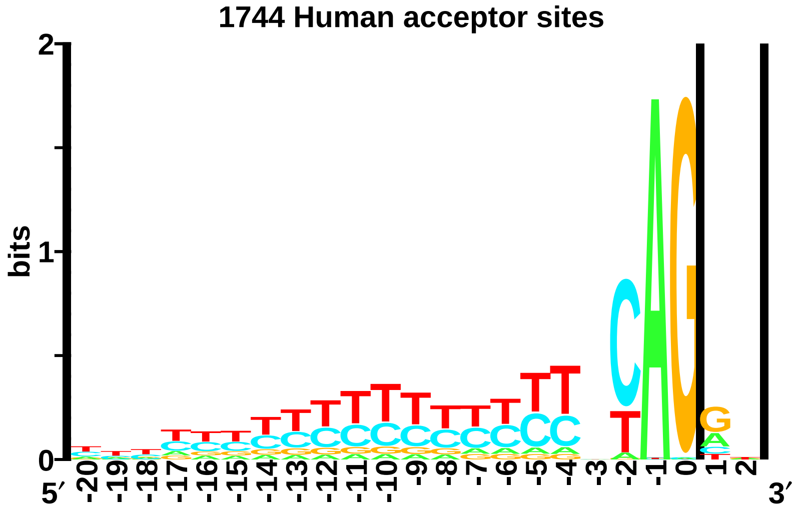
- $\boxed{\text{If } P_y/N_y = 1 \text{ the efficiency is 70\%!}}$

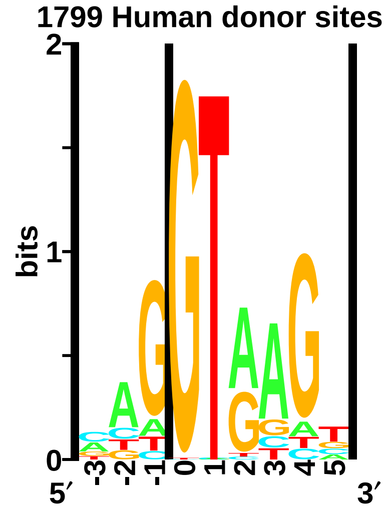
Acknowledgements

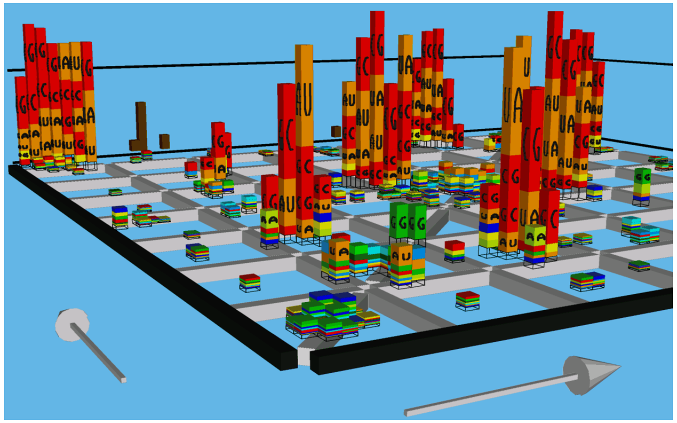
- Mentors:
  - Larry Gold (graduate school mentor)
  - Andrej Ehrenfeucht (information theory)
- Students:
  - Mike Stephens (logos, splicing)
  - John Spouge (mathematics)
  - Paul C. Anagnostopoulos (Evolution model)
  - Bruce Shapiro and Eckart Bindewald (3D logo)
  - Kevin Franco (σ^{38})
  - Ding Jin (σ^{38})
- Useful discussions with
  - Jeff Strathern
  - Amar Klar
  - Kemi Abolude
  - Susan Lauffer
  - Cedric Cagliero
  - Amar Klar
  - Zhi-Ming Zheng
  - Mark Lewandoski
- This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

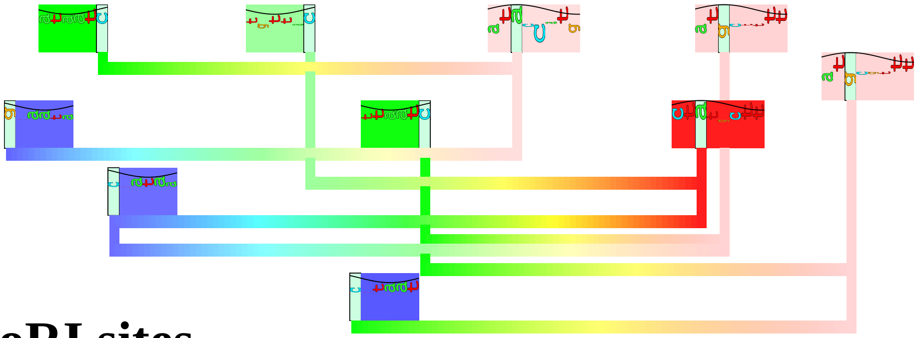


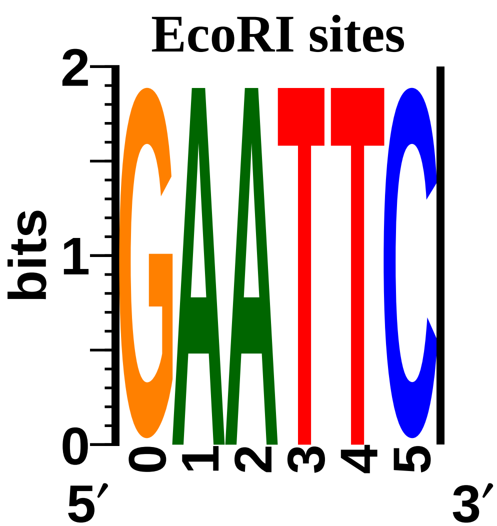
Information theory: the mathematics of biology

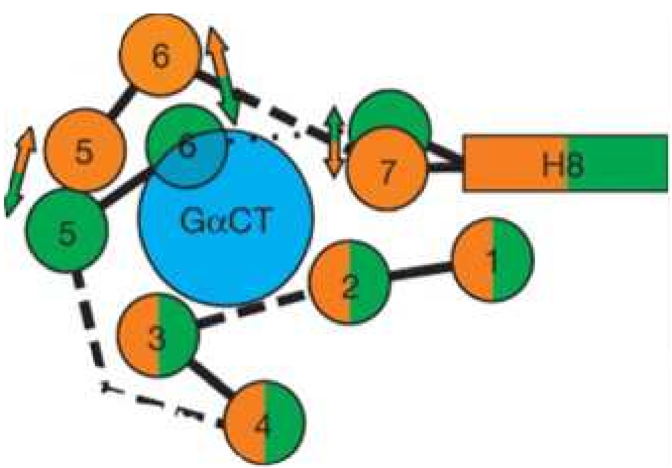












Version

version = 1.89 of kanpurtalk.tex 2016 Oct 16