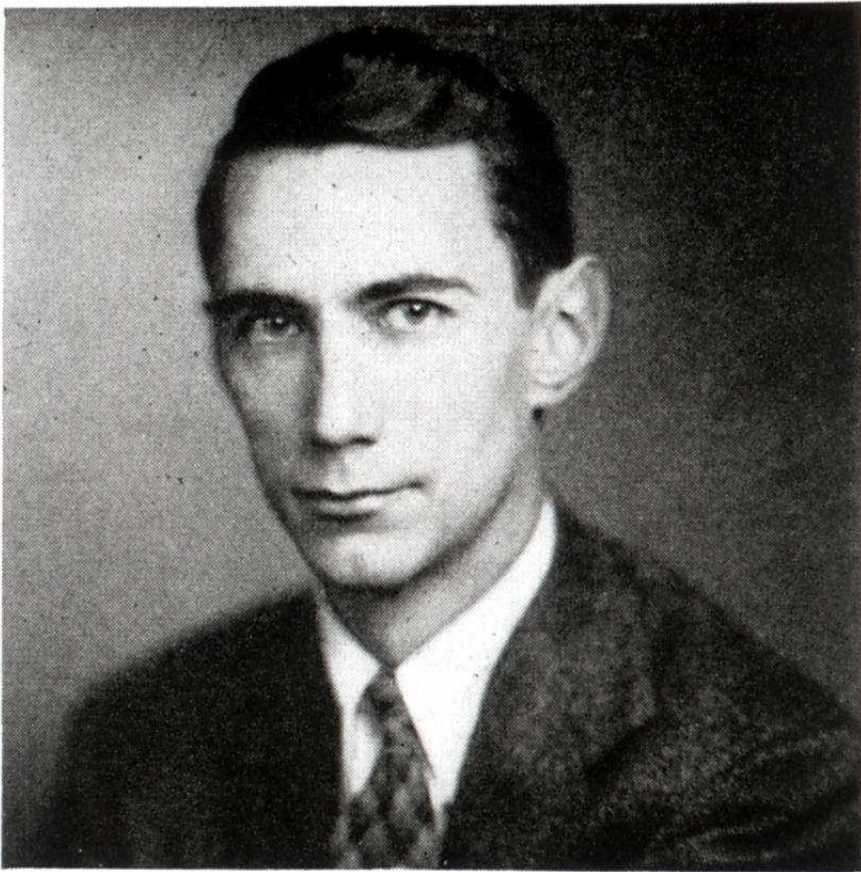


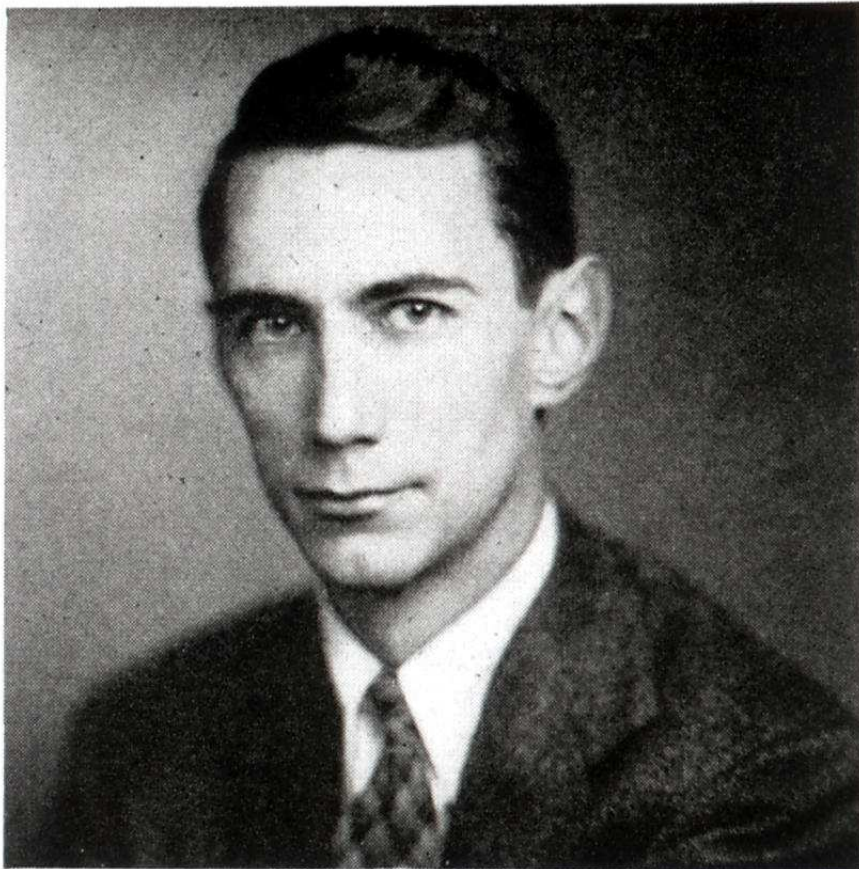
CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory
- Important papers: 1948

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

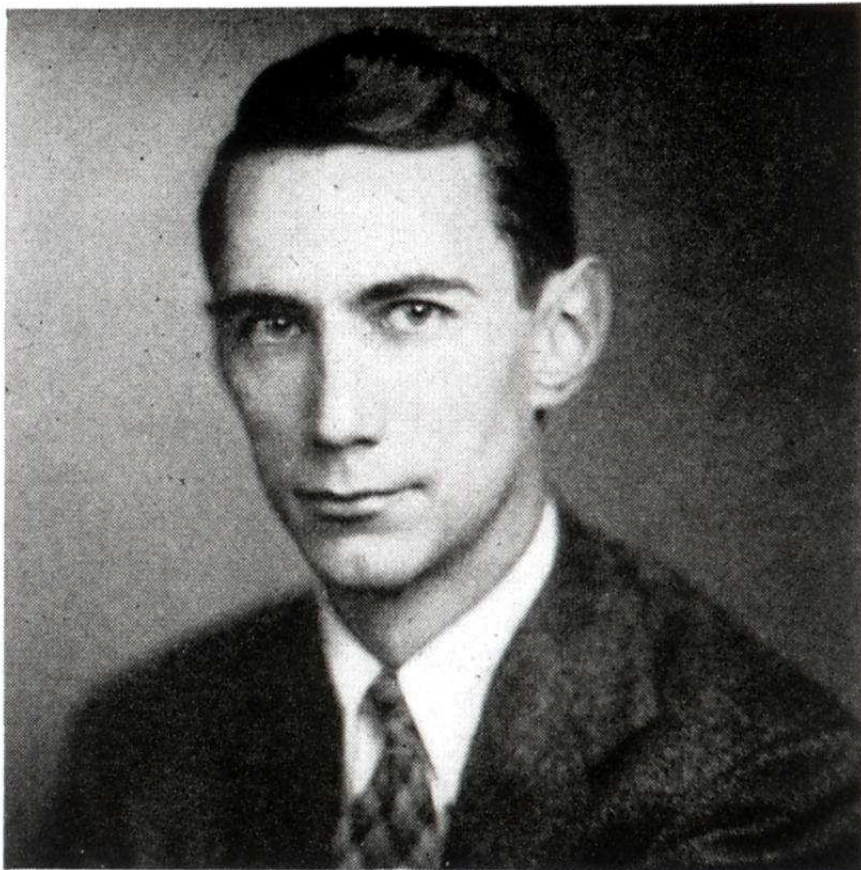
1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.
2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measure entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.
3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\begin{aligned}\log_2 M &= \log_{10} M / \log_{10} 2 \\ &= 3.32 \log_{10} M,\end{aligned}$$

¹Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A.I.E.E. Trans.*, v. 47, April 1928, p. 617.

²Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory
- Important papers: 1948 , 1949

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental process of communication is approximately a message to or are correlated aspects of communication message is one selection, not selection, not selection.

If the number of messages can be regarded as a function of choices being equally function. Although the statistics of the message are essentially logarithmic. The logarithmic message

1. It is practically the number of relays, etc., adding one relay logarithm of this doubles the logarithm.
2. It is nearer to our intuitive measure of two punched card channels twice the logarithm.
3. It is mathematical but would

The choice of a logarithm base 2 is used the rest of the paper. A device information. N such devices. If the base 10 is used the

¹Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell Syst. Tech. J.*, vol. 3, p. 324, Apr. 1924.
²Hartley, R. V. L., "Transmission of Information," *Bell Syst. Tech. J.*, vol. 3, p. 535-564, July 1928.

Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Classic Paper

A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two "function spaces," and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concerning expansion and compression of bandwidth and the threshold effect. Formulas are found for the maximum rate of transmission of binary digits over a system when the signal is perturbed by various types of noise. Some of the properties of "ideal" systems which transmit at this maximum rate are discussed. The equivalent number of binary digits per second for certain information sources is calculated.

I. INTRODUCTION

A general communications system is shown schematically in Fig. 1. It consists essentially of five elements.

1) *The Information Source:* The source selects one message from a set of possible messages to be transmitted to the receiving terminal. The message may be of various types; for example, a sequence of letters or numbers, as in telegraphy or teletype, or a continuous function of time $f(t)$, as in radio or telephony.

2) *The Transmitter:* This operates on the message in some way and produces a signal suitable for transmission to the receiving point over the channel. In telephony, this operation consists of merely changing sound pressure into a proportional electrical current. In telegraphy, we have an encoding operation which produces a sequence of dots, dashes, and spaces corresponding to the letters of the message. To take a more complex example, in the case of multiplex PCM telephony the different speech functions must be sampled, compressed, quantized and encoded, and finally interleaved properly to construct the signal.

3) *The Channel:* This is merely the medium used to transmit the signal from the transmitting to the receiving point. It may be a pair of wires, a coaxial cable, a band of radio frequencies, etc. During transmission, or at the receiving terminal, the signal may be perturbed by noise or distortion. Noise and distortion may be differentiated on the basis that distortion is a fixed operation applied to the signal, while noise involves statistical and unpredictable

This paper is reprinted from the PROCEEDINGS OF THE IRE, vol. 37, no. 1, pp. 10-21, Jan. 1949.
Publisher Item Identifier S 0018-9219(98)01299-7.

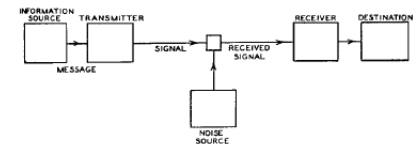


Fig. 1. General communications system.

perturbations. Distortion can, in principle, be corrected by applying the inverse operation, while a perturbation due to noise cannot always be removed, since the signal does not always undergo the same change during transmission.

4) *The Receiver:* This operates on the received signal and attempts to reproduce, from it, the original message. Ordinarily it will perform approximately the mathematical inverse of the operations of the transmitter, although they may differ somewhat with best design in order to combat noise.

5) *The Destination:* This is the person or thing for whom the message is intended.

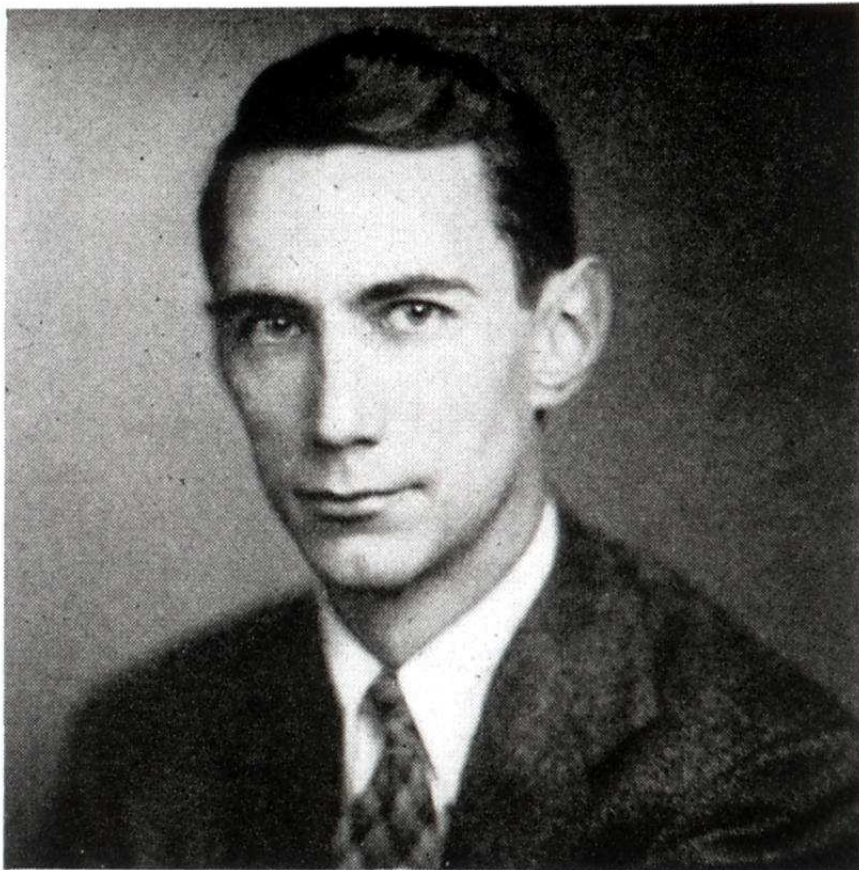
Following Nyquist¹ and Hartley,² it is convenient to use a logarithmic measure of information. If a device has n possible positions it can, by definition, store $\log_2 n$ units of information. The choice of the base b amounts to a choice of unit, since $\log_b n = \log_2 n \log_2 b$. We will use the base 2 and call the resulting units binary digits or bits. A group of m relays or flip-flop circuits has 2^m possible sets of positions, and can therefore store $\log_2 2^m = m$ bits.

If it is possible to distinguish reliably M different signal functions of duration T on a channel, we can say that the channel can transmit $\log_2 M$ bits in time T . The rate of transmission is then $\log_2 M/T$. More precisely, the channel capacity may be defined as

$$C = \lim_{T \rightarrow \infty} \frac{\log_2 M}{T} \quad (1)$$

¹H. Nyquist, "Certain factors affecting telegraph speed," *Bell Syst. Tech. J.*, vol. 3, p. 324, Apr. 1924.

²R. V. L. Hartley, "The transmission of information," *Bell Syst. Tech. J.*, vol. 3, p. 535-564, July 1928.



CLAUDE E. SHANNON

- April 30, 1916 - February 24, 2001
- Founded Information Theory
- Important papers: 1948 , 1949
- Result: modern communications!

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem is to represent a message to or are correlated aspects of communication message is one selection, not a function.

If the number of messages can be regarded as a function of the number of choices being equally function. Although the statistics of the message are essentially logarithmic, the logarithmic measure of the number of choices is

1. It is practically the number of relays, etc., adding one relay logarithm of this doubles the logarithm.
2. It is nearer to our intuitive measure of two punched card channels twice the logarithm.
3. It is mathematically simpler but would

The choice of a logarithm base 2 is used the rest of the paper. A device information. N such devices. If the base 10 is used it

¹Nyquist, H., "Certain Factors Affecting Telegraph Transmission," *Bell Syst. Tech. J.*, vol. 1, pp. 10-21, Jan. 1924.
²Hartley, R. V. L., "Transmission of Information," *Bell Syst. Tech. J.*, vol. 3, p. 535-564, July 1928.

Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Classic Paper

A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two "function spaces," and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concerning expansion and effect. Formulas are given for the number of binary digits of various types of noise which transmit at the same rate as a signal. The number of binary digits is calculated.

I. INTRODUCTION

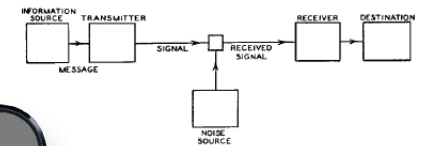
A general communication system is shown schematically in Fig. 1. It consists of an information source, a transmitter, a channel, a receiver, and a destination.

1) An information message from a set of messages is transmitted through the channel. The channel may be a wire, a radio, or a telephone line. The channel may also be a relay or a switch. The channel may be noisy or it may be noiseless.

2) The transmitter converts the message into a signal. The signal is then transmitted through the channel. The signal may be a voltage, a current, or a light beam. The signal may also be a sound wave or a radio wave. The signal may be modulated or it may be unmodulated.

3) The receiver converts the signal back into a message. The message is then delivered to the destination. The receiver may be a telephone, a radio, or a computer. The receiver may also be a relay or a switch. The receiver may be noisy or it may be noiseless.

This paper is a reprint of the original paper published in the *Bell System Technical Journal*, vol. 1, pp. 10-21, Jan. 1924. The original paper is available at <http://www.bell-labs.com/doc/1924/192401010-1021.pdf>. The original paper is also available at http://www.ieee.org/publications_standards/publications/details_standards.cfm?pnumber=9880129&frompage=1&frompage=1&frompage=1.



General communications system.



Distortion can, in principle, be corrected by using the inverse operation, while a perturbation due to noise cannot always be removed, since the signal does not undergo the same change during transmission. The receiver: This operates on the received signal to reproduce, from it, the original message. Ideally it will perform approximately the mathematical inverse of the operations of the transmitter, although they differ somewhat with best design in order to combat

The destination: This is the person or thing for whom the message is intended.

Following Nyquist¹ and Hartley,² it is convenient to use a logarithmic measure of information. If a device has n possible positions it can, by definition, store $\log_2 n$ units of information. The choice of the base b amounts to a choice of units, since $\log_b n = \log_b c \log_c n$. We will use the base 2, so that the resulting units are binary digits or bits. A group of relays or flip-flop circuits has 2^m possible sets of positions, and can therefore store $\log_2 2^m = m$ bits. If it is possible to distinguish reliably M different signal sets of duration T on a channel, we can say that the channel can transmit $\log_2 M$ bits in time T . The rate of transmission is then $\log_2 M/T$. More precisely, the channel capacity may be defined as

$$C = \lim_{T \rightarrow \infty} \frac{\log_2 M}{T} \quad (1)$$

¹Nyquist, "Certain factors affecting telegraph speed," *Bell Syst. Tech. J.*, p. 324, Apr. 1924.

²R. V. L. Hartley, "The transmission of information," *Bell Syst. Tech. J.*, vol. 3, p. 535-564, July 1928.

Information Theory: One-Minute Lesson

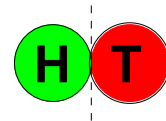
number of symbols	number of bits	example
-------------------	----------------	---------

M

B

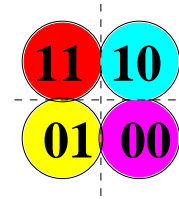
2

1



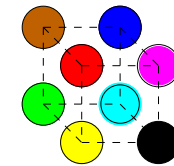
4

2



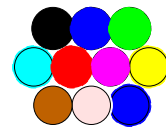
8

3



$$M=2^B$$

$$B=\log_2 M$$



Information Theory: One-Minute Lesson

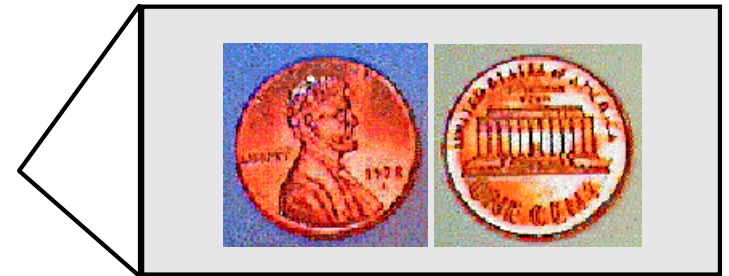
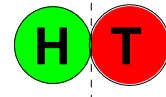
number of symbols	number of bits	example
-------------------	----------------	---------

M

B

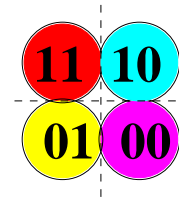
2

1



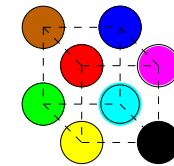
4

2



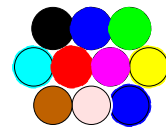
8

3



$$M=2^B$$

$$B=\log_2 M$$



Information Theory: One-Minute Lesson

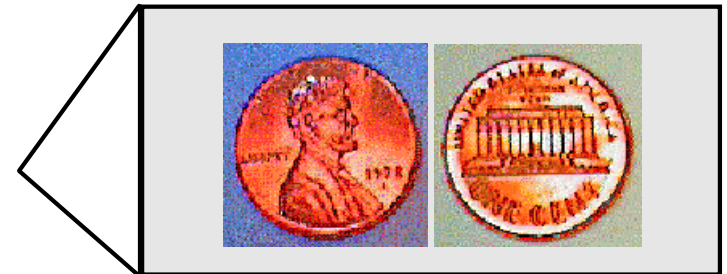
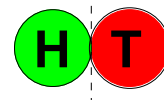
number of symbols	number of bits	example
-------------------	----------------	---------

M

B

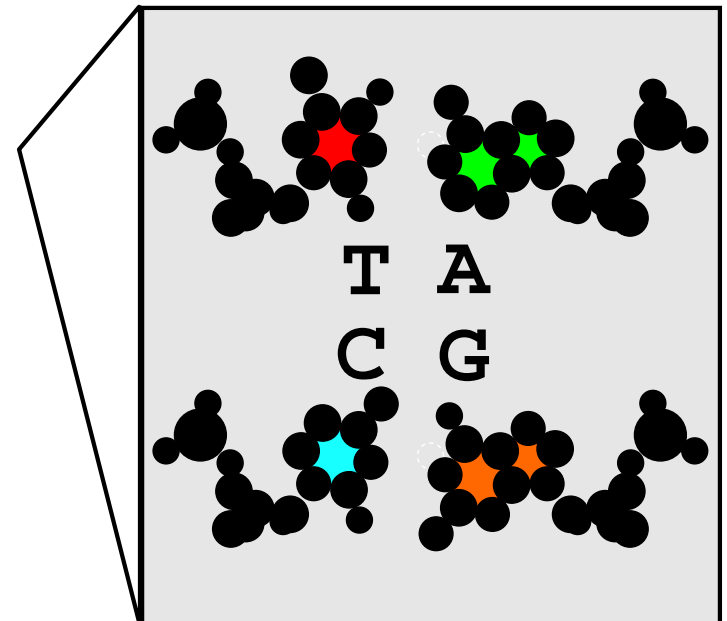
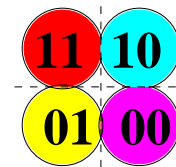
2

1



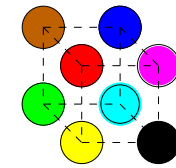
4

2



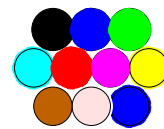
8

3



$$M=2^B$$

$$B=\log_2 M$$



Information Theory: One-Minute Lesson

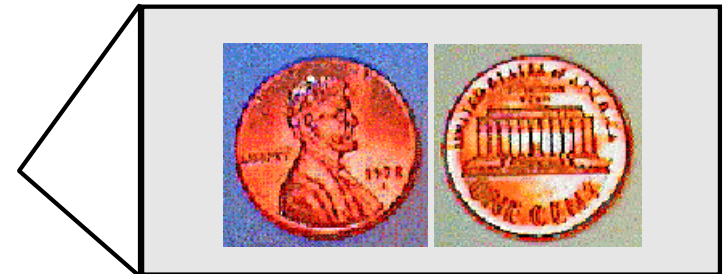
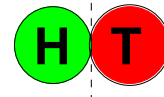
number of symbols	number of bits	example
-------------------	----------------	---------

M

B

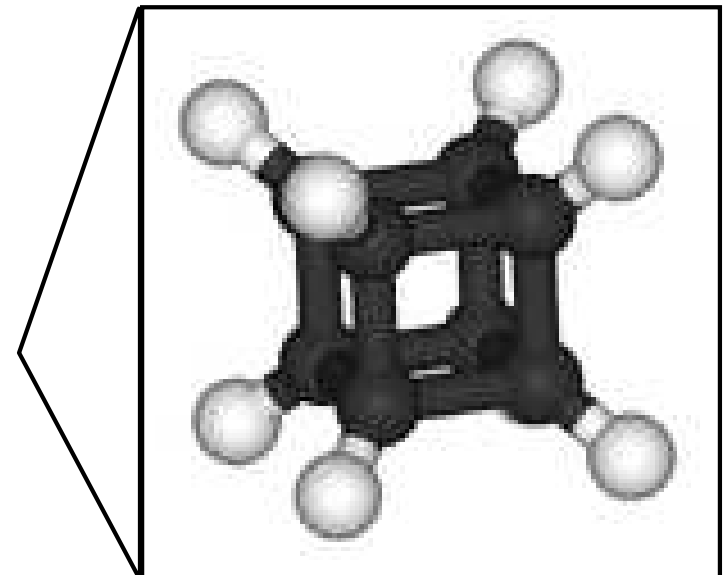
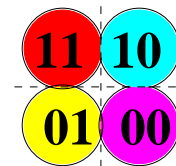
2

1



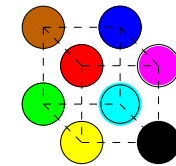
4

2



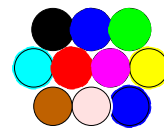
8

3



$$M=2^B$$

$$B=\log_2 M$$



Information Theory: One-Minute Lesson

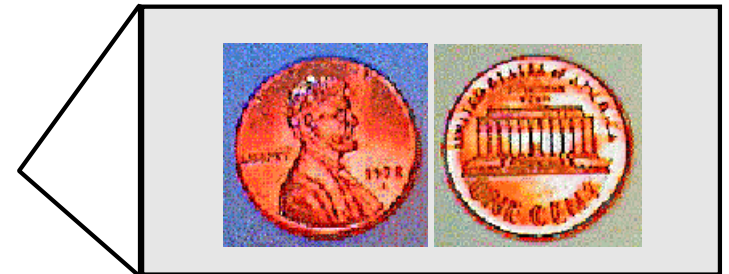
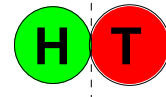
number of symbols	number of bits	example
-------------------	----------------	---------

M

B

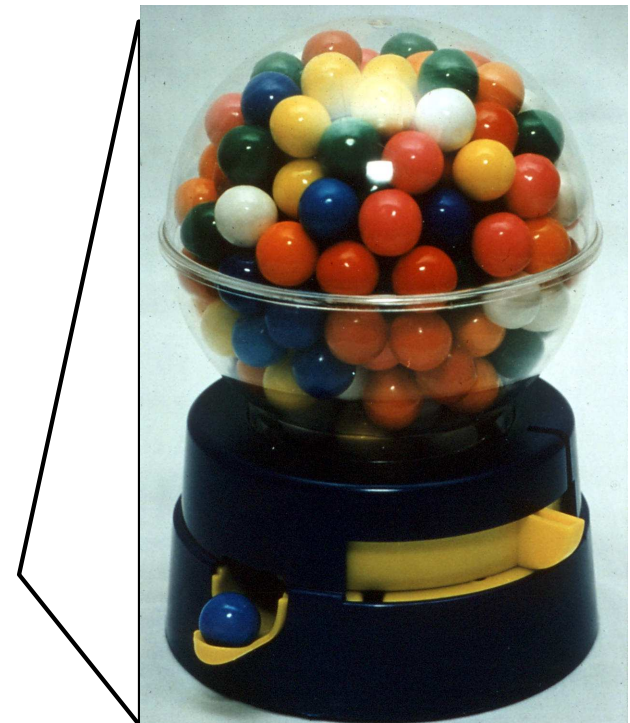
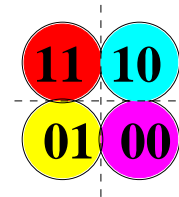
2

1



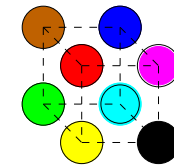
4

2



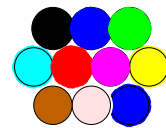
8

3



$$M=2^B$$

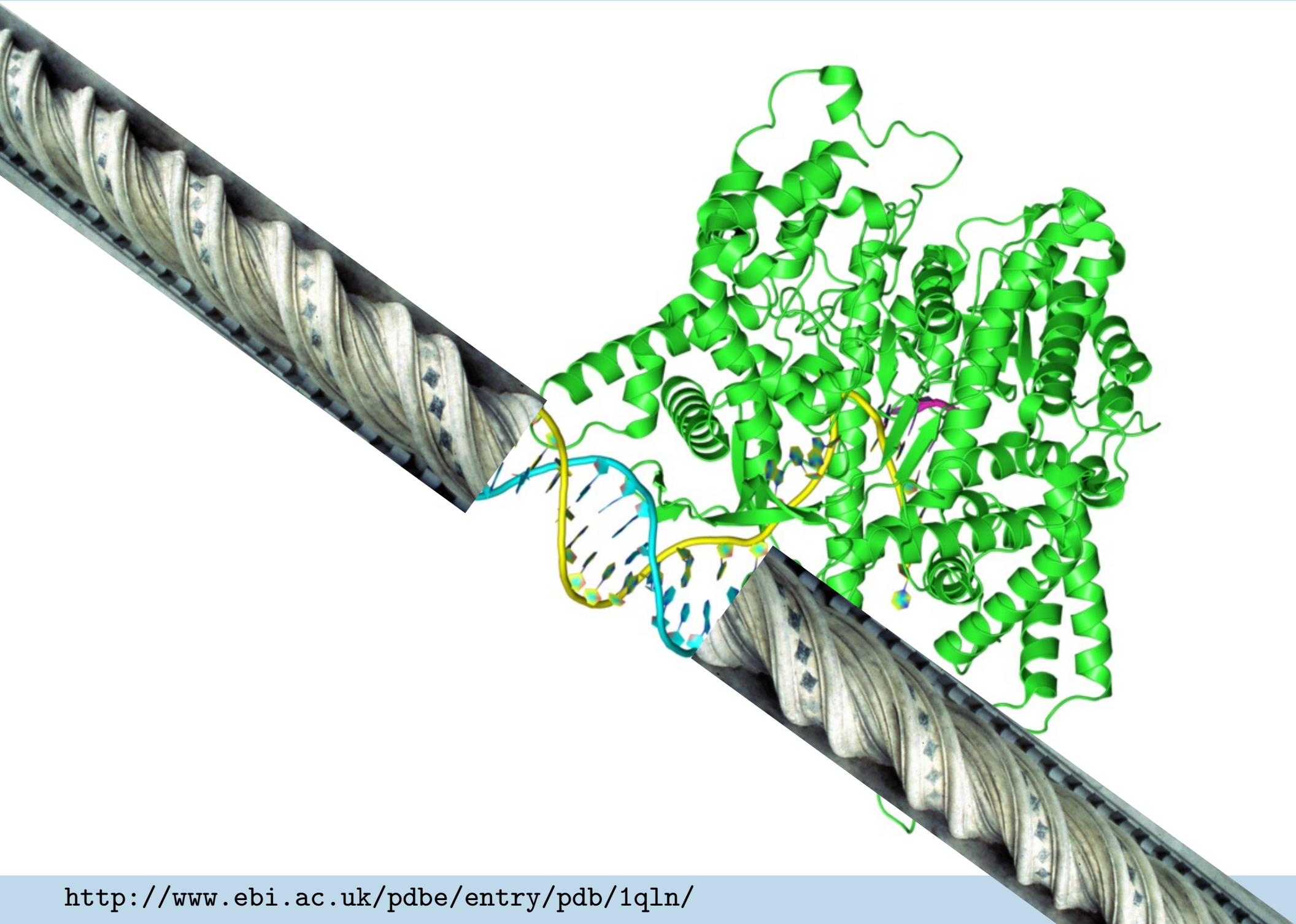
$$B=\log_2 M$$



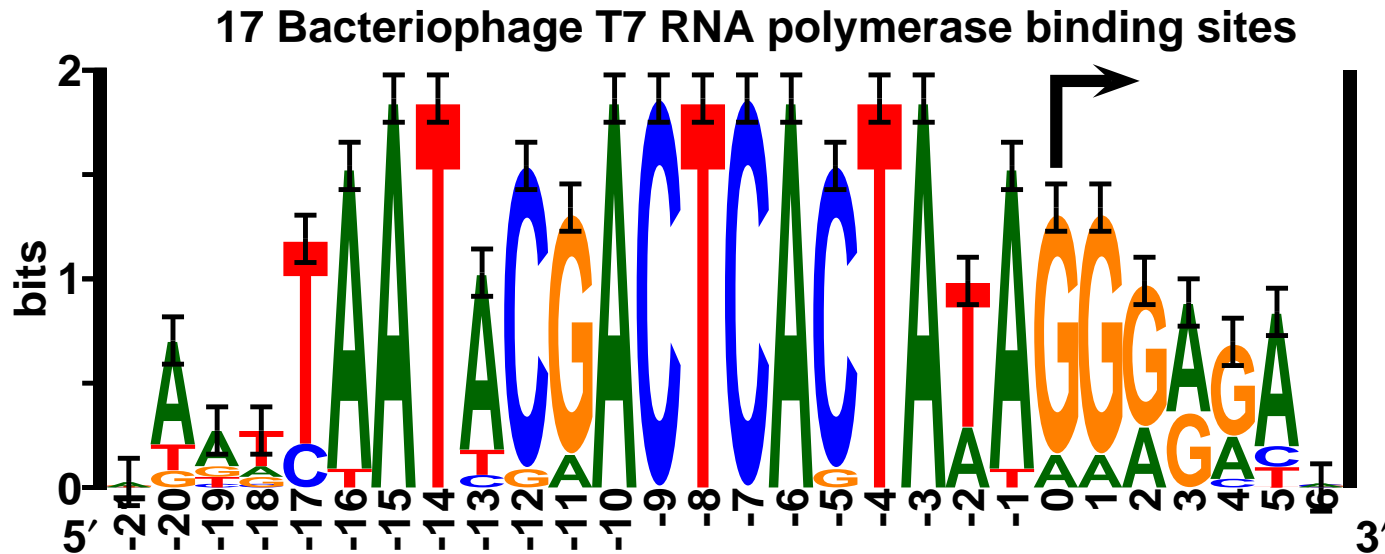
El Duomo, Florence, Italy



T7 RNA polymerase + DNA



Sequence Logo

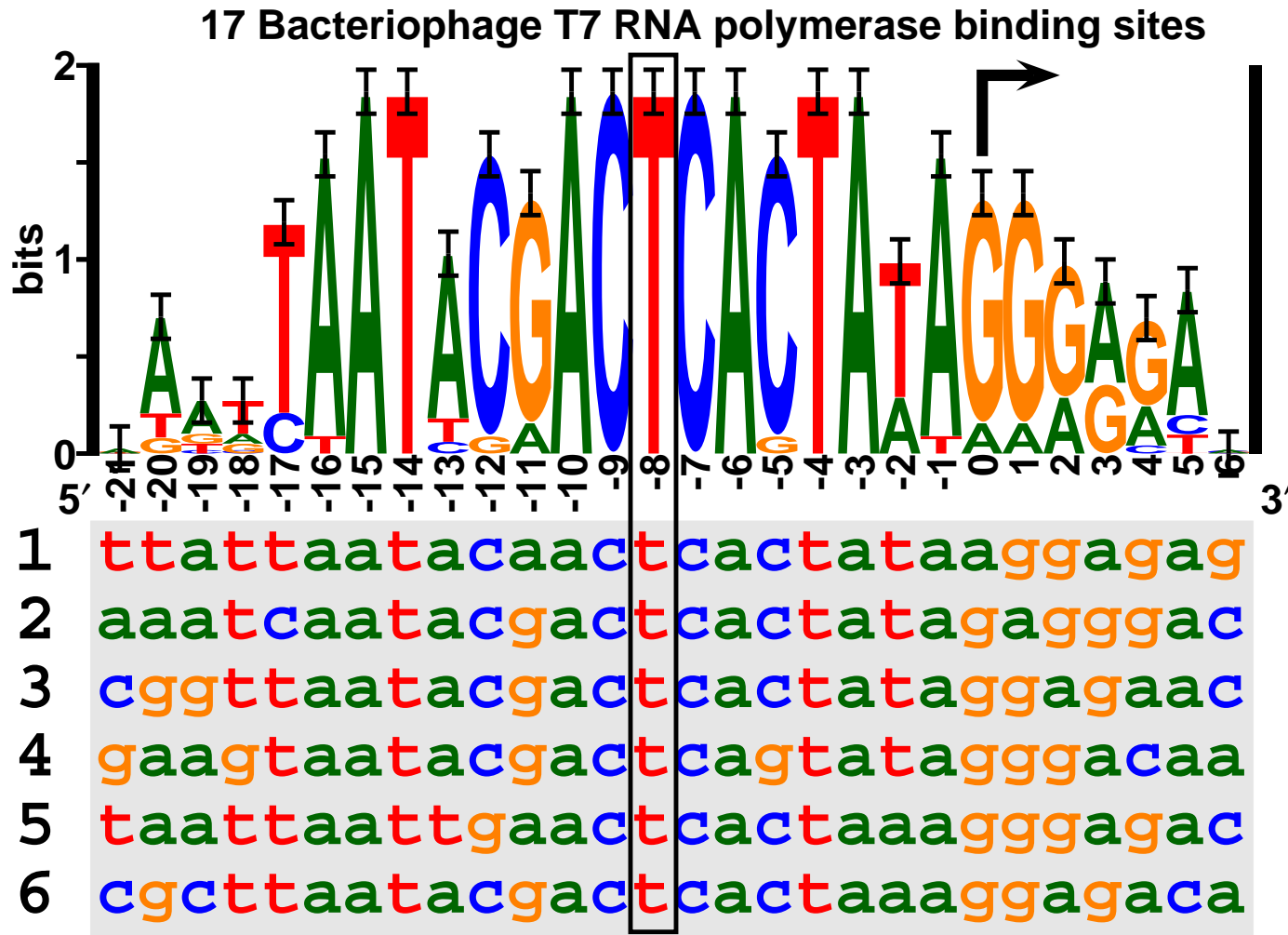


Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

```
1 ttattaatacaactcactataaggagag
2 aaatcaatacgaactcactatagaggac
3 cggttaatacgaactcactataggagaac
4 gaagtaatacgaactcagtatagggacaa
5 taattaattgaactcactaaaggggagac
6 cgcttaatacgaactcactaaaggagaca
```

6 of 17 sites

Sequence Logo

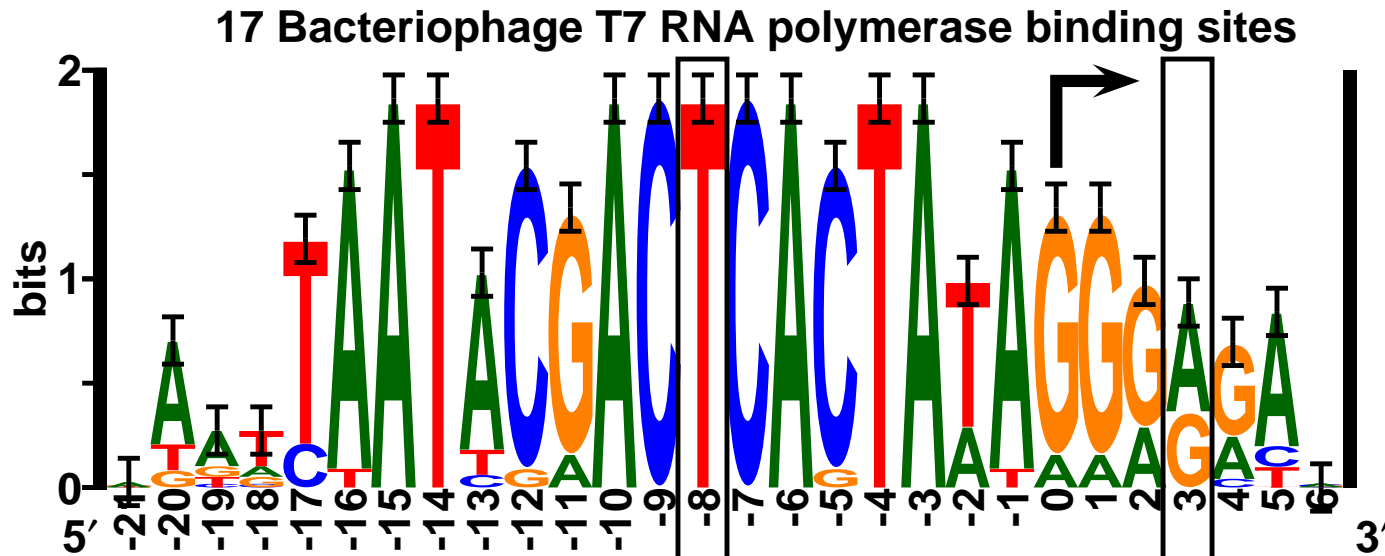


Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

6 of 17 sites

2 bits/base

Sequence Logo



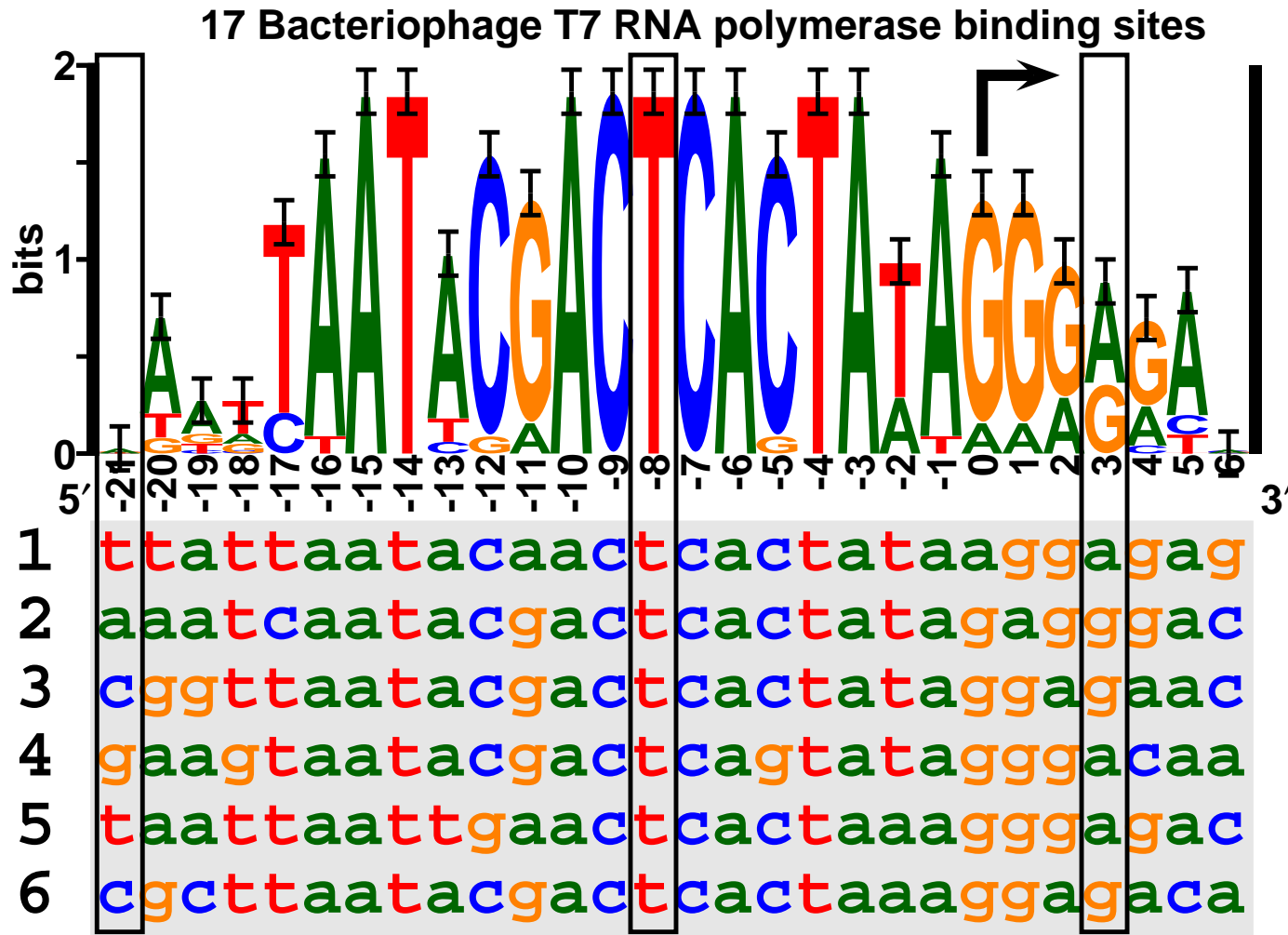
Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

1	t	t	a	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	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	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

1 bit/base

Sequence Logo

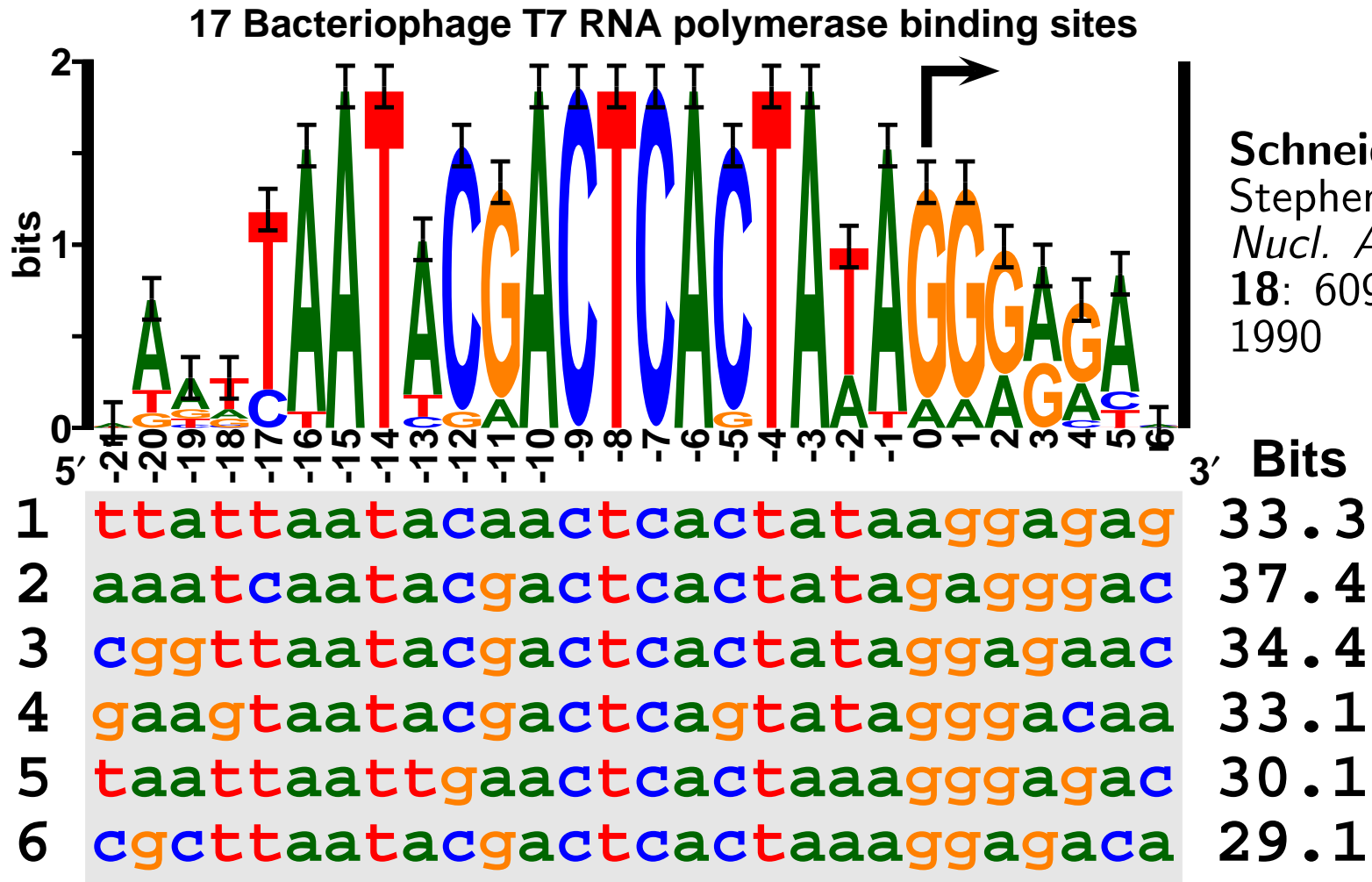


Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

6 of 17 sites

0 bits/base

Sequence Logo

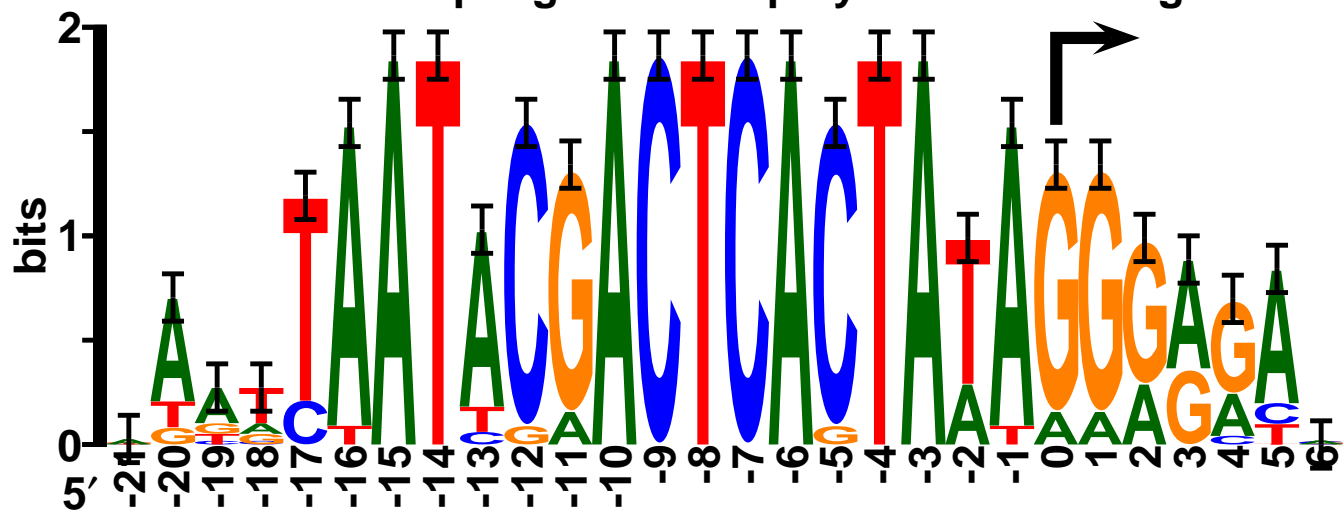


Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

Individual Information

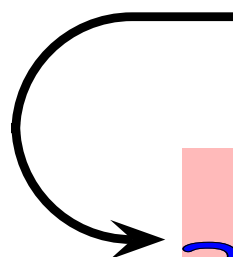
Sequence Logo and Sequence Walker

17 Bacteriophage T7 RNA polymerase binding sites



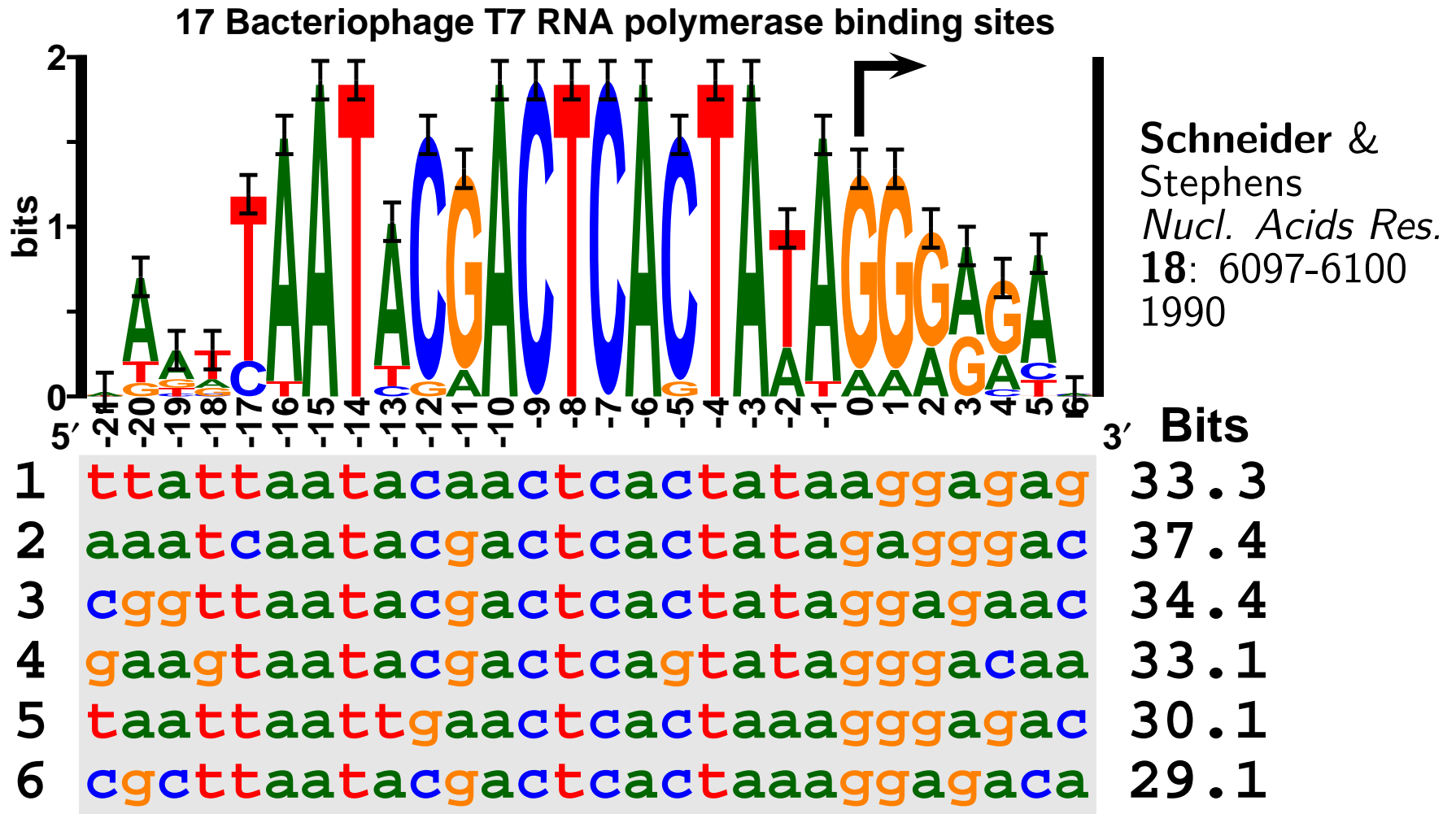
Schneider &
Stephens
Nucl. Acids Res.
18: 6097-6100
1990

	Sequence	Bits
1	ttattaatacaactcactataaggagag	33.3
2	aatcaatacgactcactatagagggac	37.4
3	ggttaatacgactcactataggagaac	34.4
4	gaagtaatacgactcagtatagggacaa	33.1
5	taattaattgaactcactaaaggggagac	30.1
6	cgcttaatacgactcactaaaggagaca	29.1



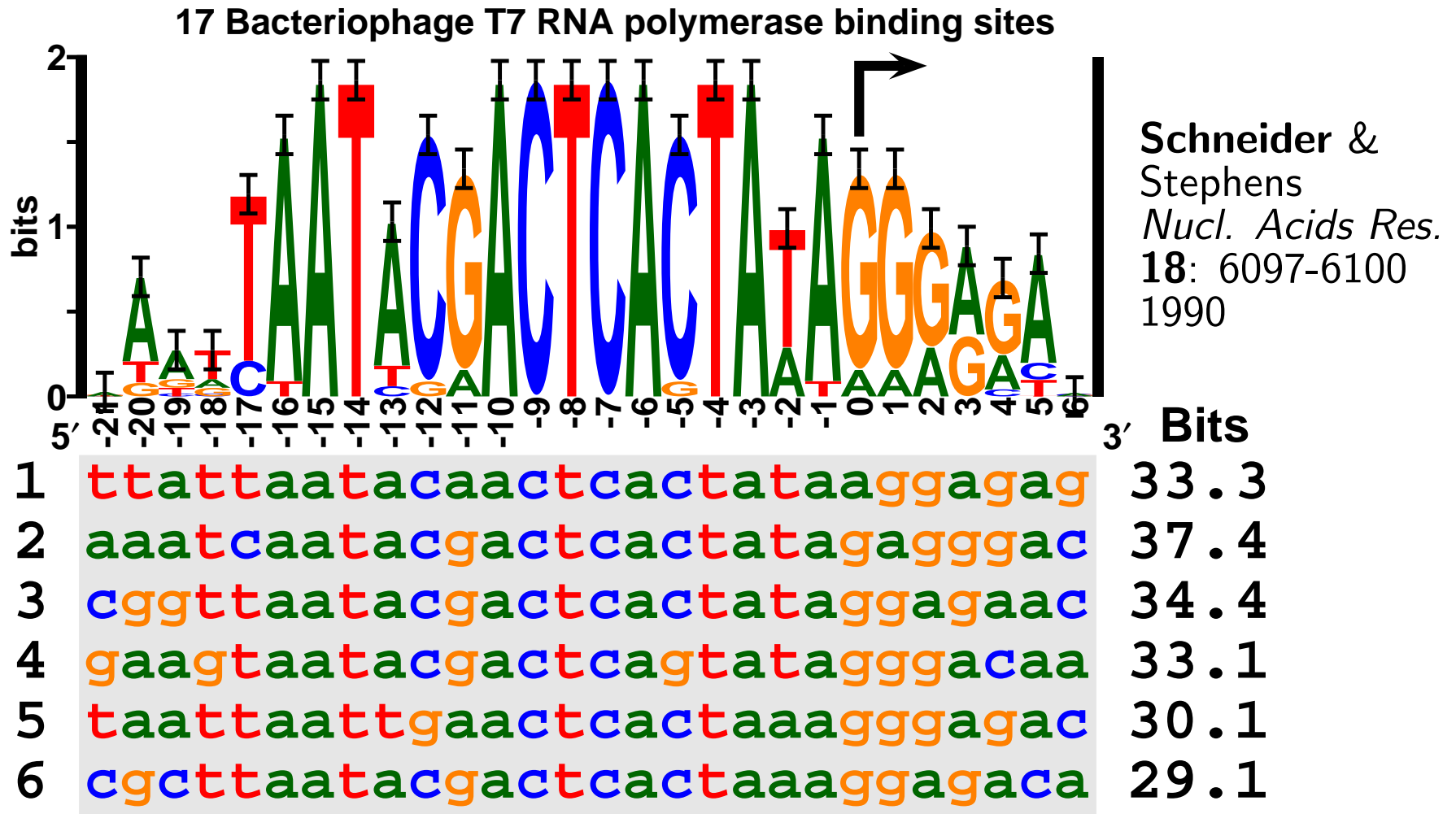
Sequence
Walker
Patent
5,867,402

Sequence Logo and Sequence Walker and Rsequence



Rsequence is the average: 35.0 ± 0.6 bits

Sequence Logo and Sequence Walker and Rsequence



Rsequence is the average $= \frac{\text{Area Under the logo}}{35}$

An Intuitive Approach

Information to chose one symbol from M symbols:

$$\log_2 M \quad (1)$$

An Intuitive Approach

Information to chose one symbol from M symbols:

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

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

An Intuitive Approach

Information to choose one symbol from M symbols:

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

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

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.



$$P_{\text{ring}} = 1/1024 \quad (3)$$

$$P_{\text{silent}} = 1023/1024 \quad (4)$$

EXAMPLE



A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \quad (3)$$

$$P_{\text{silent}} = 1023/1024 \quad (4)$$

Surprisal:

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

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

EXAMPLE



A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \quad (3)$$

$$P_{\text{silent}} = 1023/1024 \quad (4)$$

Surprisal:

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

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

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

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

EXAMPLE



A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \quad (3)$$

$$P_{\text{silent}} = 1023/1024 \quad (4)$$

Surprisal:

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

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

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

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}} + P_{\text{silent}} \times \text{surprisal}_{\text{silent}} \quad (7)$$

More Information Theory - 2

EXAMPLE



A phone rings once every 1024 seconds.

$$P_{\text{ring}} = 1/1024 \quad (3)$$

$$P_{\text{silent}} = 1023/1024 \quad (4)$$

Surprisal:

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

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

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

$$H = P_{\text{ring}} \times \text{surprisal}_{\text{ring}} + P_{\text{silent}} \times \text{surprisal}_{\text{silent}} \quad (7)$$

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

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

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

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

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

$$= \sum_{i=1}^M P_i \times (-\log_2 P_i) \quad (10)$$

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

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

$$= \sum_{i=1}^M P_i \times (-\log_2 P_i) \quad (10)$$

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

More Information Theory - Example

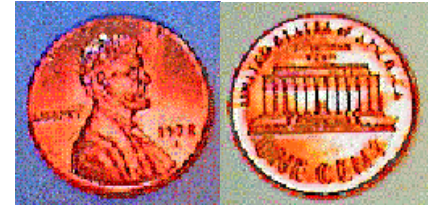
$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$



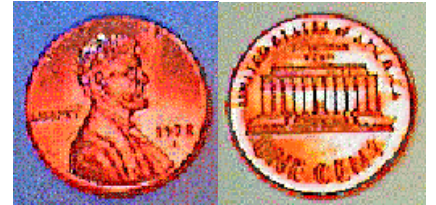
More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

$$P_2 = 1 - P_1 \quad (3)$$



More Information Theory - Example

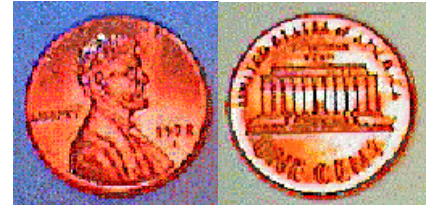
$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

$$P_2 = 1 - P_1 \quad (3)$$

$$H = -P_1 \log_2 P_1 \\ + \\ -P_2 \log_2 P_2 \quad (4)$$



More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

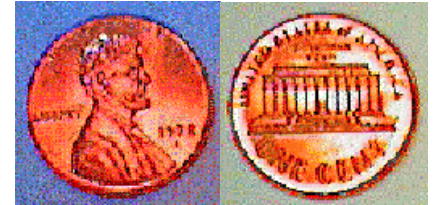
$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

$$P_2 = 1 - P_1 \quad (3)$$

$$H = -P_1 \log_2 P_1 \quad (4)$$

$$+ \\ -P_2 \log_2 P_2$$

$$H = -P_1 \log_2 P_1 + (-(1-P_1) \log_2(1-P_1)) \quad (5)$$



More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

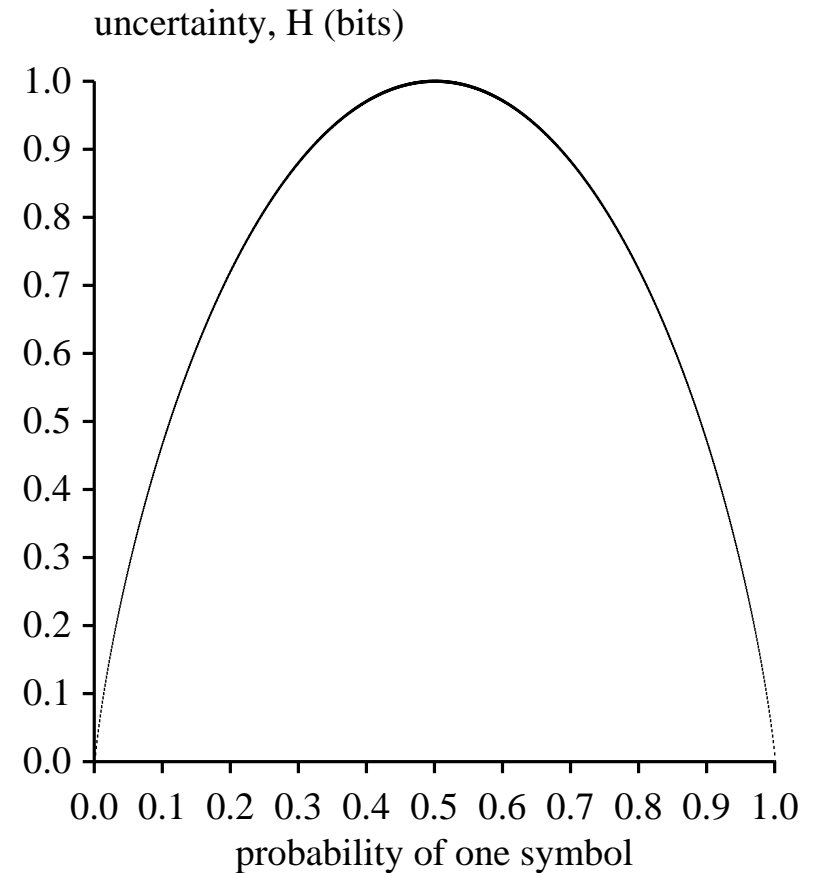
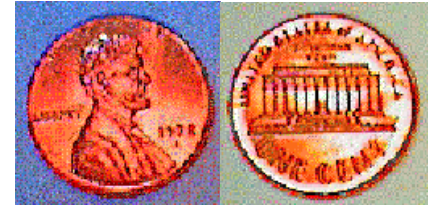
$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

$$P_2 = 1 - P_1 \quad (3)$$

$$H = -P_1 \log_2 P_1 \quad (4)$$

$$+ \\ -P_2 \log_2 P_2$$

$$H = -P_1 \log_2 P_1 + (-(1-P_1) \log_2(1-P_1)) \quad (5)$$



More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

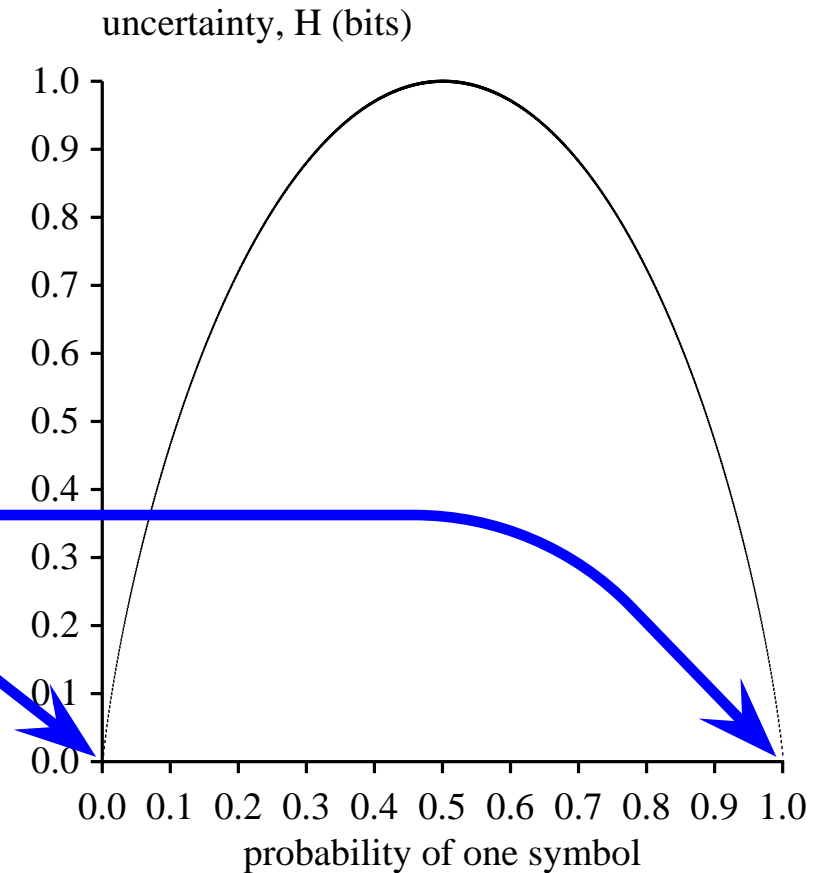
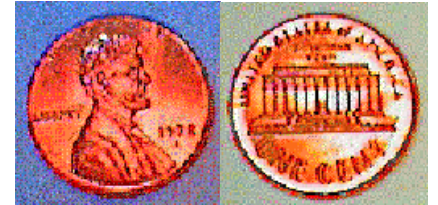
$$P_2 = 1 - P_1 \quad (3)$$

$$H = -P_1 \log_2 P_1 \quad (4)$$

$$+ \\ -P_2 \log_2 P_2$$

$$H = -P_1 \log_2 P_1 + (-(1-P_1) \log_2(1-P_1)) \quad (5)$$

NOTE 1: $\lim_{P \rightarrow 0} P \log_2 P = 0$



More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

$$P_2 = 1 - P_1 \quad (3)$$

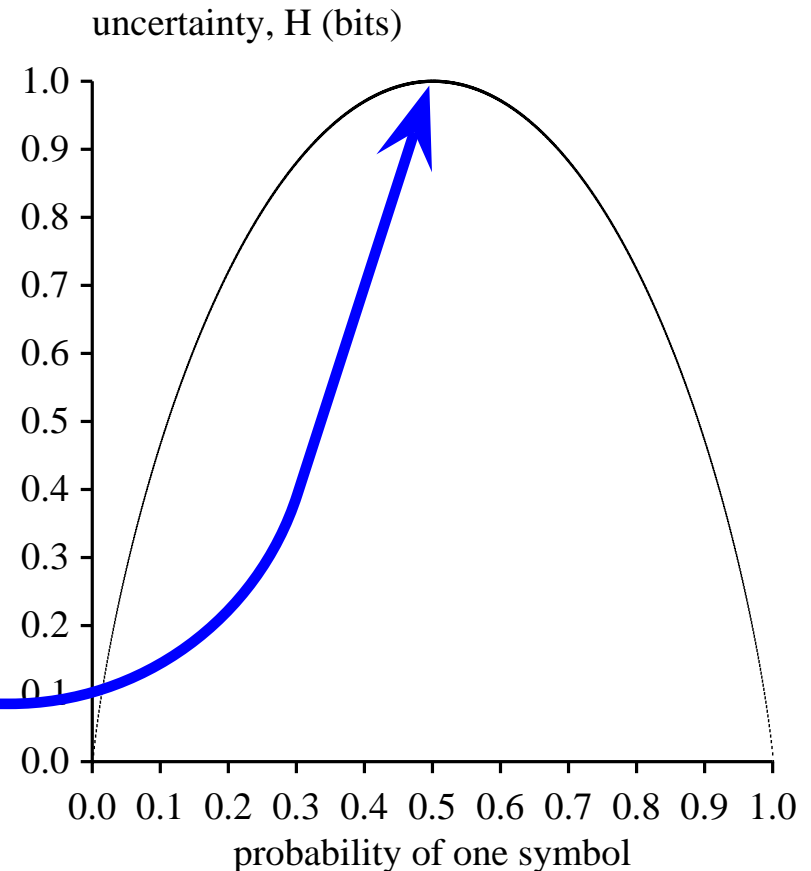
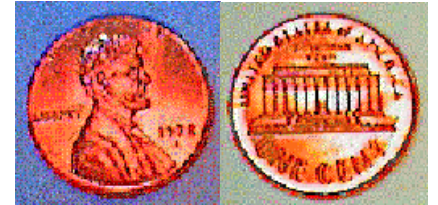
$$H = -P_1 \log_2 P_1 \quad (4)$$

$$+ \\ -P_2 \log_2 P_2$$

$$H = -P_1 \log_2 P_1 + (-(1-P_1) \log_2(1-P_1)) \quad (5)$$

NOTE 1: $\lim_{P \rightarrow 0} P \log_2 P = 0$

NOTE 2: The curve peaks at $P_1 = 0.5$ when $P_1 = (1 - P_1) = P_2$.



More Information Theory - Example

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

Simplified Example For two symbols, plot the uncertainty

$$M = 2, \quad P_1 + P_2 = 1 \quad (2)$$

$$P_2 = 1 - P_1 \quad (3)$$

$$H = -P_1 \log_2 P_1 \quad (4)$$

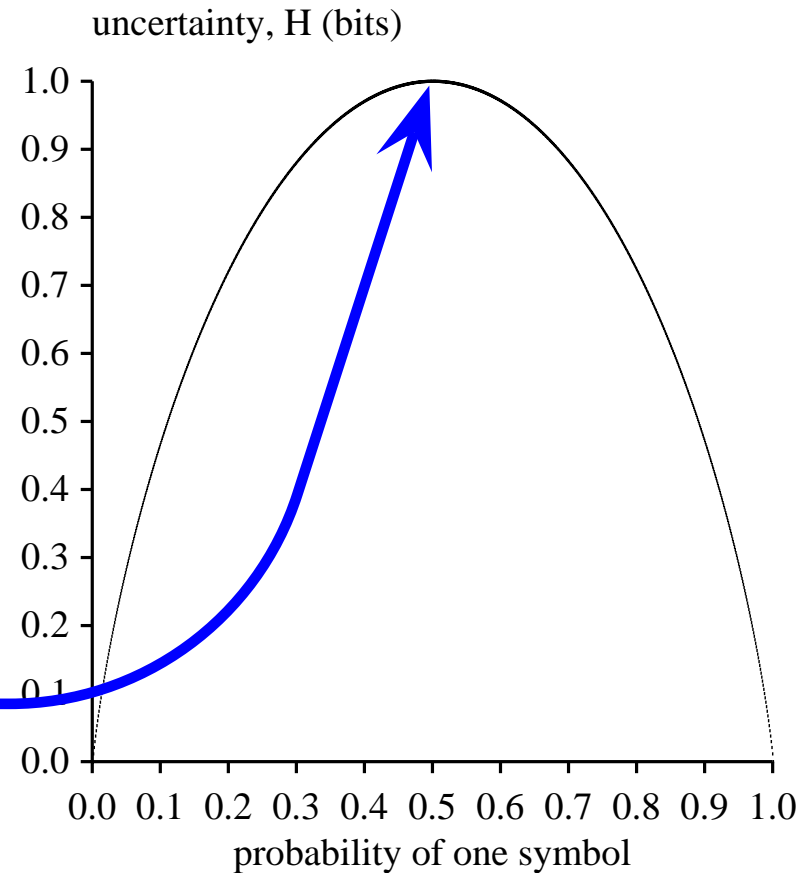
$$+ \\ -P_2 \log_2 P_2$$

$$H = -P_1 \log_2 P_1 + (-(1-P_1) \log_2(1-P_1)) \quad (5)$$

NOTE 1: $\lim_{P \rightarrow 0} P \log_2 P = 0$

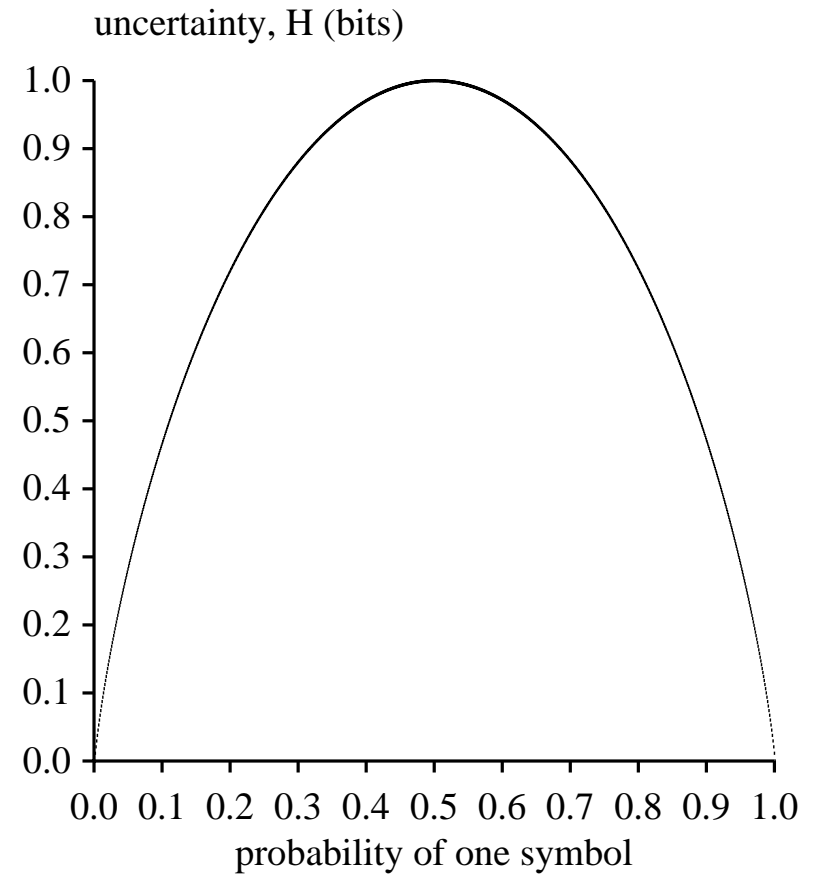
NOTE 2: The curve peaks at $P_1 = 0.5$ when $P_1 = (1 - P_1) = P_2$.

Maximum uncertainty is at equal probability.



More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

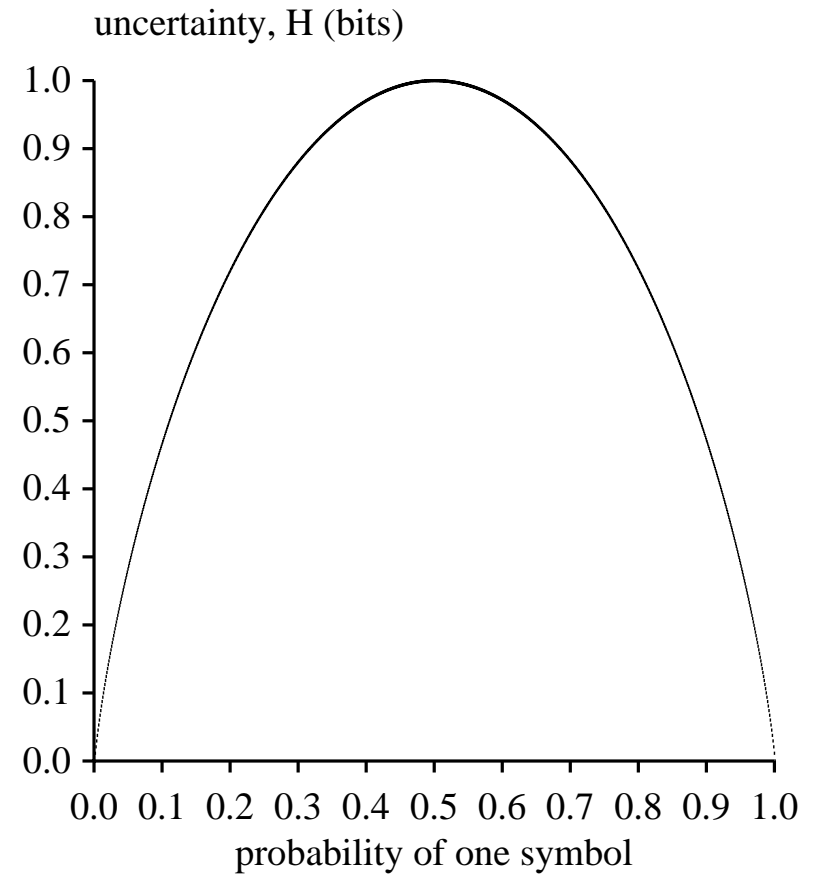


More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

All Equiprobable Symbols

$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$



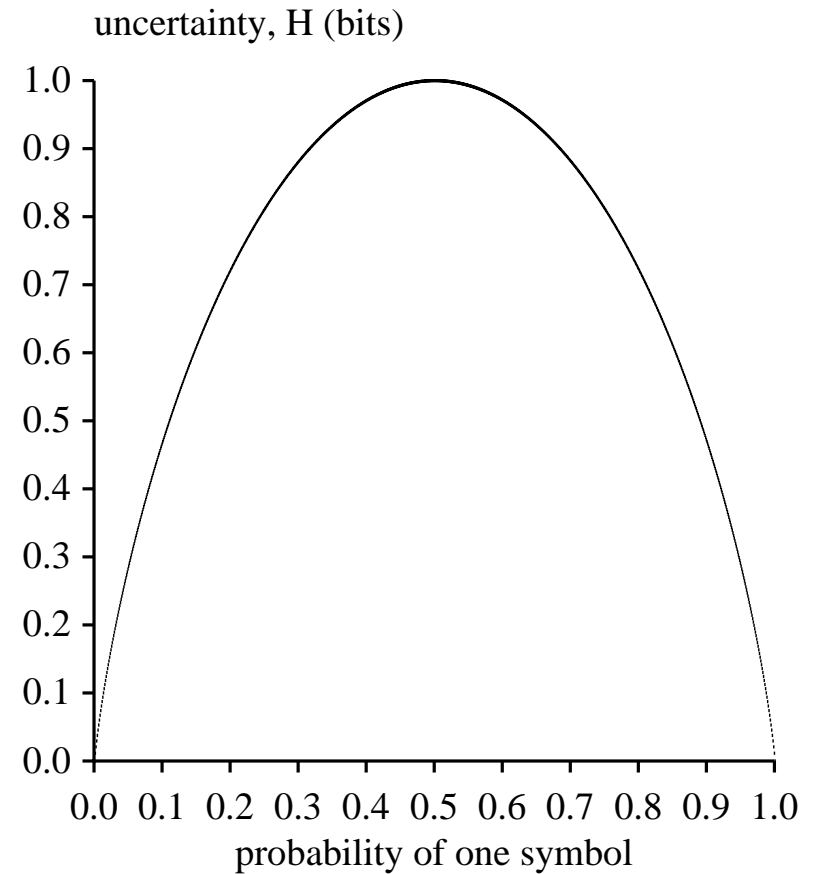
More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

All Equiprobable Symbols

$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$

$$H = - \sum_{i=1}^M \frac{1}{M} \log_2 \frac{1}{M} \quad (3)$$



More Information Theory - Maximum

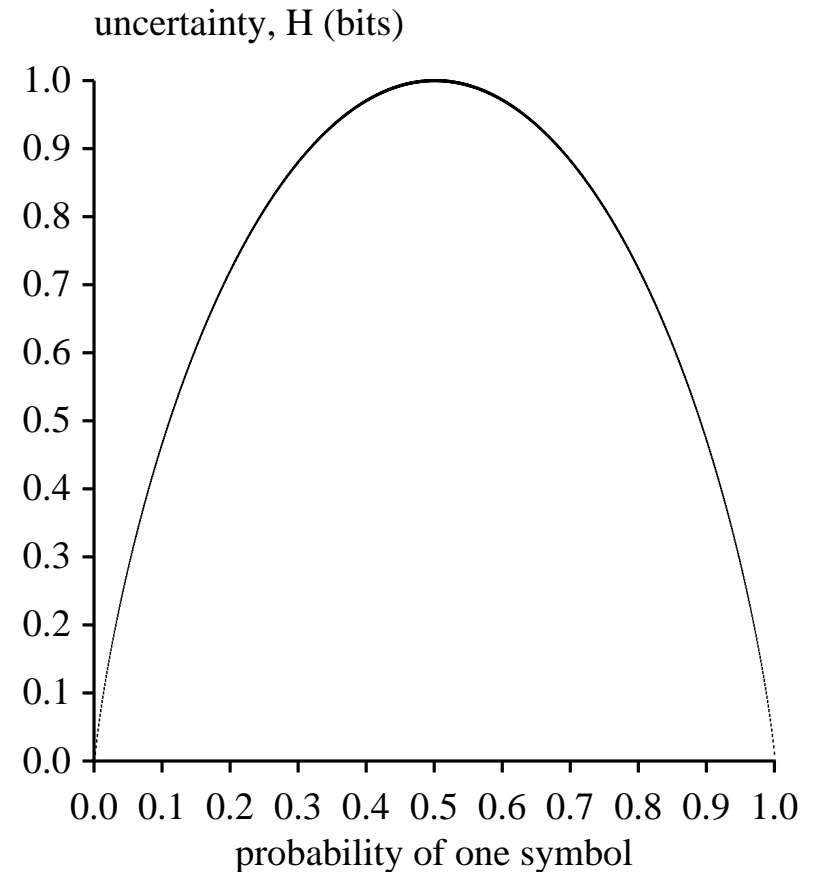
$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

All Equiprobable Symbols

$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$

$$H = - \sum_{i=1}^M \frac{1}{M} \log_2 \frac{1}{M} \quad (3)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) \sum_{i=1}^M 1 \quad (4)$$



More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

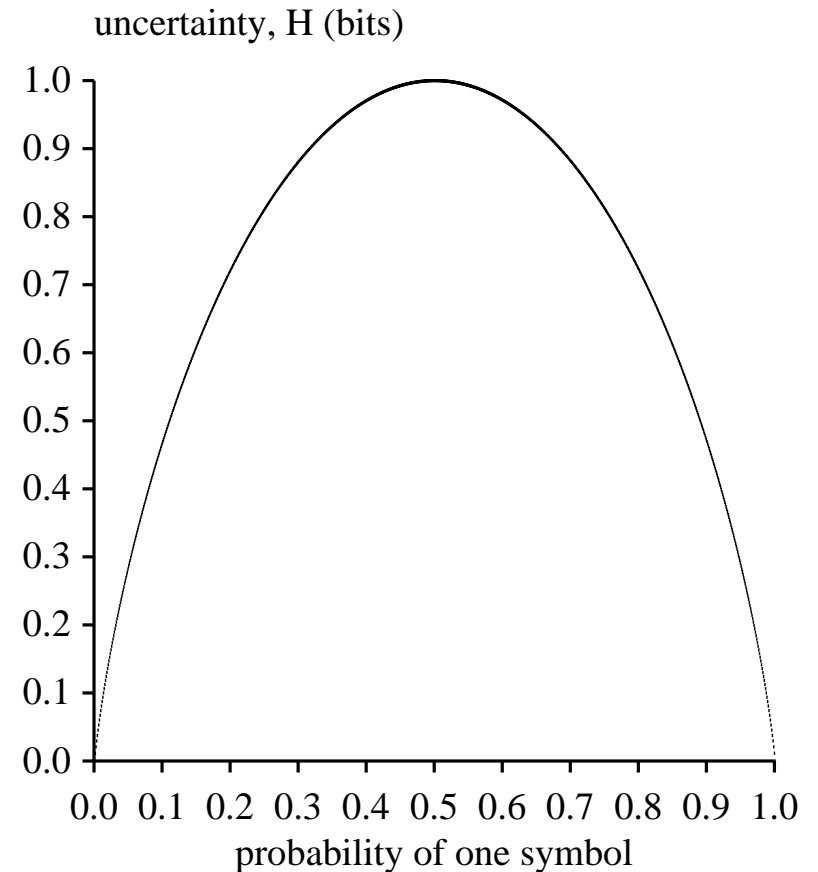
All Equiprobable Symbols

$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$

$$H = - \sum_{i=1}^M \frac{1}{M} \log_2 \frac{1}{M} \quad (3)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) \sum_{i=1}^M 1 \quad (4)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) M \quad (5)$$



More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

All Equiprobable Symbols

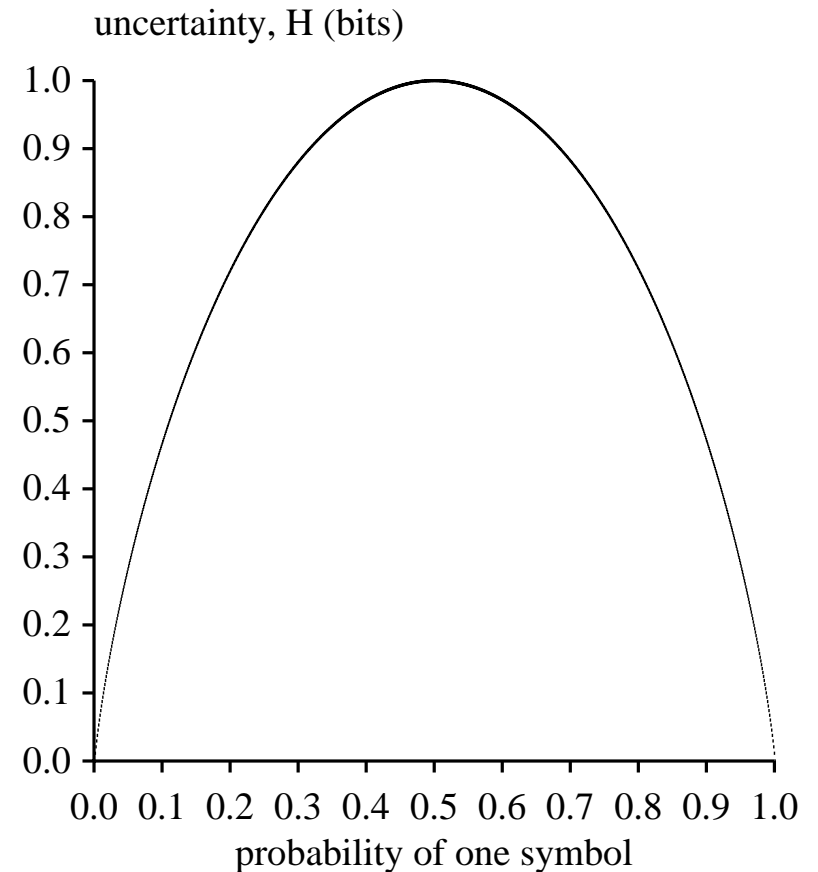
$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$

$$H = - \sum_{i=1}^M \frac{1}{M} \log_2 \frac{1}{M} \quad (3)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) \sum_{i=1}^M 1 \quad (4)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) M \quad (5)$$

$$H = - \log_2 \frac{1}{M} \quad (6)$$



More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

All Equiprobable Symbols

$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$

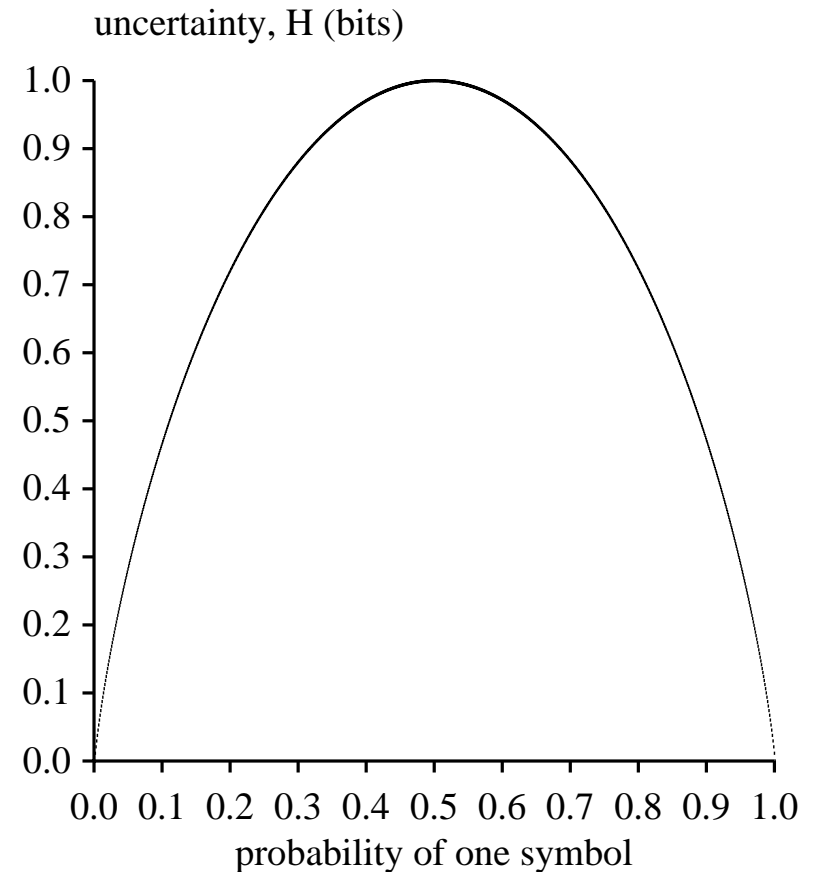
$$H = - \sum_{i=1}^M \frac{1}{M} \log_2 \frac{1}{M} \quad (3)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) \sum_{i=1}^M 1 \quad (4)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) M \quad (5)$$

$$H = - \log_2 \frac{1}{M} \quad (6)$$

$$H = \log_2 M \quad (7)$$



More Information Theory - Maximum

$$H = - \sum_{i=1}^M P_i \log_2 P_i \quad \text{bits per symbol} \quad (1)$$

All Equiprobable Symbols

$$P_i = \frac{1}{M}, \quad \text{for all } i \quad (2)$$

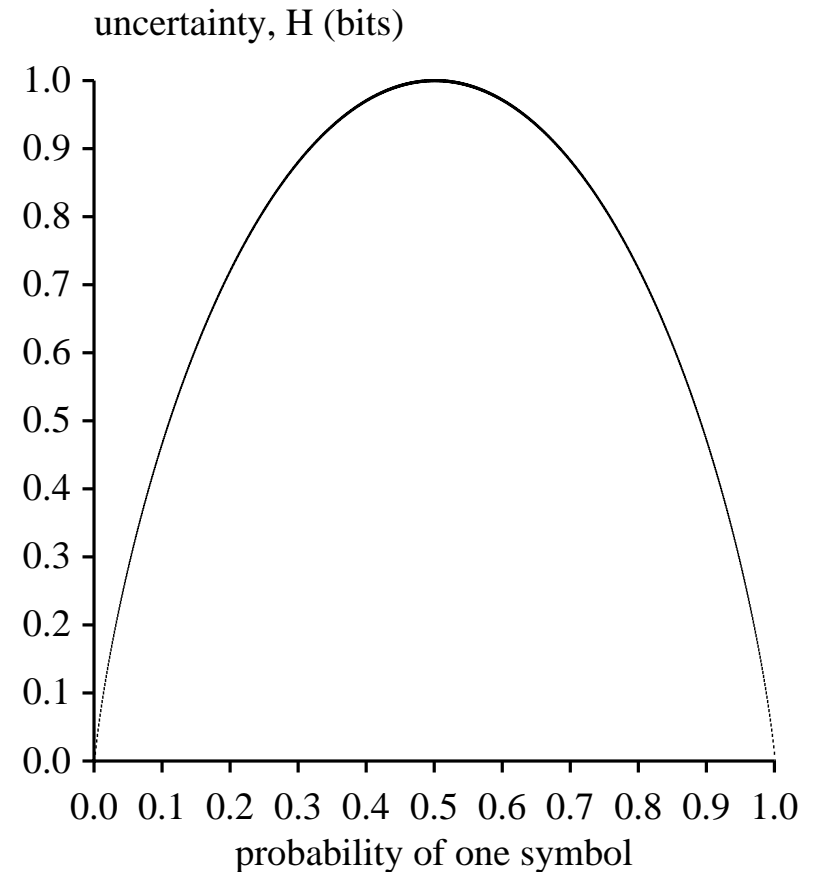
$$H = - \sum_{i=1}^M \frac{1}{M} \log_2 \frac{1}{M} \quad (3)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) \sum_{i=1}^M 1 \quad (4)$$

$$H = - \left(\frac{1}{M} \log_2 \frac{1}{M} \right) M \quad (5)$$

$$H = - \log_2 \frac{1}{M} \quad (6)$$

$$H = \log_2 M \quad (7)$$



Maximum uncertainty is at equal probability.

Information required to find a set of binding sites

$G = \#$ of potential binding sites

Information required to find a set of binding sites

G = # of potential binding sites
= genome size in some cases

Information required to find a set of binding sites

G = # of potential binding sites
= genome size in some cases

γ = number of binding sites on genome

Information required to find a set of binding sites

G = # of potential binding sites
= 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

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

γ = number of binding sites on genome

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

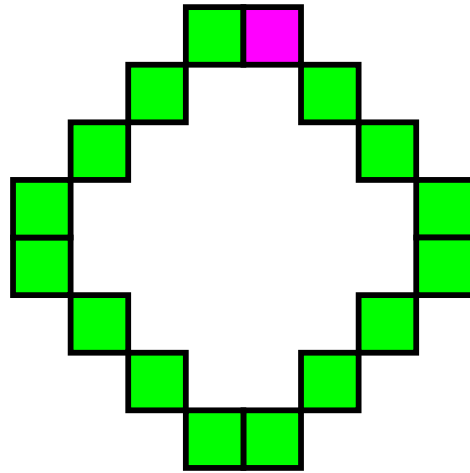
Information required to find a set of binding sites

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

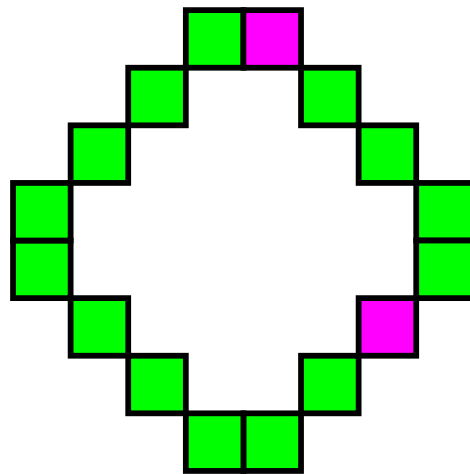
γ = number of binding sites on genome

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

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



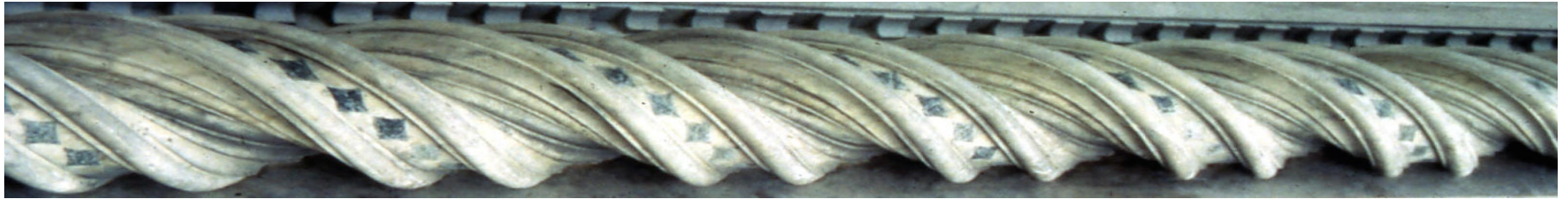
16 positions
1 site
 $\log_2 16/1 = 4$ bits



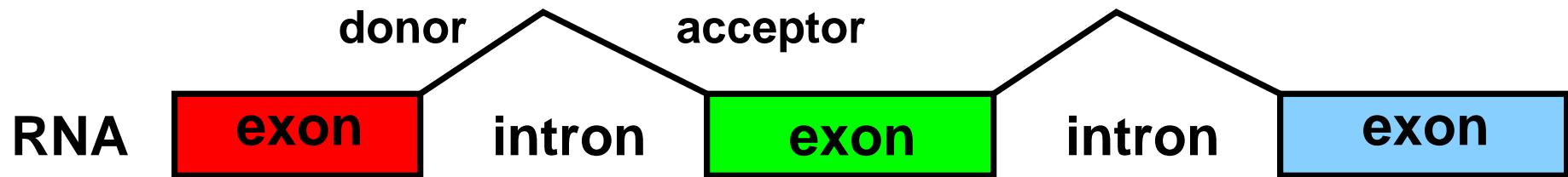
16 positions
2 sites
 $\log_2 16/2 = 3$ bits

RNA Splicing

DNA



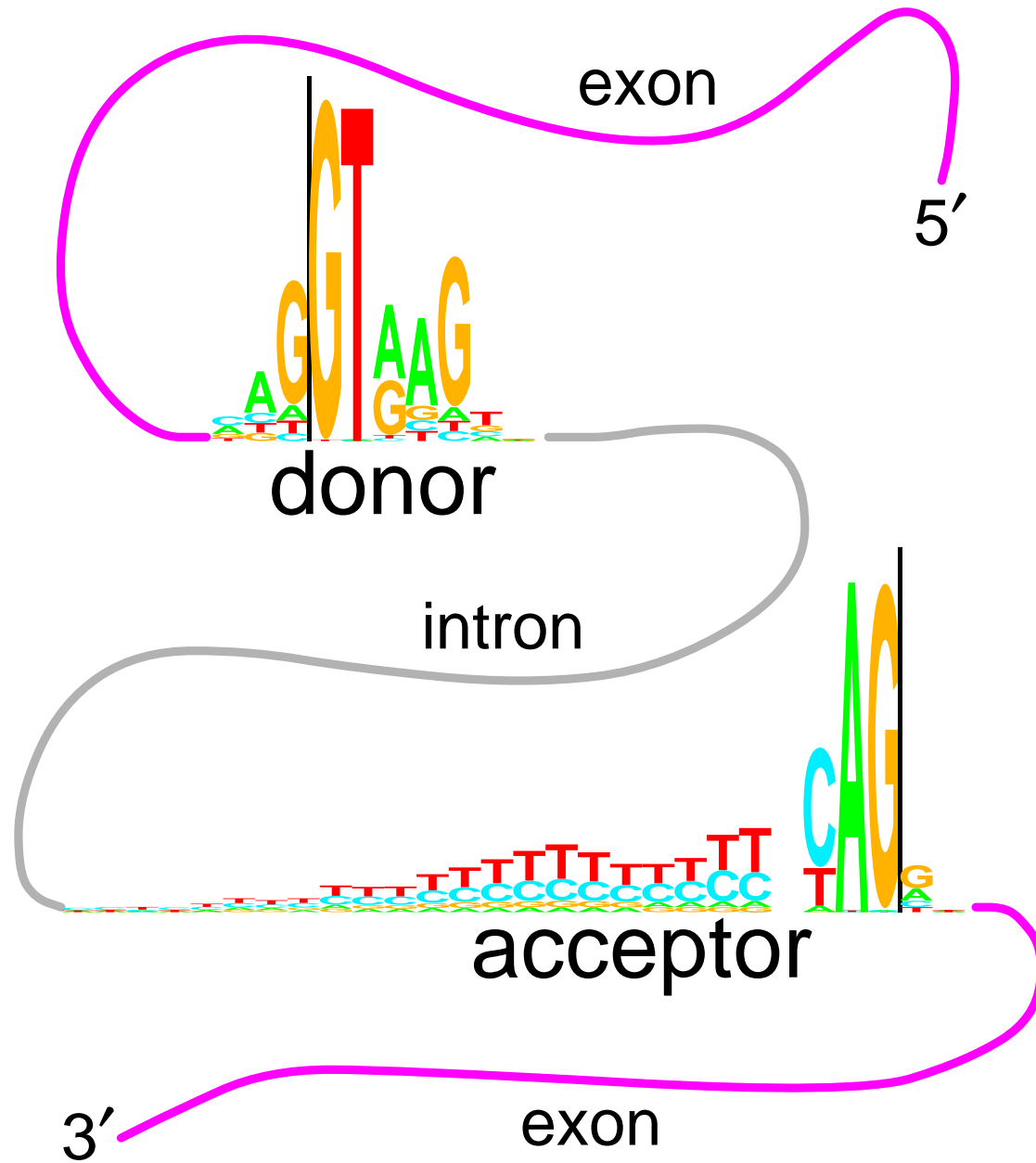
↓ Copy DNA (transcription)



↓ RNA Splicing



Donor and acceptor logos



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 = $R_{\text{frequency}}$ (bits)	$\frac{\text{Pattern Info}}{\text{Location Info}}$ = $\frac{R_{\text{sequence}}}{R_{\text{frequency}}}$
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 cl/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

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

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

³F. 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

$R_{sequence}$ must evolve towards $R_{frequency}$!

Evolution of Binding Sites

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

Evolution of Binding Sites

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

Evolution of Binding Sites

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

Setup a Computer Model, 'Ev':

A population of "creatures" with

Evolution of Binding Sites

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

Setup a Computer Model, 'Ev':

A population of "creatures" with

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

Evolution of Binding Sites

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

Setup a Computer Model, 'Ev':

A population of "creatures" with

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

Evolution of Binding Sites

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

Setup a Computer Model, 'Ev':

A population of "creatures" with

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

Evolution of Binding Sites

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

Setup a Computer Model, 'Ev':

A population of "creatures" with

- genomes containing 4 bases (A, C, G, T)
 - a defined genome size (G)
 - predetermined binding site locations (γ)
(to fix the frequency of sites)
- } $R_{frequency}$
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)
 - a recognizer gene encoded in the sequence:
use a weight matrix
- } $R_{frequency}$
is fixed

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

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

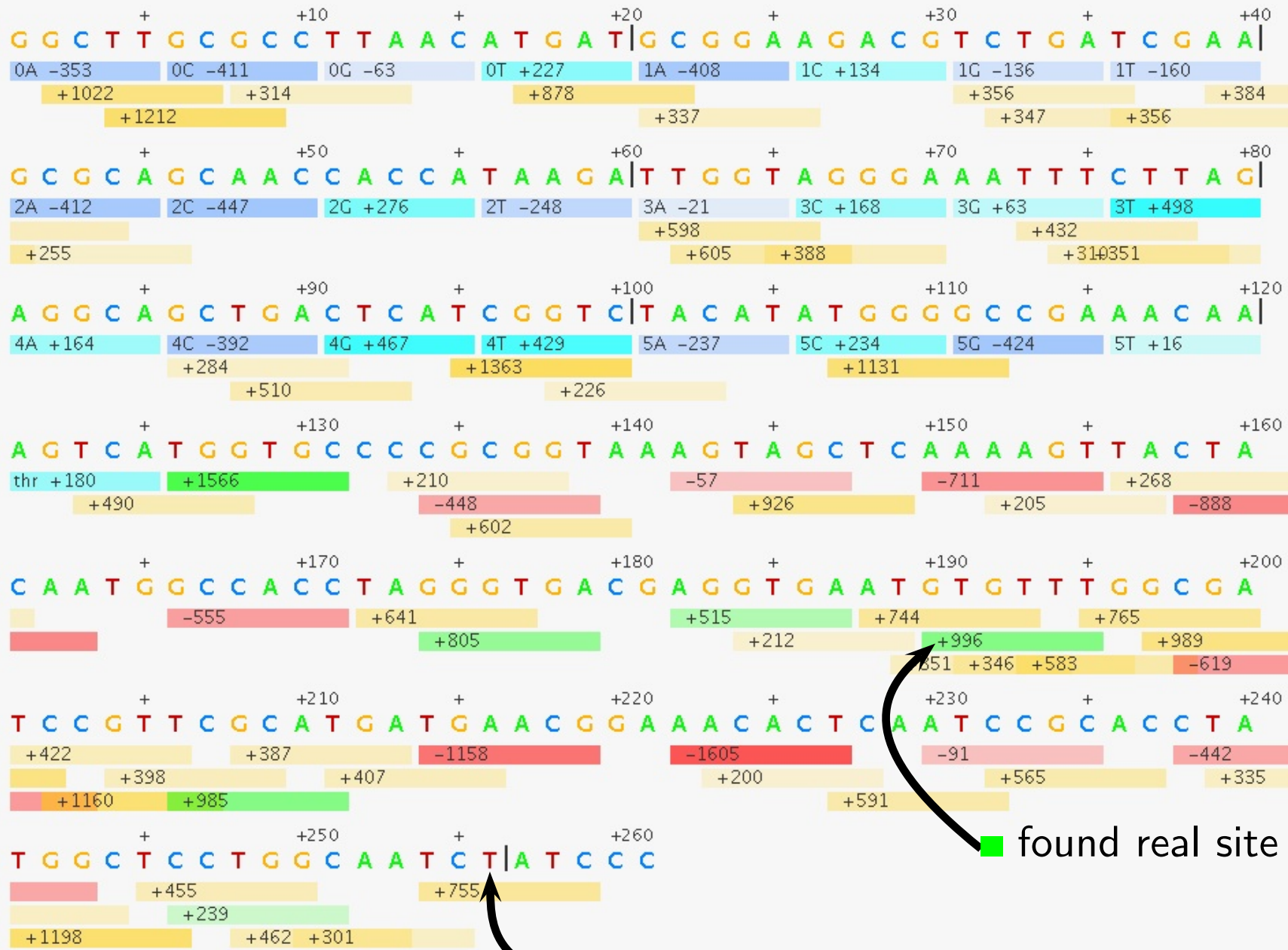


“blue”
gene
weight
matrix:
6 bp
wide

$\gamma = 16$
binding
sites

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

Unevolved Ev Creature



“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



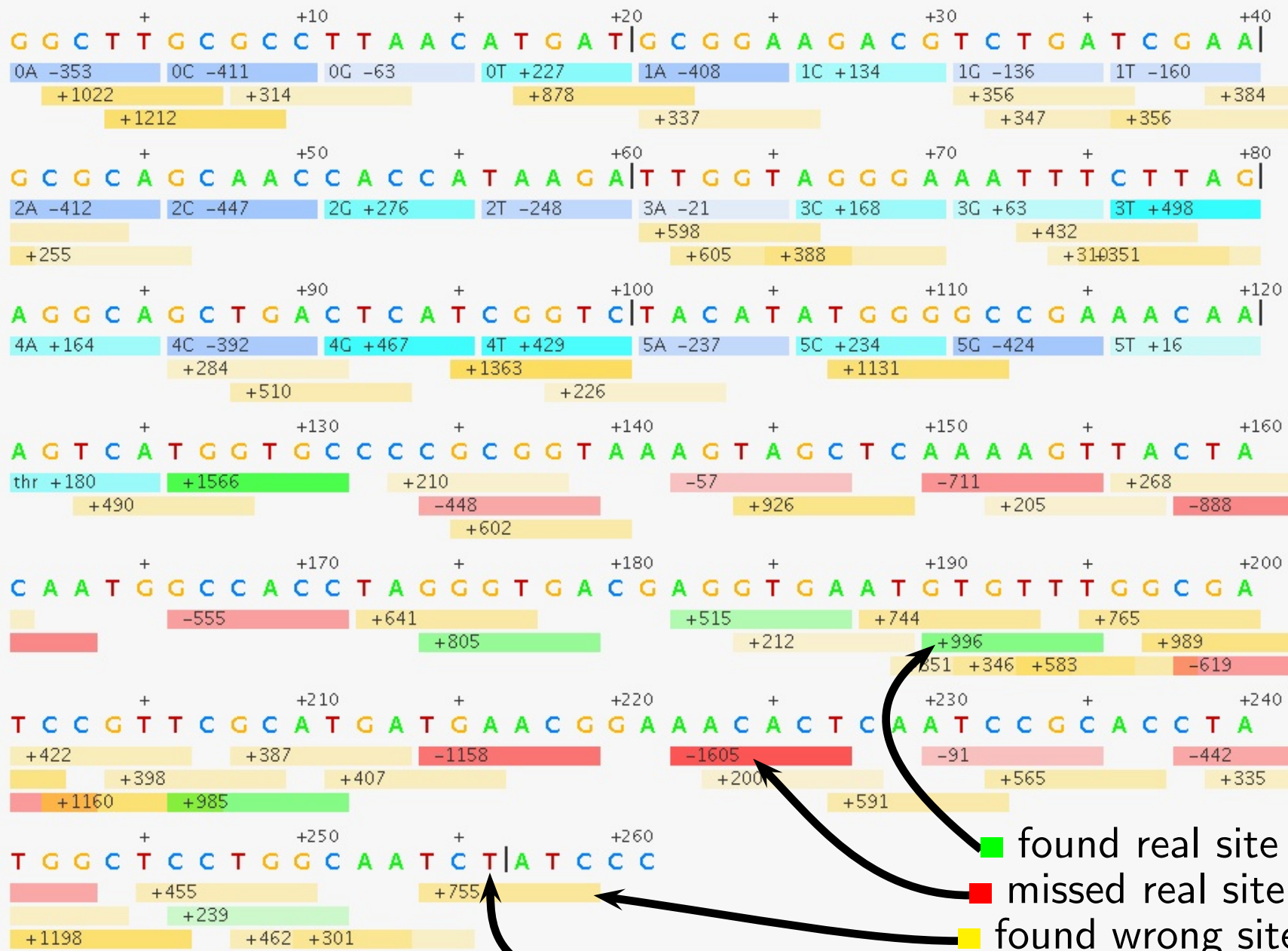
“blue” gene weight matrix: 6 bp wide

$\gamma = 16$ 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

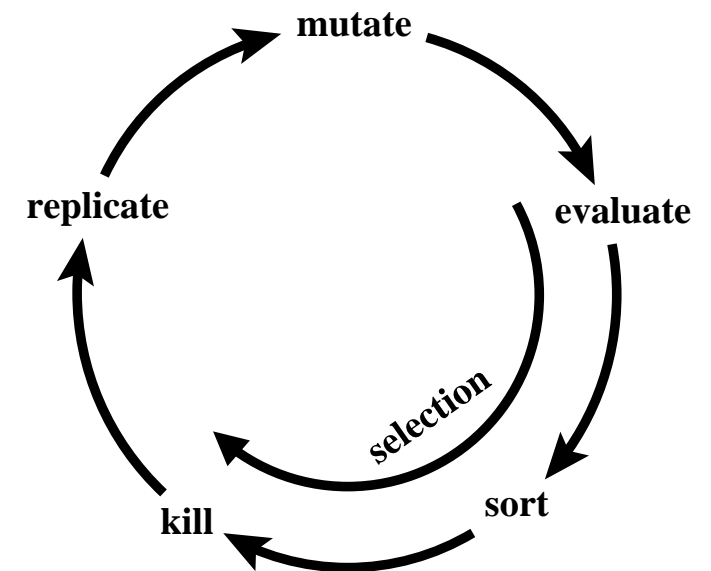
$\gamma = 16$
binding
sites

■ found real site
■ missed real site
■ found wrong site

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

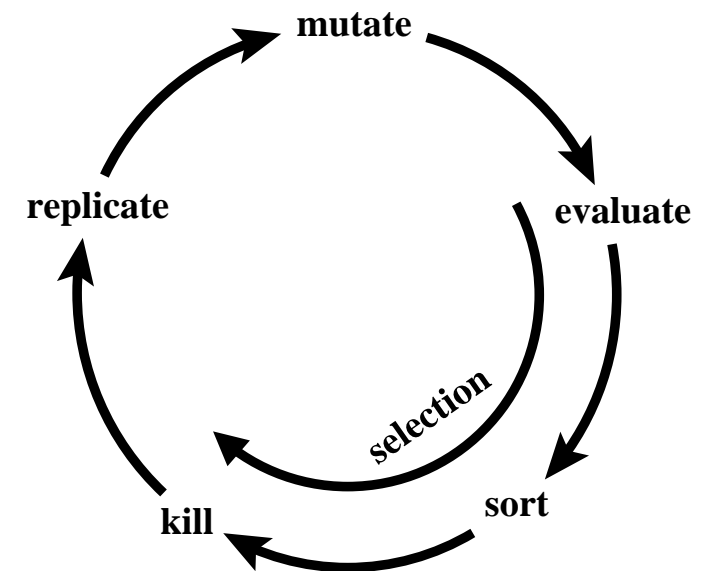
Evolution Cycle

- EVALUATE each creature



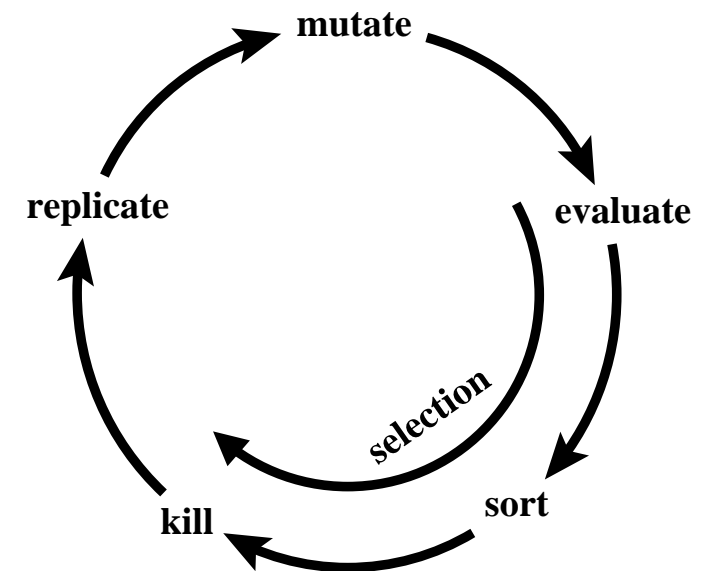
Evolution Cycle

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



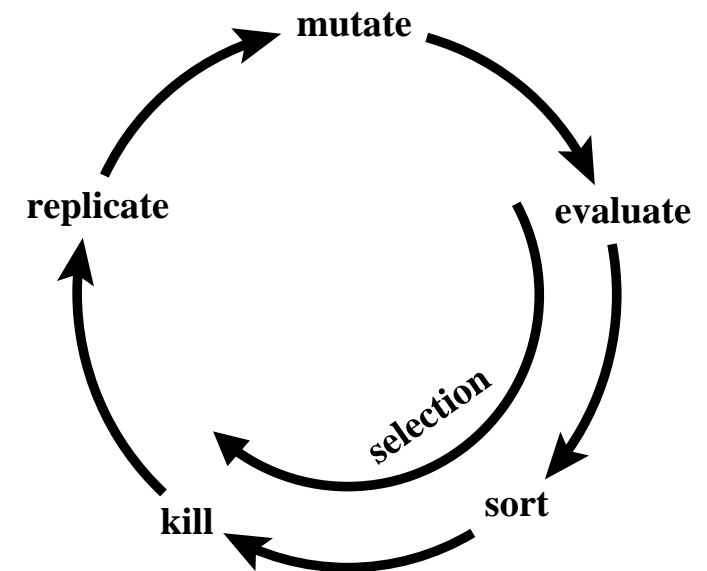
Evolution Cycle

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



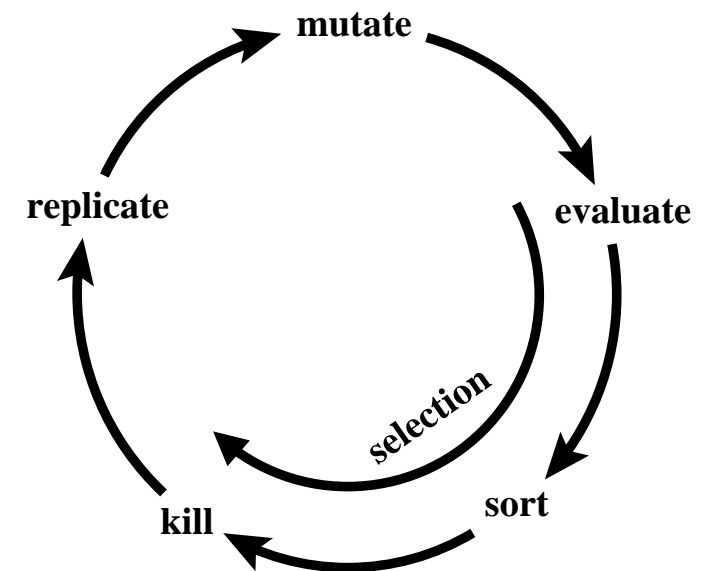
Evolution Cycle

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



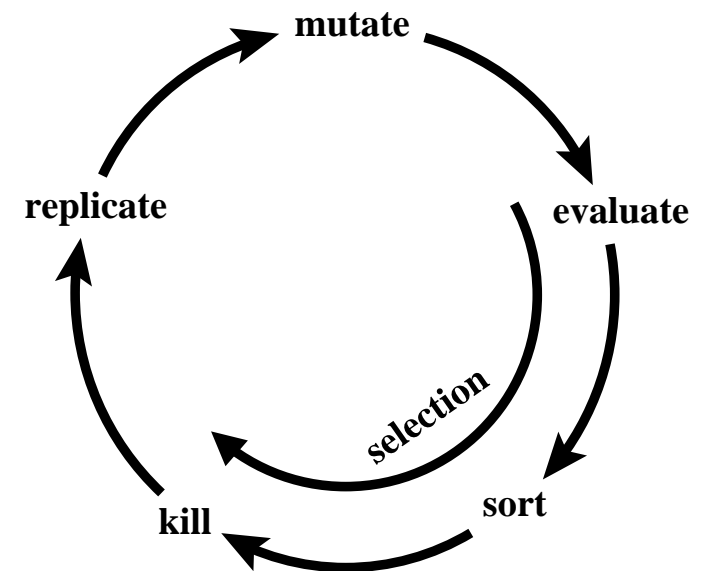
Evolution Cycle

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



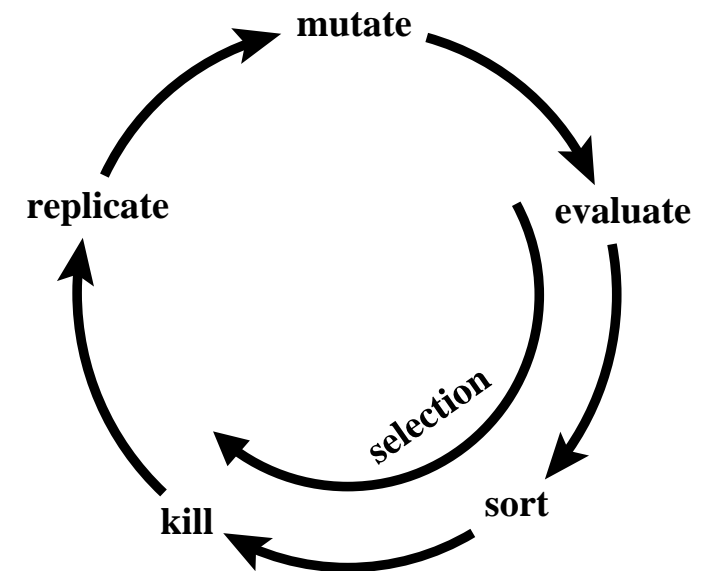
Evolution Cycle

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



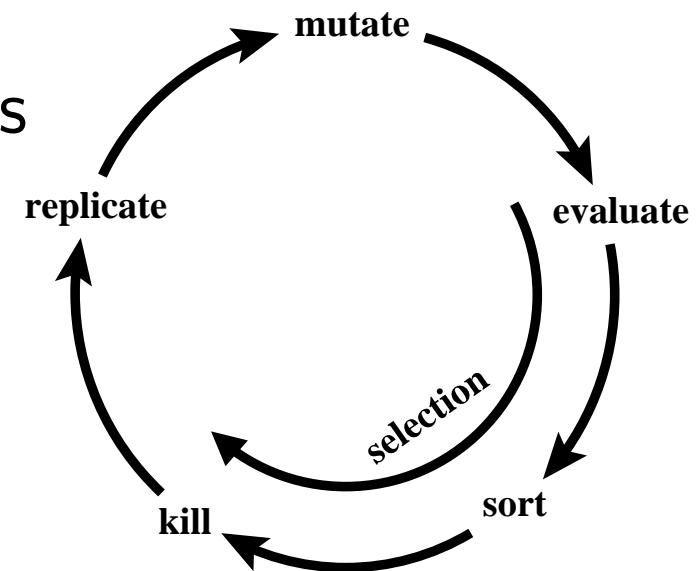
Evolution Cycle

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



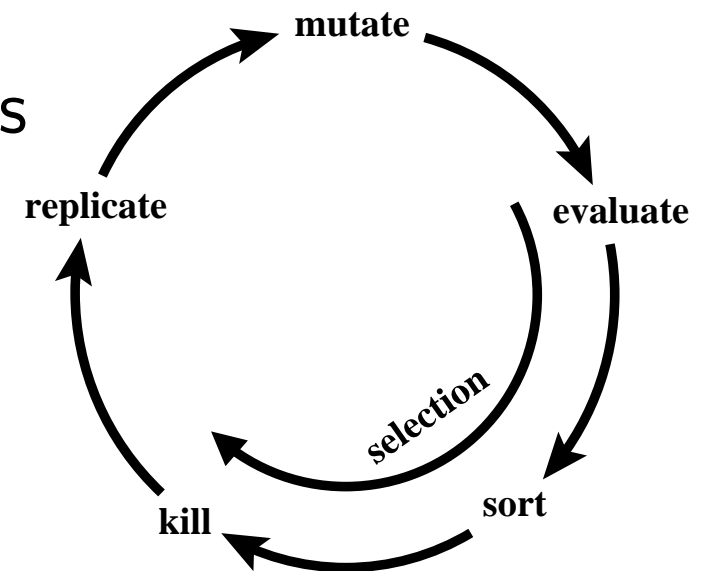
Evolution Cycle

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

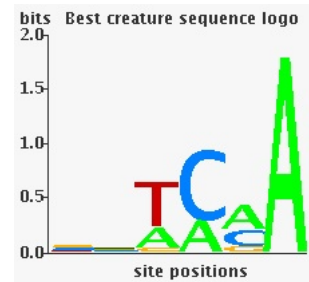


Evolution Cycle

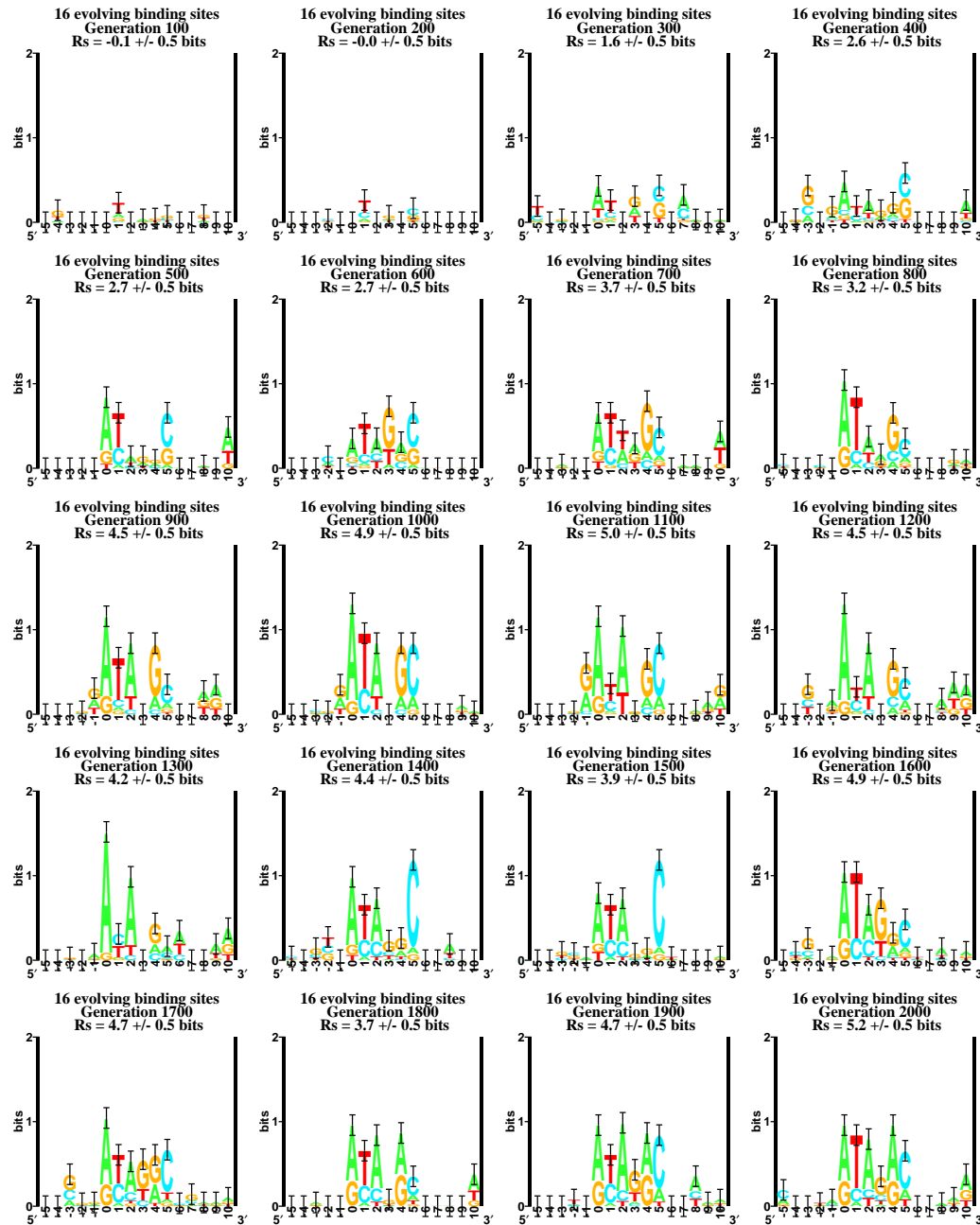
- EVALUATE each creature
 - translate the recognizer gene into a weight matrix
 - scan the weight matrix across the genome
 - count the number of mistakes:
 - missing a site at a right place
 - finding a site at a wrong place
 - Sort the creatures by their mistakes
- REPLICATE: the best creatures are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



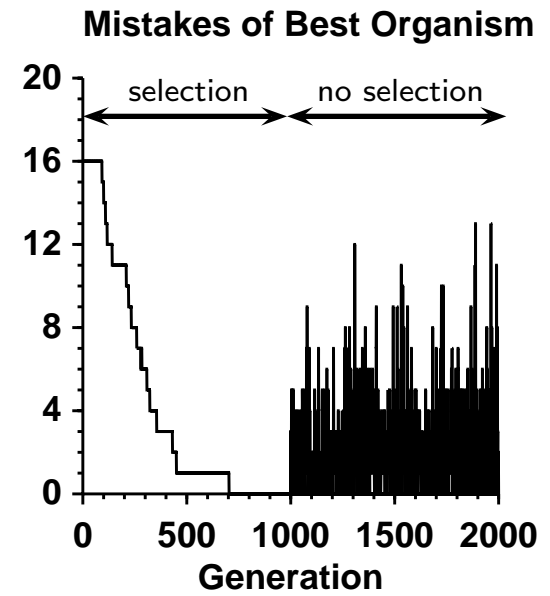
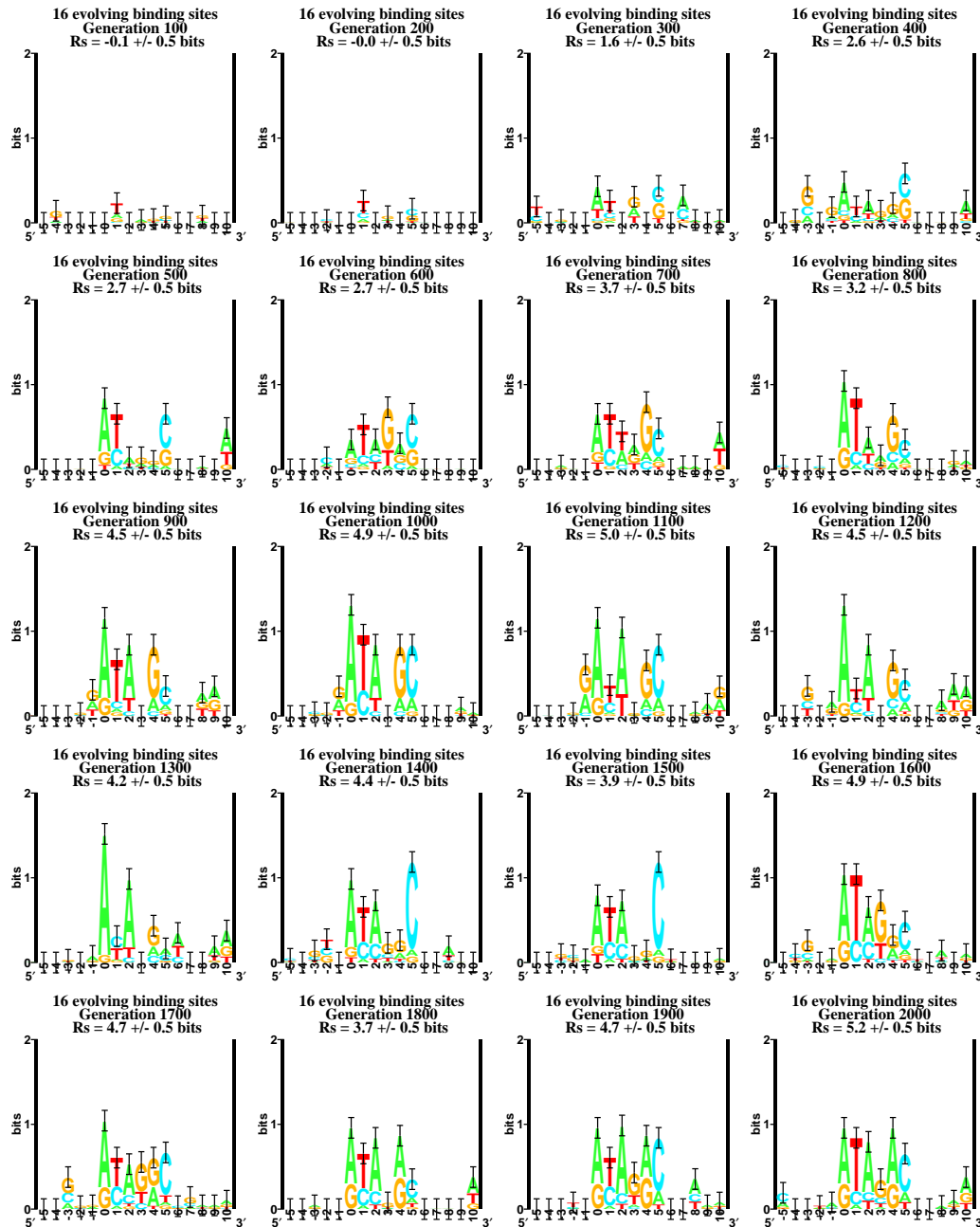
Evolved Ev Creature



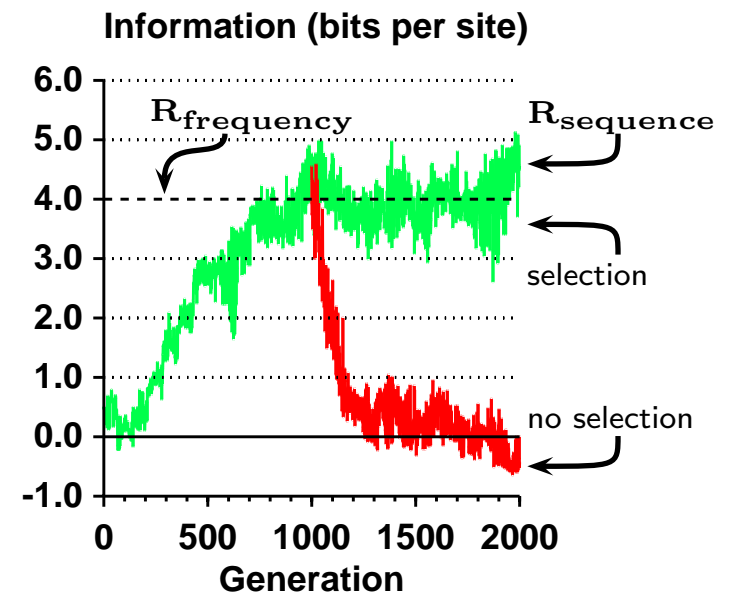
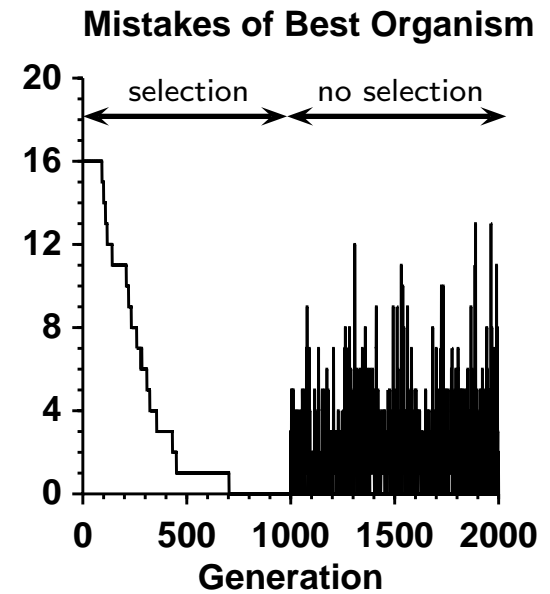
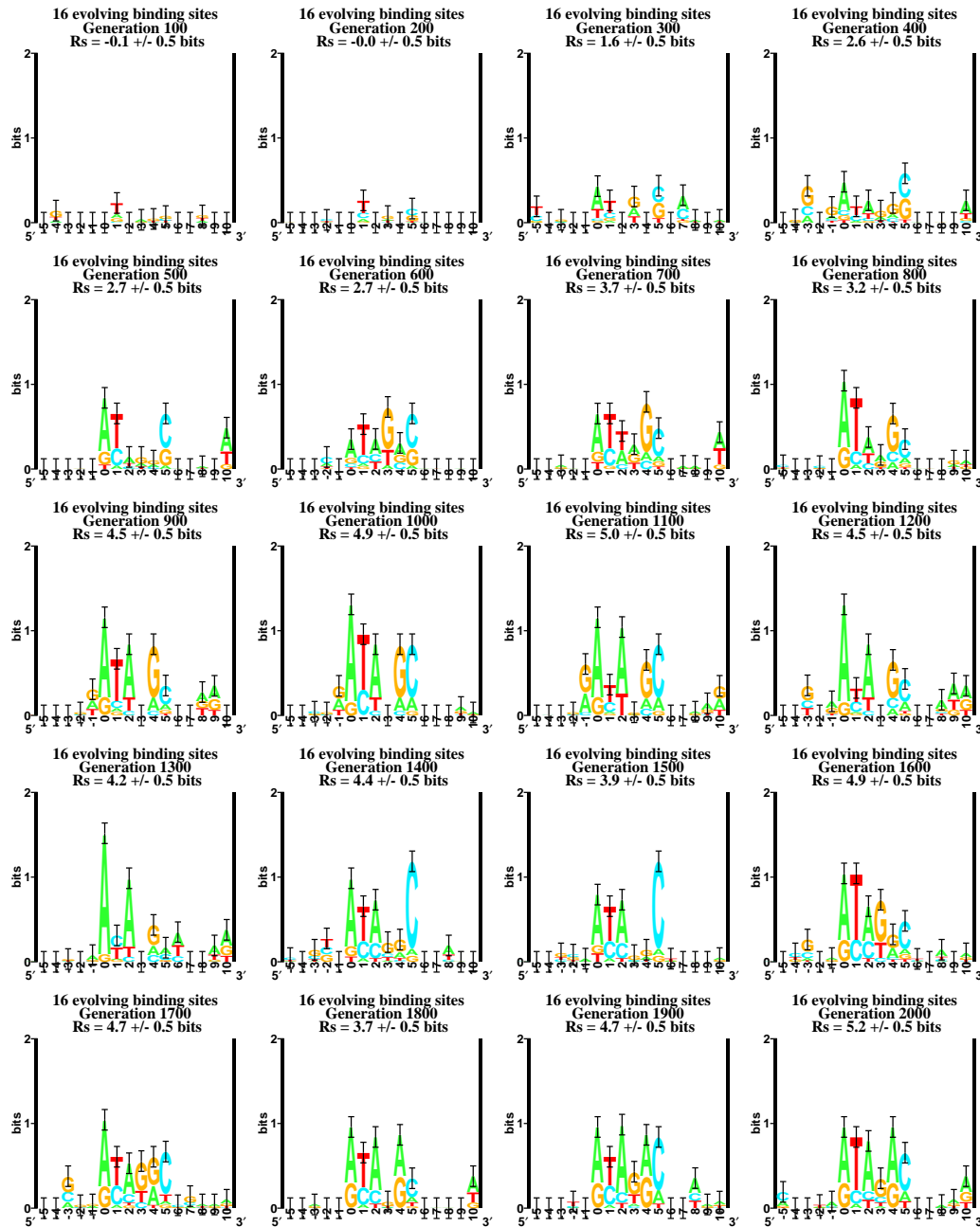
Evolution of Binding Sites



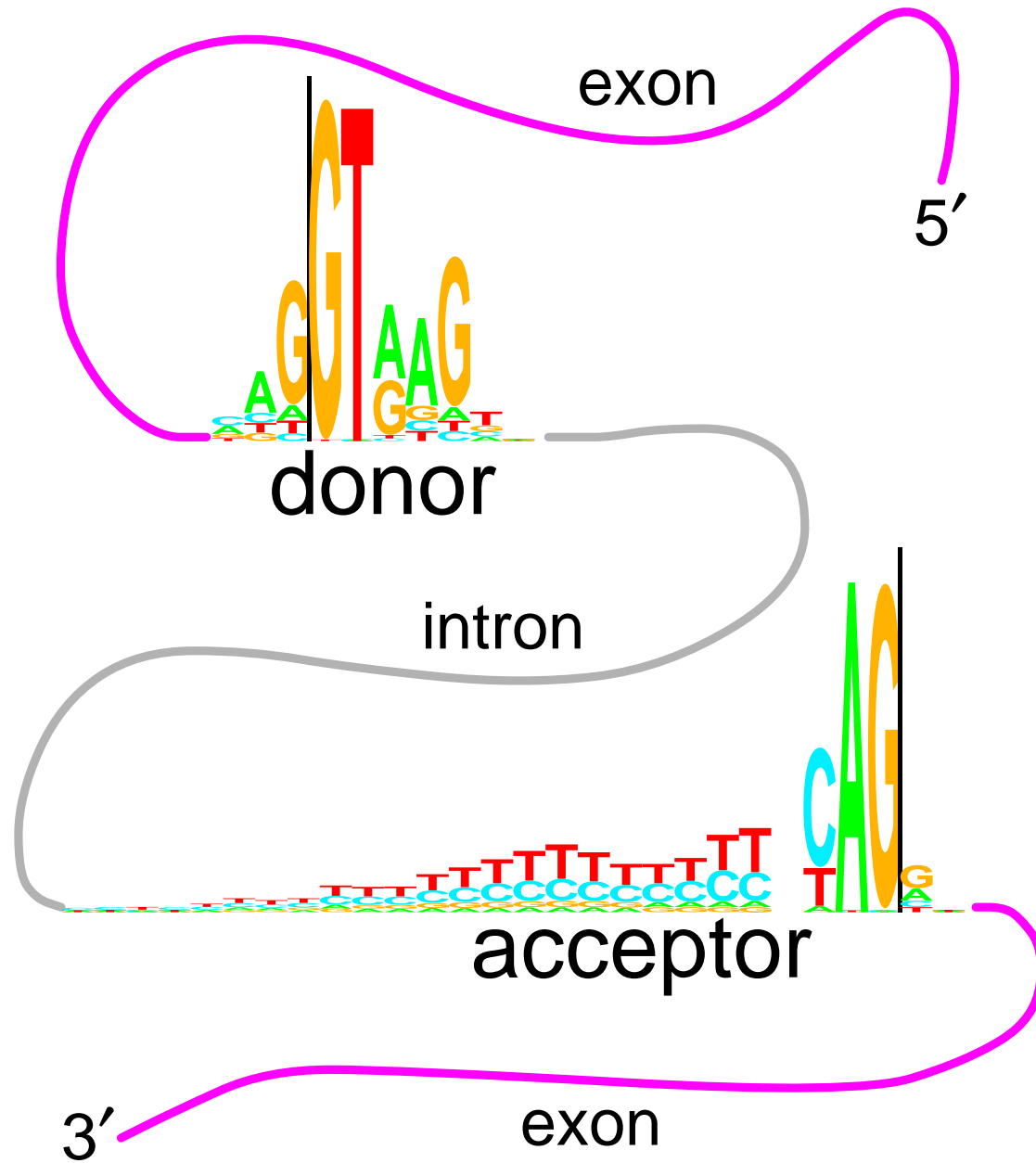
Evolution of Binding Sites



Evolution of Binding Sites

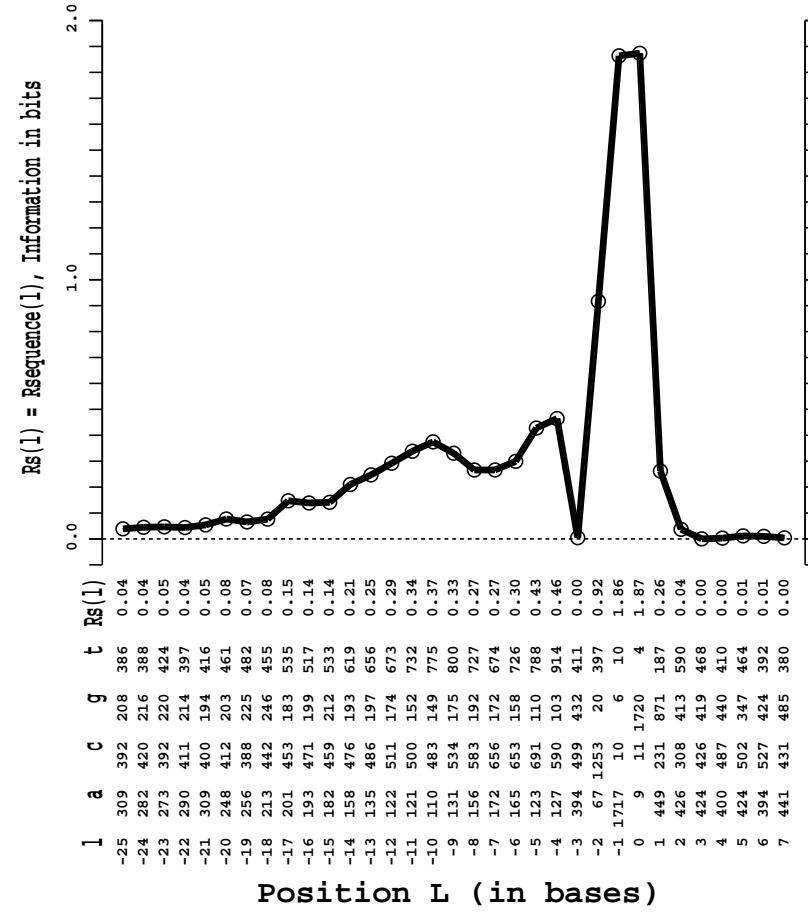
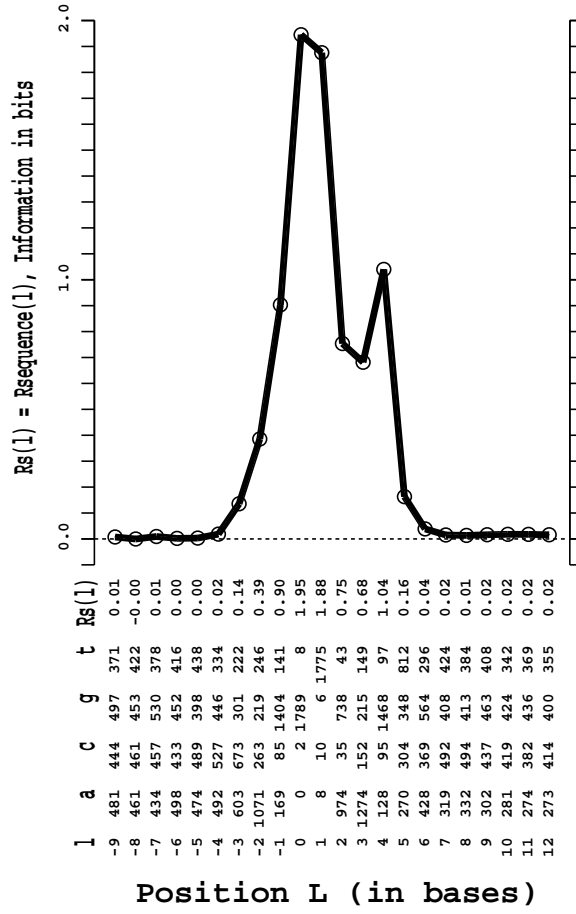


Donor and acceptor logos



Human Splice Junction Information Curves

Sequence Conservation \rightarrow
in bits per base **Donor**

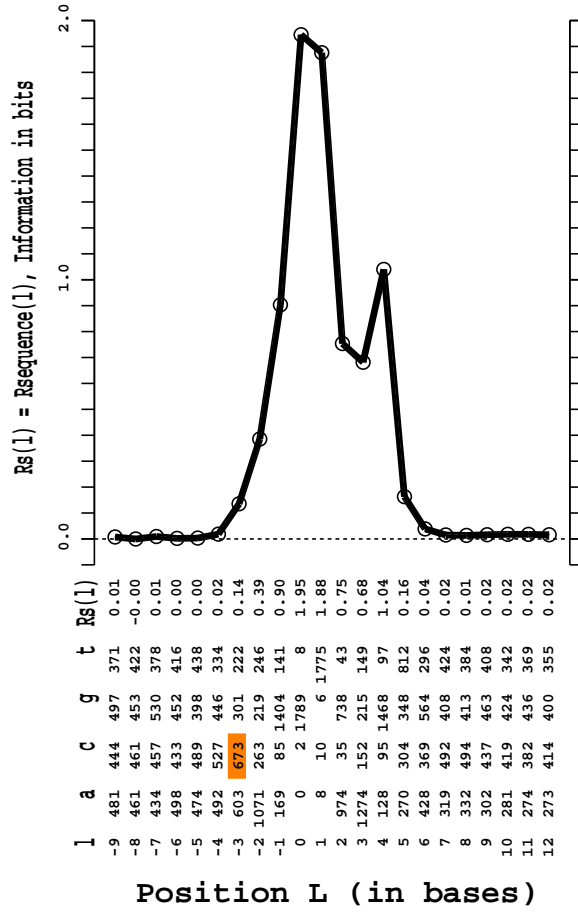


Acceptor

- The consensus sequences match ...

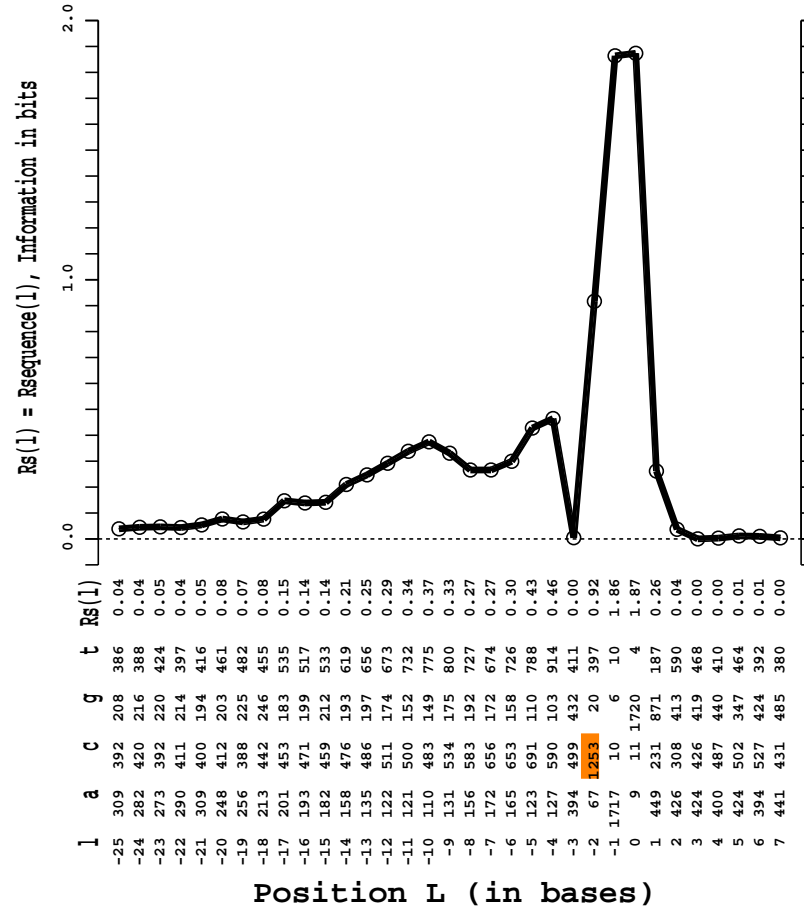
Human Splice Junction Information Curves

Sequence Conservation →
in bits per base **Donor**



C

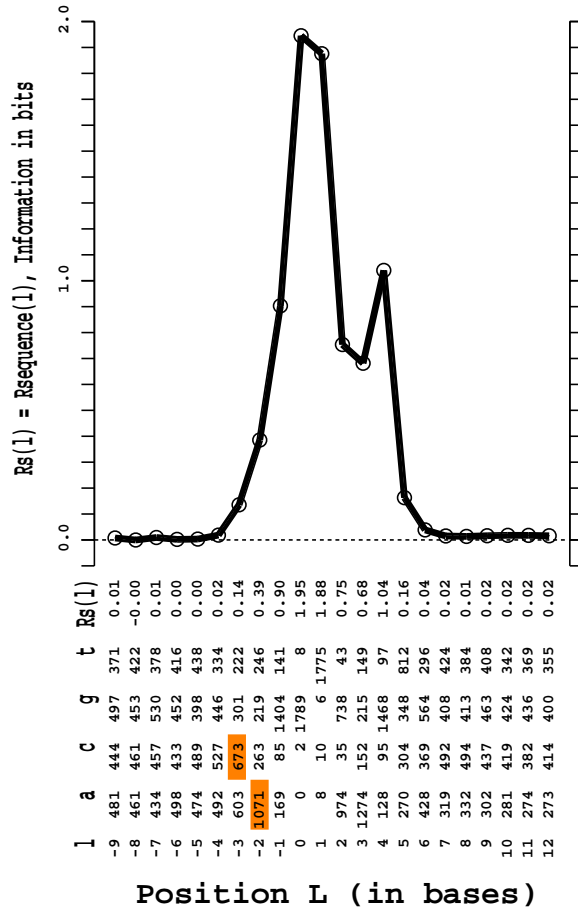
Acceptor



- The consensus sequences match ...

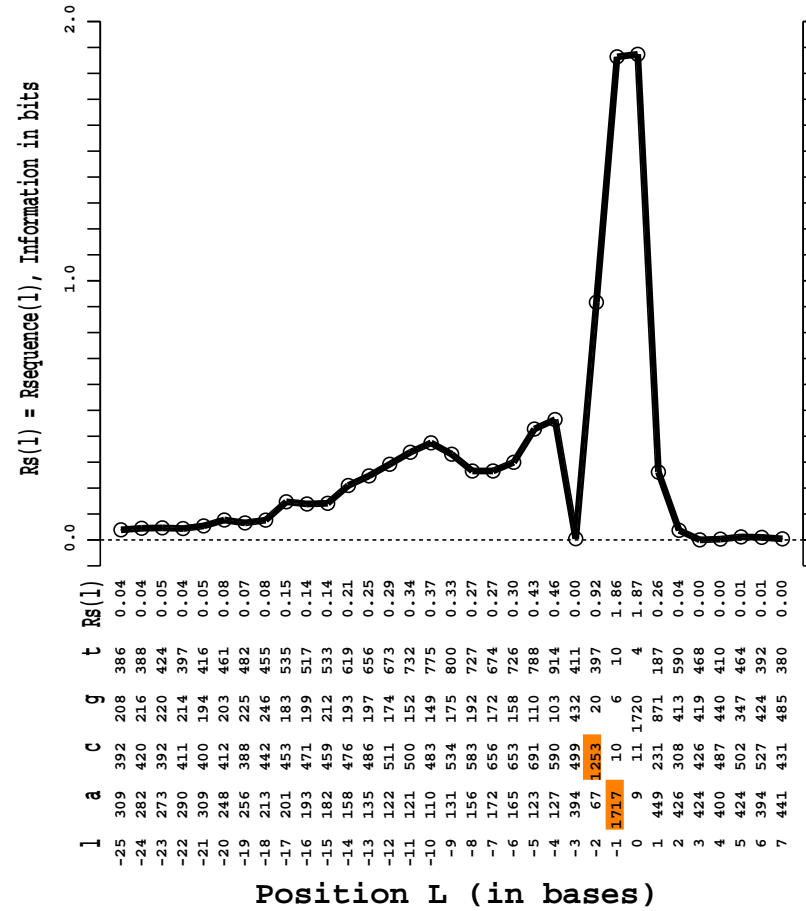
Human Splice Junction Information Curves

Sequence Conservation →
in bits per base **Donor**



C A

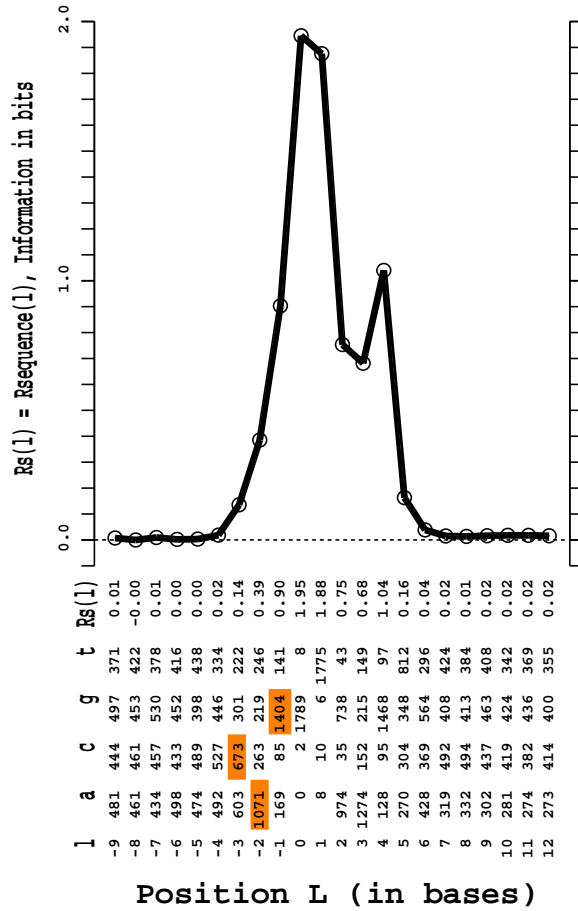
Acceptor



- The consensus sequences match ...

Human Splice Junction Information Curves

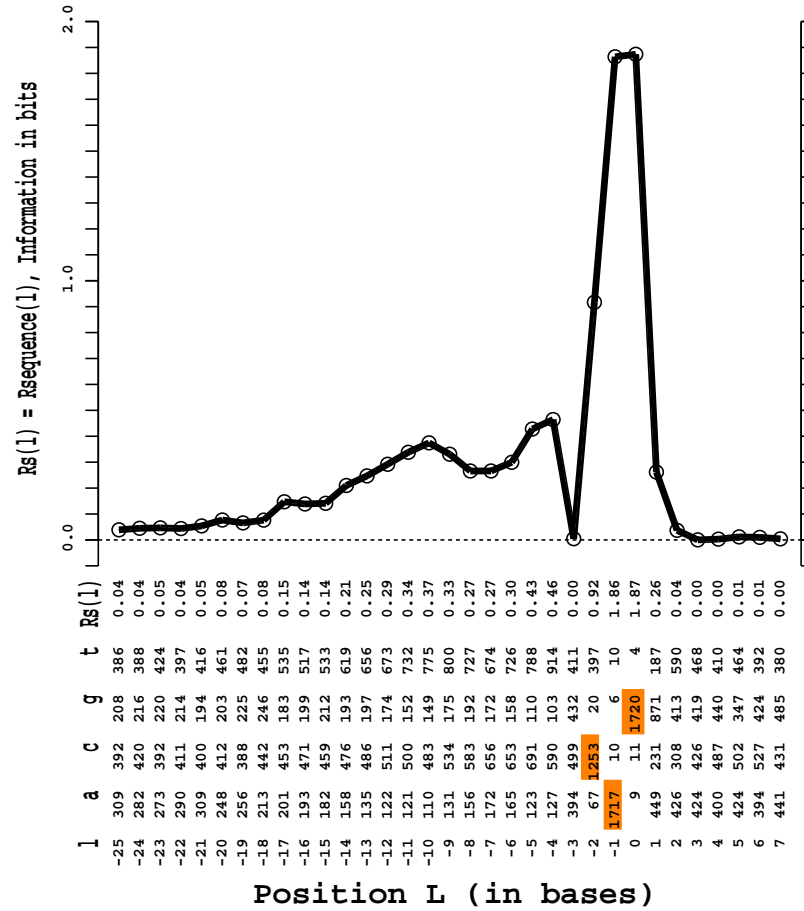
Sequence Conservation →
in bits per base **Donor**



C A G

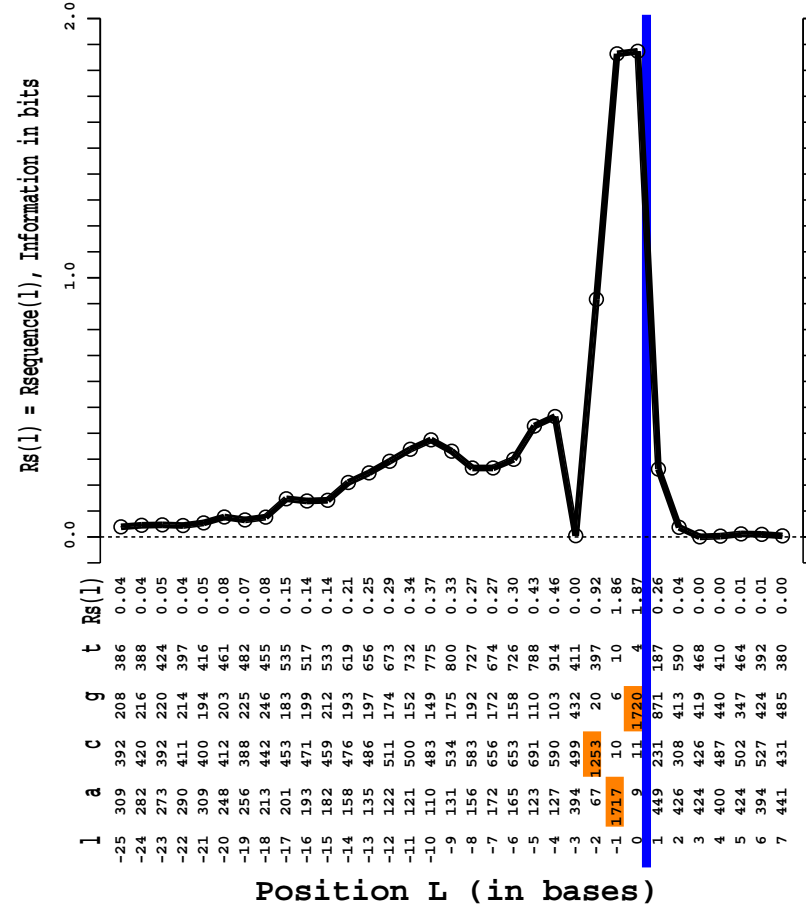
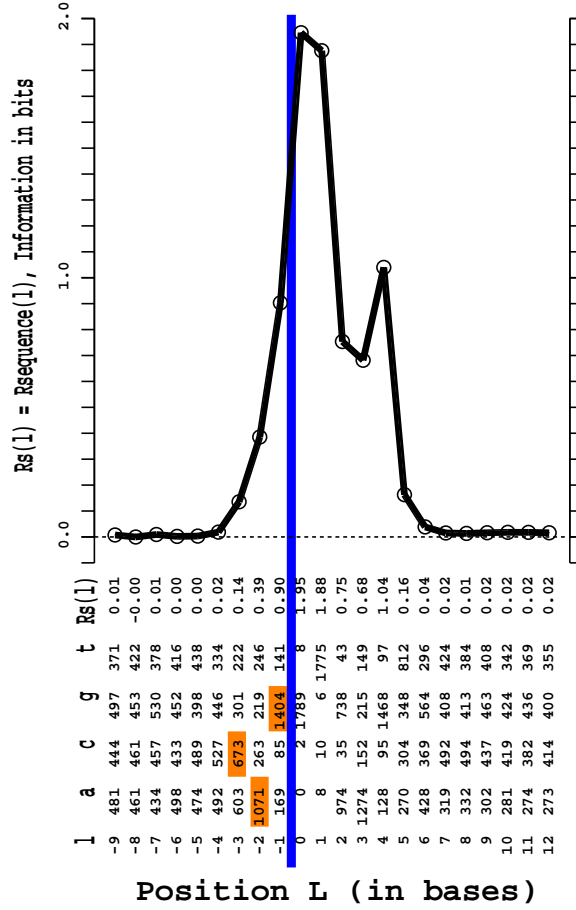
- The consensus sequences match ...

Acceptor



Human Splice Junction Information Curves

Sequence Conservation →
in bits per base **Donor**



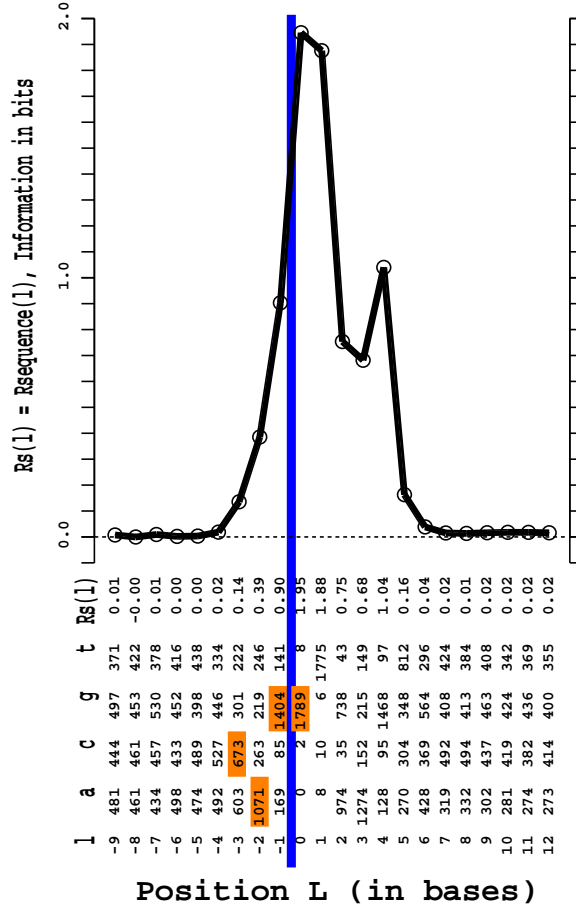
Acceptor

C A G —

- The consensus sequences match ...

Human Splice Junction Information Curves

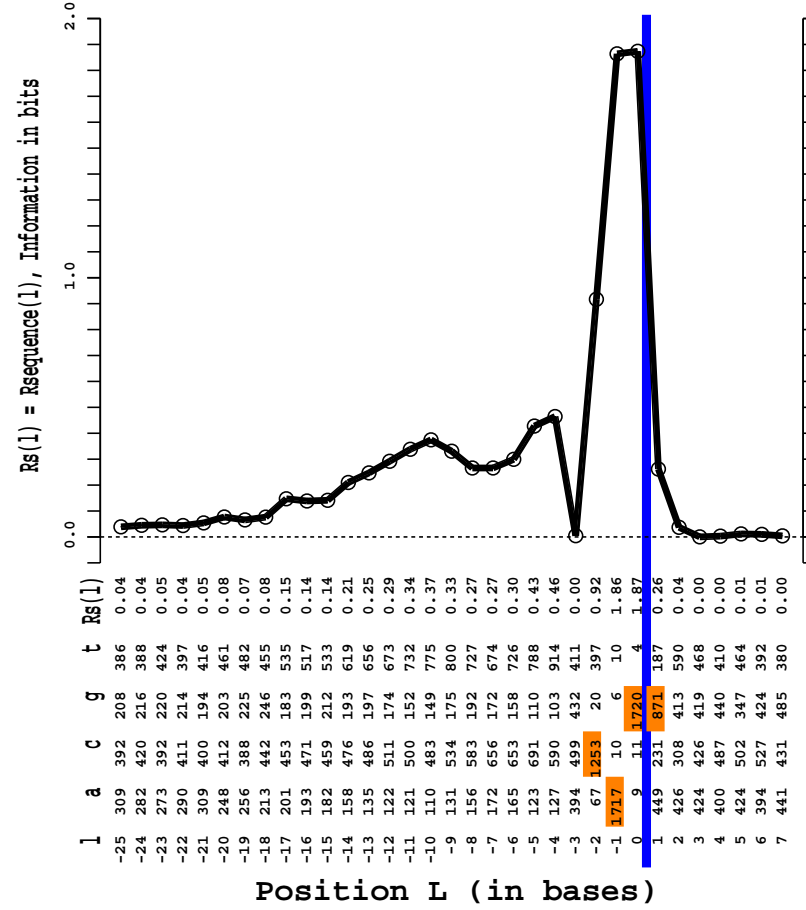
Sequence Conservation →
in bits per base **Donor**



C A G — G

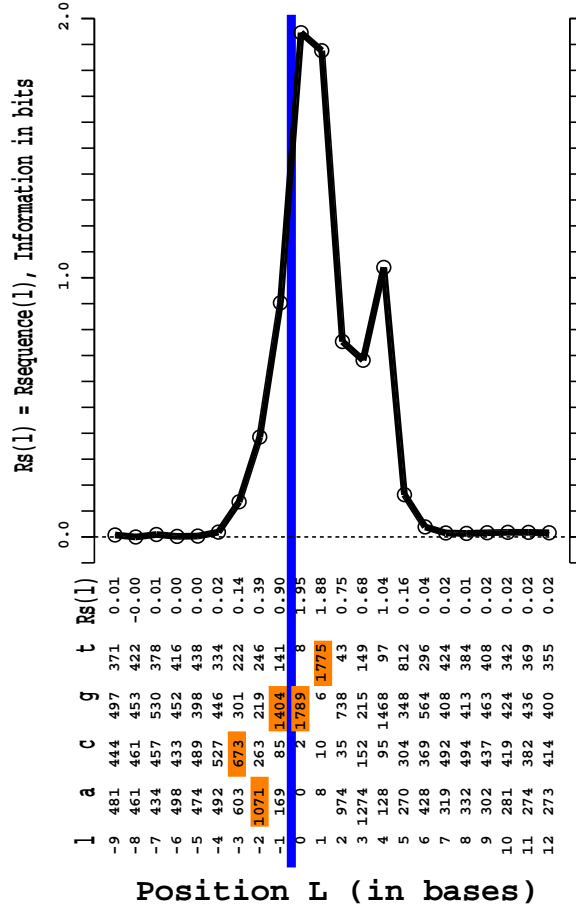
- The consensus sequences match ...

Acceptor



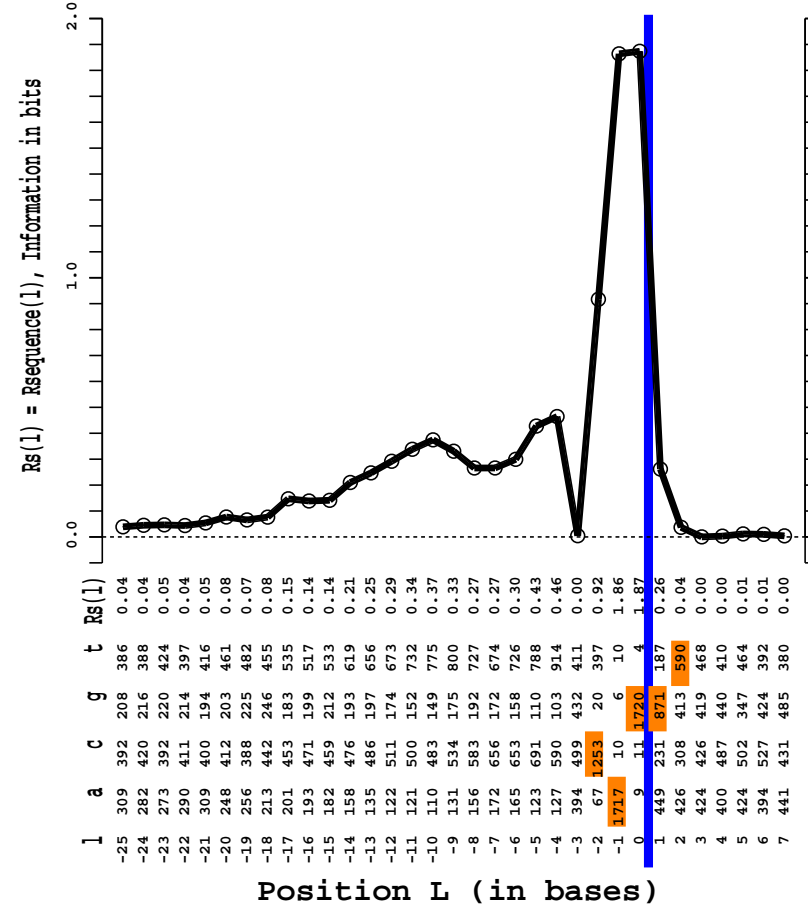
Human Splice Junction Information Curves

Sequence Conservation →
in bits per base **Donor**



C A G — G T

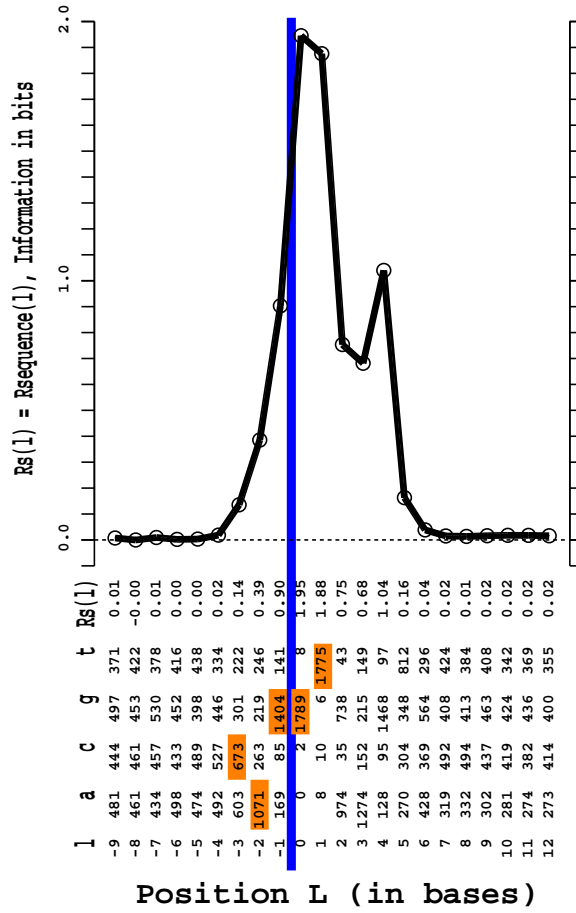
- The consensus sequences match ...



Acceptor

Human Splice Junction Information Curves

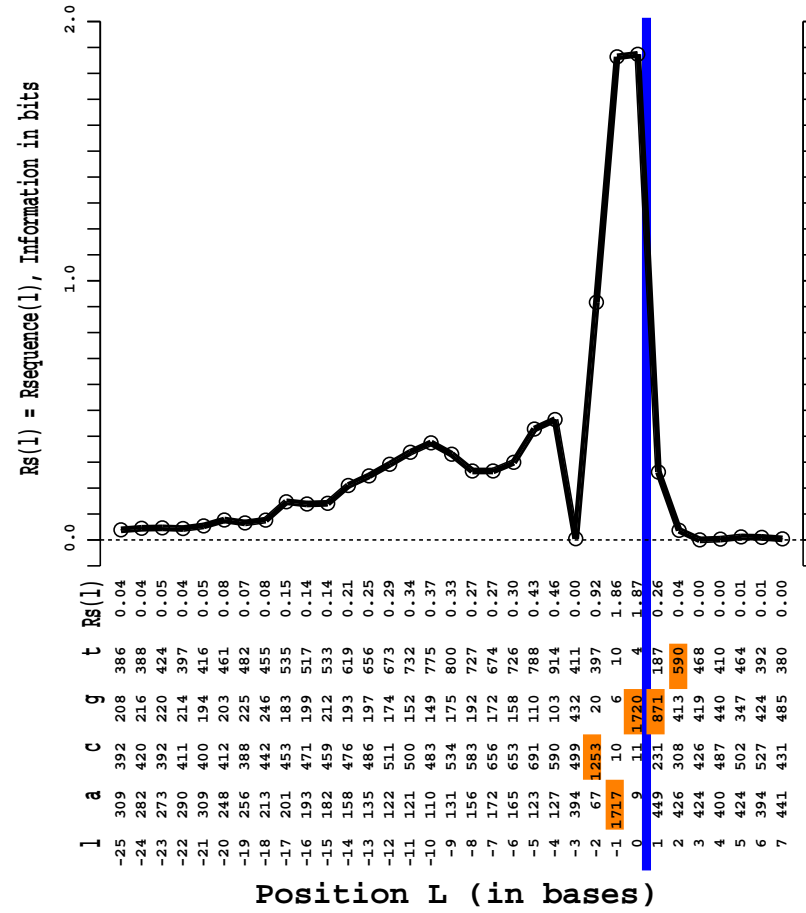
Sequence Conservation →
in bits per base **Donor**



C A G — G T

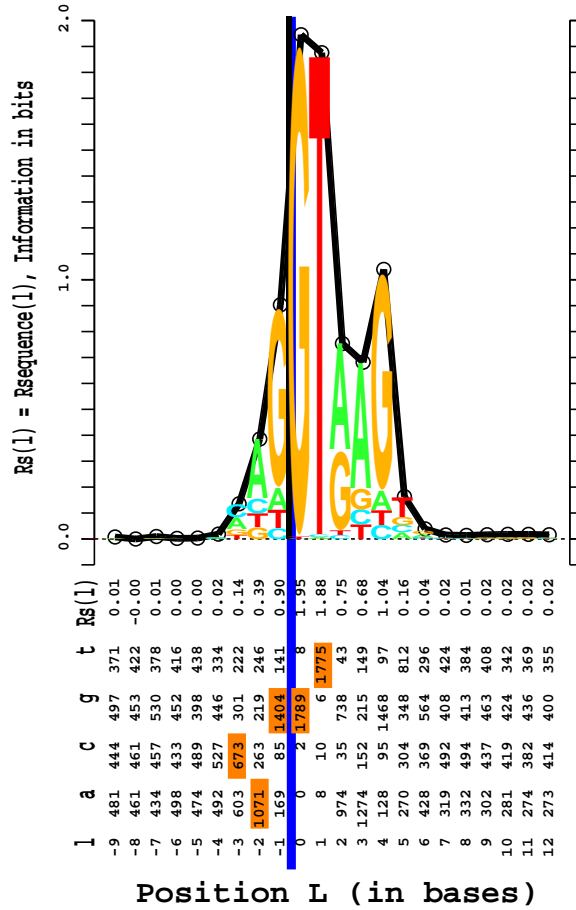
- The consensus sequences match ...
- BUT the information curves (sequence conservation) differ!

Acceptor



Human Splice Junction Information Curves

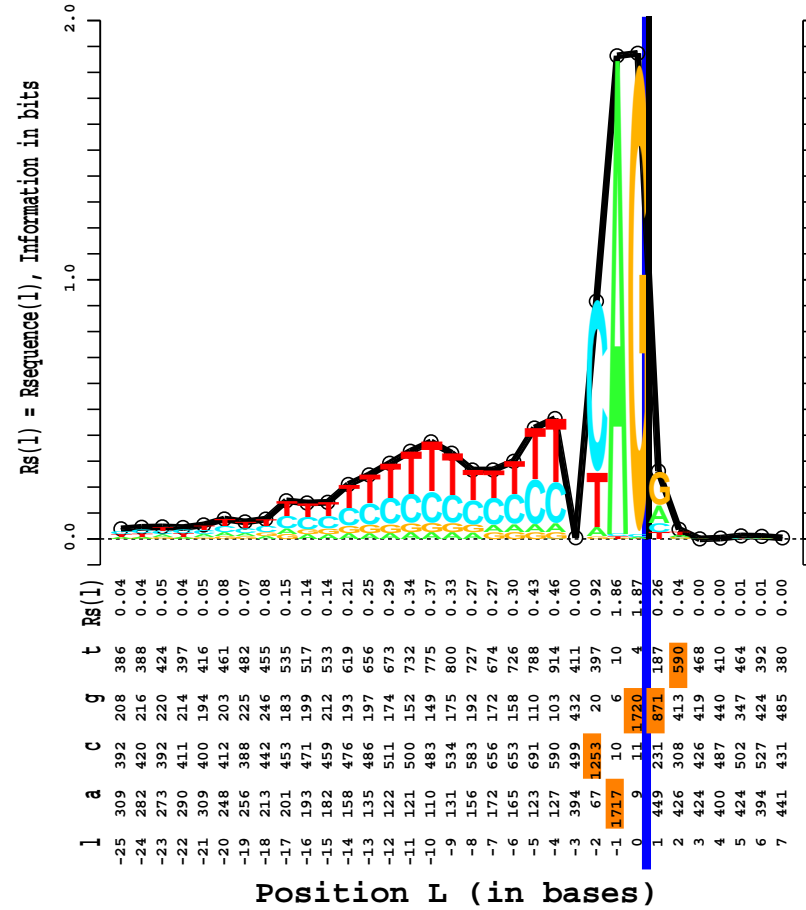
Sequence Conservation →
in bits per base **Donor**



C A G — G T

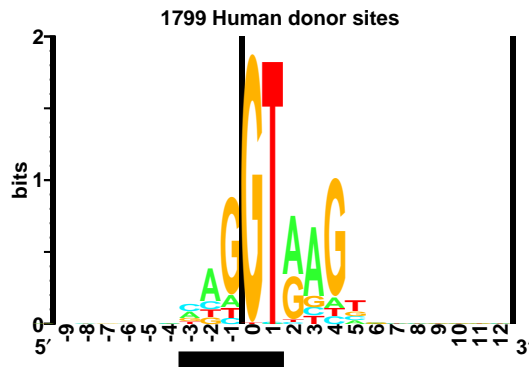
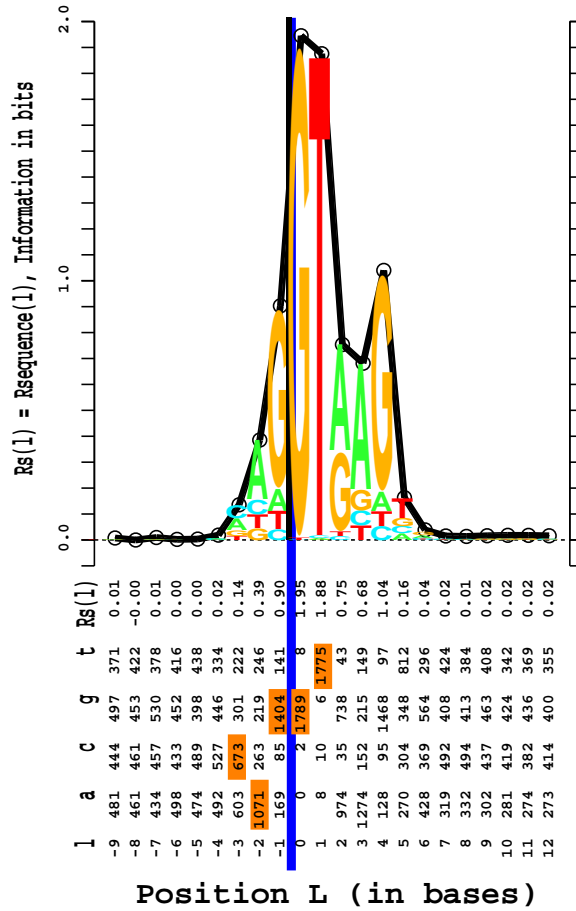
- The consensus sequences match ...
- BUT the information curves (sequence conservation) differ!
- Put letters into the graph proportional to their frequency!

Acceptor

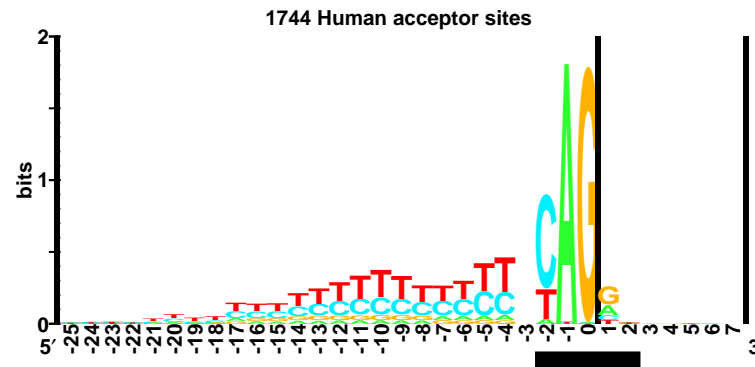
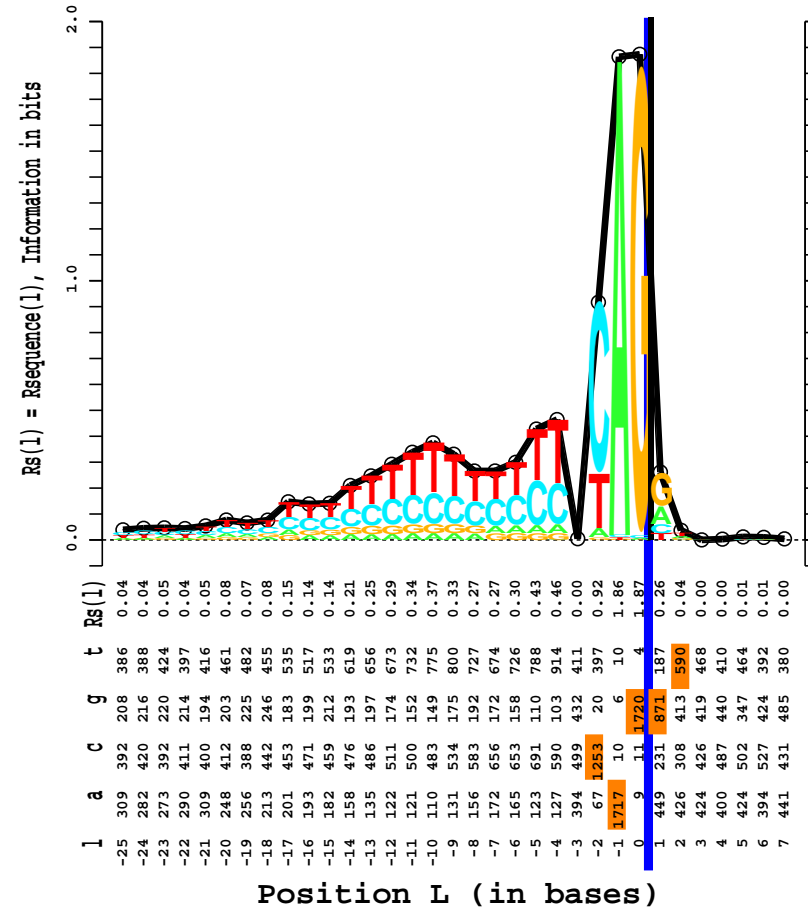


Human Splice Junction Information Curves

Sequence Conservation →
in bits per base **Donor**



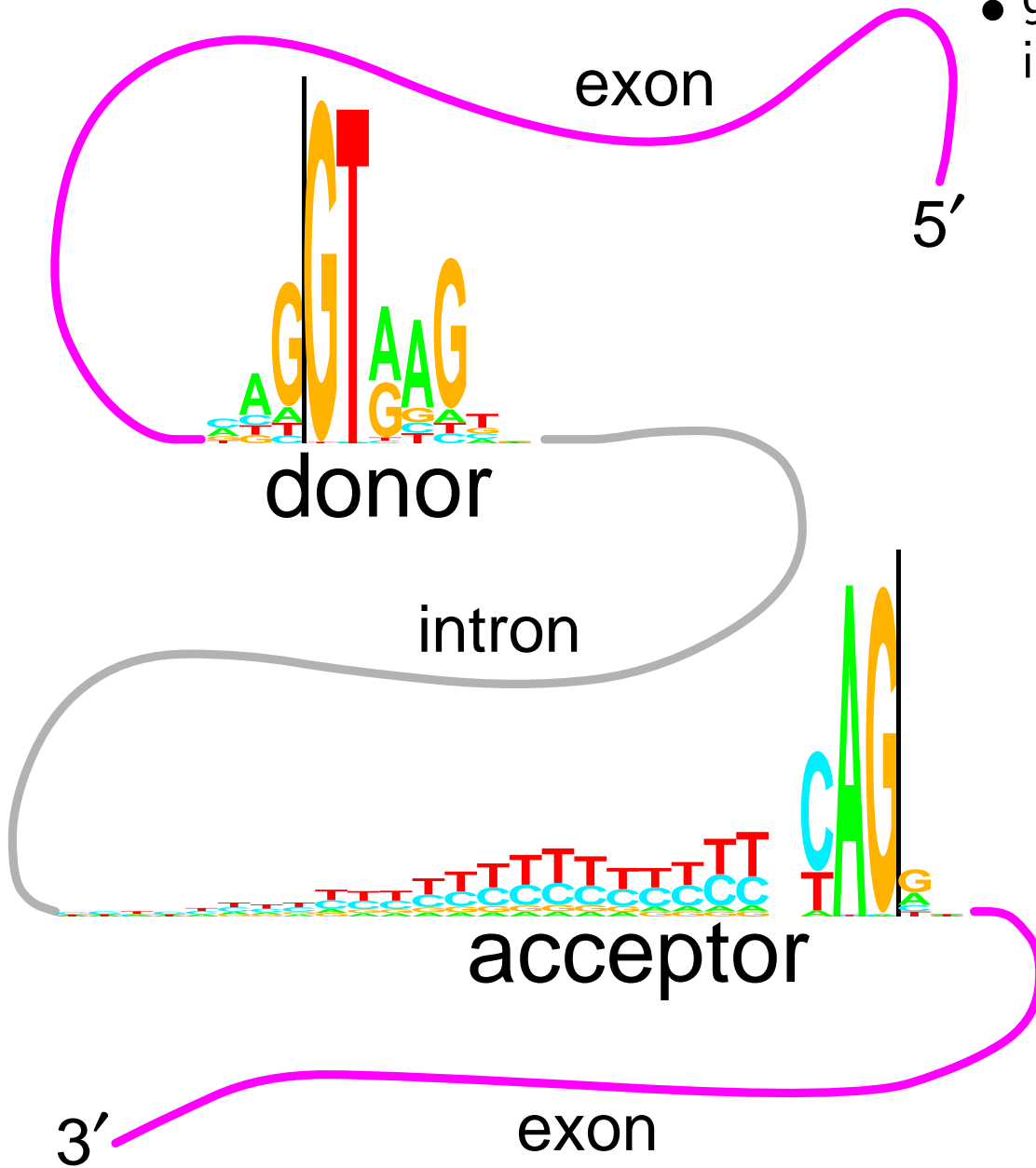
Acceptor



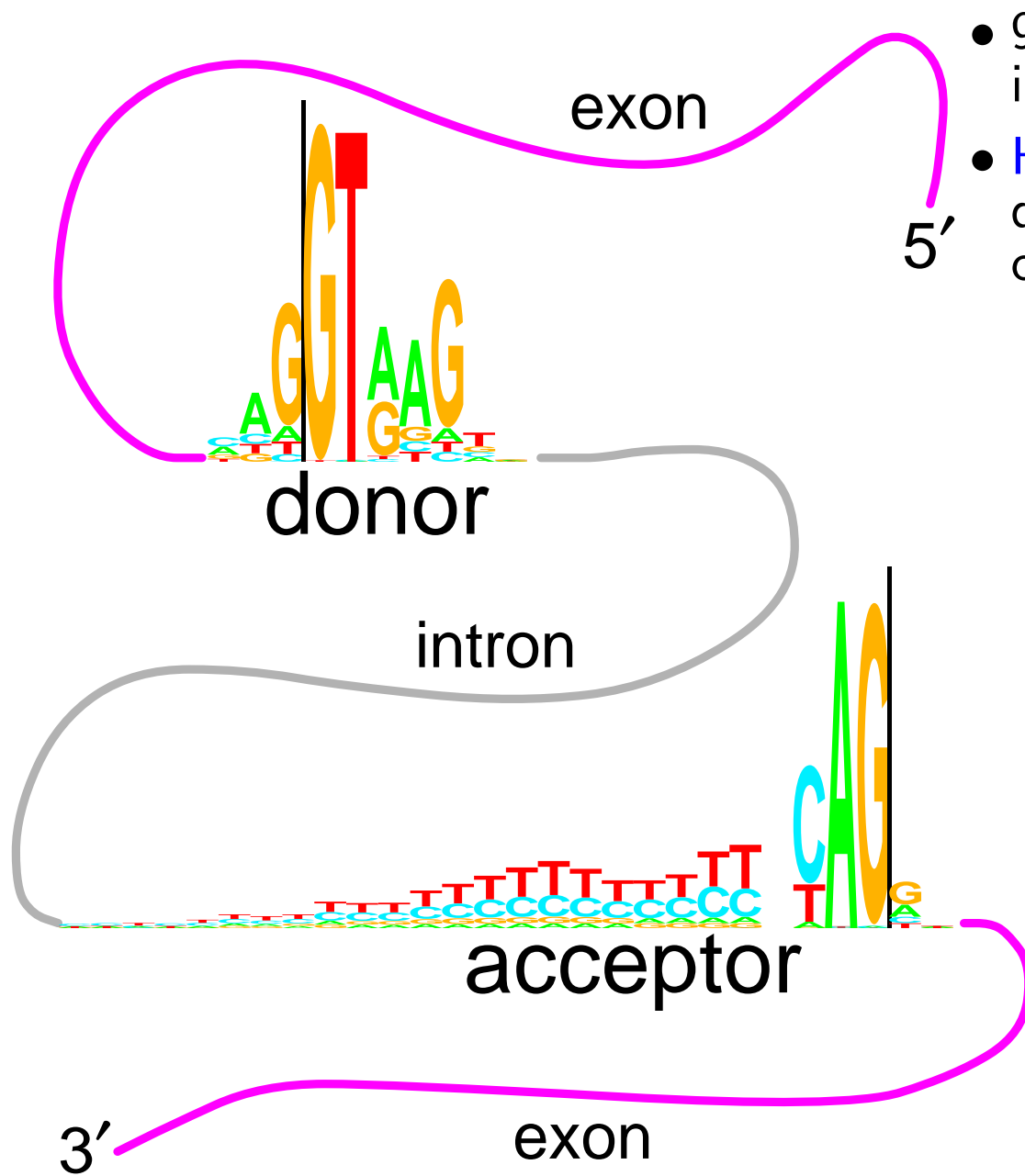
- That's how and why we invented sequence logos!

Splice Junction Sequence Logos

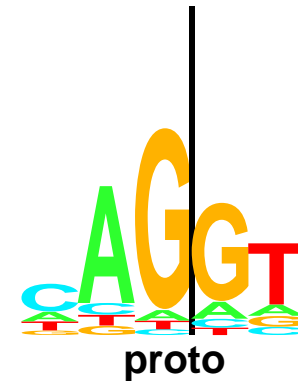
- 90% of the splice junction information is on the intron side



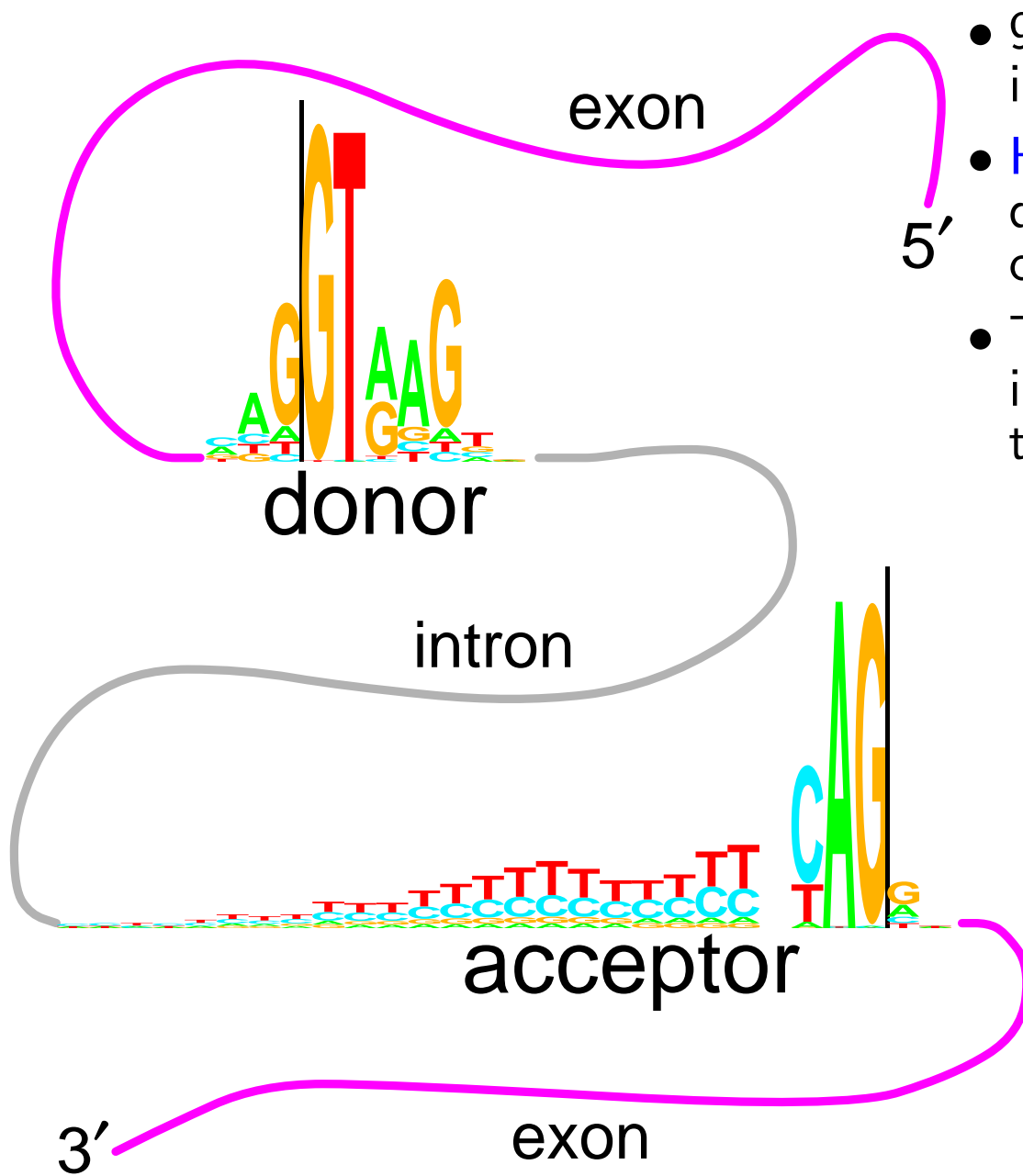
Splice Junction Sequence Logos



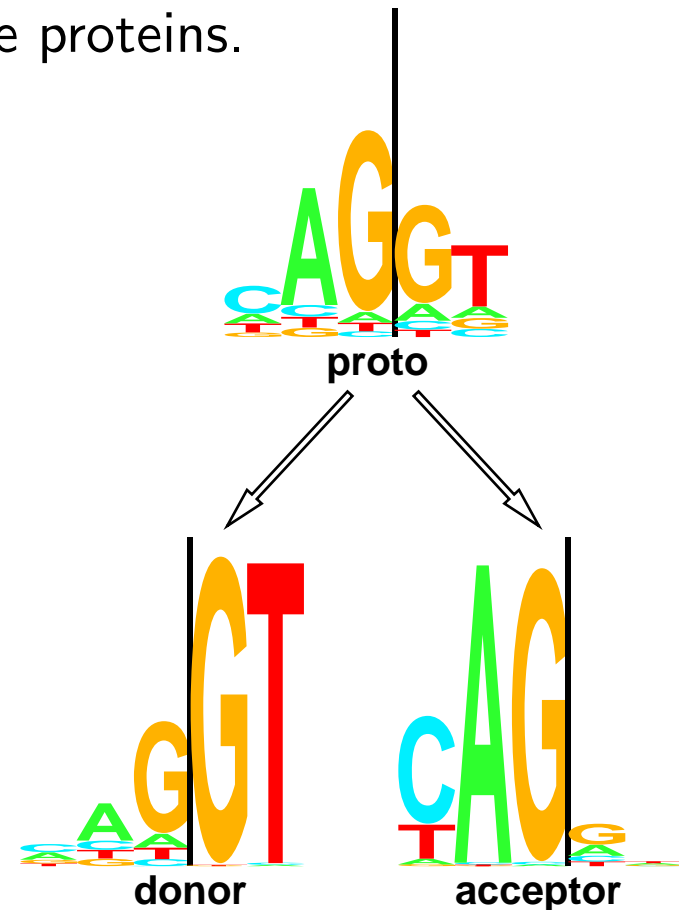
- 90% of the splice junction information is on the intron side
- **Hypothesis:** donor and acceptor sites had a common ancestor that duplicated



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.



- **Before** binding a recognizer is anywhere on the nucleic acid

Information is a Decrease of Uncertainty

- **Before** binding a recognizer is anywhere on the nucleic acid
- **After** binding a recognizer is at its binding sites

Information is a Decrease of Uncertainty

- **Before** binding a recognizer is anywhere on the nucleic acid
- **After** binding a recognizer is at its binding sites
- Information R is a decrease of uncertainty:

$$R = -\Delta H = H_{before} - H_{after}$$

Information is a Decrease of Uncertainty

- **Before** binding a recognizer is anywhere on the nucleic acid
- **After** binding a recognizer is at its binding sites
- Information R is a decrease of uncertainty:

$$R = -\Delta H = H_{before} - H_{after}$$

- This is how both $R_{sequence}$ and $R_{frequency}$ were defined.

- Information at position l as in a sequence logo:

$$\begin{aligned} R_{sequence}(l) &= H_{before} - H_{after}(l) \\ &= 2 - H_{after}(l) \end{aligned}$$

- Information at position l as in a sequence logo:

$$\begin{aligned} R_{sequence}(l) &= H_{before} - H_{after}(l) \\ &= 2 - H_{after}(l) \end{aligned}$$

- Individual Information matrix (difference of surprisals) for base b at position l , based on frequency of bases $f(b, l)$:

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

$e(l)$ = small sample correction.

- Information at position l as in a sequence logo:

$$\begin{aligned}R_{sequence}(l) &= H_{before} - H_{after}(l) \\ &= 2 - H_{after}(l)\end{aligned}$$

- Individual Information matrix (difference of surprisals) for base b at position l , based on frequency of bases $f(b, l)$:

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

$e(l)$ = small sample correction.

- Applied and averaged over a set of sequences $Ri(b, l)$ gives the area under the logo

- Information at position l as in a sequence logo:

$$\begin{aligned}R_{sequence}(l) &= H_{before} - H_{after}(l) \\ &= 2 - H_{after}(l)\end{aligned}$$

- Individual Information matrix (difference of surprisals) for base b at position l , based on frequency of bases $f(b, l)$:

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

$e(l)$ = small sample correction.

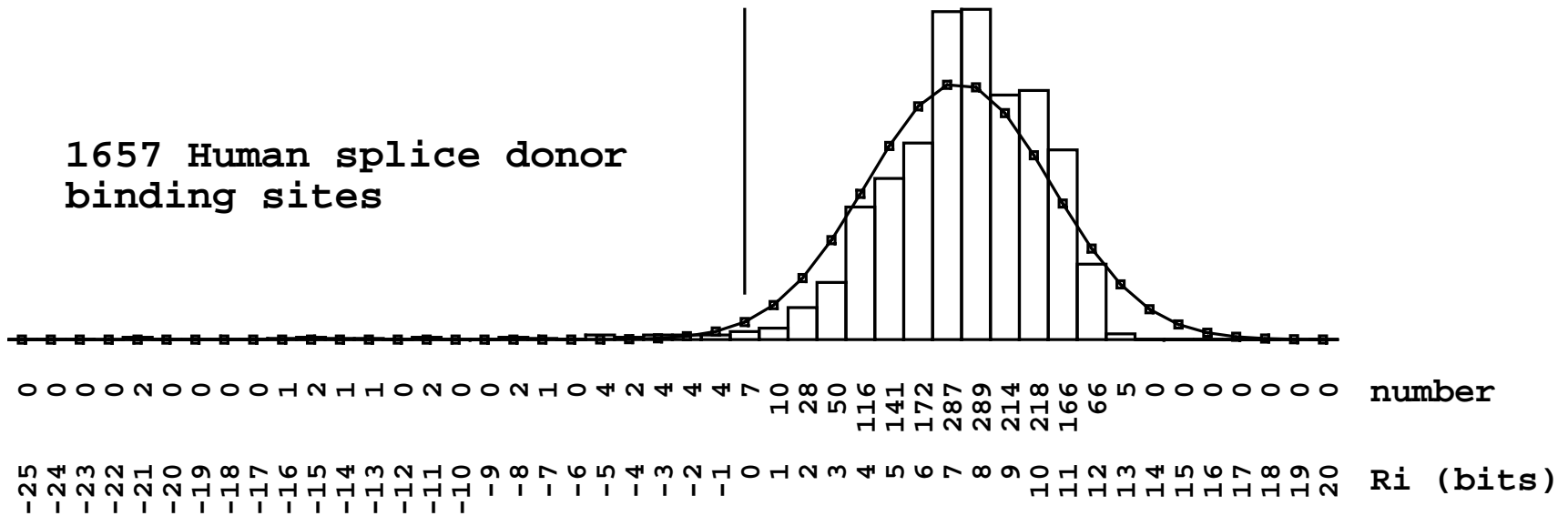
- Applied and averaged over a set of sequences

$Ri(b, l)$ gives the area under the logo

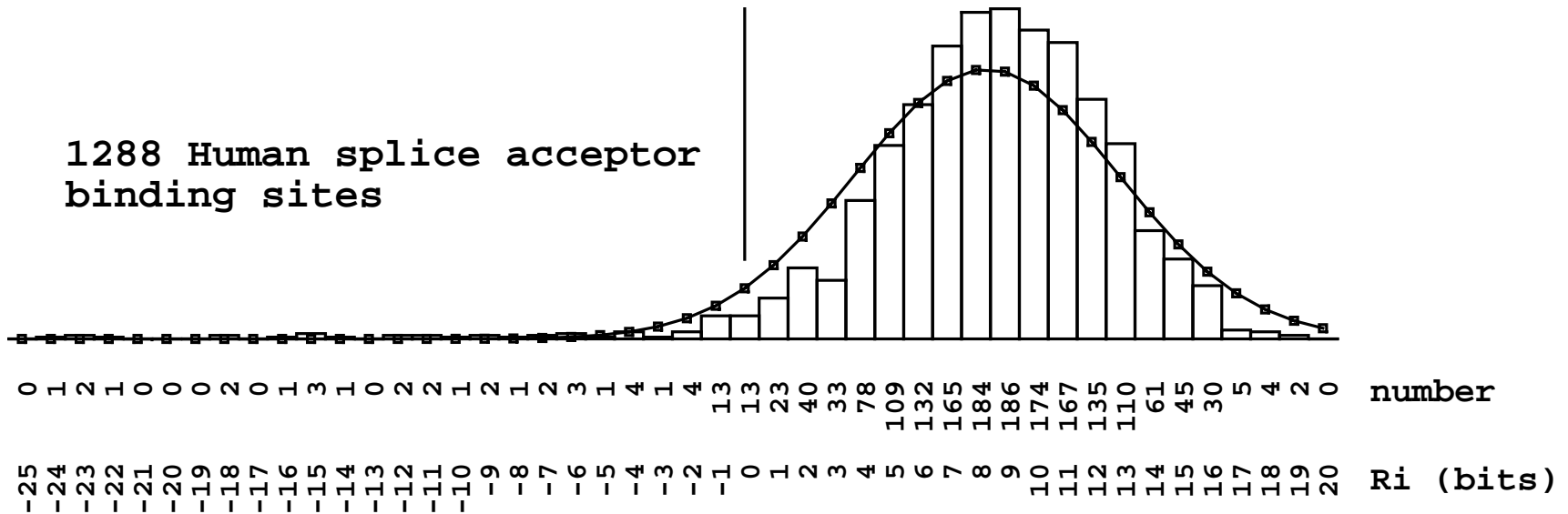
- Proven by John Spouge (NIH, NLM) to be unique

Individual Information Curves

1657 Human splice donor binding sites



1288 Human splice acceptor binding sites



- Uncertainty with probabilities P_i and M states:

$$H \equiv - \sum_{i=1}^M P_i \log_2 P_i \quad (\text{bits per state})$$

Relating Uncertainty H to physical entropy S

- Uncertainty with probabilities P_i and M states:

$$H \equiv - \sum_{i=1}^M P_i \log_2 P_i \quad (\text{bits per state})$$

- Boltzmann-Gibbs entropy for Ω microstates:

$$S \equiv -k_B \sum_{i=1}^{\Omega} P_i \ln P_i \quad \left(\frac{\text{joules}}{\text{K} \cdot \text{microstate}} \right)$$

Relating Uncertainty H to physical entropy S

- Uncertainty with probabilities P_i and M states:

$$H \equiv - \sum_{i=1}^M P_i \log_2 P_i \quad (\text{bits per state})$$

- Boltzmann-Gibbs entropy for Ω microstates:

$$S \equiv -k_B \sum_{i=1}^{\Omega} P_i \ln P_i \quad \left(\frac{\text{joules}}{\text{K} \cdot \text{microstate}} \right)$$

- H relates to S when symbols $M =$ microstates Ω :

$$S = k_B \ln(2) H$$

- Thermal noise hitting a molecule is sound in all directions, at all frequencies (up to a cutoff)

- Thermal noise hitting a molecule is sound in all directions, at all frequencies (up to a cutoff)
- The speed of sound is 1.5 nm per picosecond in sea water at 25°C

- Thermal noise hitting a molecule is sound in all directions, at all frequencies (up to a cutoff)
- The speed of sound is 1.5 nm per picosecond in sea water at 25°C
- Sound crosses the 2 nm diameter of DNA in 1.3 picoseconds

- Thermal noise hitting a molecule is sound in all directions, at all frequencies (up to a cutoff)
- The speed of sound is 1.5 nm per picosecond in sea water at 25°C
- Sound crosses the 2 nm diameter of DNA in 1.3 picoseconds
- So heat leaves a binding DNA or RNA recognizer in picoseconds!

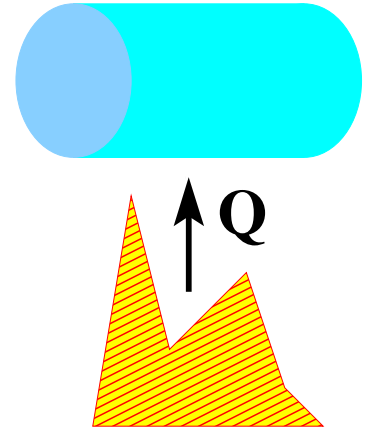
- Thermal noise hitting a molecule is sound in all directions, at all frequencies (up to a cutoff)
- The speed of sound is 1.5 nm per picosecond in sea water at 25°C
- Sound crosses the 2 nm diameter of DNA in 1.3 picoseconds
- So heat leaves a binding DNA or RNA recognizer in picoseconds!
- Equilibration is so fast that the **Before** and **After** states are at the same temperature

- Thermal noise hitting a molecule is sound in all directions, at all frequencies (up to a cutoff)
- The speed of sound is 1.5 nm per picosecond in sea water at 25°C
- Sound crosses the 2 nm diameter of DNA in 1.3 picoseconds
- So heat leaves a binding DNA or RNA recognizer in picoseconds!
- Equilibration is so fast that the **Before** and **After** states are at the same temperature
- T is constant, binding is isothermal

Second Law of Thermodynamics

- Clausius inequality for heat Q into the system:

$$dS \geq \frac{dQ}{T}$$



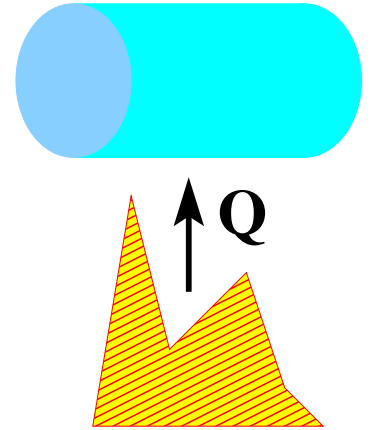
Second Law of Thermodynamics

- Clausius inequality for heat Q into the system:

$$dS \geq \frac{dQ}{T}$$

- Integrate with T constant for isothermal biological processes:

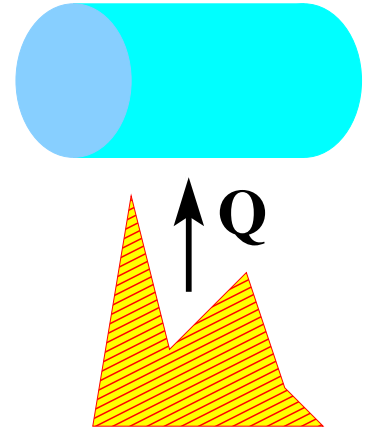
$$\Delta S \geq \frac{Q}{T}$$



Second Law of Thermodynamics

- Clausius inequality for heat Q into the system:

$$dS \geq \frac{dQ}{T}$$



- Integrate with T constant for isothermal biological processes:

$$\Delta S \geq \frac{Q}{T}$$

- Substitute for ΔS to ΔH and then to R :

$$k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

Minimum Energy Dissipated per Bit

- Minimum energy dissipated to get a bit:

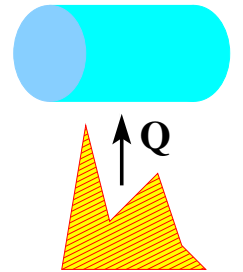
$$E_{min} \equiv k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

Minimum Energy Dissipated per Bit

- Minimum energy dissipated to get a bit:

$$E_{min} \equiv k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

- Q is heat going into the protein-DNA system

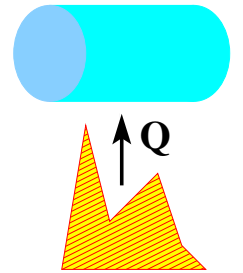


Minimum Energy Dissipated per Bit

- Minimum energy dissipated to get a bit:

$$E_{min} \equiv k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

- Q is heat going into the protein-DNA system


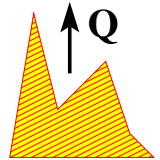


- If $R > 0$ then $Q < 0$ heat goes OUT = **BINDING**

Minimum Energy Dissipated per Bit

- Minimum energy dissipated to get a bit:


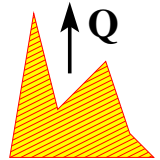
$$E_{min} \equiv k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

- Q is heat going into the protein-DNA system 
- If $R > 0$ then $Q < 0$ heat goes OUT = BINDING 
- If $R < 0$ then $Q > 0$ heat goes IN = UNBINDING

Minimum Energy Dissipated per Bit

- Minimum energy dissipated to get a bit:


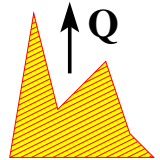
$$E_{min} \equiv k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

- Q is heat going into the protein-DNA system 
- If $R > 0$ then $Q < 0$ heat goes OUT = BINDING 
- If $R < 0$ then $Q > 0$ heat goes IN = UNBINDING
- When $R < 0$ one would have to force the recognizer to bind

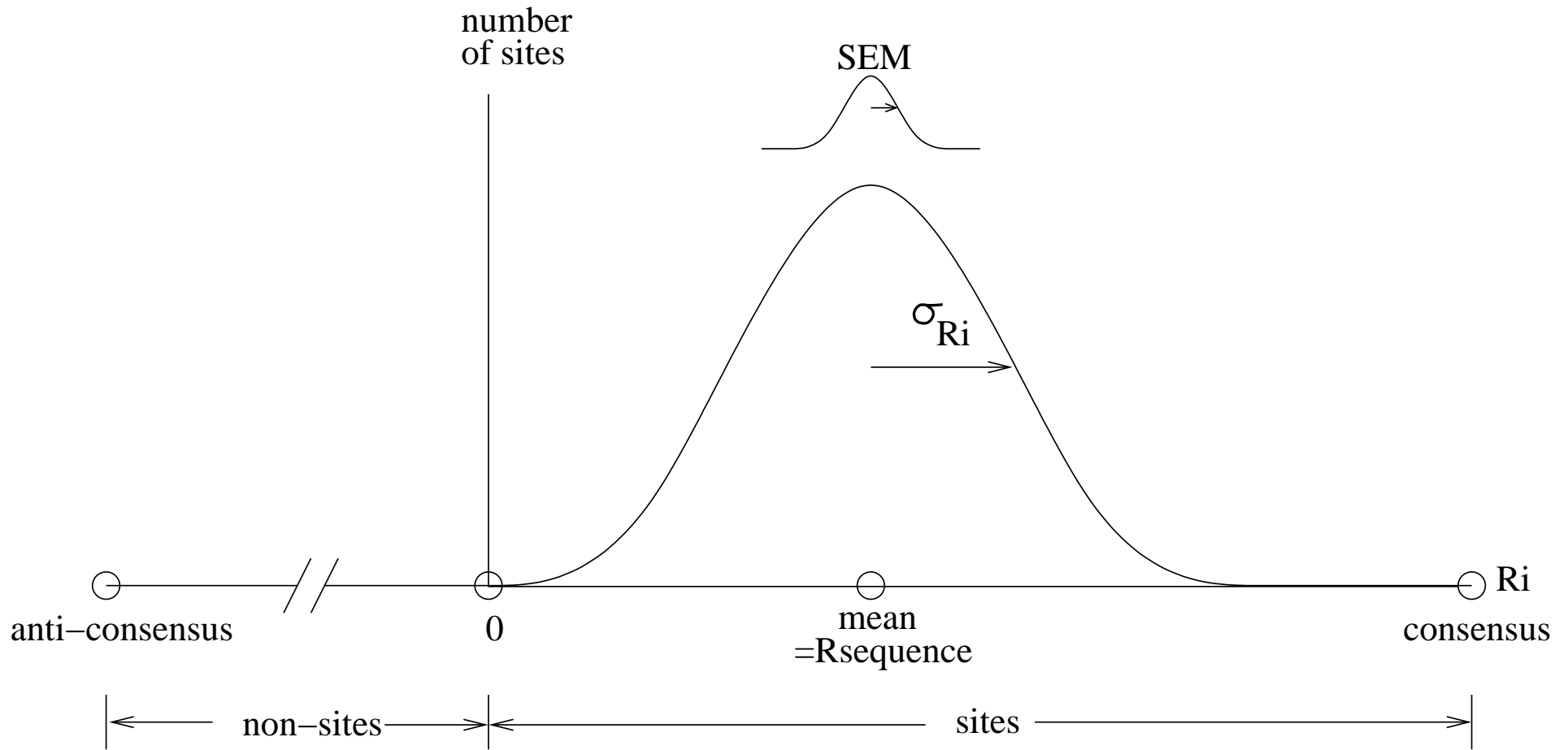
Minimum Energy Dissipated per Bit

- Minimum energy dissipated to get a bit:

$$E_{min} \equiv k_B T \ln(2) \leq \frac{-Q}{R} \quad (\text{joules per bit})$$

- Q is heat going into the protein-DNA system 
- If $R > 0$ then $Q < 0$ heat goes OUT = BINDING 
- If $R < 0$ then $Q > 0$ heat goes IN = UNBINDING
- When $R < 0$ one would have to force the recognizer to bind
- Sequences with $R_i < 0$ are not binding sites.

Individual Information Density Curve



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

How A Weight Matrix Works

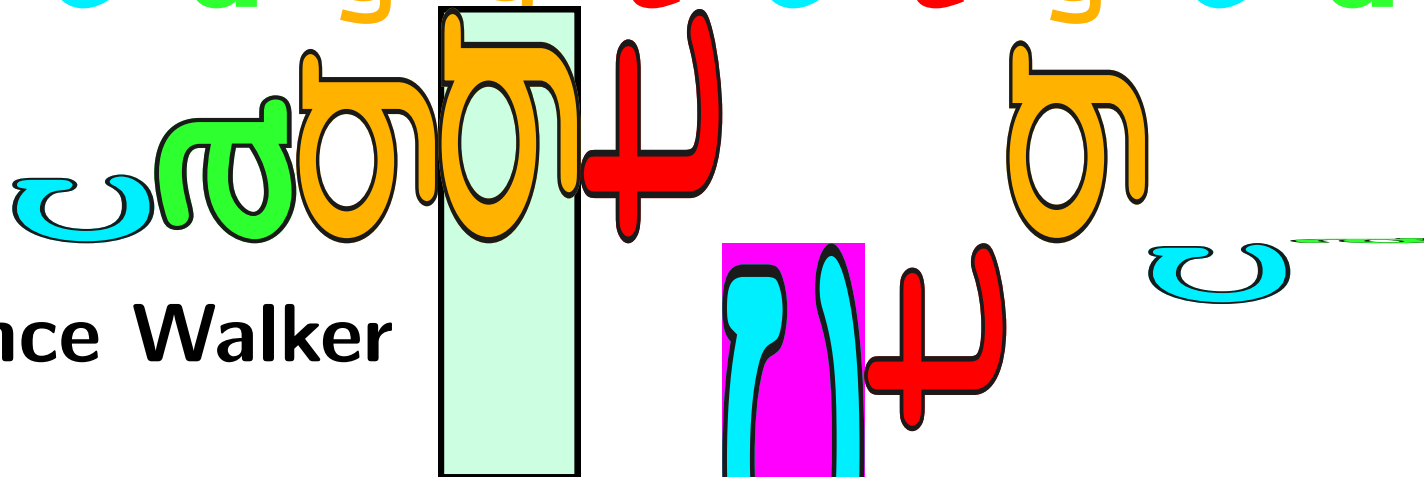
Sequence matrix, $s(b, l, j)$ for sequence j

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'



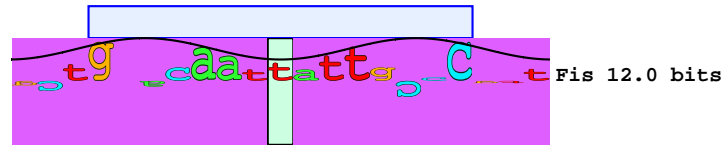
Sequence Walker

Sequence Walker example: *rrnB* P1

***rrnB* P1**

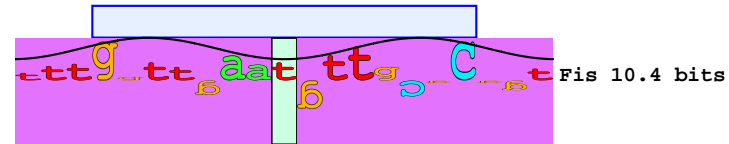
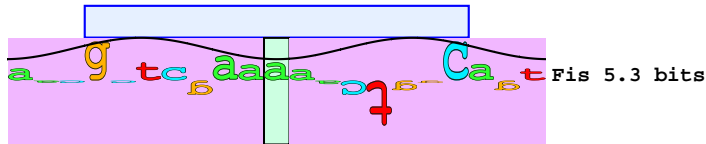
5' g g a g c t g a a c a a t t a t t g c c c g g t t t t a c a g c g t t a c g g c t t c g a 3'

3' c c t c g a c t t g t a a t a a c g g g c a a a a t g t c g c a a t g c c g a a g c t 5'



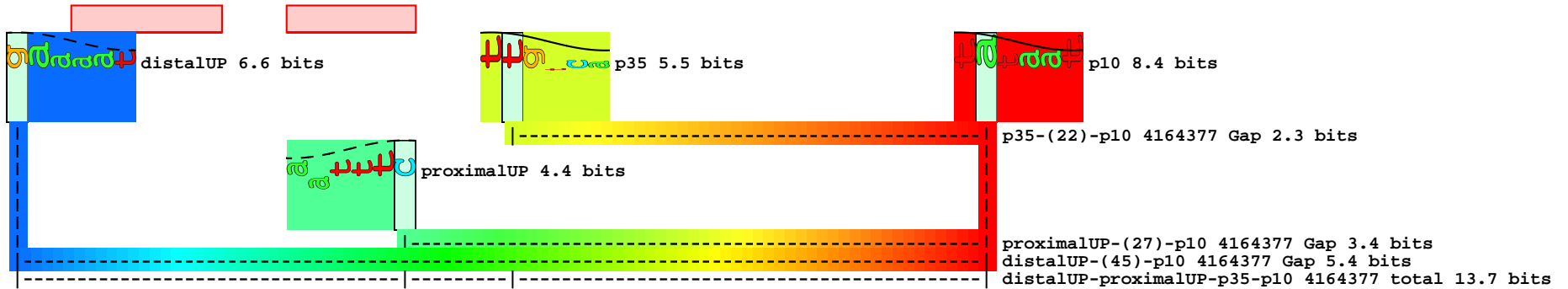
5' a a c g c t c g a a a a a c t g g c a g t t t t t a g g c t g a t t t g g t t g a a t g t t g c g c g g t c a 3'

3' t t g c g a g c t t t t g a c c g t c a a a a t c c g a c t t a c a a c g c g c a g t 5'



5' g a a a a t t a t t t t t t t t t t c c t t g t c a g g c c g g a a t a a c t c c t a t a a t g 3'

3' c t t t t a a t a a a a t t t a a g g a g a a c a g t c c g g c c t t a t t g a g g g a t a t t a c 5'



Complex Sequence Walker Example

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

Complex Sequence Walker Example

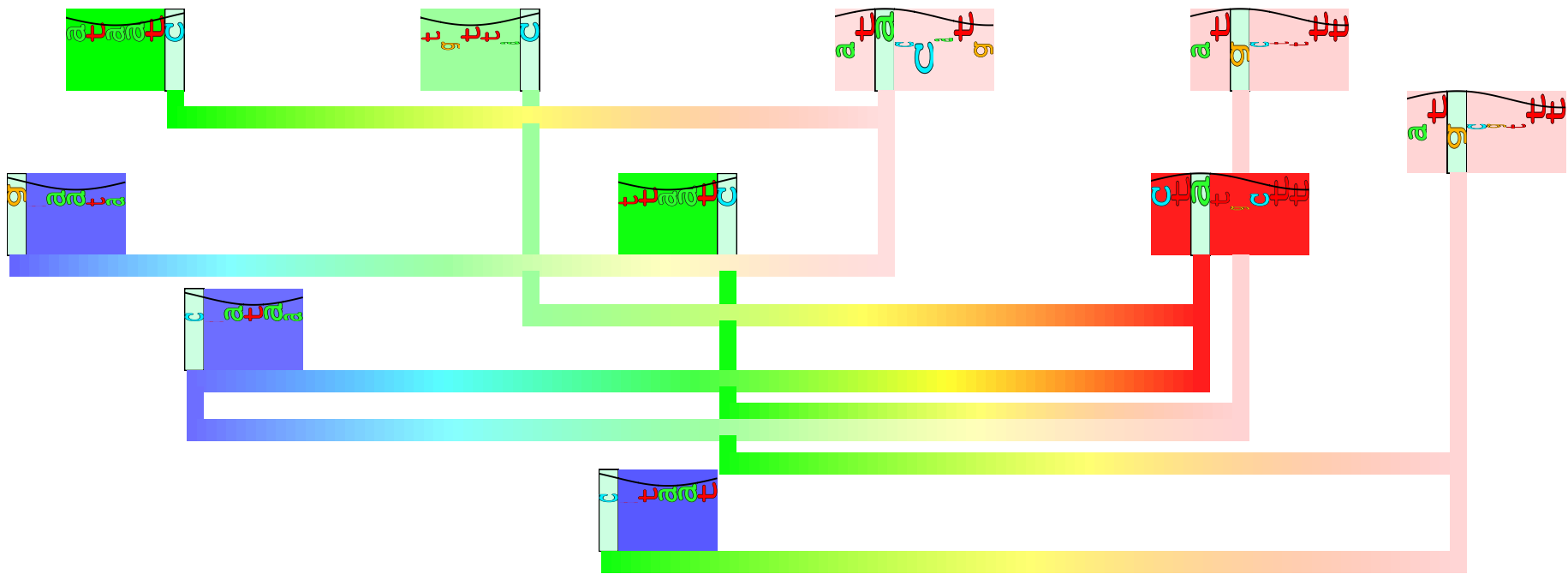
- σ^{70} promoters have a -35 and a -10
- Using information theory we discovered that stress-response σ^{38} promoters do not have a -35

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

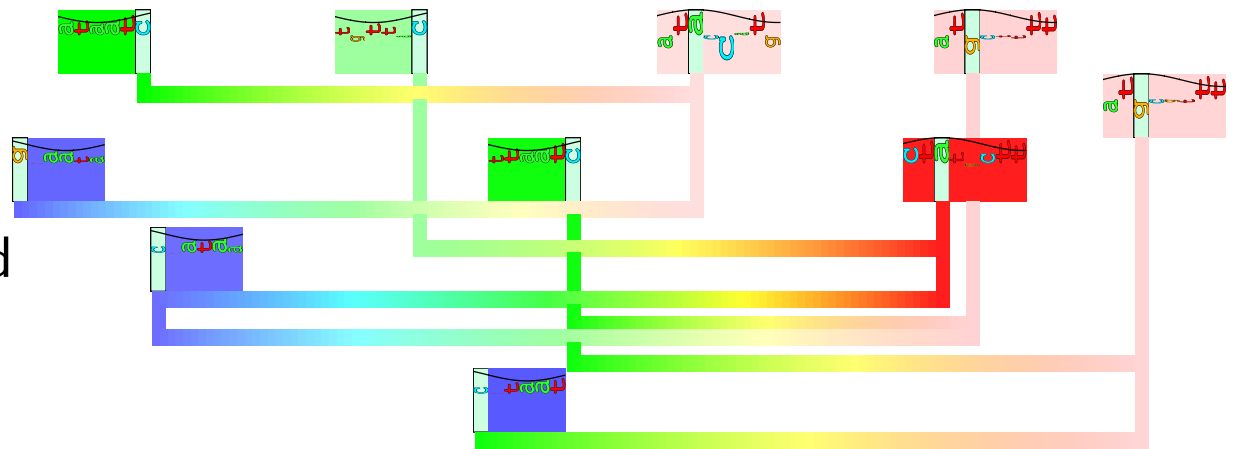
Complex Sequence Walker Example

- 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!

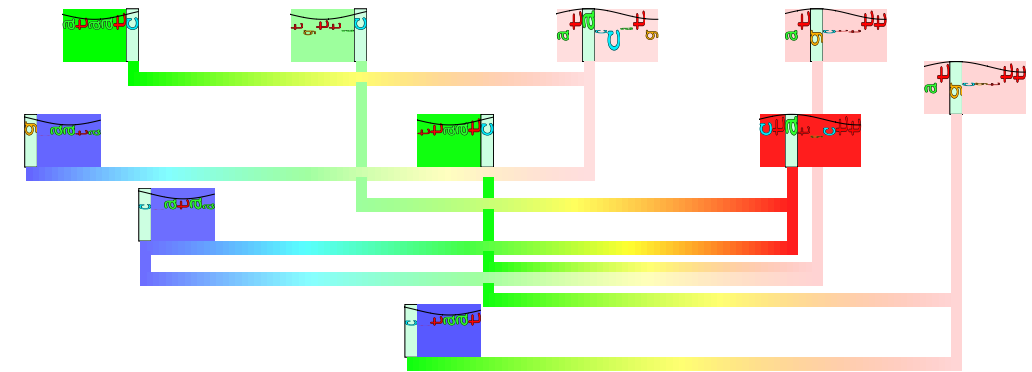
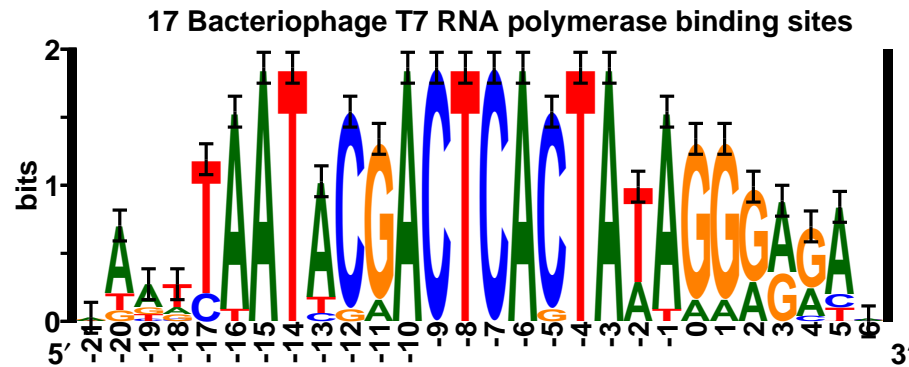
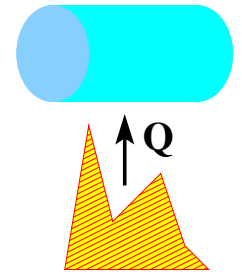
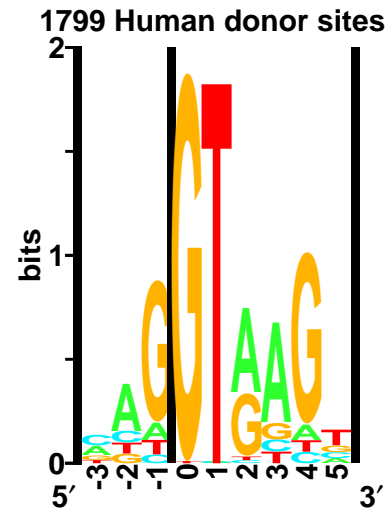
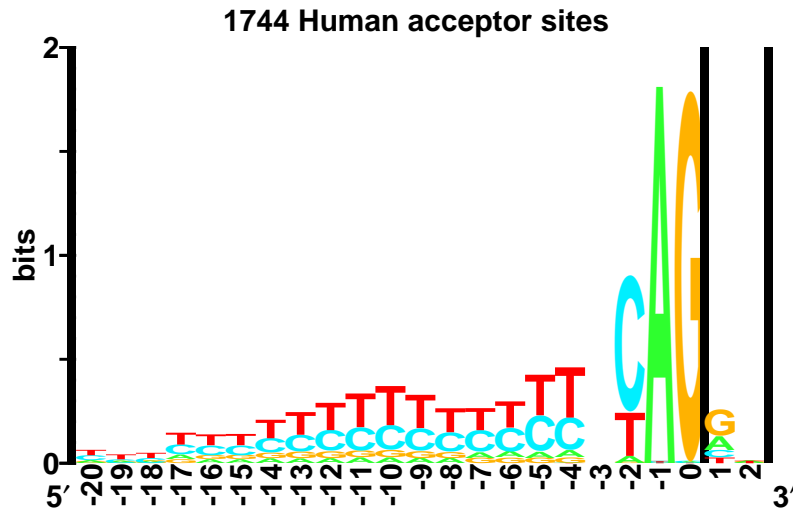


Acknowledgements

- Mentors:
 - Larry Gold (graduate school mentor)
 - Andrej Ehrenfeucht (information theory)
- Collaborators:
 - John Spouge (mathematics, proved individual info is unique)
 - Mike Stephens (logos, splicing)
 - Paul C. Anagnostopoulos (Java version of the evolution program, Evj)
 - Ryan Shultzaberger (flexible walkers)
 - σ^{38} : Kevin Franco, Zhe Sun, Ding Jin, Yixiong Chen, Cedric Cagliero, Yuhong Zuo, Yan Ning Zhou, Mikhail Kashlev
- Useful discussions with
 - Jeff Strathern
 - Amar Klar
 - Mark Lewandoski
- This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.



Biological Information theory: the mathematics of biology



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

$$H = - \sum_{i=1}^M p_i \log_2 p_i$$

Version

version = 1.09 of cutofftalk.tex 2021 Aug 27