Today's Lecture

- Information theory

Information Theory

- Gives useful concepts & terminology for describing how much “better” one probability model is than another.
- Gives interesting way to think about 2d law of thermodynamics
- Important in coding theory / data compression
- Suggests a useful approach (Minimum Description Length principle) to avoid overfitting data

From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

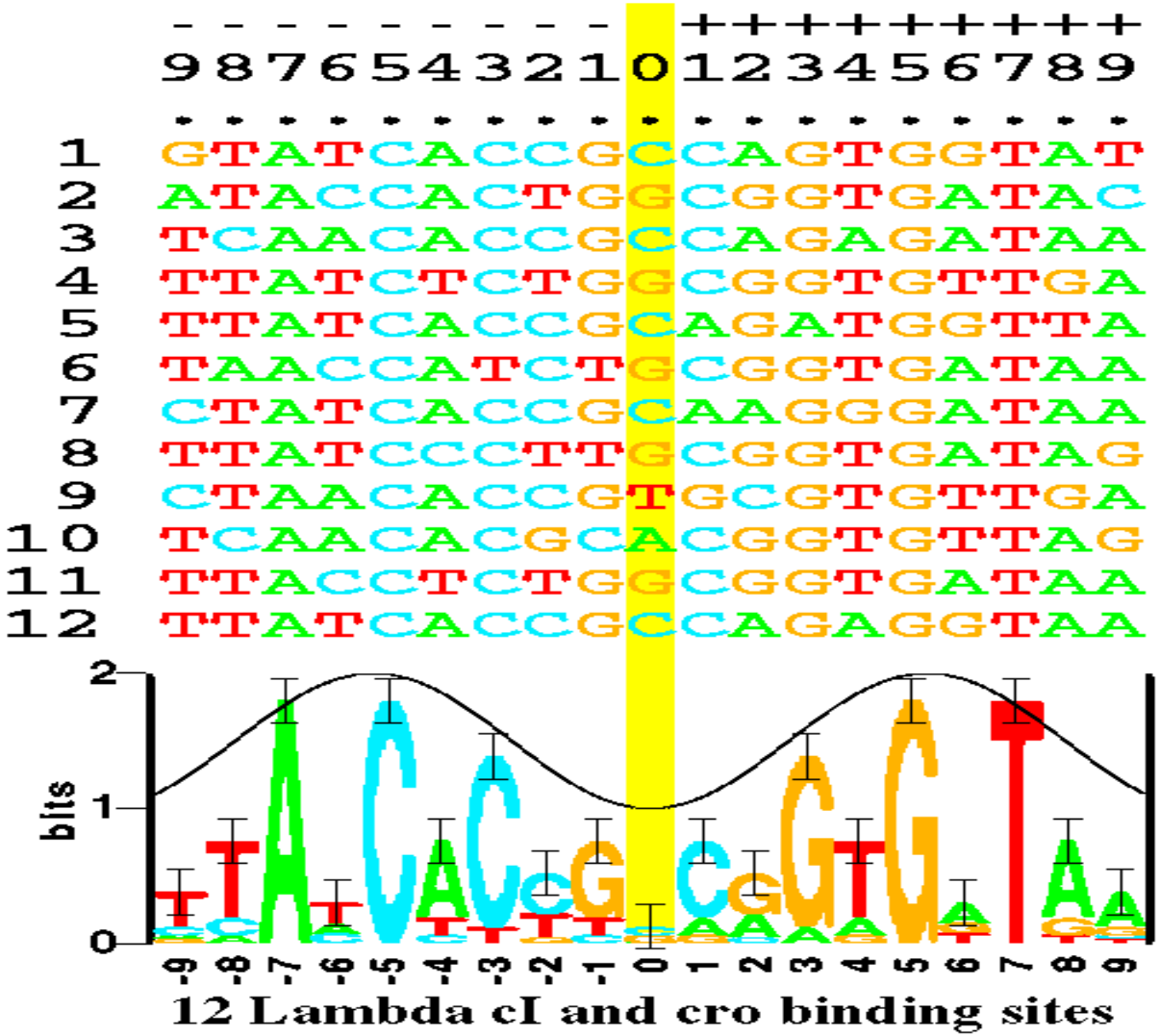


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

Entropy

- The *information theoretic entropy*
  - or *Shannon entropy*

of a probability space (S,P) is

$$H_b(P) = \Sigma_{s\in S} P(s)\log_b(1/P(s)) = -\Sigma_{s\in S} P(s)\log_b(P(s))$$

  - Terms with $P(s) = 0$ are set $= 0$
  - We usually take $b = 2$
    - in which case entropy is in "bits"
- $H_b(P) \geq 0$
    - because each term $P(s)\log_b(1/P(s)) \geq 0$

$H_b(P) = 0$ only for trivial dist'n concentrated in single point

Entropy (cont'd)

- Intuitively, the entropy measures how "spread out" the probability distribution is.
  - for $P(s)$ close to 0, or to 1, $P(s)\log_b(1/P(s))$ is close to 0.

Relative Entropy

- The *relative entropy* or *Kullback-Leibler distance* for two dist'ns P and Q on S is
$$D_b(P \| Q) \equiv \Sigma_{s \in S} P(s) \log_b(P(s) / Q(s))$$
(the expected value of the loglikelihood ratio).
  - if $P(s) = 0$, set corresponding term = 0
  - if $P(s) \neq 0$ but $Q(s) = 0$, $D_b(P \| Q)$ is taken to be $+\infty$.
- By information inequality, $D_b(P \| Q) \geq 0$, with equality only if $P = Q$.
- In general
$$D_b(P \| Q) \neq D_b(Q \| P)$$

Information Inequality

(Let $p_s = P(s)$, for $s \in S$). For any

- prob dist'n $\{p_s\}_{s \in S}$, and
- $\{q_s\}_{s \in S}$ satisfying $q_s \geq 0$ and $\Sigma_s\, q_s \leq 1$
  - e.g. $\{q_s\}$ a probability distribution

we have

$$\Sigma_s\, p_s\, ln(q_s) \leq \Sigma_s\, p_s\, ln(p_s)$$

with equality only if $q_s = p_s$ for all s ('$\forall\ s$')

Proof. $ln(x) \leq x - 1$ for all $x > 0$, with equality only for $x = 1$. (See next slide).

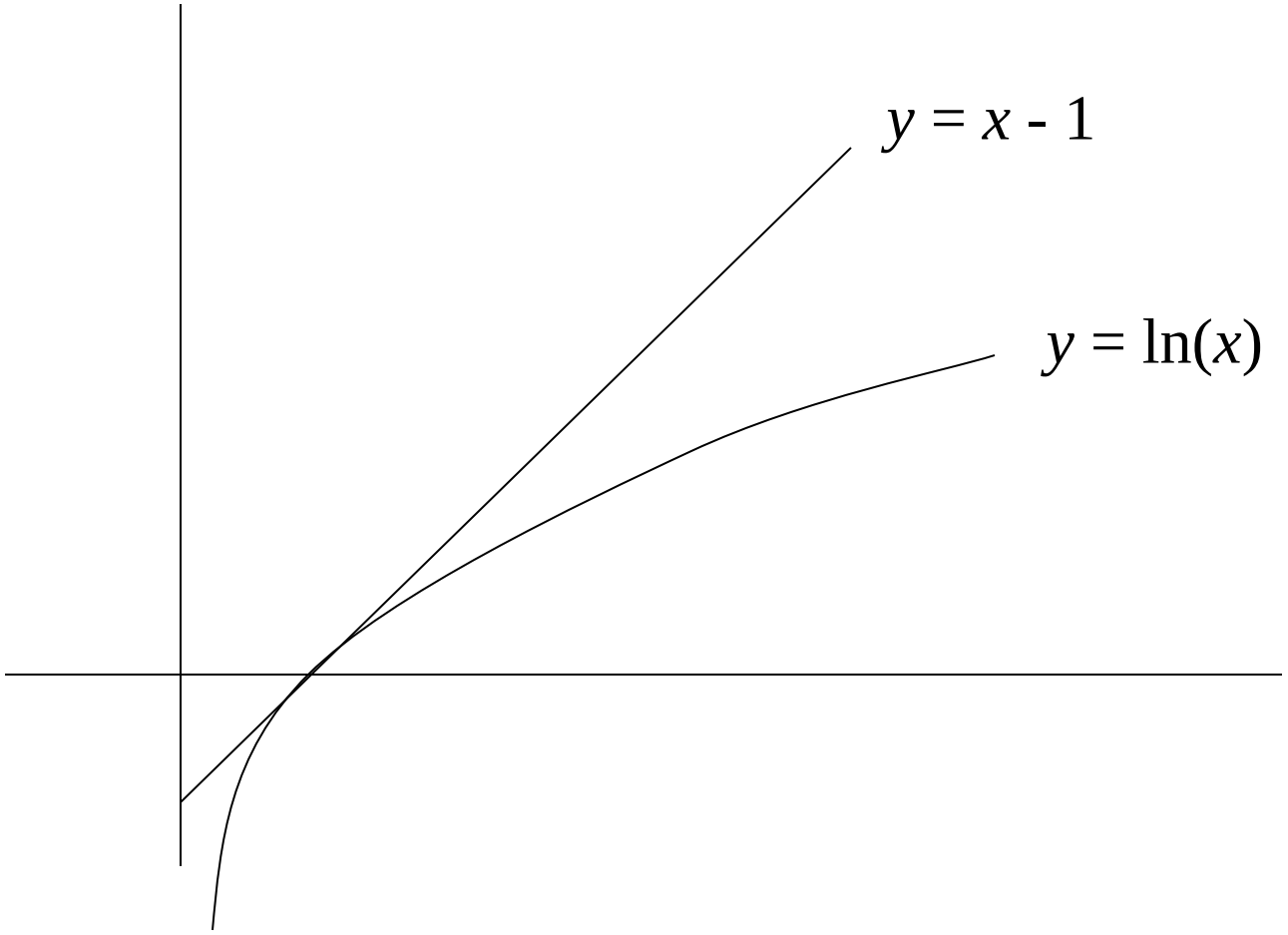
$$\therefore \quad \Sigma_s\, p_s\, ln(q_s) - p_s\, ln(p_s) = \Sigma_s\, p_s\, ln(q_s / p_s)$$

$$\leq \Sigma_s\, p_s\, (q_s / p_s - 1) \text{ (with equality only if } q_s = p_s\ \forall s)$$

$$= \Sigma_s\, q_s - \Sigma_s\, p_s \leq 1 - 1 = 0.$$

So $\Sigma_s\, p_s\, ln(q_s) \leq \Sigma_s\, p_s\, ln(p_s)$, with equality only if $q_s = p_s\ \forall\ s$.

$ln(x) \leq x - 1$



Information Inequality (cont'd)

- Since $\log_b$ for any base b is related to ln by

$$\log_b(x) = ln(x)/ln(b)$$

the information inequality holds for $\log_b$ as well:

$$\Sigma_s\, p_s \log_b(q_s) \le \Sigma_s\, p_s \log_b(p_s)$$

- Equivalent formulation: the entropy $H_b(\{p_s\})$ satisfies

$$H_b(\{p_s\}) = -\Sigma_s\, p_s \log_b(p_s) \le -\Sigma_s\, p_s \log_b(q_s) = \Sigma_s\, p_s \log_b(1/q_s)$$

for any dist'n $\{q_s\}$.

Distributions with Maximum Entropy

- For a sample space with n elements,
  - largest possible entropy (of any prob dist'n) is $\log_b(n)$, and
  - this attained only for prob dist'n $q_s = 1/n$ for each s :
- *Proof*. Take arbitrary prob dist'n $\{p_s\}$, and $\{q_s\}$ as above. Then

$$H_b(\{p_s\}) \le \Sigma_s\, p_s \log_b(1/q_s) = \Sigma_s\, p_s \log_b(n) = \log_b(n)$$

and

$$H_b(\{q_s\}) = \Sigma_s\, q_s \log_b(1/q_s) = \Sigma_s\, q_s \log_b(n) = \log_b(n)$$

Maximum Entropy Subject to Constraint: Boltzmann Distribution

- In physics,
  - S may correspond to the fixed set of *states* of a physical system,
  - the prob dist'n $P = \{p_s\}_{s \in S}$ may vary, subject to a *constraint* of the form

    $$\Sigma_s \, p_s \, E(s) = E$$

    where E and $\{E(s)\}$ are fixed (e.g. the expected energy of the system, and the energies of individual states respectively).
  - Note that

    $$\min_s E(s) = \Sigma_{t \in S} \, p_t \, (\min_s E(s)) \leq \Sigma_t \, p_t \, E(t) \leq \Sigma_t \, p_t \max_s E(s) = \max_s E(s).$$

    So (since the middle term = E)

    $$\min_s E(s) \leq E \leq \max_s E(s)$$
- We seek $\{p_s\}$ constrained as above for which the entropy $H(\{p_s\})$ is maximized.

Boltzmann Distribution (cont'd)

- Consider $\{q_s\} = \{q_s^{(r)}\}$ of the form $q_s = c_r e^{-rE(s)}$ where r is a constant and $c_r = 1/ (\Sigma_s e^{-rE(s)})$ is determined by the requirement that $\{q_s\}$ be a prob dist'n.
- We first want to show that there exists an r such that $\{q_s^{(r)}\}$ satisfies the above constraint on p, i.e. $\Sigma_s\, q_s^{(r)}\, E(s) = E$
- Write $q_s^{(r)} = c_r e^{-rE(s)} = c_r\, e^{-r\,(\min E(s))}\, e^{-r\,(E(s)-\min E(s))}$. As

  $r \to +\infty$, the last factor $e^{-r\,(E(s)-\min E(s))}$

  $= 1$ if $E(s) = \min_s E(s)$

  $\to 0$ if $E(s) \neq \min_s E(s)$ since then the exponent of e becomes large and negative.
- Consequently $\{q_s^{(r)}\}$ converges to a dist'n $\{q_s^{(\infty)}\}$ which satisfies $q_s^{(\infty)} = 0$ for any s for which $E(s) \neq \min_s E(s)$. Then $\Sigma_s\, q_s^{(\infty)}\, E(s) = \min_s E(s)$.

Boltzmann Distribution (cont'd)

- By a similar argument, as $r \to -\infty$, $\{q_s^{(r)}\}$ converges to a dist'n $\{q_s^{(-\infty)}\}$ which satisfies $q_s^{(-\infty)} = 0$ for any s for which $E(s) \neq \max_s E(s)$; and $\Sigma_s\, q_s^{(-\infty)}\, E(s) = \max_s E(s)$.
- Therefore since $\Sigma_s\, q_s^{(r)}\, E(s)$ is continuous in r it takes on all values between $\min_s E(s)$ and $\max_s E(s)$. In particular $\min_s E(s) \leq E \leq \max_s E(s)$, so we can find a value of r such that

$$\Sigma_s\, q_s^{(r)}\, E(s) = E$$

  i.e. $\{q_s^{(r)}\}$ satisfies the constraint.
- Then by the information inequality and the constraint on $\{p_s\}$,

$$H(\{p_s\}) \leq \Sigma_s\, p_s \log(1/q_s) = \Sigma_s\, p_s\, (r\, E(s) - \log(c_r))$$
$$= r\, E - \log(c_r)$$

Boltzmann Distribution (cont'd)

- But also $H(\{q_s^{(r)}\}) = \Sigma_s\, q_s^{(r)} \log(1/q_s^{(r)})$
  $= \Sigma_s\, q_s^{(r)} (r\, E(s) - \log(c_r)) = r\, E - \log(c_r) \geq H(\{p_s\})$

- So $\{q_s\}$ of the form $q_s = c_r e^{-rE(s)}$ (for an appropriate r which we have not computed explicitly!) has the maximum entropy of all prob dist'ns $\{p_s\}$ satisfying the constraint $\Sigma_s\, p_s\, E(s) = E$.
- For this distribution, the probability associated to the state s declines exponentially in $E(s)$. This is sometimes called the *Boltzmann distribution*, after its discoverer in the context of classical thermodynamics.

Basic Coding Theory/ Data Compression

- a *binary source code* for a prob space (S,P) is a mapping C: $S \to$ {strings of 0's and 1's}
  - $C(s)$ is called the *codeword* corresponding to s.
- Given C, and any "text" or string $s_1 s_2 \cdots s_n$ of elements in S
  - $s_i \in S$ for each i

  can create an encoded string $C(s_1)C(s_2) \cdots C(s_n)$ (of 0's and 1's)
  - i.e. replace each s_i by its codeword.

Uniquely Decodable Codes

- *C* is *uniquely decodable* if distinct strings from *S* always give distinct encoded strings

  $\Rightarrow$ can uniquely reconstruct the original message from the encoded message

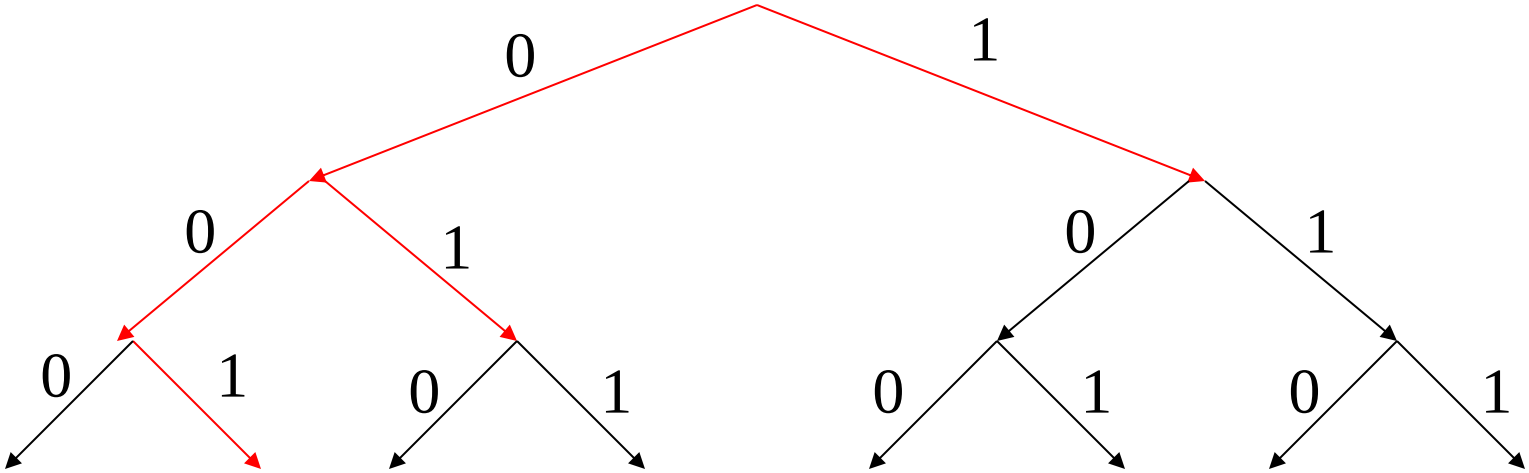
- *C* is a *prefix code* or *instantaneous code* if no codeword is a prefix of any other codeword.

- Examples: let S have three elements: 1,2,3. Then
  - $C(1) = 001$, $C(2) = 1$, $C(3) = 01$ is a prefix code on S.
  - $C(1) = 0$, $C(2) = 1$, $C(3) = 01$ is not a prefix code, because $C(1)$ is a prefix of $C(3)$.
    - Is it uniquely decodable?
  - Is $C(1) = 001$, $C(2) = 1$, $C(3) = 10$ a prefix code?
    - Is it uniquely decodable?
  - ASCII 8-bit code for representing alphabet & symbols is prefix code
    - because all codewords have same length!
    - UTF-8 is variable-width (one to four bytes) encoding of Unicode characters that includes ASCII & is a prefix code

- Prefix codes are uniquely decodable:
  - can decode the prefix-coded text by
    - reading through it in order, and
    - replacing each codeword by its corresponding *s* as soon as its end is recognized (whence “instantaneous”).
- For other types of uniquely decodable codes, may need to read whole text before decoding is possible.

Codewords as Paths

- Codewords correspond to paths from root in a *full* binary rooted tree of sufficient depth.
  - Each such path is uniquely determined by its end node.
- Code is a prefix code ⇔ no end node is ancestor or descendant of any other end node:



The three codewords are 001, 01, and 1

- Codewords in a prefix code are like the series of yes-no answers to “20 questions”, that uniquely determine a particular $s \in S$

Code Lengths

- For a code C, let $l_C(s)$ = length of $C(s)$, for $s \in S$.
- Equivalently, $l_C(s)$ = *depth* of the end node v_s of the corresponding path.

Kraft Inequality

- Let $l(s)$ assign positive integer to each $s \in S$. Then

$$l = l_C \text{ for some prefix code } C$$

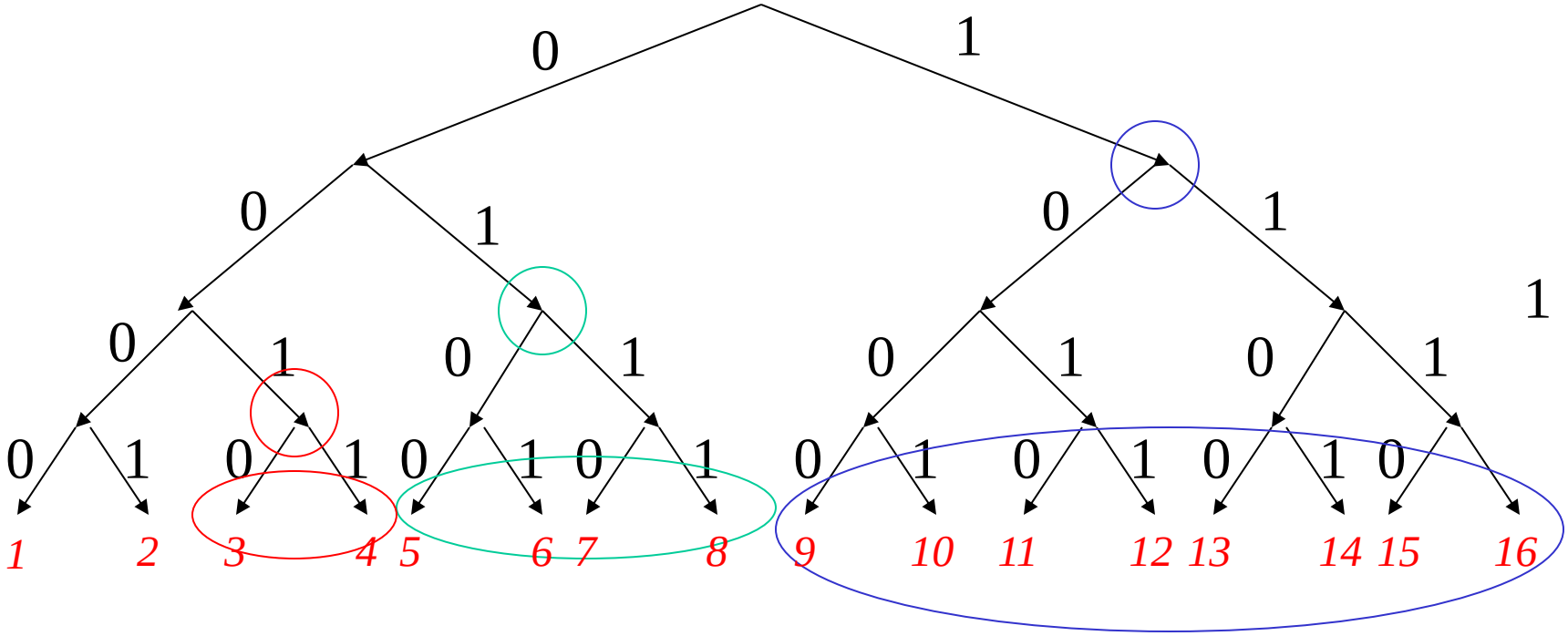
$$\Leftrightarrow \quad \Sigma_{s \in S} 2^{-l(s)} \leq 1$$

- *Example*: let $S = \{a,b,c\}$. Then can the following correspond to prefix codes?
  - $l(a) = 1, l(b) = 1, l(c) = 1$?
  - $l(a) = 1, l(b) = 1, l(c) = 2$?
  - $l(a) = 1, l(b) = 2, l(c) = 2$?

Proof of Kraft inequality

Consider full binary rooted tree of depth $n \geq \max_{s \in S} l(s)$.

Number *leaves* (= nodes of depth n) consecutively from left to right starting with 1:



Proof of Kraft inequality (cont'd)

- For each node v in the tree, if depth(v) = m then
  - v has 2^{n-m} descendants among the 2^n leaves; and
  - these are numbered consecutively from c to d, such that d is divisible by 2^{n-m}
- Conversely,

  a set of 2^{n-m} leaves consecutively numbered from c to d, & such that d is divisible by 2^{n-m}

  is the set of depth n descendants for a unique node v of depth m.
- If neither v_1 and v_2 is an ancestor of the other, then descendants of v_1 and v_2 are disjoint sets.

Proof of Kraft inequality (cont'd)

$\Rightarrow$: Assume $l = l_C$ for a prefix code C.

- C a prefix code $\Rightarrow$ end nodes v_s for the corresponding paths have disjoint sets of descendants
- Since v_s has $2^{n-l(s)}$ descendants in n^{th} row, $\Sigma_{s \in S} 2^{n-l(s)} \leq 2^n$.
- Cancelling 2^n , get $\Sigma_{s \in S} 2^{-l(s)} \leq 1$.

Proof of Kraft inequality (cont'd)

$\Leftarrow$: Conversely suppose $\Sigma_{s \in S} 2^{-l(s)} \leq 1$.

- Then $\Sigma_{s \in S} 2^{n-l(s)} \leq 2^n$.
- Arrange $l(s)$'s in increasing order
- Choose successive contiguous subsets V_s among leaves, starting from far left, such that $|V_s| = 2^{n-l(s)}$.
- Each such subset = {depth n descendants} for a unique node v_s in the tree, with depth(v_s) = $l(s)$.
- The mapping $s \rightarrow v_s$ then defines a prefix code C with $l = l_C$

Entropy & Expected Code Length

- The *expected length* $L(C)$ of a code C is given by
$$L(C) = \Sigma_s\, p_s\, l_C(s)$$
i.e. the expected value of the random variable l_C
- $L(C)$ = "expected # yes-no questions necessary to specify $s \in S$ using C"

  = avg # bits needed to encode a "character" $s \in S$, for text where each s used with freq p_s

Entropy & Expected Code Length (cont'd)

- For any prefix code, $L(C) \geq H_2(P)$:

  Proof. Define $q_s = 2^{-l(s)}$.

  – from Kraft inequality, $\Sigma_{s \in S} q_s \leq 1$, so

  – apply information inequality:

$$H_2(\{p_s\}) \leq \Sigma_s \, p_s \log_2(1/q_s) = \Sigma_s \, p_s \, l(s) = L(C)$$

- Conversely, can find prefix code C such that $L(C) < H_2(P) + 1$:

  Proof. Let $l(s)$ = smallest integer $\geq \log_2(1/\, p_s)$.

  – Then $2^{-l(s)} \leq p_s$, so $\Sigma_{s \in S}\, 2^{-l(s)} \leq \Sigma_s\, p_s = 1$.

  – By Kraft inequality $\exists$ prefix code C with $l = l_C$ Then

  $L(C) - H_2(P) = \Sigma_s\, p_s\, (l(s) - \log_2(1/\, p_s)) < \Sigma_s\, p_s\, (1) = 1$

– N.B.

- *C* chosen as above (the *Shannon code*) need not be optimal, in sense of having lowest possible $L(C)$.
- A construction of an optimal code is due to Huffman.

Interpretation of Entropy

- $\therefore$ $H_2(P)$ is (approximately!) the expected code length for an optimal prefix encoding of the probability space (S, P)

Uniquely Decodable Codes (cont'd)

- All uniquely decodable codes C satisfy Kraft inequality
  - for proof, see e.g. Cover & Thomas, *Elements of Information Theory*, sec. 5.5.
- Therefore $\exists$ prefix code D with the same codeword lengths as C:

$$l_C(s) = l_D(s) \text{ for all } s \in S.$$

- $\therefore$ expected codeword length $L(C)$ is same as for optimal prefix code
- in particular

$$L(C) = \Sigma_s p_s l_C(s) \geq H_2(P).$$

- $\therefore$ $H_2(P) \cong$ *minimum* avg # bits (0's and 1's), needed per character $s \in S$ to encode texts
  - for the *best possible uniquely decodable code*.
  - the relation becomes *exact* if more general codes (arbitrary invertible maps from texts to bit strings) are allowed

Entropy and Information

- By above, $H_2(P) \approx$ # bits needed "on average" to unambiguously specify elements of S.
- $\therefore$ Entropy = average "uncertainty" before an element of (S,P) is specified.
- *Information* corresponds to *reduction in uncertainty*.
  - Before elt of S is specified, the uncertainty is $H(P)$;
  - after it is specified, uncertainty is 0.
  - So the amount of information gained is $H(P) - 0 = H(P)$.
  - So entropy happens to equal information in this instance;
    - not in general though!

- So $H_2(P)$ = *avg amount of information per character* in a text based on (S, P).

Minimum Description Length Principle (MDL)

- Method for choosing among probability models
  - suggested by coding theory & parsimony principle (Occam's razor)
  - intent is to avoid overfitting
- idea: minimize total # bits needed to describe data, *including bits necessary to represent the model* (parameter values)
- 'best' model for data is one with minimum # bits

MDL

– Avg # bits needed to represent data, given model:

$$B_{data} = H_2(P) = \Sigma_{s \in S} P(s) \log_2(1/P(s))$$ (i.e. entropy)

- to represent a specific dataset s, given the prob model P:

$$\log_2(1/P(s)) = -\log_2(P(s)) \text{ bits}$$

  – (Shannon encoding – which is close to optimal)

– # bits needed to specify model:

$$B_{param} \approx (\#\text{ parameters}) \times \text{precision}$$

- some non-trivial issues here: can be many possible ways of 'specifying' parameters!

– *Minimize* $B_{data} + B_{param}$ over prob models & precisions

$\Leftrightarrow$ *maximizing* the (adjusted) relative entropy.

Avoiding overfitting – other approaches

- Most methods to avoid overfitting involve similar tradeoff:
  - in choosing among models, balance
    - goodness of fit to training data

      against

    - penalty for complexity of the model
- Other such methods (besides MDL) include:
  - AIC (Akaike information criterion)
  - BIC (Bayesian information criterion)

- A different, commonly used approach:
  - train multiple models on the ‘training’ data
  - then choose one that does best on separate (‘test’) data
- This is wrong: *test* data is being used for *training* !!
  - ‘training’ is *any procedure* for choosing among models, not only ‘estimating parameters’ (a *particular* type of choice)

  So still ∃ major risk of overfitting
- Can hold out part of test set for final, indep test
  - but performance in final test likely not as good

Relative Entropy

- The *relative entropy* or *Kullback-Leibler distance* for two dist'ns P and Q on S is
$$D_b(P \| Q) \equiv \Sigma_{s \in S} P(s) \log_b(P(s) / Q(s))$$
(the expected value of the loglikelihood ratio).
  - if $P(s) = 0$, set corresponding term = 0
  - if $P(s) \neq 0$ but $Q(s) = 0$, $D_b(P \| Q)$ is taken to be $+\infty$.
- By information inequality, $D_b(P \| Q) \geq 0$, with equality only if $P = Q$.
- In general
$$D_b(P \| Q) \neq D_b(Q \| P)$$

- For site dist'n P and background dist'n Q,
  - $D(P \,||\, Q)$ = the *mean* of site score distribution

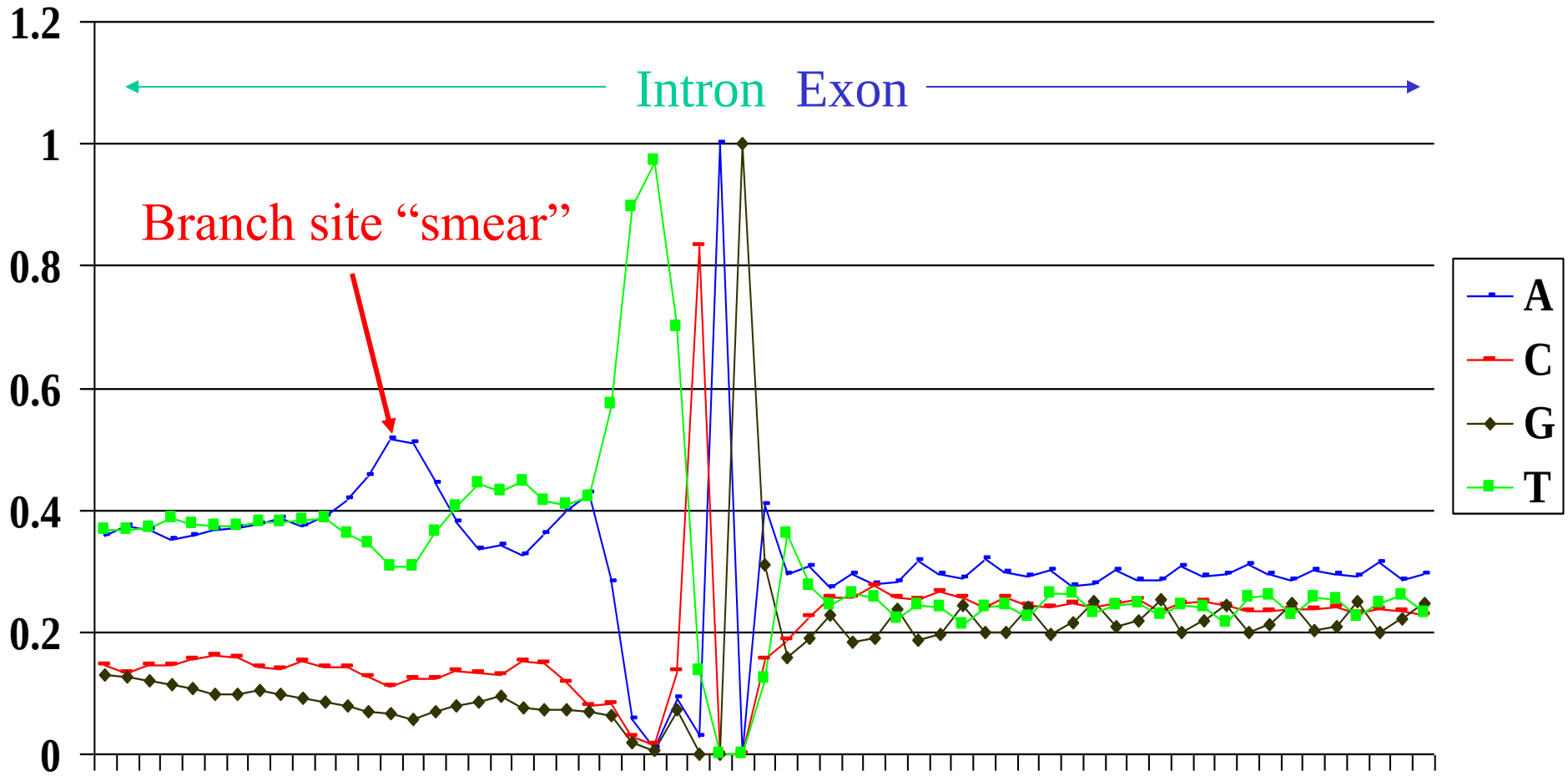
    i.e. the sum, over sequences, of prob of seq times its LLR weight.
- Since $P(s) = \prod_{1 \le i \le n} P_i(s_i)$ and $Q(s) = \prod_{1 \le i \le n} Q_i(s_i)$,

$$D(P \,||\, Q) = \sum_{s \in S} \left(\prod_{1 \le i \le n} P_i(s_i) \right) \sum_{1 \le j \le n} \left(\log(P_j(s_j)) - \log(Q_j(s_j)) \right)$$

which simplifies to

$$\sum_{1 \le i \le n} \left(\sum_{r \in A} P_i(r) \left(\log(P_i(r)) - \log(Q_i(r)) \right) \right) = \sum_{1 \le i \le n} D(P_i \,||\, Q_i)$$

3’ Splice Sites – *C. elegans*



Weight Matrix – 3' Splice Sites (*C. elegans*)

SITE FREQUENCIES:

A	0.400	0.429	0.282	0.058	0.008	0.092	0.029	1.000	0.000	0.410	0.293	0.307
C	0.118	0.079	0.081	0.029	0.016	0.135	0.834	0.000	0.000	0.156	0.187	0.225
G	0.072	0.070	0.063	0.018	0.005	0.073	0.001	0.000	1.000	0.310	0.159	0.191
T	0.409	0.422	0.574	0.896	0.971	0.700	0.135	0.000	0.000	0.124	0.361	0.276

BACKGROUND FREQUENCIES:

A	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321
C	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179
G	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179
T	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321

WEIGHTS:

A	0.32	0.42	-0.18	-2.46	-5.29	-1.79	-3.45	1.64	-99.00	0.36	-0.13	-0.06
C	-0.60	-1.18	-1.15	-2.64	-3.51	-0.41	2.22	-99.00	-99.00	-0.20	0.06	0.33
G	-1.31	-1.35	-1.51	-3.35	-5.23	-1.30	-6.93	-99.00	2.48	0.79	-0.17	0.10
T	0.35	0.39	0.84	1.48	1.60	1.12	-1.24	-99.00	-99.00	-1.37	0.17	-0.22

3’ Splice Sites

WEIGHTS:

A	0.32	0.42	-0.18	-2.46	-5.29	-1.79	-3.45	1.64	-99.00	0.36	-0.13	-0.06
C	-0.60	-1.18	-1.15	-2.64	-3.51	-0.41	2.22	-99.00	-99.00	-0.20	0.06	0.33
G	-1.31	-1.35	-1.51	-3.35	-5.23	-1.30	-6.93	-99.00	2.48	0.79	-0.17	0.10
T	0.35	0.39	0.84	1.48	1.60	1.12	-1.24	-99.00	-99.00	-1.37	0.17	-0.22

Position-specific Relative Entropy:

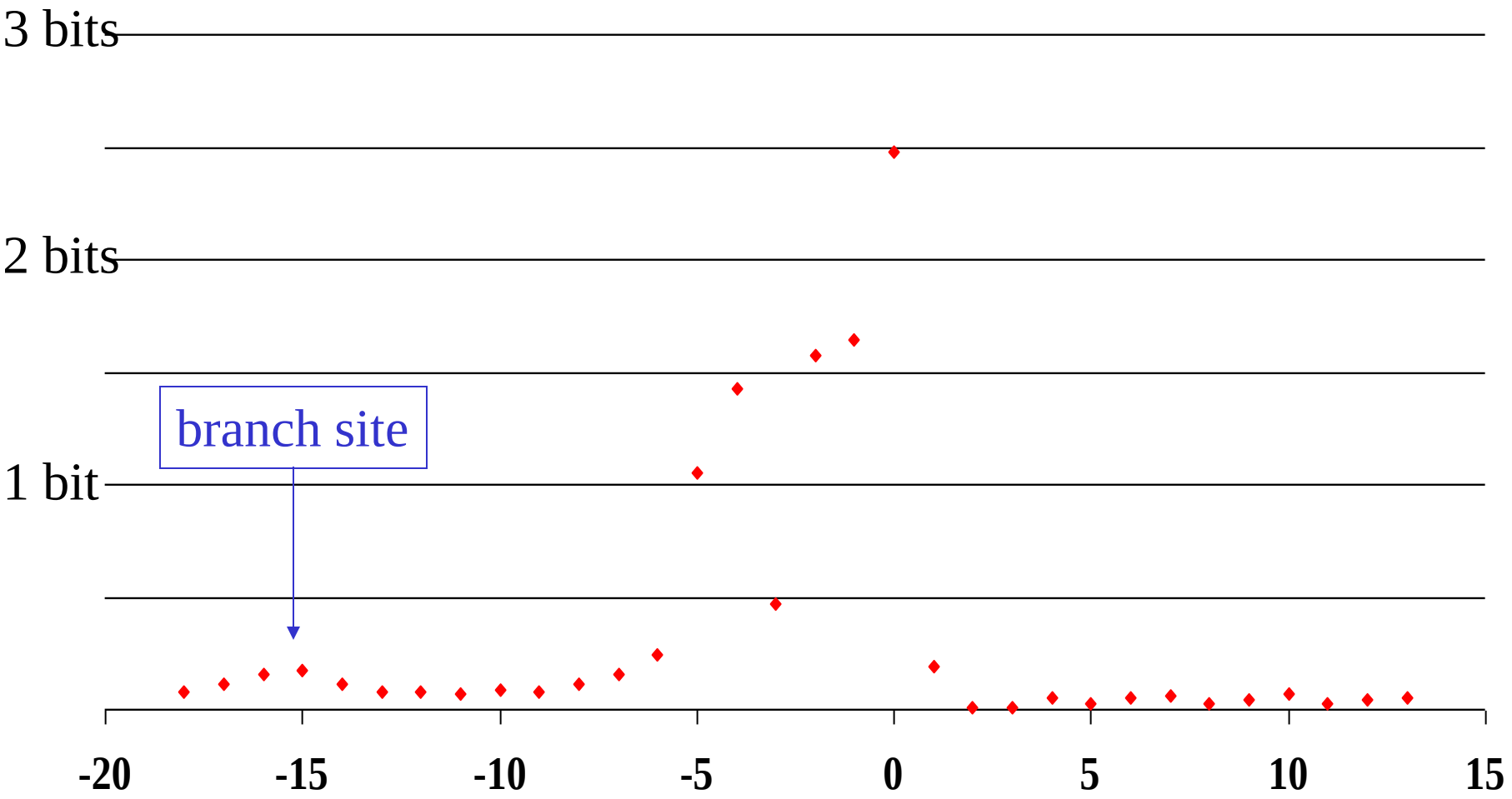
0.11	0.16	0.24	1.05	1.43	0.47	1.57	1.64	2.48	0.19	0.01	0.01

e.g. 0.11 = .400 (.32) + .118 (-.60) + .072 (-1.31) + .409 (.35)

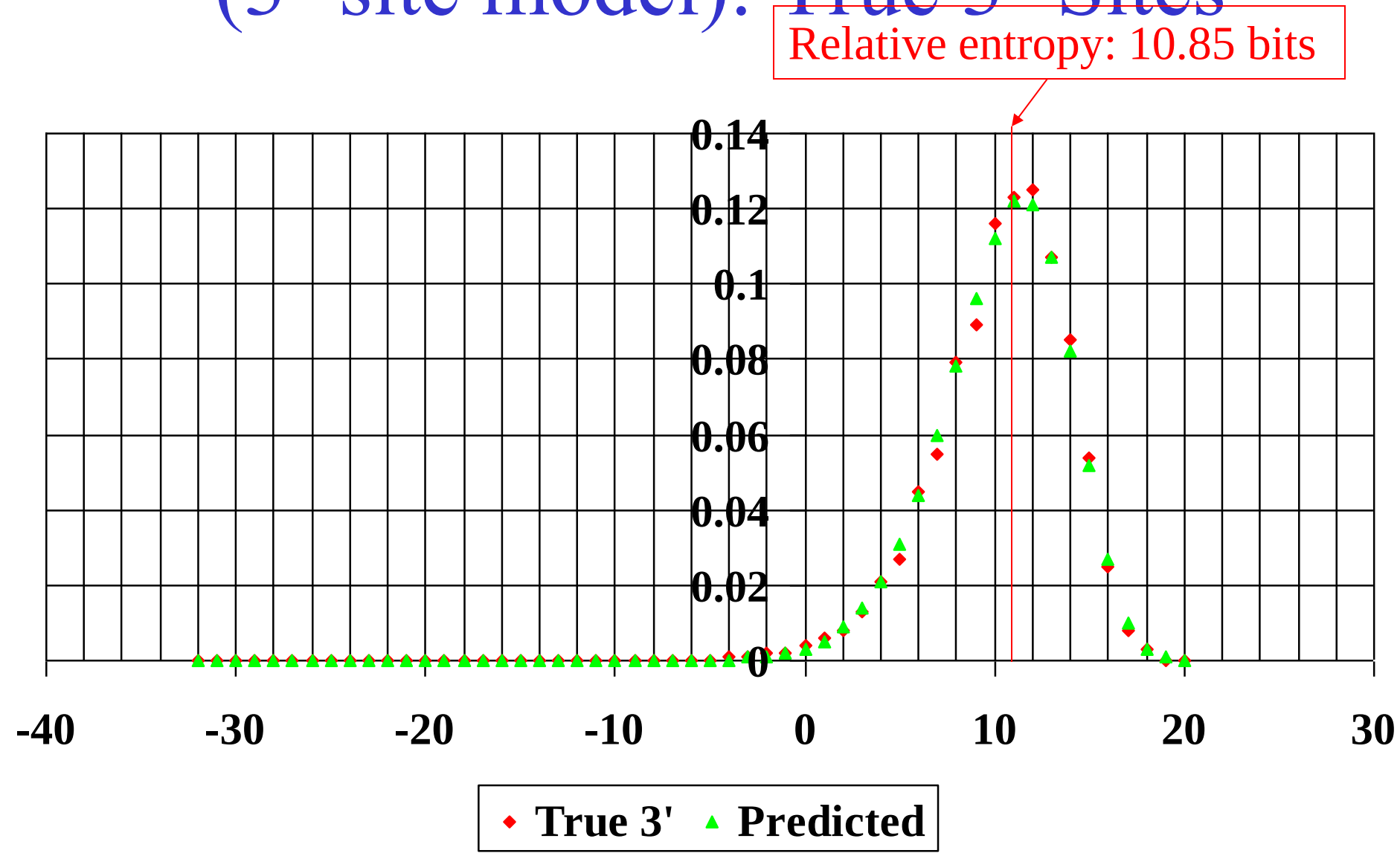
Total Relative Entropy (Sum of position-specific values) = 9.35

- Note pos-specific relative entropy always ≥ 0
  = 0 only if site freqs *exactly* equal backgd freqs.
    - will rarely happen, even far from site (when we're in backgd).
- So rel entropy increases indefinitely as window size increases
  – even when no biological information being added.
- For large enough window get spuriously clean score separation between training seqs and other seqs
  – *overfitting*.

Position-Specific Relative Entropy: 3’ Splice Sites



Predicted vs. Observed Distributions (3’ site model): True 3’ Sites



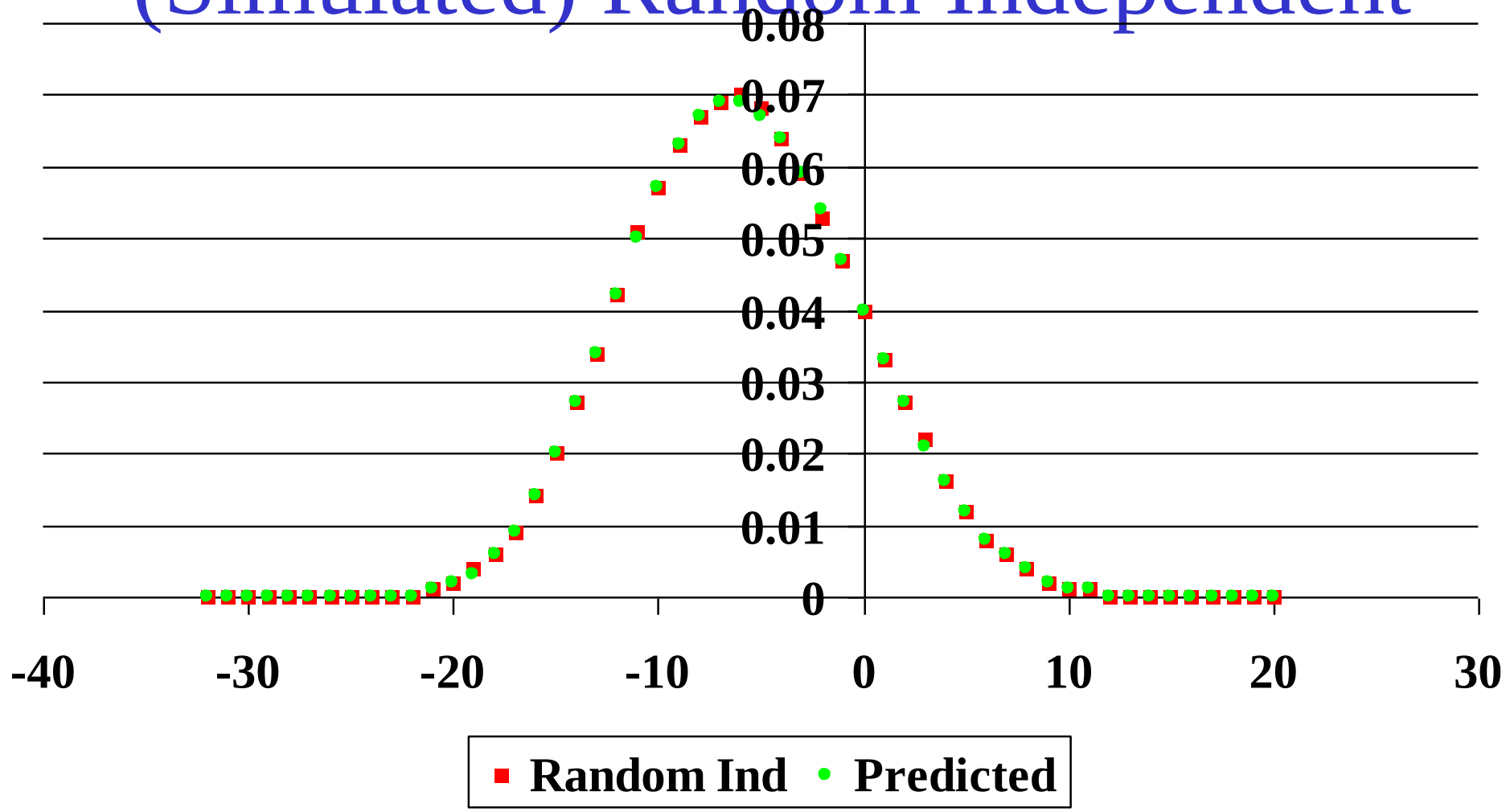
- Similarly,

$$D_b(Q \,\|\, P) = \Sigma_{s \in S} Q(s) \log_b(Q(s) / P(s))$$
$$= - \Sigma_{s \in S} Q(s) \log_b(P(s) / Q(s))$$

= *negative* of the mean of the dist'n of the LLR scores in background sequence (the "null distribution");

  – but must eliminate s for which $P(s) = 0$.

Predicted vs. Observed Distributions (3’ site model): (Simulated) Random Independent



Sequence Logos

- Schneider and Stephens (NAR 18, 6097-6100, 1990)– see http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html
- At i^{th} position, each residue r gets height
  $P_i(r)D(P_i \,||\, Q_i)$
- Schneider
  - takes Q_i to be the equal-frequency model
  - subtracts small-sample correction from $D(P_i \,||\, Q_i)$
- Gorodkin, Heyer, Brunak and Stormo (CABIO 13, 583-586, 1997)
  - use unequal frequency Q_i
  - allow for gaps
  - take height either proportional to $P_i(r)$ (as above) or to $P_i(r)/ Q_i(r)$, letter upside down if $P_i(r) < Q_i(r)$.

From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

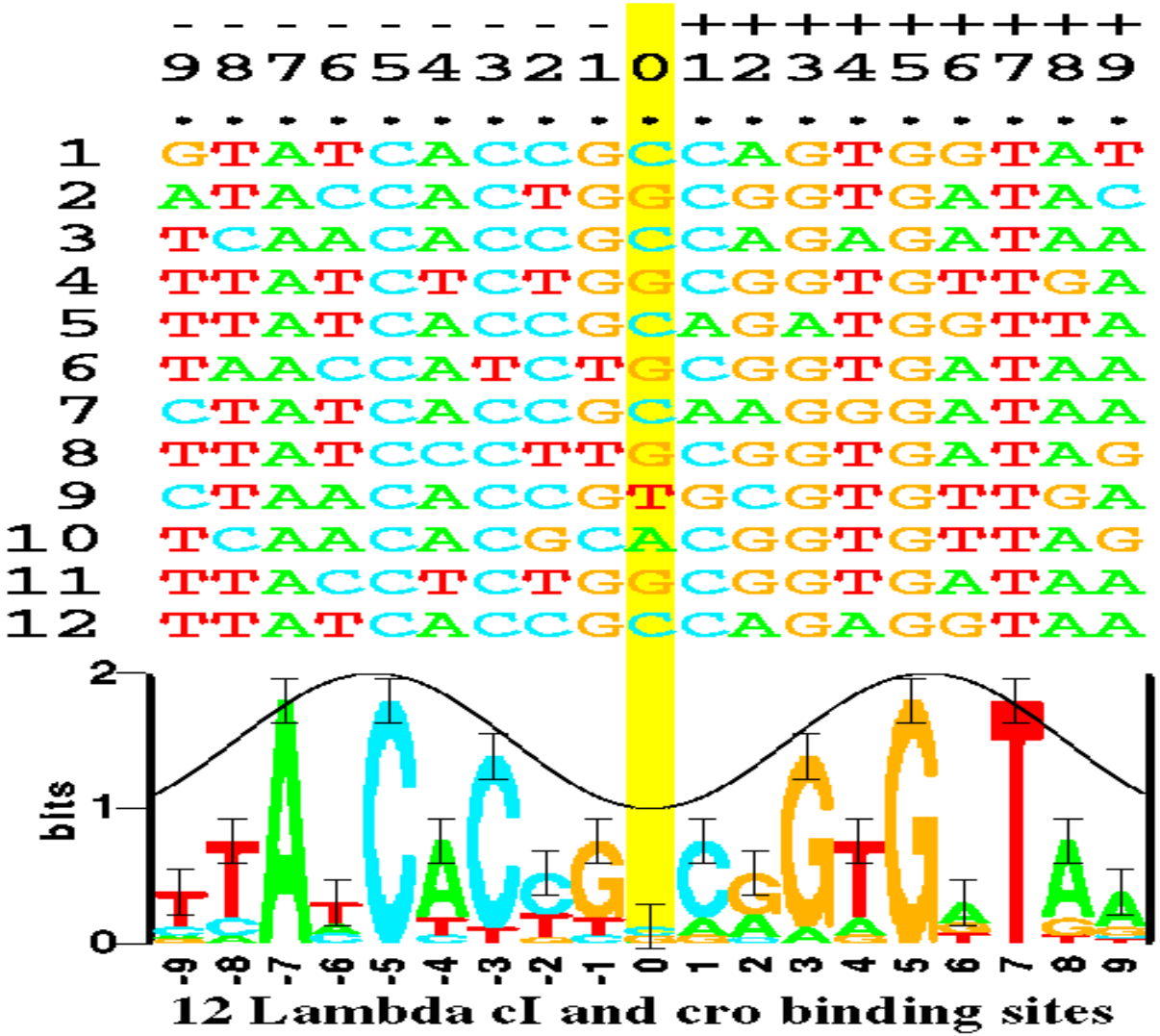
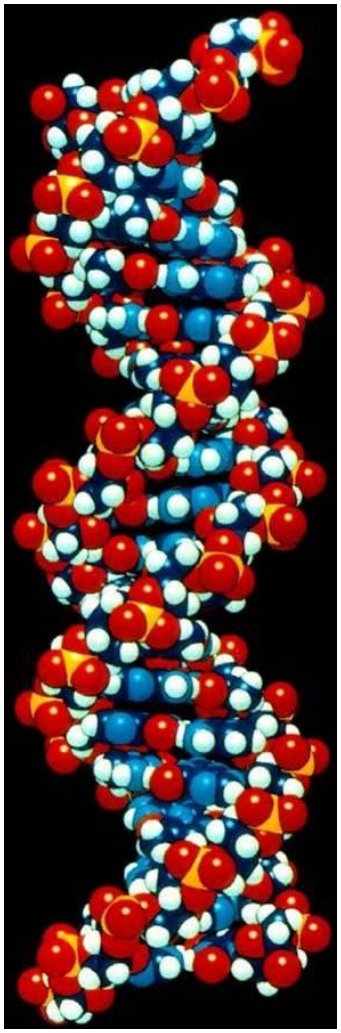
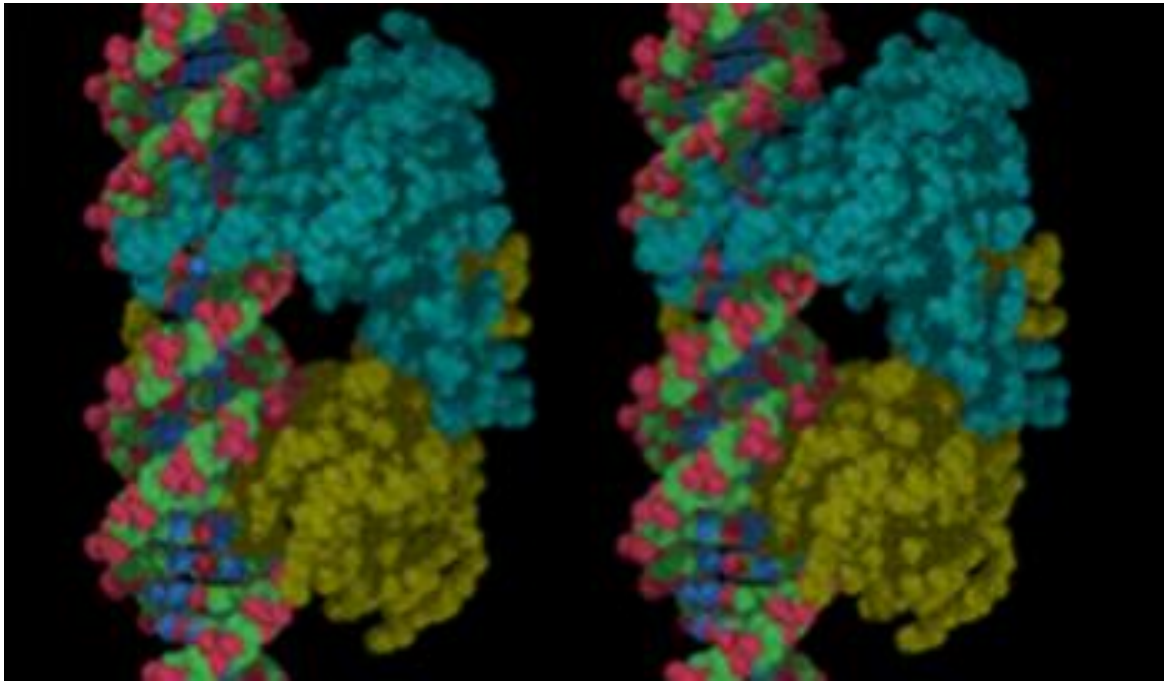


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

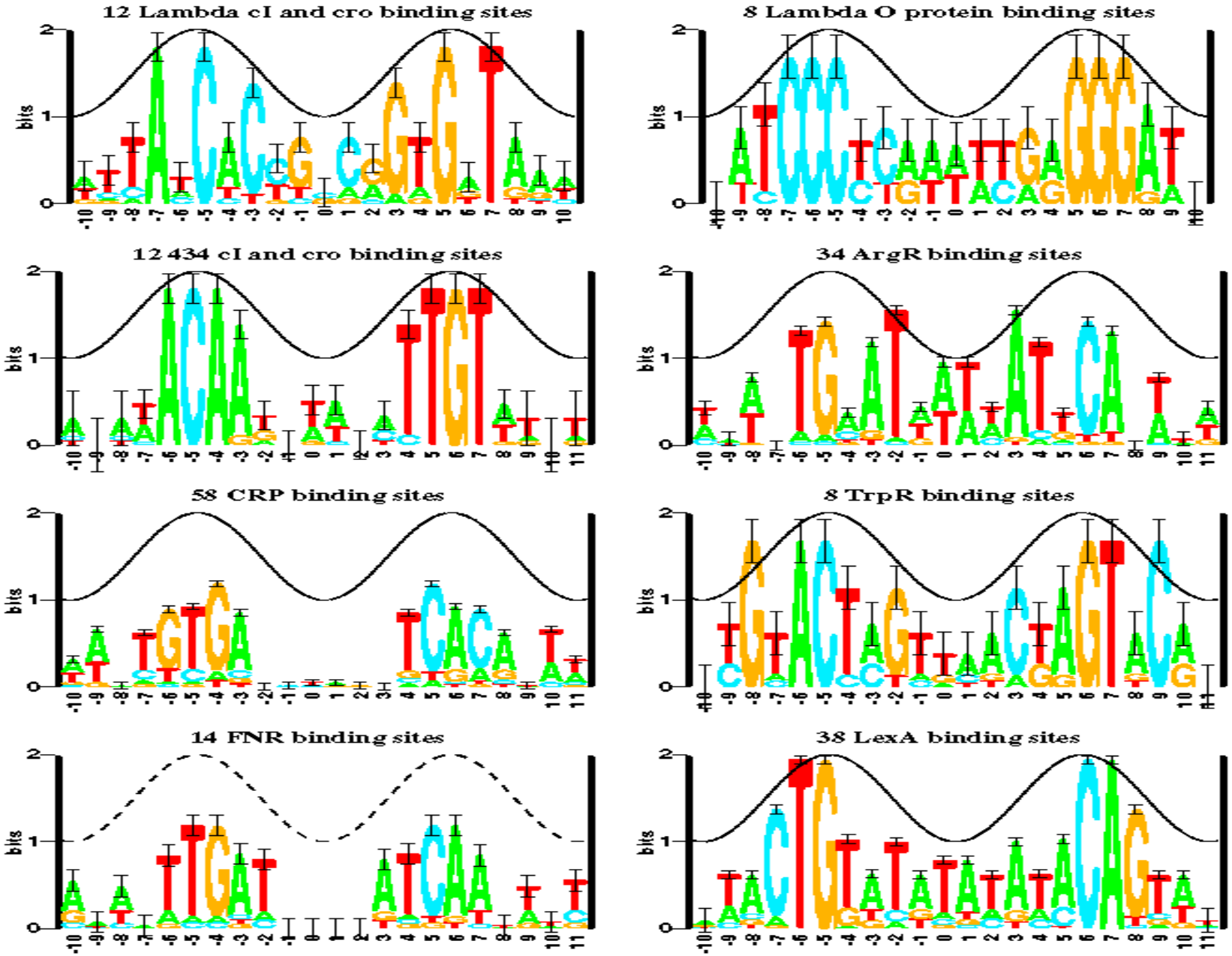


from http://www.dna-dna.net/

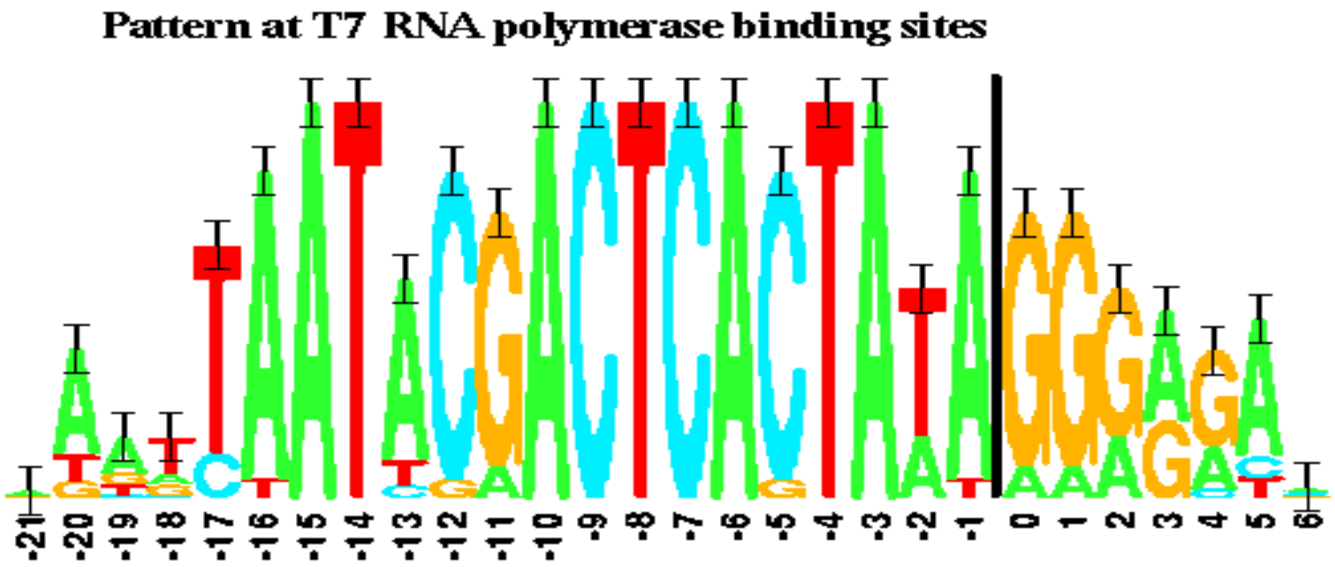


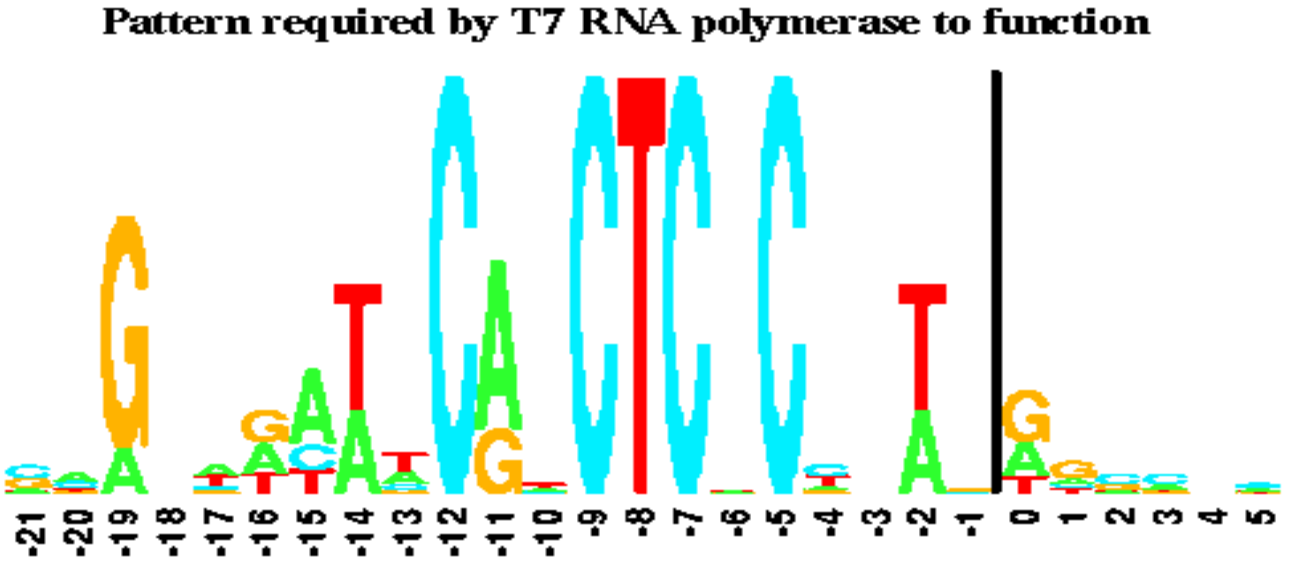
from http://gibk26.bse.kyutech.ac.jp

From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

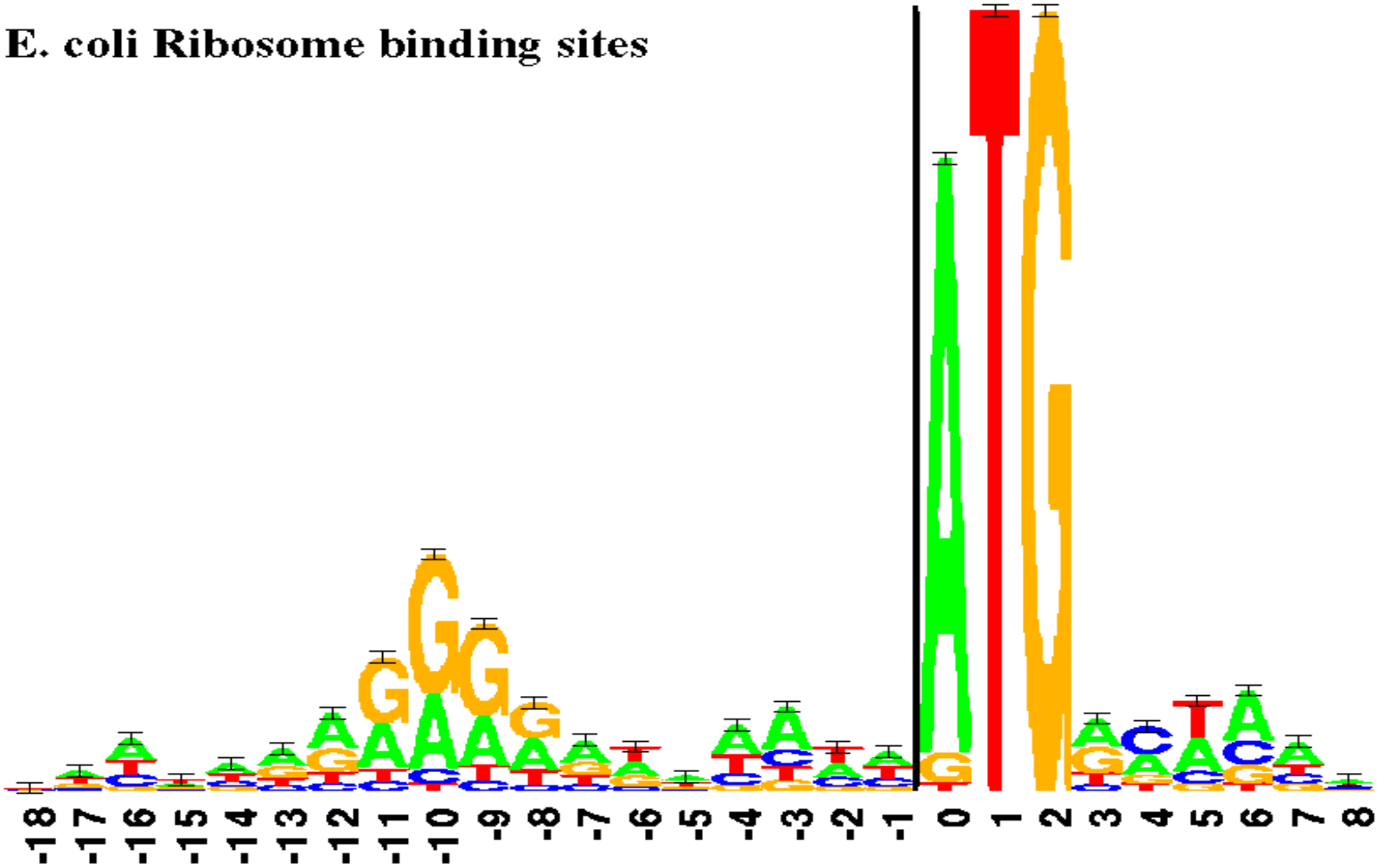


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



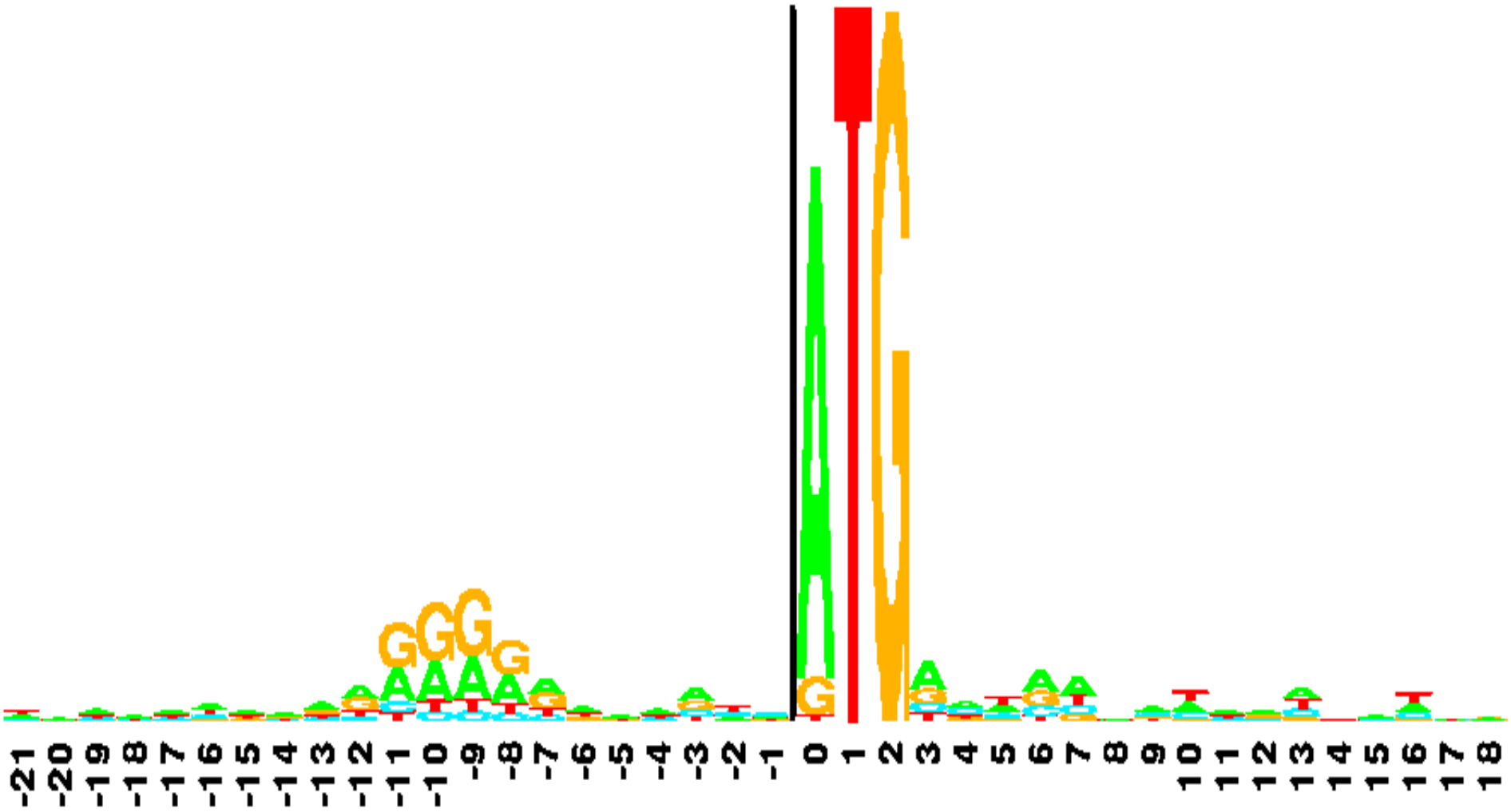


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



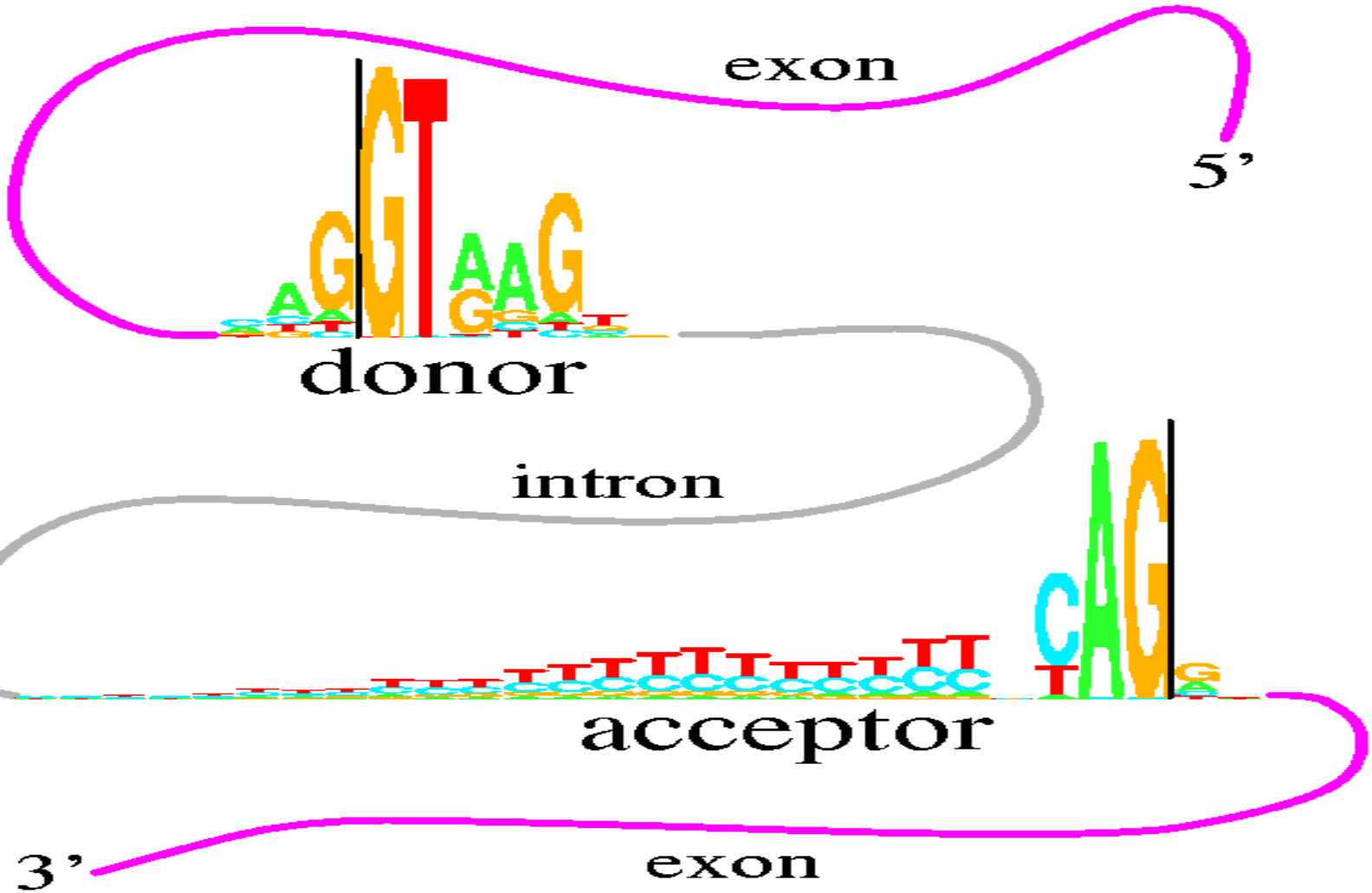
From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

1055 E. coli Ribosome binding sites listed in the Miller book

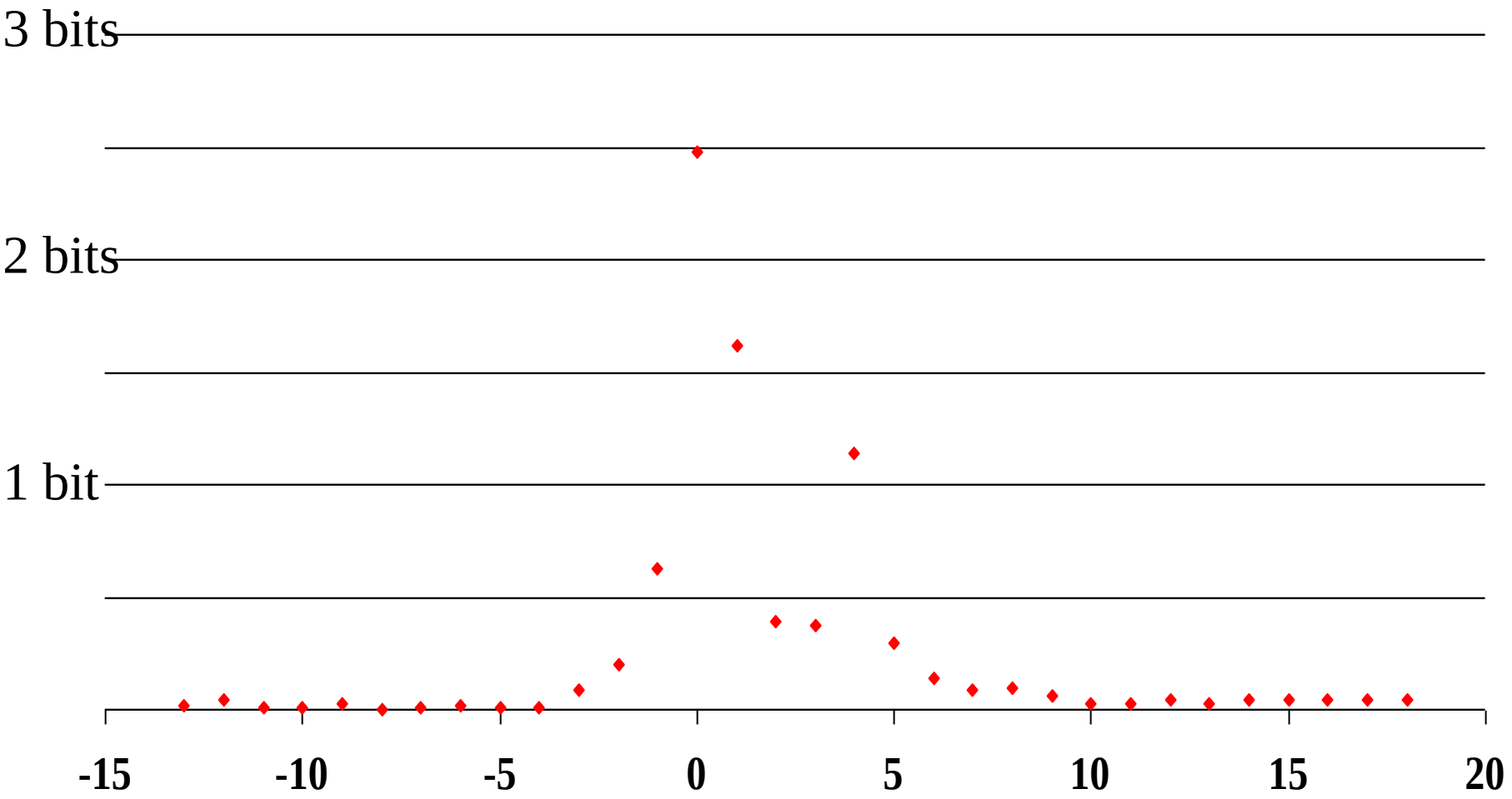


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

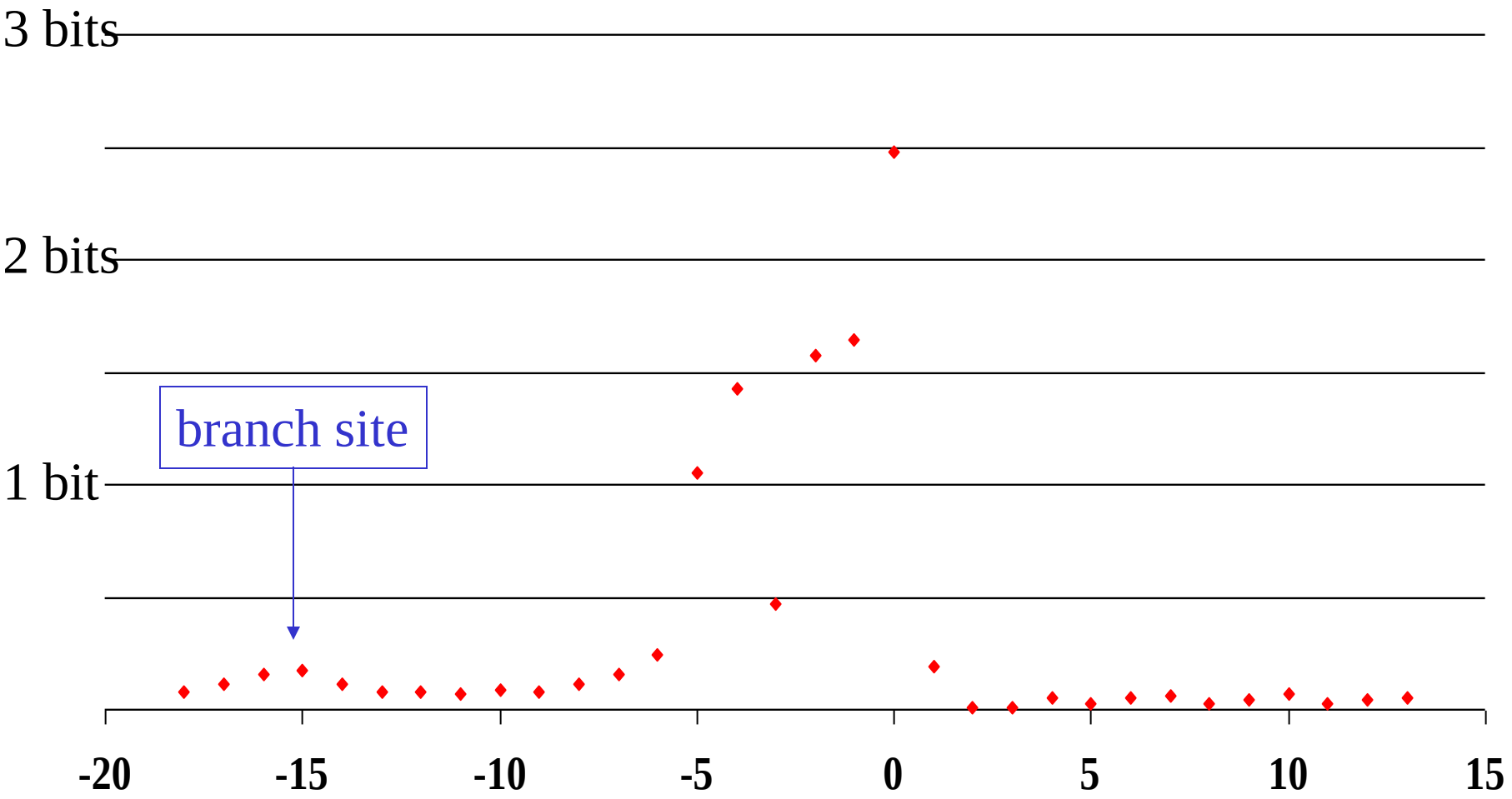
This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



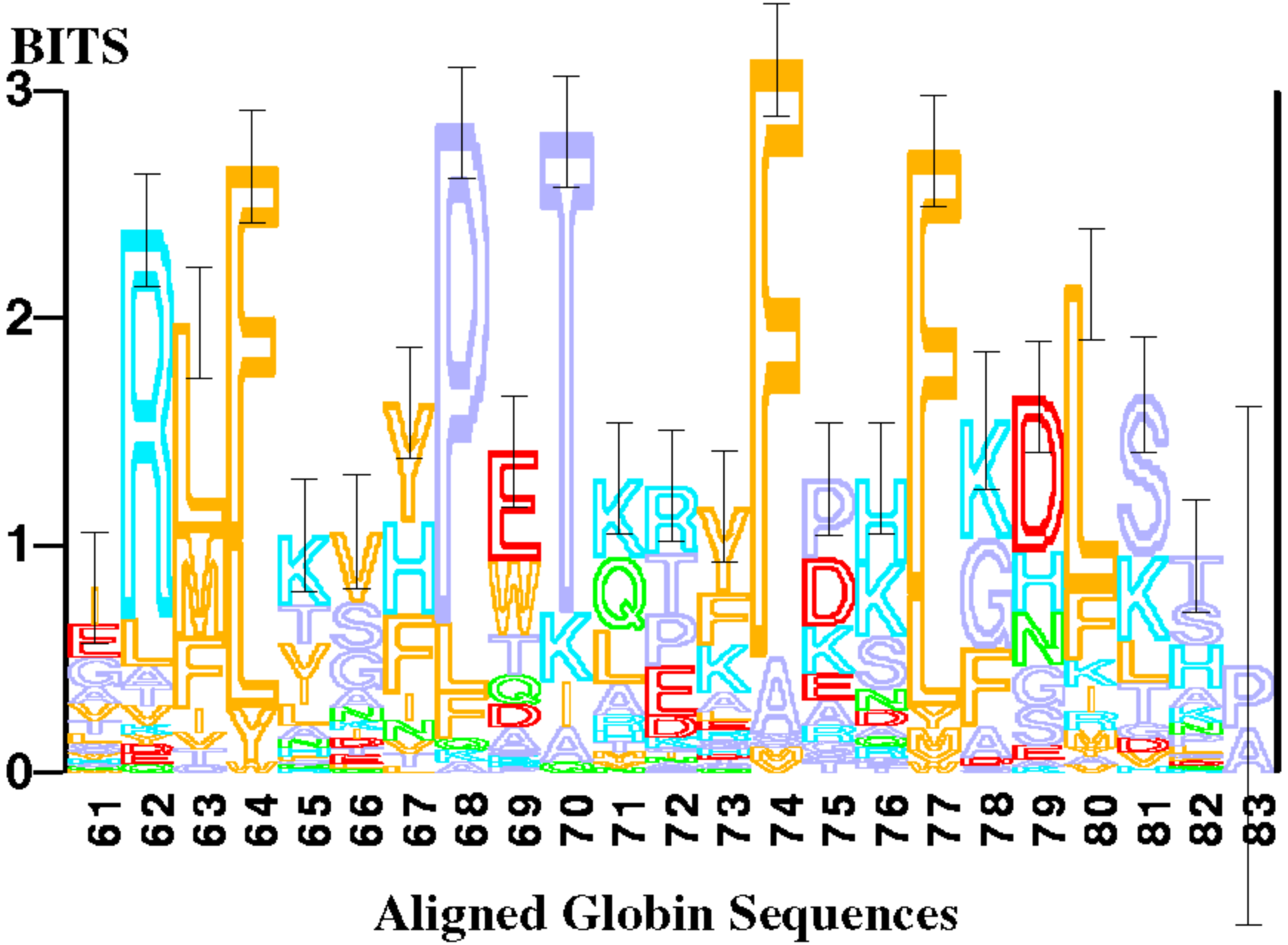
Position-Specific Relative Entropy: *C. elegans* 5' Splice Sites



Position-Specific Relative Entropy: 3’ Splice Sites



From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

Logo of Gibbs Block D (Tc1) 9 sequences