Lecture 5

- Average LRs & frequencies of site-like sequences in background

- Relative entropies (average LLRs)

- Sequence logos

Average likelihood ratios

- ***average LR*** (for sites) ≈ ***average spacing*** between occurrences of ‘site-like’ sequences ***in background***
- So e.g. for 3’ splice sites
  - – if the average *LR* is 1000, then one expects ‘splice-site-like’ sequences to occur on average once per kb *in background sequence*
  - – ***N.B.*** This says nothing about the frequency of *actual* splice sites! (which could be greater or smaller than 1 per kb), and so doesn’t by itself provide the probability that an *apparent* splice site is an *actual* site.

“*Proof*” :

- *Notation*
  - – S = seqs of length n, P = site prob dist’n, Q = bkgd dist’n
  - – T = ‘site-like’ sequences: s for which $P(s) >> Q(s)$
    - Assume ‘most’ site sequences are in T
- Simplest case: a *unique* site sequence s. Then
  - – $P(s) = 1$
  - – $LR(s) = 1 / Q(s)$ = avg spacing between occurrences of s

- More generally:

(weighted) avg $LR \approx \Sigma_{s\in T} P(s)\ (P(s)\ /\ Q(s))$

$\approx 1\ /\ \Sigma_{s\in T} P(s)\ (Q(s)\ /\ P(s))$

(‘Avg of reciprocals ≈ reciprocal of avg’ – true if the $P(s)\ /\ Q(s)$ have similar sizes)

$= 1\ /\ \Sigma_{s\in T} Q(s)$

= avg spacing (in bkgd) between seqs in *T*. *QED*

- Have *exact* equality only when all site sequences have the *same LR*

- Similar intuition – made mathematically precise – underlies *Karlin-Altschul theory* (for BLAST scores)
  - Query = cluster of (overlapping) ‘sites’, of varying lengths
  - Database = ‘genome’
  - K-A showed any ‘reasonable’ scoring scheme for alignments is rescalable to LLR
  - Look for matches (= ‘site-like’ sequences in database) for which the corresponding LR is much bigger than the *size of the database**, so unlikely to be a chance match to background
    - *Actually, the *product* of the query & database sizes, to correct for multiple testing

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 the ***information inequality***, $D_b(P \,||\, Q) \geq 0$, with equality only if $P = Q$.
- In general
  $$D_b(P \,||\, Q) \neq D_b(Q \,||\, P)$$

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

- 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.

- For site dist’n P and background dist’n Q,

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

independence assumption

$$= \sum_{1\le j\le n} \sum_{s\in S} P(s)(\log\left(P_j\left(s_j\right)\right) - \log\left(Q_j\left(s_j\right)\right))$$

summation order

$$= \sum_{1\le j\le n} \sum_{r\in A} \sum_{s|s_j=r} P(s) \ (\log\left(P_j(r)\right) - \log\left(Q_j(r)\right))$$

grouping by r at j-th position

$$= \sum_{1\le j\le n} \sum_{r\in A} (\sum_{s|s_j=r} P(s) \ ) \ (\log\left(P_j(r)\right) - \log\left(Q_j(r)\right))$$

factoring out constant

$$= \sum_{1\le j\le n} \sum_{r\in A} P_j(r)(\log\left(P_j(r)\right) - \log\left(Q_j(r)\right)$$

independence assumption

$$= \sum_{1\le j\le n} D(P_j \,||Q_j)$$

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
```

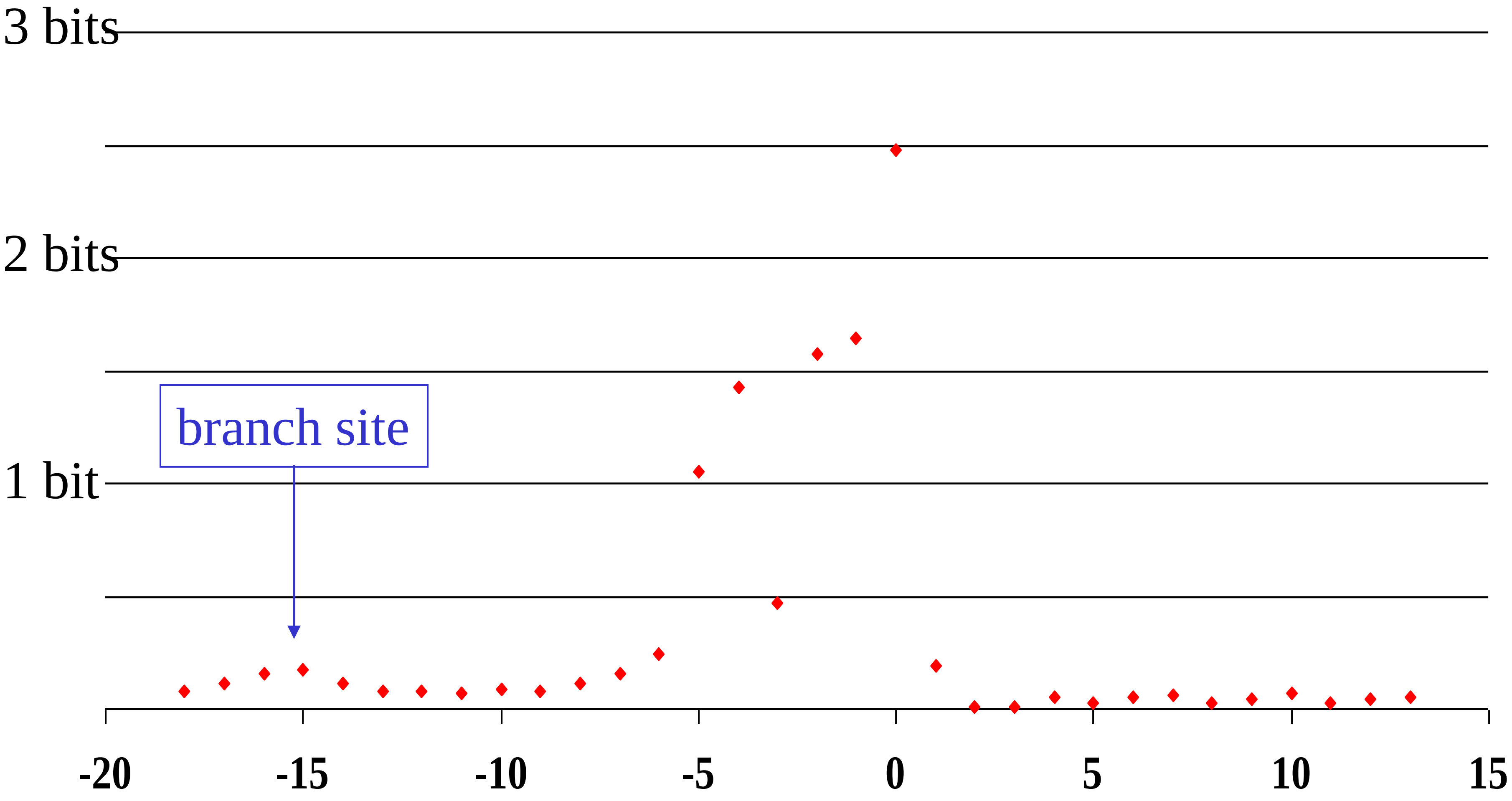
```
WEIGHTS:
A   0.32   0.42  -0.18  -2.46  -5.29  -1.79  -3.45   1.64 -99.00   0.36  -0.13  -0.06
C  -0.60  -1.18  -1.15  -2.64  -3.51  -0.41   2.22 -99.00 -99.00  -0.20   0.06   0.33
G  -1.31  -1.35  -1.51  -3.35  -5.23  -1.30  -6.93 -99.00   2.48   0.79  -0.17   0.10
T   0.35   0.39   0.84   1.48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

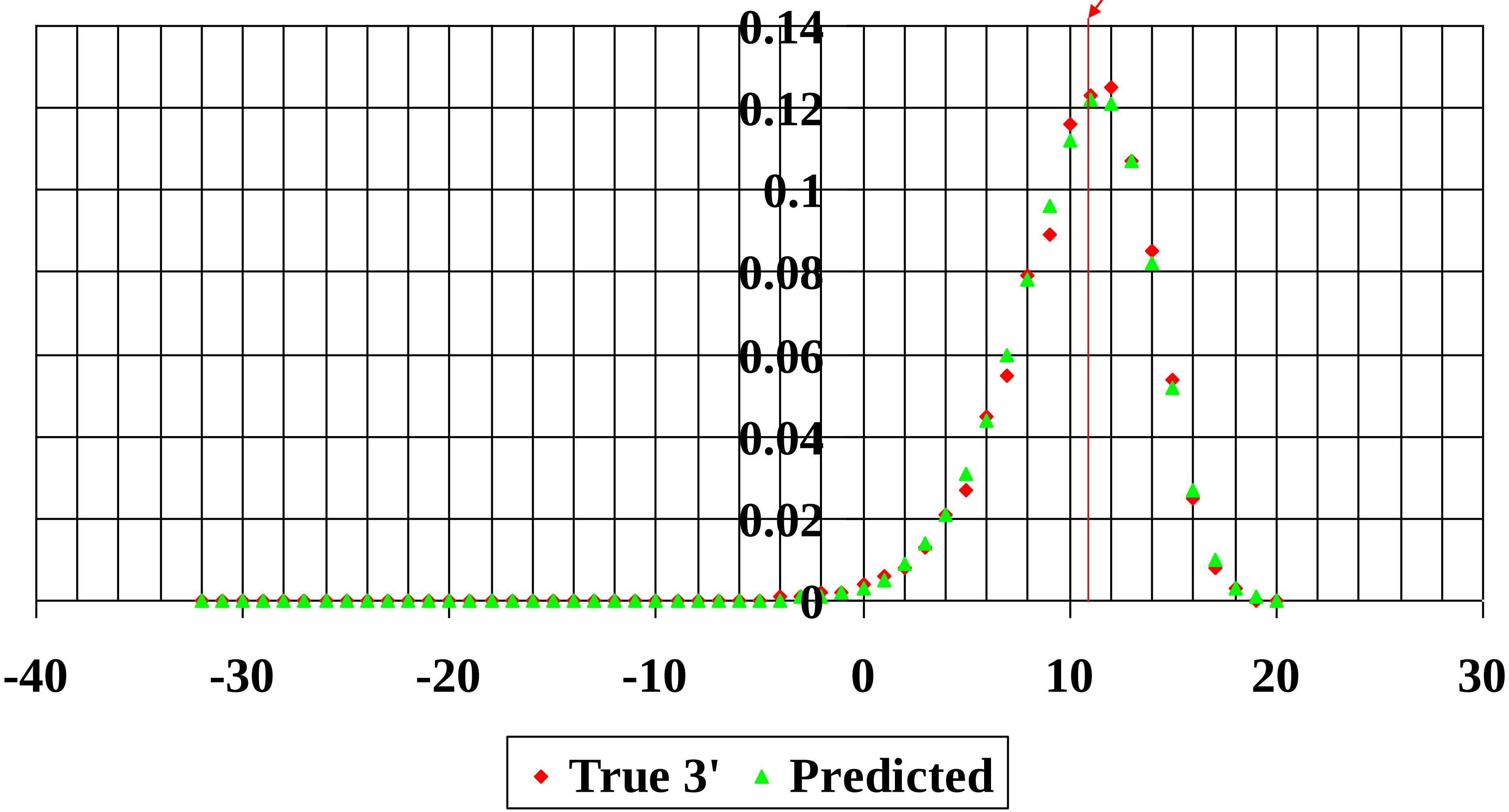
Position-Specific Relative Entropy: 3’ Splice Sites



- Note that $D(P \parallel Q)$ is the *mean* of site score distribution

  i.e. the sum, over sequences, of prob of seq times its LLR score.

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


Relative entropy: 10.85 bits

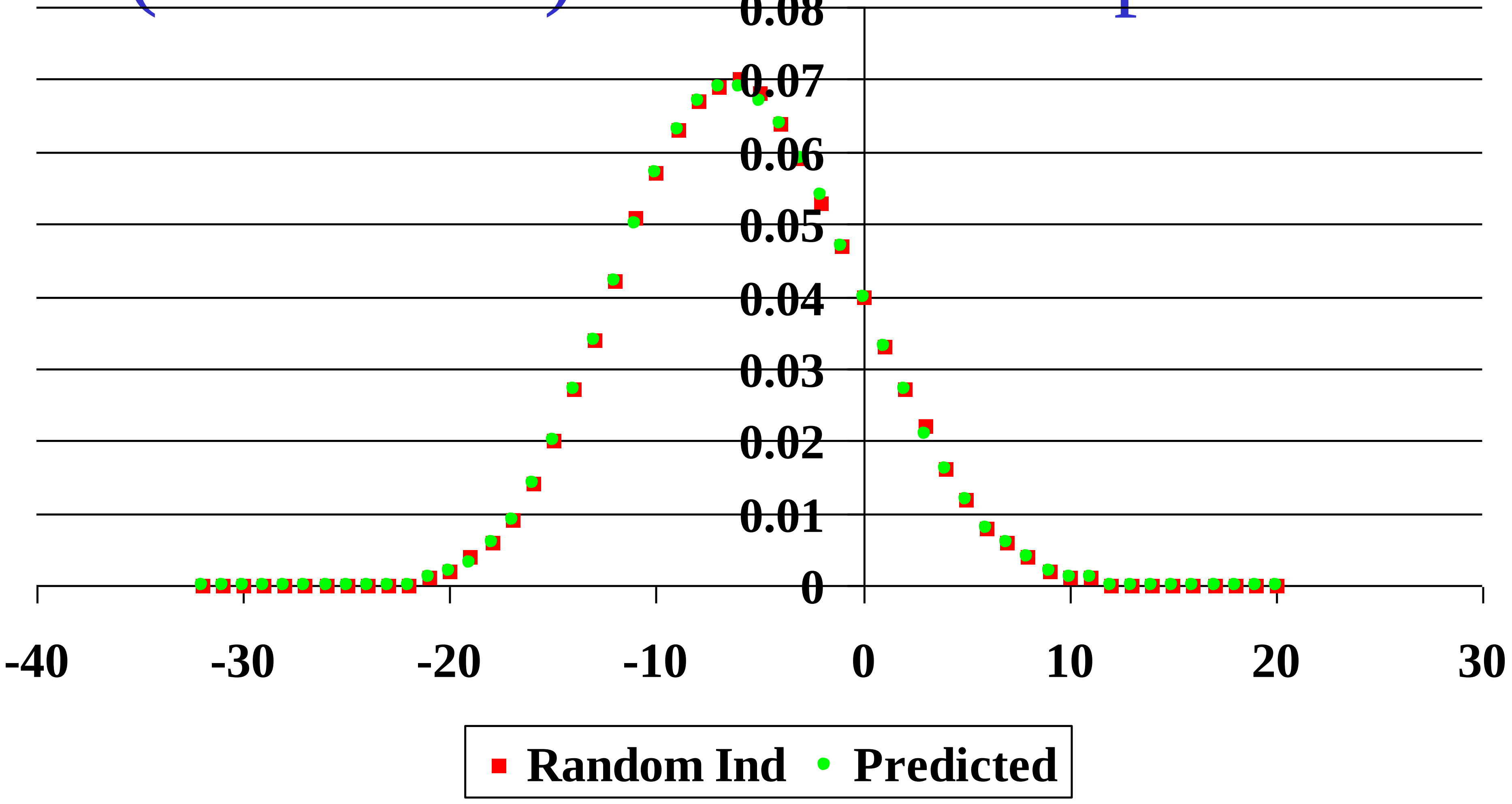
- Similarly,

$$\begin{aligned} D_b(Q \parallel P) &= \Sigma_{s \in S} Q(s) \log_b(Q(s) \,/\, P(s)) \\ &= -\, \Sigma_{s \in S} Q(s) \log_b(P(s) \,/\, Q(s)) \end{aligned}$$

  = *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



- 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*.

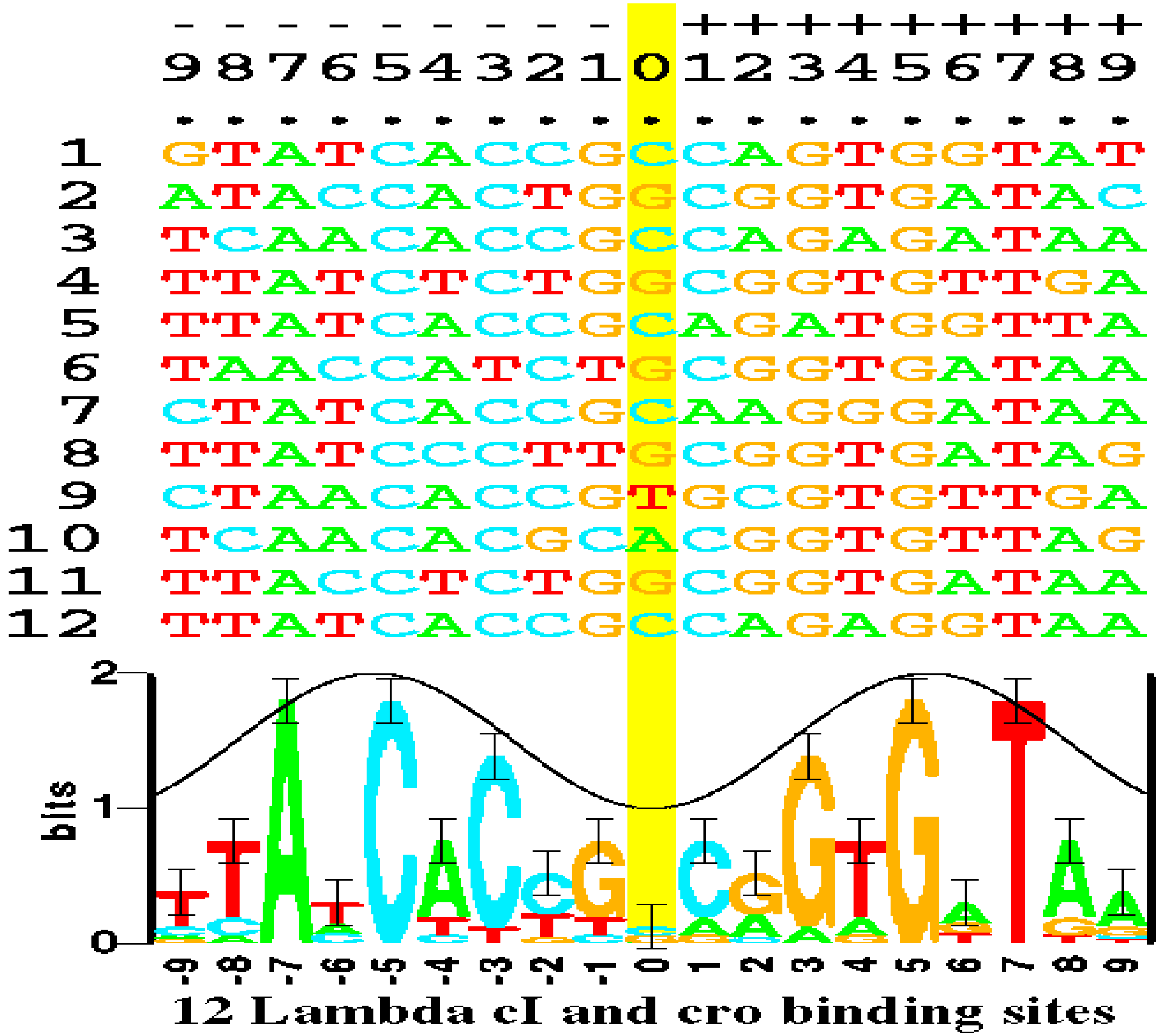
Sequence Logos

- 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 $P_i(r) < Q_i(r)$.

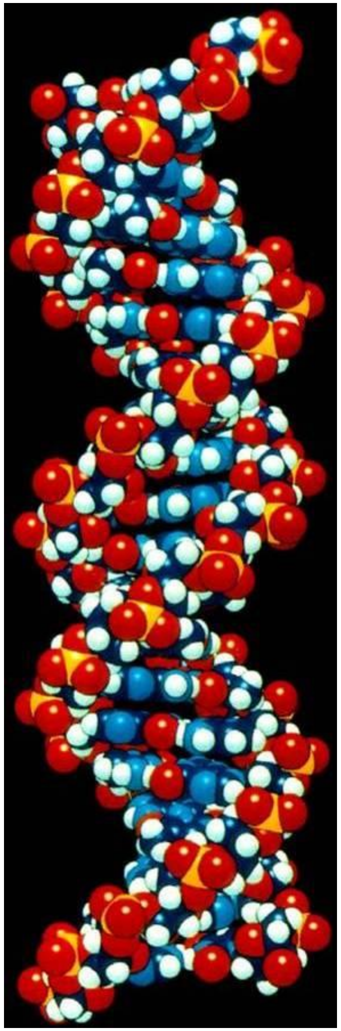
From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

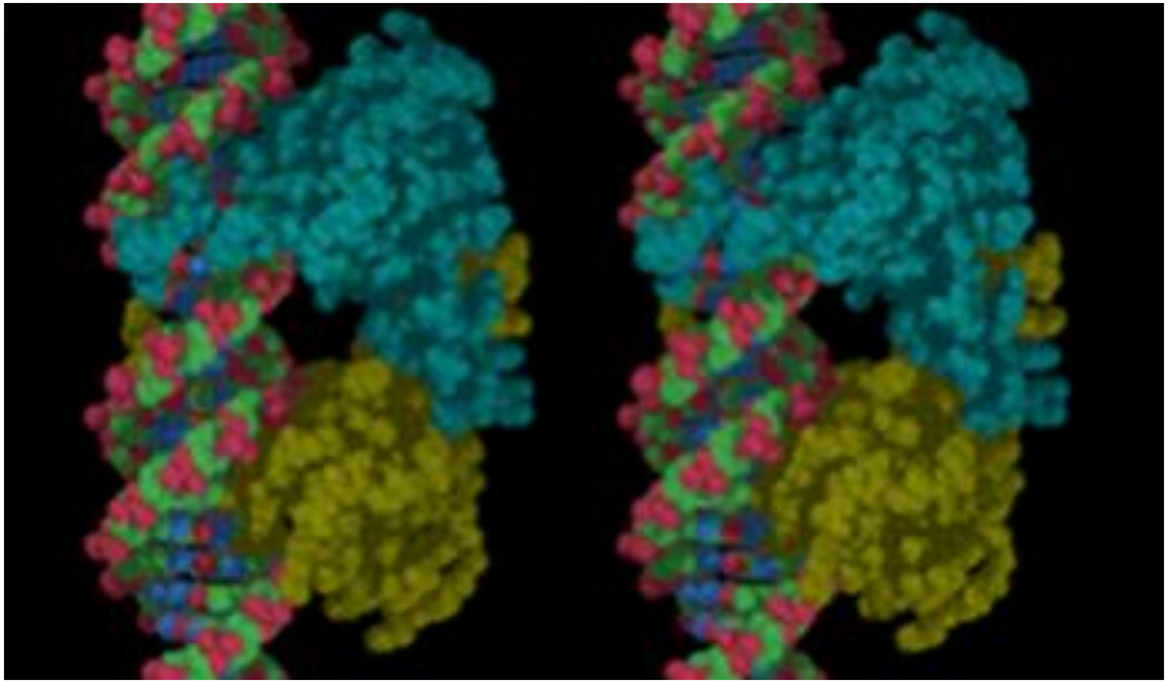


```
     ---------+++++++++
     9876543210123456789
     ...................
 1   GTATCACCGCCAGTGGTAT
 2   ATACCACTGGCGGTGATAC
 3   TCAACACCGCCAGAGATAA
 4   TTATCTCTGGCGGTGTTGA
 5   TTATCACCGCAGATGGTTA
 6   TAACCATCTGCGGTGATAA
 7   CTATCACCGCAAGGGATAA
 8   TTATCCCTTGCGGTGATAG
 9   CTAACACCGTGCGTGTTGA
10   TCAACACGCACGGTGTTAG
11   TTACCTCTGGCGGTGATAA
12   TTATCACCGCCAGAGGTAA
```

12 Lambda cI and cro binding sites

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].

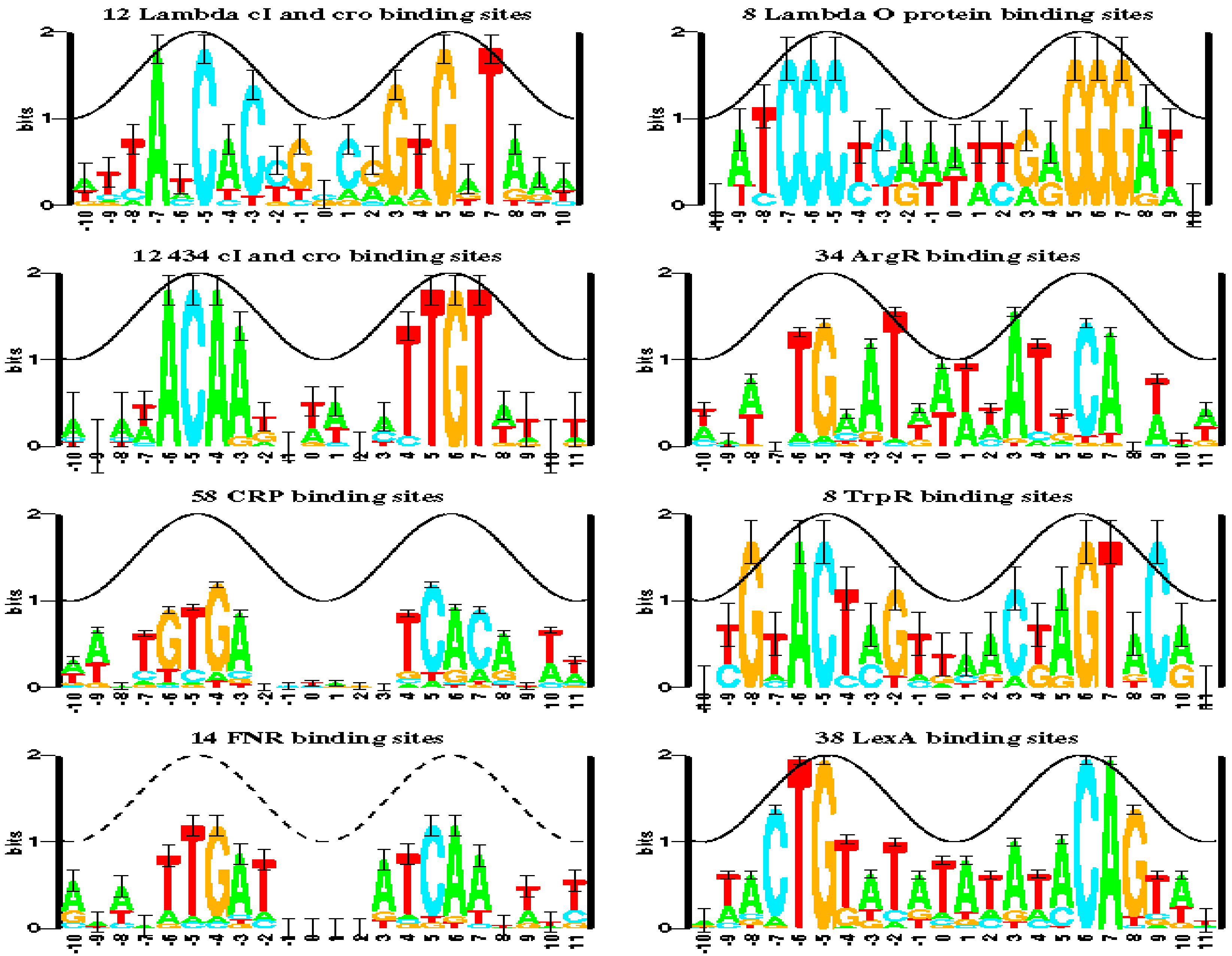




from http://gibk26.bse.kyutech.ac.jp

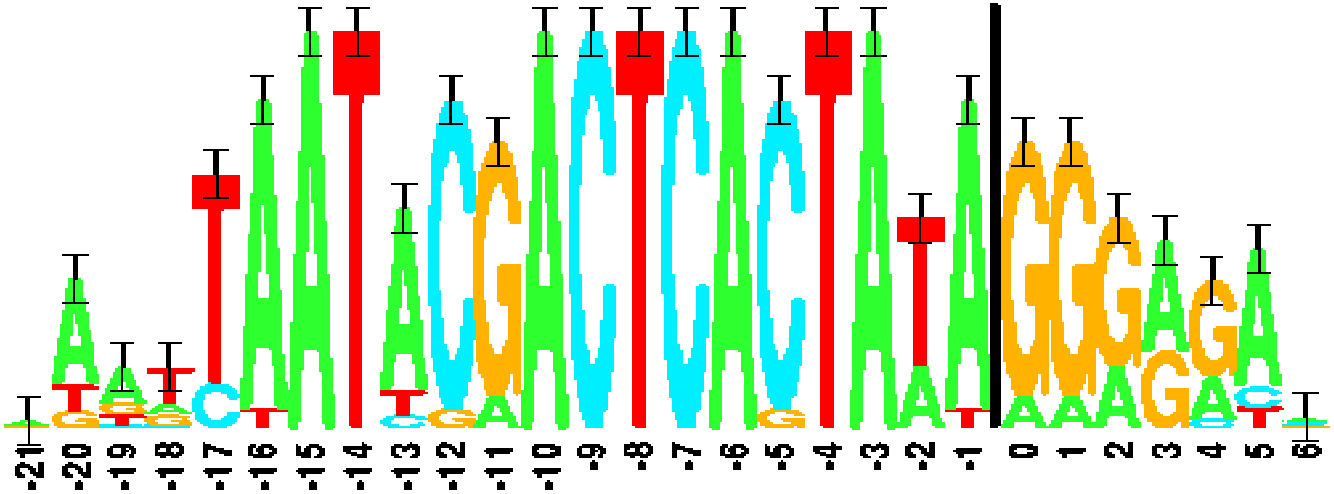
from http://www.dna-dna.net/

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

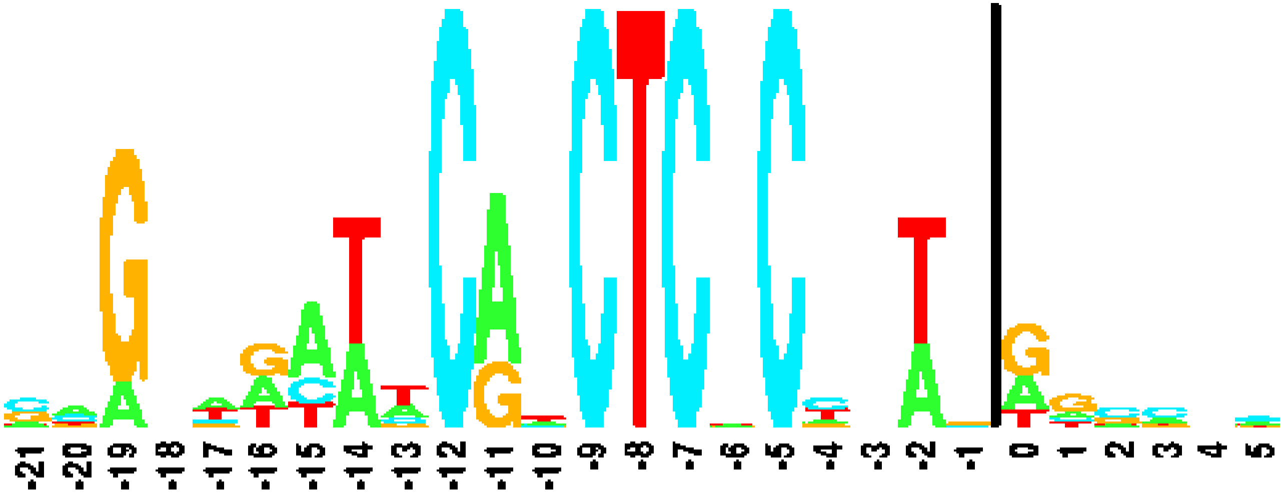


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

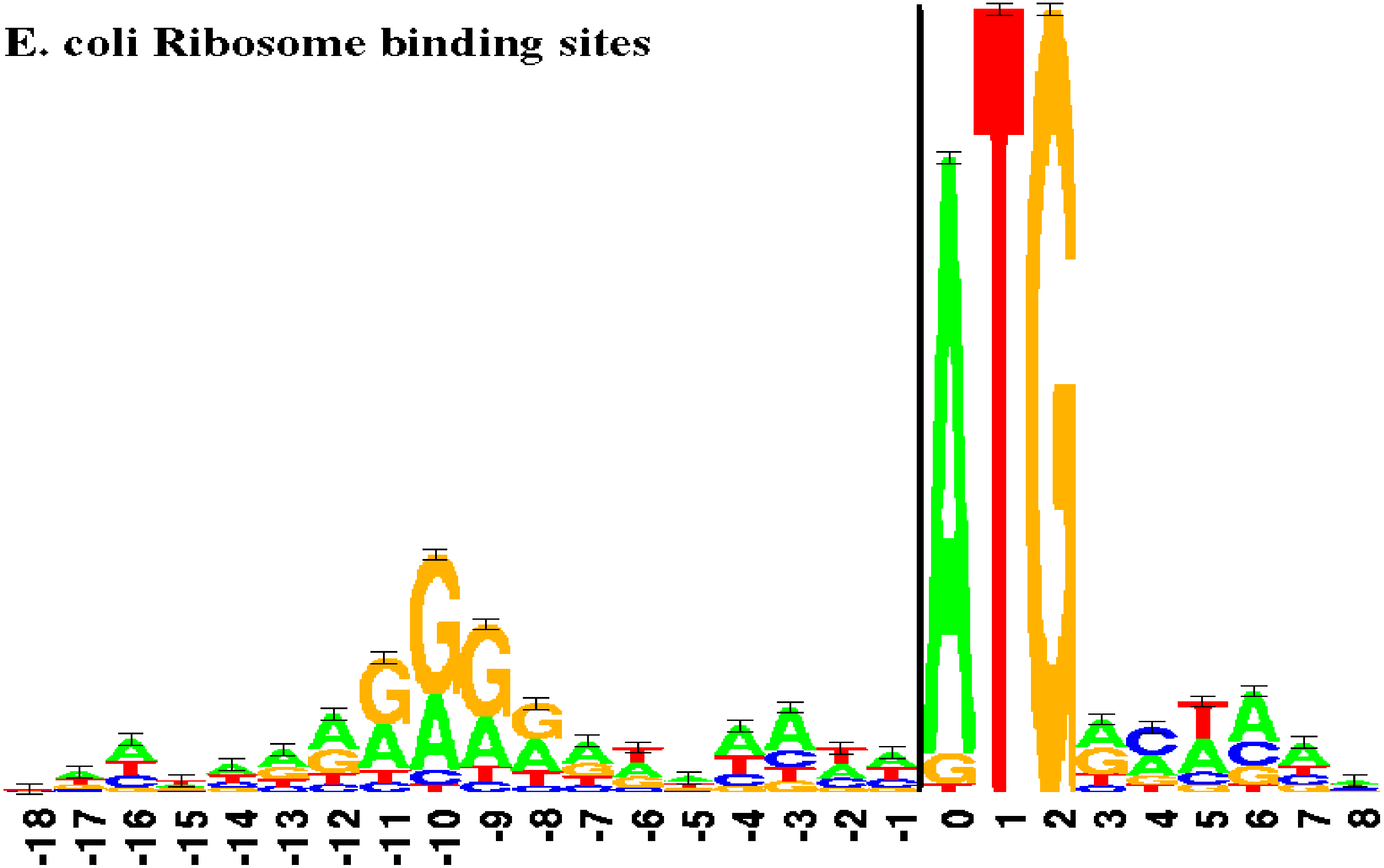
Pattern at T7 RNA polymerase binding sites



Pattern required by T7 RNA polymerase to function

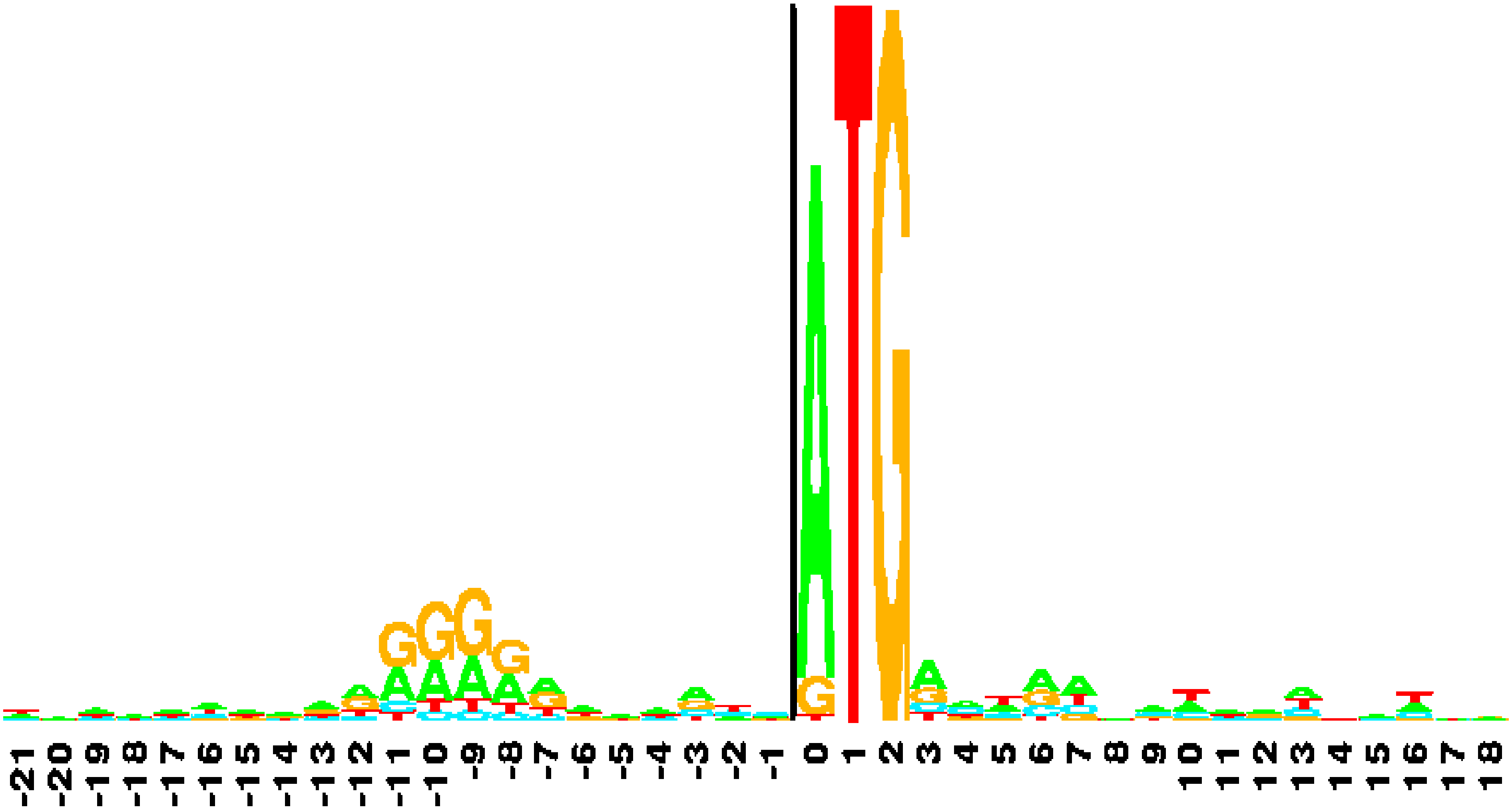


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



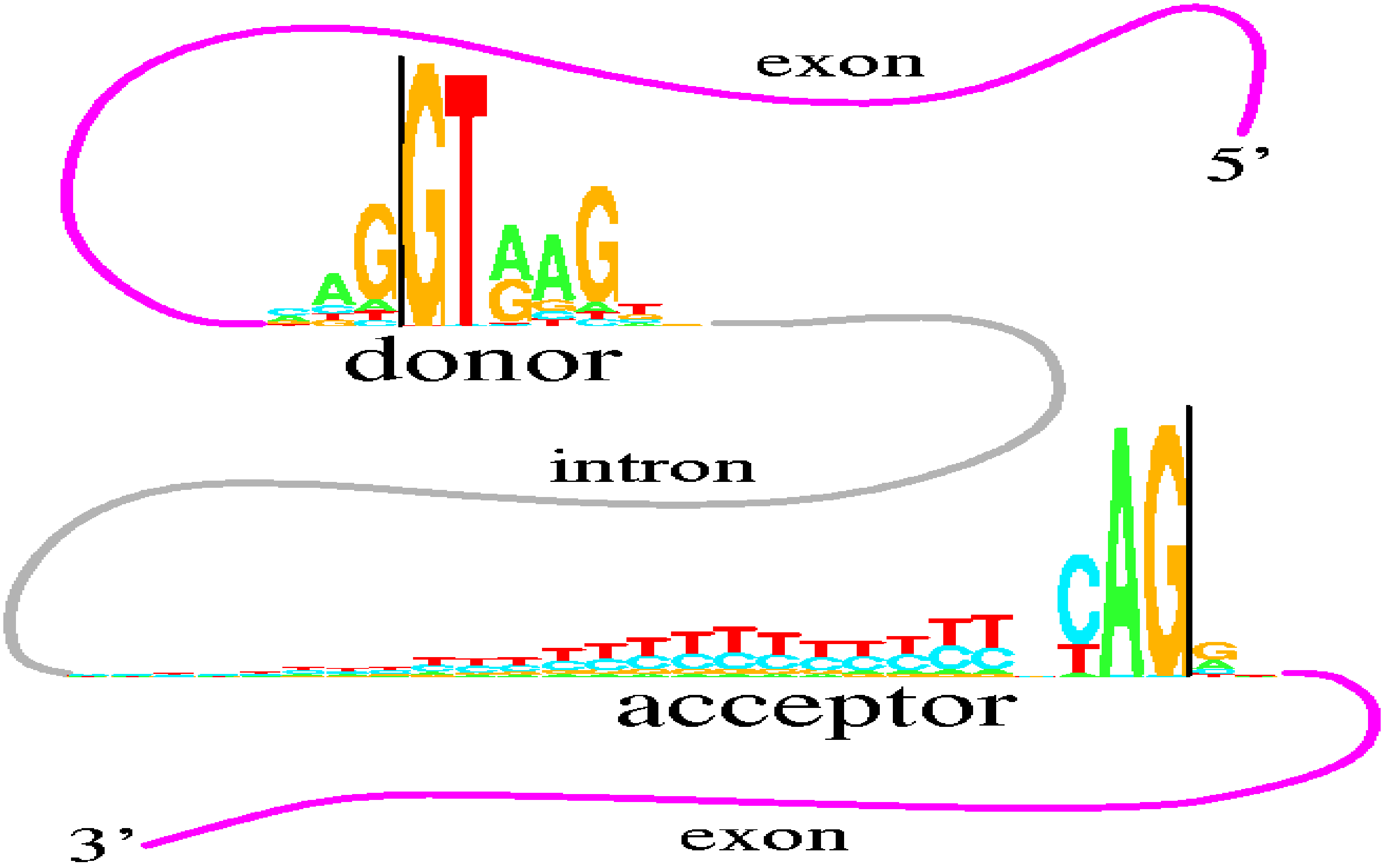
From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

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

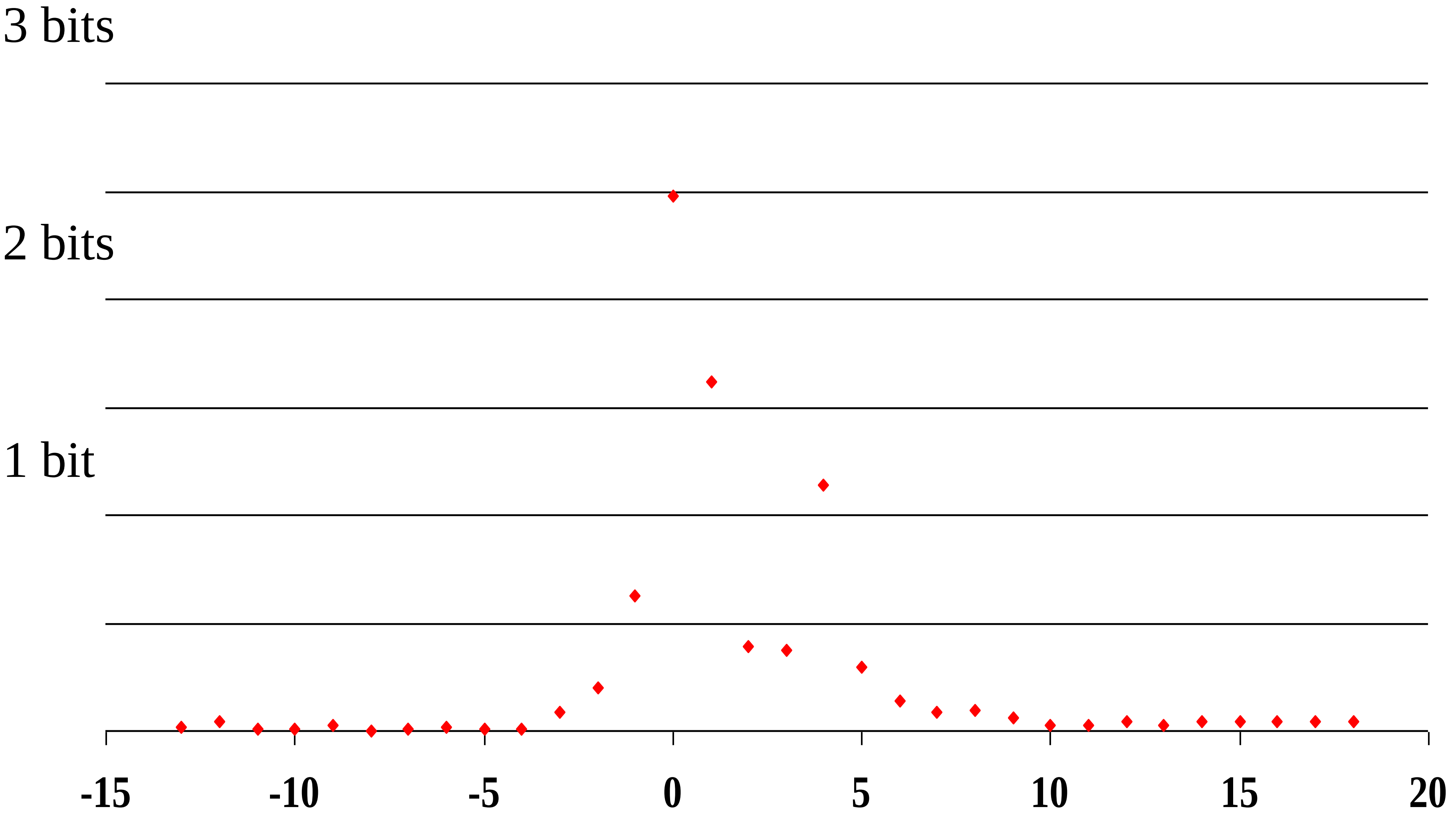


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

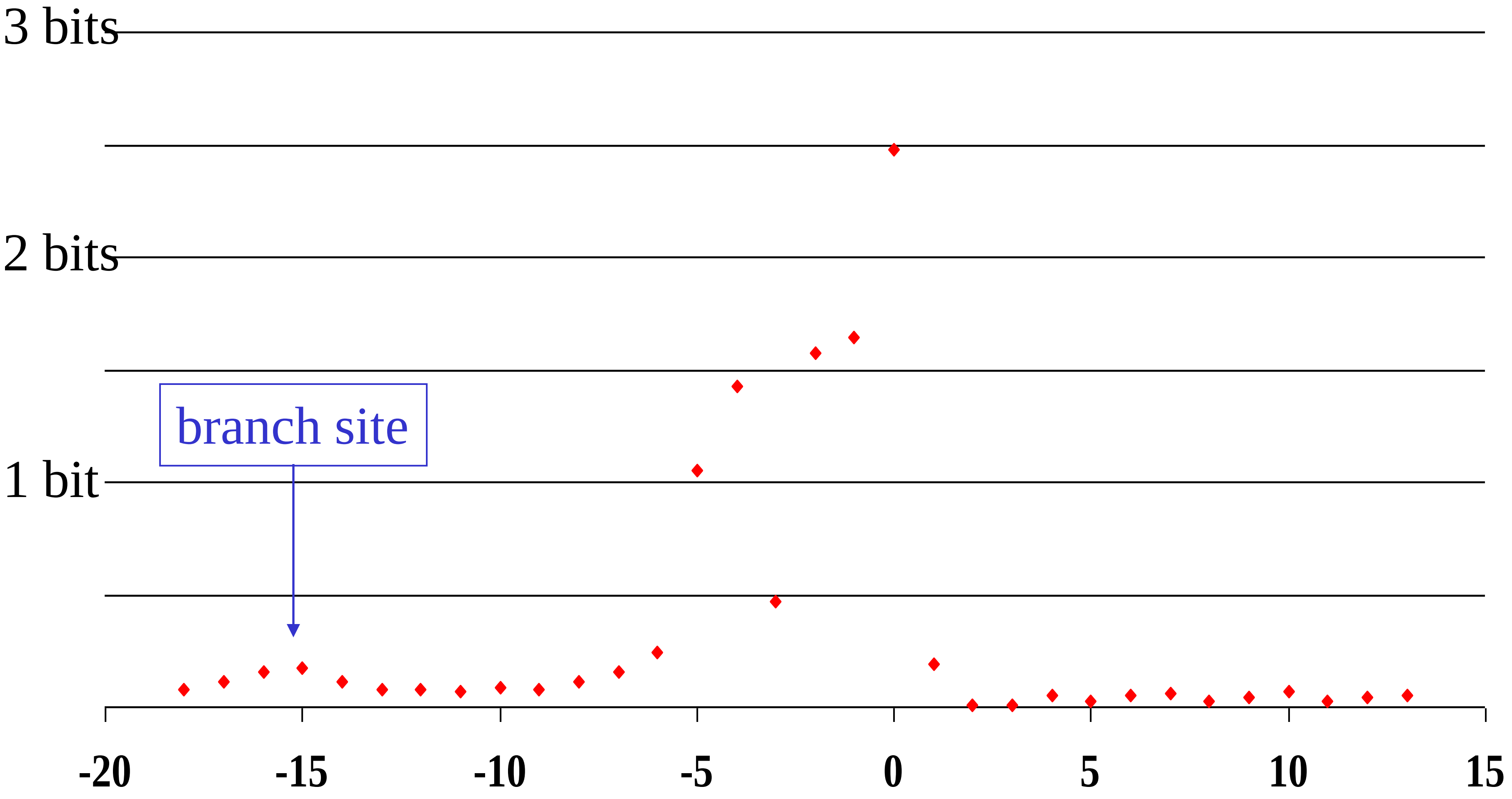
This figure shows two "sequence logos" which represent sequence conservation at the 5'(donor) and 3'(acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides 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)



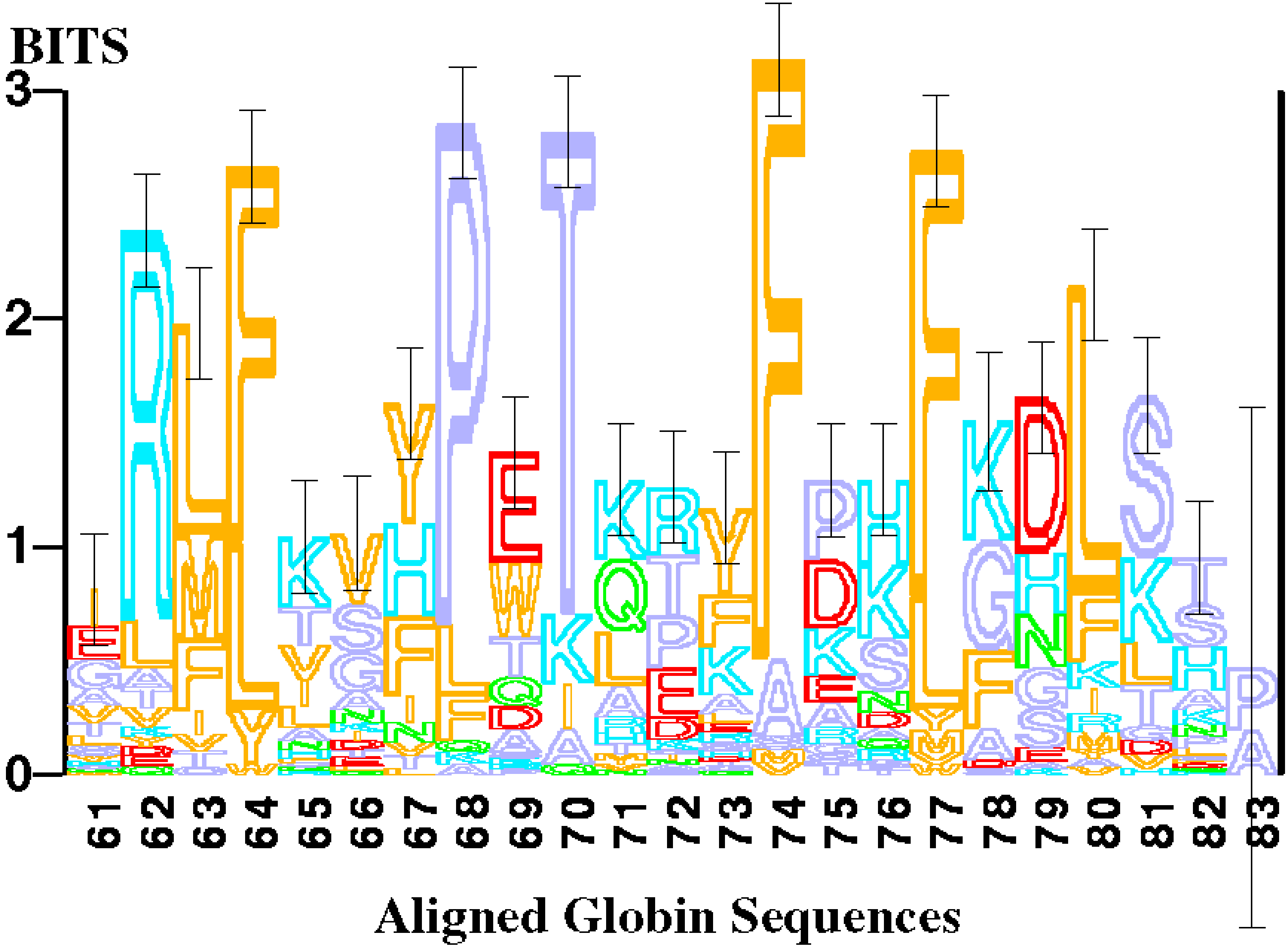
Position-Specific Relative Entropy: *C. elegans* 5’ Splice Sites



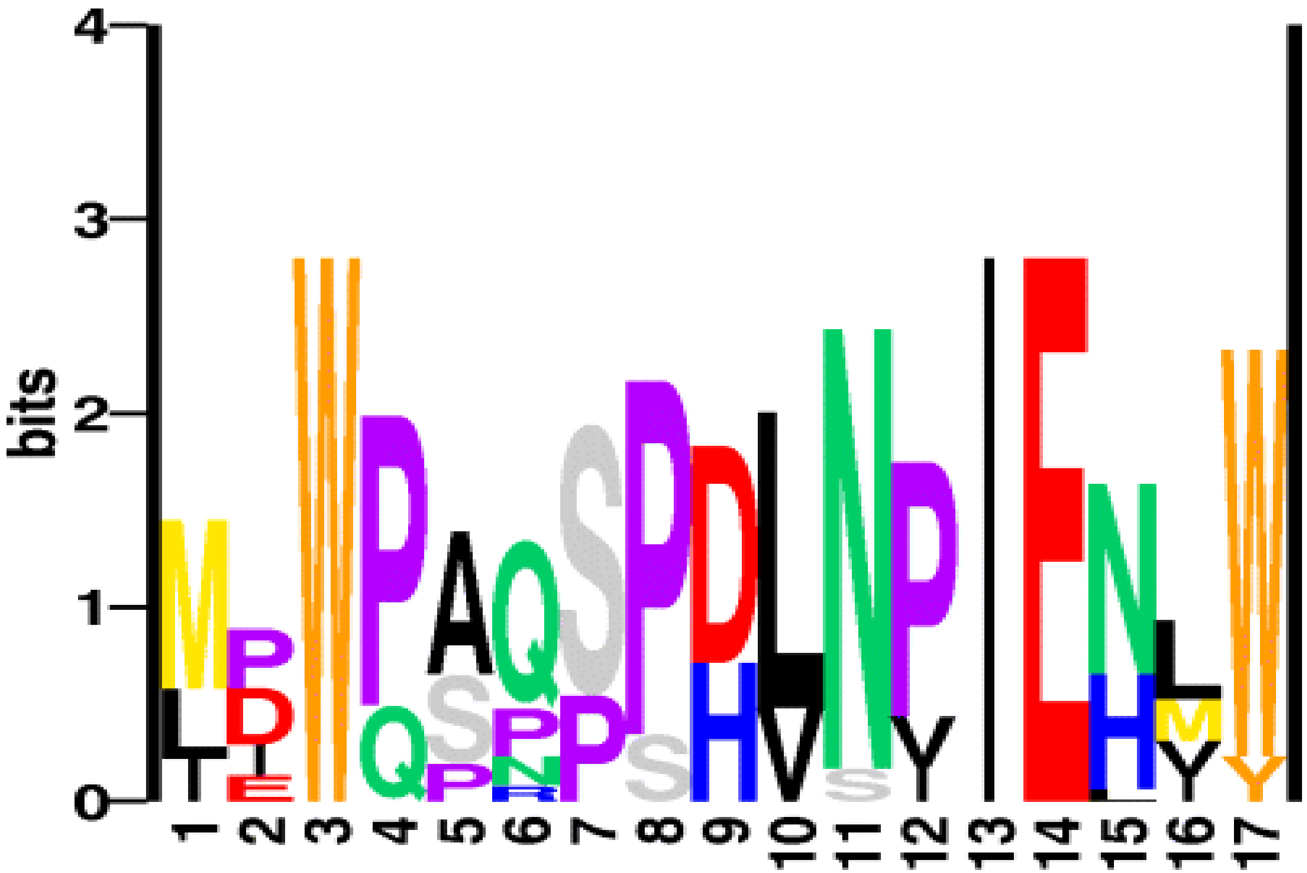
Position-Specific Relative Entropy: 3’ Splice Sites



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