Today's Lecture

- Sequence logos
- Limitations of site models
  - Gaps
  - Failure of independence assumption

Sequence Logos

- Schneider and Stephens (NAR 18, 6097-6100, 1990)– see

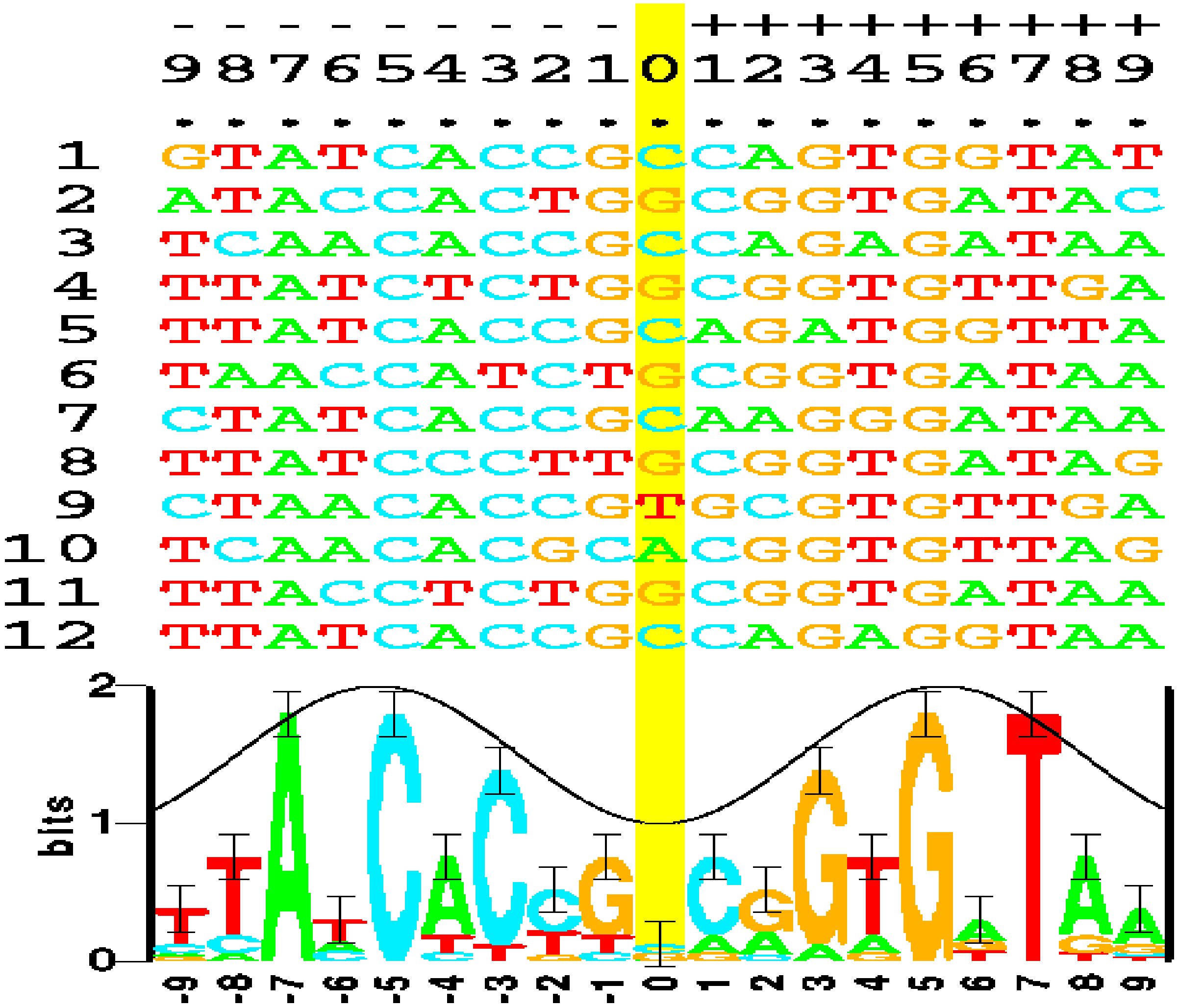
  http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

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

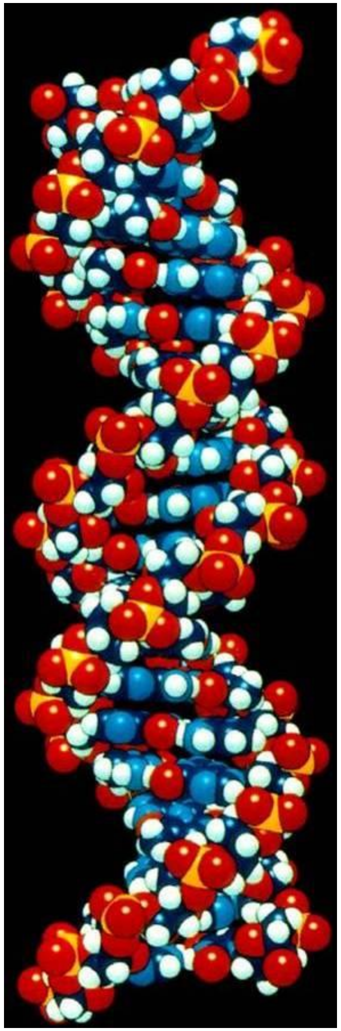
From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

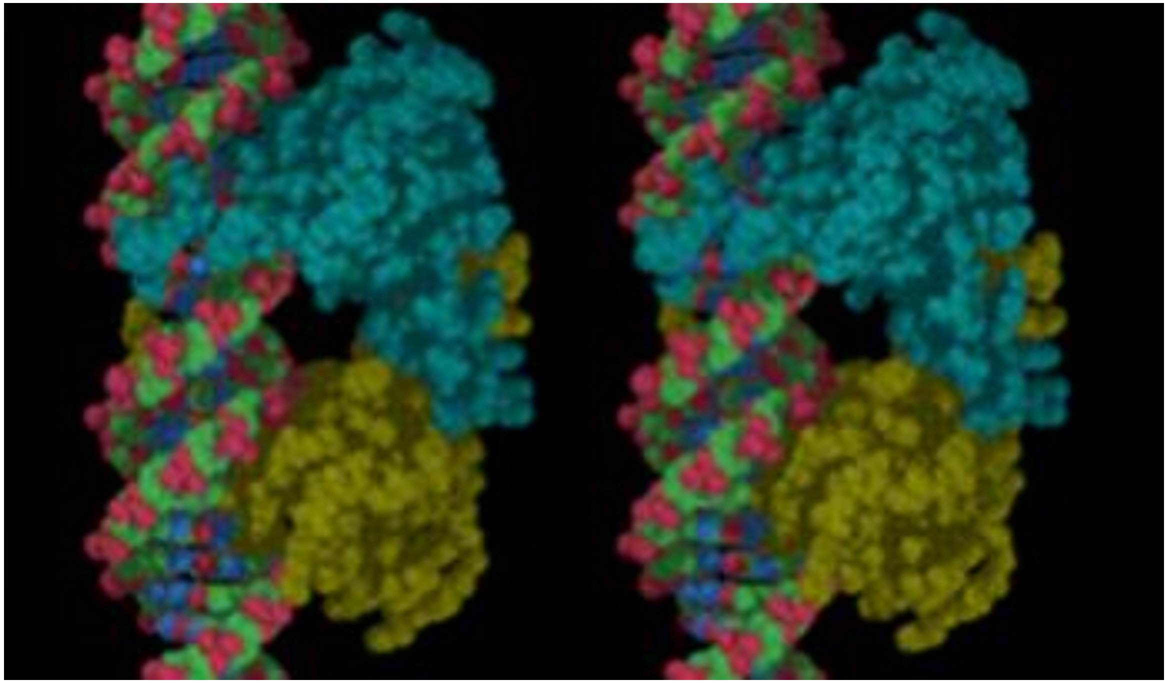


```
    ---------+++++++++
    9876543210123456789
    ...................
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG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

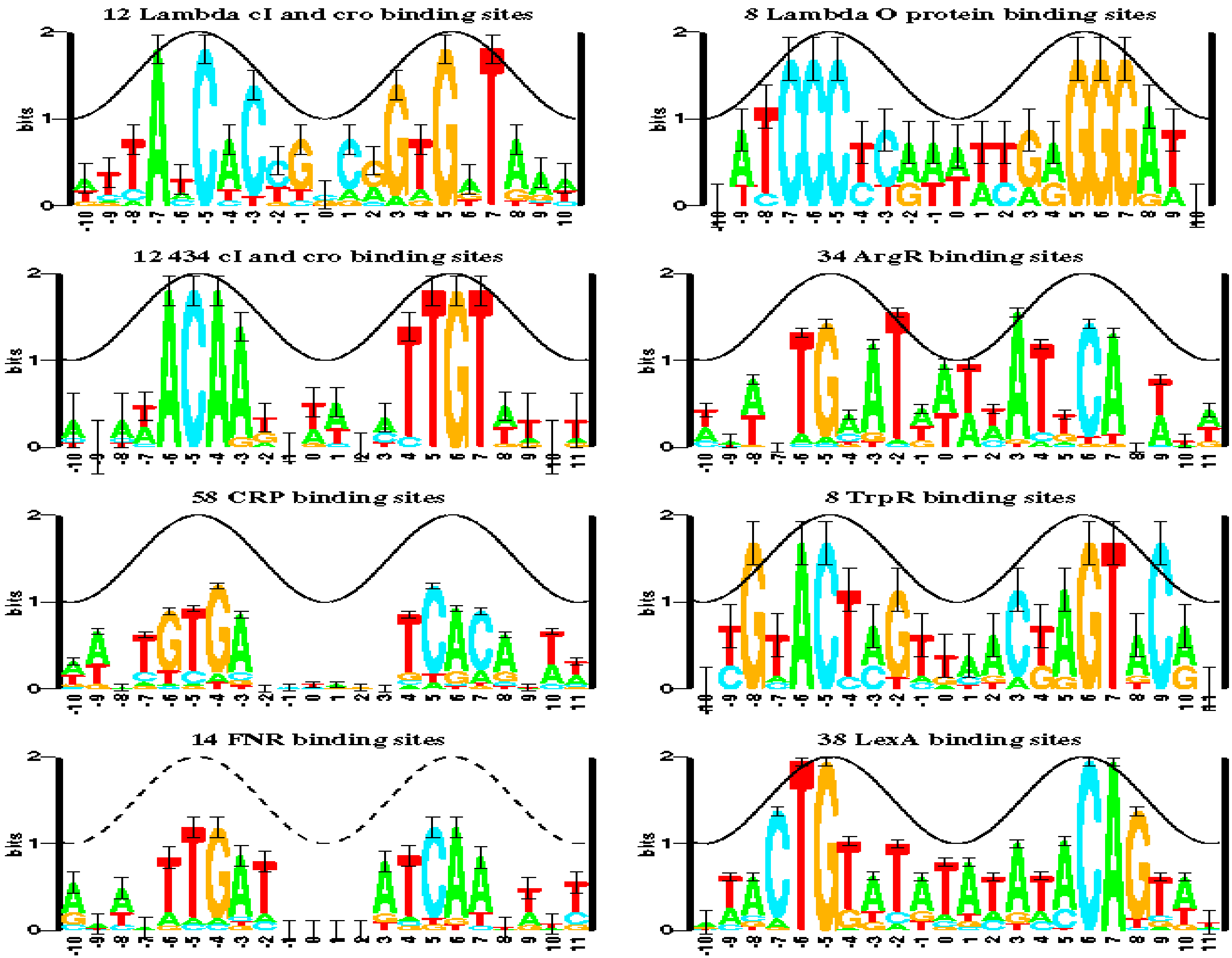




from http://gibk26.bse.kyutech.ac.jp

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

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



12 Lambda cI and cro binding sites

8 Lambda O protein binding sites

12 434 cI and cro binding sites

34 ArgR binding sites

58 CRP binding sites

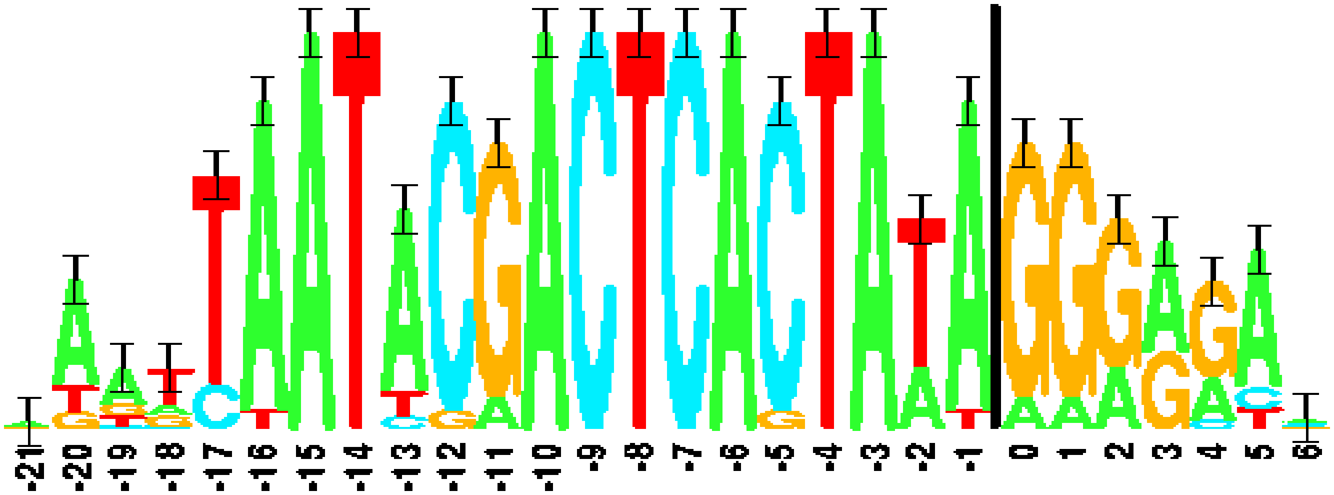
8 TrpR binding sites

14 FNR binding sites

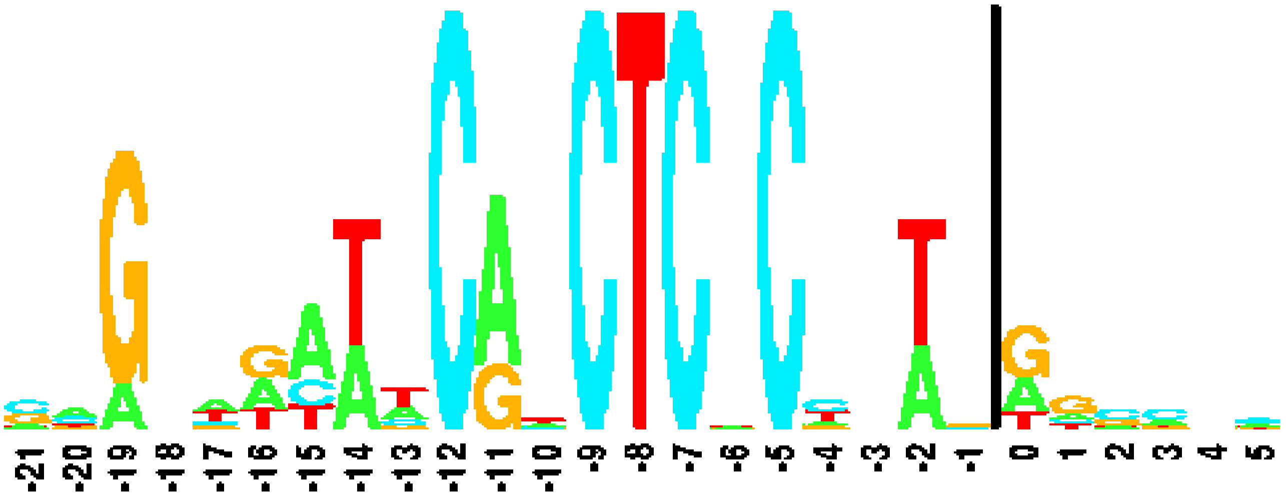
38 LexA binding sites

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

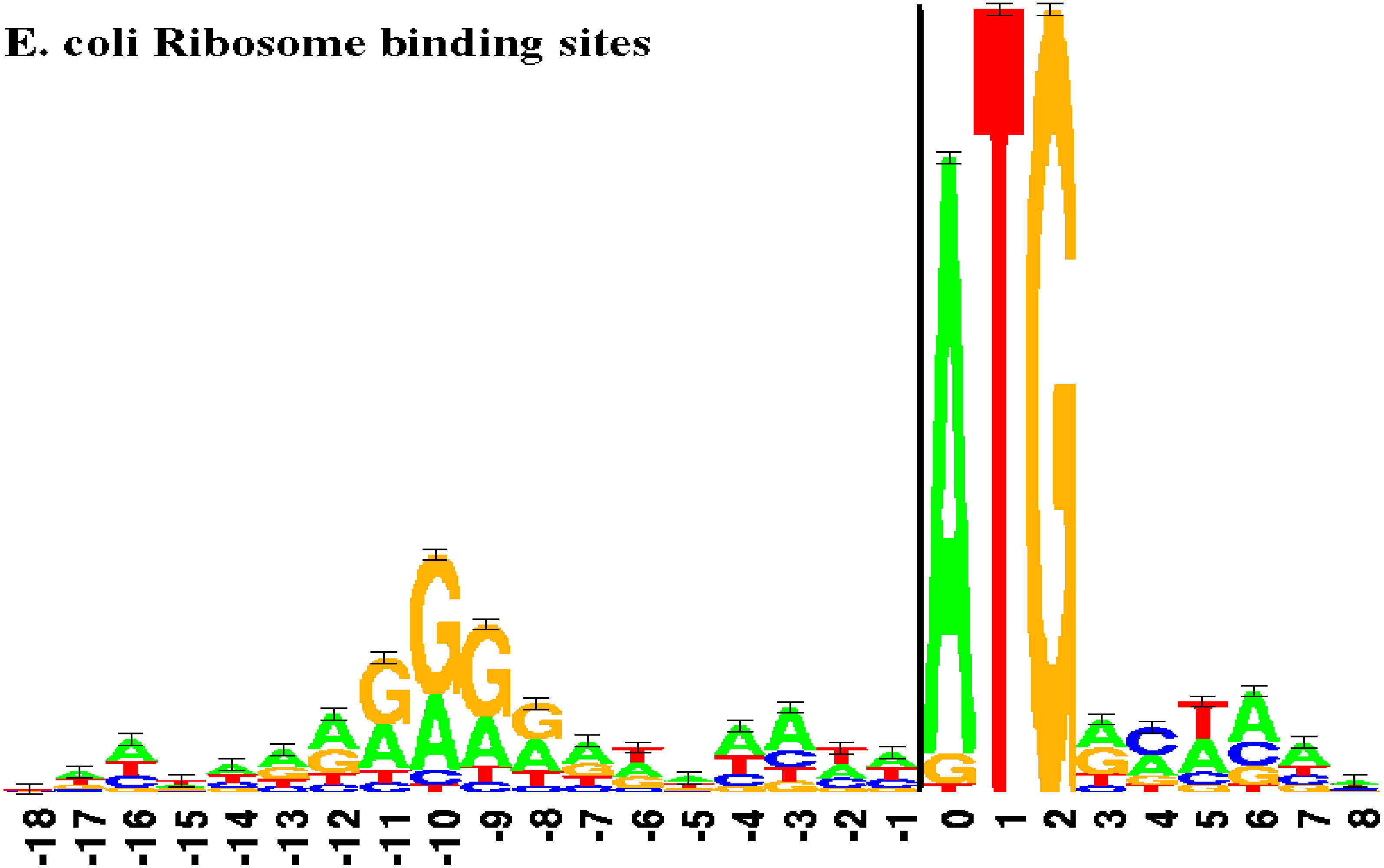
Pattern at T7 RNA polymerase binding sites



Pattern required by T7 RNA polymerase to function

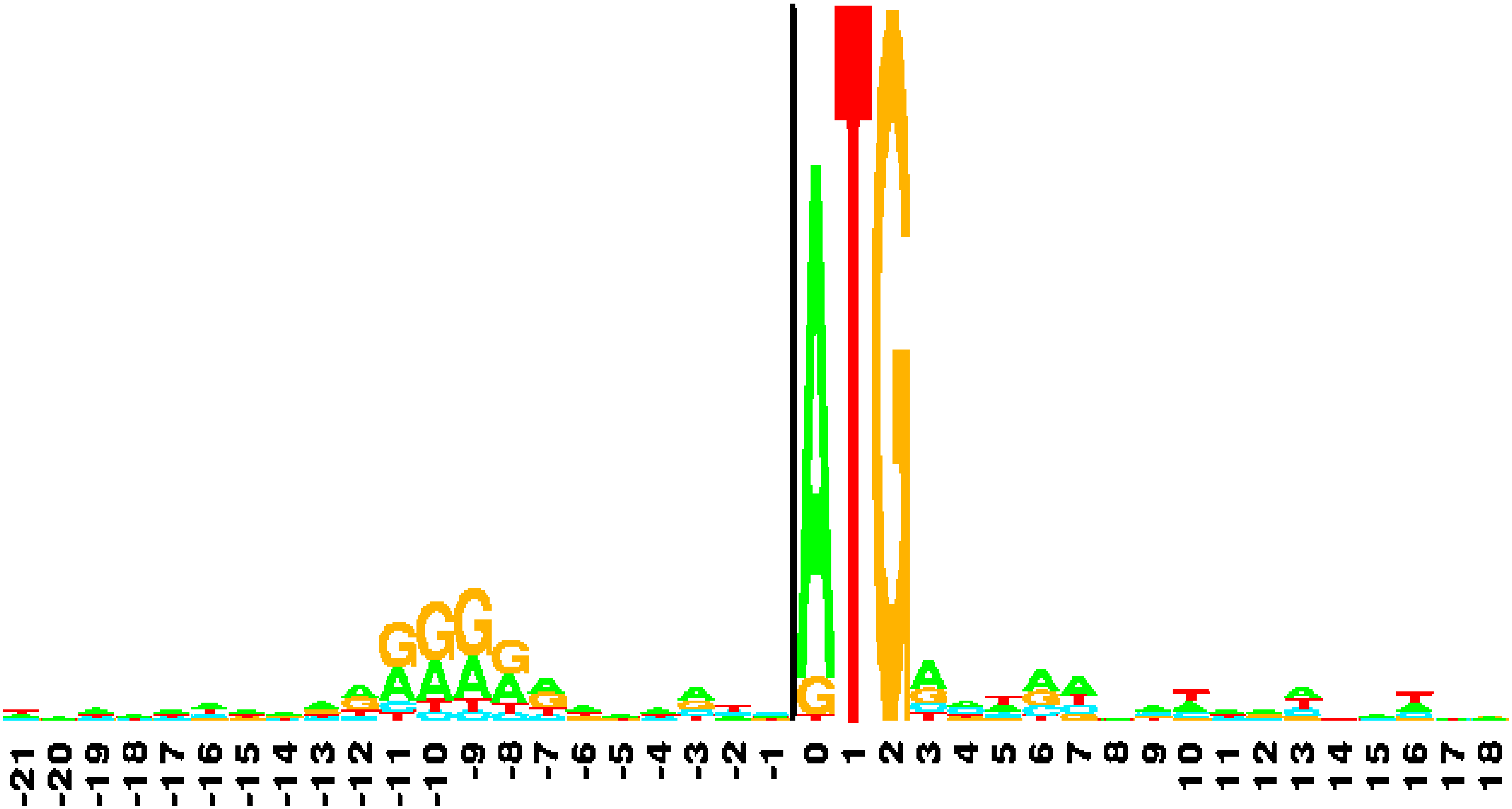


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



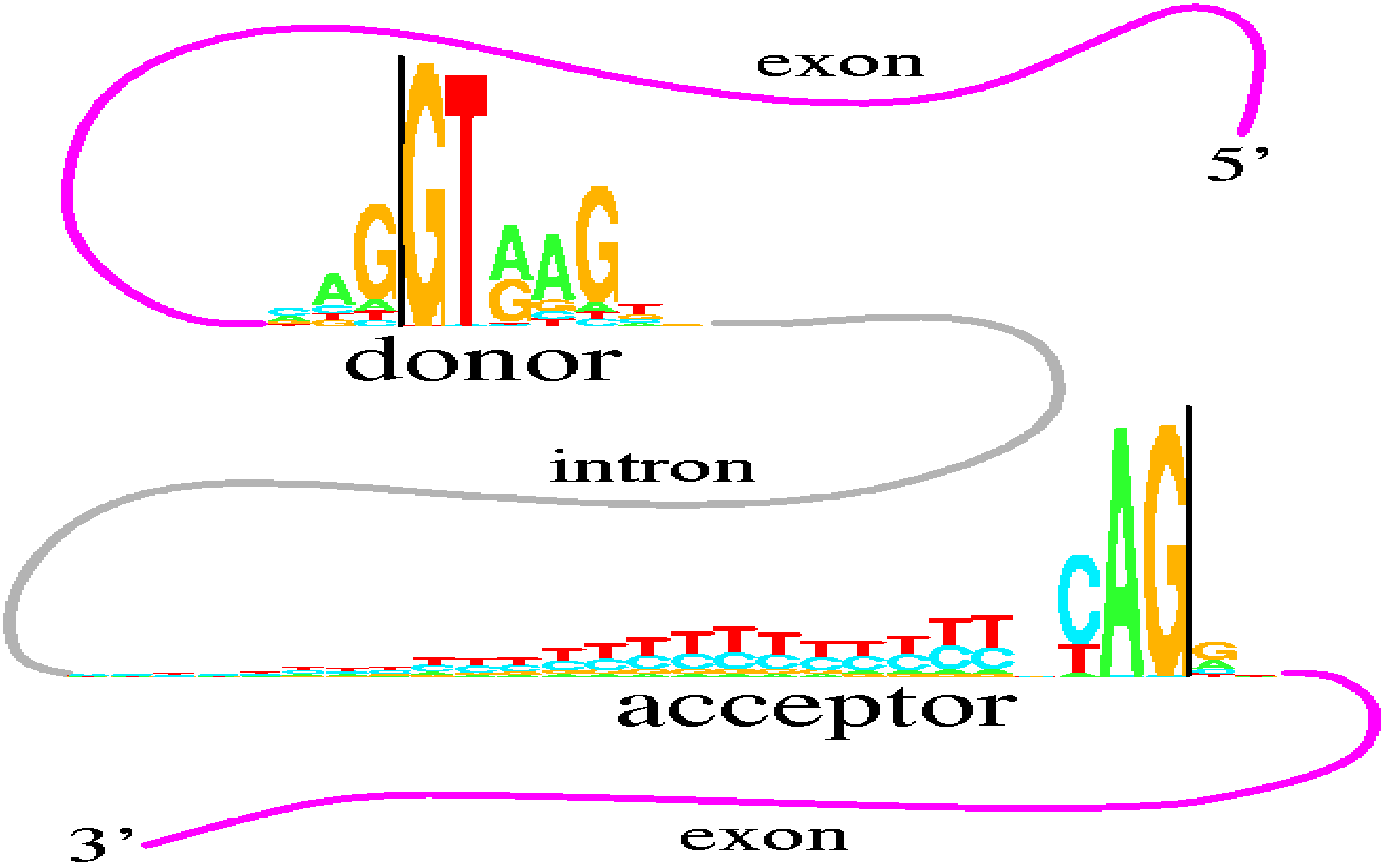
From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

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

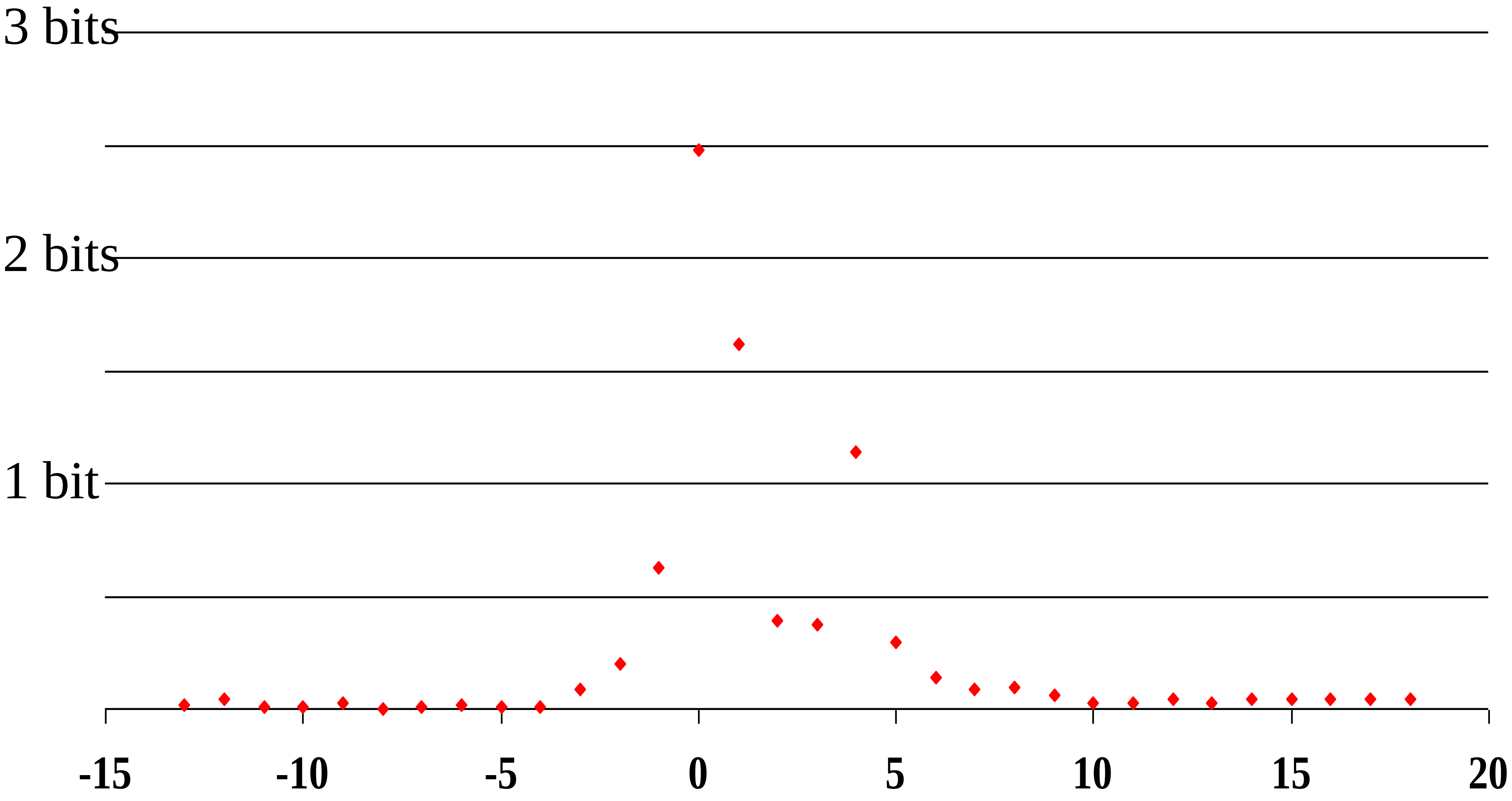


From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

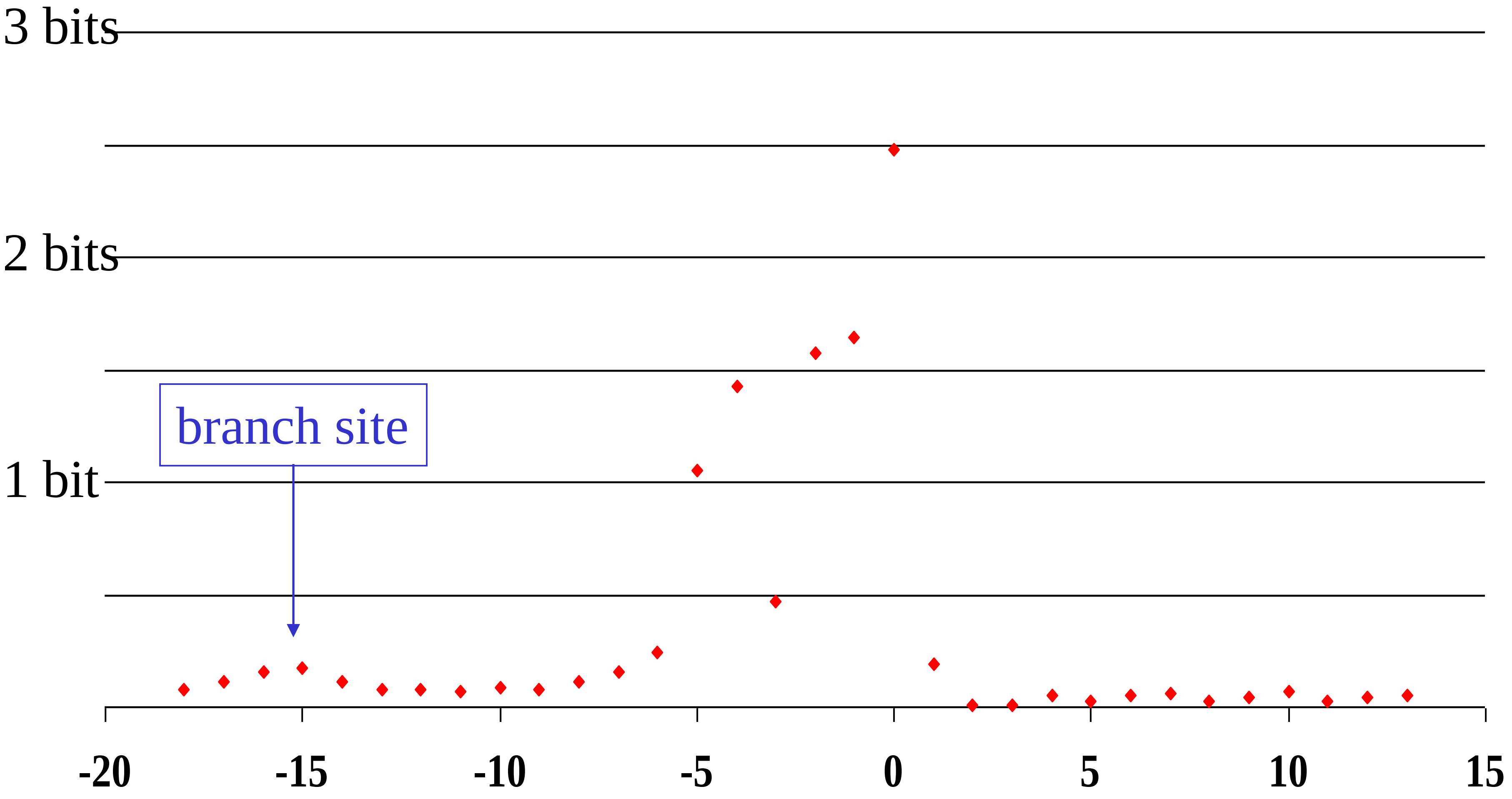
This figure shows two “sequence logos” which represent sequence conservation at the 5’(donor) and 3’(acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides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



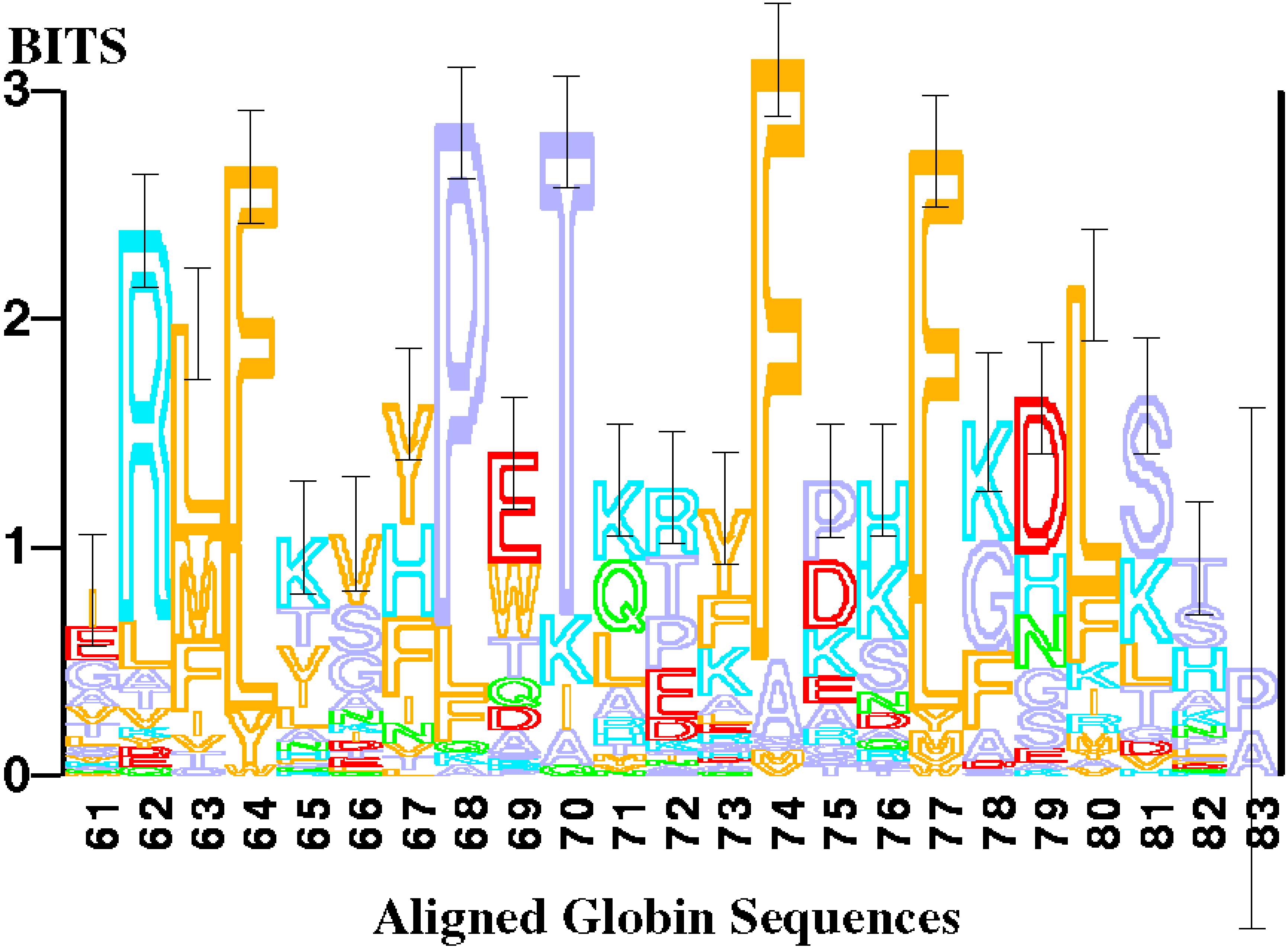
Position-Specific Relative Entropy: *C. elegans* 5’ Splice Sites



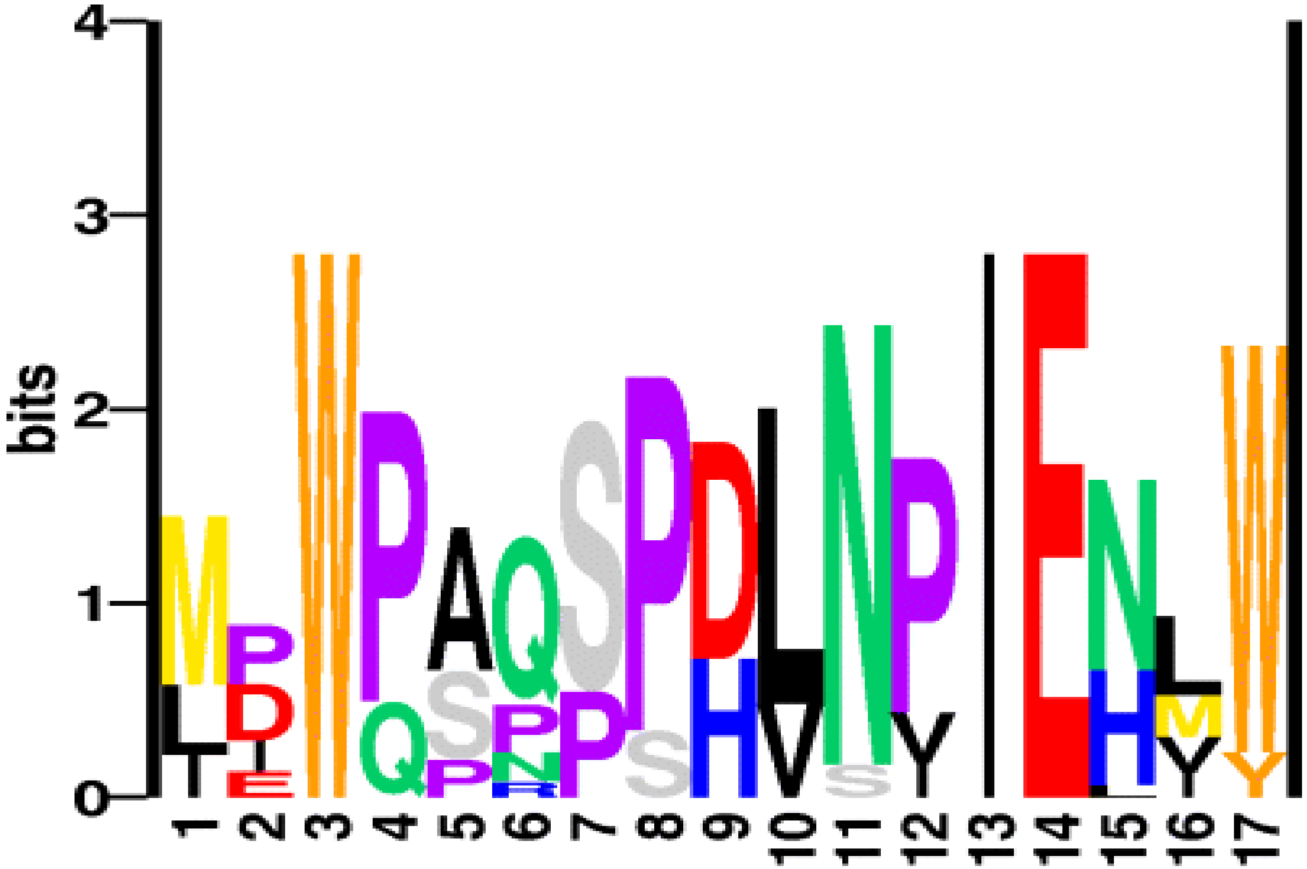
Position-Specific Relative Entropy: 3’ Splice Sites



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



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

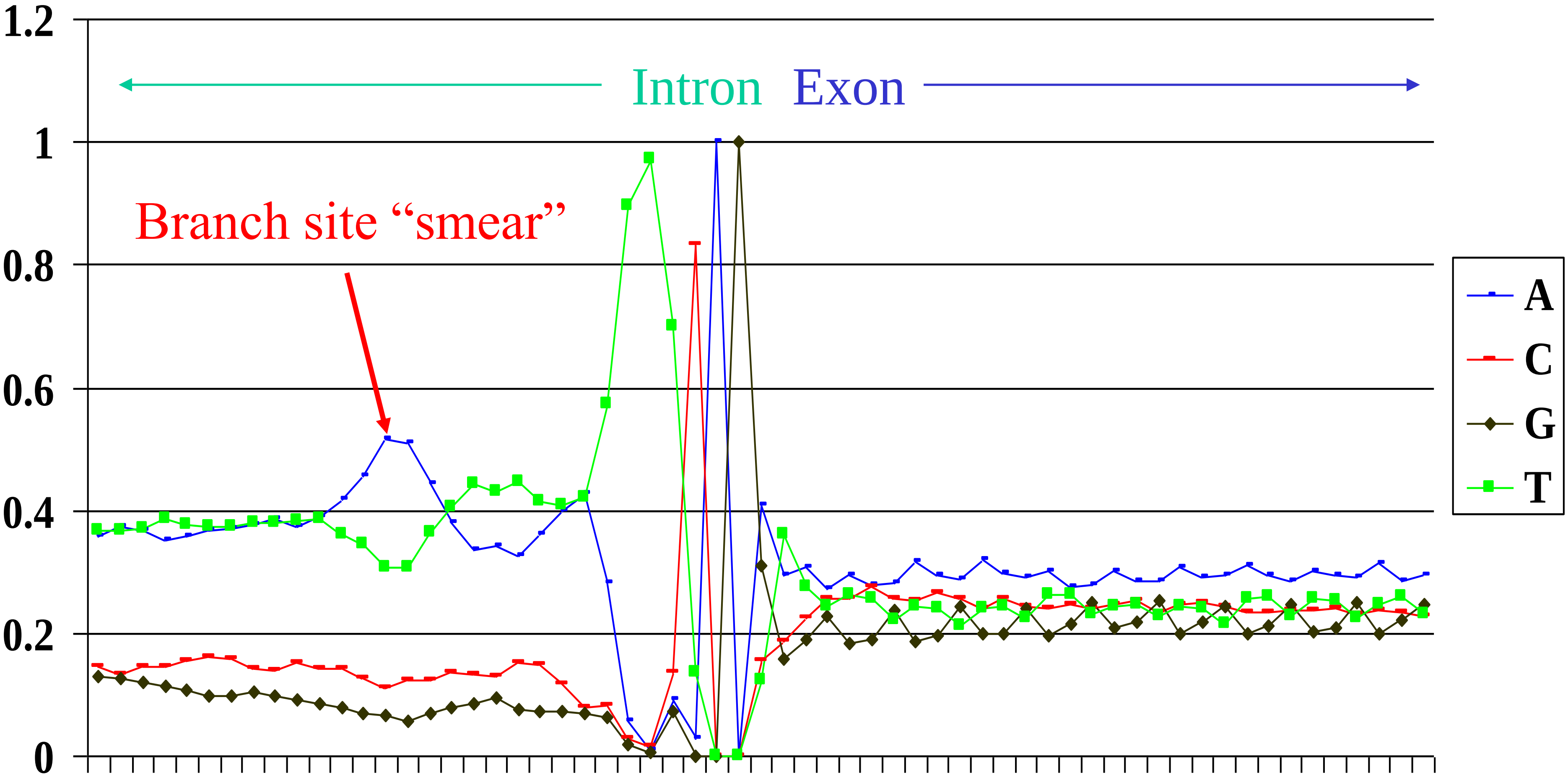


Logo of Gibbs Block D (Tc1) 9 sequences

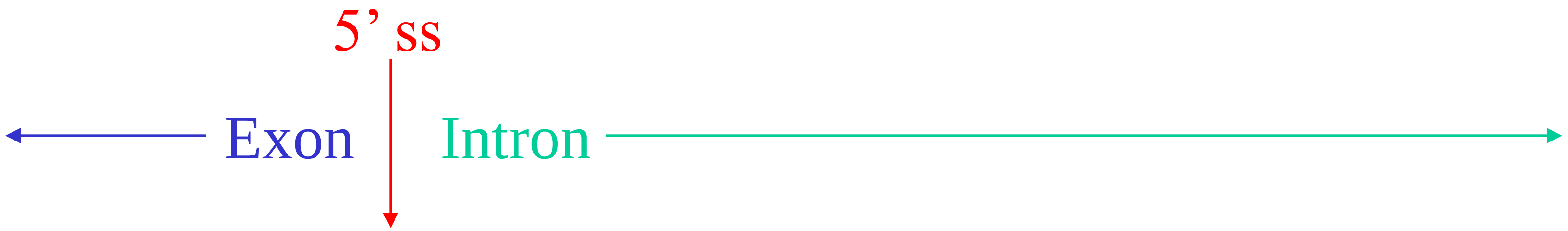
Limitations of Site Models

- Failure to allow indels means variably spaced subelements are “smeared”, e.g.:
  - branch site, for 3’ splice sites;
  - coding sequence, for both 3’ and 5’ sites
    - not really an indel issue -- could make reading-frame-specific matrices
- Independence assumption
  - usually OK for protein sequences (after correcting for evolutionary relatedness)
  - often fails for nucleotide sequences: examples:
    - 5’ sites (Burge-Karlin observation);
    - background (dinucleotide correlation)

3' Splice Sites – *C. elegans*



Nucleotide Counts for 8192 *C. elegans* 5’ Splice Sites



A	3404	4644	1518	0	0	4836	5486	837	1632	2189	2278	2355
C	1850	1224	583	0	14	118	588	237	801	771	889	986
G	1562	912	4891	8192	0	1890	672	6164	589	962	1056	827
T	1376	1412	1200	0	8178	1348	1446	954	5170	4270	3969	4024

CONSENSUS x a g G T a a g t t w t

A	0.416	0.567	0.185	0.000	0.000	0.590	0.670	0.102	0.199	0.267	0.278	0.287
C	0.226	0.149	0.071	0.000	0.002	0.014	0.072	0.029	0.098	0.094	0.109	0.120
G	0.191	0.111	0.597	1.000	0.000	0.231	0.082	0.752	0.072	0.117	0.129	0.101
T	0.168	0.172	0.146	0.000	0.998	0.165	0.177	0.116	0.631	0.521	0.484	0.491

Failure of independence for 5’ splice sites: G vs. H (‘not G’) at position -1

H in position –1 :

A	1434	1664	1518	0	0	2032	2662	98	479	694	783	912
C	633	546	583	0	5	36	177	22	225	250	350	393
G	628	553	0	3301	0	943	187	3063	134	329	405	279
T	606	538	1200	0	3296	290	275	118	2463	2028	1763	1717

A	0.434	0.504	0.460	0.000	0.000	0.616	0.806	0.030	0.145	0.210	0.237	0.276
C	0.192	0.165	0.177	0.000	0.002	0.011	0.054	0.007	0.068	0.076	0.106	0.119
G	0.190	0.168	0.000	1.000	0.000	0.286	0.057	0.928	0.041	0.100	0.123	0.085
T	0.184	0.163	0.364	0.000	0.998	0.088	0.083	0.036	0.746	0.614	0.534	0.520

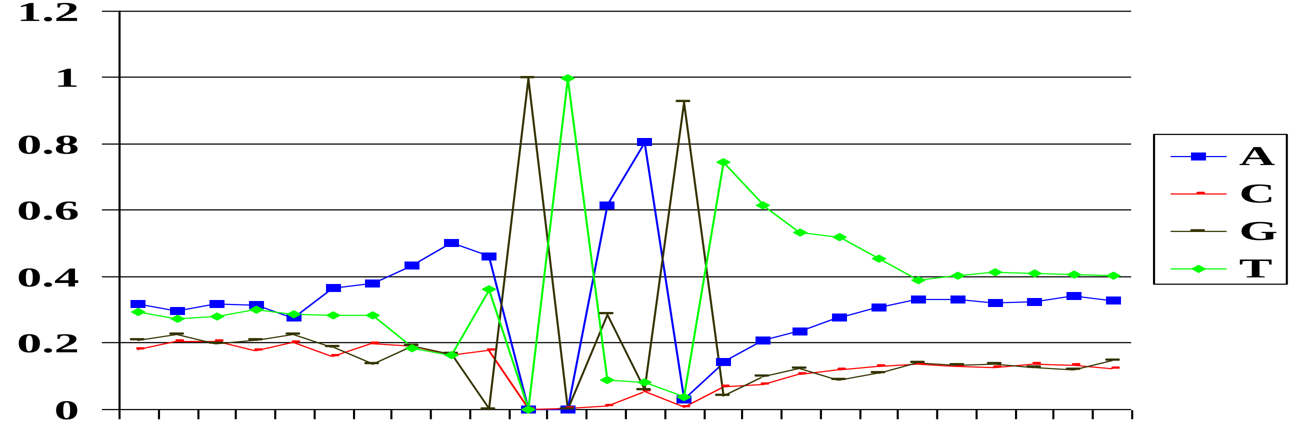
G in position –1 :

A	1970	2980	0	0	0	2804	2824	739	1153	1495	1495	1443
C	1217	678	0	0	9	82	411	215	576	521	539	593
G	934	359	4891	4891	0	947	485	3101	455	633	651	548
T	770	874	0	0	4882	1058	1171	836	2707	2242	2206	2307

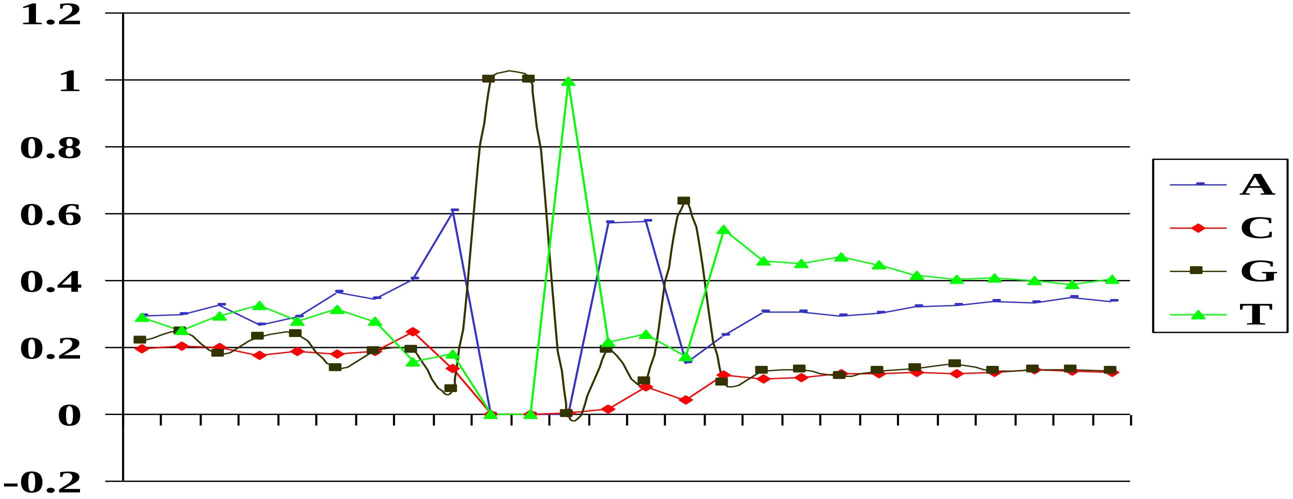
A	0.403	0.609	0.000	0.000	0.000	0.573	0.577	0.151	0.236	0.306	0.306	0.295
C	0.249	0.139	0.000	0.000	0.002	0.017	0.084	0.044	0.118	0.107	0.110	0.121
G	0.191	0.073	1.000	1.000	0.000	0.194	0.099	0.634	0.093	0.129	0.133	0.112
T	0.157	0.179	0.000	0.000	0.998	0.216	0.239	0.171	0.553	0.458	0.451	0.472

5’ Splice Sites – *C. elegans*

H at –1:



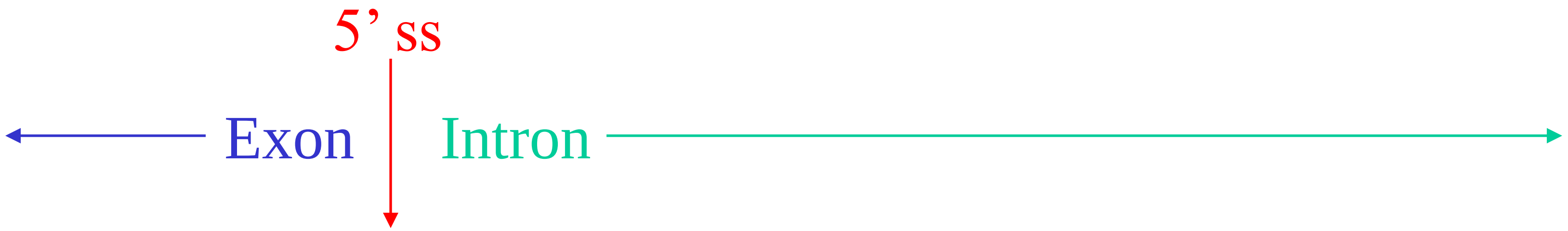
G at –1:



Why the correlation?

- Splicing involves pairing of a small RNA (U1 RNA) with the transcript at the 5’ splice site (positions -2 to +7).
- The RNA is complementary to the 5’ ss consensus sequence.
- A mismatch at position –1 tends to destabilize the pairing, & makes it more important for other positions to be correctly paired.

Nucleotide Counts for 8192 *C. elegans* 5’ Splice Sites


5’ ss

Exon

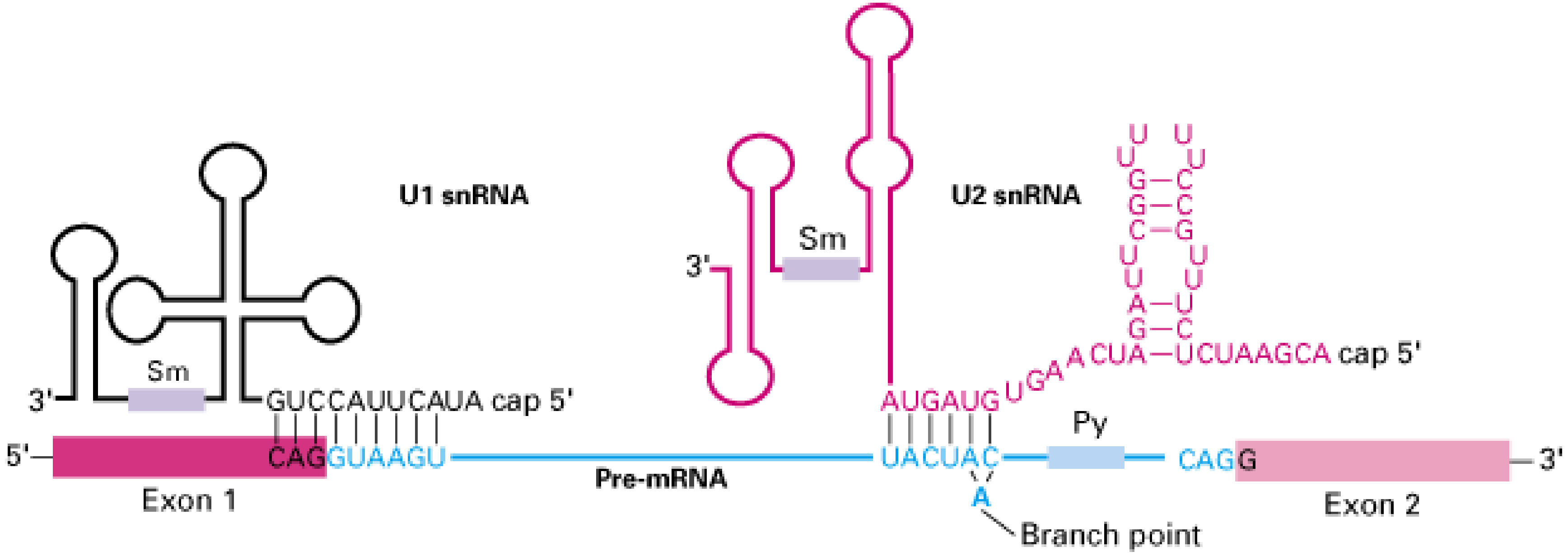
Intron

A	3404	4644	1518	0	0	4836	5486	837	1632	2189	2278	2355
C	1850	1224	583	0	14	118	588	237	801	771	889	986
G	1562	912	4891	8192	0	1890	672	6164	589	962	1056	827
T	1376	1412	1200	0	8178	1348	1446	954	5170	4270	3969	4024

CONSENSUS x a g G T a a g t t w t

A	0.416	0.567	0.185	0.000	0.000	0.590	0.670	0.102	0.199	0.267	0.278	0.287
C	0.226	0.149	0.071	0.000	0.002	0.014	0.072	0.029	0.098	0.094	0.109	0.120
G	0.191	0.111	0.597	1.000	0.000	0.231	0.082	0.752	0.072	0.117	0.129	0.101
T	0.168	0.172	0.146	0.000	0.998	0.165	0.177	0.116	0.631	0.521	0.484	0.491

complementary to portion of U1 RNA



from http://departments.oxy.edu/biology/Stillman/bi221/111300/processing_of_hnrnas.htm

(Jonathon Stillman, Grace Fisher-Adams)

Failure of independence for ‘background’

Nucleotide Freqs (*C. elegans* chr. 1):
A 4575132 (.321) ; C 2559048 (.179) ; G 2555862 (.179); T 4582688 (.321)

dinucleotide frequencies (5′ nuc to left, 3′ nuc at top – e.g. obs freq of *ApC* is .047): (Note “symmetry”!)

Observed

	A	C	G	T
A	0.135	0.047	0.051	0.088
C	0.061	0.035	0.033	0.051
G	0.063	0.034	0.034	0.047
T	0.061	0.064	0.061	0.135

Expected (*under independence*)

	A	C	G	T
A	0.103	0.057	0.057	0.103
C	0.057	0.032	0.032	0.058
G	0.057	0.032	0.032	0.057
T	0.103	0.058	0.057	0.103

Observed / Expected

	A	C	G	T
A	1.314	0.818	0.885	0.853
C	1.055	1.075	1.031	0.886
G	1.106	1.062	1.074	0.818
T	0.597	1.105	1.056	1.313

Failure of independence for background (cont'd)

Conditional probability (in *C. elegans*) of a given nucleotide (top) occurring, given the preceding nucleotide (left)

	A	**C**	**G**	**T**
A	0.421	0.147	0.159	0.274
C	0.338	0.193	0.185	0.284
G	0.355	0.190	0.192	0.263
T	0.191	0.198	0.189	0.421

Deviations From Expectation

- Underrepresentation of *T*p*A*: found in nearly all genomes;
  - reason unknown:
    - neutral (mutation patterns)?
    - selection?
- Overrepresentation of *A*p*A*, *T*p*T*, *C*p*C*, *G*p*G* – also frequently observed in other organisms.
- Unlike mammalian genomes, no underrepresentation of *C*p*G*
  - *C*p*G* not methylated in *C. elegans* (or most other non-vertebrates).