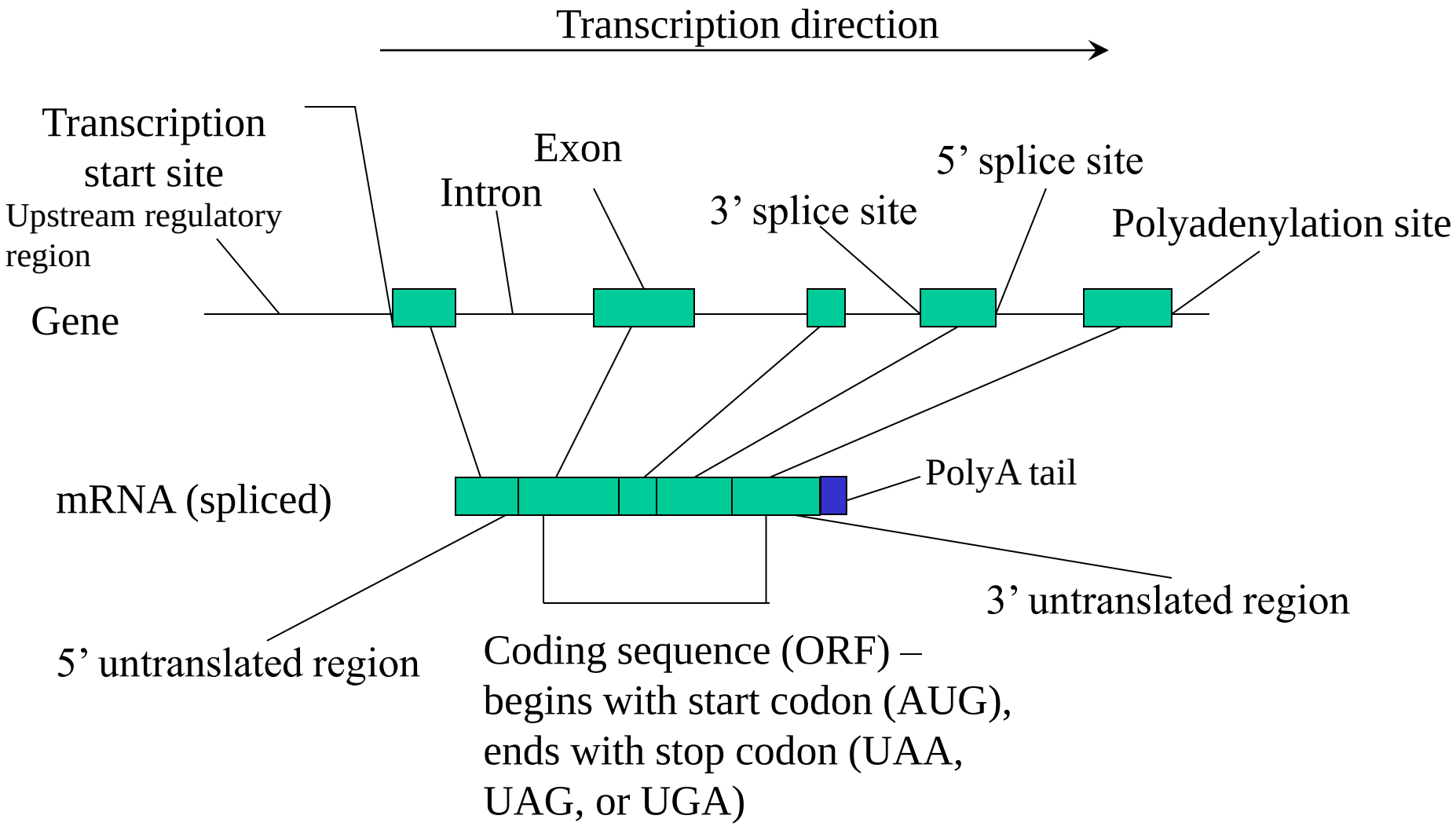
Lecture 4:
Probability Models for Sites

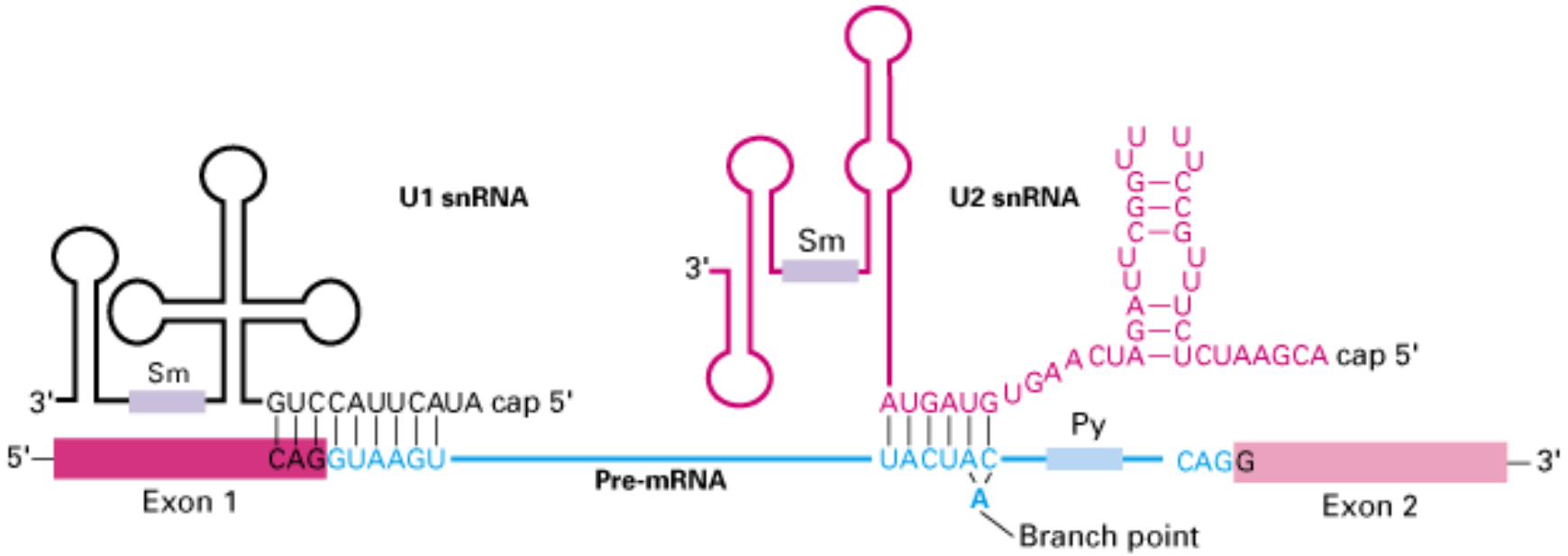
- Assumptions
- Construction
- Examples
  - Splice sites
  - Codons
- Model ‘failures’
  - independence
  - 3’ splice ‘sites’

Site Models

- Probability models for short sequences of some type for which:
  - different examples can be aligned *without gaps* (indels) such that tend to have same residues in same positions
- Applies when
  - precise residue spacing is structurally or functionally important, and
  - certain positions are highly conserved
- Examples:
  - (Genomic 'sites'): DNA/RNA sequences binding a single protein or RNA molecule (the 'reader')
  - Protein 'motifs'

(Protein-coding) Gene Structure in Eukaryotes





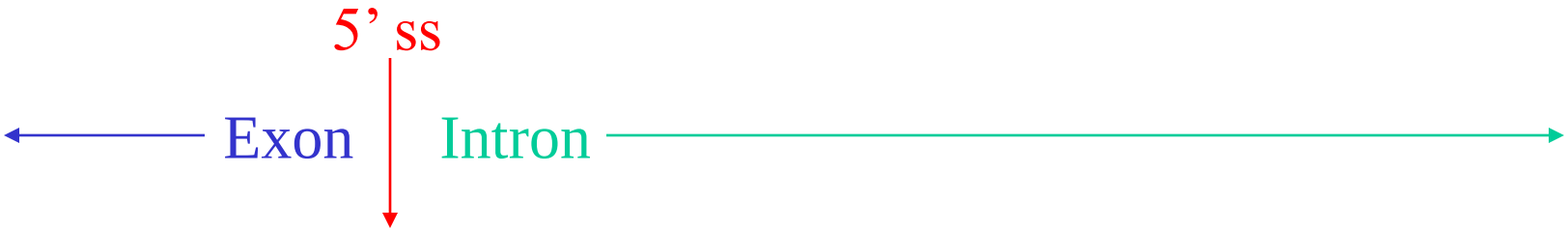
from http://departments.oxy.edu/biology/Stillman/bi221/111300/processing_of_hnrnas.htm

(Jonathon Stillman, Grace Fisher-Adams)

Construction of Site Models

- Collect examples of site (‘training data’)
- Align (without gaps)
- Count occurrences of residues at each position
- Convert to *position-specific* frequencies
- Compute sequence probabilities using *independence* assumption

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

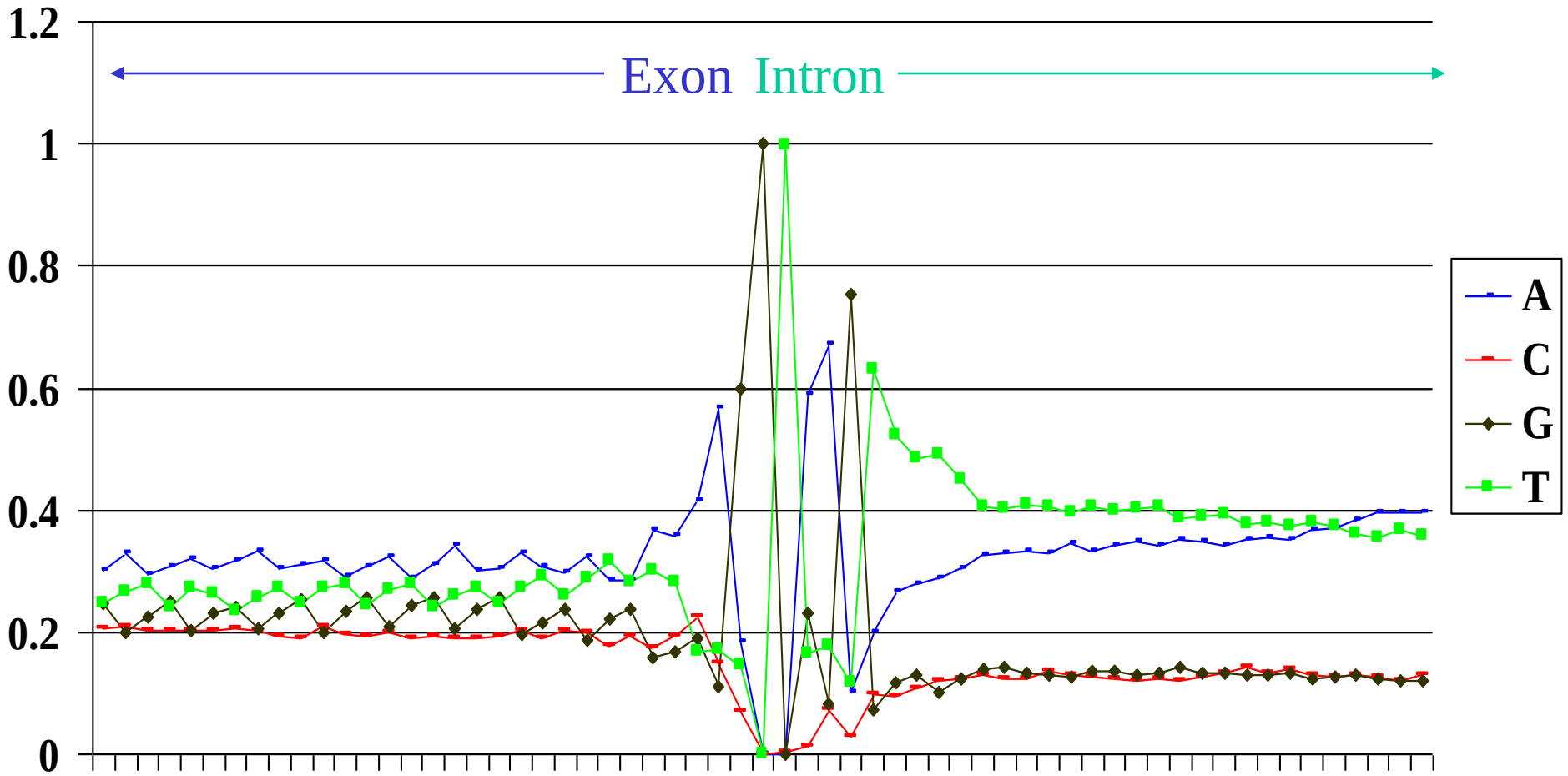


A	3404	4644	1518	0	0	4836	5486	837	1632	2189	2278	2355
C	1850	1224	583	0	14	118	588	237	801	771	889	986
G	1562	912	4891	8192	0	1890	672	6164	589	962	1056	827
T	1376	1412	1200	0	8178	1348	1446	954	5170	4270	3969	4024

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

5’ Splice Sites – *C. elegans*



Probabilities for site sequences (assuming independence!)

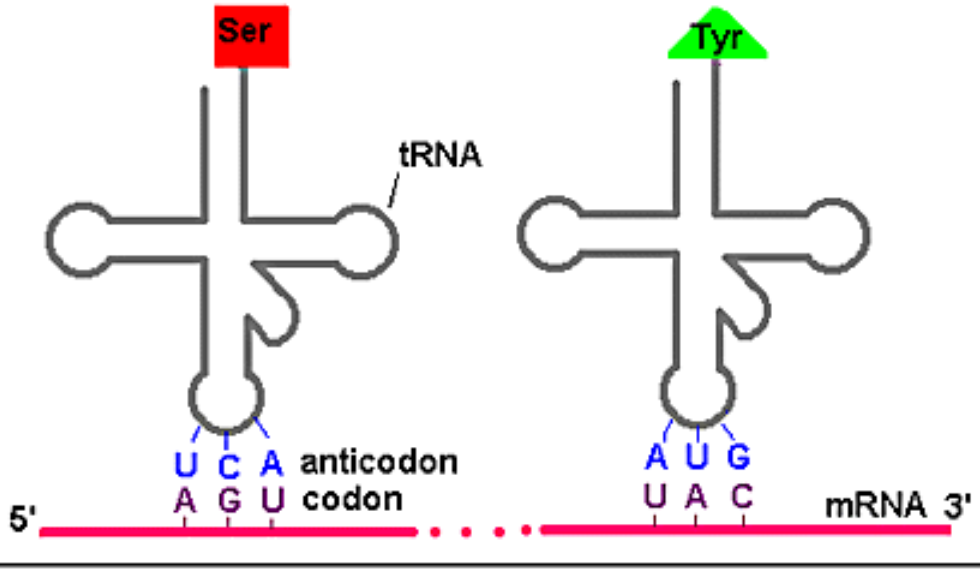
- For each position i, $1 \le i \le n$, let P_i be a prob dist'n on the alphabet of residues
  - e.g. constructed using counts at that position in a sample of sites.
  - $P_i(r)$ for each residue r is the probability that r occurs at position i in a sequence.
- Prob dist'n P on the space S of sequences of length n is defined by

$$P(s) = \prod_{1 \le i \le n} P_i(s_i)$$

where $s = s_1 s_2 \ldots s_n$

Zero Probabilities

- If $P_i(r) = 0$ for some i and r, then $P(s) = 0$ for some sequences.
  - may or may not be desirable
- If due to failure to observe residue because of small sample size,
  - should perform "small-sample correction" to change $P_i(r)$ to a small non-zero value.
  - usually done by adding 'pseudocounts' to each value in the counts matrix;
    - e.g. add 1 to each cell (has justification in Bayesian statistics)
  - Particularly an issue with proteins, due to larger alphabet size.
- If reflects real biological constraints
  - then leave as 0.
  - e.g. requirement for G at position +1 (first intronic base) in 5'ss



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

Codon Usage

- In most organisms, the codons for an amino acid are not used with equal frequency – “synonymous codon bias”.
- For many organisms this may reflect differences in translational efficiency & accuracy
  - more highly expressed genes have stronger biases
- For mammals codon usage mainly reflects the GC content of the region in which the gene is found
  - GC content variation probably reflects *GC-biased gene conversion*

From *Initial sequencing and analysis of the human genome*, International Human Genome Sequencing Consortium, *Nature* 409, 860-921 (2001)

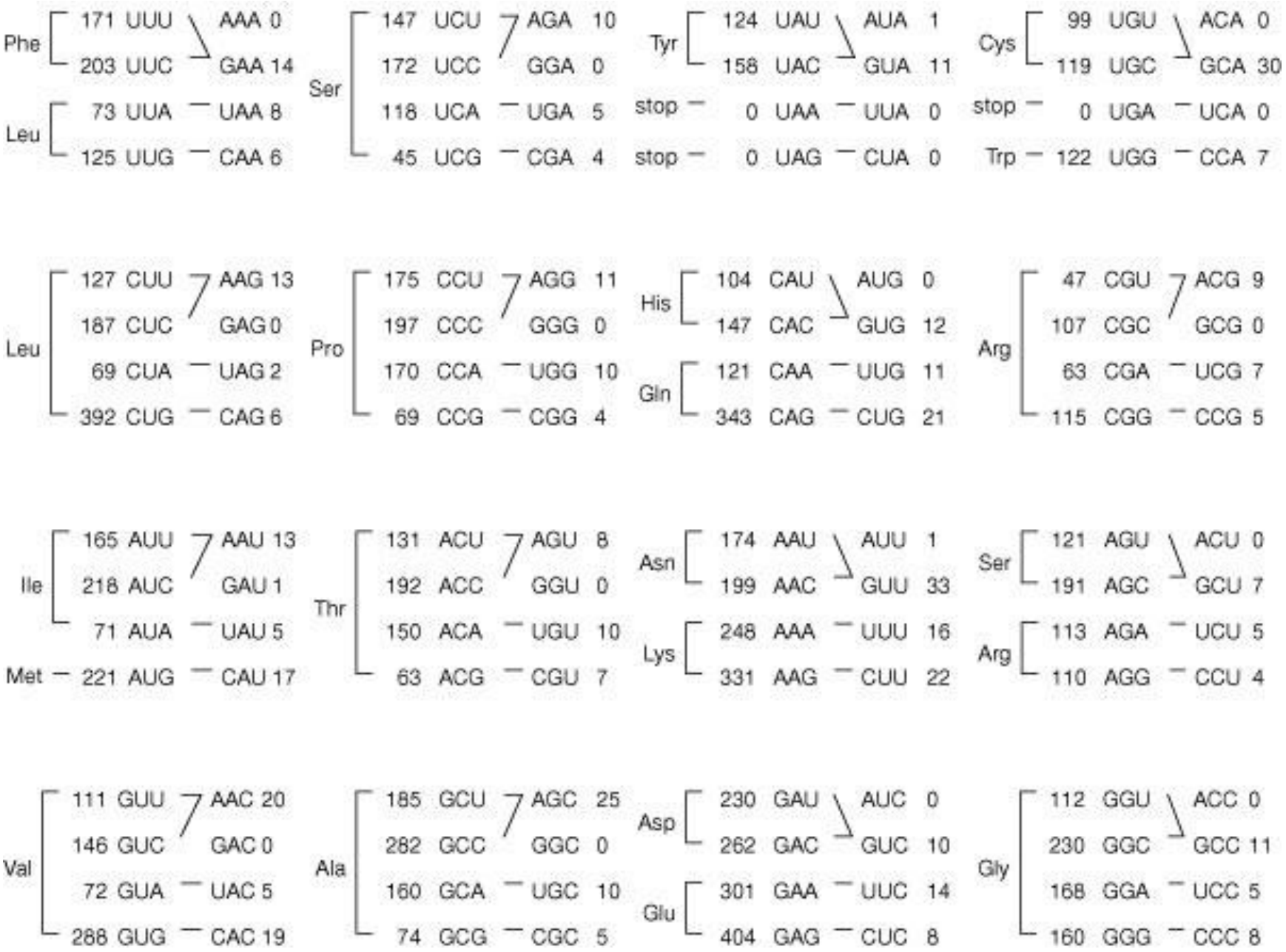
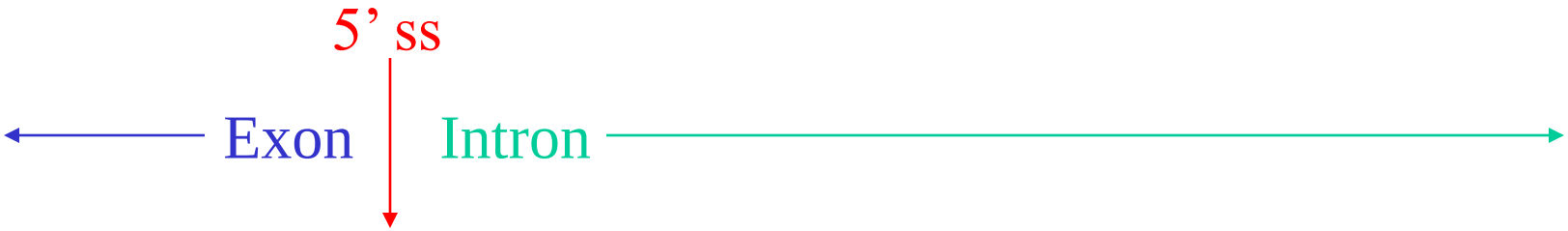


Figure 34 The human genetic code and associated tRNA genes. For each of the 64 codons, we show: the corresponding amino acid; the observed frequency of the codon per 10,000 codons; the codon; predicted wobble pairing to a tRNA anticodon (black lines); an unmodified tRNA anticodon sequence; and the number of tRNA genes found with this anticodon. For example, phenylalanine is encoded by UUU or UUC; UUC is seen more frequently, 203 to 171 occurrences per 10,000 total codons; both codons are expected to be decoded by a single tRNA anticodon type, GAA, using a G/U wobble; and there are 14 tRNA genes found with this anticodon. The modified anticodon sequence in the mature tRNA is not shown, even where post-transcriptional modifications can be confidently predicted (for example, when an A is used to decode a U/C third position, the A is almost certainly an inosine in the mature tRNA). The Figure also does not show the number of distinct tRNA species (such as distinct sequence families) for each anticodon; often there is more than one species for each anticodon.

Independence assumption failures

- 5’ sites (Burge-Karlin observation)
- Offsetting changes for interacting residues
  - RNA stems,
  - protein motifs

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

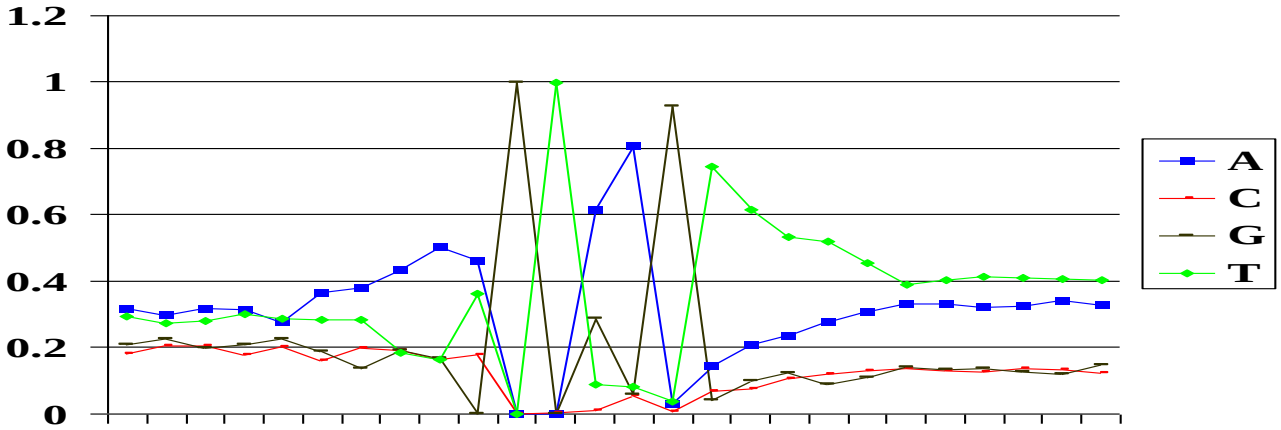
G in position –1 :

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

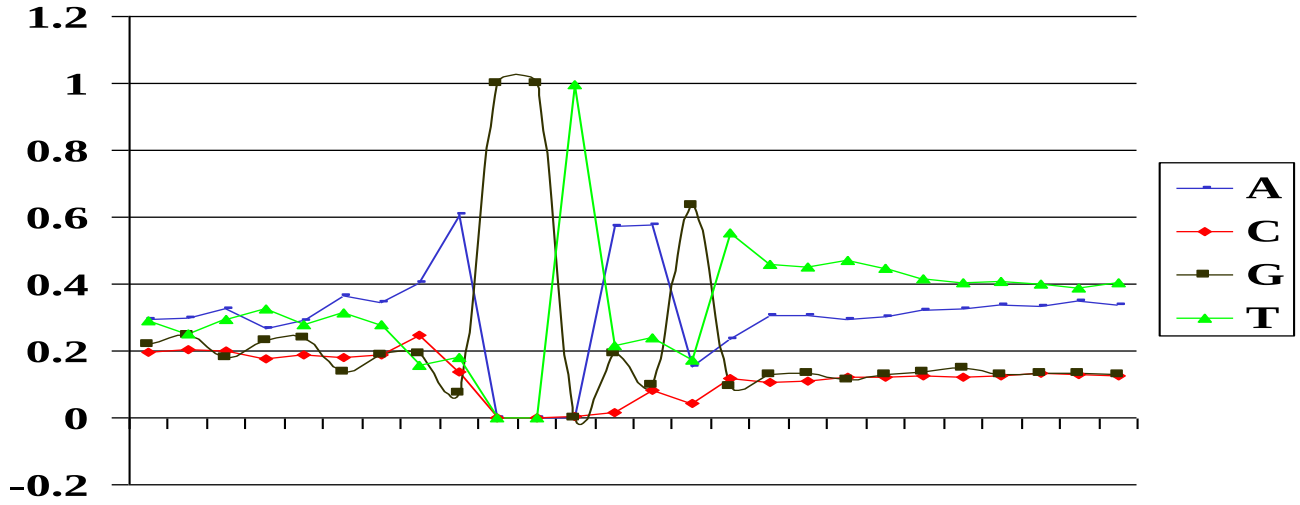
A	0.403	0.609	0.000	0.000	0.000	0.573	0.577	0.151	0.236	0.306	0.306	0.295
C	0.249	0.139	0.000	0.000	0.002	0.017	0.084	0.044	0.118	0.107	0.110	0.121
G	0.191	0.073	1.000	1.000	0.000	0.194	0.099	0.634	0.093	0.129	0.133	0.112
T	0.157	0.179	0.000	0.000	0.998	0.216	0.239	0.171	0.553	0.458	0.451	0.472

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.

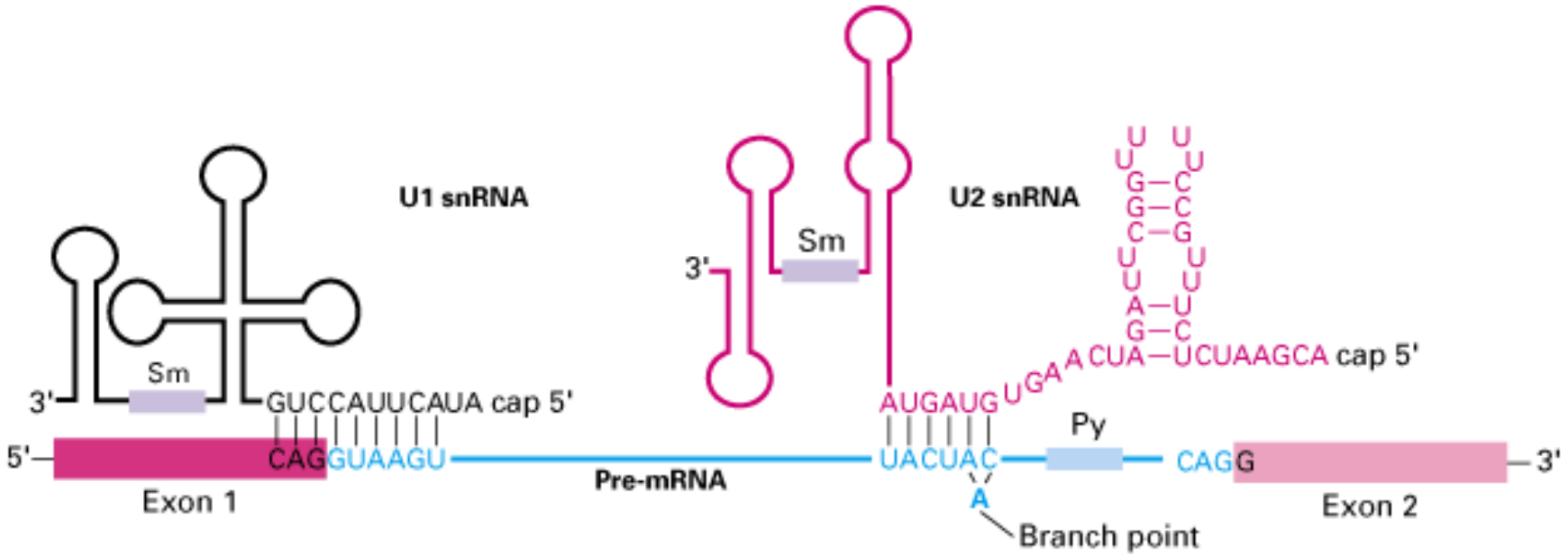
Nucleotide Counts for 8192 *C. elegans* 5' Splice Sites

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



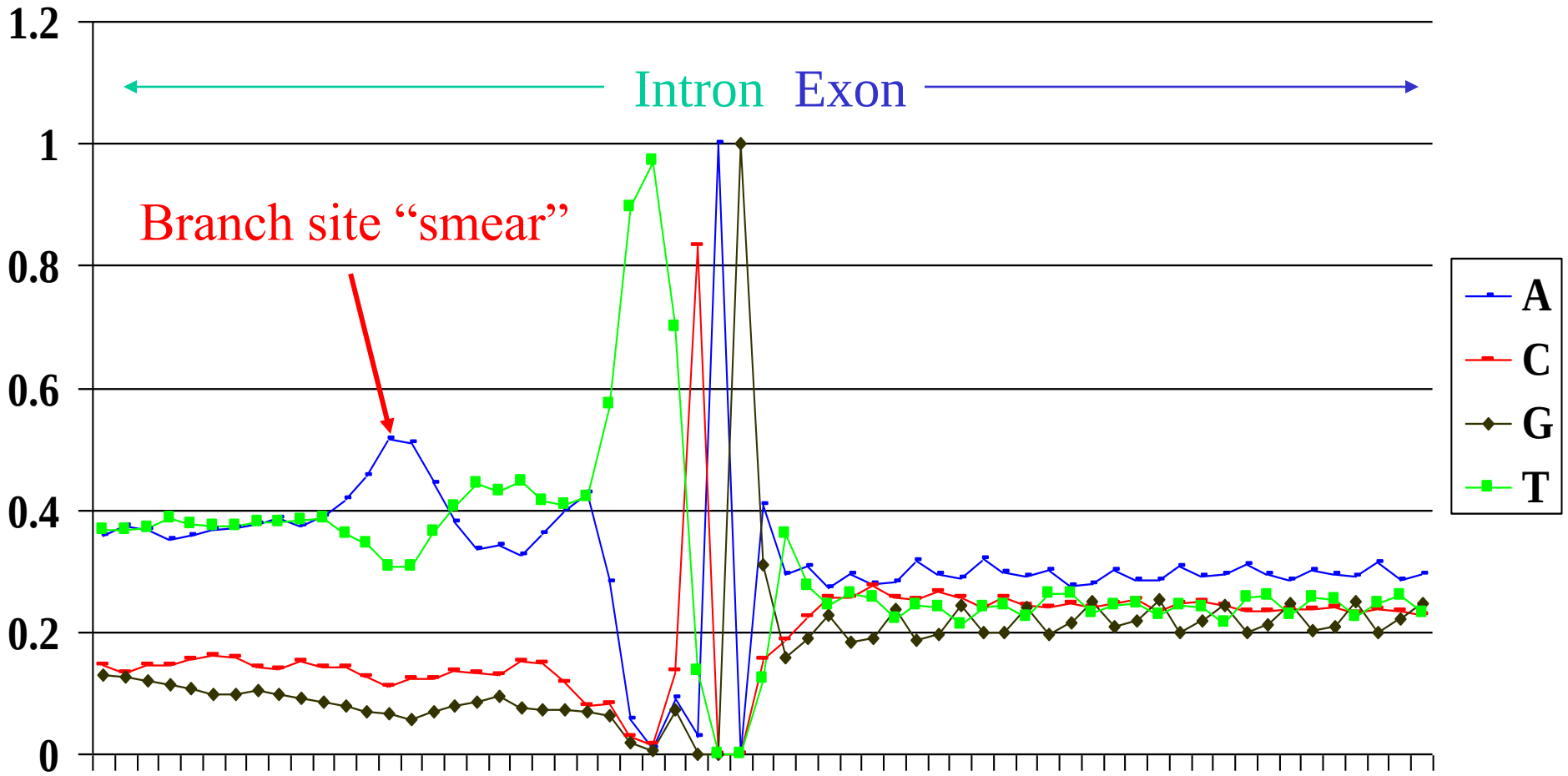
from http://departments.oxy.edu/biology/Stillman/bi221/111300/processing_of_hnrnas.htm

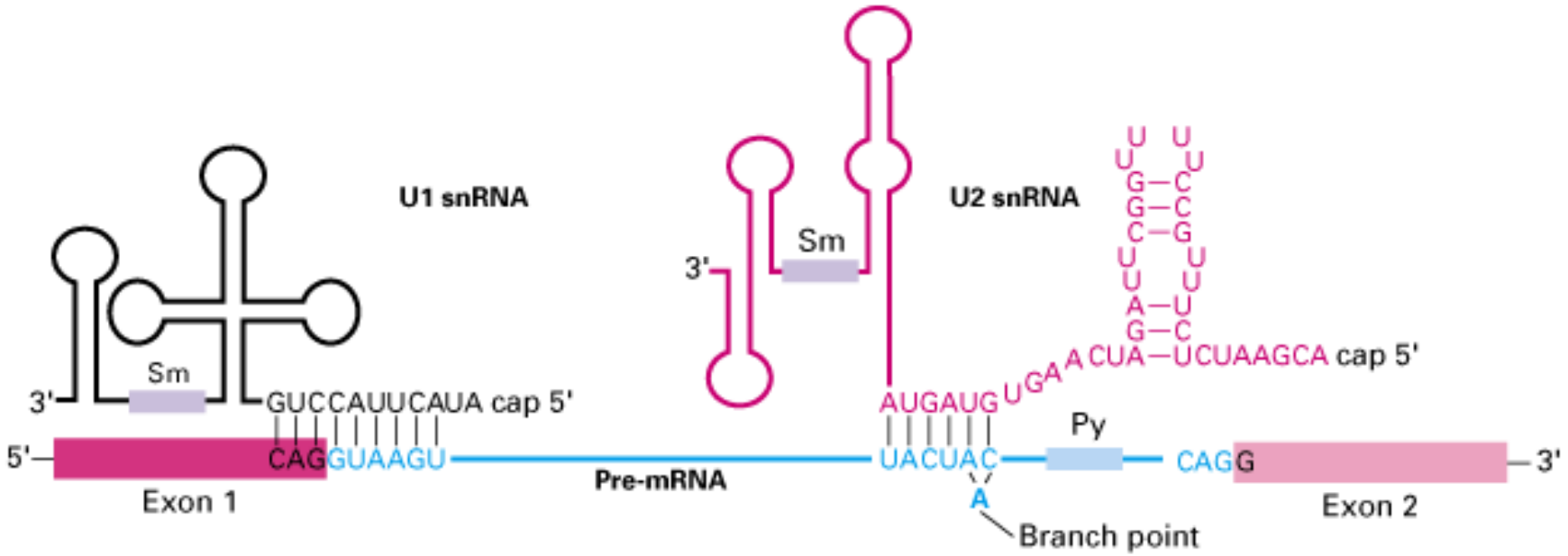
(Jonathon Stillman, Grace Fisher-Adams)

Nucleotide Counts for 8192 *C. elegans* 3' Splice Sites

A	3276	3516	2313	476	67	757	240	8192	0	3359	2401	2514
C	970	648	664	236	129	1109	6830	0	0	1277	1533	1847
G	593	575	516	144	39	595	12	0	8192	2539	1301	1567
T	3353	3453	4699	7336	7957	5731	1110	0	0	1017	2957	2264
CONSENSUS	W	W	W	T	T	t	C	A	G	r	w	w
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

3’ Splice ‘Sites’ – *C. elegans*





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

(Jonathon Stillman, Grace Fisher-Adams)

- a 3’ splice site includes more than one ‘site’ (as we originally defined it)!