# Lecture 3

- Background sequence models
  - Markov models
  - proteins
- Site models

# Failure of independence assumption

**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 |

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 |

# Markov models

- Conditional probabilities (as on the previous slide) can be used to define a ***first-order Markov model*** (or ***Markov chain model***) for sequence probabilities:

$$P(s_1\, s_2\, s_3 \cdots s_n)$$

$$\equiv P(s_1)\, P(s_2 \mid s_1)\, P(s_3 \mid s_2) \;\cdots\; P(s_n \mid s_{n\,\text{-}1})$$

- Similarly, one can define an a ***order-k Markov model*** in which the probability of $s_i$ is conditional on $s_{i-k} \cdots s_{i-2}\, s_{i-1}$

  (i.e. the *k* preceding residues)
- Note that the required number of parameters is exponential in *k*
- The ***independence model*** (which is usually good enough for us!) = the ***order-0 Markov model***

# Background models for *protein* sequences

- The *independence* assumption is (usually) OK

- The *equal frequency* assumption is not

# Failure of equal frequency assumption for proteins

| AMINO ACID | FREQUENCY. | # SYNON CODONS. |
|---|---|---|
| L | .093 | 6 |
| A | .075 | 4 |
| S | .072 | 6 |
| G | .069 | 4 |
| V | .065 | 4 |
| E | .063 | 2 |
| K | .059 | 2 |
| T | .058 | 4 |
| I | .057 | 3 |
| D | .053 | 2 |
| R | .052 | 6 |
| P | .049 | 4 |
| N | .045 | 2 |
| F | .041 | 2 |
| Q | .040 | 2 |
| Y | .032 | 2 |
| M | .024 | 1 |
| H | .022 | 2 |
| C | .017 | 2 |
| W | .013 | 1 |

# Hypotheses to explain correlation between frequency and # codons

- (*Neutralist*):
  - Nucleotide sequences that encode proteins are on average close to random,
  - so amino acid freqs are proportionate to codon freqs in random DNA.
- (*Selectionist*):
  - The genetic code evolved concurrently with early proteins, and
  - is adapted so that the most useful amino acids are encoded by the most codons.
- The truth is probably some combination of these!
  - Dependence of aa composition on genomic G+C content is consistent with neutralist hypothesis

# Deviations from randomness

- Compute, for each residue *r*, the ratio obs$_r$ /exp$_r$ of
  - the observed frequency obs$_r$, to
  - the expected frequency exp$_r$ if coding sequences were random:

$$\mathrm{exp}_r = (\#\text{codons encoding } r) / 61$$

| Amino Acid | Obs/Exp | 1st codon base | 2nd codon base | 3rd codon base | # codons |
|---|---|---|---|---|---|
| E | 1.92 | G | A | R | 2 |
| K | 1.80 | A | A | R | 2 |
| D | 1.62 | G | A | Y | 2 |
| M | 1.46 | A | T | G | 1 |
| N | 1.37 | A | A | Y | 2 |
| F | 1.25 | T | T | Y | 2 |
| Q | 1.22 | C | A | R | 2 |
| I | 1.16 | A | T | Not G | 3 |
| A | 1.14 | G | C | N | 4 |
| G | 1.05 | G | G | N | 4 |
| V | .99 | G | T | N | 4 |
| Y | .98 | T | A | Y | 2 |
| L | .95 | C(T) | T | N | 6 |
| T | .88 | A | C | N | 4 |
| W | .79 | T | G | G | 1 |
| P | .74 | C | C | N | 4 |
| S | .73 | T(A) | C(G) | N | 6 |
| H | .67 | C | A | Y | 2 |
| R | .53 | C(A) | G | N | 6 |
| C | .52 | T | G | Y | 2 |

# Obs/Exp Ratios

- All observed values are within factor of 2 of expected;
  - last column suggests trend towards “correcting” disparate # codons
- At codon position 1,
  - purines (*A* and *G*) predominate among over-represented amino acids,
  - pyrimidines (*C* and *T*) among under-represented amino acids.
- At codon position 2,
  - *A* and *T* predominate among over-represented amino acids,
  - *C* and *G* among under-represented amino acids.
- Hypotheses to explain *RWR* codon preference:
  - Vestige of ancestral code? (Shepherd)
  - Over-represented pattern more efficiently translated?

# Site Models

- Probability models for short sequences, such as:
  - splice sites
  - translation start sites
  - promoter elements
  - protein “motifs”

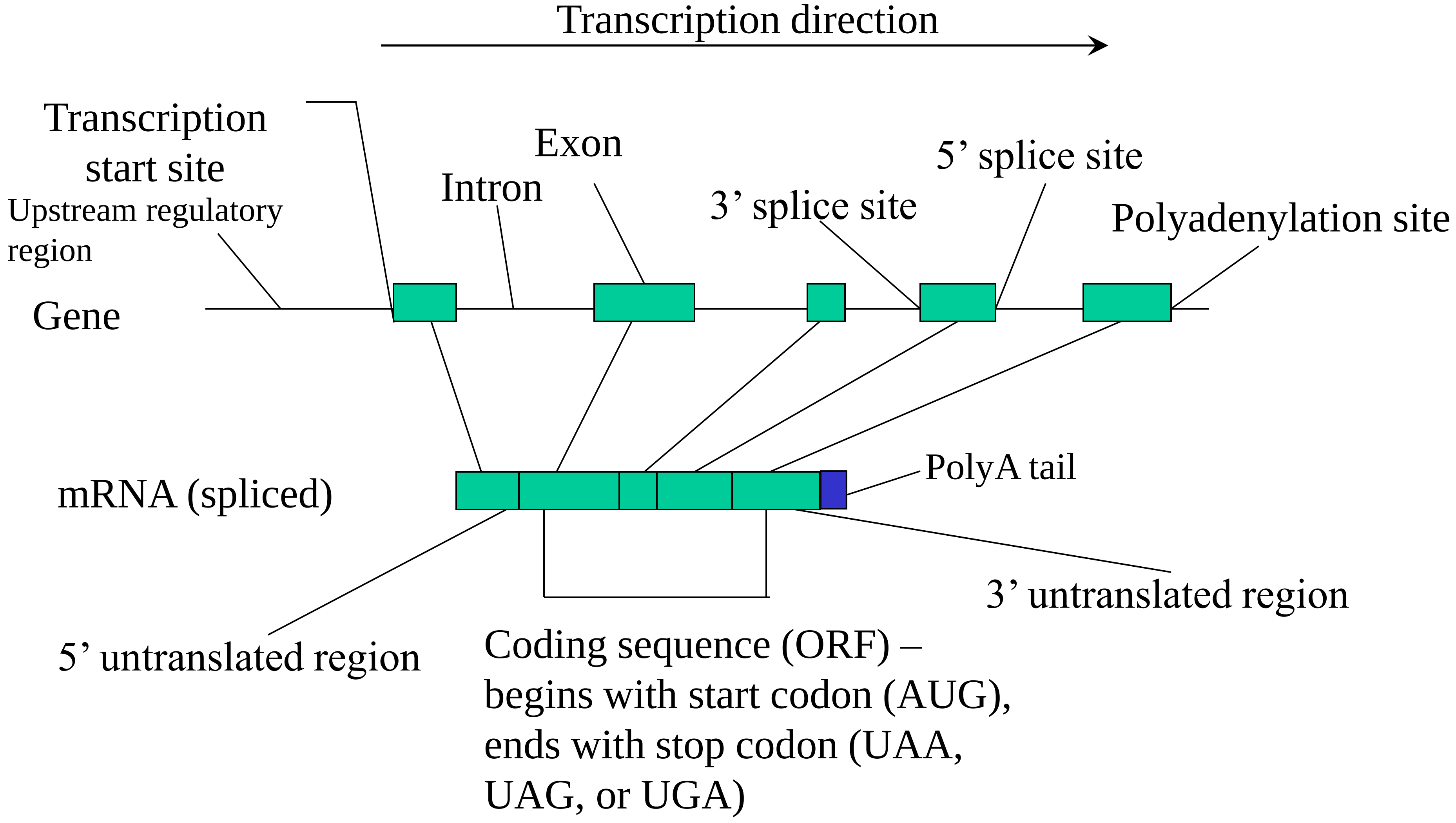
- Assumptions:
  - different examples of site can be aligned *without gaps* (indels) such that tend to have same residues in same positions
  - drop equal freq assumption: allow *position-specific freqs*
  - retain *independence* assumption (for now)

- Applies to short segments (< 30 residues) where
  - precise residue spacing is structurally or functionally important, and
  - certain positions are highly conserved
- Examples:
  - DNA/RNA sequences binding a single protein or RNA molecule
  - Protein internal regions structurally constrained due to folding requirements; or
  - protein surface regions constrained because bind certain ligands

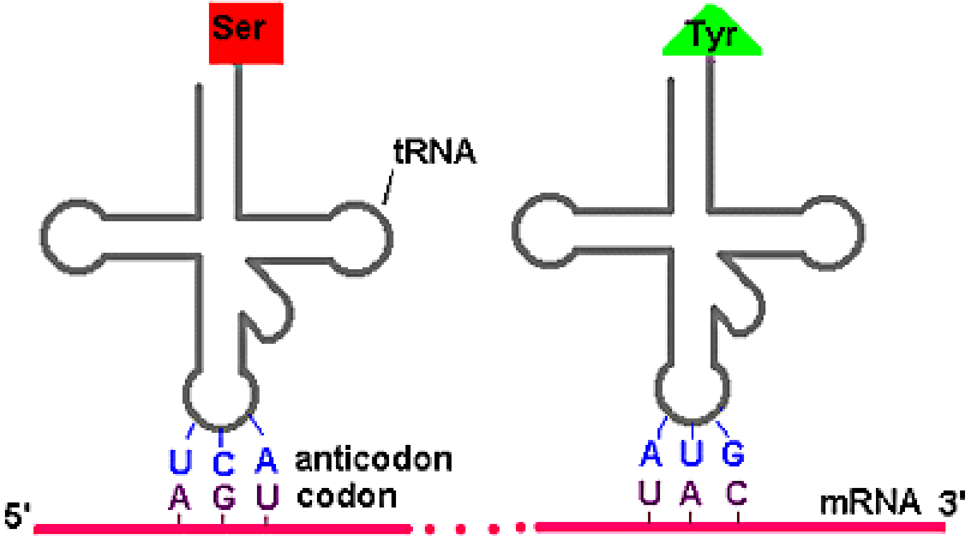
# Construction of Site Models

- Collect examples of site
- Align (without gaps)
- Count occurrences of residues at each position
- Convert to frequencies

# (Protein-coding) Gene Structure in Eukaryotes



Coding sequence (ORF) – begins with start codon (AUG), ends with stop codon (UAA, UAG, or UGA)



2nd base in codon

1st base in codon

3rd base in codon

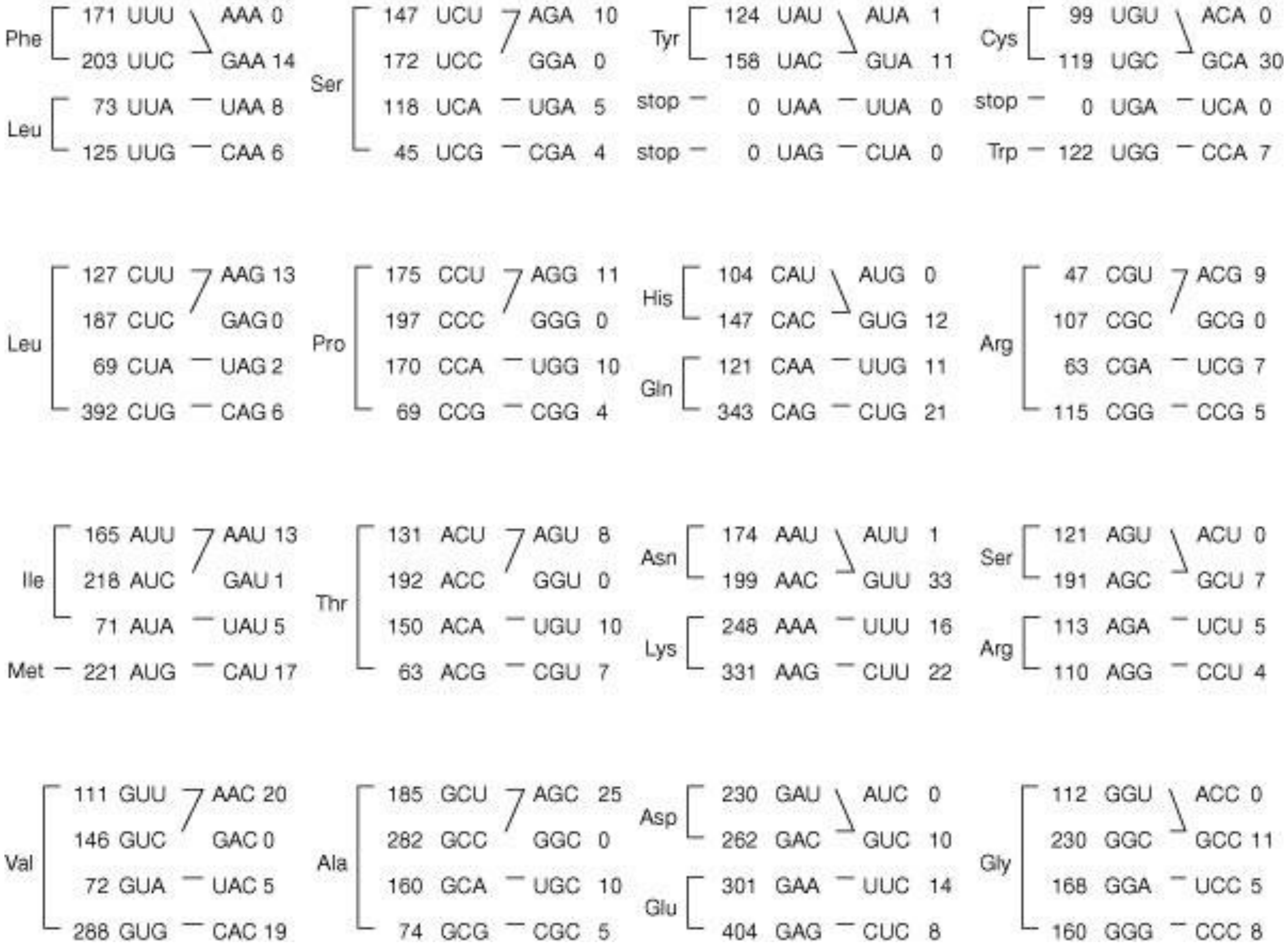
| | U | C | A | G | |
|---|---|---|---|---|---|
| U | Phe<br>Phe<br>Leu<br>Leu | Ser<br>Ser<br>Ser<br>Ser | Tyr<br>Tyr<br>STOP<br>STOP | Cys<br>Cys<br>STOP<br>Trp | U<br>C<br>A<br>G |
| C | Leu<br>Leu<br>Leu<br>Leu | Pro<br>Pro<br>Pro<br>Pro | His<br>His<br>Gln<br>Gln | Arg<br>Arg<br>Arg<br>Arg | U<br>C<br>A<br>G |
| A | Ile<br>Ile<br>Ile<br>Met | Thr<br>Thr<br>Thr<br>Thr | Asn<br>Asn<br>Lys<br>Lys | Ser<br>Ser<br>Arg<br>Arg | U<br>C<br>A<br>G |
| G | Val<br>Val<br>Val<br>Val | Ala<br>Ala<br>Ala<br>Ala | Asp<br>Asp<br>Glu<br>Glu | Gly<br>Gly<br>Gly<br>Gly | U<br>C<br>A<br>G |

# The Genetic Code

# Codon Usage

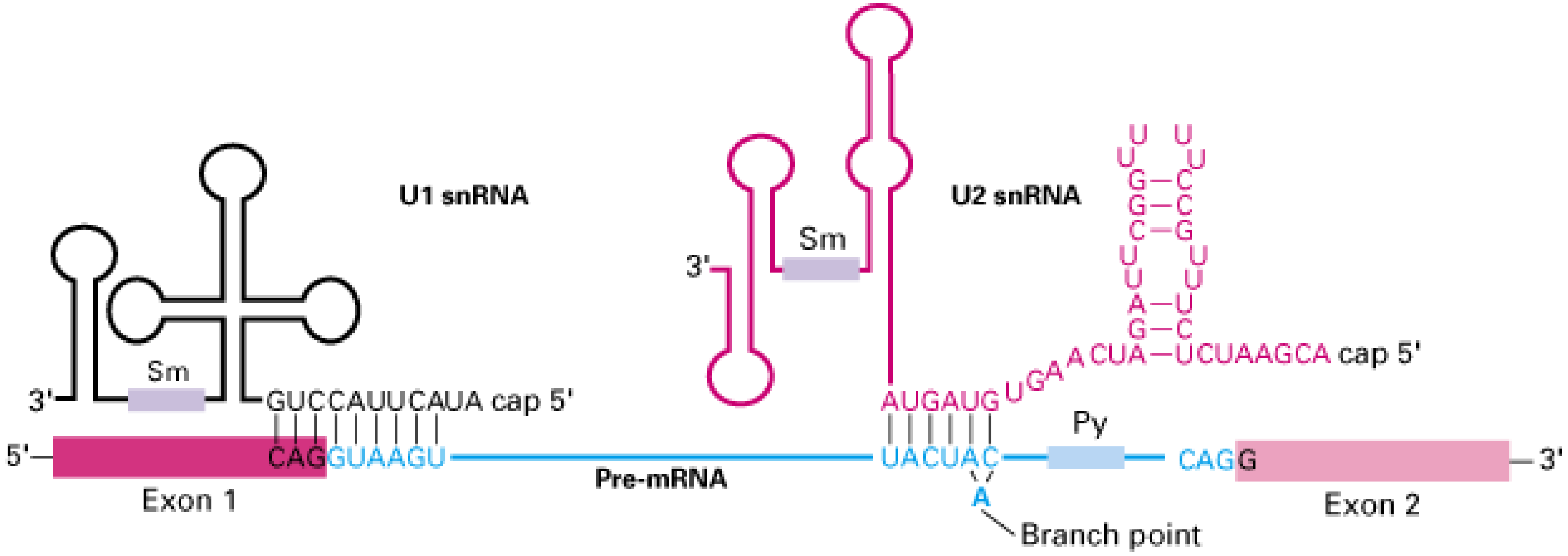
- In most organisms, the codons for an amino acid are not used with equal frequency – “synonymous codon bias”.
- For many organisms this may reflect differences in translational efficiency & accuracy: more highly expressed genes have stronger biases.
- For mammals codon usage mainly reflects the GC content of the region in which the gene is found; reasons for GC content variation unknown.
- *Even though we don’t fully understand the biological basis for this bias, it provides a powerful tool for gene identification!*

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



| | | | | |
|---|---|---|---|---|
| Phe | 171 | UUU | AAA | 0 |
| | 203 | UUC | GAA | 14 |
| Leu | 73 | UUA | UAA | 8 |
| | 125 | UUG | CAA | 6 |
| Ser | 147 | UCU | AGA | 10 |
| | 172 | UCC | GGA | 0 |
| | 118 | UCA | UGA | 5 |
| | 45 | UCG | CGA | 4 |
| Tyr | 124 | UAU | AUA | 1 |
| | 158 | UAC | GUA | 11 |
| stop | 0 | UAA | UUA | 0 |
| stop | 0 | UAG | CUA | 0 |
| Cys | 99 | UGU | ACA | 0 |
| | 119 | UGC | GCA | 30 |
| stop | 0 | UGA | UCA | 0 |
| Trp | 122 | UGG | CCA | 7 |
| Leu | 127 | CUU | AAG | 13 |
| | 187 | CUC | GAG | 0 |
| | 69 | CUA | UAG | 2 |
| | 392 | CUG | CAG | 6 |
| Pro | 175 | CCU | AGG | 11 |
| | 197 | CCC | GGG | 0 |
| | 170 | CCA | UGG | 10 |
| | 69 | CCG | CGG | 4 |
| His | 104 | CAU | AUG | 0 |
| | 147 | CAC | GUG | 12 |
| Gln | 121 | CAA | UUG | 11 |
| | 343 | CAG | CUG | 21 |
| Arg | 47 | CGU | ACG | 9 |
| | 107 | CGC | GCG | 0 |
| | 63 | CGA | UCG | 7 |
| | 115 | CGG | CCG | 5 |
| Ile | 165 | AUU | AAU | 13 |
| | 218 | AUC | GAU | 1 |
| | 71 | AUA | UAU | 5 |
| Met | 221 | AUG | CAU | 17 |
| Thr | 131 | ACU | AGU | 8 |
| | 192 | ACC | GGU | 0 |
| | 150 | ACA | UGU | 10 |
| | 63 | ACG | CGU | 7 |
| Asn | 174 | AAU | AUU | 1 |
| | 199 | AAC | GUU | 33 |
| Lys | 248 | AAA | UUU | 16 |
| | 331 | AAG | CUU | 22 |
| Ser | 121 | AGU | ACU | 0 |
| | 191 | AGC | GCU | 7 |
| Arg | 113 | AGA | UCU | 5 |
| | 110 | AGG | CCU | 4 |
| Val | 111 | GUU | AAC | 20 |
| | 146 | GUC | GAC | 0 |
| | 72 | GUA | UAC | 5 |
| | 288 | GUG | CAC | 19 |
| Ala | 185 | GCU | AGC | 25 |
| | 282 | GCC | GGC | 0 |
| | 160 | GCA | UGC | 10 |
| | 74 | GCG | CGC | 5 |
| Asp | 230 | GAU | AUC | 0 |
| | 262 | GAC | GUC | 10 |
| Glu | 301 | GAA | UUC | 14 |
| | 404 | GAG | CUC | 8 |
| Gly | 112 | GGU | ACC | 0 |
| | 230 | GGC | GCC | 11 |
| | 168 | GGA | UCC | 5 |
| | 160 | GGG | CCC | 8 |

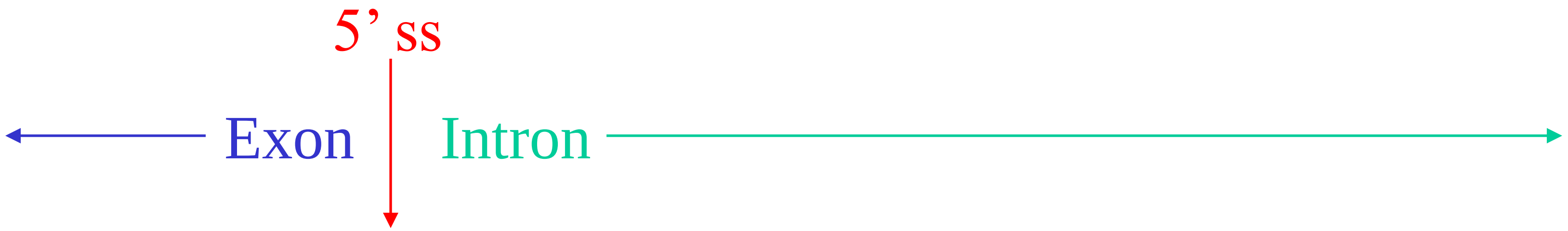
**Figure 34** The human genetic code and associated tRNA genes. For each of the 64 codons, we show: the corresponding amino acid; the observed frequency of the codon per 10,000 codons; the codon; predicted wobble pairing to a tRNA anticodon (black lines); an unmodified tRNA anticodon sequence; and the 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.



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

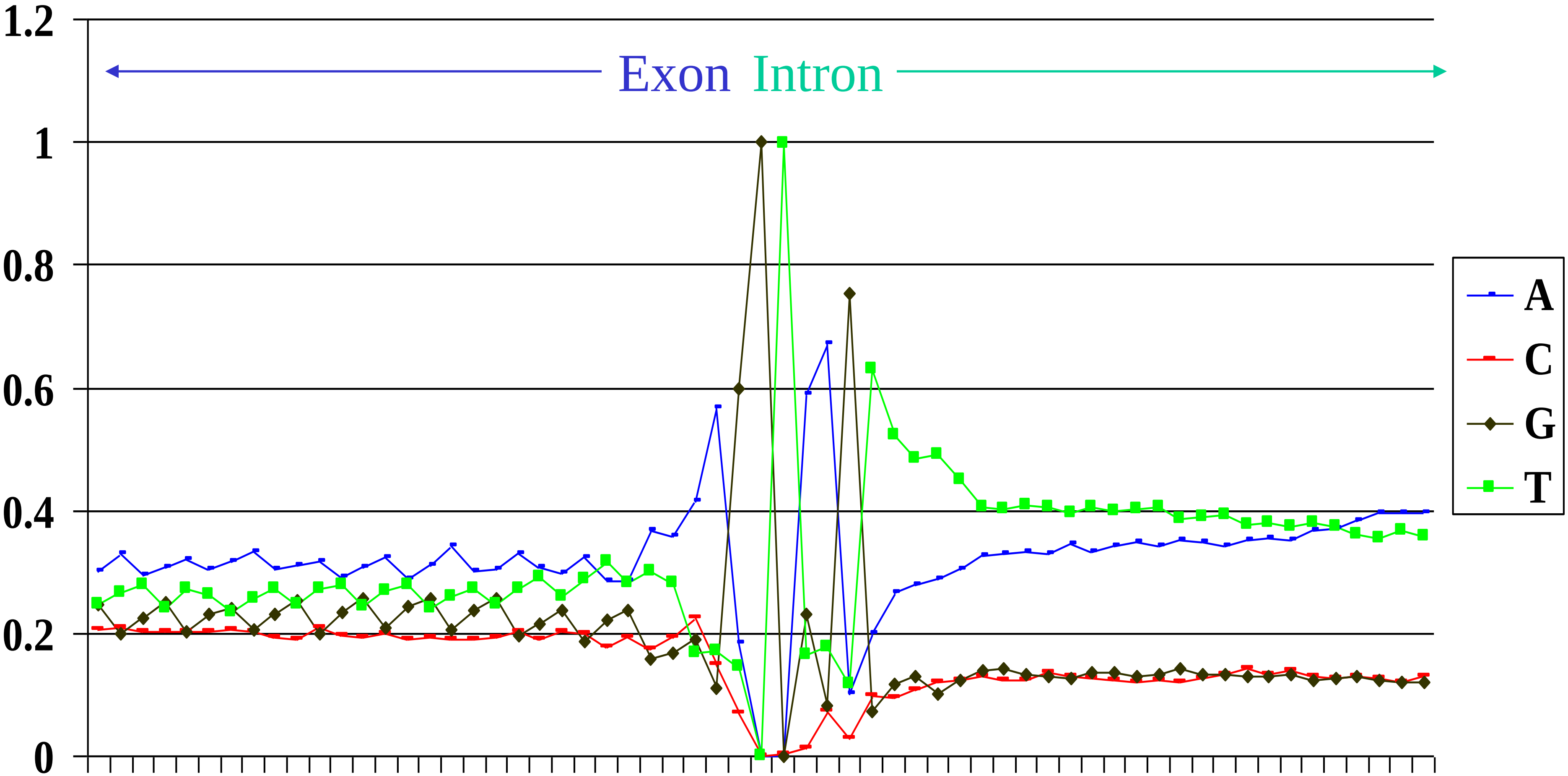
**(Jonathon Stillman, Grace Fisher-Adams )**

# 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*



# Conserved Domain in RecR and Class I Topisomerases

|  |  |
|---|---|
| RecR | RLAEEKITEVILATNPTVEGEATANYIAELC |
| RecM | RLQDDQVTEVILATNPNIEGEATAMYISRLL |
| RecR | RVDDVGITEVIIATDPNTEGEATATYLVRMV |
| TrsI | IFKENKIDEVIIATDPAREGENIAYKILNQL |
| TOP1 | KQLAEKADHIYLATDLDREGEAIAWRLREVI |
| ORF1 | AELLKQANTIIVATDSDREGENIAWSIIHKA |
| TOP1 | KDALKDADELILATDEDREGKVISWHLLQLL |
| TOP1 | TIFDKRVKTIILATDAAAEGEYIGRNILYRL |
| TOP3 | KREARNADYLMIWTDCDREGEYIGWEIWQEA |
| TOP3 | KRFLHEASEIVHAGDPDREGQLLVDEVLDYL |
| RGYR | RNLAVEADEVLIGTDPDTEGEKIAWDLYLAL |
| **CONSENSUS** | **xxxxxxxxxU&uatDxxxEGexxxxxUxxxu** |

*Consensus key*:

Uppercase: all residues chemically similar

lowercase: most are

U,u: bulky aliphatic (I,L,V)

&: bulky hydrophobic (I,L,V,M,F,Y,W)

From RL Tatusov, SF Altschul, and EV Koonin, PNAS 91: 12091-12095

# Probability Models for Sites (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

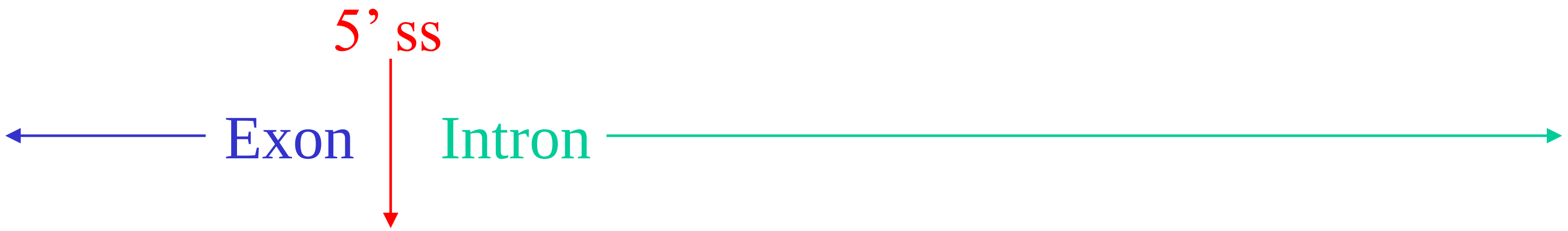
- If $P_i(r) = 0$ for some $i$ and $r$, then $P(s) = 0$ for some sequences.
  - may or may not be desirable
- If due to failure to observe residue because of small sample size,
  - should perform "small-sample correction" to change $P_i(r)$ to a small non-zero value.
  - usually done by adding 'pseudocounts' 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

# Independence assumption failures for Site Models

- 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 |

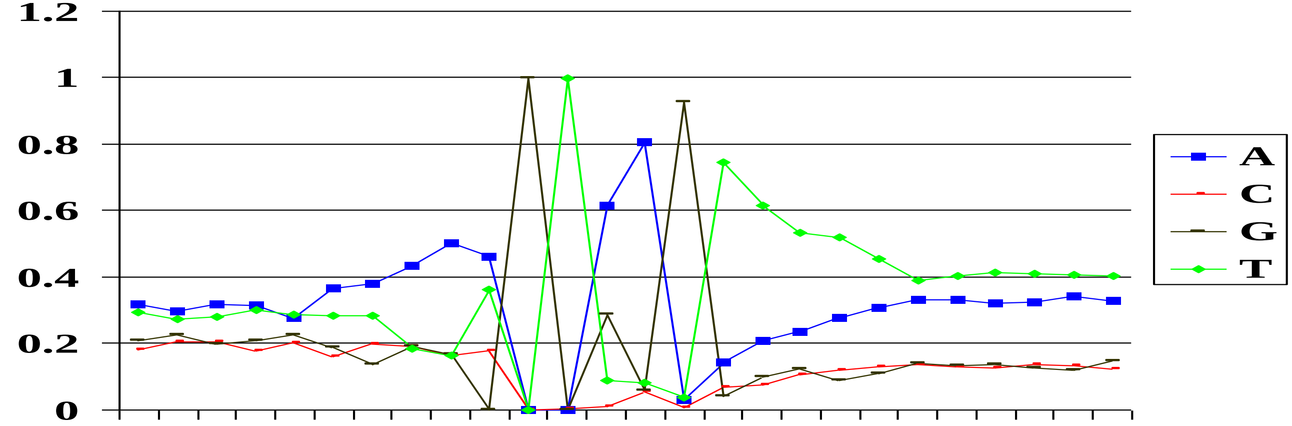
## G in position –1 :

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

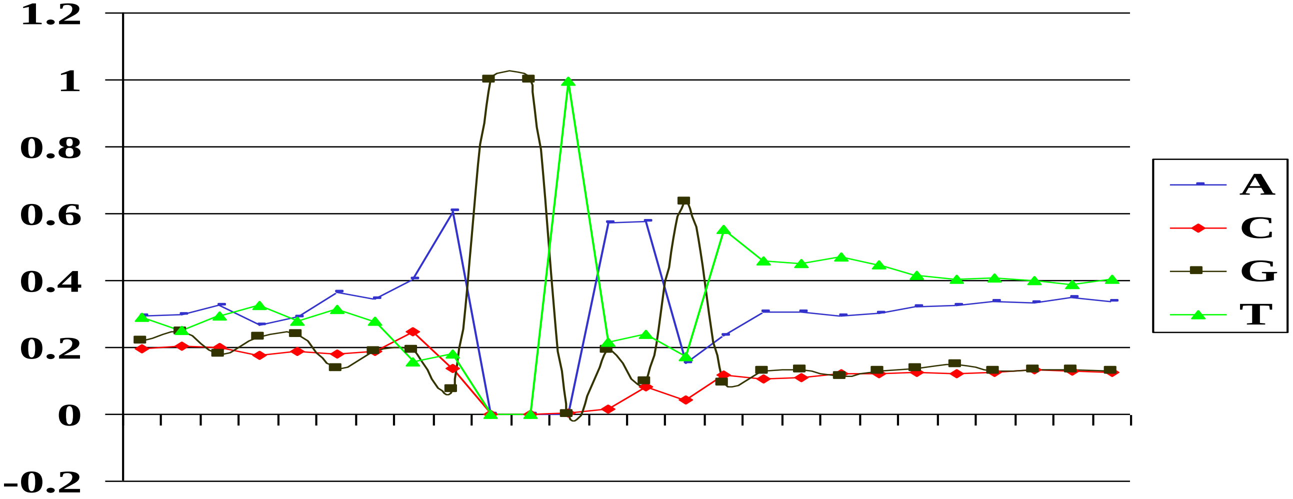
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.403 | 0.609 | 0.000 | 0.000 | 0.000 | 0.573 | 0.577 | 0.151 | 0.236 | 0.306 | 0.306 | 0.295 |
| C | 0.249 | 0.139 | 0.000 | 0.000 | 0.002 | 0.017 | 0.084 | 0.044 | 0.118 | 0.107 | 0.110 | 0.121 |
| G | 0.191 | 0.073 | 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:



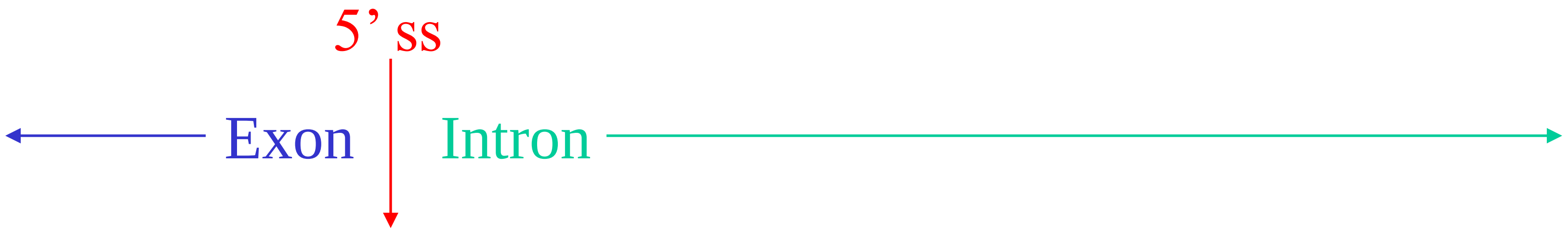
G at –1:



# 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



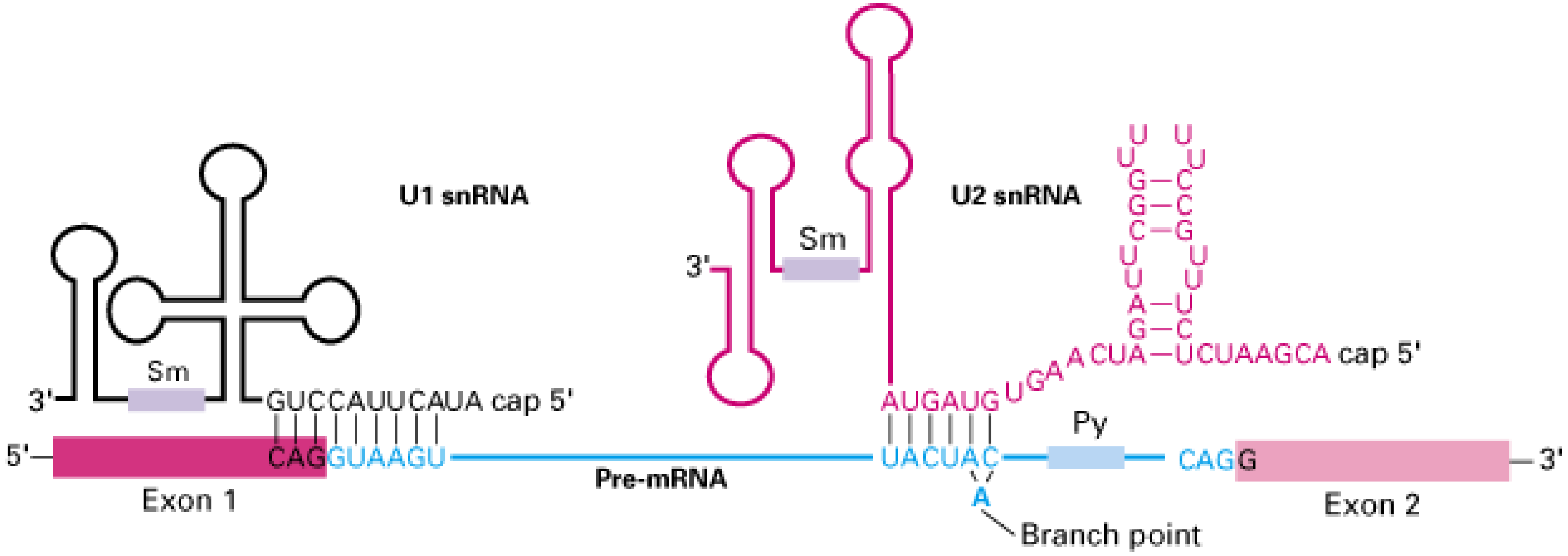
5’ ss

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 |

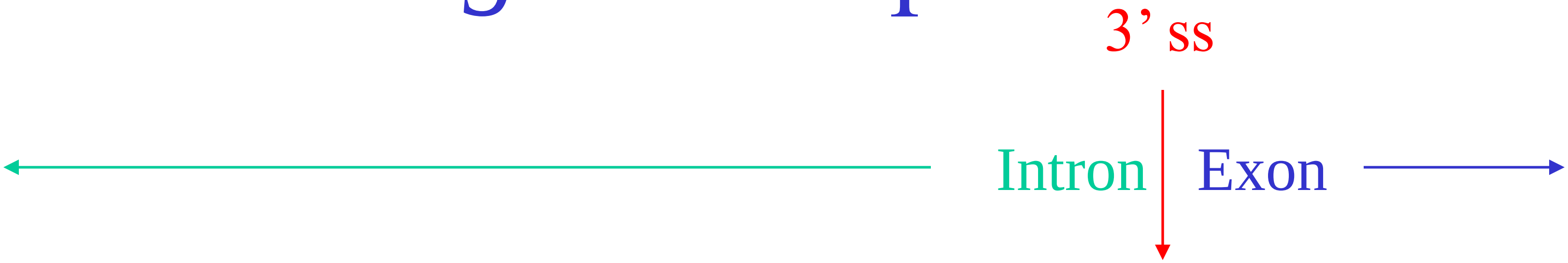
complementary to portion of U1 RNA



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

**(Jonathon Stillman, Grace Fisher-Adams )**

# Nucleotide Counts for 8192 *C. elegans* 3’ Splice Sites



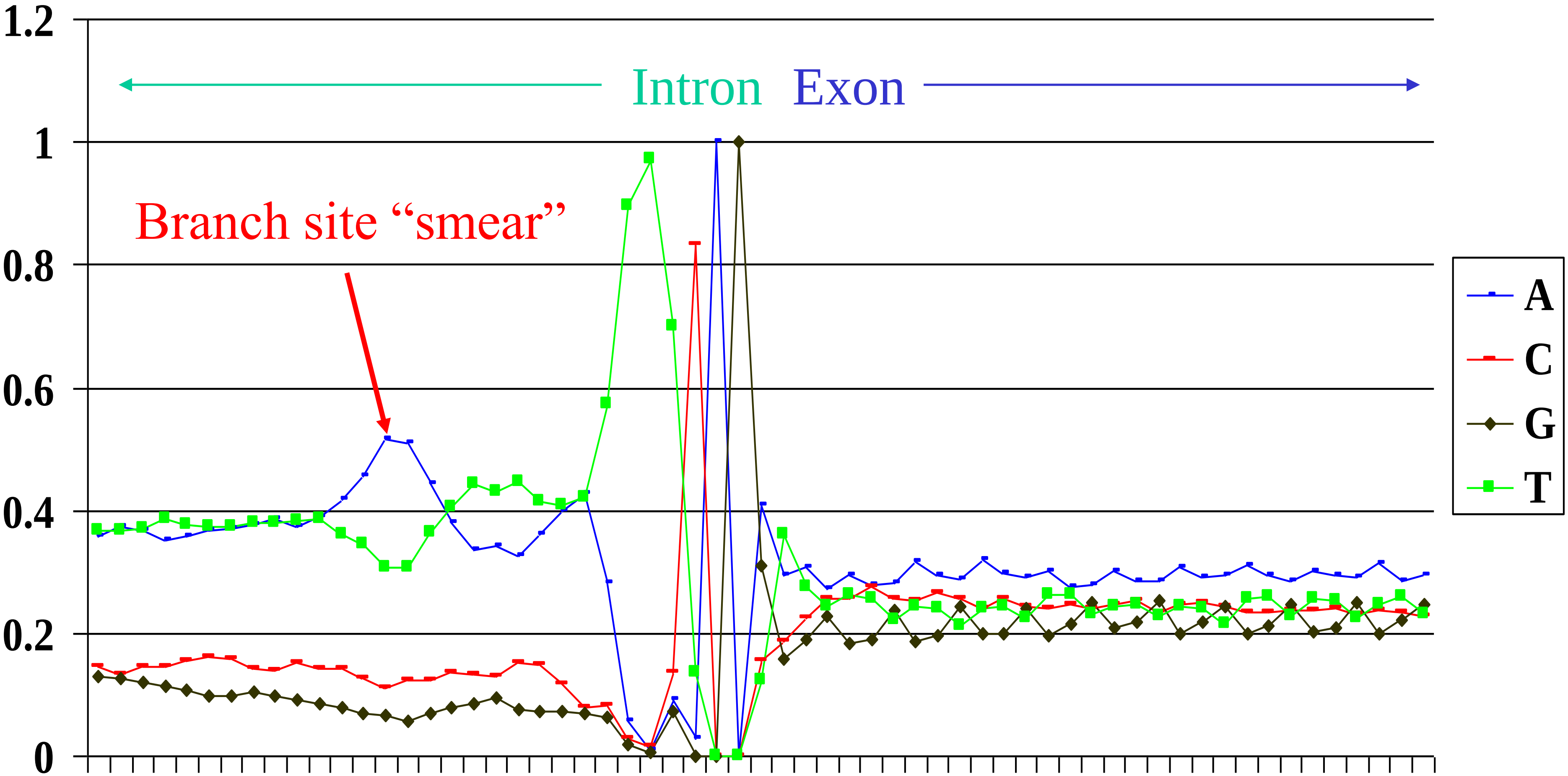
3’ ss

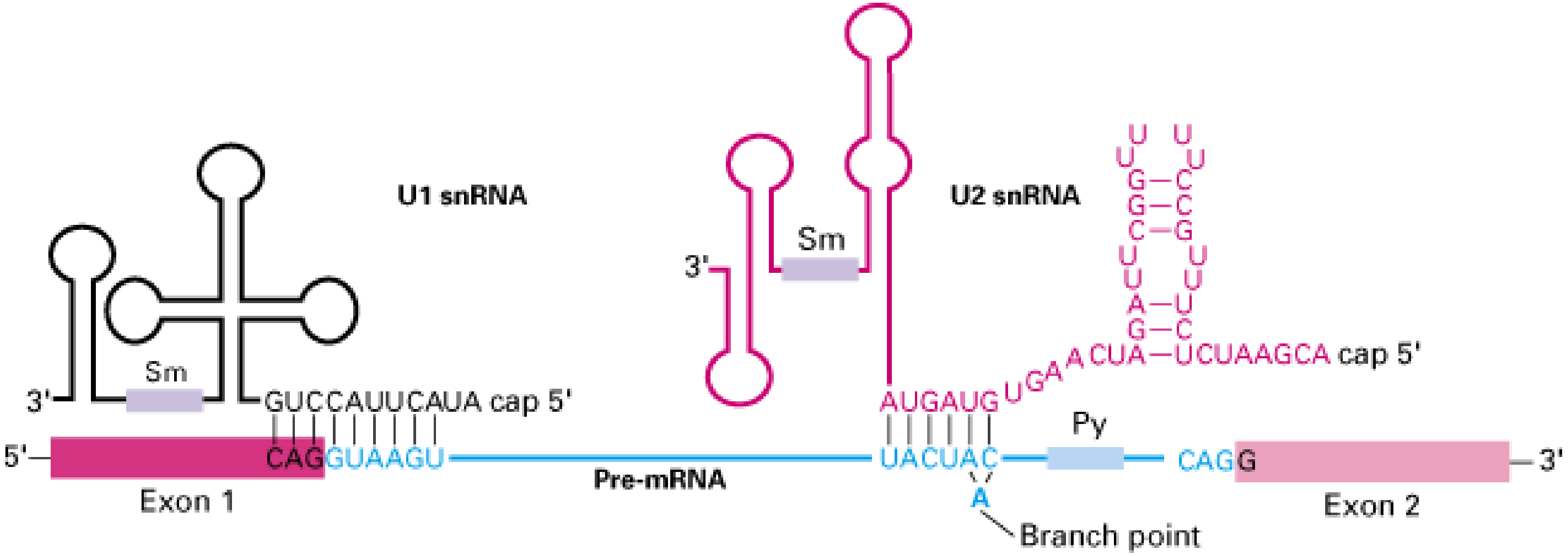
Intron

Exon

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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)!