

Today's Lecture

- Comparing models: Likelihood ratios & weight matrices
 - (Hypothesis testing & Neyman-Pearson lemma)
- Score distributions
- Limitations of site models
 - Gaps
 - Failure of independence assumption
- Hidden Markov Models

Likelihood Ratios

- The *likelihood* of a model M given an observation s is

$$L(M | s) = P(s | M)$$

This is *not* the *probability* of the model! – (the sum over all models is not 1).

- The *likelihood ratio* (LR) of two models M_a and M_0 is given by

$$LR(M_a, M_0 | s) = \frac{L(M_a | s)}{L(M_0 | s)}$$

The numerator and denominator may both be very small!

- The *log likelihood ratio* (LLR) is the logarithm of the likelihood ratio.

Weight Matrices for Site Models

- LR for sites: (prob under site model) / (prob under non-site (background) model)

$$\frac{P(s | M_{\text{site}})}{P(s | M_{\text{background}})} = \frac{\prod_{1 \leq i \leq n} P_i(s_i | M_{\text{site}})}{\prod_{1 \leq i \leq n} P_i(s_i | M_{\text{background}})}$$

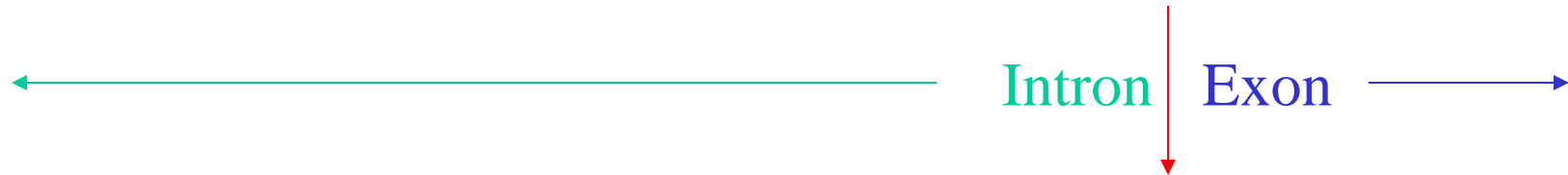
- $\text{LLR} = \sum_{1 \leq i \leq n} \log(P_i(s_i | M_{\text{site}})) - \log(P_i(s_i | M_{\text{background}}))$
 - compute by reading from a *matrix* whose i -th column contains values $\log(P_i(r | M_{\text{site}})) - \log(P_i(r | M_{\text{background}}))$ for each residue r (with r labelling the rows).
 - We use \log_2 .

Example: 3' splice sites in *C. elegans*

- For *background distribution* take
 - genomic residue freqs computed from *C. elegans* chrom. I:

A	4,575,132:	0.321
C	2,559,048:	0.179
G	2,555,862:	0.179
T	4,582,688:	0.321
 - other choices are possible, e.g. composition of *transcribed regions*
- For the *site distribution* we take
 - site residue freqs from 8192 sites:

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

Weight Matrix – 3' Splice Sites

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

Scoring a Candidate 3' Splice Site

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

T T C T T A C A G A A T

$$0.35 + 0.39 + -1.15 + 1.48 + 1.60 + -1.79 + 2.22 + 1.64 + 2.48 + 0.36 + -0.13 + -0.22 = 7.23$$

- General def.: a *weight matrix* W has entries w_{rj} indexed by residues $r \in A$, and $1 \leq j \leq n$
- *score* of a sequence $s = (s_1 s_2 \dots s_n)$ is

$$\sum_{1 \leq j \leq n} w_{s_j j}$$

- In the site case,

$$w_{rj} = \log(P_j(r | M_{\text{site}})) - \log(P_j(r | M_{\text{background}}))$$

Simple Hypothesis Testing

- Suppose we wish to decide between two models:
 - M_a (the *alternative hypothesis*), and
 - M_0 (the *null hypothesis*)

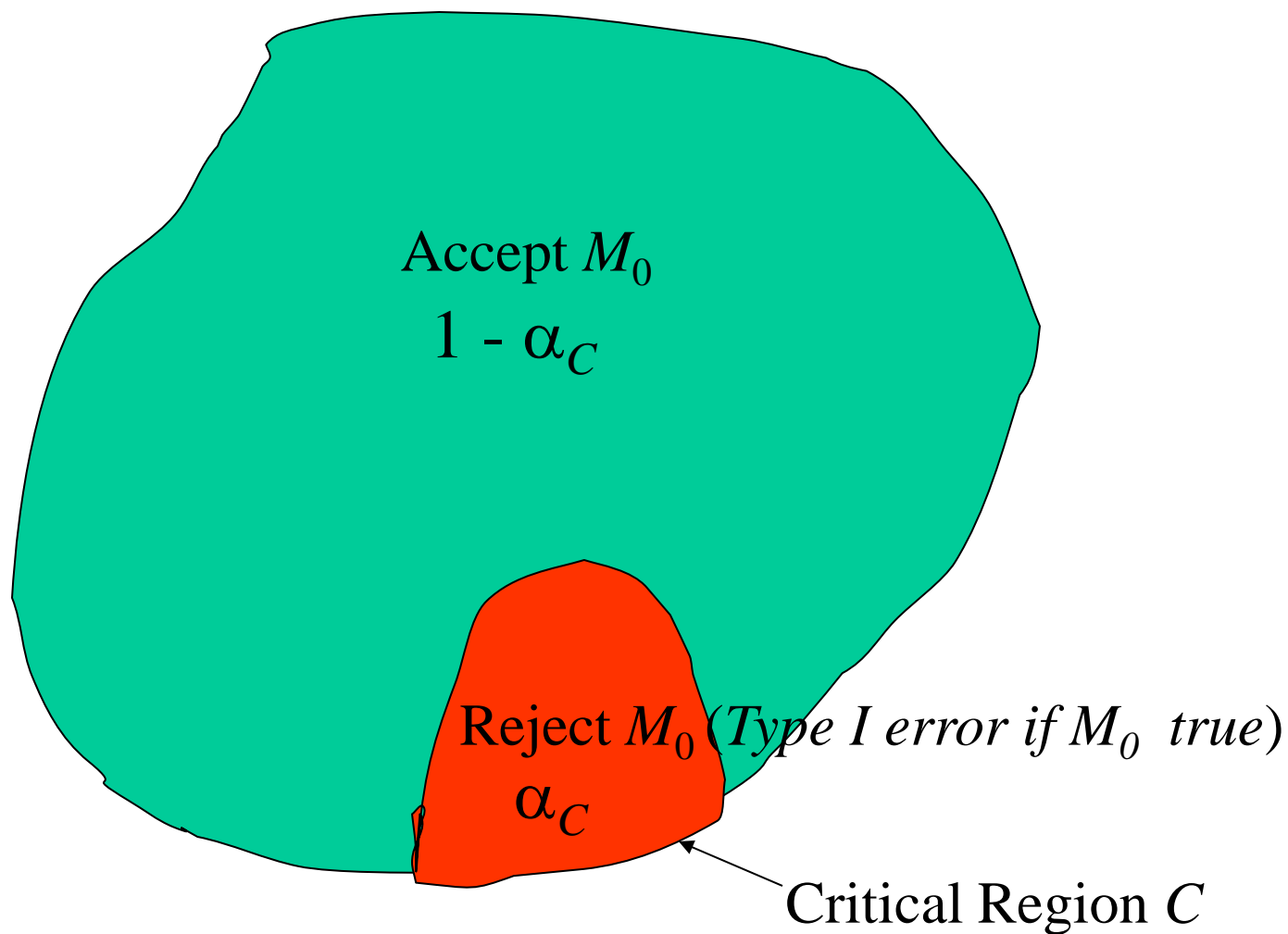
using an observation s from a sample space S . (e.g.

- s a sequence,
 - M_a a site model
 - M_0 a “background” (non-site) model.
- Strategy:
 - choose a subset $C \subset S$, called the *critical region* for the comparison.
 - If s falls within C , reject M_0 (accept M_a),
 - otherwise accept M_0 (reject M_a).

Types of Errors with Hypothesis Test

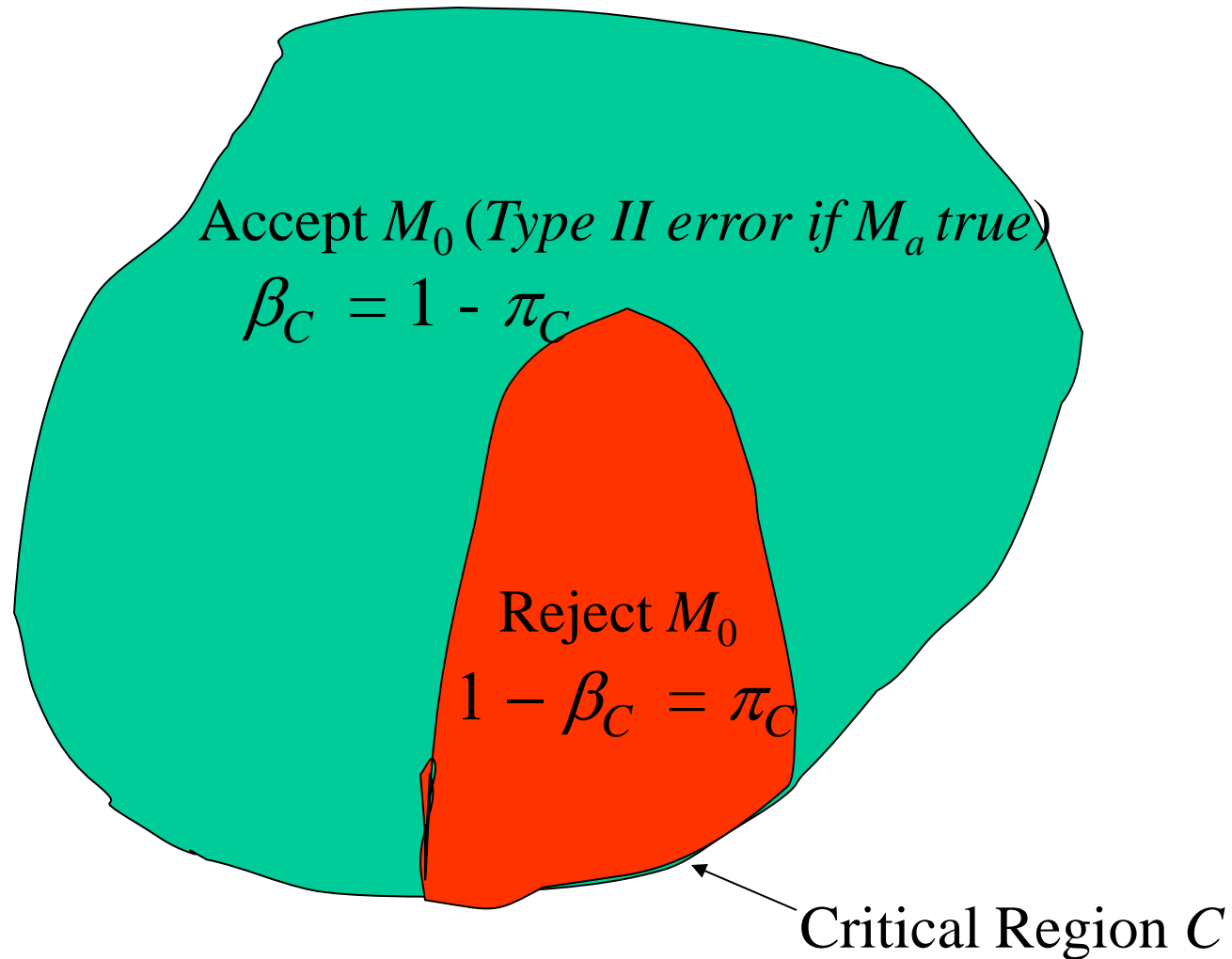
- a *Type I error* occurs if we reject M_0 when it is true.
 - For a given critical region C , the prob of committing a Type I error is denoted α_C
$$\alpha_C = P(C | M_0) = \sum_{s \in C} P(s | M_0)$$
- α_C is called the *significance level* of the test

Sample Space S – probabilities under M_0



- a *Type II error* occurs if we accept M_0 when it is false.
 - For a given C , prob of committing a Type II error is denoted β_C
$$\beta_C = \sum_{s \notin C} P(s | M_a) = 1 - P(C | M_a)$$
- $\pi_C = 1 - \beta_C$ is called the *power* of the test.

Sample Space S – probabilities under M_a



- Designing a test involves a tradeoff between significance and power
 - smaller C gives smaller Type I error but larger Type II error (lower power).

Likelihood Ratio Tests

- A *likelihood ratio test* of models M_a and M_0 is a hypothesis test of the two models, with critical region C defined by

$$C = C_\Lambda = \{s \mid LR(M_a, M_0 \mid s) \geq \Lambda\}$$

for some non-negative constant Λ , the *cutoff value*.

- Neyman-Pearson lemma motivates use of the *likelihood ratio* as an optimal *discriminator*, or “score”
 - even in contexts where we aren’t explicitly testing hypotheses.
- any monotonic function $f(LR)$ of likelihood ratio has equivalent optimality properties
 - because defines the same set of critical regions:

$$LR(M_a, M_0 | s) \geq \Lambda \Leftrightarrow f(LR(M_a, M_0 | s)) \geq f(\Lambda)$$
- convenient to take f to be the log function, in which case we get the *log likelihood ratio*.

Neyman-Pearson lemma

Let M_a and M_0 be two models, and C_A the critical region defined by a likelihood ratio test of M_a vs. M_0 with

- cutoff value Λ ,
- significance level α_A , and
- power $\pi_A = 1 - \beta_A$.

Then if C is any other critical region, we have

- If $\alpha_C < \alpha_A$, then $\pi_C < \pi_A$ (and $\beta_C > \beta_A$)
- If $\alpha_C = \alpha_A$, then $\pi_C \leq \pi_A$ (and $\beta_C \geq \beta_A$)

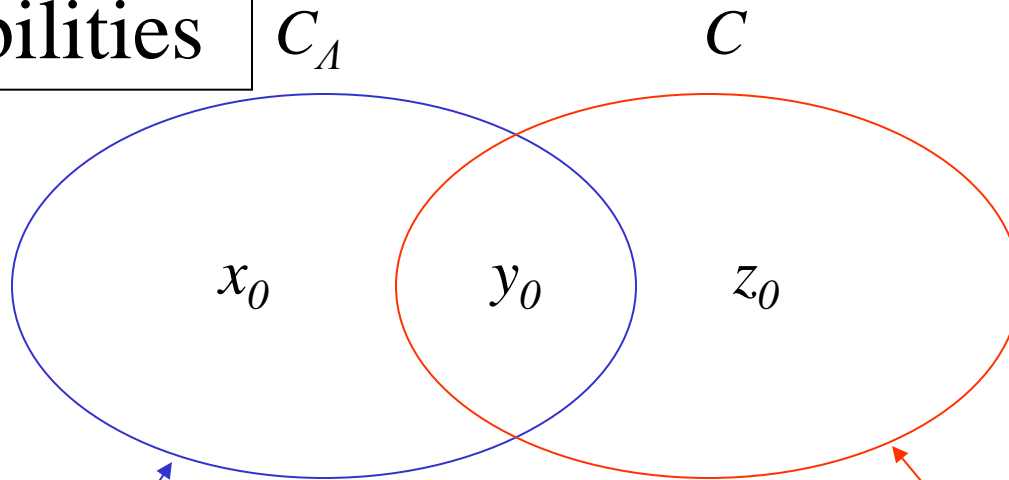
In other words, the likelihood ratio test with significance level α_A is the most powerful test

- (has the lowest type II error rate)

with that significance level.

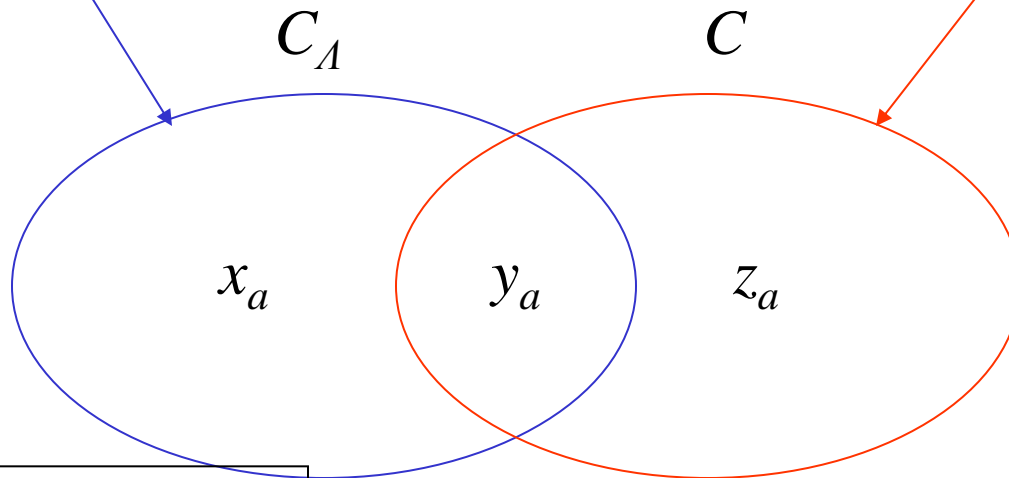
Idea of Neyman-Pearson lemma *proof*:

M_0 probabilities



$$x_a \geq \Lambda x_0$$

$$z_a < \Lambda z_0$$



M_a probabilities

$$\begin{aligned} \alpha_C &< \alpha_A \\ \Rightarrow z_0 &< x_0 \\ \Rightarrow \Lambda z_0 &< \Lambda x_0 \\ \Rightarrow z_a &< x_a \\ \Rightarrow \pi_C &< \pi_A \end{aligned}$$

- ***Proof:*** Suppose $\alpha_C < \alpha_A$. Then

$$\sum_{s \in C} P(s | M_0) < \sum_{s \in C_A} P(s | M_0)$$

Subtract from both sides the terms involving $s \in C \cap C_A$. This leaves

$$(1) \quad \sum_{s \in C \setminus C_A} P(s | M_0) < \sum_{s \in C_A \setminus C} P(s | M_0)$$

- By definition of the likelihood ratio test, for any observation s ,

$$s \in C_\Lambda \Leftrightarrow P(s | M_a) \geq \Lambda P(s | M_0)$$

- From this, it follows that

$$(2) \quad \sum_{s \in C \setminus C_\Lambda} \frac{1}{\Lambda} P(s | M_a) < \sum_{s \in C \setminus C_\Lambda} P(s | M_0)$$

and

$$(3) \quad \sum_{s \in C_\Lambda \setminus C} P(s | M_0) \leq \sum_{s \in C_\Lambda \setminus C} \frac{1}{\Lambda} P(s | M_a)$$

- Combining (2), (1), and (3)

$$\sum_{s \in C \setminus C_A} \frac{1}{\Lambda} P(s | M_a) < \sum_{s \in C \setminus C_A} P(s | M_0) < \sum_{s \in C_A \setminus C} P(s | M_0) \leq \sum_{s \in C_A \setminus C} \frac{1}{\Lambda} P(s | M_a)$$

so (cancelling the common factor $1 / \Lambda$)

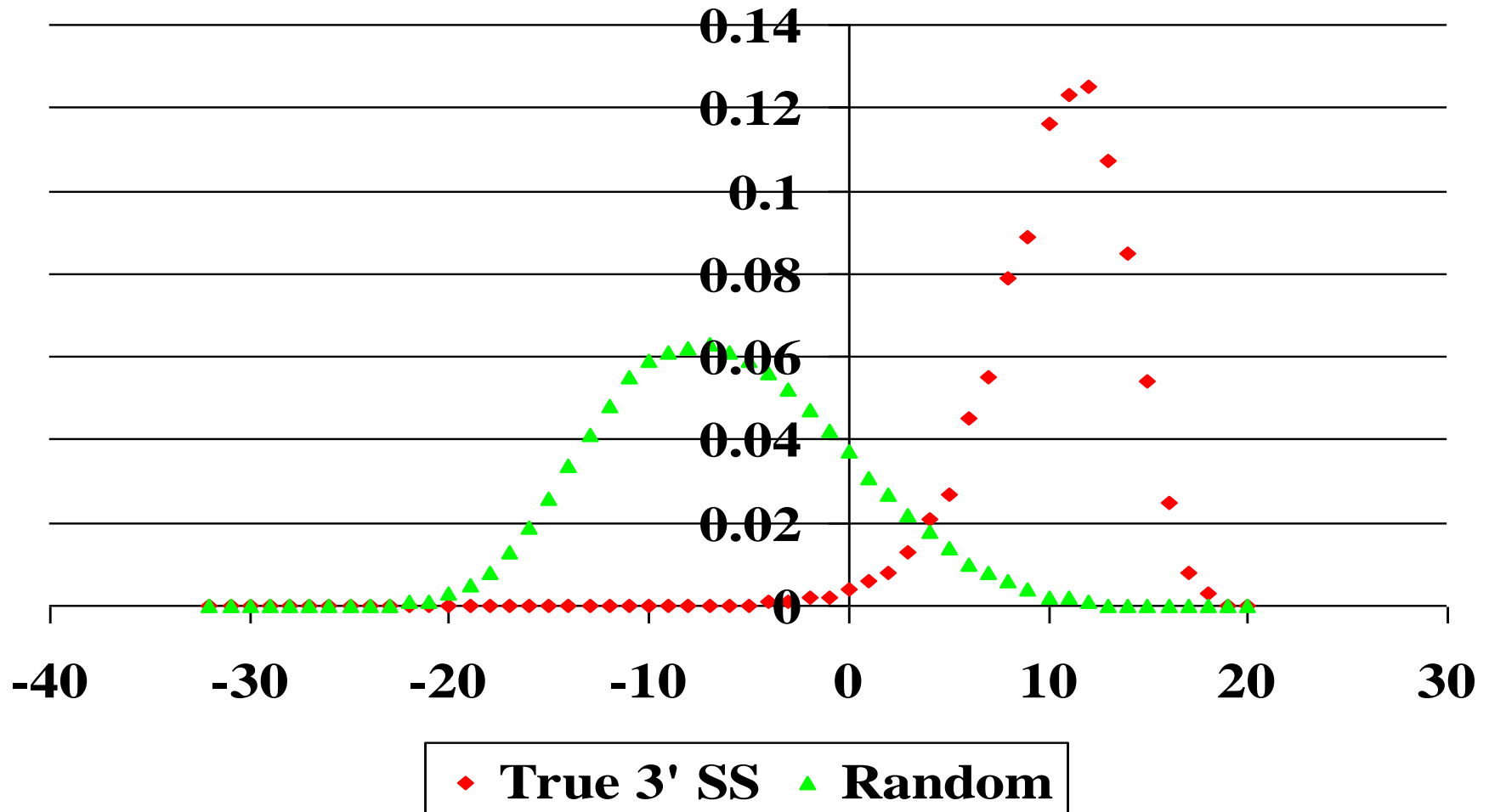
$$\sum_{s \in C \setminus C_A} P(s | M_a) < \sum_{s \in C_A \setminus C} P(s | M_a)$$

so, adding in the terms corresponding to $s \in C \cap C_A$

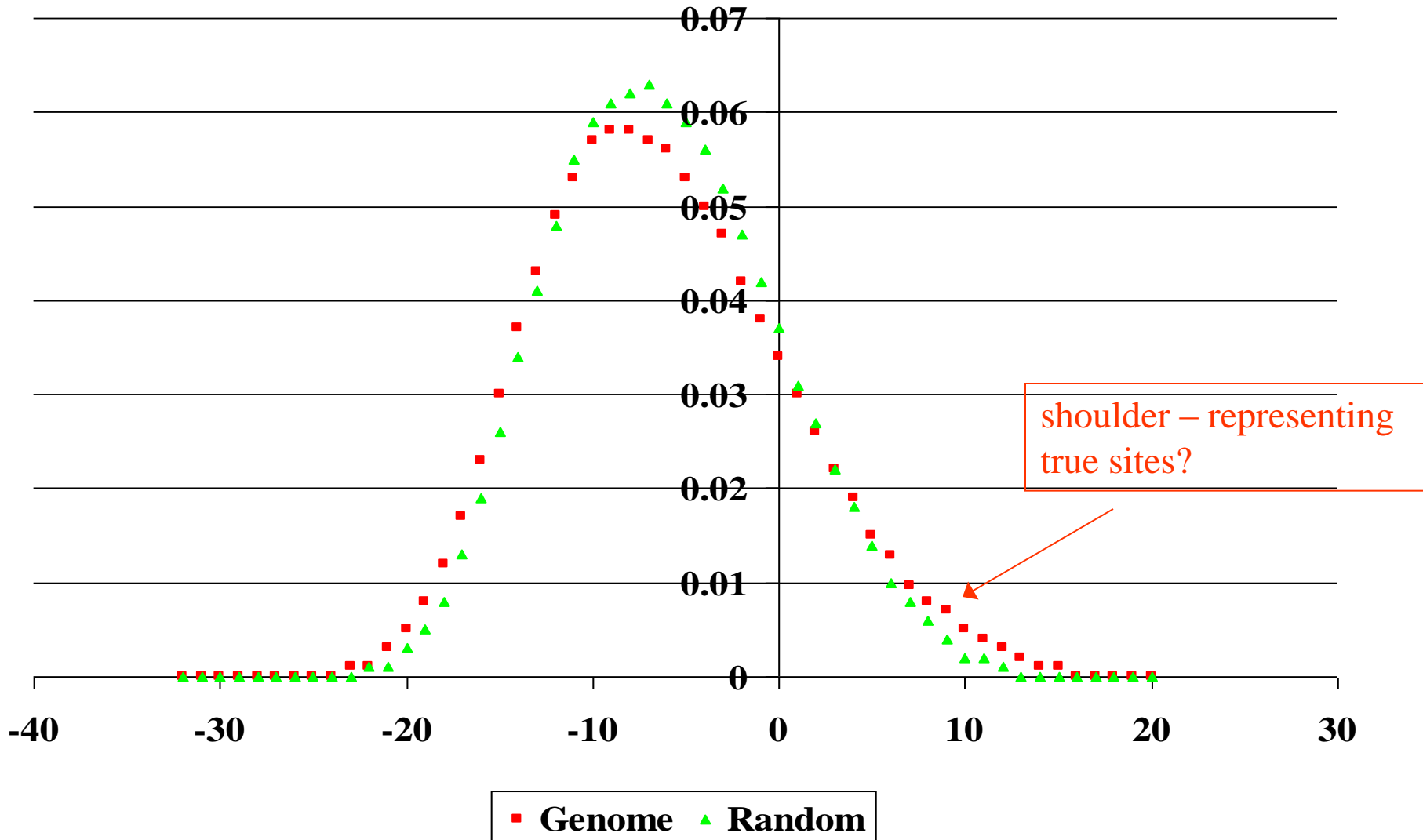
$$\sum_{s \in C} P(s | M_a) < \sum_{s \in C_A} P(s | M_a)$$

i.e $\pi_C < \pi_A$ The other part of the lemma ($\pi_C \leq \pi_A$ if $\alpha_C = \alpha_A$) is proved similarly.

Score Distributions (AG sites)– 3' SS Weight Matrix



Score Distributions (AG sites)– 3' SS Weight Matrix



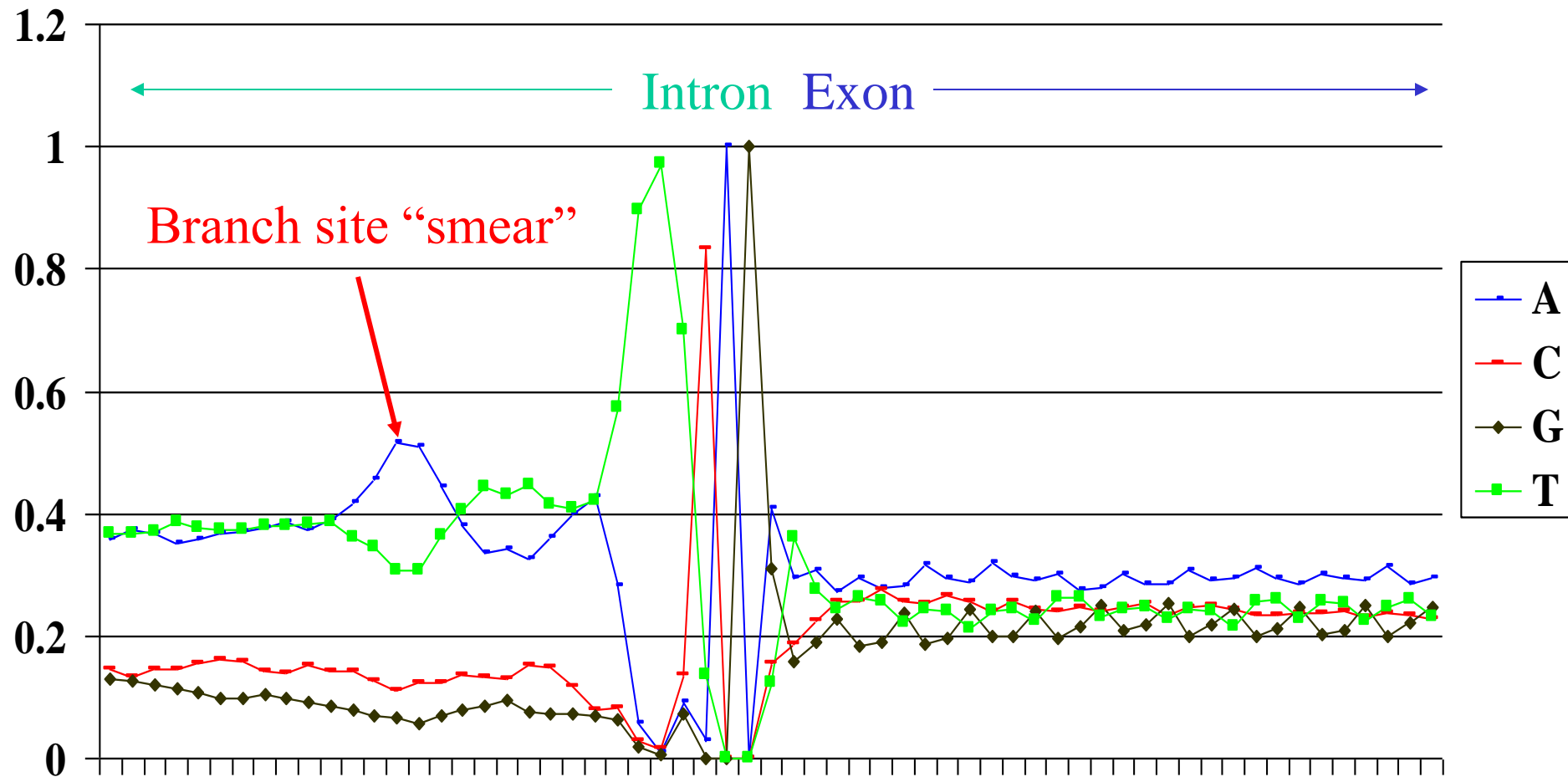
Some Issues for Site Weight Matrices (to be discussed later)

- Can derive *theoretical* probability distribution for scores, and compare with above *empirical* distributions
- Small sample correction to frequencies: pseudocounts
- Avoiding *overfitting* (e.g. using too large a window)

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 :

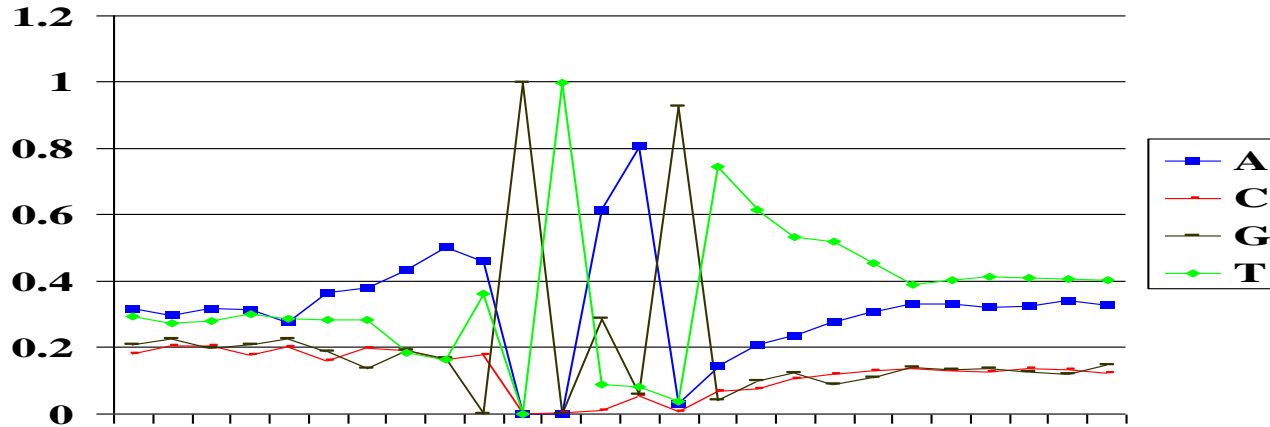
A	1434	1664	1518	0	0	2032	2662	98	479	694	783	912
C	633	546	583	0	5	36	177	22	225	250	350	393
G	628	553	0	3301	0	943	187	3063	134	329	405	279
T	606	538	1200	0	3296	290	275	118	2463	2028	1763	1717
A	0.434	0.504	0.460	0.000	0.000	0.616	0.806	0.030	0.145	0.210	0.237	0.276
C	0.192	0.165	0.177	0.000	0.002	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 :

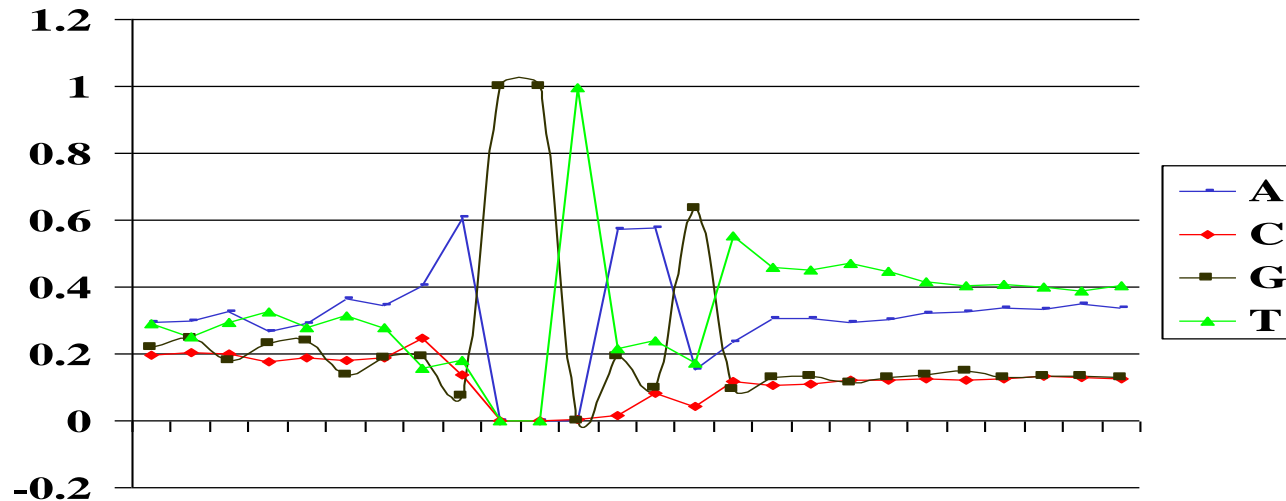
A	1970	2980	0	0	0	2804	2824	739	1153	1495	1495	1443
C	1217	678	0	0	9	82	411	215	576	521	539	593
G	934	359	4891	4891	0	947	485	3101	455	633	651	548
T	770	874	0	0	4882	1058	1171	836	2707	2242	2206	2307
A	0.403	0.609	0.000	0.000	0.000	0.573	0.577	0.151	0.236	0.306	0.306	0.295
C	0.249	0.139	0.000	0.000	0.002	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.

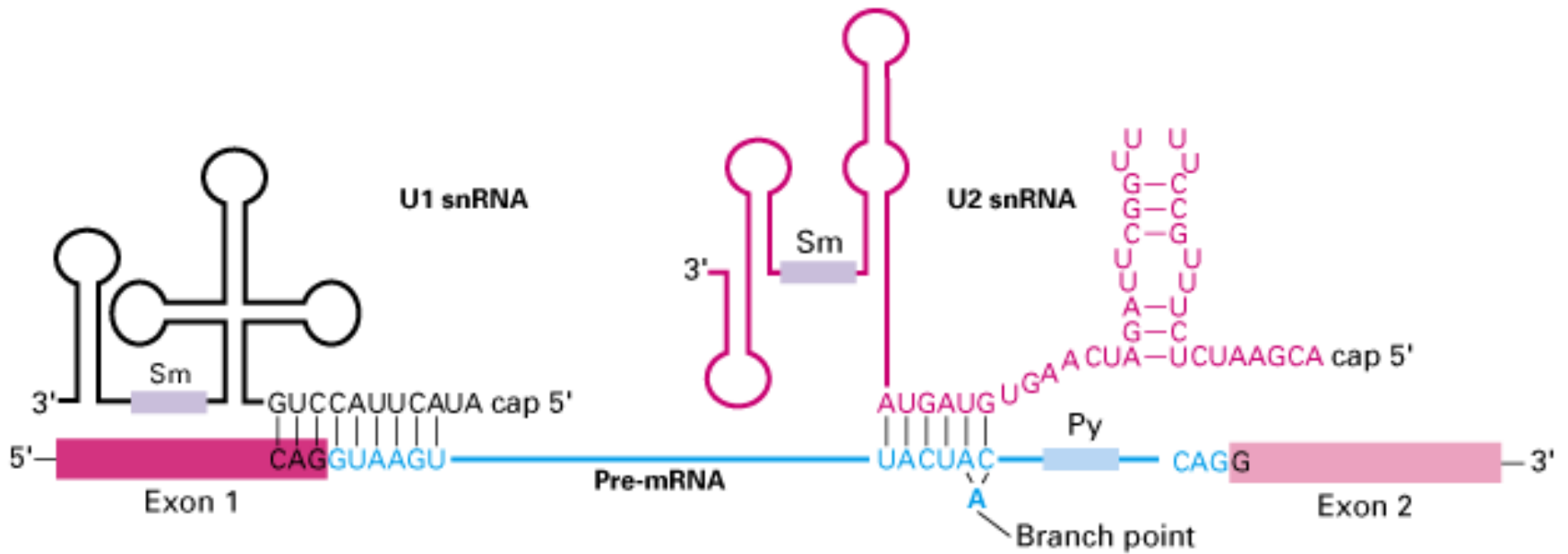
Nucleotide Counts for 8192 *C. elegans* 5' Splice Sites



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				Expected (under independence)			
	A	C	G	T	A	C	G	T
A	0.135	0.047	0.051	0.088	0.103	0.057	0.057	0.103
C	0.061	0.035	0.033	0.051	0.057	0.032	0.032	0.058
G	0.063	0.034	0.034	0.047	0.057	0.032	0.032	0.057
T	0.061	0.064	0.061	0.135	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 *TpA*: found in nearly all genomes;
 - reason unknown:
 - neutral (mutation patterns)?
 - selection?
- Overrepresentation of *ApA*, *TpT*, *CpC*, *GpG* – also frequently observed in other organisms.
- Unlike mammalian genomes, no underrepresentation of *CpG*
 - *CpG* not methylated in *C. elegans* (or most other non-vertebrates).

Dinucleotide Freqs – *H. sapiens* Chr.21

Nucleotide Freqs:

A 10032226 0.297; T 9962530 0.295

G 6908202 0.204; C 6921020 0.205

Entropy: 1.976 bits

Observed Dinuc Freqs

Expected (*under independence*)

	A	C	G	T	A	C	G	T
A	0.099	0.051	0.069	0.078	0.088	0.061	0.061	0.087
C	0.073	0.052	0.011	0.069	0.061	0.042	0.042	0.060
G	0.059	0.043	0.052	0.050	0.061	0.042	0.042	0.060
T	0.066	0.059	0.072	0.098	0.087	0.060	0.060	0.087

Observed / Expected

	A	C	G	T
A	1.124	0.839	1.139	0.891
C	1.204	1.243	0.260	1.139
G	0.974	1.025	1.245	0.839
T	0.752	0.976	1.204	1.125

Dinucleotide Freqs – *H. sapiens* Chr.22

Nucleotide Freqs:

A 8745910 0.261; T 8720493 0.261

G 7999585 0.239; C 7997931 0.239

Entropy: 1.999 bits

Observed Dinuc Freqs

Expected (*under independence*)

	A	C	G	T	A	C	G	T
A	0.077	0.051	0.075	0.058	0.068	0.062	0.062	0.068
C	0.077	0.071	0.016	0.075	0.062	0.057	0.057	0.062
G	0.061	0.057	0.071	0.051	0.062	0.057	0.057	0.062
T	0.047	0.061	0.077	0.076	0.068	0.062	0.062	0.068

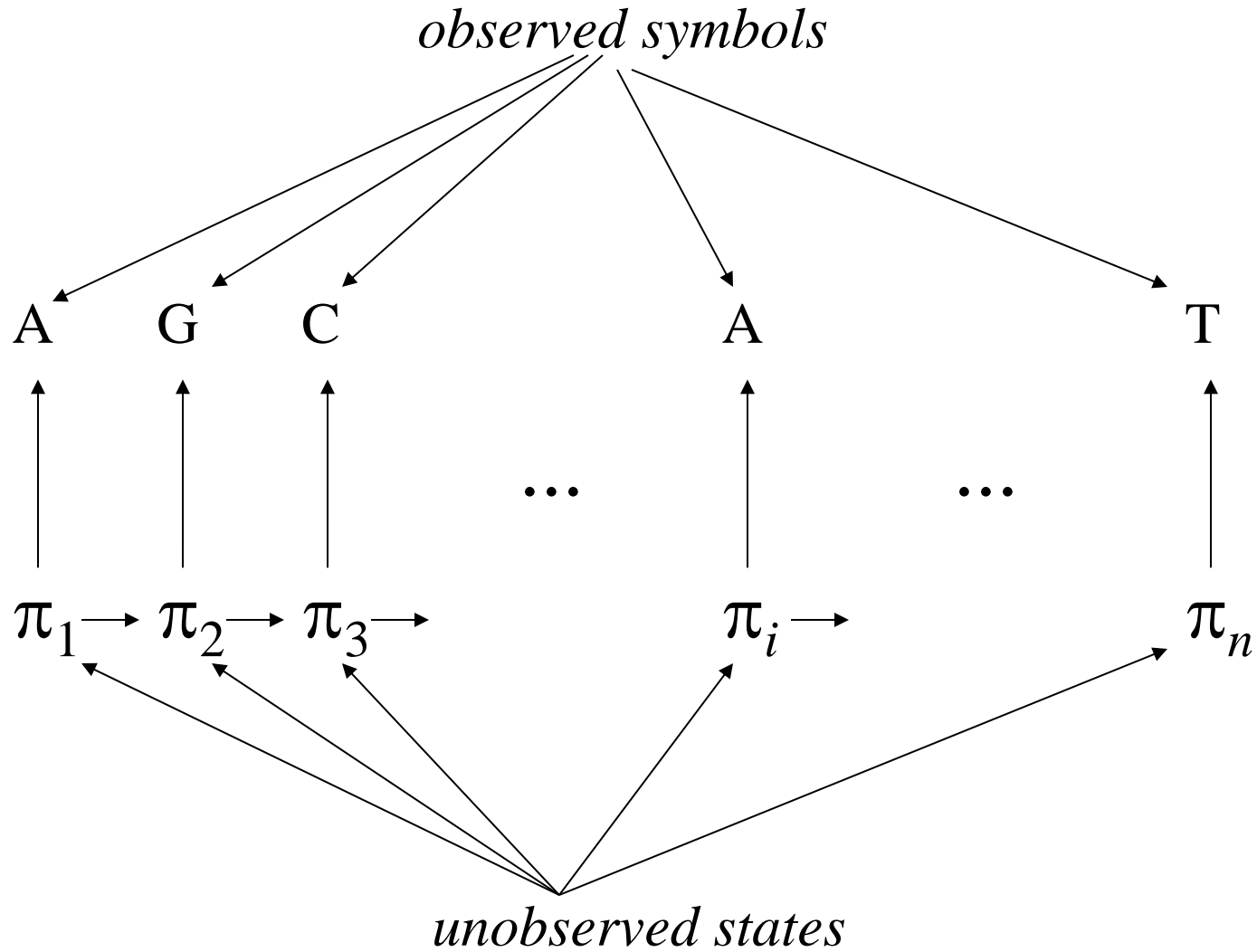
Observed / Expected

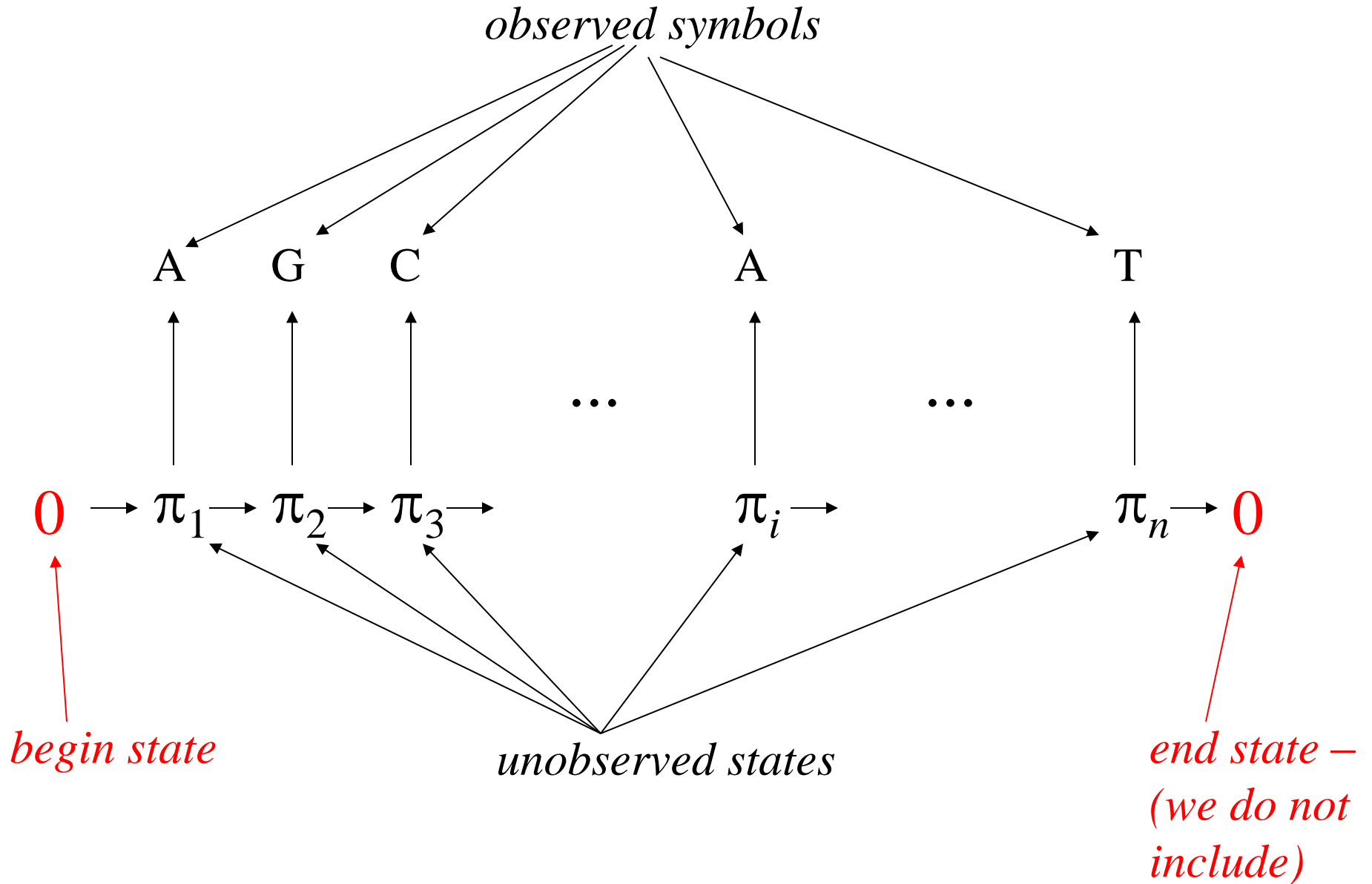
	A	C	G	T
A	1.125	0.817	1.205	0.855
C	1.233	1.236	0.285	1.206
G	0.975	0.989	1.237	0.818
T	0.684	0.977	1.233	1.124

Hidden Markov Models

- Probability models for sequences of *observed symbols*, e.g.
 - nucleotide or amino acid residues
 - aligned pairs of residues
 - aligned set of residues corresponding to leaves of an underlying evolutionary tree
 - angles in protein chain (structure modelling)
 - sounds (speech recognition)

- Assume a sequence of “*hidden*” (unobserved) *states* underlies each observed symbol sequence
- Each state “*emits*” symbols (one symbol at a time)
- States may correspond to underlying “reality” we are trying to infer, e.g.
 - unobserved biological feature:
 - (positions within) site,
 - coding region of gene
 - rate of evolution
 - protein structural element
 - speech phoneme





Advantages of HMMs

- Flexible –gives reasonably good models in wide variety of situations
- Computationally efficient
- Often interpretable:
 - hidden states can correspond to biological features.
 - can find most probable sequence of hidden states
= biological “parsing” of residue sequence.