Lecture 4

- Comparing probability models: *likelihood ratios*
  - Hypothesis testing
  - Neyman-Pearson lemma
- Weight matrices
- Score distributions

Comparing Alternative Probability Models

- We will want to consider more than one model at a time, in following situations:
  - To differentiate between two or more hypotheses about a sequence
  - To generate increasingly refined probability models that are progressively more accurate

- First situation arises in testing biological assertion, e.g. “is this a coding sequence?”
  - Compare two models:
  1. model associated with a hypothesis H_{coding},
     - assigns each sequence the prob of observing it under expt of drawing a coding sequence at random from genome
  2. model associated with a hypothesis $H_{noncoding}$,
     - assigns each sequence the prob of observing it under expt of drawing a non-coding sequence at random

Likelihood Ratios

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

$$L(M \mid s) = P(s \mid M)$$

  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

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

  The numerator and denominator may both be very small!

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

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).

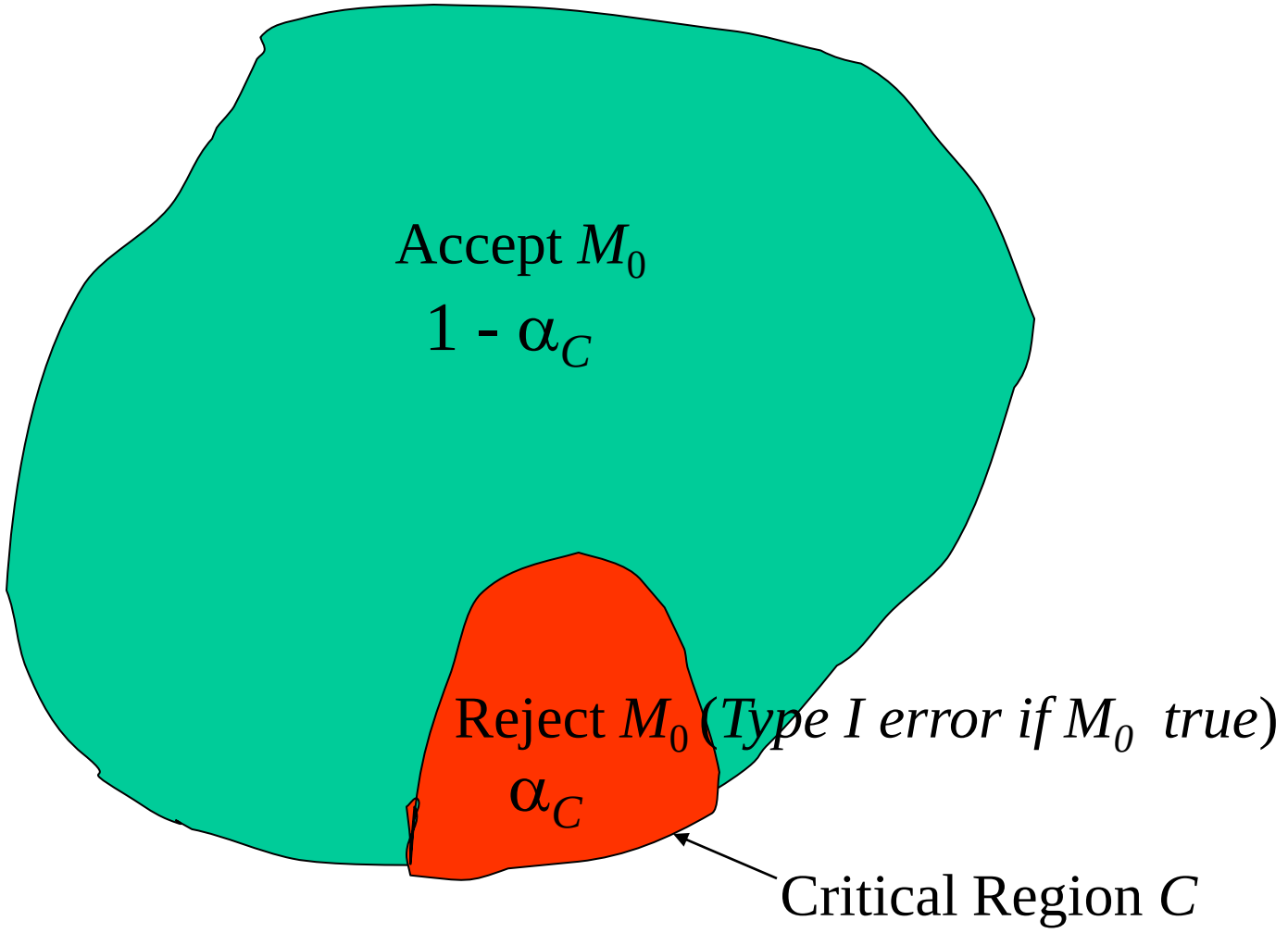
Types of Errors with Hypothesis Test

- 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

$$\alpha_C = P(C \mid M_0) = \Sigma_{s \in C} P(s \mid M_0)$$

- α_C is called the *significance level* of the test

Sample Space S – probabilities under M_0

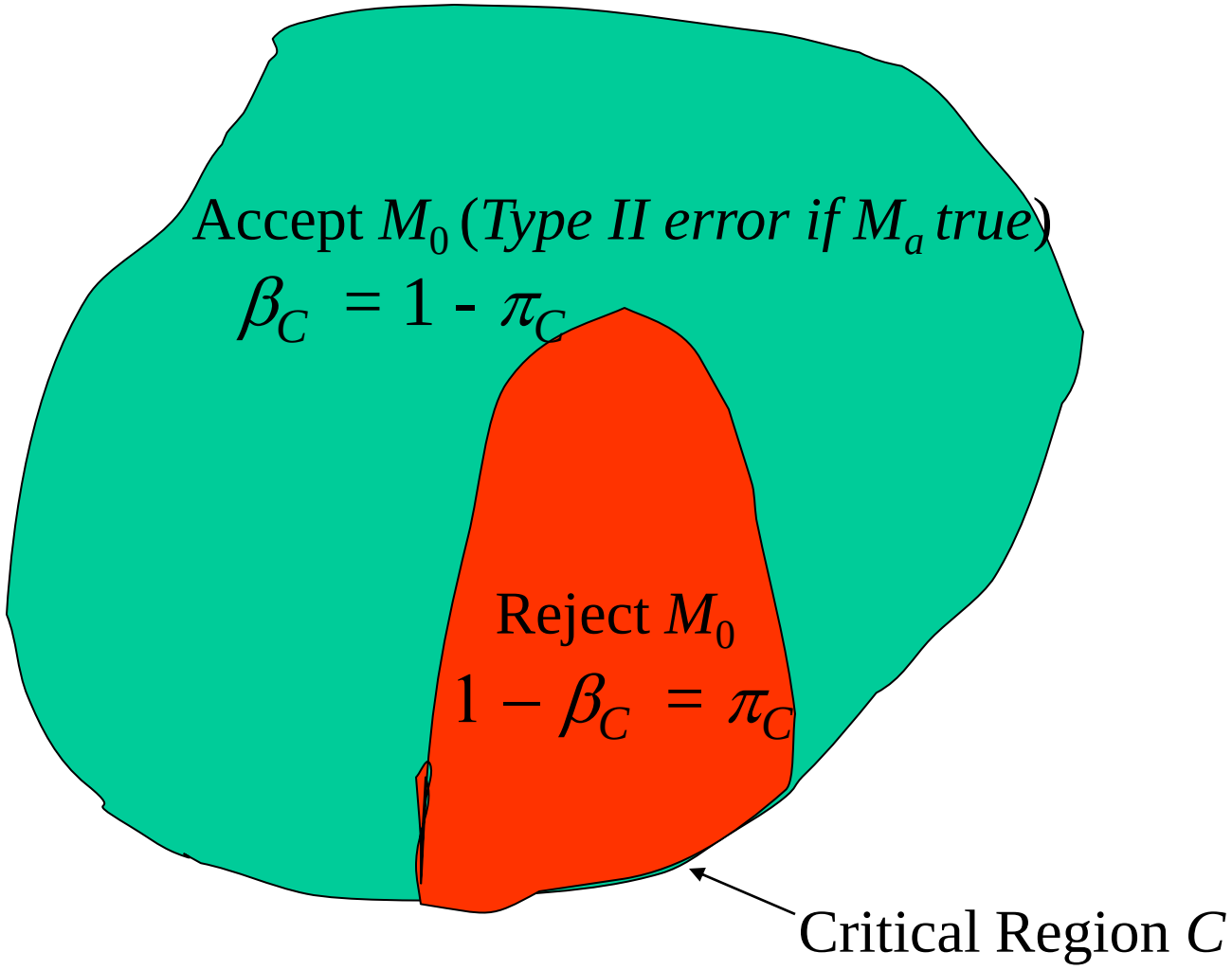


- a *Type II error* occurs if we accept M_0 when it is false.
  - For a given C, prob of committing a Type II error is denoted β_C

$$\beta_C = \Sigma_{s \notin C} P(s \mid M_a) = 1 - P(C \mid M_a)$$

- $\pi_C = 1 - \beta_C$ 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:

$$LR(M_a, M_0 \mid s) \geq \Lambda \Leftrightarrow f(LR(M_a, M_0 \mid s)) \geq f(\Lambda)$$

- 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_Λ the critical region defined by a likelihood ratio test of M_a vs. M_0 with

- cutoff value Λ,
- significance level α_Λ, and
- power $\pi_\Lambda = 1 - \beta_\Lambda$.

Then if C is any other critical region, we have

- If $\alpha_C < \alpha_\Lambda$, then $\pi_C < \pi_\Lambda$ (and $\beta_C > \beta_\Lambda$)
- If $\alpha_C = \alpha_\Lambda$, then $\pi_C \leq \pi_\Lambda$ (and $\beta_C \geq \beta_\Lambda$)

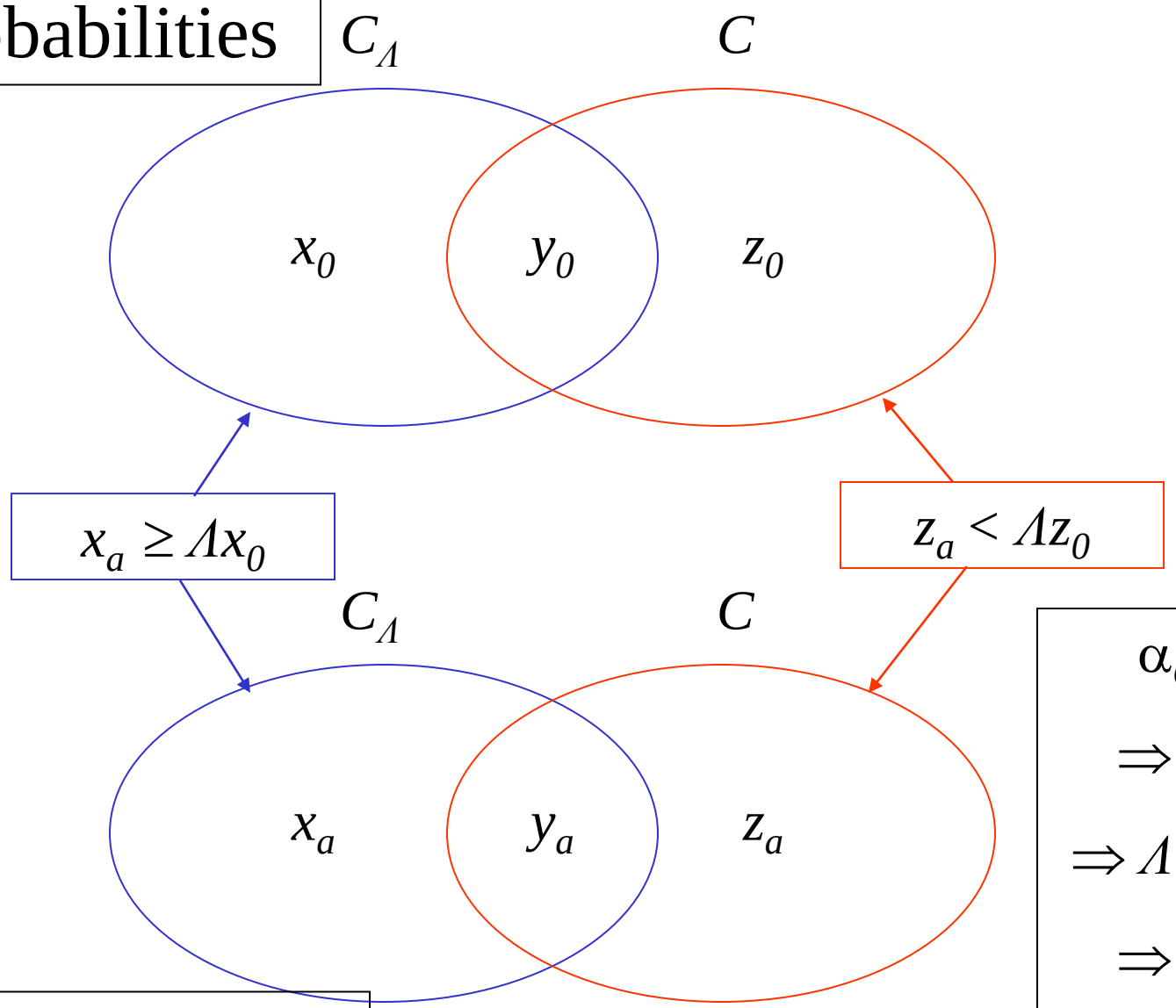
In other words, the likelihood ratio test with significance level α_Λ is the most powerful test

- (has the lowest type II error rate)

with that significance level.

Idea of Neyman-Pearson lemma *proof*:

M_0 probabilities



M_a probabilities

$$\alpha_C < \alpha_\Lambda$$

$$\Rightarrow z_0 < x_0$$

$$\Rightarrow \Lambda\, z_0 < \Lambda\, x_0$$

$$\Rightarrow z_a < x_a$$

$$\Rightarrow \pi_C < \pi_\Lambda$$

- ***Proof***: Suppose $\alpha_C < \alpha_\Lambda$ .Then

$$\sum_{s \in C} P(s \mid M_0) < \sum_{s \in C_\Lambda} P(s \mid M_0)$$

Subtract from both sides the terms involving $s \in C \cap C_\Lambda$ This leaves

$$(1) \quad \sum_{s \in C \setminus C_\Lambda} P(s \mid M_0) < \sum_{s \in C_\Lambda \setminus C} P(s \mid M_0)$$

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

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

- From this, it follows that

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

and

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

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

$$\sum_{s\in C\setminus C_\Lambda}\frac{1}{\Lambda}P(s\mid M_a)<\sum_{s\in C\setminus C_\Lambda}P(s\mid M_0)<\sum_{s\in C_\Lambda\setminus C}P(s\mid M_0)\le\sum_{s\in C_\Lambda\setminus C}\frac{1}{\Lambda}P(s\mid M_a)$$

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

$$\sum_{s\in C\setminus C_\Lambda}P(s\mid M_a)<\sum_{s\in C_\Lambda\setminus C}P(s\mid M_a)$$

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

$$\sum_{s\in C}P(s\mid M_a)<\sum_{s\in C_\Lambda}P(s\mid M_a)$$

i.e $\pi_C<\pi_\Lambda$ The other part of the lemma ($\pi_C\le\pi_\Lambda$ if $\alpha_C=\alpha_\Lambda$) is proved similarly.

Weight Matrices for Site Models

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

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

- $\text{LLR} = \sum_{1 \le i \le n} \log(P_i(s_i \mid M_{\text{site}})) - \log(P_i(s_i \mid M_{\text{background}}))$
  - compute by reading from a *matrix* whose *i*-th column contains values $\log(P_i(r \mid M_{\text{site}})) - \log(P_i(r \mid M_{\text{background}}))$ 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:

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 $1 \le j \le n$

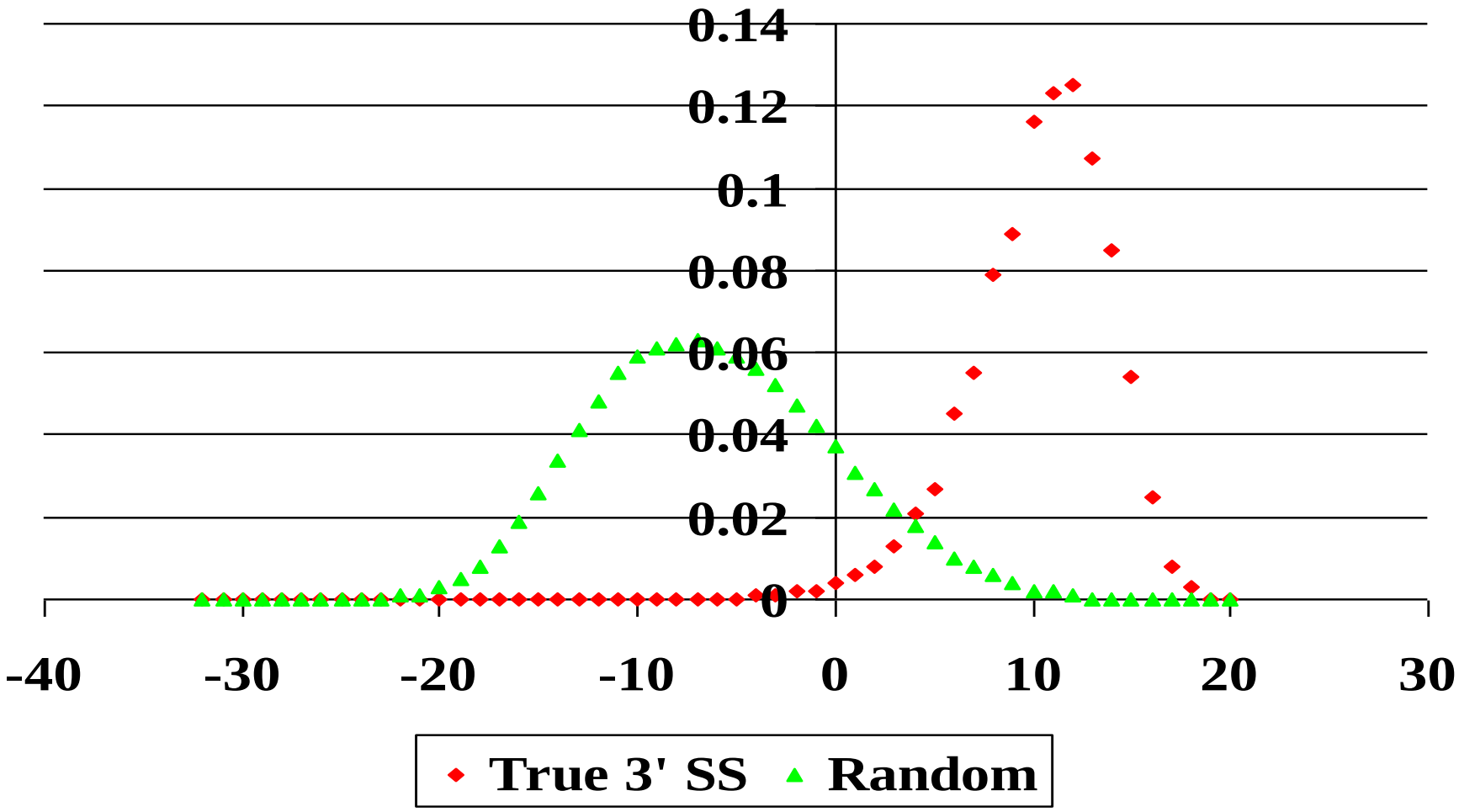
- *score* of a sequence $s = (s_1 s_2 \ldots s_n)$ is

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

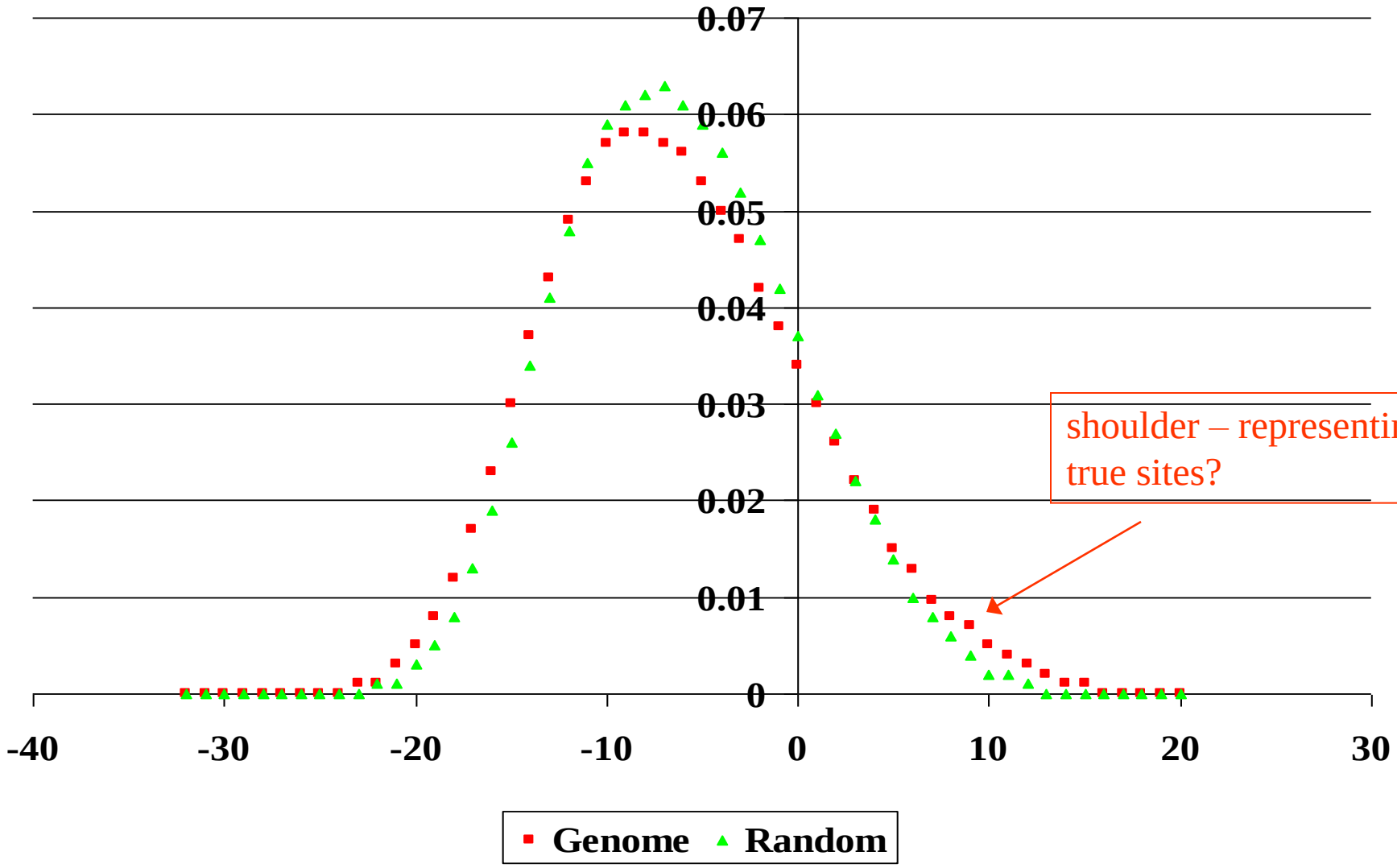
- In the site case,

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

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)