Today's Lecture

- Probability models for sequences
- Failure of equal frequency assumption
- Neutralist vs selectionist interpretations
- Comparing probability models: likelihood ratios
  - Hypothesis testing

Basic Probability Theory Concepts

- A *sample space* S is set of all possible outcomes of a conceptual, repeatable experiment.
  - $|S| < \infty$ in most of our examples.
  - e.g. S = all possible sequences of a given length.
- Elements of S are called *sample points*.
  - e.g. a particular seq = outcome of "experiment" of extracting seq of specified type from a genome.
- A *probability distribution* P on S assigns non-neg real number $P(s)$ to each $s \in S$, such that

$$\Sigma_{s \in S} P(s) = 1$$

(So $0 \leq P(s) \leq 1 \quad \forall s$)

  - Intuitively, $P(s)$ = fraction of times one would get s as result of the expt, if repeated many times.

- A *probability space* (S,P) is a sample space S with a prob dist'n P on S.
- Prob dist'n on S is sometimes called a *probability model* for S, particularly if several dist'ns are being considered.
  - Write models as M_1, M_2 , probabilities as $P(s \mid M_1)$, $P(s \mid M_2)$.
  - e.g.
    - M_1 = prob dist'n for splice site seqs,
    - M_2 = prob dist'n for "background" (arbitrary genomic) seqs.

Basic Probability Theory Concepts (cont'd)

- An *event* E is a criterion that is true or false for each $s \in S$.
  - defines a subset of S (sometimes also denoted E).
  - $P(E)$ is defined to be $\Sigma_{s|E \text{ is true}} P(s)$.
- Events $E_1, E_2, \ldots, E_n$ are *mutually exclusive* if no two of them are true for the same point;
  - then $P(E_1 \text{ or } E_2 \text{ or } \ldots \text{ or } E_n) = \Sigma_{1 \le i \le n} P(E_i)$.
- If $E_1, E_2, \ldots, E_n$ are also *exhaustive*, i.e. every s in S satisfies E_i for some i, then $\Sigma_{1 \le i \le n} P(E_i) = 1$.

- For events E and H, the *conditional probability* of E given H, is

$$P(E \mid H) \equiv P(E \text{ and } H) / P(H)$$

(= prob that both E and H are true, given H is true)

  – undefined if $P(H) = 0$.

- E and H are (*statistically*) *independent* if

$$P(E) = P(E \mid H)$$

(i.e. prob. E is true doesn't depend on whether H is true);

or equivalently

$$P(E \text{ and } H) = P(E)P(H).$$

Probabilities on Sequences

- Let S = space of DNA or protein sequences of length n. Possible assumptions for assigning probabilities to S:
  - *Equal frequency assumption:* All residues are equally probable at any position;
    - $P(E_r^{(i)}) = P(E_q^{(i)})$ for any two residues r and q,
      - where $E_r^{(i)}$ means residue r occurs at position i, then
    - Since for fixed i the $E_r^{(i)}$ are mutually exclusive and exhaustive,

      $$P(E_r^{(i)}) = 1 / |A|$$

      where A = residue alphabet

      $P(E_r^{(i)}) = 1/20$ for proteins, 1/4 for DNA).
  - *Independence assumption*: whether or not a residue occurs at a given position is independent of residues at other positions.

- Given above assumptions, the probability of the sequence

  $s = ACGCG$

  (in the space S of all length 5 sequences) is calculated by considering 5 events:

  - Event 1 is that first nuc is A. Probability = .25.
  - Event 2 is that 2^d nuc is C. Probability = .25.
  - Event 3 is that 3^d nuc is G. Probability = .25.
  - Event 4 is that 4^{th} nuc is C. Probability = .25.
  - Event 5 is that 5^{th} nuc is G. Probability = .25.

  By independence assumption, prob of all 5 events occurring is the product $(.25)^5 = 1/1024$.

  Since s is the only sequence satisfying all 5 conditions, $P(s) = 1/1024$.

- More generally, under equal freq and indep assumptions,

  prob of nuc sequence of length n = $.25^n$,

  prob of protein sequence of length n = $.05^n$

  in the space S of length n sequences.

Failure of Equal Frequency Assumption for (Real) DNA

- For most organisms, the nucleotide composition is significantly different from .25 for each nucleotide, e.g.:
  - *H. influenza* .31 A, .19 C, .19 G, .31 T
  - *P. aeruginosa* .17 A, .33 C, .33 G, .17 T
  - *M. janaschii* .34 A, .16 C, .16 G, .34 T
  - *S. cerevisiae* .31 A, .19 C, .19 G, .31 T
  - *C. elegans* .32 A, .18 C, .18 G, .32 T
  - *H. sapiens* .29 A, .21 C, .21 G, .29 T

- Note approximate symmetry: A ≅ T, C ≅ G,
  - even though we're counting nucs on just one strand.
  - Expect *exact* equality when counting both strands
- Explanation:
  - Although individual biological features may have non-symmetric composition (local *asymmetry*),
  - usually features are distributed approx *randomly* w.r.t. strand,
  - so local asymmetries *cancel*, yielding overall symmetry.

General Hypotheses Regarding Unequal Frequency

- Neutralist hypothesis: *mutation bias*
  - e.g. due to nucleotide pool composition
- Selectionist hypothesis: *selection*
  - selection on (many) particular nucleotides
  - selection on mutational bias mechanisms
  - …

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).

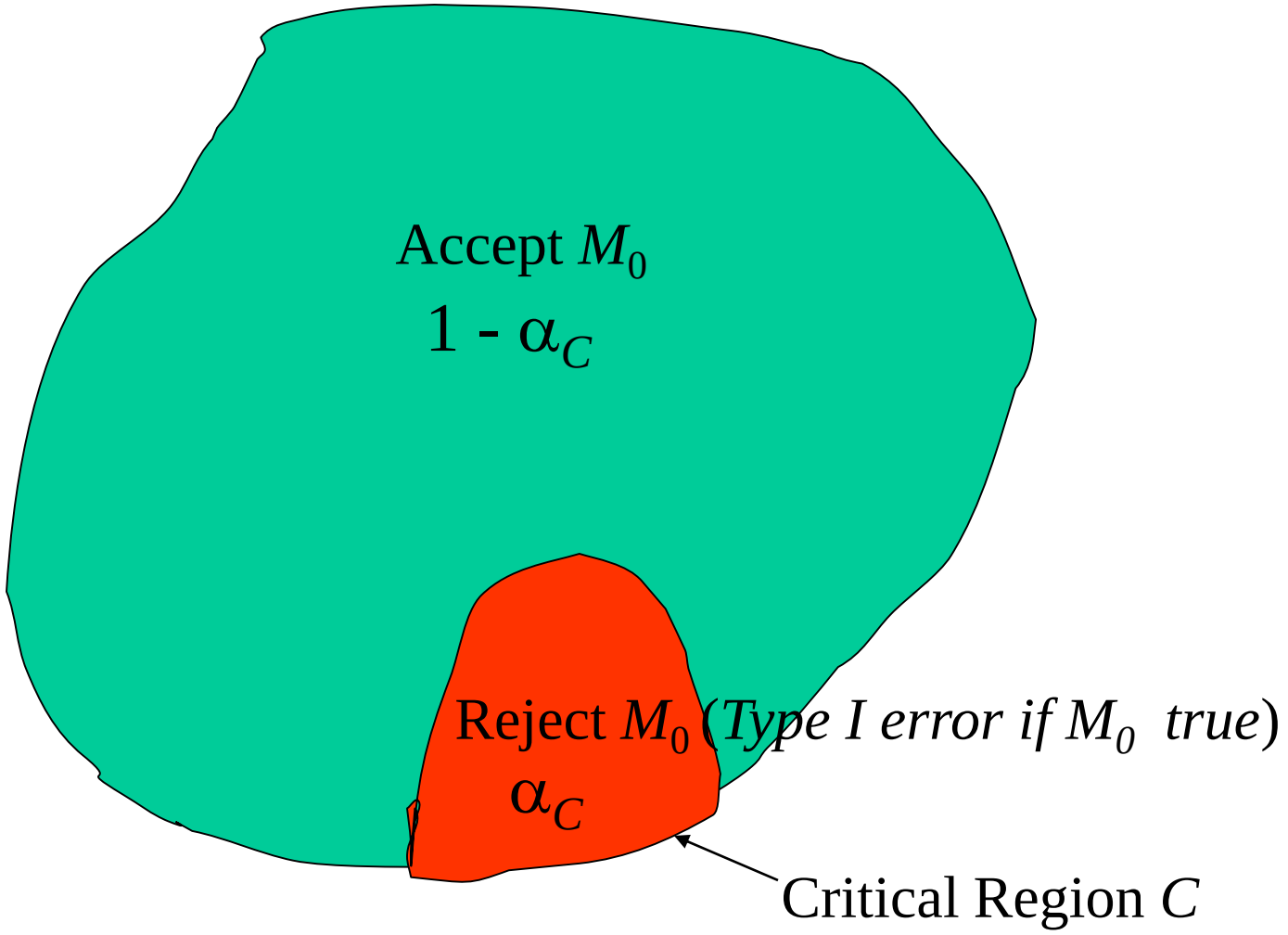
Types of Errors with Hypothesis Test

- 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

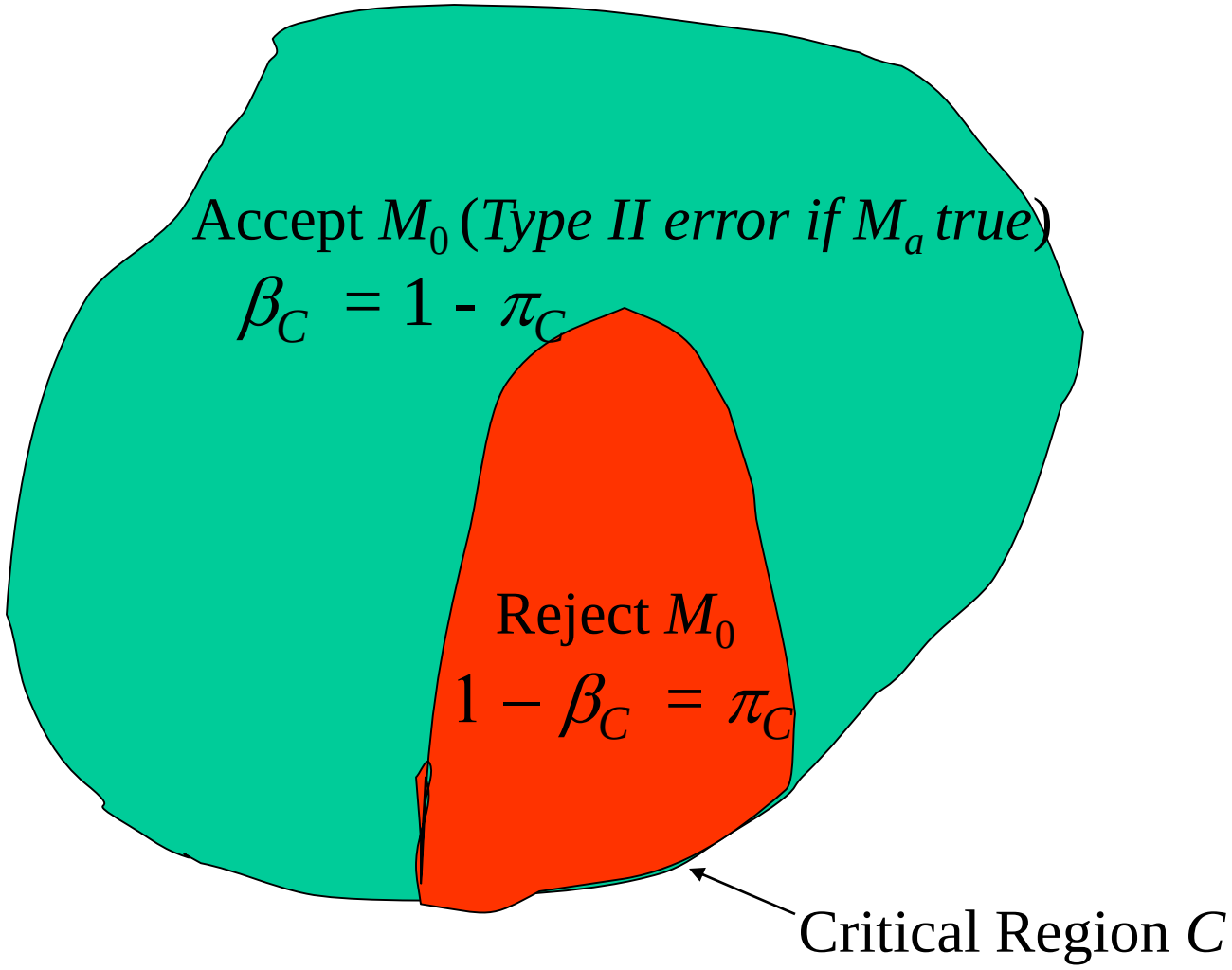


- a *Type II error* occurs if we accept M_0 when it is false.
  - For a given C, prob of committing a Type II error is denoted β_C

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

- $\pi_C = 1 - \beta_C$ 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).