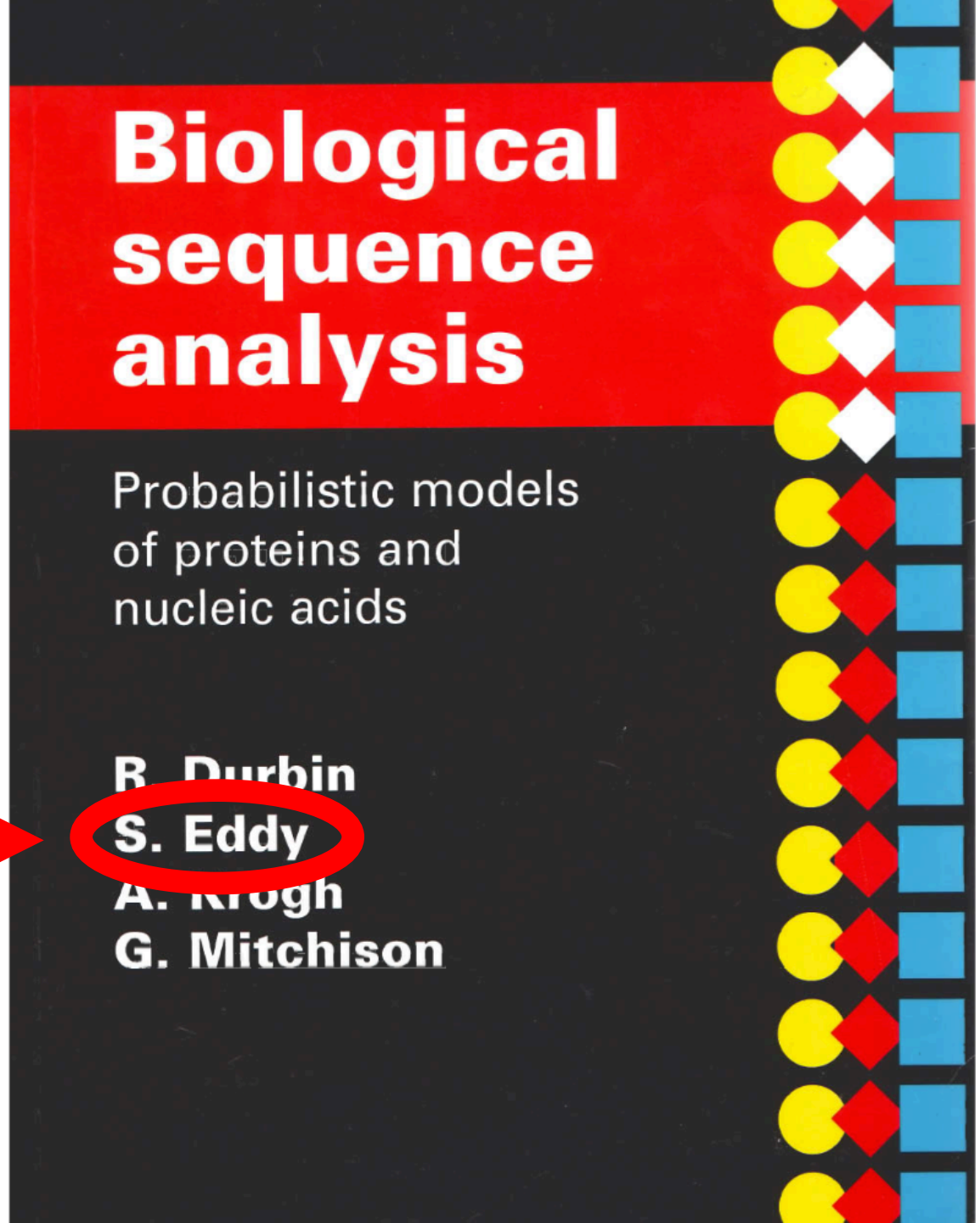
Basic Probability

in the context of molecular sequences

Erin Molloy

February 17, 2017

Example problems are adapted from



Creator of HMMER
http://hmmer.org

What is the probability that we observe the sequence

CGG

Consider a four sided die

with the sides labeled by the alphabet: A, C, G, and T.

Let the probability of rolling an A be denoted P(A) and so on so forth with

$$P(A),\ P(C),\ P(G),\ P(T) >= 0$$

and

$$P(A) + P(C) + P(G) + P(T) = 1$$

What is the probability that we observe the following sequence given a fair die?

CGG

Random sequence model

A fair die will have the following probabilities

$$P(A) = P(C) = P(G) = P(T) = 0.25$$

Each role of the die is treated as an independent event. Then the probability of rolling C **AND** G **AND** G, call P(CGG)

$$P(C)* P(G) * P(G) = 0.25^3 = 0.016$$

Random sequence model

A fair die will have the following probabilities

$$P(A) = P(C) = P(G) = P(T) = 0.25$$

Each role of the die is treated as an independent event. Then the probability of rolling C **AND** G **AND** G, call P(CGG)

$$P(C)*\ P(G)\ *\ P(G) = 0.25\text{^}3 = 0.016$$

Rule: Suppose A and B are independent events.
Then P(A AND B) = P(A) * P(B)

Now consider two different dies

D1 is fair:

$$P(A) = P(C) = P(G) = P(T) = 0.25$$

D2 is biased towards G's and C's:

$$P(C) = P(G) = 0.45$$

$$P(A) = P(T) = 0.05$$

What is the probability that we observe the following sequence given that we pick up D1? What about D2?

CGG

Recall: D1 is fair. D2 is biased towards G's and C's:

$$P(C) = P(G) = 0.45$$

$$P(A) = P(T) = 0.05$$

Conditional probability

The probability or rolling CGG given that we picked up D1, call P(CGG|D1) is

$$P(C) * P(G) * P(G) = 0.25^3 = 0.016$$

The probability of rolling CGG given that we picked up D2, call P(CGG|D2) is

$$P(C) * P(G) * P(G) = 0.45^3 = 0.091$$

What is the probability that we observe the following sequence rolling D1 when there is a 0.6 probability of picking up D1, call P(D1)?

CGG

Recall: D1 is fair. D2 is biased towards G's and C's:

$$P(C) = P(G) = 0.45$$

$$P(A) = P(T) = 0.05$$

Joint probability

The probability or rolling CGG with D1 is

$$P(CGG|D1) * P(D1) = (0.25^3) * 0.6 = 0.009$$

The probability or rolling CGG with D2 is

$$P(CGG|D2) * P(D2) = (0.45^3) * 0.4 = 0.037$$

This is the product rule!

What is the probability that we observe the sequence given our two dies?

CGG

Recall: D1 is fair. D2 is biased towards G's and C's:

$$P(C) = P(G) = 0.45$$

$$P(A) = P(T) = 0.05$$

Marginal probability

The probability of observing CGG, call P(CGG) is

$$P(CGG|D1) * P(D1) + P(CGG|D2) * P(D2)$$

$$0.016*0.5 + 0.091*0.5 = 0.054$$

Suppose we observe the following sequence. We are suspicious that we are using a loaded die. How can we test our hypothesis?

CGG

Recall: D1 is fair. D2 is biased towards G's and C's:

$$P(C) = P(G) = 0.45$$

$$P(A) = P(T) = 0.05$$

Posterior probability

The posterior probability of our hypothesis (i.e., we have are rolling the loaded die or D2) given the observed data (CGG) is written

$$P(D2|CGG)$$

Bayes' Theorem

We can use Bayes Theorem to calculate the posterior probability

$$P(X|Y) = (P(Y|X) * P(X)) / P(Y)$$

where

X is the loaded die or D2

Y is the observed data or CGG

Posterior probability

$$P(D2|CGG) = P(CGG|D2) * P(D2) / P(CGG) = 0.675$$

So yes, it is slightly more likely that we have picked up a loaded die. But not by much. We would want to do more rolls of the die to confirm our hypothesis.

Posterior probability

$$P(D2|CGG) = P(CGG|D2) * P(D2) / P(CGG) = 0.675$$

So yes, it is slightly more likely that we have picked up a loaded die. But not by much. We would want to do more rolls of the die to confirm our hypothesis.

Note: P(CGG|D2) is the likelihood of the hypothesis (picked up D2) given the observed data (CGG).

Some good terms to google

- Probability rules (e.g., AND, OR)
- Conditional probability
- Joint probability
- Marginal probability
- Posterior probability
- Prior probability
- Likelihood
- Bayes' Theorem

What is the likelihood that we observe some sequence data given a model of evolution, i.e., compute P(AGGA|M)?