MATH 448 — Lecture 11
§4.1–4.3: DNA, probability, and conditional probability

Terms

- **Base**: one of the four letters in a DNA sequence: A, G, C, T. A and G are **purines**; C and T are **pyrimidines**.
- **Site**: one position in a sequence. In **aligned** sequences, site i of one sequence corresponds to site i of the other.
- **Base substitution**: one base replaced by another at a site. A **transition** stays within a class (A↔G or C↔T); a **transversion** crosses classes.
- **Outcome** and **event**: an outcome is one possible result of a trial. An event is a set of outcomes, and it occurs if any outcome in it is observed.
- **Mutually exclusive**: events with no outcome in common.

Problem 11.1. An ancestral sequence S_0 gives rise to S_1, which gives rise to S_2:

$$\begin{aligned} S_0 &: \texttt{TACGGATCCA} \\ S_1 &: \texttt{TATGGGTCCA} \\ S_2 &: \texttt{TATGGATCTA} \end{aligned}$$

(a) Compare S_0 and S_2 only. What fraction of sites differ? (b) List every substitution from S_0 to S_1 and from S_1 to S_2, and label each as a transition or transversion. What fraction of sites substituted along the way? (c) Compare (a) and (b).

Solution. (a) Sites 3 and 9 differ, so $2/10 = 0.2$. (b) From S_0 to S_1: site 3, C→T (transition); site 6, A→G (transition). From S_1 to S_2: site 6, G→A (transition); site 9, C→T (transition). That is 4 substitutions in 10 sites, 0.4 per site. (c) Site 6 changed and then changed back, so comparing S_0 with S_2 misses both of its substitutions. The direct comparison gives 0.2 where 0.4 substitutions per site occurred.

Now do Exercise 11.1 from Homework 7.

Probability from data

Treat each repetition as a trial. Then

$$P(\text{outcome}) \approx \frac{\text{number of trials with that outcome}}{\text{number of trials}}, \qquad 0 \le P \le 1.$$

Addition rule. If E and F are mutually exclusive, $P(E \text{ or } F) = P(E) + P(F)$. The probability of an event is the sum of the probabilities of the outcomes in it.
Complement. $P(\text{not } E) = 1 - P(E)$.

Problem 11.2. A 40-base sequence is

```
CACGCCAACC TGTAGATTTA TGAAAGAACC CTTGGTTTTT
```

Treating each site as a trial, estimate $P(\text{A})$, $P(\text{G})$, $P(\text{C})$, $P(\text{T})$ for the base at site 41. Then estimate the probability that it is a purine.

Solution. Counting gives 11 A's, 7 G's, 9 C's, and 13 T's, so

$$P(\text{A}) \approx .275, \quad P(\text{G}) \approx .175, \quad P(\text{C}) \approx .225, \quad P(\text{T}) \approx .325,$$

which add to 1. The purine event is $\{\text{A}, \text{G}\}$, and A and G are mutually exclusive, so $P(\text{purine}) \approx .275 + .175 = .450$. This agrees with 18/40.

Problem 11.3. Use the estimates from Problem 11.2. (a) Are the events "purine" and "pyrimidine" mutually exclusive? (b) Are "pyrimidine" and "not A" mutually exclusive? (c) Find $P(\text{pyrimidine})$ and $P(\text{not A})$.

Solution. (a) Yes: $\{\text{A}, \text{G}\}$ and $\{\text{C}, \text{T}\}$ have no outcome in common. (b) No: $\{\text{C}, \text{T}\}$ and $\{\text{G}, \text{C}, \text{T}\}$ both contain C and T. (c) By the complement rule, $P(\text{pyrimidine}) = 1 - .450 = .550$ and $P(\text{not A}) = 1 - .275 = .725$.

Now do Exercise 11.2 from Homework 7.

Independence and conditional probability

- **Independent events**: knowing that one occurred gives no information about whether the other occurred. If E and F are independent, $P(E \text{ and } F) = P(E)\,P(F)$.
- **Conditional probability** $P(F \mid E)$: the probability of F given that E has occurred. It is computed as $P(F \mid E) = \dfrac{P(F \text{ and } E)}{P(E)}$.
- E and F are independent exactly when $P(F \mid E) = P(F)$.

Problem 11.4. At one site, track only whether the base is a purine or a pyrimidine. In each generation the class changes with probability 0.02, independently of every other generation. (a) Find the probability of each of the four possible sequences of events over two generations (change or no change, in each). (b) Find the probability that, after two generations, the site is in its original class.

Solution. (a) Using the multiplication rule:

$$\begin{array}{lll} \text{no change, no change} & (.98)(.98) & = .9604 \\ \text{no change, change} & (.98)(.02) & = .0196 \\ \text{change, no change} & (.02)(.98) & = .0196 \\ \text{change, change} & (.02)(.02) & = .0004 \end{array}$$

These add to 1. (b) The site is in its original class after "no change, no change" or after "change, change." These are mutually exclusive, so the probability is $.9604 + .0004 = .9608$.

Now do Exercise 11.3 from Homework 7.

Problem 11.5. Two fair dice are rolled in order. Let A be "the sum is 7," B be "the first die shows 4," and C be "the sum is 9." (a) Are A and B independent? (b) Are C and B independent? (c) Find $P(C \mid B)$ and compare it with $P(C)$.

Solution. There are 36 equally likely outcomes. (a) $P(A) = 6/36 = 1/6$ and $P(B) = 6/36 = 1/6$. The event "A and B" is $\{(4,3)\}$, with probability $1/36 = (1/6)(1/6)$. They are independent. (b) $P(C) = 4/36 = 1/9$, and "C and B" is $\{(4,5)\}$, with probability $1/36$. But $P(C)P(B) = (1/9)(1/6) = 1/54 \neq 1/36$, so they are not independent. (c) $P(C \mid B) = \dfrac{1/36}{1/6} = \dfrac{1}{6}$, which is larger than $P(C) = \dfrac{1}{9}$. Knowing the first die shows 4 makes a sum of 9 more likely.

Problem 11.6. Two aligned 40-base sequences, ancestor S_0 and descendant S_1:

S_0 : `GATATGGGCA AGGATGACTA GCGGTGTAGT TCCCGTCCCA`
S_1 : `GACACGGACA GGGATACTTA ACGGTCTAAT GCTCATCCTC`

Each column of the table counts sites with a given base in S_0, and each row gives the base in S_1 at those sites.

$S_1 \backslash S_0$	A	G	C	T
A	6	5	0	0
G	1	7	0	1
C	2	1	6	2
T	0	0	3	6
column sum	9	13	9	9

(a) Estimate $P(S_1 = \text{T} \text{ and } S_0 = \text{C})$ and $P(S_1 = \text{T} \mid S_0 = \text{C})$. Why are they different? (b) Find $P(S_1 = i \mid S_0 = \text{G})$ for each base i, and add them.

Solution. (a) Three of the 40 sites have C in S_0 and T in S_1, so $P(S_1 = \text{T} \text{ and } S_0 = \text{C}) \approx 3/40 = .075$. Nine sites have C in S_0, and 3 of them have T in S_1, so $P(S_1 = \text{T} \mid S_0 = \text{C}) \approx 3/9 \approx .333$. The conditional probability divides by the sites with C in S_0, not by all 40 sites. (b) The G column has sum 13, so

$$P(\text{A} \mid \text{G}) \approx \tfrac{5}{13} = .385, \quad P(\text{G} \mid \text{G}) \approx \tfrac{7}{13} = .538, \quad P(\text{C} \mid \text{G}) \approx \tfrac{1}{13} = .077, \quad P(\text{T} \mid \text{G}) = 0,$$

and these add to 1: given the base in S_0, the base in S_1 is one of the four.

Now do Exercise 11.4 from Homework 7.