COM SCI C121 Week 2

Aidan Jan

April 11, 2024

Conditional Probability

Know these:

- Conditional Probability formulas
- Bayes Theorem

Why does conditional probability matter?

- In the abstract sense, knowing one event tells us something about another event
- Once you get proficient with modeling, it will become "easy" to write down a model
  - I didn't say it woudl be a good model, just *a* model
- Consider: $P(\text{parameters}|\text{data}) = \frac{P(\text{parameters, data})}{P(\text{data})}$
  - Using Bayes Theorem to invert the conditional, this gives $P(\text{data}|\text{parameters})$.
  - This is the condition on the *likelihood* of the data!

Errors in Reads

Short read sequence alignment: the process of finding the putative source of reads. Suppose your genome is the following:

`CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG`

and you get the following read: `CGCGATTGCATTGCG`.

- Is that underlined T (incorrect base pair in the read) genetic variation or an error?

Genetic Variation:

- Humans have genomes 3.2B bases long
- Any two humans are 99.9% similar
- Many types of genetic variation exist:
  - Insertions, deletions, substitutions, inversions, etc.
- We will be focusing on **single nucleotide polymorphisms (SNPs)**, or single base mutations.

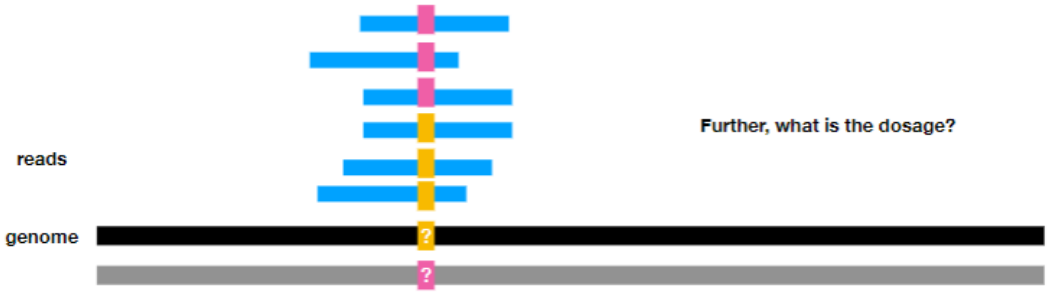
Why is genetic variation important to study?

- Many *traits* have a genetic component
  - traits can be height, skin color, disease, etc.
- Understanding which genetic variation is important enables us to understand the biology and potentially treat diseases
  - see: all of statistical genetics, population genetics, human genetics

High-Level Problem Setup

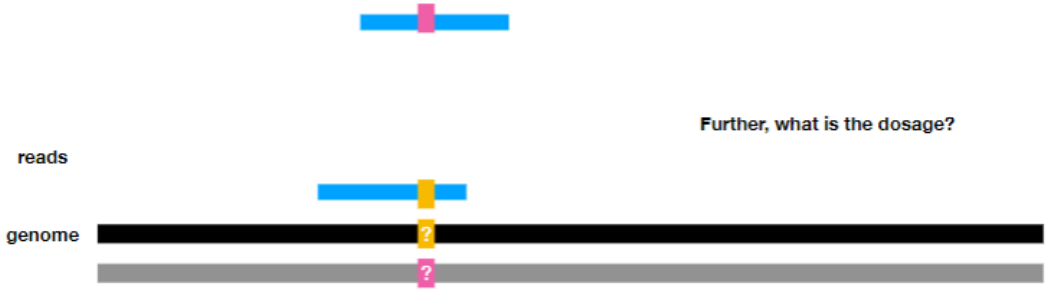
- We have sequencing *reads* that align to a genome with some *mismatches*
- Are tehse mismatches errors from the sequencer, or are they genetic variants?
- Furthermore, there are two chromosomes, so are there one or two copies of the variant?
- Jargon:
  - The "common" variant is often called the **major allele** and the "less common" variant is often called the **minor allele**
  - When we talk about genetic variants, we often talk about the **dosage** of the minor allele (how many times the minor allele occurs)
- E.g., 95% of people have an "A" at site chr1:45280934, and 5% have a "T". If Bob has "A" from his mom and "T" from his dad, his **genotype** is "AT" and his **dosage** is 1.

Specific Problem Setup



In modern sequencing techniques, each nucleotide gets an average of less than 1 read. (We use data from the general population and statistical techniques to guess nucleotides instead.) However, ten to twenty years ago, the state of the art sequencing would read each nucleotide multiple times.

From this data, what is the genotype of the person? AA, AB, or BB? (Remember that both alleles are possible due to errors.) Now, what if we only had the following data:



Now what is the genotype of the person? Assume A is the major allele and B is the minor allele.

- We have reads that overlap a position in the genome
- For simplicity, we are assuming we know this position varies in the human population
- We are restricting ourselves to **biallelic** single nucleotide polymorphisms
  - They can only have two possibilities, genotype A or B
- Further, you can assume you know what the frequency of this variant is in the population
- Finally, remember that the sequencer (e.g., the base caller) tells us what the certainty there is of a base at any given position (the probability of an error at a position)

Rigorous Problem Setup

- There are three possible states: AA, AB, BB
- We need probabilities for each state, given the data
  - i.e., we need $P(AA|\text{data}) = \frac{P(\text{data}|AA)P(AA)}{P((data))}$

The oracle: How is the data generated?

- Let's pretend we know the genotype of the person. Suppose it is AA.
- To generate reads (as the oracle):
  1. Pick a chromosome (and a starting position) and start generating bases.
  2. Is the read an error?
  3. What is the observation?

  Let i be the index for the read, and let c_i be the underlying "true" base.
  Then,

$$P(c_i = A|G = AA) = 1$$
$$P(c_i = B|G = AA) = 0$$

  This is because if the oracle generates AA, then it is impossible for the truth to be a B.
  Additionally, $P(E_i = 1) = \epsilon_i$ and $P(E_i = 0) = 1 - \epsilon_i$.

  - G = what the oracle picked (ground "truth"), which is assumed to be true
  - E_i = random variable that represents error.
  - ϵ_i = some probability
  - $P(E_i)$ essentially represents the chance that an error was made when picking 1 or 0.
- Note that at this point, we have probability of data generated, and probability of a read.
- Now, let O represent the observation of read i.

$$P(O_i = A|E_i, c_i) \rightarrow \begin{cases} P(O_i = A|E_i = 0, c_i = A) \\ P(O_i = A|E_i = 1, c_i = B) \end{cases}$$
$$P(O_i = B|E_i, c_i) \rightarrow \begin{cases} P(O_i = B|E_i = 0, c_i = B) \\ P(O_i = B|E_i = 1, c_i = A) \end{cases}$$

  In this model, if the true genotype is A, it is impossible for a B to be picked, since $c_i = B$ and an observed A has an $E_i = 1$.

Example:

Suppose you got an A from both mom and dad. Then,

$$P(O_i = A|G = AA) = P(O_i = A|E_i = 0, G = AA)P(E_i = 0) + P(O_i = A|E_i = 1, G = AA)P(E_i = 1)$$
$$P(O_i = B|G = AA) = P(O_i = B|E_i = 0, G = AA)P(E_i = 0) + P(O_i = B|E_i = 1, G = AA)P(E_i = 1)$$

Then,

$$P(O_i = A|E_i = 0, G = AA) = P(O_i = A|E_i = 0, c_i = A, G = AA)P(c_i = A) + P(O_i = A|E_i = 0, c_i = B, G = AA)P(c_i = B)$$
$$P(O_i = A|E_i = 1, G = AA) = P(O_i = A|E_i = 1, c_i = A, G = AA)P(c_i = A|G|AA) + P(O_i = A|E_i = 1, c_i = B, G = AA)P(c_i$$

- Notice that $P(O_i = A|E_i = 0, c_i = B, G = AA)P(c_i = B) = 0$ because it is impossible for there to be zero error while reading two A's, if the ground truth is a B.
- Similarly, $P(O_i = A|E_i = 1, c_i = A, G = AA) = 0$ because it is impossible to have a 100% chance of error when reading two A's, observing an A, and having a ground truth of A.

Finally,

$$P(O_i = A|G = AA) = P(O_i = A|E_i = 0, c_i = A, G = AA)P(c_i = A|G = AA)(P(E_i = 0))$$

Example:

Suppose now you read a genotype of AB. We can start with finding the probability the observed allele is A, while the genotype is AB. Then,

$$\begin{aligned} P(O_i = A|G = AB) = {} & P(O_i = A|E_i = 0, c_i = A, G = AB)P(E_i = 0)P(c_i = A|G = AB) && (1)\\ & + P(O_i = A|E_i = 0, c_i = B, G = AB)P(E_i = 0)P(c_i = B|G = AB) && (2)\\ & + P(O_i = A|E_i = 1, c_i = A, G = AB)P(E_i = 1)P(c_i = A|G = AB) && (3)\\ & + P(O_i = A|E_i = 1, c_i = B, G = AB)P(E_i = 1)P(c_i = B|G = AB) && (4) \end{aligned}$$

We know that

$$P(c_i = A|G = AB) = \frac{1}{2}$$
$$P(c_i = B|G = AB) = \frac{1}{2}$$

therefore, (1) would evaluate to $\frac{1}{2} \cdot (1 - \epsilon_i)$.
Evaluating (2) and (3) result in 0, and evaluating (4) results in $\frac{1}{2} \cdot \epsilon_i$.
Therefore,

$$P(O_i = A|G = AB) = \frac{1}{2}(1 - \epsilon_i) + \frac{1}{2}(\epsilon_i) = \frac{1}{2}$$

Similarly to above, $P(O_i = B|G = AB) = \frac{1}{2}$.

We need to find the likelihood the guess on the allele is correct given on our observations, or in mathematical terms, $P(G|O)$.

Example with multiple reads:

Suppose we have $O = \{B, A, B, B\}$. For readability, label the indices, so $O_0 = B$, $O_1 = A$, etc. To decide the most likely genotype, we consider all possiblities:

$$P(G = AA|O)$$
$$P(G = AB|O)$$
$$P(G = BB|O)$$

for each one, we would calculate ϵ. For the case, $P(G = AA|O)$, we would do

$$\begin{aligned} P(O|AA) &= P(O_1 = B|AA) \cdot P(O_2 = A|AA) \cdot P(O_3 = B|AA) \cdot P(O_4 = B|AA) \\ &= \epsilon_0 \cdot (1 - \epsilon_1) \cdot \epsilon_2 \cdot \epsilon_3 \end{aligned}$$

and repeat for each. We would choose the most likely sequence.

For this example,

$$P(O|G = AB) = \prod_i P(O_i|AB) = P(O_i = B|G = AB) \cdot P(O_i = A|G = AB) \cdot \cdots$$

In this case, if the sequencer is completely accurate, then we get a epsilon product of 0 for AA and BB, and a epsilon product of $\left(\frac{1}{2}\right)^4$ for AB. Therefore, we would assume AB to be the correct genotype.
However, if the sequencer is inconsistent, the epsilons would not be as likely to give 'perfect' values leading to 0 and $\frac{1}{2}$, which may lead to inconsistent genotype guesses.