Probability Lab for Biol-360 Bioinformatics

Introduction

Today we're focusing on several core concepts of probability, and applying these concepts to bioinformatics. Most people know informally that the probability of flipping a coin and getting a head is $1/2 = 0.5$. By the same idea, the probability of picking a random nucleotide on a chromosome and hitting a guanine is $1/4 = 0.25$.

First we have to define a few terms. These terms show up throughout probability, so it's good to learn them now.

Many organisms don't have 1:1:1:1 proportions of A:T:G:C in their genomes. For example, the unicellular alga *Chlamydomonas reinhardtii* has an unusually high GC content of $\approx 62\%$. For our simple examples, we're assuming that the four nucleotides have equal proportions in a genome.

Sample Space, Outcomes, and Events

First a quick review from basic math: a **set** is a collection of things. Here's a set of four nucleotides: `{A,C,G,T}`. Notice that we wrote the set with curly braces, which is the standard math notation.

> **Group Question 1:** Write out a set of 3 organelles found in eucaryotic cells. Be sure to use the standard math notation.

Sample space is the set of all possible outcomes of an experiment. For example, if we ran an experiment that picked a single random nucleotide from a genome, we could get any of the 4 nucleotides, so the sample space is the set `{A,C,G,T}`.

> **Group Question 2:** Really, really dumb question: How many entries are in the sample space `{A,C,G,T}`?

Don't overthink this question! It's really dumb, but you'll soon see where this is going.

> **Group Question 3:** What is the sample space of all possible dinucleotides? (Write it out.) In other words, what are all the possible dinucleotides you might find in a big genome?

A **dinucleotide** is a pair of nucleotides, for example, `TC`

> **Group Question 4:** How big is the sample space of dinucleotides?

> **Group Question 5:** Now, Google to find a chart of the standard triplet genetic code, like we discussed in class last week.

What is the relationship between the standard genetic code, and the sample space of trinucleotides?

Group Question 6: How large is the sample space of trinucleotides?

Group Question 7: Using the ideas from above, derive a simple math function that takes a number of nucleotides and gives you the sample space size.

Check your formula on the single, di-, and tri-nucleotide sample spaces in the earlier questions. Now check your formula for a 20-long nucleotide chain. The right answer is a sample space size of 1099511627776.

Biologist call this a **20-mer oligonucleotide**.

You should get these results:

Nucleotide count	Sample space size
1	4
2	16
3	64
⋮	⋮
20	1099511627776

Another important math word: A **subset** is contained within a set. For example, a set of UMB buildings is `{Campus, Clark, Healey, McCormack, Quinn, Science, Wheatley}`. A subset of UMB buildings is `{Wheatley, McCormack}`.

Figure 1 shows a set and a subset.

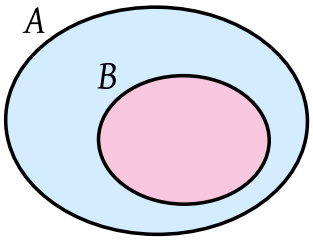


Figure 1: B is a subset of A. Figure derived from `http://upload.wikimedia.org/wikipedia/commons/9/9c/Subset-2.svg`

In probability, an **event** is the result of an experiment, like picking a random nucleotide. Events are subsets of the sample space

For example, if you randomly pick a single nucleotide from a genome, the sample space is `{A,C,G,T}`, and the four possible events are: `{A}`, `{C}`, `{G}`, and `{T}`.

Probability

Definition of Probability

There are several definitions of probability that mean wildly different things.

Believe it or not, until fairly recently there was a very long and very ugly fight between statisticians on the *meaning* of probability.

For us here today, we're focusing on the following definition, which gives the probability of A occurring:

$$P(A) = \frac{\text{Number of ways event A can occur}}{\text{Size of sample space}} \qquad \text{(Equation 1)}$$

For example, we can analyze the probability of a having a baby girl. The sample space is `{girl, boy}`, and the event is the set `{girl}`. That gives:

Assume just one baby. No twins, no triplets, no quadruplets, ...

$$P(girl) = \frac{1}{2}$$

You can write probability in several ways. The following are equivalent and all are fine to use:

$$P(girl) = \frac{1}{2} = 0.5 = 50\%$$

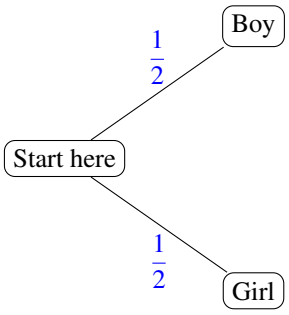
Notice that from (Equation 1) above, any probability must be between 0 and 1. If you calculate a probability and it's outside this range, you know that something went wrong.

Group Question 8: What's an example of a situation with a probability of 0? Show the answer using (Equation 1).

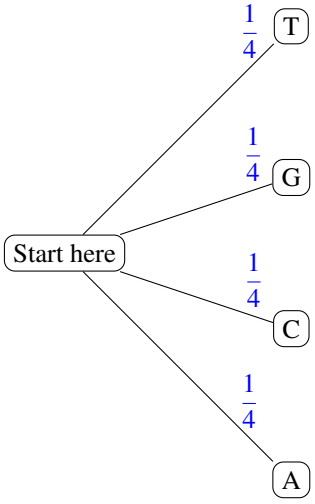
Group Question 9: What's an example of a situation with a probability of 1? Show the answer using (Equation 1).

Probability Trees

We can visualize probability with a tree diagram. Here's the probability tree for having a baby girl or boy. The blue fractions show probabilities:



Here's a probability tree of picking a random nucleotide :



Group Question 10: Draw a probability tree of picking a random UMB building. (See above for list of buildings)

Probabilities of Mutually Exclusive Events

If events are mutually exclusive, you can **add the probabilities** to find the probability of **any of the events occurring**. For example, in the nucleotide probability tree above, you can compute the probability of hitting `{A}` or `{G}` or `{C}` by adding $\left(\frac{1}{4}+\frac{1}{4}+\frac{1}{4}\right)=\frac{3}{4}$.

Mutually exclusive means that only one of the events can occur. For example, if you flip a coin, the events `{head}` and `{tail}` are mutually exclusive.

Group Question 11: What's the probability of randomly picking a UMB building whose name starts with the letter *C*? (See above for the set of buildings. Draw the probability tree to visualize the problem.)

Group Question 12: I'm thinking of a number from 1 to 7. What's the probability that my number is 3, 4, or 5? (Show the probability tree and compute the answer.)

.

Complementary Probability

If the probability event of some event A is $P(A)$, then the probability of A *not occurring* is $1-P(A)$.

For example, if the probability of getting hit by lightning is $\frac{1}{1000}$, then the probability of *not* getting by lightening is $1 - \frac{1}{1000} = \frac{999}{1000}$.

Group Question 13: If you pick a random nucleotide in a genome, what's the probability that it's *not* G? (Show calculation.)

Group Question 14: The subway runs on schedule 92% of the time. You have an important job interview downtown and take the T. What's the probability that you'll have bad luck and be delayed? (Show tree and calculation.)

Conditional Probability

Group Question 15: Think carefully about this… In an abandoned lab in the basement of McCormack, you find an ancient dusty box labeled Amino Acids, containing 20 bottles that are so old you can't read the labels. If you randomly pick one of the 20 bottles, what's the probability that it's glutamate?

Next, to narrow down the possibilities, you do some pH titration experiments, and determine that this is one of the two acidic amino acids. Given the results of the pH titration experiment, what's the probability that this bottle contains glutamate?

Recall there are two acidic amino acids: aspartate and glutamate.

Group Question 16: Here's a very similar question. Again, think carefully about this. Suppose I pick a random nucleotide from a genome and I *also tell you that it's a purine*. Now, using everything you know about the situation, what's the probability that the nucleotide is G?

Recall: the purines are A and G.
A great mnemonic is *Pure **AG**ua.*

The last two questions show the idea of **conditional probability**, which means the probability *given some prior event or information*. In Question 15 above, we're really asking, "What's the probability that this bottle contains glutamate *given* that it's acidic?"

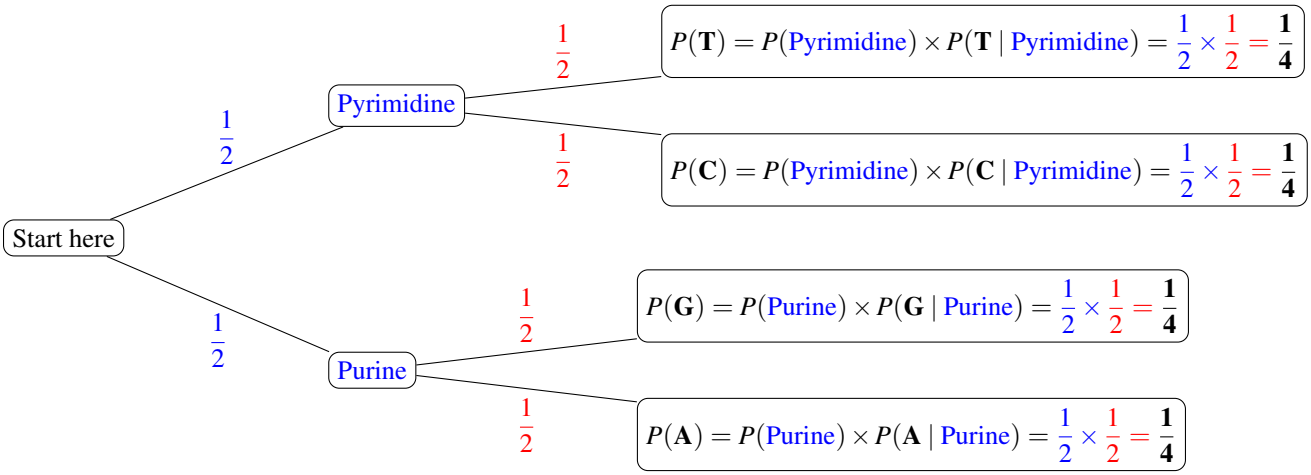
We write conditional probability with a vertical bar $P(A \mid B)$, and say, "the probability of A *given* B. For Question 15 we can write

$$P(\text{glutamate} \mid \text{it's an acid})$$

Group Question 17: Rewrite Question 16 in English using "*given*", and also rewrite it as a formula using the vertical bar notation.

Conditional Probability Trees

It's good to visualize conditional probability as a tree. Here's a tree representing the situation in Question 16. The blue numbers represent the first piece of information; the red numbers show probability *given* the first information; the **bold** numbers show the final probability.



Joint *Probability That Two Events Occur Together*

Often we need to know the **joint** probability of two events A **and** B both happening. Figure 2 shows this idea.

For example, in genetics, suppose there's a autosomal recessive trait **f** for obsessive Facebook usage. Mom's a heterozygote (**Ff**) and dad's also a heterozygote (**Ff**), so neither cares about Facebook. We want to know the probability that little Johnny inherited **f** from mom **and** inherited **f** from dad. In other words, we want to know the probability that Johnny is homozygous (**ff**) for the recessive trait.

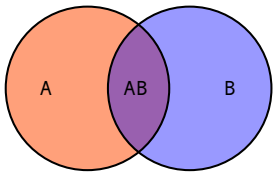


Figure 2: AB is the joint probability of both A and B occurring. Figure derived from `http://upload.wikimedia.org/wikipedia/commons/d/da/Set_intersection.svg`

To represent the joint probability of two events A **and** B both, we write $P(AB)$. To compute the joint probability, use this equation

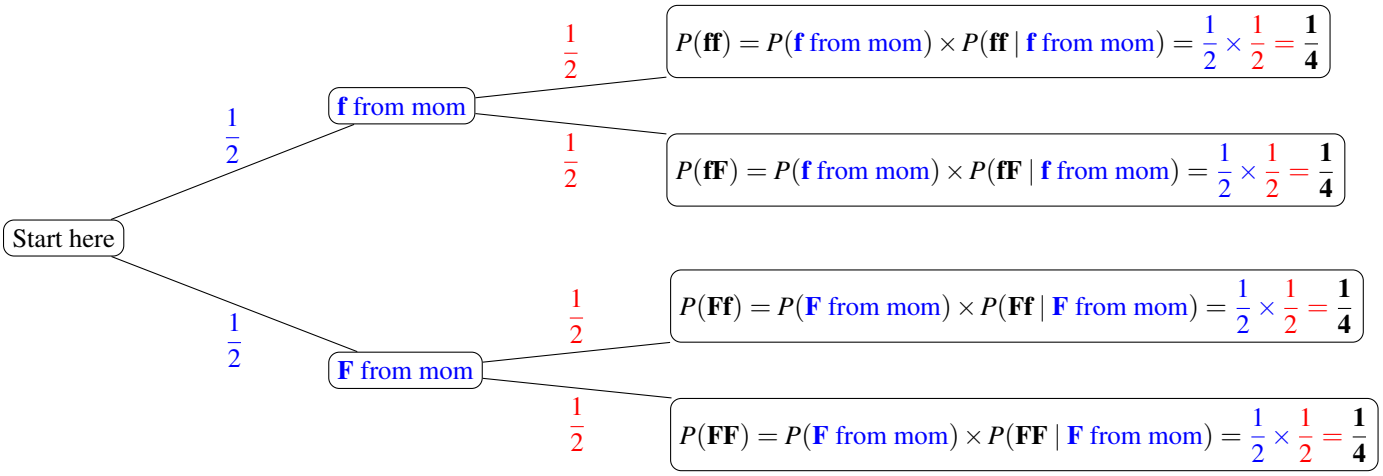
Some books also use set notation $P(A \cap B)$, some books write $P(A\&B)$, and some books write $P(A \text{ and } B)$. They all mean the same thing as $P(AB)$.

$$P(AB) = P(A) \times P(B \mid A) \qquad \text{(Equation 2)}$$

In the Facebook example, we'll compute the joint probability of **ff** given the mother's genotype:

$$P(\mathbf{ff}) = P(\text{got } \mathbf{f} \text{ from mom}) \times P(\text{child is } \mathbf{ff} \mid \text{got } \mathbf{f} \text{ from mom})$$

Here's the probability tree. The red probabilities indicate the probability of inheriting a particular allele from dad.



Notice that the conditional probability tree gave us the classic Mendelian genotypic 1:2:1 Mendelian ratios, and the classic Mendelian phenotypic 1:3 ratios! Is that cool or what?

Group Question 18: Uncle Ted is a severe hypochondriac who believes that he has **both** malaria **and** HIV. You have a brilliant idea to cure Ted's hypochondria: Give him a probability tree that shows the minuscule probability that he has both diseases. Sketch out a probability tree for Ted's two imagined diseases, and make up some tiny probabilities along the edges of the tree. Given those tiny probabilities, compute the final joint probability that Ted has both diseases. What are the numbers you show Ted?

Hypochondriacs believe they have a disease when they're actually healthy.

Bayes Theorem

Suppose you take a blood test and the results come back positive for a serious disease. Don't panic yet. It's possible that the result is a **false positive**. The are lots of examples of false positives:

- The weather forecast predicted rain, but there was no rain.
- The jury convicted an innocent person based on a false fingerprint match.
- You wrote a computer algorithm to predict protein α-helices. You put in a sequence, the program predicted it was an α-helix, but in real biology it's a β-pleated sheet.
- You took a mandatory drug test for employment, and even though you've never used illegal drugs, the test came back positive.

Group Question 19: Find three other examples of false positives?

Obviously, after getting a positive test result for a disease, we really want to know the probability of actually having the disease. We can represent that as a conditional probability:

$$P(\text{actually having the disease} \mid \text{positive on test})$$

We can represent that conditional probability as a probability tree. Figure 3 shows a probability tree of breast cancer false positives. First look at Part A of Figure 3, which shows the results of screening 1000 women for breast cancer.

Group Question 20: In Part A of Figure 3, what percentage of women *actually have* breast cancer? (show your calculation)

Hint: Focus just on the disease status, and ignore the test results

Group Question 21: In Part A of Figure 3, what percentage of women *with* breast cancer test positive? (Show your work.) On the surface, does that seem like a pretty good diagnostic test?

Group Question 22: In Part A of Figure 3, what percentage of women *without* breast cancer test negative? (Show your work.) On the surface, does that seem like a pretty good diagnostic test?

The big surprise comes in Part B of Figure 3. We want to know *Of the women who test positive, what percentage actually have breast cancer?* To find that, we compute

$$P(\text{breast cancer} \mid \text{positive test}) = \frac{\text{number of women with breast cancer } \textbf{and} \text{ test positive}}{\text{total number of women who test positive}}$$

Group Question 23: Compute $P(\text{breast cancer} \mid \text{positive test})$. Now, what do you think about this medical screening test?

Group Question 24: Look at In Part A of Figure 3 again and think carefully about this... We know that most women with breast cancer test positive (left half of Part A), and most healthy women test negative (right half of Part A), so the test seems pretty good. So, why is there such a high false positive rate?

Big hint: in general, $P(A \mid B) \neq P(B \mid A)$. $P(\text{cancer} \mid \text{test} +) \neq P(\text{test} + \mid \text{cancer})$

Closing Thought on Bayes Theorem

If you've taken a standard probability course, you learned Bayes theorem written algebraically as

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}.$$

A 1998 study showed that only 10% of physicians, using this standard algebraic Bayes formula, got the right answers to probability problems like those in Figure 3 [1, 2]! So, if that 1998 study reflects the population of today's physicians, you have a 90% chance of getting wrong information from your doctor!

The problem goes way beyond doctors. Lawyers and judges are also confused by Bayes theorem, check out the *Prosecutor's Fallacy*.

In analyzing Figure 3, we took a visual approach to Bayes theorem. Later this semester, in a regular lecture, we'll see another simple, non-algebraic approach to computing Bayes theorem using a table instead of a tree.

References

[1] G. Gigerenzer and U. Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological review*, 102(4):684, 1995.

[2] U. Hoffrage and G. Gigerenzer. Using natural frequencies to improve diagnostic inferences. *Academic medicine : journal of the Association of American Medical Colleges*, 73(5):538–540, May 1998.

[3] David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, Sep 2011.

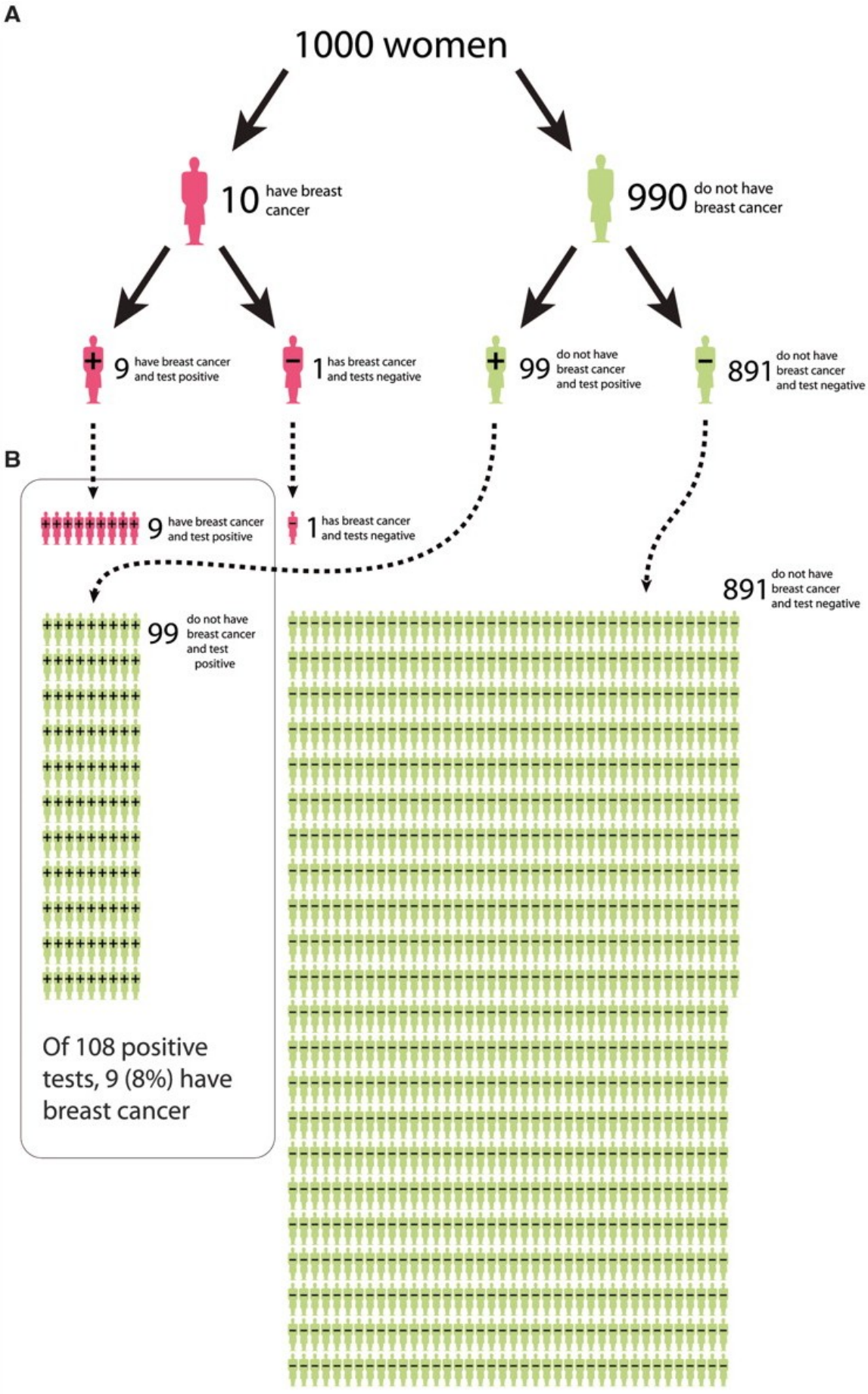


Figure 3: Image source: [3]