What is information?: We all have an intuitive idea about what information is. If we ask a kid what their favorite food is and they say 'chocolate' or 'ice cream', it's not very informative. We expected an answer like that. If they say 'broccoli and cheese' we would probably remember that weird kid. Information is a degree of surprise. The more surprising the **message**, the informative the message. The mathematical definition of information assumes there is an information **source** that emits messages. The information of any particular message is given in equation 1. Information is almost always calculated in base 2 and therefore given the unit **bits** (binary digits). Sometimes you may see **nats**, which corresponds to the log base e. For our purposes, we will always use base 2 for log, and bits will always be the units. As an example, let's say that the answer to an average kid's favorite food is 'chocolate' 50% of the time. The information of 'chocolate' is simply the -log(0.5) or 1.0. Similarly, if the probability of 'ice cream' is 0.25, this is 2.0 bits. Any easy way to think about this is as powers of two: 0.5 is 2^{-1} and 0.25 is 2^{-2}. A message occurring ~1/1000 times has 10 bits of information (2^{10} = 1024).

Equation 1

$$I(m) = -\log_2 P(m)$$

I: information
m: message
P(m): probability of message

Most programming languages have a log function that returns values in log base e. To convert to base 2, simply divide by log(2).

What does this have to do with genomics?: In order to determine if two sequences are related to each other, we need (a) some way to align sequences (b) some way to determine if the alignments are significant. It is common to use information theory to determine the significance.

Information content: Generally, we are more interested in the information content of a source rather than the information of any particular message. A rich source of information provides you with a lot of surprise *on average*. Information content is simply the average information per message (equation 2). Information content is also called **entropy** or Shannon's Entropy because it was invented by Claude Shannon. An information source can be thought of as a frequency distribution (histogram). Some histograms are more *predictable* than others. For example, consider two coins, one is fair, the other a trick coin. The fair coin comes up heads 50% of the time. The trick coin is heads 90% of the time. Which one is more predictable? As an information source, which one has higher information content? What about a fair vs loaded die? What about DNA? What about DNA with highly biased composition? Try working some examples.

Equation 2

$$H = -\sum P_i \log_2 P_i$$

H: information content
Pi: probability of message i

Relative entropy: Let's say you have nucleotide frequencies from several different genomes and you want to know which ones are the most similar to each other. How might you compare them? If you consider the sequences to be information sources and nucleotides to be messages, then you can use relative entropy to measure the similarity. This is also called **Kullback-Leibler distance** (equation 3). Strictly speaking, D(P||Q) is not always D(Q||P) but it's usually close. Not that if P or Q contains any zero probability values, you may get numerical errors (divide by zero or log zero).

Equation 3

$$D(P \,||\, Q) = \sum P_i \log_2 \left(\frac{P_i}{Q_i} \right)$$

D: distance
P: some frequency distribution
Q: other frequency distribution
Pi: probability of message i in P
Qi: probability of message i in Q

Codon bias: In the genetic code, some triplets code for the same amino acid, but not all codons are used with the same frequency. In different genomes the biased codons may not be the same. Codon bias probably exists because of translational efficiency. Not all tRNAs are expressed at the same level. As a result, highly expressed genes are optimized to use abundant tRNAs so they don't have to "wait" for rare tRNAs. Given that the codon bias of an organism is a kind of signature, if you found a gene with very different codon usage, you might expect horizontal gene transfer. You could use K-L distance to find outliers in a genome. In such an experiment, the source is the codon usage, and each message is a triplet.