

A PRIMER ON INFORMATION THEORY*

Definition: information = reduction in ambiguity

In loose terms, the information content (I) of a message sent down a channel from a sender to a receiver can be quantified as follow:

$$I \approx \frac{N_{\text{before}}}{N_{\text{after}}} \quad (1)$$

where N_{before} refers to the number of macroscopic states available to a system before the receiver receives the information, and N_{after} refers to the number after the information is received. More specifically, we need to deal with probabilities, not absolute numbers. So let P_{before} = the probability of an event occurring before the message is received, and P_{after} = the probability after the message has been received. (Recall that the probability of an event is always calculated as the number of ways the event could happen divided by the total number of ways anything could happen.) Also, we want to scale information to a log scale. Then, the information content can be defined more sharply:

$$I = \ln \frac{P_{\text{after}}}{P_{\text{before}}} \quad (2)$$

Note that the “before” and “after” terms need to be inverted when we switch to probabilities. When log base 2 (\ln) is used, the units of information are what you are familiar with from computers: bits!

This definition of information bears a strong similarity to the concept of entropy as defined in statistical mechanics: $S = k \ln W$, where S is the entropy of the system, k is the Boltzmann constant (1.38×10^{-23} J/K), and W is the number of possible microstates that the system could be in to comprise its particular macrostate. Consider here “Maxwell’s Demon”, a hypothetical beast that sits atop a gas-filled box with a partition between two sides. The Demon would let only fast molecules through to the other side, thus raising the temperature in one place and lowering it elsewhere without doing any work ... and hence violating the Second Law of thermodynamics. The way out of the dilemma, as seen by Szilard in 1929 and formalized by Brillouin in 1951, is to realize that information must be used by the Demon, and this information is numerically related to the entropy change: the amount of entropy lost (or negative entropy = “negentropy” gained) is directly proportional to the amount of information used:

$$I_{\text{needed}} \approx S_{\text{lost}} \quad (3)$$

Returning to $S = k \ln W$, we can write this as $S = -k \ln P$, where P now is the probability that the macrostate is in the particular configuration that it is in, provided that all configurations are equally probable (i.e., $P = 1/W$). Now we are in a position to make a formal measurement of information:

$$I = -k \ln P \quad (4)$$

Where I is the information stored in a system, P is the probability that the system is in its particular configuration, and k can either be the Boltzmann constant or any other constant we choose to get information to scale in the way we want. For the transmission of a message can also write this as:

$$I = k \ln(n_{\text{before}} / n_{\text{after}}) = -k \ln(n_{\text{after}} / n_{\text{before}}) \quad (5)$$

where n is the number of possible configurations that a system can adopt. If however, not all configurations of a system are equally probable, then we need to account for this by the Gibbs expression:

$$I = S = -k \sum p_i \ln p_i \quad (6)$$

where p_i is the probability of the i th state. With the same relationship between entropy and information as given above, this leads us to the Shannon equation:

$$H = -K \sum p_i \ln p_i \quad (7)$$

where I has been replaced by H , which is called the complexity of the system, and k has been replaced by K , which is generally taken to be 1. Shannon was one of the founders of the field of information theory in the late 1940's and the Shannon complexity is a common way to mathematically describe the complexity of a DNA sequence, a population of DNA sequences, and even the diversity of species in a biological ecosystem.

Consider the information contained in nucleic acids. Now there is some debate over whether "complexity" and "information" mean the same thing, especially when it comes to DNA. You can think about the complexity of DNA very simply as follows. Each nucleotide can take on any one of four possible configurations (G, A, T, or C) and thus if you know the identity of a particular nucleotide at a particular spot in the genome, then you have reduced the ambiguity four-fold, and using equation (5) above for information, we get:

$$I = k \ln(4/1) = k \ln 4 = 1.39k$$

if we choose k to be $1/\ln 2$ ($= 1.44$), then we get $I = 2$ bits, like a computer

each nucleotide then, can contain 2 bits of information, so 3 nucleotides contain 6 bits of info (1 byte = 8 bits) if they are completely free to vary. By this method, a small gene contains about 300 nt, or 600 bits of information... human genome contains 3.3×10^9 nt = 6.6×10^9 bits or about 7 GB of information (1 GB = 10^9 bytes). However, many have argued that the complexity of DNA is not the same as the information content stored in (and thus available from) a DNA sequence. You can see this easily when you consider that by the complexity definition, there is as much information in a 100 b.p. stretch of "junk" (random) DNA as there is in a 100 b.p. gene critical for cell function. A better measure of the information content in a gene has to take into context the "language" of gene expression. There is something called the "grammar complexity" of DNA, but we will not consider how to measure this parameter in Chem 492.

*Sources:

1. Jacobson H (1955). Information, reproduction, and the origin of life. *Amer. Sci.* **43**: 119-127.
2. Setlow RB and Pollard EC (1962). *Molecular Biophysics*. Addison-Wesley.
3. Gatlin LL (1972). *Information Theory and the Living System*. Columbia.
4. Wicken JS (1987). *Evolution, Thermodynamics, and Information*. Oxford.