Lecture No. 6
Information Theory

January 26, 2003

- Information: Knowledge or intelligence communicated, received or gained.

- Information Theory: Indication of number of possible choice.

- Information theory deals with transmission of messages.

- Sequence data can be thought of as messages.

- Bits units: From information theory, a bit denotes the amount of information required to distinguish between two likely possibilities.

- Bit, the minimal amount of structural complexity needed to encode a given piece of information.

- The number of bits of information, N required to convey a message that has M possibilities is:

$$\log_2 M = N\, bits$$

- Consider a number 1224 an equivalent 12 digit binary number made of $\{0,1\}$ is 010011001000
  $0 + 1024 + 0 + 0 + 128 + 64 + 0 + 0 + 8 + 0 + 0 + 0$
  Here each residue is one bit of information.

- A sequence of 0's and 1's or any sequence of any alphabet or symbols need not represent a number. It may just be a code that needs to be translated into some action.

- How many bits of information there are in a sequence of given length composed of symbols drawn from a chosen alphabet. E.g., {A,G,C,T}.

- A number of bits of information in a sequence of length 12 with symbol {0,1} is exactly 12.

- The RNA alphabet is {A,G,C,U}.
  There are 4^{12} words of length 12 that can be formed from these 4 letter alphabet.

- On the assumption each letter occurs in nature with equal probability ($p = 0.25$), the information content associated with such a word
  $\log_2(4^{12}) = \log_2(2^{24}) = 24$ bits

- Actual genetic instructions arise from the 20-letter amino acid code, each letter of which is associated with an ordered triple from the RNA alphabet.

- Since there are $4^3 = 64$ residue triples from the alphabet {A,G,C,U}.Information in an amino acid gene word of length 4 is

$$\log_2 (20^4) = 17.3$$

This is only true if each of the letter occurs with equal probability ($p = 1/20$), which they do not. There are reasons for this:

1). Some amino acids are coded for by more triples than others.

2). Some amino acids are biologically more important than others.

- In a perfect world every symbol in a alphabet occurs with equal probability. In this way the information content is maximized.

1). Example Suppose two sequence symbols $\{0,1\}$ with equal probability for each symbol to occur, i.e., $P_0 = P_1 = .5$. In this case each residue is one bit.
Thus equal probability yield one bit/residue-maximum.

2). Example Suppose $P_0 = 1$ and $P_1 = 0$
The information for each residue in this case is zero bit.
Probabilities of 0's and 1 yield zero bits/residue–minimum.

5). What if $p_0 = 0.8$ and $p_1 = 0.2$. In this case number of bits/residue is between one and zero. In particular it is:

$$-[0.8 \log_2 (0.8) + 0.2 \log_2 (0.2)] = 0.722$$

- The information/residue of n letters alphabet, such that the probability that the ith letter will appear in any given residue is p_i is:

$$H = -\sum_{i=1}^{n} P_i \log_2 p_i$$

Where $H =$ Entropy
It is a measure of disorder.

- Example: If all the p_i's are the same, i.e., $P_i = 1/n$ for $i = 1, 2, \ldots, n$, then

$$\begin{aligned} H &= -\sum_{i=1}^{n} 1/n \log_2(1/n) \\ &= -\log_2 (1/n) \\ &= \log_2 n \end{aligned}$$

- Example: The nucleotide alphabet contains 4 symbols with probability 1/4, has Entropy

$$H = \log_2 4 = 2$$

i.e., average information content = 2 for random DNA.
i.e., the bits/residue is 2. We need to ask two (yes/no) questions to determine a base A or G to match the column to a position in a test sequence.
A nucleotide string of length 900 is equivalent to 1800 bits which can code for $4^{900} = 2^{1800}$.

- Calculating log base 2

$$\log_2 a = \frac{\log_z a}{\log_z 2}$$

where $z =$ any base

$$\begin{aligned} \log_2 a &= \frac{\log_1 0a}{\log_{10} a} \\ \log_2 a &= 3.32 \log_{10} a \end{aligned}$$

- How does the amount of information change after we have more information? This is considered a decrease in uncertainty.

$$\text{Information} = H_{\text{before}} - H_{\text{after}}$$

$$\begin{aligned} \text{Random DNA} &= H_{\text{before}} \\ &= 2 \end{aligned}$$

- Example: If we note that one region has $P_A = 0.7$, $p_G = 0.4$ then

$$\begin{aligned} H_{\text{after}} &= -.7 \log_2 .7 - .3 \log_2 .3 \\ &= 0.88 \text{ bits} \\ \text{Information} &= 2 - 0.88 \\ &= 1.12 \text{ bits} \end{aligned}$$

  This is an increase in information content.