

## **Appendix 2: Relative Entropy and Information**

---

This section is best read immediately after chapter 4. The techniques used in chapter 4 to calculate information are different from those used by most authors. Most authors use relative entropy.

Suppose that a sequence alignment for a protein gives the following results (only the first two amino acids in the sequence are shown):

Chicken	AlaVal.....
Man	AlaVal.....
Dog	AlaVal.....
Lizzard	AlaAla.....
Fish	AlaAla.....
JellyFish	AlaAla.....

The information in the first position is easy to calculate. Four out of 64 possible codons specify alanine. So the information is by equation 2 in chapter 1 as follows:

$$\text{Information} = 3.32 \times \log ( 64/4) = 4 \text{ bits.}$$

The information of the second position is also easy to calculate. Four codons specify alanine and 4 specify valine. The information is as follows:

$$\text{Information} = 3.32 \times \log ( 64/ 8) = 3 \text{ bits.}$$

These calculations are in full agreement with the techniques described in chapter 4.

Now consider a different alignment.

Chicken	AlaVal.....
Man	AlaVal.....
Dog	AlaVal.....
Lizzard	AlaVal.....
Fish	AlaVal.....
JellyFish	AlaVal.....
Oak Tree	AlaVal.....
E coli	AlaAla.....

The information content of the first position has not changed. It is still 4 bits, but what about the second position? The same two amino acids are present, but valine is found in 7 of the 8 sequences. Intuitively, it would seem that the second position in this alignment contains more information, and it does. To calculate the information for the second position, the formula for relative entropy must be used. The formula for relative entropy when two amino acids appear in the same alignment column is as follows:

$$\begin{aligned}
 \text{Relative Entropy} &= \text{Frequency amino acid 1} \times 3.32 \times \log \left[ \frac{\text{Frequency amino acid 1}}{\text{Expected Frequency of amino acid 1}} \right] \\
 &+ \text{Frequency amino acid 2} \times 3.32 \times \log \left[ \frac{\text{Frequency amino acid 2}}{\text{Expected Frequency of amino acid 2}} \right]
 \end{aligned}$$

In this case, the relative entropy is as follows:

$$\text{Relative Entropy} = 1/8 \times 3.32 \times \log [(1/8)/(4/64)] + 7/8 \times 3.32 \times \log [(7/8)/(4/64)]$$

$$\text{Relative Entropy} = .125 + 3.33 = 3.46 \text{ bits.}$$

Relative entropy is a measure of information. In fact, the actual information at position 2 is the relative entropy.

$$\text{Relative Entropy} = \text{Actual Information} = 3.46 \text{ bits.}$$

Now consider what happens when the equation for relative entropy is applied to the second position in the first set of sequences. In this set of sequences, valine occurs  $1/2$  of the time, and alanine occurs  $1/2$  of the time.

$$\text{Relative entropy} = 1/2 \times 3.32 \times \log [(1/2)/(4/64)] + 1/2 \times 3.32 \times \log [(1/2)/(4/64)]$$

$$\text{Relative entropy} = 1.5 + 1.5 = 3 \text{ bits.}$$

So relative entropy gives the same results as the technique used in chapter 4 when the amino acids in a column all occur at the frequency that would be expected by the underlying probabilities where the underlying probability is given by the number of codons that specify each amino acid.

Why not use relative entropy? Relative entropy is always greater than or equal to the information calculated with the techniques of chapter 4. Thus, the techniques used in chapter 4 always calculate the minimum possible information in the gene or protein today. The true information which can only be found by using relative entropy will always be higher.

The figure below depicts these concepts. The actual information in a gene today must be calculated using the formula for relative entropy. This formula takes the frequency of the various amino acids in an alignment column into consideration. This value can never be related to a probability. The technique used in chapters 4 and 5 calculates the minimum possible information by assuming that all allowed amino acids in any given column are found at the expected frequency. Under some circumstances, this value can be related to a probability for evolution (chapter 5). But most of the time, this should never be done. Molecular knowledge defines the minimum information required for a selective advantage to be realized. It can almost always be related to a probability for evolution. The size of this first vertical step determines whether or not chance will create the required knowledge.

