

Appendix 4: A Review of Yockey's Approach

Today, many scientists apply information theory to molecular biology, but only a few have tried to use information theory to answer the most important question. Is evolution possible? Yockey was probably the first. Because so few scientists are trying to answer this question, a consensus as to how to assign information has not been reached.

Yockey assigns information to proteins by a technique that is very different from the one developed in chapter 4. His technique relies on both Shannon entropy and conditional entropy. He models the transfer of information from DNA to proteins and assigns information to this transfer by treating mutations as a source of noise. Unfortunately, his analysis fails to consider natural selection, and so the information that he calculates is incorrect. Natural selection weeds out harmful mutations. In a communication system, this is analogous to a person receiving an unintelligible message and then asking for the message to be re-sent. The equations that Shannon developed to model noise in communication systems do not apply if a person on the receiving end inspects the incoming messages for errors and discards any messages that contain errors.

This book did not use Yockey's technique because the information assigned by his technique cannot be related to a probability for protein evolution. For a full development of his technique, refer to the two references at the end of this appendix.

Yockey's technique divides every sites in a protein into two categories, absolutely conserved and not absolutely conserved. An absolutely conserved amino acid is one that never changes. For example, if column 20 in a multiple sequence alignment is always a glycine, then position 20 in the alignment is absolutely conserved. If more than one amino acid is found in column 20, then it is not absolutely conserved. Yockey's technique sets the information content of any absolutely conserved amino acid equal to the Shannon entropy, 4.14 bits.

So using this technique, a peptide composed of 10 methionines has the same chance of evolving as one composed of 10 serines. This is the drawback of using Shannon entropy. Methionine is only specified by 1 codon, and serine is specified by six. Assuming random mutations, serine should arise by chance six times as often as methionine.

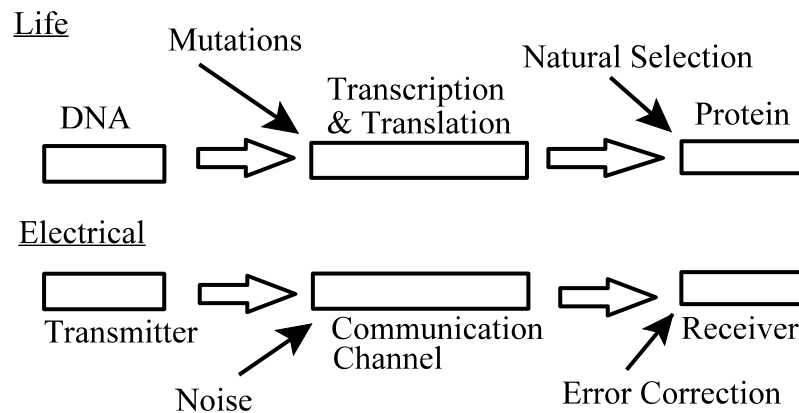
For reference, the Shannon entropy, assuming the genetic code, is calculated below. To follow Yockey's method exactly only the 61 codons that do not terminate the peptide chain are allowed.

amino acid group	expected frequency	information	total (bits)
6 codons (3 amino acids)	29.5%	3.34	0.99
4 codons (5 amino acids)	32.8%	3.93	1.29
3 codons (1 amino acid)	4.9%	4.34	0.213
2 codons (9 amino acids)	29.5%	4.93	1.45
1 codon (2 amino acids)	3.3%	5.93	0.195
			4.14

Example calculation: Three amino acids are specified by 6 codons. The information acquired when one of these is observed is as follows: $\text{information} = 3.32 \times \log(61/6) = 3.34$ bits. Because these amino acids should arise by chance 6 times in every 61 tries and because there are 3 of these amino acids, the expected frequency is $(3 \times 6)/61 = .295$ or 29.5%. The last column is the product of the expected frequency and the information for each amino acid group (each row). The sum of all entries in the last column is the Shannon entropy.

Because proteins are short messages, Shannon entropy is almost never a true measure of a specific protein's information. A typical protein may only have 30 absolutely conserved amino acids, and the probability for these amino acids arising by chance might be very different from the number calculated using Shannon entropy.

Yockey's calculations for amino acids that are not absolutely conserved run into another issue. He models mutations as noise, and then applies the techniques developed by Shannon to calculate information transfer through a noisy communication channel. The figure below illustrates the similarity between information transfer in life and in electrical communication systems.



In electrical systems, conditional entropy models the amount of information lost due to noise. For example, suppose the transmitter transmits the results of a trapped scientist experiment. The two results are heads or tails. If the channel is noisy, when heads is transmitted tails might be received. This error reduces the rate of information transfer. Conditional entropy models how much information is lost. Mutations in life have the same effect as noise in communication systems, so mutations can be modeled as a noise source.

The problem with this approach is that natural selection does not allow harmful mutations to survive. So if a mutation creates a non-functional protein, the mutation will be removed from the population by natural selection. In communication systems, error correction is responsible for this function. With error correction, conditional entropy can no longer be used to accurately model the information lost due to noise. Likewise, because of natural selection, conditional entropy does not model the effect of mutations on information transfer.

Natural selection cannot ensure that only allowed messages are transmitted, but it does ensure that only allowed messages survive. And the neutral theory of evolution (see chapter 4) predicts that many if not all of the allowed amino acid substitutions in the final protein will be observed if the same proteins in many diverse species are analyzed. Thus, the information content of the final protein does not depend on the information transferred from DNA, and conditional entropy is not needed for this analysis. The equations to calculate information in chapter 1 are applicable, and they may be applied using the techniques introduced in chapters 4 and 5.

Furthermore, the techniques used in this book always maintain a strict one to one relationship between probability space and information space. In other words, probability theory and information theory both yield the same results. Yockey's approach does not preserve this one to one mapping. So the odds that he calculates for protein evolution are different than the odds that one would calculate using probability theory.

References:

- 1) Yockey, Information Theory, Evolution and the Origin of Life, 2005.
- 2) Yockey, Information Theory and Molecular Biology, 1992.
- 3) Shannon, A Mathematical Theory of Communications, 1948.