

Appendix 5: A Review of Schneider's Approach

Tom Schneider has a very good web site of information theory. He has proposed a different way to assign information to a protein. His suggestion is to use relative entropy with the assumption that all amino acids are equally probably in the final protein. The resulting equation is shown below.

Let F_1 = frequency of amino acid 1, F_2 = frequency of amino acid 2 and so on. The last amino acid found at a given position is thus F_n .

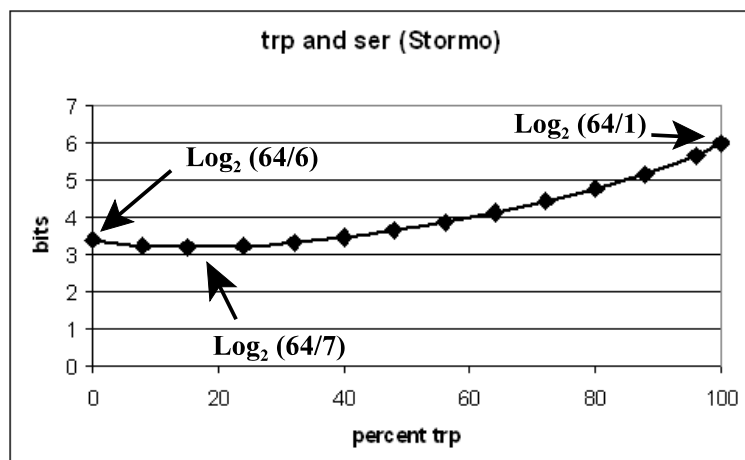
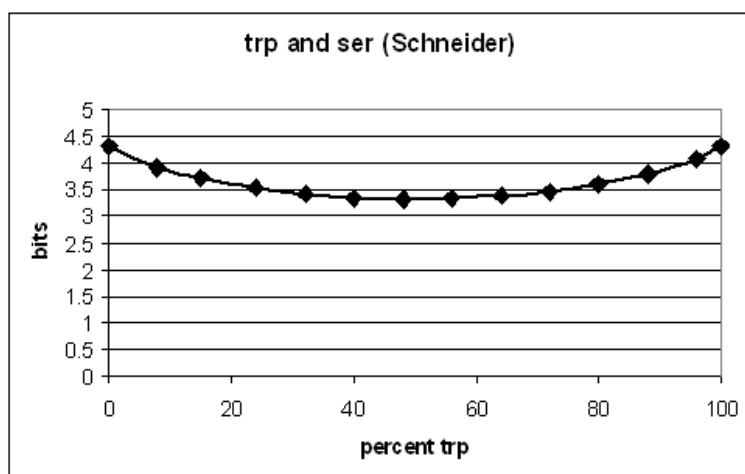
$$\text{Information} = \log_2 (20) + F_1 \times \log_2 (F_1) + F_2 \times \log_2 (F_2) \\ \dots + F_n \times \log_2 (F_n)$$

The first term is the maximum possible average information for each site within the protein. The terms that follow all are negative, and so these reduce the amount of information at each site (except when they are zero).

Using this equation, any column in a multiple sequence alignment that contains 1 and only one amino acid is assigned 4.32 bits. This assignment does not depend on the amino acid. So a column with only serine (6 codons) is assigned 4.32 bits, as is a position with only methionine (1 codon). Clearly, there is a problem with this approach. Serine is 6 times as likely to arise by chance as methionine. Assigning both of these columns the same information destroys any hope of relating the information in a protein to a probability of evolution. This problem is very similar to what was observed in appendix 4 with Yockey's technique.

Information theory works best when the messages are very long, and proteins are not long messages. Therefore, the average information per column (in the multiple sequence alignment) for all proteins in life is not necessarily equal to the information found in the average for a single protein. The equation above predicts that a peptide composed entirely of methionine has the same chance of evolving as one composed entirely of serine. This is not correct.

The following graphs consider two amino acids as the frequency of each changes. The first graph uses relative entropy. Because Stormo has suggested that this is the correct measure of information, his name appears in the title. The second graph uses Schneider's equation. Notice that the Schneider's equation never calculates the information in such a way that it can be related to a probability of evolution. In contrast, relative entropy gives exactly the correct solution at the three distinct points shown in the second graph.



The minimum observed in the second graph is precisely the information predicted by using equations 1 and 2 introduced in chapter one. That is if trp and ser are found in a specific column in the multiple sequence alignment, then the column contains $3.32 \times \log_2(64/7)$ bits of information which of course is equal to $\log_2(64/7)$. It is a trivial exercise to show that the two endpoints in the second graph are also correct. Thus, relative entropy is the best measure of protein information. It is the only available approach that preserves the relationship between information and probability. It is preferable to both Schneider's and Yockey's techniques as these other two techniques can assign more information to a protein that is more likely to evolve. This is clearly contrary to the definition of information as originally proposed by Shannon.

References:

- 1) Hertz, Stormo, Identifying DNA and protein Patterns with Statistically Significant Alignments of a Multiple Sequences, Bioinformatics, 1999.
- 2) <http://www-lmmb.ncifcrf.gov/~toms/>
- 3) Schneider, Stormo, Gold, Ehrenfeucht, "Information Content of binding Sites on Nucleotides Sequences," J. Mole Biology, 1986.
- 4) Schneider, Stormo, "Excess Information at Bacteriophage T7 Genomic Promoters Detected by Random Cloning Techniques," Nucl. Acids Res, 1989.
- 5) Schneider, "Measuring Molecular Information," Journal of Theoretical Biology, 1999.
- 6) Shannon, Weaver, The Mathematical Theory of Communication, 1964.

<http://www.theory-of-evolution.net>