ELSEVIER

Information Sciences 141 (2002) 219–225

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

Information theory, evolution and the origin of life

Hubert P. Yockey [*]

1507 Balmoral Drive, Bel Air, MD 21014-5638, USA

Received 8 October 2000; received in revised form 15 March 2001; accepted 25 April 2001

1. Introduction

It is almost universally believed that the number of possible sequences in polypeptide chains of length N can be calculated by the following expression:

$$(20)^N. \tag{1}$$

Expression (1) gives the total number of sequences we must be concerned with, if and only if, all events are equally probable. However, many events in general and amino acids in particular do not have the same probability. Let us consider a long sequence of N symbols selected from an alphabet of A letters or events [10]. In the present case the letters or events will be the alphabet of either codons or amino acids. The sequence will contain $Np(i)$ of the ith symbol. Let P be the probability of the sequence. Then by elementary probability theory

$$P = \prod_i p(i)^{p(i)N}. \tag{2}$$

Taking the logarithm of both sides:

$$\log_2 P = N \sum_i p(i) \log_2 p(i), \tag{3}$$

$$\log_2 P = N \sum_i p(i) \log_2 p(i) = -NH, \tag{4}$$

[*] Tel.: +1-401-879-1805.
E-mail address: hpyockey@aol.com (H.P. Yockey).

PII: S0020-0255(02)00173-1

where

$$H = -\sum_i p(i)\log_2 p(i), \tag{5}$$

where H is the information or Shannon entropy of the probability space containing the events i.

Accordingly, the probability of a sequence of N symbols or events is very nearly

$$P = 2^{-NH}. \tag{6}$$

The number of sequences of length N is very nearly

$$2^{NH}. \tag{7}$$

We have approached the calculation of the number of sequences of length N in two apparently correct ways. What happened to the sequences left out of the total possible by expression (7)? This is explained by the Shannon–McMillan–Breiman theorem [10]:

> For sequences of length N being sufficiently long, all sequences being chosen from an alphabet of A symbols, the ensemble of sequences can be divided into two groups such that:
> 1. The probability P of any sequence in the first group is equal to 2^{-NH}.
> 2. The sum of the probabilities of all sequences in the second group is less than ε, a very small number.

The Shannon–McMillan–Breiman theorem tells us that the number of sequences in the first or high probability group is 2^{NH} and they are all nearly equally probable. We can ignore all those in the second or low probability group because, if N is large, their *total probability* is very small. Thus the Shannon entropy, H, controls the number of sequences we need to be concerned with. The number of sequences in the high probability group is almost always many orders of magnitude smaller than that given by expression (1) which contains an enormous number of "junk" sequences.

2. Shannon's Channel Capacity theorem and the role of error in protein function and specificity

Shannon's Channel Capacity theorem proved that codes exist such that sufficient redundance can be introduced so that a message can be sent from source to receiver with as few errors as may be specified. Error detecting and correction codes are formed by going to higher extensions and using the redundance of the extended symbols for error detection and correction. Since Shannon regarded the generation of a message to be a Markov process, it was natural to measure the effect of errors due to noise by the conditional entropy, $H(x|y)$, between the source probability space $(\Omega, A, \mathbf{p})$ with an input alphabet A with elements x, and the receiving probability space $(\Omega, B, \mathbf{p})$ with receiving alphabet B with elements y. The conditional probability matrix $\mathbf{P}$ with matrix elements $p(i|j)$ gives the probability that if letter y_j appears at the receiver that letter x_i was sent. The probabilities p_i in $(\Omega, A, \mathbf{p})$ and p_j in $(\Omega, B, \mathbf{p})$ are related by the following equation:

$$p_j = \sum_i p_i p(i|j). \tag{8}$$

The conditional entropy, $H(x|y)$ is written in terms of the components of $\mathbf{p}$ and the elements of $\mathbf{P}$:

$$H(x|y) = -\sum p_j p(i|j) \log_2 p(i|j). \tag{9}$$

The mutual entropy $I(A;B)$ that measures the amount of shared or mutual information of the input sequence and the output sequence is

$$I(A;B) = H(x) - H(x|y). \tag{10}$$

It proves more convenient to deal with the message at the source and therefore with the p_i and the $p(j|i)$ [13,14]. From Bayes' theorem on conditional probabilities we have

$$p(i|j) = p_i p(j|i)/p_j. \tag{11}$$

Substituting this expression for $p(i|j)$ in Eq. (9) we have

$$I(A;B) = H(x) - H(y|x) - \sum p_i p(j|i)[\log_2(p_j/p_i)], \tag{12}$$

where

$$H(y|x) = -\sum p_i p(j|i) \log_2 p(j|i). \tag{13}$$

$H(y|x)$ vanishes if there is no noise because the matrix elements $p(j|i)$ are all either 0 or 1 $(0 \log 0 = 0)$. The third term in Eq. (12) is the information that cannot be transmitted to the receiver if the entropy of $(\Omega, A, \mathbf{p})$ is greater than the entropy of $(\Omega, B, \mathbf{p})$. For illustration we may set all the matrix elements of $\mathbf{P}$, $p(j|i)$ to the values given in Table 5.1 in [14,15] where α is the probability of misreading one nucleotide. Substituting these matrix elements in Eqs. (12) and (13) and, replacing the logarithm by its expansion, keeping only terms of second degree we have

$$I(A;B) = H(x) - 1.7915 - 9.815\alpha + 34.2018\alpha^2 + 6.803\alpha \log_2 \alpha. \tag{14}$$

The genetic code cannot transfer 1.7915 bits of its 6-bit alphabet to the protein sequences even when free of errors. The misreading causes a decrease in the mutual entropy $I(A;B)$. The information content of any message, genetic messages included, decreases gradually until the redundance of the message is exhausted.

Just as some error can be tolerated in human languages, some error can be tolerated in the process of protein formation. Specific protein molecules having amino acids that differ from those coded for in DNA may have full specificity if the mutation is to a functionally equivalent amino acid. It is only when the supply of essential proteins decays below a critical level that protein error becomes lethal.

Eigen [5] and Eigen and Schuster [4], addressed the question of the effect of errors in the formation of proteins from considerations of the errors themselves rather than in terms of Shannon entropy. These authors find an "error threshold" to apply to the transfer of information from DNA through mRNA to protein. When calculated correctly, there is no "error catastrophe".

3. The similarity of protein sequences

We can calculate the mutual entropy of any sequences or families of sequences however they may have been generated. The mutual entropy will give us a measure of the "similarity" of the sequences or families of sequences [10,14,15]. The measure of the similarity of sequences is given in the literature as "per cent identity". "Per cent identity" is only an ad hoc score of similarity for the same reasons that error frequency is not an acceptable measure of protein error. Given a specific site in an alignment of two sequences or families of sequences, one must consider the closely related functionally equivalent amino acids. "Per cent identity" does not take into account that amino acids are almost always not equally probable and for this reason leads to illusions. Mutual entropy is the correct measure of "similarity".

4. The Central Dogma of molecular biology

Crick [3] suggested that information could flow from DNA to mRNA and from mRNA to protein but not from protein to DNA or mRNA or from protein to protein. Kolmogorov [6] proved that two sequences are not isomorphic unless they have the same entropy [9,11]. The entropy of the DNA sequence is $\log_2 61$. The entropy of the protein sequence is $\log_2 20$. These two sequences are not isomorphic. Therefore a mapping or a several-to-one code must exist to send information from DNA and mRNA to protein. For this reason information cannot be communicated from protein sequences to DNA or mRNA.

This point is important in speculations on the origin of life. Life cannot be "protein first", chemical speculations on the appearance of DNA, RNA and protein notwithstanding. The notion that life is just complicated chemistry and emerged from a "primeval soup" is one of the more distracting red herrings in the origin of life field. The evidence for a "primeval soup" has often been questioned [8]. The absence of evidence is evidence of absence.

5. The "order" and "complexity" of DNA and protein sequences

The question of whether "complexity" increases along a phylogenetic chain can be addressed only when the well-established definition and quantitative measure of "complexity" such as that given by Chaitin and Kolmogorov [14,16,17] is adopted in molecular biology. A DNA sequence of an alphabet of four letters A, C, G, T looks like a very long computer program [7].

A sequence is "highly ordered" only if it has regularities *and can be described by a much shorter sequence.* Nevertheless, some sequences of symbols that exhibit no orderly pattern from which the rest of the sequence can be predicted may have low complexity because they can be computed from an algorithm of finite information content. For example, π and e have been calculated to more than a billion digits. These two numbers satisfy all the classical conditions for a random sequence – yet each digit in turn is uniquely computable. A short computer program exists that carries all the information contained in these infinite sequences even though there is no discernible pattern.

A sequence of symbols is highly "complex" when it has little or no redundance or "order" and cannot be calculated by an algorithm of finite length. A random sequence has the highest degree of complexity, has no redundance and cannot be described except by the sequence itself. Thus π and e, although their digits have no orderly pattern, are neither complex nor random. Chaitin [1,2] has proved that no procedure exists to determine whether a given sequence can be calculated from a computer program. Consequently, it is impossible to determine whether a given sequence is random or not. The units of measurement of information content, orderliness, complexity and randomness are the bit and the byte, which are familiar to computer users as a gauge of how much information their computers can process or store.

The universal phylogenetic tree exhibits the relationship of all organisms, those from which extant organisms evolved and those to be evolved in the future [12]. The root of the phylogenetic tree represents the first stage in molecular evolution. As one attempts to follow the tree to its root by vertically derived sequences one encounters the effects of horizontal gene transfer. Horizontal gene transfer was pervasive and dominating in the early history of life and its vagaries limit the ability of genomic sequencing to follow the phylogenetic tree to the universal ancestor. The nearly universal structure of the genetic code and the handedness of proteins and nucleic acids is preserved in horizontal gene transfer and attests to a universal ancestor. Nevertheless, horizontal gene transfer has substantially erased the record of the earliest genetic sequences. This means that the earliest branches of the tree are not knowable.

6. Conclusion

The segregated, linear and digital character of the genome has allowed us to apply information theory and other mathematical theorems about sequences or strings of symbols to make a quantitative rather than an anecdotal and ad hoc discussion of significant problems in molecular biology. This procedure has led us to avoid a number of illusions common in the literature. The application of these mathematical procedures will play a role in molecular biology analogous to that of thermodynamics in chemistry.

References

[1] G.J. Chaitin, The Limits of Mathematics – A Course on Information Theory and the Limits of Formal Reasoning, Springer, New York, 1998.

[2] G.J. Chaitin, The Unknowable, Springer, New York, 1999.

[3] F.H.C. Crick, The origin of the genetic code, J. Mol. Biol. 22 (1968) 361–363.

[4] M. Eigen, P. Schuster, The hypercycle: a principle of natural self-organization. Part A: emergence of the hypercycle, Naturwissenschaften 64 (1977) 541–565.

[5] M. Eigen, Self-organization of matter and the evolution of biological macromolecules, Naturwissenschaften 58 (1971) 465–523.

[6] A.N. Kolmogorov, A new metric of invariants of transitive dynamical systems and automorphisms in Lebesgue spaces, Dokl. Akad. Nauk SSSR. 119 (1958) 861–864.

[7] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and its Applications, second ed., Springer, New York, 1997.

[8] S.J. Mojzsis, R. Krishnamurthy, G. Arrhenius, Before RNA and after: geophysical and geochemical constraints on molecular evolution, in: The RNA World, second ed., Cold Spring Harbor Laboratory Press, Cold Spring, 1999.

[9] K. Petersen, Ergodic Theory, Cambridge University Press, Cambridge, 1983.

[10] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 17 (1948) 379–424, 623–656.

[11] Y.A.G. Sinai, The notion of entropy of a dynamical system, Dokl. Akad. Nauk. 125 (1959) 768–771.

[12] C. Woese, Interpreting the universal phylogenetic tree, Proc. Natl. Acad. Sci. USA 97 (2000) 8392–8396.

[13] H.P. Yockey, An application of information theory to the Central Dogma and the sequence, J. Hypothesis Theor. Biol. 46 (1974) 369–406.

[14] H.P. Yockey, Information Theory and Molecular Biology, Cambridge University Press, Cambridge, 1992.

[15] H.P. Yockey, Origin of life on earth and Shannon's theory of communication, Comput. Chem. 24 (2000) 105–123.

[16] H.P. Yockey, A calculation of the probability of spontaneous biogenesis by information theory, J. Theoret. Biol. 67 (1977) 377–398.

[17] H.P. Yockey, Self-organization origin of life scenarios and information theory, J. Theoret. Biol. 91 (1981) 13–31.