ELSEVIER

Computers and Chemistry 24 (2000) 105–123

Computers & Chemistry

www.elsevier.com/locate/compchem

Origin of life on earth and Shannon's theory of communication

Hubert P. Yockey *

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

Received 22 December 1998; accepted 8 May 1999

Abstract

The *genetic information system* is segregated, linear and digital. It is astonishing that the technology of information theory and coding theory has been in place in biology for at least 3.850 billion years (Mojzsis, S.J., Kishnamurthy, Arrhenius, G., 1998. Before RNA and after: geological and geochemical constraints on molecular evolution 1–47. In: Gesteland, R.F. (Ed.), The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA, second ed. Cold Spring Harbor Laboratory Press, Boca Raton, FL). The genetic code performs a mapping between the sequences of the four nucleotides in mRNA to the sequences of the 20 amino acids in protein. It is highly relevant to the origin of life that the genetic code is constructed to confront and solve the problems of communication and recording by the same principles found both in the genetic information system and in modern computer and communication codes. There is nothing in the physico-chemical world that remotely resembles reactions being determined by a sequence and codes between sequences. The existence of a genome and the genetic code divides living organisms from non-living matter. If the historic process of the origin and evolution of life could be followed, it would prove to be a purely chemical process (Wächtershäuser, G., 1997. The origin of life and its methodological challenge. J. Theor. Biol. 187, 483–694). The question is whether this historic process or any reasonable part of it is available to human experiment and reasoning; there is no requirement that Nature's laws be plausible or even known to mankind. Bohr (Bohr, N., 1933. Light and life. Nature 308, 421–423, 456–459) argued that life is *consistent* with but *undecidable* by human reasoning from physics and chemistry. Perhaps scientists will come closer and closer to the riddle of how life emerged on Earth, but, like Zeno's Achilles, never achieve a complete solution. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Information; Genetic; Code; Byte; Entropy; Nucleotide; Probability

Socrates: "Every sort of confusion like these is to be found in our minds; and it is this weakness in our nature that is exploited, with a quite magical effect, by many tricks of illusion, like scene-painting and conjuring."

Glaucon: "True."

Socrates: "But satisfactory means have been found for dispelling these illusions by measuring, counting and weighing. We are no longer at the mercy of apparent differences of size and quantity and weight; the faculty which has done the counting, measuring or weighing takes control instead. And this can only be the work of the calculating or reasoning element in the soul."

The Republic, Book X, Plato (428–348 BC), translated by Francis M. Cornford, Oxford University Press, London.

* Tel.: + 1-410-879-1805.

E-mail address: hpyockey@aol.com (H.P. Yockey)

0097-8485/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.

PII: S0097-8485(99)00050-9

1. Introduction

Socrates (The Republic, Book VI, p. 745) had noted, in an earlier conversation with Glaucon, that students of geometry and reckoning first set up postulates appropriate to each branch of science, treating them as known absolute assumptions, taking it for granted that they are obvious to everybody. Thus, as Socrates taught us, the essence of science is measuring, counting and weighing together with reasoning from postulates or axioms. When science attempts to proceed from qualitative arguments, there is a danger that we may be looking at Rorschach ink blots, so to speak, and seeing what we want to see. The more discussions can be made quantitative and avoid ad hoc explanations the better theoretical biology is served. Only by expressing our knowledge in numbers and employing calculation and reasoning can we separate fact from the tricks of illusion in determining whether any proposal on the nature, origin and evolution of life can be retained or discarded. This breaks molecular biology out of sophisticated Kipling's Just So Stories into the quantitative mode, used by natural scientists (Wolynes, 1998).

Charles Darwin (1809–1882) and almost all naturalists of his time, believed in the blending theory of genetics (Jenkin, 1867). However, Gregor Johann Mendel (1822–1884), published a paper in 1865 (Mendel, 1865) that proved inherited characteristics are segregated and do not blend. It was not until 1930 that Fisher (Fisher, 1930) showed that the blending theory is incorrect and that natural selection proceeds according to the particulate laws of Gregor Mendel. The next important discovery in genetics was that the mechanism of inheritance is linear. El momento de la verdad came in 1953, when Watson and Crick discovered that the formation of biomolecules is controlled by the sequences of nucleotides recorded in the double helix of DNA (or RNA in some viruses). Thus the genetic information system is segregated, linear and digital. It is astonishing, furthermore, that the technology of information theory and coding theory, just now being discovered, has been in place in biology for at least 3.850 billion years. Information theory and coding theory and their tools of measuring information in sequences are essential to understanding the crucial questions of the nature and origin of life.

The development of a human being is guided by just 750 megabytes of digital information (Olson, 1995). It could be stored on a single CD-ROM in the biologist's personal computer. The genome in principle contains all the information necessary to bridge the gap between genotype and phenotype. All information necessary to determine the three-dimensional structure of proteins lies in the form of amino acid sequences, yet we remain unable to predict their tertiary structures (Huynen and Bork, 1998).

Chemical or physical reactions in non-living matter are not controlled by a message. If the genetical processes were purely chemical, as mechanists–reductionists believe, the law of mass action and thermodynamics would govern the placement of amino acids in the protein sequences according to their concentration. Therefore, if a scenario for the origin of life is to be acceptable, it must show how a genetic message was generated and attained a minimum threshold of complexity — which is measurable in bits and bytes — in a molecule characteristic of life and needed for the assimilation of carbon dioxide and nitrogen by the protobiont.

First, how large an information content does a genetic message require to be characteristic of life? Second, do the laws of physics and chemistry have enough information content so that non-living matter may organize itself and become living matter, say in the manner that crystals are formed? Computer users are familiar with the rather large memory requirements, measured in bytes, of complicated programs and the capacities of their hard drives. The same applies to programs that purport to generate the genome of the protobiont.

2. The mathematical theory of sequences

I shall therefore approach the subject from the mathematical theory of sequences and the codes between sequences. First we must establish a mathematical meaning and a measure of the *information* in a message. Shannon (1948) established information theory as a mathematical discipline in his classic paper and pointed out, in the second paragraph:

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering (biological) problem. The significant aspect is that the actual message is *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will be actually chosen since this is unknown at the time of design."

Similarly, the genetic logic system must be capable of accommodating the genetic messages of all organisms that have ever lived, live now or will be evolved in the future. The DNA sequences that make up the genome of any organism are selected from a set of possible genetic messages. Thus the biological information system or genetic logic system is independent of the specificity or meaning of the genetic message in the genome. Likewise, communication systems will process meaningless noise as well as a play by Sophocles.

To some people it may seem counter-intuitive that the information in a sequence or a message can be measured, while there is no measure for *meaning, knowledge, erudition or specificity*. We have accepted this as true in our daily life in many ways. As computers, hard drives, modems and other devices that measure information in bits and bytes are becoming increasingly commonplace, one becomes aware that information and meaning are separate concepts.

The meaning, if any, of words, that is, a sequence of letters, is arbitrary. It is determined by the natural language and is not a property of the letters or their arrangement. For example, the English word 'hell' means bright in German, 'fern' means far, 'Gift' means poison, 'bald' means soon, 'Boot' means boat, "singe" means sing. In French 'pain' means bread, 'ballot' means a bundle, 'coin' means a corner or a wedge, 'chair' means flesh, 'cent' means hundred, 'son' means his, 'tire' means a pull, 'ton' means your. This confusion of meaning goes as far as sentences. For example, 'O singe fort!' has no meaning as a sentence in English, although each is an English word, yet in German it means 'O sing on!' and in French it means 'O strong monkey.' However, while the meaning of a sequence of letters is arbitrary, the meaning or specificity of a sequence of amino acids composing a protein is not arbitrary, but rather is determined by nature.

In formal probability theory, one must establish the probability sample space Ω consisting of the set A, of events, x, under discussion, which are random variables with the corresponding probabilities, p_i. The set of probabilities p_i form a probability vector **p**. The probability sample space is designated (Ω, A, **p**). All messages are formed from a sequence of the members of a finite alphabet. For example, the primary alphabet used in electronic communication is [0, 1]; in molecular biology they are the nucleotide bases in DNA, RNA and the amino acids in protein. The members of these alphabets are the events in the probability sample space.

Gilbert (1987) and Gilbert et al. (1997) calculated the number of possible sequences in polypeptide chains of length N by the following expression:

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

Eq. (1) gives the total number of sequences we must be concerned with only if all events are equally probable. However, many events in general and amino acids in particular do not have the same probability. Does neglecting that fact affect our analysis? Shannon (1948) addressed this problem as follows: Let us consider a long sequence of N symbols selected from an alphabet of A letters or events. In the present case the letters or events will be the alphabet of either codons or amino acids. With high probability the sequence will contain $Np(i)$ of the ith symbol. Let P be the probability of the sequence. Then:

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

Taking the logarithm of both sides:

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

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

Where

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

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 approximately:

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

The number of sequences of length N is approximately:

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

Notice that the expression for H was not introduced ad hoc; rather it comes out of the woodwork, so to speak.

Let us apply Eq. (7) to calculating the number of sequences in 100 throws of a dice game where the probabilities of all events are known exactly and are not all equal. For a given throw the probability of 2 and 12 is 1/36 whereas the probability of 7 is 6/36 because there are six ways to roll a 7 and only one way to roll 2 or 12. Substituting the probabilities of each of the possible events in Eq. (7) to calculate H one finds that the number of sequences is only 2.69×10^{-6} of that calculated from Gilbert's expression (Eq. (1)) 11^N.

We have approached the calculation of the number of sequences of length N in two apparently correct ways and the question arises as to what happened to the sequences left out of the total possible by the second method. This is explained by the Shannon–McMillan–Breiman theorem (Shannon, 1948; McMillan, 1953; Breiman, 1957):

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.

The expression for the measure of information that appears in Shannon (1948) is:

$$H = -K\Sigma_i p_i \log p_i \tag{8}$$

where the p_i are the probabilities of the ith member of the alphabet. Shannon was quick to explain that K is a positive constant that 'merely amounts to a choice of a unit of measure'. K appears from the mathematical argument that justifies Eq. (8). It has no physical dimensions and when the logarithm is taken to the base 2 and $K = 1$, H is measured in bits, a contraction of *b*inary dig*it*. Like all messages, the life message is non-material but has a measurable information content. Of course, the genetic message, although non-material, must be recorded in matter or energy. Now in this paper, the term *information* has the meaning given in Eq. (8). It does not mean *knowledge*, although a message composed of a sequence of symbols may transfer knowledge to the receiver of the message.

The expression given in Eq. (8) is very similar to that for entropy in statistical mechanics. As I explained in Yockey (1992), entropy is a general term and must be defined as a function of the probability space and the probability distribution of the elements or events to which it refers. As Hamming (1986) wrote:

> "For us entropy is simply a function of a probability distribution p_i. ...The confusion at this point has been very great for outsiders who glance at information theory."

Every probability sample space has an entropy. For example, the probability sample space in classical statistical mechanics, called *phase space* by theoretical physicists, is six-dimensional and the p_i are defined by the position and momentum vectors of the particles in the ensemble. The function for entropy in both classical statistical mechanics and quantum statistical mechanics has the dimensions of the Boltzmann constant k and has to do with energy and momentum, not information. However, entropy in information theory and probability theory has no mechanical dimensions. There are no counterparts in communication theory to temperature, energy, pressure, work or volume. There is, furthermore, no counterpart to the First Law of Thermodynamics, namely, the conservation of the energy of a system. To illustrate this point further, one may consider the probability space of a dice game that consists of the numbers 2 through 12 as random variables and calculate the corresponding entropy (Yockey, 1992, see exercise on page 88). Clearly, the entropy of a dice game has nothing whatever to do with statistical mechanics or thermodynamics. It may have something to do with information theory since a sequence of symbols selected from the alphabet is generated as a stochastic process by a series of tosses of the dice. Such a sequence of letters forms a message in which some gamblers find meaning or knowledge by which they make their bets. On the other hand, information theory is concerned with messages expressed in sequences of letters selected from a finite alphabet by a Markov process. The probability sample space is constructed of the letters of the alphabet under consideration as random variables and the p_i are defined accordingly.

3. The extensions of an alphabet: definition of the byte

A binary alphabet consisting of zero and one, each containing one bit of information, is used in the source alphabet of computers and all communication technology. The receiving alphabets in all applications have more than two letters. The binary source alphabet must be *extended* by forming pairs, triplets, quadruplets and so forth to form receiving alphabets larger than two. These alphabets are called the first, second, third and so forth extensions of the primary binary alphabet. These extensions are called a *byte*.

Packaged items in stores have a bar code attached that permits the cashier to record the price of the item. The Postal Service in the United States has established a ZIP + 4 bar code in which mailing addresses can be written. The alphabets of the postal and other bar codes are composed of short bars and long bars, that is, the source alphabet is binary. There are 32 ($2^5 = 32$) members of the fifth extension of the source binary bar code alphabet. Thus the postal bar code has a five bit byte. An important receiving alphabet is one in which the ten numbers from 0 to 9, have two ones or two zeros. The source probability space (Ω, A, $\mathbf{p}$) alphabet is mapped on the receiver probability space alphabet, (Ω, B, $\mathbf{p}$). The Postal Service has assigned arbitrarily the ten members of the fifth extension that have two ones to one of the ten digits in the decimal system. The ten members selected from the fifth extension are called *sense code letters*. The assignment of sense code letters in the fifth extension alphabet of the postal ZIP + 4 code is shown by the first row in Table 1.

The other code letters are called *non-sense* because they have been given *no sense* or *meaning* assignment in the receiving alphabet. (Remember that non-sense dose not mean nonsense or foolishness.) I have listed those non-sense code letters of the fifth extension of the binary code with just one change from the sense code letters in each column, five rows down. Notice that the postal ZIP + 4 code is an error detecting code. A single error does not change one sense code letter to another sense code letter. If, due to smudging or other malfunction, the sorting machine reads a non-sense code letter the mail piece is rejected to be examined by a postal employee.

Computer equipment uses the seventh extension, ASCII [A(merican) S(tandard) C(ode for) I(nforma-

Table 1
The postal ZIP+4 code

0	1	2	3	4	5	6	7	8	9
11000	00011	00101	00110	01001	01010	01100	10001	10010	10100
01000	10011	10101	10110	11001	11010	11100	00001	00010	00100
10000	01011	01101	01110	00001	00010	00100	11001	11010	11100
11100	00111	00001	00001	01101	01110	01000	10101	10110	10000
11010	00001	00111	00100	01011	01000	01110	10011	10000	10110
11001	00010	00100	00111	01000	01011	01101	10000	10001	10101

tion) I(nterchange)] computer code. The byte is 7 bits, the amount of information in one printed character in the receiver alphabet. There are $2^7 = 128$ members of the seventh extension alphabet. Members of this seventh extension of the primary alphabet are assigned arbitrarily to the letters in the receiving alphabet of the word processor. This assignment is appropriate to the language for which the word processor is designed.

4. The genetic code

In another corner of the groves of academe, remote from the activities of mathematicians and communication engineers, after the discoveries of Watson and Crick in 1953, it was realized that there must be a code that performs a mapping between the sequences of the four nucleotides in DNA to the sequences of the 20 amino acids in protein. George Gamow made the first suggestion of an overlapping code. It was soon proved to be incorrect, but the search was on. It is easy to see that overlapping codes impose severe restrictions on the allowed amino acid sequences. Since it was clear that there are 64 possible triplets and only 20 amino acids, the genetic code was believed to be 'degenerate' and that some codons must be 'nonsense'. Both terms are unnecessarily pejorative and non-constructive. The first sentence of Sonneborn (1965) expresses the attitude of the time: "Part I of this paper deals with the amount of degeneracy in the genetic code or, looked at in reverse, the amount of nonsense." Unfortunately, the use of the word 'nonsense' persists in some current publications (Maeshiro and Kimura, 1998).

The apartheid in the groves of academe between mathematicians, electrical engineers and molecular biologists prevented the notion of block codes, initiators and enders to be applied to the understanding of the genetic code and the genetic logic system. For example, there are three different frames in which a sequence of nucleotides in DNA can be read. Crick et al. (1957) assumed that there must be 'commas' to separate a string of nucleotides forming codons to prevent them from reading out of frame. They attempted to show that the maximum number of sense codons cannot be greater than 20 and gave a solution for the 20. Unfortunately they proved (sic) that the triplet AAA and other triplets must be nonsense (sic). Table 2 shows that AAA codes for lysine, CCC codes for proline, UUU codes for phenylalanine and GGG codes for glycine. Furthermore, a tRNA species possesses an anticodon complementary to UGA, one of the nonsense codons, which codes for selenocysteine, thus making 21, not 20 amino acids (Hawkes and Tappel, 1983; Sunde and Evenson, 1987; Leinfelder et al., 1988; Mizutani and Hitaka, 1988). The use of an initiator codon and codons to terminate the sequence and the realization that the genetic code is a block code, like the bar codes discussed above, makes it unnecessary to assume that commas are needed.

The source alphabet of the genetic code is the four nucleotides of DNA and mRNA. Thus each nucleotide has a two-bit byte. The first extension of that quaternary alphabet has sixteen letters. That is not sufficient to code the canonical 20 amino acids that are transcribed in protein; accordingly, Nature has gone to the second extension 64-letter alphabet. Thus the genetic code has a six-bit byte, usually called a codon. Sometimes the codon is called a word, but clearly this is inappropriate. The codon is a letter in the second extension of the mRNA alphabet; the protein is a word. The genetic code, shown in Table 2, shares a number of properties with the postal ZIP + 4 code and the ASCII computer codes. The genetic code is a block code because all codons are triplets. The genetic code is distinct and uniquely decodable because the single methionine codon AUG, and sometimes the leucine codons UUG and CUG, serve as a starting signal for the protein sequence and perform the same function as the long frame bar at the beginning of the postal message in the ZIP + 4 code. The non-sense codons UGA, UAA and UAG stop the translation of the protein from the mRNA and initiate the release of the protein sequence from the mRNA (Maeshiro and Kimura, 1998). They perform the same function as the long frame bar at the end of the postal bar code message.

Table 2
The mRNA genetic code

Amino acid	Triplet codons	Amino acid	Triplet codons	Amino acid	Triplet codons
Glycine	GGG	Phenylalanine	UUU	Leucine	UUA
	GGC		UUC		UUG
	GGU				
	GGA				
Proline	CCG	Cysteine	UGU	Tryptophan	UGG
	CCC		UGC	Nonsense	UGA
	CCU				
	CCA				
Leucine	CUG	Glutamine	CAA	Histidine	CAU
	CUC		CAG		CAC
	CUU				
	CUA				
Arginine	CGG	Aspartic acid	AAU	Lysine	AAA
	CGC		AAC		AAG
	CGU				
	CGA				
Threonine	ACG	Glutamic acid	GAA	Asparagine	GAU
	ACC		GAG		GAC
	ACU				
	ACA				
Valine	GUG	Isoleucine	AUU	Methionine	AUG
	GUC		AUC		
	GUU		AUA		
	GUA				
Alanine	GCG	Nonsense	UAA	Tyrosine	UAU
	GCC		UAG		UAC
	GCU				
	GCA				
Serine	UCG			Arginine	AGA
	UCC				AGG
	UCU			Serine	AGU
	UCA				AGC

5. The origin and evolution of the genetic code

The Standard Genetic Code as represented in Table 2 was once thought to be universal in all biology, indeed, in the earlier papers it was sometimes refereed to as the Universal Genetic Code. Crick (1968) proposed that the genetic code was a 'frozen accident'. It has been learned since that the genetic code is not universal, nor is it fixed or frozen. Placing this problem in the hands of the fickle goddess Fortuna begs the question. There has been considerable speculation about the 'order' in the genetic code. Any small group of chance events seems to have 'order'. This is the gambler's ruin. As the ancient Greeks and Romans presumed to see the future in the flight of birds, he sees a pattern or 'order' in the fall of the dice or the cards. Fortuna, the goddess of chance, may allow him a brief episode of luck but inevitably she turns against him to his ruin (Feller, 1957).

Jukes (1966, 1983) suggested that the second extension triplet genetic code may have evolved from a preceding first extension doublet code. Jukes' archetypal genetic code had 16 codons that translated 14 or 15 amino acids together with a stop codon (Jukes, 1966). All amino acid codons except Methionine and Tryptophan have some redundance in the third place in the codon; this lends some support for that suggestion. That redundance provides some protection from mutational error. I have given a detailed discussion of the consequences of the evolution of the standard and mitochondrial genetic codes from an archetypal doublet code in which the third nucleotide is silent (Yockey, 1992).

Crick (1968) pointed out, correctly, that any change in the code would cause errors throughout the process of protein formation, resulting in a disruption that would be lethal. However, Jukes (1993) noted that codon use varies considerably and that in some organisms certain codons fall out of use altogether. He suggested that after a time such codons had been reassigned or captured. For example, the codon in vertebrate mitochondria for methionine is AUA, and UGA for tryptophan. The most recent studies of codons that code for amino acids that differ from those in the Standard Genetic Code in Table 2 is in Osawa (1995) and Osawa et al., (1990, 1992) and Maeshiro and Kimura (1998).

6. The Central Dogma of Francis Crick

With the appearance of the experimental knowledge about the transfer of information from the four letter alphabet of DNA and of mRNA, the question emerged among molecular biologists of how a four letter alphabet could send information to a 20 letter alphabet. Crick (1968), in a remarkably prescient paper, suggested that information could be transferred from DNA to RNA and from RNA to protein, but not from protein to protein. He called this by the somewhat ecclesiastical name *The Central Dogma* of molecular biology (Crick, 1970). The reason for these statements, as we have seen above, is that Nature had extended the primary four letter alphabet to the six-bit, 64-codon alphabet of the genetic code. Thus the genetic code is a several-to-one code. Note also, that the dice game is also a several-to-one code and for that reason has a Central Dogma. The source alphabet has 36 members, namely, all pairs of the numbers 1 through 6. These pairs of numbers are analogous to codons in the genetic code. The receiver alphabet is formed by adding the two numbers and has only 11 members, the numbers 2 through 12. These numbers are analogous to the amino acids in protein sequences. Except for 2 and 12, all totals can be made by more than one combination of numbers read from the dice. Thus there is a loss of information (and knowledge) when the numbers are added because the logic operation lacks a single-valued inverse. This is known as a logic ADD gate and is irreversible. Thus the Central Dogma is a theorem in coding theory and the logic of the genetic communication system. It does not come from biochemistry. In the full glory of its mathematical generality it applies to all codes where the information entropy of the source alphabet is larger than that of the receiver alphabet (Kolmogorov, 1958; Billingsley, 1965; Shields, 1973; Ornstein, 1974; Yockey, 1992).

It is obvious that if the source and receiver alphabets have the same number of symbols, and a one-to-one correspondence between the members of the alphabets, the logic operation has a single valued inverse and information may be passed, without loss, in either direction. Thus, since RNA and DNA both have a four-letter alphabet, genetic messages may be passed from RNA to DNA. The passage of information from protein to RNA or DNA is prohibited for two reasons: (1) the logic OR gate is irreversible, and (2) there is not enough information in the 20 letter protein alphabet to determine the 64 letter mRNA alphabet from which it is translated (Yockey, 1992, 1995a). There is no mathematical restriction of the transfer of information from protein to protein. This was once thought to be the means by which prions were formed but that is no longer believed to be the case. There appears to be no biological mechanism by which information can be exchanged between two proteins.

7. The genetic code and error in the genetic message

Let us refer to the second paragraph in Shannon's 1948 paper where he wrote:

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

Shannon modeled the generation of a message as a Markov process. At all stages of the communication system the message may be acted upon by a second Markov process leading to an interchange in some of the letters in the message in a random and non-reproducible fashion. The result of this process is called noise.

Fig. 1 (from Information Theory and Molecular Biology (1992), Cambridge University Press, with permis-

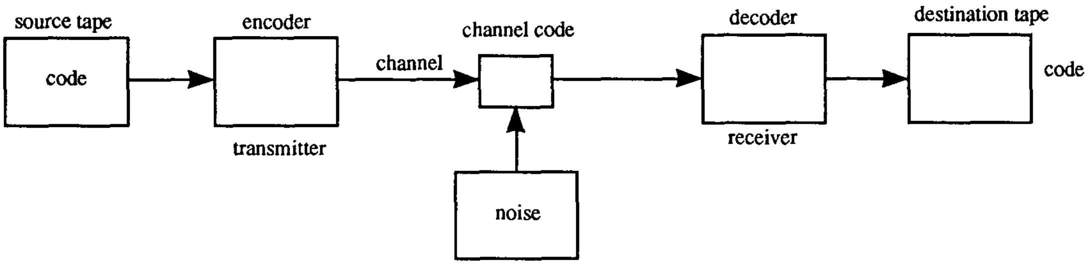

Fig. 1. The transmission of information from source to destination as conceived in electrical engineering. Noise occurs in all stages but is shown according to accepted practice in electrical engineering.

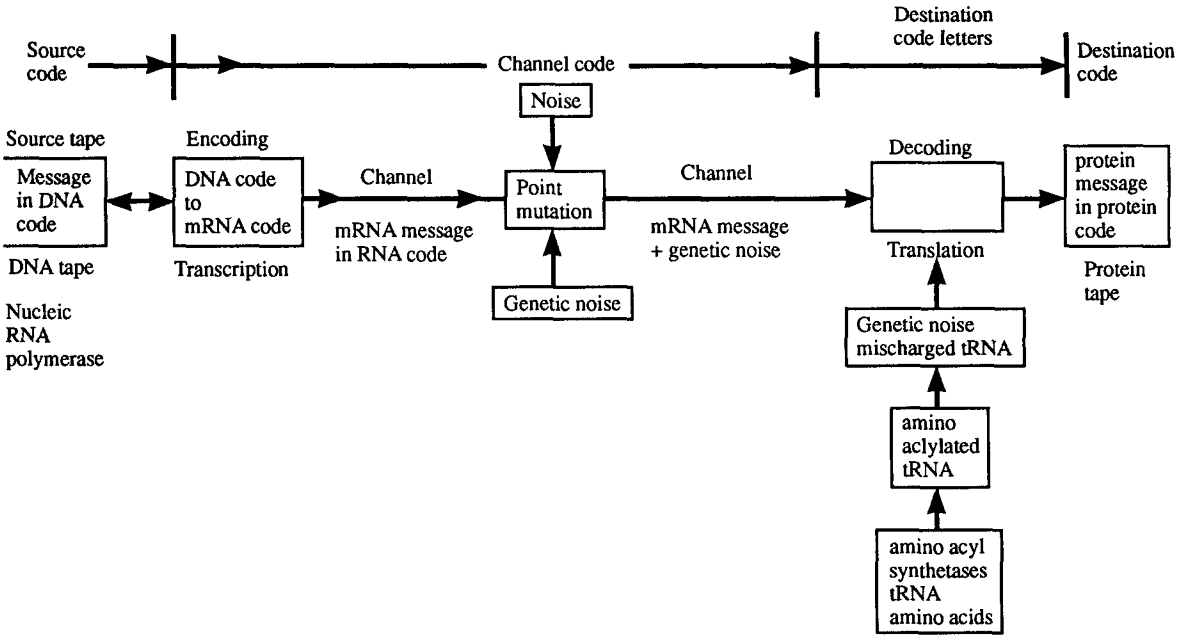

Fig. 2. The transmission of genetic messages from the DNA to the protein tape as conceived in molecular biology. Genetic noise occurs in all stages but is lumped in the figure to fix the idea.

sion) shows the basic elements of a communications system as presented by Shannon. In Fig. 2 (from Information Theory and Molecular Biology, Cambridge University Press, with permission) we see the communication of information from the genome as recorded in DNA to the message that determines the protein. The sequence of nucleotides that compose DNA plays the same role as the tape in a tape recording machine. The protein molecule, which is the destination of the message, is also a tape. Thus the one-dimensional genetic message is recorded in a sequence of amino acids, which folds up to become a three-dimensional active protein molecule. One is reminded of the linear sequence of signals from a television station that, because of the raster, defines a two-dimensional object, namely, the picture on the screen.

A reading error that afflicts reading the message in the postal code, and indeed, all codes, is called genetic noise in Fig. 2. There is very little error detection or correction attributed to the structure of the genetic code. Nevertheless, the genetic information system is extremely accurate. The error frequency in the first aminoacylation of an amino acid on its adapter tRNA is about 3×10^{-4}. It is a curious phenomenon that there is a proofreading or double sieving mechanism that checks this process by discarding 'wrong' amino acids. It also has an error frequency of about 3×10^{-4}. The error frequency after the operation of both mechanisms is the product $3 \times 10^{-4} \times 3 \times 10^{-4} = 9 \times 10^{-8}$ (Freist et al., 1998).

8. 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. The theorem is not constructive so one must use one's ingenuity to construct suitable codes. 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 (Hamming, 1986). The intense academic and industrial interest in the improvements in communication promised by Shannon's Channel Capacity Theorem led to the discovery by mathematicians and electrical engineers of a number of error correcting codes.

Biologists, working in another part of the groves of academe, confronted the problem of error in the formation of protein from the sequence of nucleotides in DNA. Although they had the idea of error accumulation and its effect on the viability of organisms, regrettably they attempted to consider the problem in terms of the errors themselves (Orgel, 1963, 1970). This led many of them to speculate that an 'error catastrophe' would occur inevitably as errors accumulated, leading to the death of the organism. Eigen (1971, 1992) and Eigen and Schuster (1977), also address the question of the effect of errors in the formation of proteins from considerations of the errors themselves. These authors find an 'error threshold' to apply to the transfer of information from DNA through mRNA to protein.

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. Protein molecules rendered inactive because of mischarged amino acids that are not functionally equivalent are degraded by proteolytic enzymes. It is only when the supply of essential proteins decays below a critical level that protein error becomes lethal. Although other considerations on the aging process have been made (Holliday, 1986), the accuracy of protein biosynthesis is still of topical interest (Freist et al., 1998).

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, $H(x \mid y)$, 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 $p(i \mid j)$ 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:

$$p_j = \Sigma_i p_i p(i \mid j) \tag{9}$$

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

$$H(x \mid y) = \Sigma_i p_j p(i \mid j) \log_2 p(i \mid j) \tag{10}$$

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

$$I(A; B) = H(x) - H(x \mid y) \tag{11}$$

It proves more convenient to deal with the message at the source and therefor with the p_i and the $p(j \mid i)$ (Yockey, 1974, 1992). From Bayes' theorem on conditional probabililities, we have (Feller, 1957):

$$p(i \mid j) = p_i p(j \mid i)/p_j \tag{12}$$

Substituting this expression for $p(i \mid j)$ in Eq. (10) we have:

$$I(A; B) = H(x) - H(y \mid x) - \Sigma_i p_i p(j \mid i)\ [\log_2 (p_j/p_i)] \tag{13}$$

where

$$H(y \mid x) = -\Sigma p_i p(j \mid i) \log_2 p(j \mid i) \tag{14}$$

$H(y \mid x)$ vanishes if there is no noise because the matrix elements $p(j \mid i)$ are all either 0 or 1. The third term in Eq. (13) 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 $\mathbf{P}$, $p(j \mid i)$ to the values given in Table 3 where α is the probability of misreading one nucleotide. Substituting these matrix elements in Eqs. (13) and (14) and, replacing the logarithm by its expansion, keeping only terms of second degree we have:

$$I(A; B) = H(x) - 1.7915 + 34.2018\alpha^2 + 6.803\alpha \log_2 \alpha \tag{15}$$

The genetic code cannot transfer 1.7915 bits of its six-bit alphabet to the protein sequences even when free of errors. The reader may find it instructive and amusing to carry out these calculations for the dice game. How much information is lost by adding the numbers on the dice? One must write zero instead of α for the matrix element in Table 3 if the replacement amino acid is functionally equivalent at each site in the protein sequence. I have pursued this discussion to calculate the information content of the cytochrome c molecule to be 374 bits (Yockey, 1992).

We can replace the sender and receiver and 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 (Shannon, 1948; Yockey, 1992; Tononi et al., 1996, 1999). Perhaps because of the apartheid in the groves of academe, the measure of the similarity of sequences is given in the literature as 'per cent identity'. 'Per cent identity' (Altschul et al., 1997; Levitt and Gerstein, 1998) 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. Since as Socrates said, "...but satisfactory means have been found for dispelling these illusions by measuring, counting and weighing." the mutual entropy should be used as the only correct measure of 'similarity'.

According to mechanists–reductionists, non-living matter can self-organize, and so life and evolution are manifestations of the spontaneous emergence of 'order'. Living matter, they say, has both 'order' and 'complexity.' Information theory shows that the ideas of 'complexity' and 'order' are contradictory and mutually exclusive. Crystals are highly ordered due to the regularities in the placement of their molecules as directed by the laws of physics and chemistry. However, these regularities, which allow crystals to be described by a

Table 3
Genetic code transition probability matrix elements $p(j/i)^a$

Amino acid y_j	Leu	Ser	Arg	Ala	Val	Pro	Thr	Gly	Ile	Term	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Phe	Trp	Met
Codon x_i																					
UUA	$(1-7\alpha)$	α			α				α	2α									2α		
UUG	$(1-7\alpha)$	α			α					α									2α	α	α
CUU	$(1-6\alpha)$		α		α	α			α			α							α		
CUC	$(1-6\alpha)$		α		α	α			α			α							α		
CUA	$(1-5\alpha)$		α		α	α			α				α					α			
CUG	$(1-5\alpha)$		α		α	α							α					α			α
UCU		$(1-6\alpha)$		α		α	α				α								α		
UCC		$(1-6\alpha)$		α		α	α				α								α		
UCA	α	$(1-6\alpha)$		α		α	α			2α		α		α				α			
UCG	α	$(1-6\alpha)$		α		α	α			α		α		α				α			
AGU		$(1-8\alpha)$	3α				α	α	α				α					α		α	
AGC		$(1-8\alpha)$	3α				α	α	α				α					α			
CGU	α	α	$(1-6\alpha)$			α		α							α						
CGC	α	α	$(1-6\alpha)$			α		α							α					α	
CGA	α		$(1-5\alpha)$			α		α								α				α	α
CGG	α		$(1-5\alpha)$				α	α		α						α					
AGG		2α	$(1-7\alpha)$				α	α									α				
AGA		2α	$(1-7\alpha)$				α	α	α	α							α				
GCU		α		$(1-6\alpha)$	α	α	α	α								α			α		
GCC		α		$(1-6\alpha)$	α	α	α	α								α			α		
GCA		α		$(1-6\alpha)$	α	α	α	α									α				
GCG		α		$(1-6\alpha)$	α	α	α	α									α				
GUU	α			α	$(1-6\alpha)$			α	α												
GUC	α			α	$(1-6\alpha)$			α	α												α
GUA	2α			α	$(1-6\alpha)$			α	α												
GUG	2α			α	$(1-6\alpha)$			α													
CCU	α	α	α	α		$(1-6\alpha)$	α					α									
CCC	α	α	α	α		$(1-6\alpha)$	α					α									
CCA	α	α	α	α		$(1-6\alpha)$	α						α								
CCG	α	α	α	α		$(1-6\alpha)$	α						α								

Table 3 (Continued)

Amino acid y_j	Leu	Ser	Arg	Ala	Val	Pro	Thr	Gly	Ile	Term	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Phe	Trp	Met
ACU		2α		α		α	$(1-6\alpha)$		α					α							
ACC		2α		α		α	$(1-6\alpha)$		α					α							
ACA		α	α	α		α	$(1-6\alpha)$		α						α						α
ACG		α	α	α		α	$(1-6\alpha)$								α	α		α			
GGU		α	α	α	α			$(1-6\alpha)$									α				
GGA			2α	α	α			$(1-6\alpha)$		α						α		α		α	
GGC		α	α	α	α			$(1-6\alpha)$									α		α		α
GGG			2α	α	α			$(1-6\alpha)$											α		α
AUU	α	α			α		α		$(1-7\alpha)$												α
AUC	α	α			α		α		$(1-7\alpha)$					α							
AUA	2α		α		α		α		$(1-7\alpha)$					α							
UAA	α	α			α					$(1-7\alpha)$	2α				α					α	
UAG	α	α						α		$(1-8\alpha)$	2α		α		α		α			α	
UGA	α	α	2α							$(1-8\alpha)$			α		α		α	2α			
UAU		α	α							2α	$(1-8\alpha)$	α		α		α		α	α		
UAC		α								2α	$(1-8\alpha)$	α		α		α		α	α		
CAU	α		α			α					α	$(1-8\alpha)$	2α	α		α					
CAC	α		α			α					α	$(1-8\alpha)$	2α	α		α					
CAA	α		α			α				α		2α	$(1-8\alpha)$		α		α				
CAG	α		α			α				α		2α	$(1-8\alpha)$		α		α				
AAU		α					α		α		α	α		$(1-8\alpha)$	2α	α					
AAC		α					α		α		α	α		$(1-8\alpha)$	2α	α					
AAA			α				α		α	α			α	2α	$(1-8\alpha)$		α				α
AAG			α				α			α			α	2α	$(1-8\alpha)$		α				
GAU				α	α			α			α	α		α		$(1-8\alpha)$	2α				
GAC				α	α			α			α	α		α		$(1-8\alpha)$	2α				
GAA				α	α			α		α			α		α	2α	$(1-8\alpha)$				
GAG				α	α			α		α			α		α	2α	$(1-8\alpha)$				
UGU		2α	α					α		α	α							$(1-8\alpha)$	α	α	
UGC		2α	α					α		α	α							$(1-8\alpha)$	α	α	
UUU	3α	α			α				α		α							α	$(1-8\alpha)$		
UUC	3α	α			α				α		α							α	$(1-8\alpha)$		
UGG	α	α	2α					α		2α								2α		$(1-9\alpha)$	
AUG	2α		α		α		α		3α						α						$(1-9\alpha)$

[a] Reformatted from Yockey HP. An application of information theory to the central dogma and the sequence hypothesis. J Theor Biol 1974;46:369–406. Published with permission.

short sequence or algorithm, mean that crystals have a very low information content. Information theory narrows this definition to say that we may only call a sequence 'highly ordered' if it has regularities and can be described by a much shorter sequence or algorithm. The information content of a sequence is measured in the bits (or bytes).

Sequences of symbols that exhibit no orderly pattern from which the rest of the sequence can be predicted may, nevertheless, have low complexity because they can be computed from an algorithm of finite information content (Pincus and Kalman, 1997; Pincus and Singer, 1996). For example, π and e have been calculated to sequences of more than a billion digits. They exhibit no orderly pattern — yet each digit in turn is computable. We may conclude, in general, that a finite message, specified by a computer program, may exist that carries all the information contained in an infinite sequence even though there is no discernible pattern in the sequence of symbols.

A sequence of symbols is highly 'complex' when it has little or no redundance or 'order' and cannot be described by a much shorter sequence or 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. Thus when we speak of the amount of complexity in a sequence we are speaking about the amount of its randomness as well (Yockey, 1990).

The principle, that computer data compression software may reduce the order, redundancies, patterns and regularities peculiar to natural languages, is based on a theorem in information theory by Shannon (1948). The result, achieved after the compression, is nearly indistinguishable from a random sequence since most of the patterns and regularities have been removed.

9. Information theory and the origin of life: the gap between living and non-living

It is a curious fact, highly relevant to the origin of life, that the genetic code is constructed to confront and solve the problems of communication and recording by the same principles found both in the genetic information system and in modern computer and communication codes. The origin of a rather accurate genetic code is a pons asinorum that must be crossed to pass over the abyss that separates crystallography, high polymer chemistry and physics from biology. As I mentioned above, there is nothing in the physico-chemical world that remotely resembles reactions being determined by a sequence and codes between sequences. The existence of a genome and the genetic code divides living organisms from non-living matter.

The origin of life has concerned philosophers and theologians since antiquity, for example, the idealist philosopher, Immanuel Kant (1724–1804) in his Critique of Judgment. (A quotation in English translation can be found in Wächtershäuser (1997).) It was not until the nineteenth century that these considerations can be called scientific. There were, in general, three philosophical approaches to the origin of life. The first was vitalism, the belief there is a metaphysical, supernatural, non-material, idealist élan vital, a vital principle or life force, which distinguishes living from non-living matter. Vitalism has its roots in the German idealist philosophy of F.W.A. Schelling (1775–1854) and others in the nineteenth century, members of a romantic philosophic movement, Naturphilosophie, who believed all creation was a manifestation of a World Spirit. They believed all matter possessed this Spirit and organized bodies had it to an intense degree. In the nineteenth century it was quite possible to be a vitalist, believing in a vital force or élan vital, without thinking of the vital force being supernatural. At the time it was as valid to attribute the laws and effects of vitality to a nonmaterial vital force as it was to attribute the laws and effects of gravity to a nonmaterial gravitational force. As Darwin put it, "Who can explain what is the essence of the attraction of gravity? No one now objects to following out the results consequent on this unknown element of attraction." (Origin of Species, Chapter XV) Vitalism is no longer considered since the discovery by Watson and Crick of the role played by DNA and RNA in the formation of protein. The second was mechanist–reductionism. A group of young physiologists at an 1869 meeting in Innsbruck, Austria, declared: "The ultimate objective of the natural sciences is to reduce all processes in Nature to the movements that underlie them and to find their driving forces, that is, to reduce them to mechanics (Mayr, 1982)." Wächtershäuser (1997) is a current supporter of mechanist–reductionism, the view that if the historic process of the origin and evolution of life could be followed it would prove to be a purely chemical process. The third was the dialectical materialism of Friedrich Engels (1954), supported by Oparin (1957). According to dialectical materialism, the appearance of life is achieved, not through the laws of physics and chemistry, but through the Law of the Transformation of Quantity into Quality. For that reason the proponents of dialectical materialism are extremely hostile to mechanist–reductionism. Engels (1820–1895) had observed in one of the fragments of Dialectics of Nature:

> "Mechanism applied to life is a hopeless category, at the most we could speak of chemism, if we do not want to renounce all understanding of names."

Oparin, upon reading that, abandoned reductionism–mechanism and became an immediate convert to dialectical materialism. Oparin, in his later books, went into great detail to denounce reductionism–mechanism and support dialectical materialism. Western intellectual apologists for Oparin who are not trained in the intricacies of philosophy may find the relationship between reductionism–mechanism and dialectical materialism to be a distinction without a difference (Lazcano, 1997; Miller et al., 1997).

I shall pursue those proposals that purport to generate a genome and a genetic code. A central speculation involves the premise that life originated from a primeval soup of organic substances in the ocean by prebiotic or chemical evolution, a belief that life would arise spontaneously in flagrante delicto from this non-living matter and that it would almost inevitably arise on 'sufficiently similar young planets elsewhere'. George Gaylord Simpson (1964) pointed out that the National Aeronautics and Space Administration (NASA) has a program in 'space bioscience'. "There is even increasing recognition of a new science of extraterrestrial life, sometimes called exobiology — a curious development in view of the fact that this 'science' has yet to demonstrate that its subject matter exists!" Thirty five years later that statement is still true.

Stanley Miller (1953) is usually given credit for being the first to generate amino acids in a prebiotic atmosphere. However, Walther Löb (1913) and Oskar Baudisch (1913) anticipated him by 40 years. These two workers, both addressing the question of the assimilation of nitrogen to form protein, showed that amino acids are generated in the silent electrical discharge (Löb, 1913) only in a reducing atmosphere, and by UV (Baudisch, 1913), also only in a reducing atmosphere. I have given a review of the views of Darwin, Oparin and Miller together with a criticism that there is no geological evidence that such a primeval soup ever existed (Yockey, 1992, 1995a,b, 1997). Miller et al. (1997) and Lazcano (1997) admit that there is no geological or geochemical evidence that a primeval soup ever existed. Mojzsis et al. (1998) remarked: "However, it is now held highly unlikely that the conditions used in these experiments (silent electrical discharge) could represent those in the Archaen atmosphere. Even so, scientific articles still occasionally appear that report experiments modelled on these conditions and explicity or tacitly claim the presence of resulting products in reactive concentrations 'on the primordial Earth' or in a 'prebiotic soup.' (See for example, Deamer, 1997; Smith, 1998, 1999). The idea of such a 'soup' containing all the desired organic molecules in concentrated form in the ocean has been a misleading concept against which objections were raised early (Sillén, 1965)." This remark belies the title of the book in which it appears.

Exactly by whom and when the modern notion that life originated from colloids or coacervates generated from the organic substances in the early ocean was first expounded in specific form is hard to find. Haeckel (1834–1919) claimed priority. He wrote: "The monistic hypothesis of abiogenesis, or autogeny in the strictly scientific sense of the word, was first formulated by me in 1866 in the second book of the General Morphology." (Today autogeny is called self-organization.) As Haeckel (1866) explained:

> "The chemical processes which first set in at this stage of development must have been catalysis, which led to the formation of albuminous combinations, and eventually of plasm. The earliest organisms to be thus formed can only have been plamodomous Monera, structureless organisms without organs; the first forms in which living matter individualized were probably homogeneous globs of plasm, like certain of the actual chromacea (*chroococcus*). The first cells were developed secondarily from these primitive Monera, by separation of the central caryoplasm (nucleus) and peripheral cytoplasm (cell body)."

Haeckel's views were well-known in the nineteenth and early twentieth centuries for he was widely published in professional books and journals and in best-selling popular books translated from the German to several languages. Haeckel's discussion of the origin of life from a primordial soup in the early ocean was so well known among scientists, theologians and the theater-going public in 1878 that Sir William Schwenck Gilbert (1836–1911) had Pooh-Bah, a comic, greedy and conceited character in The Mikado, introduce himself as follows:

> I am in point of fact, a particularly haughty and exclusive person, of pre-Adamite ancestral descent. You will understand this when I tell you that I can trace my ancestry back to a protoplasmal primordial atomic globule (Gilbert, 1878).

Louis Pasteur (1822–1895) showed that properly sterilized cultures always became infected when exposed to air. Alpine air, which was almost free of germs, seldom produced a growth of organisms. Thoroughly sterilized cultures remained so, for years, in the absence of germs added from without. Life, therefore, must come only from life. Haeckel maintained that Pasteur had only settled the negative in certain circumstances. It being very difficult or impossible to prove a negative, many scientists in the nineteenth century, and many today, support spontaneous origin of life from a primeval chemical soup in the early ocean.

The 'warm little pond' quotation (Darwin, 1898) from Darwin's private correspondence does not reflect Charles Darwin's views on the origin of life. Had he thought the 'warm little pond' idea worthy of publication, he would certainly have done so. Significantly, Darwin avoided the origin of life controversy in the sixth edition (1872) of The Origin of Species:

> It is no valid objection that science as yet throws no light on the far higher problem of the *essence or origin of life*. Who can explain what is the *essence of the attraction of gravity*? No one now objects to following out the results consequent on this unknown element of attraction; not withstanding that Leibnitz formerly accused Newton of introducing 'occult qualities into philosophy' (Chapter XV). (My italics)

Those who speculated about the origin of life proposed various models similar to Haeckel's. Before the discovery of the genetic code, Jacques Loeb in 1906 objected to the proposal that life emerged from colloids through catalysis (Loeb, 1906). Loeb was an expert in the chemistry of colloids and very well-known in the first part of the twentieth century. In his book, The Dynamics of Living Matter, published in 1906, Loeb wrote:

> But we see that plants and animals during their growth continually transform dead into living matter, and that the chemical processes in living matter do not differ in principle from those in dead matter. There is, therefor, no reason to predict that abiogenesis is impossible and I believe that it can only help science if younger investigators realize that experimental biogenesis is the goal of biology. On the other hand, our lectures show clearly that we can only consider the problem of abiogenesis solved when the artificially produced substance is capable of development, growth, and reproduction. It is not sufficient for this purpose to make protein synthetically, or to produce in gelatin or other colloidal material round granules that have an external resemblance to living cells.

Loeb rejected the colloid or Pooh Bah's *protoplasmal primordial atomic globule* model of the origin of life; however, much it may have appealed to his mechanist ideology. He saw that colloids lack the characteristic chemical processes, namely enzymes, by which organisms make sugars, fats, proteins and other molecules essential to their metabolism. In the terms I am discussing, they have no genome to control the formation of these critical compounds. It is a travesty that Loeb's comments are not now mentioned in the literature.

Self-replication is a fundamental requirement in speculations on the origin of life. Self-replication mechanisms, at the origin of life, must be accomplished without the action of enzymes because enzymes themselves are produced by instructions in the genetic message. All schemes for non-enzymatic self-replication must have a skilled chemist acting as a deus ex machina using pure chemicals in a well-equipped chemical laboratory. Attempts to discover non-enzymatic self-replicating chemical systems is a current field of biochemistry. The function of enzymes is to speed up or catalyze chemical reactions. On the biologically significant time scale of seconds to years, some enzymes may catalyze as many as a million reactions per minute, whereas in the absence of the enzyme the reaction may take many years.

Directed Panspermia, the suggestion that life was introduced from outer space, raises its head from time to time. Hermann Ludwig Ferdinand von Helmholtz (1821–1894) and Lord Kelvin, William Thompson (1824–1907), proposed the Directed Panspermia theory in the nineteenth century. Helmholtz and Arrhenius (1908) proposed that life is eternal and that meteors that roam about the solar system might contain germs of organisms that, under favorable conditions, might reach the Earth and other planets. Although Crick (1968) suggested that: "Possibly the first enzyme was an RNA molecule with replicase properties", this proposal is now known as *The RNA World* as coined by Gilbert (1986). The discovery of ribozymes, in which RNA plays both roles, namely, that of recording and transfer of information and the catalysis role of enzymes proved this prophetic (Cech, 1986; Zaug and Cech, 1986). Some thought ribozymes provided the pathway from the primeval soup to the protobiont. It was proposed as a self-replicating system preceding the present system, which did not depend on the enzyme action of proteins but rather on a pre-enzyme action of nucleic acids (Joyce, 1998). This suggestion seems to move the problem a step nearer to the protobiont but still encounters the primary questions of the generation of the genetic message and of the origin and evolution of the genetic code.

Taken at best, however, this proposal simply moved the search for an origin of life scenario one step from the primeval soup, where it encounters the need for a chemical procedure for the prebiotic formation of the sugar ribose, a component of RNA. There is only one plausible synthesis of ribose that may be considered in the prebiotic milieu, namely, the polymerization of formaldehyde. Ribose is only one of a number of sugars and never the primary product (Orgel, 1986). Furthermore, the condensation of adenine or guanine with ribose leads to mixtures of the optical isomers that are very complex. From the point of view of the biochemist, this problem is as difficult as the one the discovery of ribozymes was supposed to solve. Furthermore, non-biological reactions in the prebiotic milieu lead to 2′,5′ isomers (Orgel, 1986). This does not lead in the direction of the origin of life because the reactions that are typical of biochemical products are almost always the 3′,5′ linked isomers, whether or not they are catalyzed by protein–enzymes. RNA is hydrolyzed about 100 times more rapidly than DNA and would not appear in quantity in the primeval soup. Furthermore, this proposal is not as promising as seemed at first, since RNA is not as versatile a catalyst as protein enzymes. Moreover, RNA is itself a molecule more complex than amino acids.

Even those who believe in a primeval soup do not regard it plausible that RNA existed in large quantities in the primeval soup. The nucleotide cytosine is a component of DNA and mRNA. Shapiro (1999) has examined the question of the prebiotic synthesis of cytosine. He found: "No reactions have been described thus far that would produce cytosine, even in a specialized local setting, at a rate sufficient to compensate for its decomposition. On the basis of this evidence, it appears quite unlikely that cytosine played a role in the origin of life." Even more striking is the paper by Levy and Miller (1998). Mojzsis et al. (1998) showed that life must have pullulated on Earth before 3.850 billion years ago, leaving at most 600 million years after the formation of the planet.

10. The minimum information content of the genome

Can we answer the questions posed at the beginning of this paper? First, how large an information content does a genetic message require to be characteristic of life? Second, do the laws of physics and chemistry have enough information content so that non-living matter may organize itself and become living matter, say in the manner that crystals are formed? To determine the minimum information content required for the protobiont, we may draw our estimates from the information content of the genomes of the most primitive free-living organisms. The smallest free-living organisms known today are mycoplasmas and spiroplasmas. Many spiroplasmas contain plasmids that have between 2000 and 50 000 base pairs, each carrying two bits (Bové, 1984) or between 500 and 12 500 bytes of genetic information.

Other estimates can be taken from the genomes of viruses, for example, the genome of the virus ΦX174 has 5375 nucleotides which transmit nine proteins, an information content of 1344 bytes ((Fiddes, 1977). However, viruses are not free living and are dependent on their hosts for replication. For this reason the information content of the independent protobiont must be larger than that of a virus.

One estimate came from John von Neumann (1903–1957), who, near the end of his all too short but very productive life, was interested in a robot that had a control message of sufficient information content to direct its own actions and to self-replicate (von Neumann, 1966). The robot's built-in set of instructions is isomorphic to a genetic message. Von Neumann gauged the complexity of his self-replicating machine by the number of its parts, using as a model the computing machines of his time. He thought the number of parts must be in the millions. That is, the information content for the robot must be in the millions of bytes. These estimates are, of course, highly speculative but they do establish that the minimum information content of the protobiont must be in the range of several tens of thousands to several million bytes. That is a rather modest amount of information to require, in terms of the computer technology of today, yet even this low threshold provides for the generation of an enormous number of messages. As we learned above, the nine-symbol ZIP + 4 postal message can write one billion postal addresses.

Those who believe in the origin of life by chance from a primeval soup on Earth or who believe that life swarms in the universe do not first establish that the origin of life is, indeed, a stochastic or aleatoric event. Many events are not due to chance, for example, the rising of the Sun in the morning is not a chance event. It is therefore ridiculous to speak about the 'probability' that the Sun will rise. As I pointed out above the decimal sequences that express π and e are not chance events even though there is no pattern or repetition in those sequences.

Hoyle and Wickramasinghe (1978) suggest that a protobiont would need 2000 enzymes to sustain itself and to reproduce. Choosing cytochrome c, let us put these proposals to a quantitative test, as Socrates said we should do. I calculated (Yockey, 1992) that there must be 2.3×10^{93} cytochrome c molecules, taking into account that some sites in the protein sequence may be occupied by more than one functionally equivalent amino acid and that all amino acids are not equally probable. The Shannon–McMillan–Breiman theorem of information theory provides the only way of allowing for these factors. This greatly reduces the information content required for the protobiont and therefor favors the mechanist–reductionist speculations of Hoyle and Wickramasinghe.

If each of Hoyle's 2000 enzymes has 374 bits, the same as the information content of cytochrome c, then we can estimate that the information content of the genome of the protobiont amounts to 93 500 bytes. Although this number seems small in terms of computer technology, as noted above, an enormous number of messages may be written with this modest information content.

What is the probability of writing a given message containing 93 500 eight-bit bytes? According to the Shannon–McMillan–Breiman theorem, discussed above, since cytochrome c is a long sequence, each has approximately the same probability $[2.3 \times 10^{93}]^{-1}$, and we are treating the appearance of each of the 2000 enzymes as an independent chance event. Therefore the 'probability' that all 2000 enzymes were formed by chance is the product of the separate 'probabilities': $[2.3 \times 10^{93}]^{-2000} = 10^{-186\,000}$. Thus we see that Hoyle's estimate of a 'probability' of $10^{-44\,000}$ that a protobiont of 2000 enzymes was created by chance is outrageously optimistic.

The protobiont must carry out at least two critical biochemical processes: namely, the fixation of nitrogen and carbon dioxide. Each processes is mediated by an enzyme much longer than cytochrome c; namely, the nitrogenase enzyme, and the assimilation of carbon dioxide by ribulose-1,5-bisphosphate. Nitrogenase consists of two component metalloproteins, the molybdenum protein and the iron protein, which are an α_2 β_2 tetramer. The two α and β subunits have 491 and 522 amino acids respectively, giving nitrogenase a total of 2026 amino acids. The real 'probability' must be very much smaller than $10^{-186\,000}$ because these enzymes are much longer than cytochrome c, which has only 110 amino acids, so that this estimate (of the minimum information content required for the protobiont) is very generous to the mechanist–reductionist notion that genetic messages were generated by chance.

The late Sidney W. Fox (1912–1998) and his disciples (Fox et al., 1994, 1996) have proposed their 'proteinoid microsphere' scenario for the origin of life, a form of colloids. According to the scenario, the proclivities of living organisms, including the appearance of a genetic message and a genetic code evolved from these proteinoid microspheres by unspecified gradualist and uniformitarian processes. Fox insisted that the 'non randomness' due to the self-ordering of the amino acid sequences by 'the information in the amino acids themselves' in his microspheres constitutes the initial critical steps toward molecular evolution. It is well known that in languages such as English, there is a relation between successive letters of the alphabet. This can to some small extent be used to detect and correct errors (Shannon, 1951). Fox is hoist by his own petard since, as I have discussed above, the genetic code does not allow 'information in the amino acids themselves.' Furthermore, a random sequence has the highest degree of complexity, and therefor information content. On the other hand 'non randomness' degrades the amount of information in a sequence.

Fox insisted that life is rooted in protein. As I discussed above, a message cannot be transmitted to a sequence that has an alphabet of larger entropy. Thus the origin of life cannot be based on 'protein first'. This is not mathematically impossible if the primeval genetic code had been doublet and carried 15 or 16 amino acids. I have given further reasons why Fox's 'proteinoids' are not precursors of life (Yockey, 1992).

11. Conclusion

I suggest that Wächtershäuser (1997) is correct in his view that if the historic process of the origin and evolution of the universal ancestor (Darwin, 1872; Woese, 1998) could be followed it would prove to be a purely chemical process. However, much of the historic process of the origin and evolution of life has been lost. Furthermore, following this course involves serious and perhaps insurmountable mathematical and computational problems (Schröder 1870; Edwards and Cavalli-Sforza, 1964; Diaconis and Holmes, 1998). I pointed out above that DNA, mRNA and protein are tapes similar to the classical Turing model of computation. This brings up the question of the decidability (Turing, 1936) of any attempt to follow the numerous protein sequences now available down the path of evolution. It is vitalism to suggest that the genetic logical system is not subject to the same restrictions as the $P/(NP)$ logical problems of decidability (Schulz et al., 1994; Freedman, 1998). Maximum parsimony and maximum likelihood methods attempt to reduce the computational difficulties. Since the evolutionary trace may be regarded as a random walk, such paths are seldom the most parsimonious (Mitchison and Durbin, 1995; Willson, 1998).

The question is whether this historic process or any reasonable part of it is available to human reasoning from axioms or by 'counting, measuring or weighing'. There is no requirement that Nature's laws be plausible or even known to mankind. As Hamlet said to his friend: "There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy." (Hamlet, Act 1, Scene V) With regard to the nature and existence of life, Bohr argued that life is consistent with but undecidable by human reasoning from physics and chemistry. Bohr (1933) made this point in his famous 'Light and Life' lecture:

The recognition of the essential importance of fundamentally atomistic features in the function of living organisms is by no means sufficient, however, for a comprehensive explanation of biological phenomena, before we can reach an understanding of life on the basis of physical experience. Thus, we should doubtless kill an animal if we tried to carry the investigation of its organs so far that we could describe the role played by single atoms in vital functions. In every experiment on living organisms, there must remain an uncertainty as regards the physical conditions to which they are subjected, and the idea suggests itself that the minimal freedom we must allow the organism in this respect is just large enough to permit it, so to say, to hide its ultimate secrets from us. On this view, the existence of life must be considered as an *elementary fact* (or axiom) that cannot be explained, but must be taken as a starting point in biology, in a similar way as the *quantum of action*, which appears as an *irrational element* from the point of view of classical mechanical physics, taken together with the existence of elementary particles, forms the foundation of atomic physics.

Randomness exists even in pure mathematics and therefor the solution of some problems is beyond the ability of human reasoning (Chaitin, 1985, 1987, 1998). The ancient Greek mathematicians were aware that among these problems are: the trisection of a given angle, the squaring of the circle and the doubling of the cube. There can be no doubt that three equal angles exist that compose any given angle but finding them escapes our mathematical reasoning. We must measure the given angle and construct one that is one third as large. But this is an act of physical measurement, not of mathematical reasoning. Measurement always involves error. While Sir Isaac Newton found that he could calculate exactly the motion of one planet about a central Sun attracted by the force of gravity which varies inversely as the square of the distance between them, it is now known that there is no explicit solution to the three-body problem. There are explicit solutions in terms of the coefficients of each term for equations up to the fourth degree but none for the fifth and higher. Computer programmers are aware that there is no general test to determine whether a given computer program will halt, that is, complete its calculation. Nevertheless, the halting property of a given computer program exists whether we can determine it or not. Mathematicians, who thought that all mathematical statements could be either proven or disproved, were astonished by the 'incompleteness' theorem of Kurt Gödel that for any axiom system that is consistent and can be expressed in a computer program there are statements that can be neither proved nor disproved (Smullyan, 1992). That is, they are *undecidable* to mathematicians (Turing, 1936). Thus, although some are optimistic that life may be made in the laboratory (Deamer, 1997), it may well be that scientists, by *counting, measuring or weighing and employing the calculating or reasoning element in the soul* will come closer and closer to the riddle of how life emerged on Earth but, because of the limitations of measuring and human reasoning, like Zeno's Achilles, will never achieve a complete solution.

References

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database searc programs. Nucleic Acid Res. 25, 3389–3402.

Arrhenius, S., 1908. Worlds in the Making. Harper, London.

Baudisch, O., 1913. ber Nitrat-und Nitritassimilation. Z. Angew. Chem. 26, 612.

Billingsley, P., 1965. Ergodic Theory and Information. Wiley, New York [see Theorem 15, in Chapter 5].

Bohr, N., 1933. Light and life. Nature 308, 421–423, 456–459.

Bové, J.M., 1984. Wall-less prokyrotes of plants. Annu. Rev. Phytopathol. 22, 361–396.

Breiman, L., 1957. Ann. Math. Stat. 28, 809–811.

Cech, T.R., 1986. A model for the RNA-catalyzed replication of RNA. Proc. Natl. Acad. Sci. USA 83, 4360–4363.

Chaitin, G.J., 1985. An APL2 gallery of mathematical physics — a course outline. Proceedings Japan 85 APL Symposium Publication N:GE18-9948-0, IBM Japan, pp. 1–56.

Chaitin, G.J., 1987. Algorthmic Information Theory. Cambridge University Press, Cambridge, UK.

Chaitin, G.J., 1998. The Limits of Mathematics. Springer-Verlag, Singapore.

Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. USA 43, 416–421.

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

Crick, F.H.C., 1970. Central Dogma of Molecular Biology. Nature 227, 561–563.

Darwin, C.R., 1872. The Origin Of Species By Means Of Natural Selection Or The Preservation Of Favoured Races In The Struggle For Life — A Mentor Book. New American Library, New York.

Darwin, F., 1898. The Life and Letters of Charles Darwin, vol. II. Appleton, New York, p. 202.

Deamer, D., 1997. The first living systems: a bioenergic perspective. Microbiol. Mol. Biol. Rev. 61, 239–261.

Diaconis, P.W., Holmes, S.P., 1998. Matchings and phylogenetic trees. Proc. Natl. Acad. Sci. USA 95, 14600–14602.

Edwards, A.W.F., Cavalli-Sforza, L.L., 1964. In: Heywood, V.H., McNeil, J. Phenetic and Phylogenetic Classification. Symatics Association, London, Publication No. 6, pp. 67–76.

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

Eigen, M., 1992. Steps Toward Life. Oxford University Press, Oxford, UK.

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

Engels, F., 1954. Dialectics of Nature. Foreign Language Publishing House, Moscow.

Feller, W., 1957. An Introduction to Probability Theory and its Applications. Wiley, New York.

Fiddes, J.C., 1977. The nucleotide sequences of a viral DNA. Sci. Am. 237, 55–67.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Oxford University Press, Oxford, UK.

Fox, S.W., et al., 1994. Experimantal retracement of the origins of a protocell: it was also a protoneuron. J. Biol. Phys. 20, 17–36.

Fox, S.W., et al., 1996. Experimental retracement of terrestrial origin of an excitable cell: was it predictable? In: Chela-

Flores, J., Raulin, F. (Eds.), Chemical Evolution. Kluwer, The Netherlands.

Freedman, M.H., 1998. Limit, logic and computation. Proc. Natl. Acad. Sci. USA 95, 95–97.

Freist, W., et al., 1998. Accuracy of protein biosynthesis: quasi-species nature of protein and possibility of error catastrophes. J. Theor. Biol. 193, 19–38.

Gilbert, W.S. The Mikado, Act 1 (first performed May 28, 1878).

Gilbert, W., 1986. Nature 319, 618.

Gilbert, W., 1987. The exon theory of genes. Cold Spring Harbor Symp. Quant. Biol. 52, 901–905.

Gilbert, W., de Souza, S.J., Long, M., 1997. Origin of genes. Proc. Natl. Acad. Sci. USA 94, 7698–7703.

Haeckel, E.H., 1866. Entstehung der ersten Organismen. In: Generelle Morphologie der Organismen V.I. George Reimer, Berlin, pp. 167–190.

Hamming, R.W., 1986. Coding and Information Theory. Prentice-Hall, Englewood Cliffs, NJ.

Hawkes, W.C., Tappel, A.L., 1983. In vitro synthesis of glutothione peroxidase from selenite translational incorporation of selenocysteine. Biochem. Biophys. Acta 739, 225–234.

Holliday, R., 1986. Genes, Proteins, and Cellular Aging. Van Nostrand Reinhold, New York.

Hoyle, F., Wickramasinghe, N.C., 1978. Life Cloud: The Origin of Life in the Universe. Harper & Row, New York.

Huynen, M.A., Bork, P., 1998. Measuring genome evolution. Proc. Natl. Acad. Sci. USA 95, 5849–5856.

Jenkin, H.C.F., 1867. The origin of species. North Br. Rev. 46, 277–318 (This paper may be found in 'Hull, D.L., 1973. Darwin and his Critics. Harvard University Press, Cambridge, MA.').

Joyce, G.F., 1998. Nucleic acid enzymes: playing with a fuller deck. Proc. Natl. Acad. Sci. USA 95, 5845–5847.

Jukes, T.H., 1966. Molecules and Evolution. Columbia University Press, New York.

Jukes, T.H., 1983. Evolution of the amino acid code: inferences from mitochondrial codes. J. Mol. Evol. 19, 219–225.

Jukes, T.H., 1993. The genetic code function and evolution. Cell. Mol. Biol. Res. 39, 685–688.

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

Lazcano, A., 1997. Chemical evolution and the primitive soup: did Oparin get it all right? J. Theor. Biol. 184, 219–223.

Leinfelder, W., et al., 1988. Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine. Nature 331, 723–725.

Levitt, M., Gerstein, M., 1998. A unified statistical framework for sequence comparison. Proc. Natl. Acad. Sci. USA 95, 5913–5920.

Levy, M., Miller, S.L., 1998. The stability of the RNA bases: implications for the origin of life. Proc. Natl. Acad. Sci. USA 95, 7933–7938.

Löb, W., 1913. Über das Verhalten des Formids unter der Wirkung der stillen Entladung: Ein Beitrag zur Frage der Stickstoff-Assimilation. Bericht 46, 684–697.

Loeb, J., 1906. The Dynamics of Living Matter. Macmillan, London.

Maeshiro, T., Kimura, M., 1998. The role robustness and changebility on the origin and evoluton of genetic codes. Proc. Natl. Acad. Sci. USA 95, 5088–5093.

Mayr, E., 1982. The Growth of Biological Thought. The Belknap Press of Harvard University Press, Cambridge, MA, p. 114.

McMillan, B., 1953. Ann. Math. Stat. 24, 196–219.

Mendel, G.J., 1865. Versuche über Pflantzen-Hybriden. Verh Naturf Ver Brünn, vol. X.

Miller, S.L., 1953. Production of amino acids under possible primitive Earth conditions. Science 117, 528–529.

Miller, S.L., Schopf, J.W., Lazcano, A., 1997. Oparin's 'Origin of Life': sixty years later. J. Mol. Evol. 44, 351–353.

Mitchison, G.F., Durbin, R., 1995. Tree-based maximal likelihood substitution matrices and hidden Markov models. J. Mol. Evol. 41, 1139–1151.

Mizutani, T., Hitaka, T., 1988. The conversion of phosphoserine residues to selenocysteine on an opal suppressent tRNA and casein. FEBS Lett. 232, 243–248.

Mojzsis, S.J., Kishnamurthy, Arrhenius, G., 1998. Before RNA and after: geological and geochemical constraints on molecular evolution 1–47. In: Gesteland, R.F. (Ed.), The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA, second ed. Cold Spring Harbor Laboratory Press, Boca Raton, FL.

Olson, M.V., 1995. A time to sequence. Science 270, 394–396.

Oparin, A.I., 1957. The Origin of Life on Earth, third revised ed. (translated by Synge, A.) Oliver & Boyd, London.

Orgel, L.E., 1963. The maintence of accuracy of protein synthesis and its relevance to ageing. Proc. Natl. Acad. Sci. USA 49, 517–521.

Orgel, L.E., 1970. The maintence of accuracy of protein synthesis and its relevance to ageing: a correction. Proc. Natl. Acad. Sci. USA 67, 1476.

Orgel, L.E., 1986. RNA catalysis and the origin of life. J. Theor. Biol. 123, 127–149.

Ornstein, D.S., 1974. Ergodic Theory, Randomness, and Dynamical Systems, Yale Mathematical Monographs 5. Yale University Press, New Haven, CT.

Osawa, S., 1995. Evolution of the Genetic Code. Oxford University Press, Oxford, UK.

Osawa, S., et al., 1990. Evolutionary changes in the genetic code. Proc. R. Lond. B 241, 19–28.

Osawa, S., et al., 1992. Recent evidence for evolution of the genetic code. Microbiol. Rev. 56, 229–264.

Pincus, S., Kalman, R.E., 1997. Not all (possibly) random sequences are created equal. Proc. Natl. Acad. Sci. USA 94, 32513–32518.

Pincus, S., Singer, B.H., 1996. Randomness and degrees of irregularity. Proc. Natl. Acad. Sci USA 93, 2083–2088.

Schröder, E., 1870. Z. Math. Phys. 15, 361–376.

Schulz, A.S., Shmoys, D.B., Williamson, D.P., 1994. Approximation algorithms. Proc. Natl. Acad. Sci. USA 94, 12734–12735.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–424, 623–656.

Shannon, C.E., 1951. Prediction and entropy of printed English. Bell Syst. Tech. J. 30, 50–64.

Shapiro, R., 1999. Prebiotic cytosine synthesis: a critical analysis and implications for the origin of life. Proc. Natl. Acad. Sci. USA 96, 4396–4401.

Shields, P.C., 1973. The Theory of Bernoulli Shifts. University of Chicago Press, Chicago, IL.

Sillén, L., 1965. Oxidation state of the Earth's ocean and atmosphere. IA model calculation on earlier states. The myth of the probiotic soup. Ark. Kemi. 24, 431–456.

Smith, J.V., 1998. Biochemical evolution. I. Polymerization on integral, organophilic silica surfaces of dealuminated zeolites and feldspars. Proc. Natl. Acad. Sci. USA 95, 3370–3375.

Smith, J.V., 1999. Biochemcal evolution III: polymerization on organophilic silica-rich surfaces, crystal–chemical modeling, formation of first cells, and geological clues. Proc. Natl. Acad. Sci. USA 96, 3479–3485.

Smullyan, R.M., 1992. Gödel's Incompleteness Theorems. Oxford University Press, Oxford, UK.

Sonneborn, T.M., 1965. Degeneracy of the genetic code: extent, nature and genetic implications. In: Bryson, V., Henry, J. (Eds.), Evolving Genes and Proteins. Academic Press, New York.

Sunde, R.A., Evenson, J.K., 1987. Serine incorporation into selenocysteine moity of gultathione peroxidase. J. Biol. Chem. 262, 933–937.

Tononi, G., Sporns, O., Edelman, G.M., 1996. A complexity measure for selective matcing of signals by the brain. Proc. Natl. Acad. Sci. USA 93, 3227–3244.

Tononi, G., Sporns, O., Edelman, G.M., 1999. Measures of degeneracy and redundancy in biological networks. Proc. Natl. Acad. Sci. USA 96, 3257–3262.

Turing, A.M., 1936. On computable numbers, with an application to the Entscheidungs problem. Proc. Lond. Math. Soc. Ser. 2 42, 230–265 (a correction 43, 544–546).

von Neumann, J., 1966. Theory of Self-replicating Automata. University of Illinois Press, London, UK.

Wächtershäuser, G., 1997. The origin of life and its methodological challenge. J. Theor. Biol. 187, 483–494.

Willson, S.J., 1998. Measuring inconsistency in phylogenic trees. J. Theor. Biol. 190, 15–36.

Woese, C., 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA 95, 6854–6959.

Wolynes, P.G., 1998. Computational biomolecular science. Proc. Natl. Acad. Sci. USA 95, 5848.

Yockey, H.P., 1974. An application of information theory to the Central Dogma and the sequence hypothesis. J. Theor. Biol. 46, 369–406.

Yockey, H.P., 1990. When is random random? Nature 344, 823.

Yockey, H.P., 1992. Information Theory and Molecular Biology. Cambridge University Press, Cambridge, UK.

Yockey, H.P., 1995a. Information in bits and bytes. BioEssays 17, 85–87.

Yockey, H.P., 1995b. Comments on Let There be Life: thermodynamic reflections on biogenesis and evolution by Avshalom C. Elizur. J. Theor. Biol. 176, 349–355.

Yockey, H.P., 1997. Walther Löb, Stanley L. Miller, and prebiotic 'building blocks' in the silent electrical discharge. Perspect. Biol. Med. 41, 125–131.

Zaug, A., Cech, T.R., 1986. The intervening sequence of RNA of Tetrahymena is an enzyme. Science 231, 470–475.