Critical Focus

An Information-Theoretic Analysis of the Genetic Code

Robert D. Pedersen*

Dallas, TX

Abstract

This paper focuses on Shannon information theory entropy analysis of the genetic code structure and implications of that analysis for natural genetic code origination theories. Results are derived for genetic code actual block code structure, for human Genbank data, and for protein amino acid variable lengths from 100 to 2000 amino acids. Strong correlation between the genetic code structure and human GenBank probabilities are demonstrated. Upper bounds for genetic code occurrence probability for variable protein lengths are derived based on the genetic code structure and considering the requirements for translation and the probabilistic resources of the universe. The resulting upper bound occurrence probability for an average length protein of 430 amino acids is found to be $p \approx 10^{-550}$. It has been estimated that there are about 10^{80} elementary particles in the known universe. The probability of random selection of a particular elementary particle (10^{-80}) is therefore much larger than the calculated probabilities of natural occurrence of the genetic code. The various naturalistic accounts of the origin of the genetic code, including chemical evolution, protein first, multiple protein, RNA World, multiverse, Darwinian natural selection, and accidental freeze theories, all face severe probabilistic challenges.

Cite as: Pedersen (2026) An Information-Theoretic Analysis of the Genetic Code. BIO-Complexity 2026 (5):1–12. **doi:**10.5048/BIO-C.2026.5.

Editor: Robert J. Marks II

Received: January 28, 2026; **Accepted:** July 6, 2026; **Published:** September 9, 2026

Copyright: © 2026 Pedersen. This open-access article is published under the terms of the Creative Commons Attribution License, which permits free distribution and reuse in derivative works provided the original author(s) and source are credited.

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2026.5.c.

*Email: rdptek@rpedersen.com

INTRODUCTION

The genetic code in living cells is remarkable [1– 8]. The individual code elements, called codons, specify each of the twenty different amino acids. Each codon consists of a sequence of three nucleotides selected from a set of four (designated A, C, G, and T) resulting in $4^3 = 64$ possible codons. Because the number of codons is greater than the number of amino acids, most amino acids are specified by multiple codons.

This manuscript analyzes the Genetic code information with mathematical modeling based on Shannon information theory entropy, Bayesian conditional probability theory, binomial theorem modeling, Pearson (Correl) correlation, genetic code structure, and Genbank data. Probabilities of occurrence formulas and computed results are presented based on two variables: the DNA Shannon code entropy and code protein length. Results include predicted probabilities over all time since the big-bang and in all of space in the entire known universe.

The probabilities of the occurrence for correct translation of gene sequences are important in research directed to further understanding of the origin of the DNA code. Hundreds of papers and books have been written on origin-of-life theories. It is generally conceded that after decades of work, none of the theories developed to date conclusively resolve origination questions. This paper addresses the origin probability of the genetic code and code error protection operations used in cellular gene expression. The arguments and theses of the paper relate to origin-of-life arguments to the extent that life requires DNA and the genetic code.

GENETIC CODE DESCRIPTION

Scientific research has demonstrated that the genetic code comprises a digital program that provides the genetic information required for the synthesis of functional proteins.

> "A drastic change in the life sciences was brought about by the discovery of the double helical structure of DNA by James Watson and Francis Crick in 1953, eventually leading to the deciphering of the genetic code. The elucidation of the genetic code was one of the greatest discoveries of the 20th century. The genetic code is an algorithm that connects 64 DNA triplets to 20 amino acids, and functions as the Rosetta stone of molecular biology" [6].

"In 1961, Francis Crick, Sydney Brenner, Leslie Barnett, and Richard Watts-Tobin first demonstrated the three bases of DNA code for one amino acid. That was the moment that scientists cracked the code of life." [6].

Genetic Code Overview

The genetic is shown in Table 1 below including individual names and accepted single letter identification for each of the 20 amino acids in column 1. Column 2 lists specific codons assigned to each of the 20 amino acids. Column 3 gives the number of codons used for each amino acid. Columns 4 and 5 give computed coded frequencies and NIH Human Genome database frequencies, respectively [2, 8, 9].

Within the cell, gene sequences are transcribed to form mRNA (messenger RNA) molecules. The mRNAs are transferred to the ribosome for translation/decoding and selection of amino acids used to assemble proteins. Genes are delineated by start and stop codons. The individual sections for each protein form a robust code with redundant instructions. Origination accounts must account not only for the formation of the Genetic code but also for the existence of the complex transcription and translation systems [1, 7, 8]

Each of sixty-one codons are assigned to an amino acid. TGA, TAA and TAG are end or stop codons for a protein sequence. All protein codon sequences start with ATG codon, which may also appear elsewhere in the protein amino acid sequence. The code itself provides a degree of redundancy with multiple codons specifying the same amino acid. The choice of base-four coding provides an efficient structure for identifying the 20 amino acids while minimizing the consequences of many possible codon errors.

Nonrandom Structure of the Genetic Code

Multiple references exist exploring the merits of the Genetic code, including the code representation, quantification, and nature of genetic information. Previous papers have reported investigations of its information content [1 – 12, 25 – 29, 34, 35]. The choices of three nucleotides to represent each codon, the use of 64 codons to represent 20 amino acids, the start and stop control codes and their operations, and the assignment of nucleotides to amino acids all speak to a nonrandom code structure. All 64 codons are used in the code of Table 1. As a result, with 64 possible codons and 3 nucleotides per codon, there are 64 x 3 = 192 required assignments of individual

Table 1. Summary of Redundant Codon Coding and Amino Acid Probabilities (Standard Genetic Code). Start Codon ATG, which may also be a regular codon, is denoted with an asterisk (*).

Amino Acid	Codon to Amino Acid	No. Codons	Codon Coded "C" $p(i)$	Human Genbank "G" $p(i)$
1 - Alanine (A)	GCA, GCC, GCT, GCG	4	0.0656	0.0704
2 - Arginine (R)	AGA, AGG, CGA, CGC, CGT, CGG	6	0.0984	0.0552
3 - Asaragine (N)	AAC, AAT	2	0.0328	0.0368
4 - Aspartic acid (D)	GAC,GAT	2	0.0328	0.0484
5 - Cysteine (C)	TGC, TGT	2	0.0328	0.0231
6 - Glutamic acid (E)	CAA, CAG	2	0.0328	0.0692
7 - Glutamine (Q)	GAA, GAG	2	0.0328	0.0465
8 - Glycine (G)	GGA, GGC, GGT, GGG	4	0.0656	0.0675
9 - Histidine (H)	CAC, CAT	2	0.0328	0.0256
10 - Isoleucine (I)	ATA, ATC, ATT	3	0.0492	0.0450
11 - Leucine (L)	CTA, CTC, CTT, CTG, TTA, TTG	6	0.0984	0.0984
12 - Lysine (K)	AAA, AAG	2	0.0328	0.0565
13 - Methionine (M)	ATG*	1	0.0164	0.0237
14 - Phenylalanine (F)	TTC, TTT	2	0.0328	0.0378
15 - Proline (P)	CCA, CCC, CCT, CCG	4	0.0656	0.0610
16 - Serine (S)	AGC, AGT, TCA, TCC, TCT,TCG	6	0.0984	0.0799
17 - Threonine (T)	ACA, ACC, ACT, ACG	4	0.0656	0.0534
18 - Tryptohan (W)	TGG	1	0.0164	0.0121
19 - Tyosine (Y)	TAC, TAT	2	0.0328	0.0282
20 - Valine (V)	GTA, GTC, GTT, GTG	4	0.0656	0.0613
Stop	TAA, TAG, TGA	3		
Totals		64	1.0	1.0

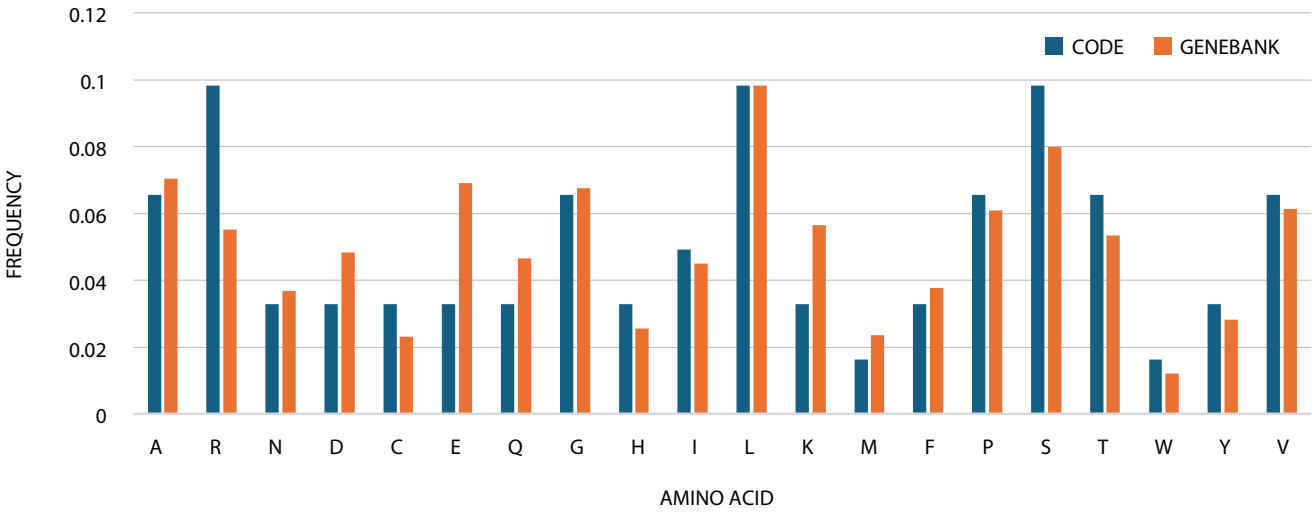

Figure 1. Occurrence frequency correlation between genetic code and GenBank. **doi:**10.5048/BIO-C.2026.5.f1

nucleotides to corresponding codons. Examination of the code of Table 1 reveals that each of the 4 nucleotides (A,C,T, and G) is assigned to exactly 48 codons as would be required in a code with exhaustive use of code symbols: 48 x 4 = 192. The use of nucleotides in the formation of the coded information of Table 1 comprises a structured, not random, code.

Note that the use of the chemical symbols A, C, T, and G could be replaced with the numerals 0, 1, 2, and 3 to form a numerical base-four codon representation containing the same information of Table 1. All 64 numerical base-four values from 000, 001, 002, to 333 would be present in a modified Table 1, column 2. For example, 000 would correspond to AAA (lysine), 221 to TTC (phenylalanine), 312 to GCT (alanine), and 333 to GGG (glycine). The DNA code may be considered a digital program comprising numerical base-four encoding of information to identify nucleotides.

Column 4 provides a calculated probability for each codon based on the ratios of the number of codons of column 4 divided by 61, the number of different codons available for protein synthesis. Column 5 of Table 1 lists the Human Genbank probability of appearance of each amino acid in DNA protein codes [2].

Information from columns 4 and 5 of Table 1 is plotted in Figure 1 below, comparing frequencies of occurrence from column 4 to those of column 5. Close correlation of the code and Genbank frequencies is apparent from the figure. Pearson (Correl) correlation calculations for Table 1 columns 4 and 5 yield a correlation of 0.79 for all 20 amino acids. Perfect correlation is 1.0. The base-four coding of Table 1, column 2, provides more codons for amino acids having the greatest probability of occurrence. The assignments of codons to amino acids are clearly not random.

The central dogma of molecular biology asserts that genetic communication information within the cell occurs from DNA to mRNA to ribosomes for amino acid selection and protein synthesis. Multiple recent references have questioned the completeness of the central dogma [30, 32]. Regardless, the central dogma presents an obstacle to "protein first" theories of DNA origination. Protein molecules perform various functions in living organisms. But, according to the central dogma, transfer of genetic information to DNA is not one of those functions. The DNA code itself does not originate from proteins [5]. The codon/amino acid relationships of Table 1 are not specified in the final protein amino acid sequence. This lack of coding information assumes and reflects the central dogma [2, 3, 5, 30, 32].

Each type of protein has a unique sequence of amino acids, the same from one molecule to the next. Many thousands of different proteins are known, each with its own particular amino acid sequence. As explained by B. Alberts, et. al., "Proteins come in a wide variety of shapes, with up to two thousand or more amino acids. Only about 1300 nucleotide pairs are required to encode a protein of average size (about 430 amino acids in humans)" [1]. More recently, based on protein size, AlphaFold2 predicts protein structures with the highest confidence for proteins with about 300 – 500 amino acids [38].

DNA Error Protection Analysis

The Genetic code is a "block code" as defined in communication and computer technology. A block code is a code that uses sequences of n symbols or n-tuples. The code blocks or code words are as shown in Table 1, column 2. In the DNA code, n = 3 corresponding to the 3-symbol codons. The coding distance or "hamming distance" between code words of Table 1 with two of the same symbols is one. Block code "error protection" refers to a code wherein particular errors do not matter – the decoded code word remains the same regardless of the error [3, 38, 39].

Table 2 below summarizes DNA codon block code single error protection capabilities. Columns 1, 2, 3, and 4 are the

same as Table 1 above. Column 5 summarizes error protection for the coding of column 2.

Many common amino acids are specified by the first two codon nucleotides (making the third nucleotide redundant). These amino acids are therefore specified by any of four codons. This applies to five frequently occurring amino acids: alanine, glycine, proline, threonine, and valine. For each of these, a change in the third nucleotide will result in another codon that specifies the same amino acid.

Three additional amino acids, leucine, arginine, and serine use a group are specified by six codons using a similar error-protection principle. Two codons with the same nucleotides in positions 1 and 2 are assigned to nine different amino acids. One amino acid uses three codons with the same position 1 and 2 nucleotides. Two amino acids use one codon. Maintaining this pattern of grouping codons with the same first two nucleotides simplifies codon translation operations. Only one or two amino acid options exists for each pair of nucleotides at positions 1 and 2.

As shown in Table 2, with this coded redundancy, error protection is provided for many possible nucleotide codon errors. With four codons (GCA, GCC, GCT and GCG) each of the third positions may take on 3 different nucleotides without changing the encoded amino acid. There are 12, (4 x 3 = 12), error protected possible changes in the four codons for these five amino acids.

Nucleotide positions with multiple possible nucleotides yielding the same amino acid are sometimes referred to as "wobble" positions [39, 40, 41]. Careful examination of the code of Table 1 column 2 reveals 130 such error protection code configurations. Codons for the eight more frequently used amino acids, $p(i)$, as indicated in column 4 are assigned to the strongest error protection codon codes. The error protection property is found for 18 of the 20 amino acids, excluding those specified by one codon. This error protection capability is based on codon assignments to amino acids with minimal coding distance between codon code elements for each individual amino acid. The particular coding assignment is selected from all codon coding possibilities of the 64 codons to the 20 amino acids. The use of codon code error protection for 62 of the 64 codon amino acid code words further validates the selection of the codon code word structure of Tables 1 and 2 for the probability of occurrence values, $p(i)$, as used in the calculations below. Beyond such error protection observations, repair

Table 2. Summary of Codon Coding error protection. Start Codon ATG, which may also be a regular codon, is denoted with an asterisk (*).

Amino Acid	Codon to Amino Acid	No. Codons	Codon Coded "C" $p(i)$	Error Protection Capabilities
1 - Alanine (A)	GCA, GCC, GCT, GCG	4	0.0656	12
2 - Arginine (R)	AGA, AGG, CGA, CGC, CGT, CGG	6	0.0984	14
3 - Asaragine (N)	AAC, AAT	2	0.0328	2
4 - Aspartic acid (D)	GAC,GAT	2	0.0328	2
5 - Cysteine (C)	TGC, TGT	2	0.0328	2
6 - Glutamic acid (E)	CAA, CAG	2	0.0328	2
7 - Glutamine (Q)	GAA, GAG	2	0.0328	2
8 - Glycine (G)	GGA, GGC, GGT, GGG	4	0.0656	12
9 - Histidine (H)	CAC, CAT	2	0.0328	2
10 - Isoleucine (I)	ATA, ATC, ATT	3	0.0492	6
11 - Leucine (L)	CTA, CTC, CTT, CTG, TTA, TTG	6	0.0984	14
12 - Lysine (K)	AAA, AAG	2	0.0328	2
13 - Methionine (M)	ATG*	1	0.0164	0
14 - Phenylalanine (F)	TTC, TTT	2	0.0328	2
15 - Proline (P)	CCA, CCC, CCT, CCG	4	0.0656	12
16 - Serine (S)	AGC, AGT, TCA, TCC, TCT,TCG	6	0.0984	14
17 - Threonine (T)	ACA, ACC, ACT, ACG	4	0.0656	12
18 - Tryptohan (W)	TGG	1	0.0164	0
19 - Tyosine (Y)	TAC, TAT	2	0.0328	2
20 - Valine (V)	GTA, GTC, GTT, GTG	4	0.0656	12
Stop	TAA, TAG, TGA	3		4
Totals		64	1.0	130

mechanisms including proof reading, error correction, and replacement of incorrect bases to protect against DNA errors are part of cellular DNA processing operations [3, 4].

SHANNON INFORMATION THEORY PROBABILITIES OF GENETIC CODE FORMATION AND TRANSLATION/DECODING

The following analyses present derivations of formulas and calculation results for probabilities of Genetic code formation and translation/decoding with multiple codon selection assumptions. Results are based on computed DNA protein code entropy as defined in Shannon information theory. The results of these calculations are summarized in Table 3 below presented as power of 10 exponents.

Column 1 indicates a range of protein lengths L varying between 100 to 2000 amino acids. The probabilities $p(L)$ of amino acid sequences of length L delimited by codon start and stop codes are indicated in column 2. The probabilities of columns 3-9 of Table 3 correspond to occurrence of respective protein codon codes selected from all codes with the $p(i)$ distributions of Table 1 columns 4 and 5

$p(G/L)$ is the conditional probability of formation of a particular coded codon sequence with a protein of length L with proper start and stop codes and calculation based on Shannon information theory code entropy and the human Genbank probabilities of Table 1 column 5 above. $p(G,L)$ is the joint probability of occurrence with the codon coded probabilities and the sequence length of L. $p(C/L)$ and $p(C,L)$ are corresponding probabilities with codon coded frequencies of Table 1, column 4. Joint probabilities of $p(C,L)$ including calculations with translation/decoding requirements $p(C,L,D)$, in all time since the big bang $p(C,L,D,\mathrm{T})$, and with all particles in the entire known university $p(C,L,D,T,U)$, are presented in Table 3, columns 6-9. Corresponding results for $p(G,L,D)$, $p(G,L,D,\mathrm{T})$, and $p(G,L,D,T,U)$ would be essentially the same.

The Genetic code of Table 1, column 2 is a generic, virtually universal code for all life forms, including humans [11]. Note that even if a billion code variations would yield the same protein sequence, the exponent values of columns 3-9 would only be increased by an inconsequential numerical value of 9. The conclusions presented above would not be changed.

Formation Probability, $p(L)$, of Codon Sequences of Length L with Start/Stop Codes

Consider the coding of individual sections of DNA specifying the amino acids for protein synthesis as shown in Table 1. All protein codon sequences start with ATG codon – which may also appear elsewhere in the protein codon sequence. TGA, TAA and TAG are end or stop codons for a protein sequence. For example, for an average 430 amino acid protein code, there will be 430 codons with one start code followed by 430 codons without a stop code and then a stop code. The start code may repeat before the occurrence of the stop code. There is a total of $4^3 = 64$ possible codons.

The probability of a start codon is 1/64. The probability of a stop codon is 3/64. The probability of L sequential codons without a stop code is $(61/64)^L$.

Let $p(L)$ designate the probability of formation of any codon sequence of length L with selection of individual codons from the set of 64 codons ($4^3 = 64$) and with DNA start and stop codes as shown in Table 1. $p(L)$ may be calculated as:

$$p(L) = (1/64) \times (61/64)^L \times (3/64). \quad (1)$$

Converting to base 10:

$$\log_{10} p(L) = \log_{10} (3/4096) + \log_{10}(0.9531)^L.$$

Then,

$$p(L) \approx 10^{-3.135 - (0.021)L}. \quad (2)$$

For example, as shown in Table 3, with 430 codons, $p(L)$ will be about $10^{-12.165}$ or 000….01 with about 12 zeros before the numeral one, or about one chance in one trillion.

Probability of Protein Synthesis Sequence with Shannon Entropy and Start/Stop Codes with L Amino Acid Sequence

It is shown in the text by Yockey [3] that a formulation based on an information theory Shannon entropy model provides probabilities of amino acid sequence formations based on probabilities of individual amino acids. Yockey contrasts this

Table 3. Summary of Shannon Codon Sequence Formation Probabilities – Power of 10 Exponents.

L	*p(L)*	*p(G/L)*	*p(G,L)*	*p(C/L)*	*p(C,L)*	*p(C,L,D)*	*p(C,L,D,T)*	*p(C,L,D,T,U)*
100	-5.235	-126.028	-131.263	-124.584	-129.819	-213.019	-195.374	-115.374
150	-6.285	-189.041	-195.326	-186.876	-193.161	-276.361	-258.716	-178.716
300	-9.435	-378.083	-387.518	-373.752	-383.187	-466.387	-448.742	-368.742
430	-12.165	-541.919	-554.084	-535.711	-547.876	-631.076	-613.431	-533.431
500	-13.635	-630.138	-643.773	-622.92	-636.555	-719.755	-702.11	-622.11
1000	-24.135	-1260.276	-1284.41	-1245.84	-1269.98	-1353.175	-1335.53	-1255.53
1500	-34.635	-1890.414	-1925.05	-1868.76	-1903.4	-1986.595	-1968.595	-1888.95
2000	-45.135	-2520.552	-2565.69	-2491.68	-2536.82	-2620.015	-2602.37	-2522.37

Shannon entropy model with the assumption of models with twenty equally likely amino acids in a protein sequence with L amino acids.

Entropy is a measure of the uncertainty in a sequence of events. A particular sequence selected from a completely random sequence of events has more uncertainty than a sequence selected from a sequence with known structure or other properties. More uncertainty results in higher entropy. A totally random sequence has the highest entropy resulting in a lower occurrence probability compared to other selected sequences. For example, for an average DNA sequence of 430 amino acids, the probability of occurrence with total random occurrence will be $(1/20)^{430} = (10^{-1.3})^{430} = 10^{-559}$. This compares to a value of $10^{-547.876}$ for the nonrandom coded sequence Shannon entropy calculation of Table 3. While still essentially zero, the coded sequence is more probable by a factor of about 10^{11}, or one-hundred billion times more probable.

The Yockey formulation may be used to calculate the probabilities of formations of the codon code of Table 1 based on the frequencies of occurrence of the coded codon sequencies of columns 4 and 5. See Ref. [3]. Following Yockey, if the probability of a particular codon coding occurrence with a sequence of length L is $p(i)$, then the sequence will contain an average of $Lp(i)$ occurrences of that codon coding. The probability of the occurrence of a particular codon coding follows as $p(i)^{Lp(i)}$. $p(i)$ corresponds to the occurrence of the assigned codons to a particular amino acid. For example, for Alanine of Table 1, "i" corresponds to the event of occurrence of one of the codons (GCA), (GCC), (GCT), or (GCG).

With independence of individual codon coding probabilities, the probability, P, of occurrence of a particular sequence of length L will be given by the product, $\prod$, of the individual probabilities over the relevant space as follows:

$$P = \prod_{[\text{therelevantspace}]} p(i)^{Lp(i)}. \tag{3}$$

As Yockey points out, using the logarithm of each side of equation (3) changes multiplication into addition. Using base 2 logarithms again with summation over the relevant space , it follows from Yockey that:

$$\log_2 P = L \sum_{[\text{therelevantspace}]} \{p(i) \times \log_2 p(i)\} = -LH \tag{4}$$

where

$$H = -\sum_{[\text{therelevantspace}]} \{p(i) \times \log_2 p(i)\}. \tag{5}$$

H is the Shannon entropy of the sequence of events, or the expected value of the information in the possible codon coding events summed over the number of amino acids. It follows that $P = 2^{-LH}$. The average number of sequences, N, of length L with the specified probabilities follows as:

$$N = 1/P = 2^{LH}. \tag{6}$$

The same calculation may be made using base 10 logarithms. Consider the above Shannon model with $p(i)$ corresponding to the codon coded frequencies from Table 1, column 4. Define the conditional probability, $p(C/L)$, of a given sequence of length L with coded codon frequencies. Based on the above Yockey analysis with base 10 algorithms and Bayesian conditional probability:

$$p(C/L) = 10^{-LH} \tag{7}$$

where the entropy H is computed with base 10 algorithms as follows:

$$H = -\Sigma\{p(i) \times \log_{10} p(i)\}. \tag{8}$$

Let $p(C,L)$ denote the joint probability computed with the above Shannon probability model of forming a codon coded amino acid protein sequence of length L with proper start and stop codes. Note that $p(L/C) = 1$. Based on conditional probability:

$$\begin{aligned} p(C,L) &= p(L) \times p(C/L) = p(L) \times 10^{-LH} \\ &= \{10^{-3.135 - 0.021L}\} \times 10^{-LH} \\ &= 10^{-3.135 - (H+0.021)L}. \end{aligned} \tag{9}$$

where the entropy H is computed with $p(i)$ from Table 1 column 4 for coded codon frequencies and base 10 logarithms. See Ref. [10].

$p(G/L)$ and $p(G,L)$ can be calculated in the same way using the Human Genbank frequencies of Column 5 of Table 1. As above, note that $p(L/G) = 1$.

The conversion from base 2 to base 10 can also be made using the standard base conversion formula: $\log_b(a) = \log_d(a)/\log_d(b)$. In the present case, a=P, b=10 and d=2. Then $\log_{10} P = \log_2 P/\log_2 10 = \log_2 P/3.322$. Of course, the numerical value P is the same.

Computed results for $p(G/L)$, $p(G,L)$, $p(C/L)$, and $p(C,L)$, and are presented in columns 3-6 of Table 3. p(C,L) is based on the actual digital codon code of Table 1 and is used in the following analyses. For example, for L=430, $p(C,L) = 10^{-547.876}$, or 000….01 with about 548 zeros before the numeral one. Similar results can be derived passed on p(G,L).

DNA Code Translation/Decoding Requirements

Consider a genetic code origination model with 20 amino acids and wherein the probability of selecting a particular amino acid for a particular codon is $1/20 \approx 10^{-1.3}$. With this assumption, the probability of selection of all 64-codon translation/decoding rules follows as:

$$p(D) = (10^{-1.3})^{64} \approx 10^{-83.2}. \tag{10}$$

Let $p(C,L,D)$ denote the joint probability of formation of the correct translation/decoding for each of all 64 codons of the code corresponding to particular amino acids or codon codes. With independence of the DNA code formation and the ribosome translation/decoding operation formations, the total probability will be the product of $10^{-83.2}$ with each of the probability values corresponding to the exponent values of Table 3.

$$\begin{aligned} p(C,L,D) &= p(C,L)^* p(D) \\ &= \{10^{-3.135 - (H+0.021)L}\}^* 10^{-83.2} \\ &= 10^{-86.34-(H+0.021)L}. \end{aligned} \tag{11}$$

Tabulated calculation results based on equation (11) above are presented in Table 3 for values of protein length from 100 to 2000 codons. For example, for L = 430, the probability of correct translation/decoding of codons based on the codon coded "C" amino acid frequencies is approximately 10^{-631}, or

0.000…001, a decimal point followed by about 631 zeroes before the numeral one. Note that $p(C,L,D)$ is less than $p(C,L)$ because of the further requirement for proper DNA code translation/decoding.

Clearly the results are an even further reduction of the overall probabilities of DNA code formation and translation/decoding to specified protein amino acid sequences. And the above assumptions do not account for the overall complexity of the cellular operations needed for DNA double helix formation, mRNA code transcription, and ribosome translation/decoding operations and resulting protein formation. Also, even if $p(D) = 1.0$ with all required DNA translation/decoding information determined by natural selection, chemical evolution reactions, or other evolution theories, the probabilities of Table 3 would not be changed in a significant way. Chemical reaction research of interest to DNA translation/decoding includes stereochemical theories, coevolution theories and adaptive theories discussed further below in Section IV-A. See Ref. [12].

Accounting for Deep Time

It has been suggested that there has been so much time since the Big Bang that certainly at least one valid protein codon sequence would be formed by natural processes during that time. Consider the probability, p, of forming a valid protein codon sequence with n attempts over all of time. For example, if the probability of finding x equals p, then the probability of not finding $x = 1 - p$. With independence of trials, after n attempts the probability of not finding $x = [1- p]^n$. Based on the Binomial series, it is well known (See Ref.[10]) that:

$$(1-p)^n = 1-np/1! + n(n-1)p^2/2! - \ldots \quad (12)$$

For small p, as is the case here, the probability of not finding x is approximately:

$$(1-p)^n \approx 1-np. \quad (13)$$

The probability of one or more occurrences of x follows as:

$$1 - [1 - n\,p] = np. \quad (14)$$

Consider the formation of a particular protein with the codon coding and amino acid probabilities of Table 1 over all of time since the Big-Bang. The new variable in question here is time. It has been estimated that the Big Bang occurred about fourteen billion years ago. It follows that there has been about $60\times60\times24\times365\times14\times10^9 = 4.42\text{x}10^{17} = 10^{17.645}$ seconds since the Big Bang.

Let $p(C,L,D,T)$ denote the joint probability of formation and translation/decoding of one or more valid codon sequences of length L with required start and stop codes for protein synthesis with selection based on codon coded "C" frequencies of Table 1, column 4, since the Big Bang. Assume further that one such formation and translation/decoding occurs during each one second time. Then:

$$p(C,L,D,T) = p(C,L,D) \times 10^{17.645} = 10^{\,-68.5-\,(H+0.021)L}. \quad (15)$$

For example, for L=430 and with the codon coded frequencies "C" of Table 1, from Table 3, $p(C,L,D,T) = 10^{-613}$, or 000….01 with about 613 zeros before the numeral one.

The manuscript conservative assumption of one second time intervals between DNA changes or mutations compares to intervals of days or even years reported in some references. See, for example, Ref. [35].

Another possible assumption would be the use of Plank's constant for the possible bound on the rate of transition between physical states, about 10^{45} times per second. Plank's time constitutes the smallest meaningful unit of time. Planck's time compares to the about 10^{18} seconds since the big bang used above. Using Plank's time would improve the values of Table 3 columns 8 and 9 by a factor of 10^{27} times per second which still would not change the conclusions of the manuscript, Table 3. Not even considering all possible changes of all elementary particles in the whole history of the universe based on Plank's time interval can account for the numbers in Table 3. See Refs. [28], [29].

Accounting for Universal Probabilistic Resources

It has been estimated that there are about 10^{80} elemental particles (protons/neutrons) in the entire known universe [13]. It follows that there are less than 10^{80} atoms and even fewer molecules. Thus, there are less than 10^{80} total nucleotide molecules in the entire universe and there would be less than 10^{80} nucleotide molecule interactions or opportunities in any time interval to form a coded codon sequence as described above.

For example, with the conservative assumption that there are 10^{80} nucleotide molecules available in each one second time interval in the entire known universe to form a codon coded sequence, there would be at most 10^{80} more opportunities in any time interval. Let $p(C,L,D,T,U)$ represent the formation probabilities of one or more protein codon sequences of length L with proper start and stop codes in all of time and space in the entire known universe. $p(C,L,D,T,U)$ would then be bound by:

$$p(C,L,D,T,U) = 10^{80} \times p(C,L,D,T) = 10^{80} \times 10^{\,-68.5-\,(H+0.021)L.} = 10^{11.5-\,(H+0.021)L}. \quad (16)$$

For example, for L=430 and with the codon coded frequencies "C" of Table 1, column 4, from Table 3, $p(C,L,D,T,U) = 10^{-533.431}$, or 000….01 with about 533 zeros before the numeral one.

The results of column 8 are based on attempts to form the universal codon code of Table 1 at the rate of 1 attempt per second since the big bang. Column 9 provides results for formation attempts once per second with each of those attempts involving every elemental particle (proton) in the entire known universe using the Edington value of 10^{80} particles. See Ref. [13]. The result is suggested as an upper bound of possible formation attempts based on the product of the number of time intervals times the number of particles in all of time (T) and space of the entire universe (U). As noted, increasing the

formation rate even to Planck time intervals would not change the conclusion.

Probabilities as a Function of Protein Amino Acid Sequence Length

Based on the above analyses, Table 3 above provides results for DNA protein codon sequence lengths varying between 100 and 2000 codons. The values given in Table 3 are the computed power of ten exponents for the respective probabilities. The results confirm that the probabilities $p(G/L)$, $p(G,L)$, $p(C/L)$, $p(C,L)$, $p(C,L,D)$, $p(C,L,D,T)$ and $p(C,L,D,T,U)$ are all basically zero. Table 4 below summarizes equations derived above and used in calculations for Table 3. All equations are expressed in terms of two variables: calculated codon code Shannon entropy "H" and protein amino acid sequence length "L".

The above analyses demonstrate the mathematical difficulties for natural formation, selection, and translation/decoding of the Genetic code of Table 1 considering all elemental particles in the entire known universe for all of time from the Big Bang forward.

COMMENTS ON ORIGIN OF THE DNA CODE

The DNA code is required for life on earth. Multiple theories have been proposed in attempts to understand the evolution of life. That work continues today with progress on multiple fronts. Proposed theories of evolution include chemical evolution, code optimization, protein first, multiple protein, RNA world, multiverse, code mutation, natural selection, accidental-freeze theories, and selected combinations of these theories. Details of these, other theories, and ongoing research are beyond the scope of this paper. Brief comments are provided below pointing out implications of the results derived here to the origin of the Genetic code of Table 1 described above.

The origin of the code translation/decoding operations has been described as the most fundamental and hardest problem in evolution biology. See Refs. [11], [12]. The above focuses on mathematical analyses based on Shannon information theory code entropy and information content of the nonrandom codon base-four DNA code for protein synthesis. Importantly, proposed origination theories must provide for the digital information found in the Genetic code of Table 1 with the information content, structural properties and processing as described above.

Chemical and Code Optimization Evolution Theories

One proposed theory for DNA evolution is "Chemical Evolution." In an article by Koonin and Novozhilov, "Origin and evolution of the genetic code: the universal enigma," physicochemical, adaptive, and coevolution are presented as important theories for the origin of the DNA code [12, 14].

Briefly, the *physicochemical* (physical and chemical) theory postulates a natural affinity or tendency of two chemical structures, DNA codons and amino acids, to attract each other and form a chemical bond. Even if such a natural occurring codon/amino acid affinity could be established, the physicochemical theory does not provide the DNA digital base-four code information, structure, and corresponding occurrence frequencies of Table 1. In addition to the DNA code structural requirements described above, the code error protection and code proof reading discussed above in Section II.B must be provided. Explanation of this Genetic code of Table 1 is lacking in physicochemical theories. In DNA translation/decoding the codons identify but are not bonded to the corresponding amino acid in the final proteins. Even if the postulated chemical bonding can be established, more research is required to develop a workable physicochemical digital code information formation and processing model [12, 14].

The *adaptive* theory presumes the existence of some "selective forces" without explanation of how such forces would produce the above-described complex nonrandom DNA digital base-four Genetic code [12, 14].

The *coevolution* theory postulates that: "the code coevolved with the amino acid biosynthetic pathways, i.e., during the code evolution, subsets of codons for precursor amino acids have been reassigned to encode the product amino acids" [12]. The requirement for "codons for precursor amino acids" does not explain origination of the Genetic code of Table 1 [12, 14].

As stated by Koonin and Novozhilov: " … despite a long history of research and accumulation of considerable circumstantial evidence, none of the three major theories on the nature and evolution of the genetic code is unequivocally supported

Table 4. Probabilities of Genetic code Occurrence as Function of DNA Code. Shannon Entropy "H" and Protein Length "L".

Probability	Description
$p(L) = 10^{-3.135 - 0.021\,L}$	Probability of codon sequence of length *L* with start/stop codes. Eq. 2
$H = -\Sigma\{p(i) \times \log_{10} p(i)\}$	Shannon Entropy - $p(i)$ Table 1 codon code (*C*) or Genbank code (*G*). Eq. 8
$p(G/L) = 10^{-LH}$	Conditional Genbank *G* codon probabilities given protein length *L*. Eq. 7,8
$p(G,L) = 10^{-3.135 - (H+0.021)L}$	Joint Human Genbank *G* codon probabilities and protein length *L*. Eq. 9
$p(C/L) = 10^{-LH}$	Conditional DNA protein code *C* probabilities given protein length *L*. Eq. 7,8
$p(C,L) = 10^{-3.135 - (H+0.021)L}$	Joint DNA protein code *C* probability and protein length *L*. Eq. 9
$p(C,L,D)=10^{-86.34-(H+0.021)}$	$p(C,L)$ conditions with mRNA translation/decoding *D* probability. Eq. 11
$p(C,L,D,T) = 10^{-68\,5-(H+0.021)L}$	$p(C,L,D)$ conditions over all time *T* since big bang. Eq. 15

by the currently available data" [12]. Additionally, Brown and Hullender calculate extremely low chemical evolution probabilities for simultaneous natural generation in time and space of different necessarily cooperating protein enzyme and coenzyme molecules [22].

Protein-First Theory

Complex life on earth requires DNA molecules. The protein first theory postulates formation of amino acids in nature and then natural combinations of those amino acids in the correct sequence to form proteins without requiring the DNA molecule codon code [19, 24, 36]. Even if such independent protein formation occurred, it would not provide the digital base-four coded codons found in the DNA code as specified in Table 1 above. Simply knowing the sequence of amino acids in a protein does not provide the necessary information to form the actual corresponding DNA base-four codon code. Unexplained issues include the use of DNA base-four coding, the use of particular nucleotides defining each codon, the use of the 64 codons with protein coding start and stop codes, the assignments of individual codons to particular amino acids, or the number of codons assigned to each amino acid as set forth in Table 1.

In living cells, proteins arise from reading the DNA digital base four code, transfer of that code to ribosomes, and decoding the code instructions to form proteins. The analogy to digital computers is striking. The output is based on the input program code. The resulting proteins are the output. And as in a digital computer, the input code cannot be reconstructed from the output. A protein first sequence does not provide the DNA base-four codon code information and structure of Table 1 described above.

And as noted above, the Central Dogma of Molecular Biology asserts that genetic communication of information within the cell occurs from DNA to mRNA to ribosomes for amino acid selection and protein synthesis. As also noted above, multiple recent references have questioned the completeness of the central Dogma [30, 32]. The lack of RNA coding information assumes and reflects the Central Dogma. Regardless, the basic Central Dogma concept presents an issue for "protein first" theories of DNA development. Protein molecules perform various functions in living organisms. But, according to the Central Dogma, transfer of genetic information to DNA is not one of those functions [5]. The Genetic code itself cannot originate from proteins.

Multiple Protein Evolution Theories

Another suggested possibility to mitigate the prohibitive Genetic code protein formation probabilities of Table 3 is consideration of protein evolution theories with a combination of gradual mutations and selection [19, 24, 36]. Multiple protein variations are possible, some of which affect health while others may not. But gradual mutations and selection assume the existence of a first protein. Regardless, for the reasons discussed in Section 4.2 above consideration of multiple proteins arising through protein independent evolution or without Genetic coded instructions does not provide the necessary information for formation of the Genetic code of Table 1.

According to Bengtson and Edstrom: "In regard to the construction of information rich (coded) sequences reticent in proteins and genetic material, there is a haunting lack of experimental evidence provided in the literature. This appears despite hundreds of published works spanning nearly four decades of work focused on prebiotic models for peptide and, to a lesser extent, nucleic acid synthesis" [24].

RNA-World Theory

Significant DNA and RNA research has furthered the understanding of cellular biology, medicine, and life itself. The RNA World hypothesis sets forth the possible role of RNA molecules in ribosome operations [15, 16]. Altman and Cech received a Nobel Prize in 1989 for research in this area.

> "This year's Nobel Prize in chemistry has been awarded to Sidney Altman, USA and Thomas Cech, USA for their discovery that RNA (ribonucleic acid) in living cells is not only a molecule of heredity but also can function as a biocatalyst. This discovery, which came as a complete surprise to scientists, concerns fundamental aspects of the molecular basis of life. Many chapters in our textbooks have to be revised" [16].

Importantly, RNA must meet the same base-four coding requirements of the Genetic code of Table 1 . The RNA first assumption of RNA World is subject to the same probability of occurrence limitations described above for DNA codon coding and requirements for ribosome translation/decoding. As explained above, translation/decoding requires knowledge of the DNA base-four coded information to properly translate DNA codons to the correct amino acids for protein synthesis. The RNA World theory requires explanation of how such translation/decoding information came about. Also, the RNA molecule itself has been shown to be unstable and will exist by itself for only a short time, making it unsuitable as a basis of DNA origination theories. In recent years, the RNA World theory has been characterized as at least being incomplete:

> "The nexus connecting pre-biotic chemistry to biology is not replication but the translation table that maps amino acid sequences of functional proteins onto nucleotide triplet codons" [18]. "RNA world theories fail to explain how "critical transformation became embedded, in parallel, into tRNA and gene sequences, together with the ribosomal read-write mechanism" [20].

> "The direct evolution of inherited genetic information coupled to encoded functional proteins, as is observed in real-world molecular biology, is far more plausible than any scenario in which there was an initial RNA World of ribozymes sophisticated enough to operate a genetic code. The preservation of encoded information processing during the historically necessary transition from any ribozymally operated code to the ancestral aaRS enzymes of molecular biology

> appears to be impossible, rendering the notion of an RNA Coding World scientifically superfluous. While this conclusion is grounded in an understanding of exactly how the dynamical architecture of molecular biology can solve the computational chicken-egg paradox of code evolution, it leaves a host of problems concerning the evolution of the complex apparatus of translation unresolved" [18].

RNA World theories fail to explain the Genetic code structure of Table 1 and the required translation/decoding rules for mapping that code to amino acids. Without this information there can be no translation/decoding operations.

Recognizing the above RNA world issues, some suggest combining RNA world with other evolution theories, for example, Darwinian evolution. But Darwinian evolution lacks an explanation of the Genetic code origin. Darwinian evolution requires replication for survival of the fittest theories. Replication requires the Genetic code with codon code translation/decoding processing as described above. Further research is required for development of a complete RNA world codon code evolution theory [15–18].

Multiverse Theories

Faced with the above difficult probabilities of natural DNA codon code formation as set forth in Tables I and II and other analyses of very precise values of multiple physical and cosmological constants necessary for life to exist, some investigations have pivoted to postulate the existence of multiple universes or multiverses [37]. It is suggested that if many (or even an infinite number) of separate universes exist (a multiverse), anything is possible including overcoming the probabilities and other issues outlined above. The argument is that we are fortunate enough to live in a "goldilocks" universe that somehow has been finely tuned and fortunate enough to support life.

Despite years of study and speculation concerning the existence of multiverses, there has been a lack of evidence developed that such multiverses exist.

Darwinian Natural Selection

Evolution theories have long been based on Darwinian concepts that life emerged spontaneously, for example from a prebiotic soup, and evolved from simple organisms to the complex life forms we have today based on natural selection and survival of the fittest [4, 21, 22, 31]. The Darwinian theory may play a significant role in the evolution of existing life forms with survival of the fittest improvements. But the DNA code itself is necessary for genetic replication in reproductive survival of the fittest biology [21].

Darwinian natural selection lacks explanation of the origin of the genetic code. Indeed, none of the above discussed theories explain the origin of the Genetic code. Multiple other Darwin theory difficulties have been identified in the literature. Brown and Hullender demonstrated the virtually zero chemical evolution probability of Darwin required natural occurrence combination of protein enzyme and coenzyme molecules [22].

Accidental-Freeze Theories

The accidental-freeze theory of evolution, attributed to Francis Crick, postulates that the DNA code is basically a "frozen accident" rather than the result of deterministic events. The theory suggests that the DNA sequence formed very early in evolution history without significant subsequent changes over time. The sequence was "frozen" because of complexity and negative effects of changes. As discussed above, the extremely low probability of such an occurrence basically assures that the accidental-freeze theory is at least incomplete with serious mathematical issues [23].

The accidental-freeze theory does not explain why a particular code was originally chosen. Rather it suggests that once that code was established, it was not changed because alterations would change defined protein aminos acid synthesis in all life forms.

Design Theories

Faced with the difficult probabilities such as those outlined above for natural genetic code formation as set forth in Tables I and III and other analyses of very precise values of multiple physical and cosmological constants necessary for life to exist, coupled with the understanding of the complexity of all life on earth, increasing numbers of investigators are considering the possibility of design to explain the origin-of-life with scientifically determined biological systems, molecular operations, and information requirements. See, for example, *The Demon in the Machine* [33]. As noted above, none of the present origin-of-life theories, or combinations of those theories, explain the origin of the Genetic code and its structured information content. Dembski suggests that the very small probabilities may justify eliminating chance occurrence of events such as the origin of the genetic code with occurrence probabilities indicated in Table 3 above as a realistic possibility [27, 28].

CONCLUSIONS

Science has determined that DNA sequences encode proteins via the genetic code. Required cellular operations that rely on that code include the formation of the digital DNA double helix coded instructions, cellular mRNA transcription, and the ribosome translation/decoding operations. The analysis presented here focuses on mathematical modeling of the Genetic code and digital processing of that code. Probability formulas and numerical results for DNA digital code formation, transcription, and translation/decoding over all universal time and space are presented. Shannon information theory, DNA code entropy, Bayesian conditional probability, binomial series expansion, and Pearson (Correl) numerical series correlation form the basis of analysis with variable codon frequencies.

It has been estimated that there are about 10^{80} protons in the entire known universe. See Refs. [13], [22]. Imagine naming just one of those protons "Lucy." The probability of finding Lucy with selection out of the entire known universe follows as 10^{-80}. As shown above, the probability of finding Lucy is more

likely than formation of the Genetic code and cellular code processing and translation/decoding operations for just one protein anywhere in the entire known universe. Furthermore, as also discussed above, assuming the universe is about four billion years old, there has been about 10^{18} seconds since the Big Bang. With the conservative assumption of every one of the 10^{80} protons in the entire universe being involved during each of the 10^{18} one second intervals since the Big Bang in codon code formation, the resulting probabilities of natural formation in column 7 of Table 3 are improved by 10^{98} with the results shown in column 9 of Table 3. Based on the results of Table 3, Dembski's "universal probability bound" concepts provide inference for negating natural formation of the universal codon code. See Refs. [28], [29].

Importantly, as noted above, the sections of code defining proteins in the human body are only a small percentage of the over three billion individual DNA code elements with each individual protein requiring a few hundred DNA instructions. The complexity of the DNA digital code as described above is important to development of theories explaining the origin of that code in nature. Comments are provided above for popular theories including chemical, code optimization, protein first, RNA world, multiverse, natural selection, and accidental-freeze theories. A successful theory of Genetic code evolution must provide a path to the structured non-random codon code of Table 1. Important research continues today with the goal of improving understandings of DNA and the origin-of-life.

The manuscript is directed to the origin of the Genetic code set forth in Table 1. Lack of an explanation of the origin of the Genetic code is an issue for naturalistic origin-of -life theories requiring-further research. See, for example, Refs. [3], [7], [11], and [12].

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular Biology of the Cell. 4th edition. Garland Science, New York, NY.
2. Shin S, Cho KH, Kim D (2006) Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences. Physica A 370(2):651–662. **doi:**10.1016/j.physa.2006.02.046
3. Yockey HP (2006) Information Theory, Evolution, and the Origin of Life. Cambridge University Press, Cambridge, UK, pp. 27–30.
4. Moorhead PS, Kaplan MM, eds (1966) Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution. Wistar Institute Press, Philadelphia, PA.
5. Cobb M (2017) 60 years ago, Francis Crick changed the logic of biology. PLOS Biology 15(9):e2003243. **doi:**10.1371/journal.pbio.2003243
6. Tamura K (2016) The genetic code: Francis Crick's legacy and beyond. Life 6(3):36. **doi:**10.3390/life6030036
7. Donohue CR (2023) Genetic Code. NIH - National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Genetic-Code
8. Balado F (2013) Capacity of DNA data embedding under substitution mutations. IEEE Transactions on Information Theory 59(2):928–941. **doi:**10.1109/TIT.2012.2223838
9. Gilis D, Massar S, Cerf NJ, Rooman M (2001) Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biology 2(11):research0049. **doi:**10.1186/gb-2001-2-11-research0049
10. Parzen E (1960) Modern Probability Theory and Its Applications. John Wiley & Sons, New York, NY.
11. Wolf YI, Koonin EV (2007) On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. Biology Direct 2:14. **doi:**10.1186/1745-6150-2-14
12. Koonin EV, Novozhilov AS (2008) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61(2):99–111. **doi:**10.1002/iub.146
13. Eddington A (1953) Fundamental Theory. Cambridge University Press, Cambridge, UK.
14. Ruiz-Mirazo K, Briones C, de la Escosura A (2017) Chemical roots of biological evolution: the origins of life as a process of development of autonomous functional systems. Open Biology 7(4):170050. **doi:**10.1098/rsob.170050
15. Muñoz-Velasco I, Hernández-Morales R, Becerra A (2024) Pioneering role of RNA in the early evolution of life. Genetics and Molecular Biology 47 (Suppl 1):e20240102. **doi:**10.1590/1678-4685-GMB-2024-0028
16. The Royal Swedish Academy of Sciences (1989) Press Release: The 1989 Nobel Prize in Chemistry (Ribonucleic acid (RNA) – a biomolecule of many functions). NobelPrize.org.
17. Xu C (2018) Saying goodbye to the RNA world hypothesis. BioTechniques 64(3):88–91. **doi:**10.2144/btn-2018-0022
18. Kulkarni P, Porter L, Chou TF, Chong S, Chiti F, Schafer JW, Mohanty A, Ramisetty S, Onuchic JN, Tuite M, Uversky VN, Weninger KR, Koonin EV, Orban J, Salgia R (2025) Evolving concepts of the protein universe. iScience 28(3):112012. **doi:**10.1016/j.isci.2025.112012
19. Carter CW Jr, Wills PR (2017) Interdependence, reflexivity, fidelity, impedance matching, and the evolution of genetic coding. Molecular Biology and Evolution 35(2):269–286. **doi:**10.1093/molbev/msx265
20. Rice JW, Warner DA, Kelly CD, et al. (2010) The theory of evolution is not an explanation for the origin of life. Evolution: Education and Outreach 3(1):141–142. **doi:**10.1007/s12052-010-0224-x
21. Brown OR, Hullender DA (2024) Biological evolution is dead in the water of Darwin's warm little pond. Progress in Biophysics and Molecular Biology 193:1–6. **doi:**10.1016/j.pbiomolbio.2024.08.003
22. Koonin EV (2017) Frozen accident pushing 50: stereochemistry, expansion, and chance in the evolution of the genetic code. Life 7(2):22. **doi:**10.3390/life7020022
23. Bengtson M, Edstrom ED (1999) A new method for testing models of prebiotic peptide assembly. In: Chemical Evolution: Exobiology, Elsevier, Amsterdam, pp. 125–134. **doi:**10.1016/S0921-0423(99)80062-8
24. Thorvaldsen S, Hössjer O (2023) Estimating the information content of genetic sequence data. Journal of the Royal Statistical Society Series C: Applied Statistics 72(5):1310–1338. **doi:**10.1093/jrsssc/qlad062
25. Thorvaldsen S, Hössjer O (2024) Use of directed quasi-metric distances for quantifying the information of gene families. BioSystems 243:105256. **doi:**10.1016/j.biosystems.2024.105256
26. Thorvaldsen S, Ohrstrom P, Hössjer O (2024) The representation, quantification, and nature of genetic information. Synthese 204(1):15. **doi:**10.1007/s11229-024-04613-z

27. Dembski WA (2006) The Design Inference: Eliminating Chance Through Small Probabilities. Cambridge University Press, Cambridge, UK.
28. Dembski WA, Ewert W (2023) The Design Inference: Eliminating Chance Through Small Probabilities. 2nd edition. Discovery Institute Press, Seattle, WA.
29. Tan CL, Anderson EH (2024) Revising the Central Dogma: Regulated, dynamic, and system-dependent information coding and decoding. BIO-Complexity 2024(3):1–21. **doi:**10.5048/BIO-C.2024.3
30. Noble D (2011) Neo-Darwinism, the Modern Synthesis and selfish genes: are they of use in physiology? Journal of Physiology 589(5):1007–1015. **doi:**10.1113/jphysiol.2010.201384
31. Shapiro JA (2009) Revisiting the Central Dogma in the 21st century. Annals of the New York Academy of Sciences 1178(1):6–28. **doi:**10.1111/j.1749-6632.2009.04990.x
32. Jablonka E, Lamb MJ (2005) Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. MIT Press, Cambridge, MA.
33. Davies P (2019) The Demon in the Machine: How Hidden Webs of Information Are Solving the Mystery of Life. Allen Lane, London, UK.
34. Wei W, et al. (2022) Rapid evolution of mutation rate and spectrum in response to environmental and population-genetic challenges. Nature Communications 13:4752. **doi:**10.1038/s41467-022-32353-6
35. Csányi V, Csermely P (2005) The origins of life – the 'protein interaction world' hypothesis: protein interactions were the first form of self-reproducing life and nucleic acids evolved later as memory molecules. Medical Hypotheses 64(4):678–688. **doi:**10.1016/j.mehy.2004.11.029
36. Wikipedia Contributors (2026) Multiverse. https://en.wikipedia.org/wiki/Multiverse
37. Abbas U, Chen J, Shao Q (2023) Assessing fairness of AlphaFold2 prediction of protein 3D structures. bioRxiv 2023.05.23.542006. **doi:**10.1101/2023.05.23.542006
38. Peterson WW (1962) Error-Correcting Codes. MIT Press, Cambridge, MA.
39. Aharon A, Polak P, Yaari G (2026) Revealing the inherent design principles of the genetic code via an error correcting code representation. Scientific Reports 16:39862. **doi:**10.1038/s41598-026-39862-0
40. McLatchie J (2026) Introducing the remarkable genetic code. Science and Culture Today. https://scienceandculture.com/2026/05/introducing-the-remarkable-genetic-code
41. Agris PF, Vendeix FA, Graham WD (2007) tRNA's wobble decoding of the genome: 40 years of modification. Journal of Molecular Biology 366(1):1–13. **doi:**10.1016/j.jmb.2006.11.046