Genetic code optimisation: Part 2

Royal Truman and Peer Terborg

In the standard genetic code, 64 codons map to 20 amino acids and a stop signal. The particular assignments show clear evidence that the coding convention used protects against the deleterious effects of mutations. Evolutionists believe the standard code went through multiple refining stages preceding the Last Universal Common Ancestor (LUCA). Variant genetic codes are usually found in mitochondria and are caused by post-transcriptional chemical modifications of the nucleotides of tRNAs and mRNAs. These are complex biochemical processes unlikely to arise naturally. During code modification, ambiguous codon translations lead to a multitude of protein variants present concurrently in a cell. Natural selection therefore would be faced with an ever-varying set of inconsistent goals. Two algorithms programmed with Java show that there are over 1.5 x 10^{84} genetic codes based on a {64 : 21} mapping, most of which are far less robust to mutations than the standard code. In the face of such a huge number of coding possibilities, and an internal and external cellular environment which is changing constantly, natural selection cannot focus on exploring one code after the other in searching for better versions. We suggest that some variants of the standard code may be degenerations, but believe alternatives may have been designed from the beginning, optimal for specific organisms.

The origin of the genetic code is a mystery to the materialist. With but one notable exception,[1] the details and complexity of the genetic machinery are glossed over in various *ad hoc* hypotheses to the point of retaining no chemical or biological relevance. It is not our purpose to analyse here the origin of genetic code theories themselves. Instead we critically examine the common assumption that the code has evolved over time, leading to an increase in sophistication and robustness.

A large number of interpretative frameworks have been published, proposing naturalistic reasons for the particular assignment of 64 codons to 20 amino acids and a stop signal. We have summarized these frameworks elsewhere[2] in three categories: (I) chemical/stereochemical theories, (II) coevolution of biosynthetically related amino acid pathways, and (III) evolution and optimisation to prevent errors. We showed that they lack explanatory substance.[2] We extend the arguments here and introduce new aspects, showing that code modification and improvement is more difficult than generally conceded.

Evolving the genetic code through constraints

Trifonov summarized 40 different theories in the literature through which amino acids were claimed to be added successively into evolving genetic codes.[3] Many interesting facts were documented in these papers. These include details about the biosynthesis of amino acids,[4–6] proportions of amino acids used by proteins,[7] thermodynamic stability of codon-anticodon interactions,[8] alternative uses of the same codon,[9] substitutability by amino acids with similar physicochemical properties,[10] etc.

But the discerning reader recognizes that these vague hypotheses are not real models because they lack sufficient detail and relevance to real chemistry and biology to permit serious discussion or testable experiments. Many of the competing proposals have a large number of supporters, and we cannot provide a detailed critique of each one here. We find that the flaws in these proposals usually fall into two categories: (i) implausible reasoning is used or (ii) alternative or better interpretations of various observations can be accommodated by an 'intelligently designed' interpretive framework. In all cases, it is critical that we make sure some data is not overrated[11] before attempting to critique or re-evaluate information presented in these papers: are the empirical facts correctly reported, have the exceptions or 'outliers' also been communicated, and are the arguments formally correct using scientific and evolutionary principles?

(i) Implausible reasoning

Some illustrative examples follow. It is true that the strength of the interactions between codons and anticodons differ depending on the nucleotides involved. Dr Xia calculated the enthalpies for RNA triplet pair interactions.[8] These ranged from 13.6 kcal/M (for codons AAA and UUU) up to 28.3 kcal/M (GCC and GGG). These facts alone say nothing about the coding convention itself, though. mRNA-tRNA codon-anticodon interactions may be strong or weak, but which amino acid is enzymatically charged to each kind of tRNA molecule is unrelated to such interactions.

The flaw in approaches such as Xia's[8] is to assume that thermodynamical considerations or any other physicochemical properties should have any influence on how a code is devised. They don't.[12] The information stored in the genetic code is a non-materialistic property represented by the order of the nucleotides. Encoded messages derive their use by the outcome which results after processing the message, and the messages are defined by the order of the symbols in the code's alphabet. The physical properties of the media upon which the message is imprinted must not determine the sequence of symbols which comprise a message. This disjoint between natural, physical causes and the messages generated is fundamental to all coded information systems, and no exception has ever been found.[12]

Only an infinitesimal minority of polypeptides fold into stable structures,[13-15] which is comparable to selecting a single kernel of sand from among all the beaches on Earth. A protein cannot be built using amorphic, randomly folded polypeptides. And the instructions to produce useful proteins, based on suitable orders of amino acids, have nothing to do with the physics of codon-anticodon (tRNA-mRNA) interactions. This information-theoretic fact is important, since it means that the messages which need to be generated (here proteins) can be freely encoded genetically. And once encoded, a lack of bias is necessary to permit the intended messages to remain constant over multiple generations. Furthermore, if thermodynamically stronger binding were to be favoured by nature, then more than three nucleotides would be yet stronger. Triplet codes would not have evolved if thermodynamic stability was a major driving force in creating a genetic code. There are incomprehensibly many kinds of interactions between polymers constructed from sugar monomers which are much stronger than the triplet nucleotide hydrogen-bonds used by the genetic code. Nature would have had to ignore all these, and concentrated on a minuscule subset involving weaker interactions using only three nucleotides. But the thermodynamic argument claims the opposite, that the stronger interactions determine the original structure of the genetic code.

Note that a weaker interaction would have led to faster translation *ceteris paribus*, and such 'organisms' or ensemble of molecules would have 'reproduced' faster. We must therefore state that there is no sound basis for reasoning that the code could have developed by beginning with the codons possessing the strongest interactions.

Some have proposed that the codon GCU, based on observed repeats, and its point change derivates, are likely to have been the first codons. These repeats are due to slippery DNA polymerases, and are actually invariably associated with various diseases.[16] Based on the 40 hypotheses for how amino acids may have been incorporated into a genetic code,[3] we read: 'Spectacularly, [the] resulting six amino acids, ala, asp, gly, pro, ser and thr, are, indeed, encoded today by the GCU derivative triplets.'[16] But was this a true prediction of evolutionary theory, or merely another example of the vast number of coincidences which an evolutionist can accommodate to any of the alternative scenarios to 'cherry pick' from? Errors in modern chromosome replication reflect today's polymerases; hugely complex machines that use many proteins which are based on *all* the amino acids available today. There is no justification for assuming any relationship to errors found in the existing genetic machinery and some theoretical totally unrelated unspecified primitive genetic precursor system.

Trifonov[3] combined the appearance order of amino acids into the genetic code as expected from the 40 theories in the evolutionary literature into a consensus order. We won't go into the details of his conclusions except for a few comments. In terms of critical thinking,[11] we were disappointed to read that 'Nine amino acids of the Miller's imitation of primordial environment are all ranked as topmost (G, A, V, D, E, P, S, L, T).' As Bergman pointed out,[17] the first experiment produced no amino acids, and only after performing hundreds of carefully planned experiments under different incompatible conditions have researchers been able to produce a handful of amino acids, in miniscule proportions. These hypothetical atmospheric compositions are not relevant to origin-of-life theories anyway.[17-19]

The amino acids not ranked in the list above are for the most part considerably more complex and therefore less likely to form in these kinds of laboratory experiments. Some simply cannot be produced under the experimental set-ups since a necessary chemical element was not made available (e.g. without a source of sulphur, amino acids cysteine (C) and methionine (M) can't form). But what is the meaning of the claim that some amino acids were coded for *first*? What purpose would a genetic scheme serve which is only able to code for say, glycine (G) and alanine (A)? Molecular machines cannot be produced using just one or two or three amino acids, as an intermediate stepping-stone during evolution. In support of this claim, one needs only to examine the 'highly conserved' proteins which show very little variability across all organisms, such as those critical to cellular information processing, which evolutionists interpret as reflecting an ancient origin with little mutation tolerated since. One quickly notes that inevitably all or most of the twenty amino acids are also used in these proteins. This is unlikely if primitive proteins once were based on a much smaller set of amino acids.

Suppose many key proteins showing little amino acid substitution, and found in all domains of life, were found to be almost only composed of the same, small subset (10 or less) amino acids. Every self-respecting evolutionist would immediately present this as conclusive and compelling proof the genetic code started simple and grew in complexity. Why the reluctance to make such a prediction *a priori*? We hope to examine these sequences in the future, and predict now, before looking at any protein sequences, that these proteins will also use the full repertoire of amino acids currently available through the standard genetic code.

(ii) Alternative explanations

It has been suggested that more complex amino acids were incorporated later into the code[20] and the latest amino acids would be largely underrepresented among modern proteins.[7,21] Is this a legitimate prediction unique to the evolutionary viewpoint? Intuitively, anyone believing the genetic code was designed would expect amino acids with complex physicochemical features to be present in lower proportions. For example, the amino acid proline introduces 'kinks' in the polypeptides formed, producing major geometric irregularities. These irregularities tend to disrupt alpha helices and beta sheets, the fundamental building stones used to create proteins. To illustrate, suppose you wish to build a vacation home. As building materials you collect many identical nails, boards and bricks. But you will only need a few locks or bathtubs. The general-purpose items are needed in large numbers, the specialized in fewer.

The statistics of the building items say nothing about how the house came to be.

Mechanisms to rewrite the genetic code. Various researchers have given thought as to how the codons have come to be distributed to represent 20 amino acids, and how coding conventions may be able to change. For the tRNA adaptor to match a codon with an amino acid, two steps are involved[22,23] (figures 1 and 2). An aminoacyl-tRNA synthetase (aaRS) must covalently link the correct amino acid to the corresponding tRNA. Then, within the ribosome machinery the anticodon of charged tRNA must base pair with the corresponding codon of an mRNA strand. The sequence of amino acids in a protein is thereby determined by the order of codons on the mRNA.

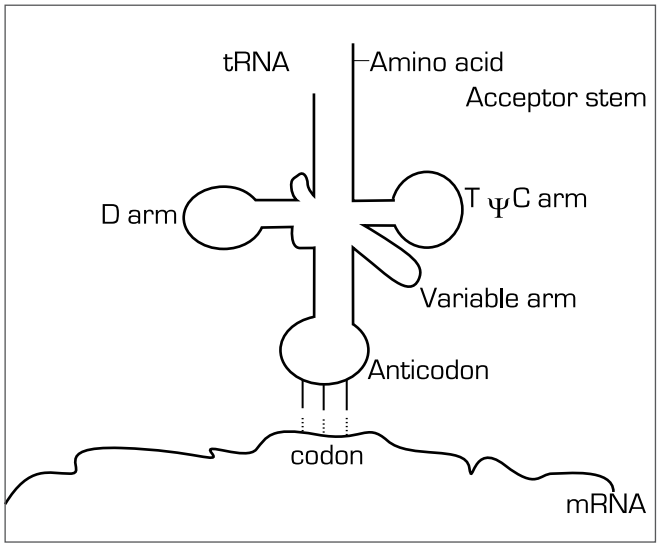

Figure 1. Structure and key parts of tRNA adaptor molecules.

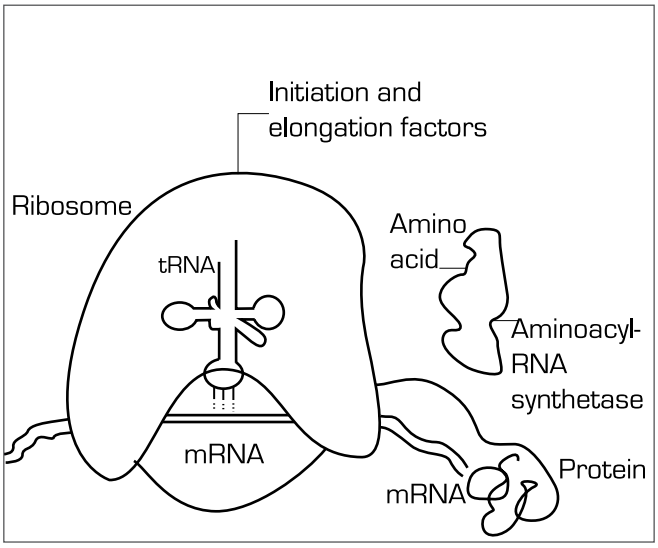

Figure 2. Some of the machines used to translate genetic information. Not drawn to scale.

There are four main mechanisms through which nature might be able modify the genetic code:[22]

Specific nucleotides could be modified on individual mRNA strands. Post-transcription chemical modifications known to all three domains of life (archae, bacteria and eukaryotes) include U –> pseudouridine and A –> inosine. This affects the codon-anticodon interactions such that a different amino acid is introduced from what is expected based on the original codon within the gene. The mRNA itself is modified before being translated.

tRNA molecules can be chemically modified. One of the three anticodon bases is sometimes chemically modified in some tRNAs[24] causing other codons in mRNAs to be recognized. Mutations elsewhere in the tRNA can also alter the usual codon recognition patterns by changing the topology in the anticodon region. Remarkably, tRNA editing can be limited to particular organelles. For example, special signals can target a tRNA from the nucleus to mitochondria in the case of the protist *Leishmania tarentolae,* where the CCA anticodon of $tRNA^{Trp}$ is modified to UCA, permitting reassignment only in the mitochondria.[25] The available data indicates that chemical modification of anticodon bases is the major cause of genetic code variation.[24]

The aminoacyl-tRNA synthetases (aaRS) could mutate. The tRNA-binding and amino acid binding domains of the aaRSs rely on many of the enzymes' amino acids. Mutations here would probably alter the translation of several codons and is therefore a very unreasonable proposal for altering the genetic code.[26]

Components of ribosome could mutate. Changes in components of the ribosome could affect precisely where the tRNAs are held and thereby the nature of codon-anticodon interactions. This would probably affect all interactions with codons and has essentially zero chance of working as this type of mutation would cause rampant nonsense decoding.

There are three well-known scenarios through which variations in the code might occur.[25]

(i) The 'codon capture' hypothesis suggests that strong changes in G+C content over time could cause some codons to disappear from the genome.[27,28] Because nucleotides A and T are complementary (as are also nucleotides C and G), bias towards high or low G+C genome content will statistically favour disappearance of some codons, affecting both protein-coding genes and the anticodons of tRNA producing genes. Should the bias direction reverse strongly at a later time, then mutations might cause the codon to reappear, and if a tRNA charged with a different amino acid should subsequently recognize these new codons, the coding convention would have changed.[24]

Objections. This is a clever argument, but resembles guiding a funnel over a hole and then pretending to be surprised when the golf ball lands there. Going from perhaps 30% G+C content to say 70% G+C and then back to 30% is not so easy to justify in free nature. During the shift in G+C content, suitable mutations must then occur,

Table 1. Distribution of codon usage for select organisms (per 1000), and C+G overall genome content (%). (From The Codon Usage Database.[27])

Codon	1 UUU	2 UUC	3 UUA	4 UUG	5 CUU	6 CUC	7 CUA	8 CUG	9 AUU	10 AUC	11 AUA	12 AUG	13 GUU	14 GUC	15 GUA	16 GUG	17 UUC	18 UCU	19 UCA	20 UCG	21 CUC	22 CCU	23 CCA	24 CCG	25 AUC	26 ACU	27 ACA	28 ACG	29 GUC	30 GCU	31 GCA	32 GCG
a)	22.6	8.8	16.6	5.3	11.5	7.4	13.3	20.8	30.3	9.2	18.3	9.3	18.4	15.6	32.3	24.7	16.3	8.8	12.1	6.4	11.0	5.6	9.6	21.1	24.6	22.8	21.1	15.8	15.1	25.2	18.8	28.6
b)	31.1	15.6	14.6	3.3	19.8	10.4	8.0	2.4	46.7	20.3	14.1	1.9	14.1	15.6	7.1	11.3	25.0	11.8	14.1	3.3	18.9	8.5	15.1	3.3	38.2	33.5	26.4	10.4	8.5	22.6	8.5	15.6
c)	13.2	7.1	10.8	5.4	9.0	6.9	10.8	8.8	16.5	9.5	21.0	11.5	11.0	14.4	27.6	13.3	21.8	19.6	18.4	13.1	13.8	18.1	16.2	18.0	22.8	21.3	26.3	20.5	13.9	33.6	24.6	26.7
d)	23.3	16.7	17.5	11.2	21.2	8.8	14.1	11.2	32.2	18.9	30.1	12.1	24.1	22.4	35.7	10.9	23.9	10.6	13.7	9.1	14.9	4.4	9.2	5.1	18.9	10.4	18.3	8.4	13.6	12.7	17.1	6.7
e)	17.6	14.6	11.6	9.8	11.1	17.7	8.4	5.5	16.1	13.4	14.6	9.9	10.6	20.7	21.6	12.6	24.7	18.2	18.0	13.2	20.1	19.1	15.3	11.0	28.1	21.9	22.9	17.5	17.9	28.5	29.8	24.2
f)	17.1	16.1	12.1	11.1	13.3	18.4	10.4	4.7	15.4	13.6	15.5	12.5	10.6	20.1	21.1	11.4	22.0	18.1	16.3	12.1	20.3	18.3	15.3	9.5	22.9	19.1	20.5	19.7	15.6	26.3	26.4	21.5
g)	13.6	16.1	11.8	9.2	11.4	16.1	10.7	4.6	13.7	12.3	13.3	9.8	9.7	20.6	17.5	9.1	22.0	18.9	19.1	13.0	22.1	21.2	15.5	13.0	23.0	25.5	18.2	18.7	15.6	29.1	26.9	23.4
h)	17.5	15.1	12.1	10.5	13.1	17.5	10.8	4.6	15.9	13.1	16.9	12.1	11.0	18.5	21.8	10.8	20.4	17.7	15.3	12.6	19.6	19.8	15.1	10.5	20.9	18.9	19.1	19.5	14.5	27.9	25.2	22.3
Min:	13.2	7.1	10.8	3.3	9.0	6.9	8.0	2.4	13.7	9.2	13.3	1.9	9.7	14.4	7.1	9.1	16.3	8.8	12.1	3.3	11.0	4.4	9.2	3.3	18.9	10.4	18.2	8.4	8.5	12.7	8.5	6.7
Max:	31.1	16.7	17.5	11.2	21.2	18.4	14.1	20.8	46.7	20.3	30.1	12.5	24.1	22.4	35.7	24.7	25.0	19.6	19.1	13.2	22.1	21.2	16.2	21.1	38.2	33.5	26.4	20.5	17.9	33.6	29.8	28.6

	33 UUA	34 UCA	36 UGA	37 CUA	38 CCA	39 CAA	40 CGA	41 AUA	42 ACA	43 AAA	44 AGA	45 GUA	46 GCA	47 GAA	48 GGA	49 UUG	50 UCG	51 UAG	52 UGG	53 CUG	54 CCG	55 CAG	56 CGG	57 AUG	58 ACG	59 AAG	60 AGG	61 GUG	62 GCG	63 GAG	64 GGG
a)	14.2	7.7	1.0	4.0	8.4	15.5	3.7	5.0	7.8	33.4	2.4	11.1	20.7	39.1	8.6	13.9	8.8	0.2	15.2	52.1	22.5	29.0	5.6	27.5	14.5	10.3	1.5	25.7	32.7	17.8	11.2
b)	35.4	33.5	25.5	65.1	32.1	25.0	10.8	52.3	44.3	23.6	0.5	17.9	27.3	16.0	28.8	9.0	1.9	0.5	0.9	8.5	1.4	1.4	0.0	10.4	0.9	1.9	0.5	8.0	0.9	5.7	8.5
c)	4.5	7.9	0.5	8.2	13.6	15.6	8.5	9.5	11.1	16.9	5.1	6.4	12.8	21.1	18.0	16.1	16.7	0.6	9.8	38.2	15.9	36.3	8.2	23.6	14.4	39.4	6.3	27.7	14.0	42.5	4.7
d)	9.8	20.6	1.4	7.9	26.1	27.4	12.0	9.4	20.0	37.5	15.4	9.8	19.8	40.8	31.7	20.0	12.2	0.6	11.1	12.1	9.7	14.3	4.7	26.1	8.9	25.9	3.9	14.3	8.2	24.5	4.4
e)	5.0	8.7	1.2	6.8	15.3	9.5	6.8	6.0	13.6	22.6	9.6	6.4	13.7	25.5	16.3	12.2	4.0	0.5	12.3	40.0	6.6	29.9	9.3	23.4	6.3	37.2	10.7	31.4	6.9	40.1	15.4
f)	6.6	11.7	1.2	8.0	17.1	11.7	6.6	7.2	15.9	21.7	11.7	7.4	15.8	26.9	16.8	13.3	4.3	0.6	12.4	40.0	6.2	34.2	10.3	22.9	5.7	33.9	12.0	28.8	6.5	39.7	15.3
g)	5.6	9.9	1.6	6.5	14.1	11.1	6.4	6.6	14.4	18.9	12.8	5.5	14.3	21.0	14.7	10.6	4.0	0.6	17.2	42.9	7.8	39.4	12.5	21.2	6.6	30.1	13.4	32.2	10.5	43.0	19.0
h)	7.6	12.2	1.5	7.2	16.9	12.2	6.2	7.4	15.0	24.3	12.1	7.1	15.9	28.8	16.5	12.9	4.4	0.8	13.2	39.8	6.9	34.2	11.5	22.1	6.1	31.9	11.9	28.2	7.4	39.6	16.4
Min:	4.5	7.7	0.5	4.0	8.4	9.5	3.7	5.0	7.8	16.9	0.5	5.5	12.8	16.0	8.6	9.0	1.9	0.2	0.9	8.5	1.4	1.4	0.0	10.4	0.9	1.9	0.5	8.0	0.9	5.7	4.4
Max:	35.4	33.5	25.5	65.1	32.1	27.4	12.0	52.3	44.3	37.5	15.4	17.9	27.3	40.8	31.7	20.0	16.7	0.8	17.2	52.1	22.5	39.4	12.5	27.5	14.5	39.4	13.4	32.2	32.7	43.0	19.0

		Coding G or C	1st letter G or C	2nd letter G or C	3rd letter G orC
a)	*Escherichia coli* 536	51.51%	58.68%	40.74%	55.11%
b)	mitochondrion *Abronia graminea*	39.27%	44.70%	40.74%	32.39%
c)	*Drosophila melanogaster*	53.87%	55.81%	41.51%	64.29%
d)	*Caenorhabditis elegans*	42.93%	50.00%	38.98%	39.82%
e)	*Rattus rattus*	52.82%	55.37%	41.43%	61.67%
f)	*Mus musculus*	52.21%	55.43%	42.21%	58.99%
g)	*Pan troglodytes*	54.78%	56.74%	43.98%	63.63%
h)	*Homo sapiens*	52.34%	55.78%	42.55%	58.67%

and the new codons generated just happen to be placed in judicious places to now code for the right kind of amino acid.

Most changes in the standard code occur in mitochondria, which are all A+T rich. Here the data shows the opposite trend than predicted by this hypothesis: only a small minority actually involve codons with C or G in the third position (which is less constrained to mutation). '*So the codon capture model does not seem to explain adequately the pattern of codon reassignments.*'[24]

A paper by Castresana[29] has often been used to support the codon capture hypothesis, although the codon which disappears, AAA, is unlikely to have done so in the A+T rich genome![24] Note that in the mitochondrion we randomly selected from a public database[27] (*Abronia graminea,* A+T content 60.7%), codon AAA is present in 23.6 out of 1000 codons (table 1), which is about 1.5 times the average value of 15.6 (1000/64 expected statistically for C+G = A+T content). Consistent with the high A+T overall content, some codons possessing two As are present in large amounts (ACA: 44.3/1000, AAC: 26.4/1000) (table 1). However, other codons having two As are completely absent on this mitochondrion's genome, or infrequently present (UAA and AAG: 1.9/1000) in spite of the high over-all A+T content. Remarkably, the AAG codon proportion of 1.9/1000 is much lower than that found in *E. coli* (10.3/1000), even though the later has a much lower A+T genome content (table 1). It does not seem that fluctuating biases towards high or low A+T content would really drive codons in and out the genome very effectively.

(ii) The 'ambiguous intermediate' hypothesis claims mutations in tRNAs distant from the anticodon can affect the nature of the codon-anticodon interactions, leading to alternative amino acid assignments. If one is particularly useful it is selected. In this hypothesis, there could be two tRNAs that decode the same codon but are charged with different amino acids. Modifications distant from the anticodon can also alter which codons are recognized by distorting the shape of the tRNA in the region of interaction. It is also known that a single tRNA can be charged with more than one amino acid. For example, the CUG codon is translated as both Ser and Leu in some organisms, including *Saccharomyces cerevisiae.*[30] This leads to misfolded peptides which stimulate heat-shock proteins which, under exceptional environmental challenges, can permit survival.

Objections. Ambiguous decoding, especially in genomes which involve hundreds or thousands of genes, would typically introduce alternative amino acids at multiple positions in most of these proteins. The evolutionary assumption is that such ambiguity is not carefully crafted via specialized molecular machines in response to external

signals. If the proportions and locations of alternative amino acids are guided in useful manners by complex machinery, this suggests deliberate design. Ambiguity may be occurring that degrades the quality of genomes, but would be a poor explanation for how a coding scheme may be changed to a new, better variant through evolutionary processes.

Examples of alternative amino acids being charged on to the same tRNA, if clearly demonstrating biological use, is compatible with the design hypothesis. Sometimes a rifle which shoots a single bullet must be designed, and sometimes a shotgun is needed. The average distribution of pellets from a shotgun needs to be guided to perform the intended purpose. For example, a shotgun barrel 3 cm long would not be terribly useful for hunting purposes, nor would such a weapon which shoots at the wrong time.

(iii) The 'genome streamlining' hypothesis suggests that especially in very small genomes such as obligate intracellular parasites and mitochondria, simplification in the coding system can occur by loss of some tRNA genes.

Evaluation. Most of the variants to the standard code are found in mitochondria. These use a vastly smaller number of proteins than coded for by nuclear genes. It would be statistically more likely that a codon reassignment would be tolerated by a few dozen mitochondria proteins than thousands in the cytoplasm. Therefore, it is possible that a process of degradation, loss of sophistication and fine-tune specificity, is occurring over time in the sense of the *'genome streamlining' hypothesis*. *Contra* this explanation is the sheer complexity of the processes involved in chemically modifying either specific mRNA codons or portions of tRNAs. It seems more likely that exceptions to the standard code needed to be intelligently planned, and the chemical processes therefore provided. This fine-tuning requires interaction with signals so that the processing equipment can identify where, and perhaps when, to make the chemical modifications.

Evaluation of such proposals

A significant observation is that the exact same kinds of putative codon reassignments are found in phylogenetic patterns which cannot be explained by common descent.[23] Since the claim of reassignment is merely an assumption and not an observation, the obvious question then arises: is not deliberate design a better explanation?

Another point is that the genomes of mitochondria vary dramatically in gene content throughout nature.[31] For a small number of proteins in such organelles, optimal designs based on different coding conventions could have been created. In the future we hope to see if separately designed families of mitochondria can be discovered through mathematical clustering analysis.

How large is the code search space?

There are many ways various codons could be assigned to a given amino acid or stop signal.[32] In one scheme a single codon could code for one amino acid, and all the remaining codons assigned to 'Stop'. There would be 64 x 63 x 62 x … 45 = 64!/(64–20)! = 4.8 x 10^{34} alternative codes in such a coding strategy. Alternatively, the number of codons could be distributed 'more fairly', using three or four codons for each amino acid, and many other schemes could be devised.

Table 2. Reported values for Stirling numbers of the second type validate results from the Java-based algorithm.[38]

n	$\left\{ {n \atop 0} \right\}$	$\left\{ {n \atop 1} \right\}$	$\left\{ {n \atop 2} \right\}$	$\left\{ {n \atop 3} \right\}$	$\left\{ {n \atop 4} \right\}$	$\left\{ {n \atop 5} \right\}$	$\left\{ {n \atop 6} \right\}$	$\left\{ {n \atop 7} \right\}$	$\left\{ {n \atop 8} \right\}$	$\left\{ {n \atop 9} \right\}$
0	1									
1	0	1								
2	0	1	1							
3	0	1	3	1						
4	0	1	7	6	1					
5	0	1	15	25	10	1				
6	0	1	31	90	65	15	1			
7	0	1	63	301	350	140	21	1		
8	0	1	127	966	1701	1050	266	28	1	
9	0	1	255	3025	7770	6951	2646	462	36	1

We and others recognized years ago that the search space of possible genetic codes is too large to be tested in an effort to find better error-minimization alternatives.[33,34] Just how many different genetic codes could exist, based on 64 codons which need to code for 20 amino acids (aa) and a stop signal?

Yockey calculated[35] that there are 1.4 x 10^{70} codes with the characteristics of the genetic code, meaning there are three amino acids represented by 6 codons, five by 4 codons, one by 3 codons, nine by 2 codons, and two amino acids covered by only 1 codon (actually the correct answer[36] is 2.3 x 10^{69}). However, there is no reason other combinations of codon–amino acids should not be permissible.

Professor Clote reported that there are 1.51 x 10^{84} possible codes based on the logic of 21!S(64,21), where S(n,m) is a Stirling number of the second kind.[37] We contacted him for details, and were told the LISP-based source code had been lost. He generously provided a good reference to Stirling's numbers and some helpful comments on the logic.[38] We wished to be sure Clote had considered all 64 ⇨ 21 possibilities and decided to program the algorithm using JAVA (Appendices 1 and 2), which anyone is free to install and use. We also need such a program to determine how many codes could exist using other assumptions such as a doublet of four nucleotides coding for less than twenty amino acids (some evolutionists claim the present code evolved from such a simpler one). The recursive algorithm (Appendix 1) was validated against known numbers (table 2),[38] but after a week, it had still not calculated S(64,21) using a PC with 3 GHz CPU and 1 GB memory.[39] In private correspondence, Clote informed us that their algorithm took a couple of weeks or so to execute.

From the recursive logic, it is obvious that the same recursive calculations are being repeated a huge number of times. We therefore used a dynamic programming bootstrap approach (Appendix 2) which took less than a second on the same PC to solve S(64,21) (the reader is free to choose which method he or she would like to use…). The bootstrap

approach gave exactly the same number of codes Clote and Schönauer had reported:[37,40] S (64,21) = 2.95572845518811 x 10^{64}, and therefore 21! x S (64,21) = 1.5 x 10^{84} alternatives in total. Additional trials in addition to these reported in table 2, using large number S(m,n) values, gave the same results for our two algorithms.

Discussion

To fold stably, proteins need a large number of well-placed hydrophilic and hydrophobic residues. As a minimal requirement, a genetic code cannot use codons which lead predominantly to one class of amino acid or the other. For example, it would be close to impossible to generate a stable, folded protein using 95% hydrophilic residues at random positions. Using too many large bulky residues or too many prolines would also not work. Permitting too many disulfide bonds is also not likely to permit very many useful proteins to form. Producing necessary secondary structural motifs, based on helices and sheets, using amino acids with suitable characteristics (size, hydrophobicity) presents additional contraints. It is also obvious that a random sequence of nucleotides in DNA or RNA is virtually never going to lead to useful proteins.[15] The evolutionist therefore argues that some kind of primitive genetic replicator was generated which improved its coding scheme over time.[2]

Evolving toward an optimal code

The evolutionist could assume either of two models:[23]

1. many codes existed at some point in competition
2. a single code changed over time.

Of course, these scenarios require at least one genetic code to have already arisen by chance, which in light of the above is fantastically unreasonable.

(i) Many codes in competition. The best code would presumably be the most error tolerant. At this time, no one has offered any reasonable model[1] for how all the necessary components for the universal code (ribosome machinery, polymerases, aminoacyl-tRNA synthetases, etc.) could have appeared at the same time and place. Producing a large number of independent versions which compete with each other only compounds the unlikelihood. To fine-tune genes a vast number of mutations would be needed, which would be facilitated if one had as large a population as possible. This model, however, would fragment the attempts into independent code-family sub-populations, decreasing yet more the opportunities of finding improved genes by trial-and-error. However, without a near-optimal ensemble of genes, natural selection would have no way of evaluating which sub-population would be 'more worthy' to be maintained. Mutations leading to changes in genes which are barely useful, instead of near optimal, would be essentially impossible for natural selection to identify.

There are many random factors which affect survival, and the relationship between genetic information and morphological outcome is highly stochastic.[41] Had there been thousands of very different genetic codes long ago, it defies logic to believe all members of each of these competing codes could have been displaced by the fortunate members of a better code, leaving no evidence of their existence. This is supposed, by evolutionary reasoning, to have occurred before the LUCA lived,[23] about 2.5 billion years ago.[42] Actually, many evolutionists believe the genetic code is almost as old as our planet, which would provide virtually no time to evolve.[43]

(ii) A single code changed over time. One has to assume that an acceptable triplet nucleotide genetic code arose somehow. The biosynthetic pathways for novel, highly precise aaRSs somehow arose, *having no selective merit until the process of development was perfected*. It is important to remember, that each codon position of the mRNA strand must have a meaning (i.e. must code for an amino acid or 'stop' signal).[33] Otherwise during translation, the decoding would stall. For every triplet combination an anti-codon on a charged tRNA must exist. There cannot be long gaps along the mRNA which cannot be translated. Also, metabolic paths to provide the amino acid 'feed stocks' need to have arisen at the same time.

According to evolutionary reasoning, a LUCA would have been a fairly sophisticated organism, with between 1,344 and 1,529 gene families[44] and containing the full set of tRNA synthetases and tRNAs.[45] Many speculate that preceding DNA-based genetics, life forms based on RNA or other chemistries had already existed. This means that the amount of time for experimenting with different DNA-based codes would have been quite limited. And the coding differences would have been tested on organisms with thousands of genes.[44] What would a code rewiring process do without intelligent guidance? The chances of mutations occurring at precisely the right locations to provide useful opportunities are very remote. For countless generations the genome would produce a large number of proteins with different sequences from the same genes. A mixture of favourable and deleterious proteins would result, at best. More realistically, survival of the lineage would be highly unlikely.

Going in every direction at the same time

We now wish to introduce the notion which prompted this paper. The scenarios discussed in the literature not only oversimplify what is involved in evolving to a new code, but also postulate guiding forces which are natural. If one believes that many mechanisms exist which could permit the code to evolve, we will see that nature has no way to stay focused on any particular evolutionary direction long enough to accomplish any substantial change.

This is a detail we believe has been overlooked in the literature on code evolution. Let's reason this out. If a gene is not expressed for several generations, nature cannot select for it. So, since natural selection is to play a role, we must assume the genes involved in an evolving code are being expressed. Each generation, an individual gene would on average generate between a few and up to millions of copies of a protein, depending on the gene.[46] During putative code modification, when codon meaning

is ambiguous, a wide variety of different protein sequences would be produced.[47] For a small 350-codon gene, each coding codon will on average be present about *5.7* times (350/61). For two meanings of the same codon this leads to $2^{5.7} = 52$ different protein variants on average, randomly produced, for each gene in the genome. The content of the cell would change every second! How is natural selection to focus on any direction in all this chaos?

Evolving a new meaning for a codon which involves a *very different* amino acid is unrealistic. On average, five to six positions of each gene would be affected. This is bound to be deleterious. But alternatively, coding for an amino acid with *very similar* physicochemical properties is hardly going to provide a recognizable selective advantage, especially in view of the cellular chaos mentioned above. At the same time a specific code is being improved by natural selection, nature must be busy carrying out a myriad of other evolutionary duties: genes must be fine-tuned, for example, to form optimised enzymes, novel metabolic networks must be generated, plus the thousands of other marvels natural selection is supposed to have performed. Natural selection would face a multitude of unrelated goals, both from the environment and the changing cellular components, and would attempt to go in every direction at the same time and end up going nowhere very fast. New, improved genetic codes in free-living organisms would not evolve in this manner.

Exploring codes by randomly changing codon assignments for amino acids in this way, one after the other, makes no sense. Jumping over the moon once is hard enough. Having to do so billions of times in an attempt to find better coding assignments is much less likely. Therefore, it is remarkable that claims of the genetic code being the most robust to random mutations, out of a million alternatives, is accepted so readily by evolutionists.[48] The distribution of 'code quality' would roughly follow a Gaussian distribution, based on the law of large numbers. This means that as coding conventions are improved, it would be ever more difficult for natural selection to generate yet better codes.

The fact that there are so many possible '64 ⇨ 21 mappings' is a major conceptual difficulty for an evolutionist. Having accepted that there are mechanisms which permit codes to evolve, one must accept that up to 10^{31} organisms[13] may suffer mutations which would permit variation in over 10^{84} possible coding directions, the great majority of which would be inferior.

Switching to a different coding scheme is going to require a vast number of mutational attempts no matter which of the above mechanisms would be used. For example, chemically modifying specific nucleotides, whether on tRNA or mRNA, requires complex new biochemical networks to be created. Nature can't know in advance that this will eventually turn out to be a great idea. Countless random mutations would have to be generated. Meanwhile, other visionary organisms would be mutating to see if another of the over 10^{84} codes might be better. If all this were true, we should find evidence of such exploratory behaviour among current prokaryotes.

Before any new coding scheme could be in place, *multiple* mutations consistent with that evolutionary direction must occur. Since the number of directions is so much greater than the number of organisms, relatively few would stay focused on one possible direction. Mutations would affect different codons for different lineages. And the probability of producing the necessary number of random mutations to finalize one exploratory coding attempt would be too small when limited to very small fractions of the 10^{31} organisms and mutation rates sufficiently low to avoid immediate destruction.

It should be clear that attempting to change *several* codon assignments at the same time, instead of replacing one codon at a time by another, would produce a much greater variety of protein sequences from most genes at the same time and would lead to utter cellular chaos.[25,47]

In two papers,[49,50] we explored the common assumption that genes can duplicate and one copy would be unconstrained to mutate and possibly discover new useful proteins. Careful mathematical analysis revealed that the extra metabolic cost and longer replication time of possessing additional genes for small genomes would have a measurable selective disadvantage. Such lineages would go extinct long before enough mutations could accumulate to fortuitously create new useful genes. The new genes necessary to chemically modify specific bases would initially present considerable selective disadvantages, even assuming they were to have a suitable duplicate gene version from which to initiate evolution. Novel aaRSs would also be strongly disadvantageous until the whole new coding scheme was in place.

The data seem to show that when a particular coding variant is used, it is used by all members of that taxa even though the organisms are present in large populations with long generation times.[23] The simplest evolutionary explanation would be that natural selection must have very strongly favoured this fitness peak for all members of these sub-populations. Given that our analysis strongly denies that natural selection is able to perform this feat, we introduce the hypothesis that coding alternatives, which are clearly not a degradation of the standard code, represent independent designs. The observation that the exact same code appears distributed throughout nature in a manner inconsistent with common descent reinforces this view. Since evolutionists recognize that mitochondria are very different from each other, and propose multiple bacterial mitochondrion capture events,[51] we suggest another research approach: there may be separate designs of mitochondria, perhaps recognizable from clusters of genes with the same coding convention. These could suggest alternative designs.

Conclusions

The processes which would modify a coding scheme without the need for new biochemical networks include primarily mutations to ribosome components and aaRSs.

These would have drastic effects on multiple codons and cannot be considered feasible. Most examples of changes in the standard code involve post-transcription, chemical changes to nucleotides. These are biochemically complex processes which require new genes and biochemical networks to be first developed. The chemical changes to various nucleotides must be precisely regulated, which implies severely deleterious effects on the organism until perfectly evolved. During these putative evolutionary processes, those lineages would also be subject to negative selection due to the energy cost of having extra genes and longer chromosome replication times.[49]

During the process of evolving to a new code, a given codon would produce any of two amino acids on average at five or six positions for all genes. This means each organism with an identical genome would produce thousands of protein variants whose composition would change every second. Along with a multitude of external environmental challenges, this means natural selection would not be able to identify an improved coding variant for the multitude of generations needed to further fine-tune the additional mutations needed. The huge number of '64 ⇨ 21 codes' possible, for which nature cannot know in advance which would be improvements until perfected, supports further the view that natural processes did not lead to the 20 or so variant codes discovered so far, except for cases of clear degradation.

The distribution of the same codes across taxa is inconsistent with the evolutionary idea of common descent, and the predominance of alternative codes in mitochondria, which differ greatly from each other, is highly suggestive of having resulted from separate designs.

Appendix 1

Recursive algorithm to calculate the number of possible S{m:n} codes, using Java

```
class Stirling_recurse
{
 double calc(int n, int k)
 {
 double index=–1;
 if ((n == 0) && (k == 0)) index=1;
 else if ((n > 0) && (k == 0)) index=0;
 else if ((n > 0) && (k == 1)) index=1;
 else if (n == k) index = 1;
 else index=k*calc(n–1,k) + calc(n–1,k–1);
 return index;
 }

 public static void main(String[] args)
 {
 Stirling_recurse stirling=new Stirling_recurse();
 int n=20, k=4;
 System.out.println("n= " + n + " k= " + k + "
stirling= " + stirling.calc(n,k));
 }
 }
```

Appendix 2

Bootstrap algorithm to calculate the number of possible S{m:n} codes, using Java

```
class Stirling_kind_2
{
 public static void main(String[] args)
 {
 int n = 65; // Nr of objects + 1;
 int k = 22; // Nr of sets + 1
 int m, i, j; // Use in indexes
 double[][] stirl = new double[n][k];

 for(i=1; i<n; i++) {stirl[i][0]=0; stirl[i][1]=1;}
 for(j=0; j<k; j++) {stirl[j][j]=1;}

 for(i=3; i<n; i++)
 {if (i < k) {m = i+1;} else {m=k;}
  for(j=2; j<m; j++)
  {stirl[i][j] = j*stirl[i–1][j] + stirl[i–1][j–1];}
 }
 System.out.println ("Stirling[n][k]=
  " + stirl[n–1][k–1]);

 // System.out.println(stirl[0][0]);
 // System.out.println(stirl[1][0]+" "+stirl[1][1]);
 // System.out.println(stirl[2][0]+"
"+stirl[2][1]+" "+stirl[2][2]);
 // System.out.println(stirl[3][0]+"
"+stirl[3][1]+" "+stirl[3][2]+" "+stirl[3][3]);
 // System.out.println(stirl[4][0]+" "+stirl[4][1]+"
 "+stirl[4][2]+" "+stirl[4][3]+" "+stirl[4][4]);
 }
 }
```

References

1. Trevors, J.T. and Abel, D.L., Chance and necessity do not explain the origin of life, *Cell Biol. Int.* **28**:729–739, 2004.
2. Truman, R. and Terborg, P., Genetic code optimisation: simplistic silly and senseless speculations, *J. Creation* **21**(2): 90–100, 2007.
3. Trifonov, E.N., Consensus temporal order of amino acids and evolution of the triplet code, *Gene* **261**:139–151, 2000.
4. Wong, J.T.-F., Co-evolution of genetic code and amino acid biosynthesis, *Trends Bioch. Sci.* **6**:33–36, 1981.
5. Wong, J.T.-F., Evolution of the genetic code, *Microbiol. Sci.* **5**:174–181, 1988.
6. Hartman, H., Speculations on the origin of the genetic code, *J. Mol. Evol.* **40**:541–544, 1995.
7. Arques, D.G. and Michel, C.J., A complementary circular code in the protein coding genes, *J. Theor. Biol.* **182**:45–58, 1996.

8. Xia, T., Santa Lucia, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X. and Cox, C., Turner, D.H., Thermodynamical parameters for an expanded nearest-neighbour model for formation of RNA duplexes with Watson-Crick base pairs, *Biochemistry* **37**:14719–14735, 1998.
9. Osawa, S., Jukes, T.S., Watanabe, K. and Muto, A., Recent evidence for evolution of the genetic code, *Microbiol. Rev.* **56**, 229–264, 1992.
10. Henikoff, S. and Henikoff, J.G., Amino acid substitution matrices for protein blocks, *Proc. Natl. Acad. Sci.* **89**:10915–10919, 1992.
11. Truman, R. and Terborg, P. Why the shared mutations in the hominidae exon X GULO pseudogene are not evidence for common descent, *J. Creation* 21(**3**):118–127, 2007.
12. Gitt, W., *Am Anfang war die Information*, hänssler Verlag, Germany, 2002.
13. Truman, R., and Heisig, M., Protein families: chance or design? *J. Creation* **15**(3):115–127, 2001; <www.creationontheweb.com/images/pdfs/tj/tjv15n3_protein_families.pdf>.
14. Truman, R., The ubiquitin protein: chance or design? *J. Creation* **19**(3):116–127, 2005; <www.creationontheweb.com/content/view/4346/>.
15. Truman, R., Searching for needles in a haystack, *J. Creation,* **20(**2):90–99, 2006.
16. Trifonov, ref. 3 p. 139.
17. Bergman, J., Why the Miller-Urey research argues against abiogenesis, *J. Creation* **18**(2):74–84, 2002; <www.creationontheweb.com/content/view/4111/>.
18. Simpson, S., Life's first scalding steps, *Science News* **155**(2):24–26, 1999, p. 26.
19. Flowers, C., *A Science Odyssey: 100 Years of Discovery*, William Morrow and Company, New York, p. 173, 1998.
20. Dufton, M.J., Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins? *J. Theor. Biol.* **187**:165–173, 1997.
21. Trifonov, ref. 3 p. 141.
22. Knight, R.D., Freeland, S.J. and Landweber, L.F., Rewiring the Keyboard: Evolvability of the Genetic Code, *Nature Reviews* **2**:49–58, 2001.
23. Knight *et al.*, ref. 22 p. 50.
24. Knight *et al.*, ref. 22 p. 55.
25. Knight *et al.*, ref. 22 p. 52.
26. Knight *et al.*, ref. 22 p. 51.
27. The Codon Usage Database shows the proportion of the 64 codons across 32 775 organisms (as of Oct. 15, 2006). See <www.kazusa.or.jp/codon/>
28. Program which calculates proportion of codons used for a known DNA coding sequence: <www.kazusa.or.jp/codon/countcodon.html>.
29. Castresana, J., Feldmaier-Fuchs and Pääbo, S., Codon reassignment and amino acid composition in hemichordate mitochondria, *Proc. Natl Acad. Sci. USA* **95**:3703–3707, 1998.
30. Santos, M.A., Cheesman, C., Costa V., Moradas-Ferreira, P. and Tuite, M.F., Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in *Candida* spp., *Mol. Microbiol.* **31**:937–947, 1999.
31. Alberts, B. *et al.*, *Molecular Biology of The Cell*, Garland Publishing, 3rd Edition, pp. 714–715, 1994.
32. There are about 17 alternative genetic codes, most of which involve mitochondria in different organisms. See <www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c#SG1>.
33. Ardell, D.H., On error minimization in a sequential origin of the standard genetic code, *J. Mol. Evol.* **47**:1–13, 1998.
34. Junck, J., The genetic code as a periodic table, *J. Mol. Evol.* **11**:211–224, 1978.
35. Yockey, H.P., *Information theory and molecular biology*, Cambridge University Press, Great Britain, 1992. See page 183.
36. There is a small error in his calculations. In using 64 codons he must include the three used as 'Stop.' The correct answer is actually 2.31628×10^{69} codes.
37. Schönauer, S. and Clote, P., How optimal is the genetic code? <www.dr-schoenauer.de/download/geneticCode.pdf>.
38. Graham, R.L., Knuth, D.E. and Patashnik, O., *Concrete Mathematics: A Foundation for Computer Science,* 2nd edition, Addison-Wesley, 1994. See Chapter 6.
39. More precisely, a certain nine-year-old boy who shall remain unnamed, otherwise known for his good judgement, turned the computer off after the program had been running for a week.
40. We also found 21! x $2.95572845518811 \times 10^{64}$ using the second, much faster algorithm.
41. Sanford, J.C., *Genetic Entropy & The Mystery of the Genome*, Ivan Press, Lima, New York, 2005.
42. Gu, X, The Age of the Common Ancestor of Eukaryotes and Prokaryotes: Statistical Inferences, *Mol. Biol. Evol.* **14**(8):861–866, 1997.
43. Eigen, M., Lindemann, B.F., Tietze, M., Winkler Oswatitsch, R., Dress, A. and Haeseler, A., How old is the genetic code? Statistical geometry of tRNA provides an answer, *Science* **244**:673–679, 1989.
44. Ouzounis, C.A., Kunin, V., Darzentas, N. and Goldovsky L., A minimal estimate for the gene content of the last universal common ancestor: exobiology from a terrestrial perspective, *Res. Microbiol*., **157**:57–68, 2006.
45. Knight, R.D., Freeland, S.J. and Landweber, L.F., Selection, history and chemistry: the three faces of the genetic code, *TIBS* **24**:241–247, 1999.
46. Each gene produces multiple mRNA copies, each of which produce multiple polypeptides.
47. Crick, F., The origin of the genetic code, *J. Mol. Biol.* **38**:367–379, 1968.
48 Freeland, S.J. and Hurst, L.D., The Genetic Code is One in a Million, *J. Mol. Evol.* **47**:238–248, 1998.
49. Truman, R., The race between genome truncation and mutational opportunity: can new genes arise naturalistically via gene duplication? Part 1, *J. Creation* 2007. In Press.
50. Truman, R. and Terborg, P., The race between genome truncation and mutational opportunity: can new genes arise naturalistically via gene duplication? Part 2 *J. Creation*, 2007. In Press.
51. Alberts, B. *et al.*, ref. 31, p. 715.

Royal Truman has bachelor's degrees in chemistry and in computer science from SUNY Buffalo, an MBA from the University of Michigan, a Ph.D. in organic chemistry from Michigan State University and post-graduate studies in bioinformatics from the universities of Heidelberg and Mannheim. He works for a large multinational in Europe.

Peer Terborg has an M.Sc. in Biology (Hons biochemistry and molecular genetics) and a Ph.D. in Medical Sciences from the University of Groningen, The Netherlands. He is currently working on the cellular and molecular aspects of pulmonary diseases, such as asthma and COPD, and is an expert on the molecular biology of signal transduction and gene expression.