Journal of
Theoretical & Computational Science

Review Article Open Access

Evolution of the Genetic Code – Some Novel Aspects

Jan C Biro*

Homulus Foundation. 612 S Flower Str. #1220. Los Angeles, CA 90017, USA

Abstract

Compilation of previously discussed main ideas about the evolution of genetic code is presented and completed with bioinformatical analyses of codon usage frequency data from 113 species.It is suggested that the recent 64/20 Genetic Code (Nirenberg) and the associated redundancy in translation developed successively from a much simpler, primitive Code containing only a few AT-rich codons. Codon boundaries were not yet defined so the codons were translated overlappingly. The subsequent addition of GC bases (especially those added at 1st and 3rd codon positions) provided the conditions for the physicochemical definition of codon boundaries and the development of non-overlapping translation. This view is supported by bioinformatics studies in the recent literature as well as by novel findings.

Keywords: Codon; Translation; Genetic code

Introduction

The Genetic Code has been known since 1962 [1]. It is largely universal, though some minor variations have been discovered. It is the logical connection between the nucleic acid and protein “worlds” [2]. Both nucleic acids and proteins have changed over evolutionary history. It is rational to assume that even the Genetic Code has an evolutionary history. Studies on the Genetic Code demand simultaneous and independent knowledge of the corresponding nucleic acid and protein sequences, for which data are usually not available. *De novo* sequencing of proteins has no scientific priority. An additional methodological difficulty is that the species that utilized early variants of Genetic Code may have been extinct for a long time, and ancient variants of proteins are no longer translated. The result of these difficulties is that studies on the evolution of the Genetic Code consist mainly of speculations that have very little chance of experimental confirmation or rejection. However, the history of the Genetic Code is not wholly beyond the reach of serious scientific study. This article reviews and analyzes the main theories about the evolutionary development of the Genetic Code and extends them with recently-discovered aspects of its function.

Results and Discussion

Assumptions

In this study we have made some necessary assumptions:

1. The development of the Genetic Code is, like any other biological development, a process from the simpler to the more complex. Therefore we accept the possibility that the present-day four-base-type nucleic acids developed from (say) two-base-type molecules; that the current triple code might have been preceded by a one- or two-letter code; and that the current 64 codons have developed from, say, four or sixteen codons.
2. Codons that have been in use for longer (in evolutionary terms) are numerically over-represented in the species-specific Codon Usage Tables (CUT). It is logical to suppose that newly-developed functions required new proteins, and coding for more proteins required a number of available amino acids and of codons that increased during the millions of years of biological evolution.
3. Some species disappear. This extinction of the “unfit” may (though not necessarily) mean the loss of all information related to their biology while they were extant. However, much of ancient biological history is preserved in modern organisms, often in hidden, non-expressed genomic sequences. It is well documented that the non-expressed part of the genome has grown (accumulated) rapidly during evolution and has reached huge proportions (~98% of the total DNA) in humans. (It is also observed that these structures sometimes become activated by mistake and this activation leads to immunodeficiency, cancerization or the appearance of bizarre body parts, a phenomenon called atavism [3]). This consideration suggests that the history of ancient proteins and the function of the primitive Genetic Code might be preserved in non-expressed genomic DNA.
4. Codons developed in a way that was compatible with the principle of base complementarity. Bases are known to form complementary pairs and nucleic acids are known to form complementary strands. Therefore, it seems inevitable that codon-anticodon pairs must have existed at every stage of codon development, and it has always been important that the meanings of codons and anticodons are not confused during translation.
5. Codons developed in close connection with their encoded amino acids. The specific (unique) interaction between nucleic acids and proteins is biologically as important and meaningful as the specific (unique) interaction between complementary nucleic acid strands. This assumption might be controversial and is known to have both supporters [4] and opponents [5] among influential scientists. However, our own development, *The Common Periodic Table of Codons and Amino Acids* [6], strongly supports it.

The structure of the codon

Nirenberg’s Genetic Code is redundant: 64 codons encode 20

***Corresponding author:** Jan C Biro, Homulus Foundation. 612 S Flower Str. #1220. Los Angeles, CA 90017, USA, E-mail: Jan.biro@att.net; Website: www.janbiro.com

Received October 31, 2013; **Accepted** December 16, 2013; **Published** December 24, 2013

Citation: Biro JC (2013) Evolution of the Genetic Code – Some Novel Aspects. J Theor Comput Sci 1: 107. doi: 10.4172/2376-130X.1000107

amino acids and a stop signal. The location of redundancy is the 3rd codon position, called the wobble base. This is the first indication of structure in the codons.

The second (central) codon bases are also clearly distinguished from the others (1st and 3rd). These central bases are undoubtedly related to the physico-chemical properties (charge, hydropathy and some structural aspects) of the encoded amino acids [6].

There is a readily detectable, periodic energy pattern along exons that is not detectable along intronic sequences. The free folding energy (-dG, Gibbs energy) is periodically lower (on average) in relation to the 1st and 3rd codon bases than the 2nd. This is possible only if G-C base pairs are preferentially located at the 1st and 3rd codon positions (also as the average distribution). This energy pattern provides a virtual physicochemical definition of codon boundaries [7].

The foregoing observations indicate that the 2nd codon letter is clearly distinguished from the 1st and 3rd, both structurally and functionally.

The possible origin of the Genetic Code

The literature is rather rich in ideas regarding the possible origin and development of the genetic code. It has been suggested that the recent Genetic Code developed from a primitive A+T-containing code [8], while others have found evidence for a primitive G+C-containing code [9]. We performed statistical analyses of Codon Usage Frequencies (CUFs) in several species in the hope of finding evidence for one or the other primitive code.

Genome-wide species sequencing projects have emerged only during the past 10 years and have provided reliable data for analysis of species-specific codon usage. These data are collected in numerous Codon Usage Tables (CUT) [10]. We examined codon usage frequencies in 113 species from different stages of evolutionary development, and found that codon usage is strongly biased in every species and shows a rather similar pattern in different organisms (Figure 1).

These large differences in codon usage frequencies not simply the result of differences in amino acid usage frequencies, which are to be expected, but are largely caused by differences in the usage of synonymous codons (codons encoding the same amino acid). Furthermore, the synonymous codon usage bias is about the same (fairly well conserved) in all species. Detailed analyses of the base compositions of codons reveal that the most frequent codons are preferentially built of A & T bases (Figure 2). The 13 most frequent codons (20% of the 64 possible) have A or T in the central position (the critical position in relation to the physicochemical properties of the encoded amino acid [6]) and provide 36% of all codon usages (100%).

The codon usage frequency pattern is strikingly similar in the seven major species categories. However, it is possible to detect some minor but statistically significant differences (Figure 3). The relative CUF pattern shows significant, systematic differences when the values of animals are compared to distant species categories such as viruses, protists and archaea. No such difference is found when the animal values are compared to phylogenetically closer relatives such as bacteria, fungi and plants.

The negative correlation between animals and archaea is most significant. This difference is clearly related to the base compositions of the codons: AT-rich codons are preferentially used by archaea, while animals prefer GC-rich codons (Figure 4).

The amino acid usage frequencies of animals and archaea are clearly different, but this difference does not correlate with the AT content of the synonymous codons and therefore fails to explain the AT-related differences in CUF patterns (Figure 5).

The dominance of A and T in the most frequently used codons, and the synonymous codon usage bias in favor of AT-rich codons in older rather than younger species, suggest to us that primitive codons were built of A and T bases, while G and C came into use during a later, second developmental stage.

The next important question about the primitive Code is the number of bases necessary in the codons. The modern triplet codon provides 64 different combinations of the four nucleic acid bases, far more than necessary to encode 20 amino acids. Two singlet codons (A and T) can theoretically encode two amino acids. The coding rules in the modern Common Periodic Table of Codons and Nucleic Acids [6] suggests that T codes for one hydrophobic and A for one hydrophilic or charged amino acid, say T>Phe [now encoded by TTT]

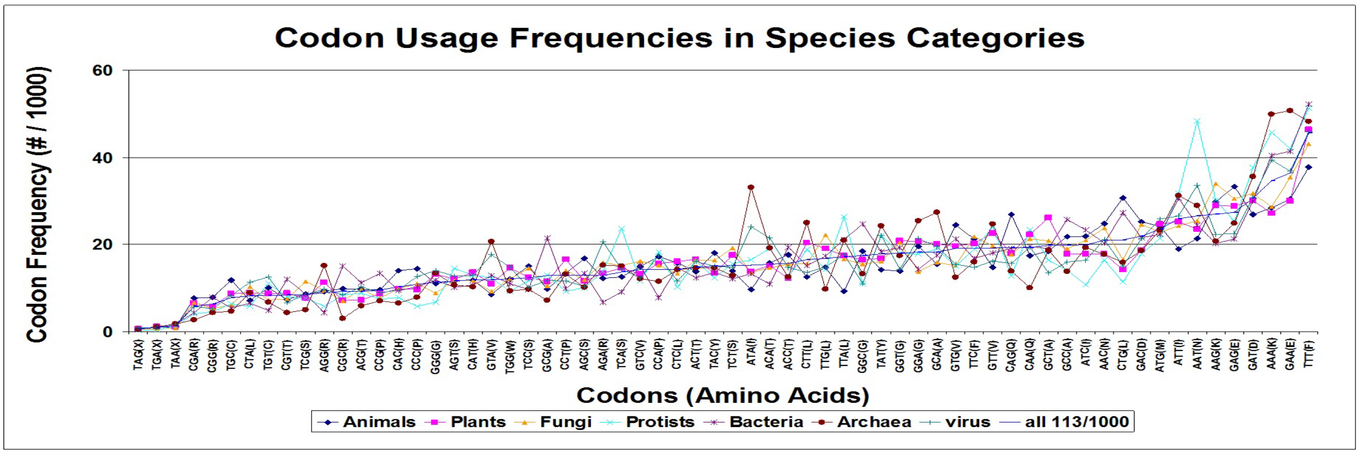


Figure 1: Codon Usage Frequencies in Species Categories. Codon usage frequency data were collected for 113 species and combined into seven species categories. Each symbol represents the mean codon usage frequency of the given codon in the indicated species category. [n= 25 (animals), 20 (plants), 18 (fungi), 8 (protists), 24 (bacteria), 9 (archaea), 9 (viruses).].

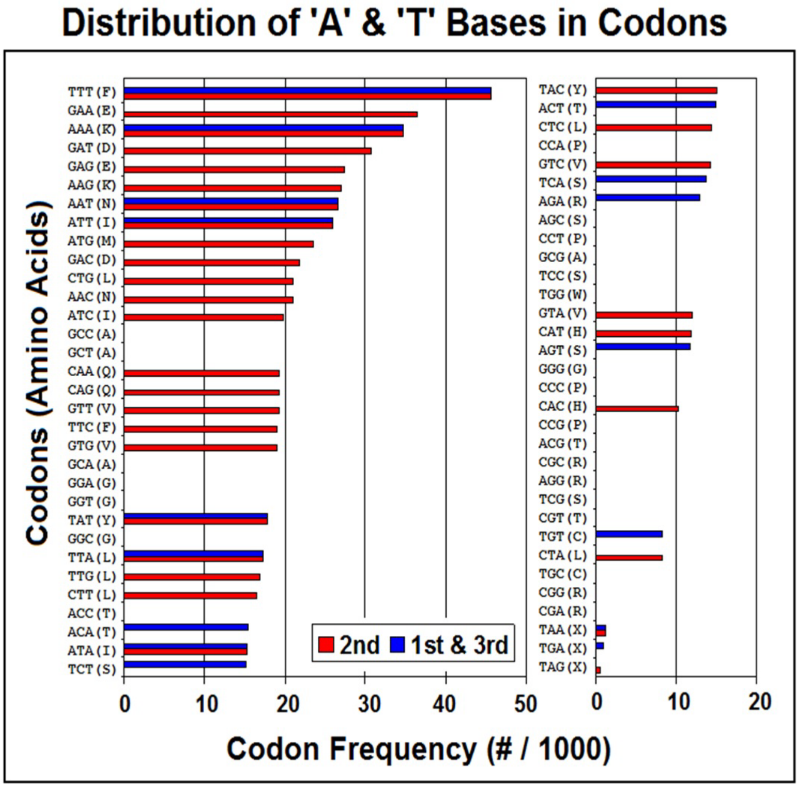


Figure 2: Distribution of 'A' & 'T' Bases in Codons. CUF data were collected for 113 species and the mean values were sorted in descending order. A or T bases at the 2nd as well as at the 1st and 3rd codon positions are indicated by colors.

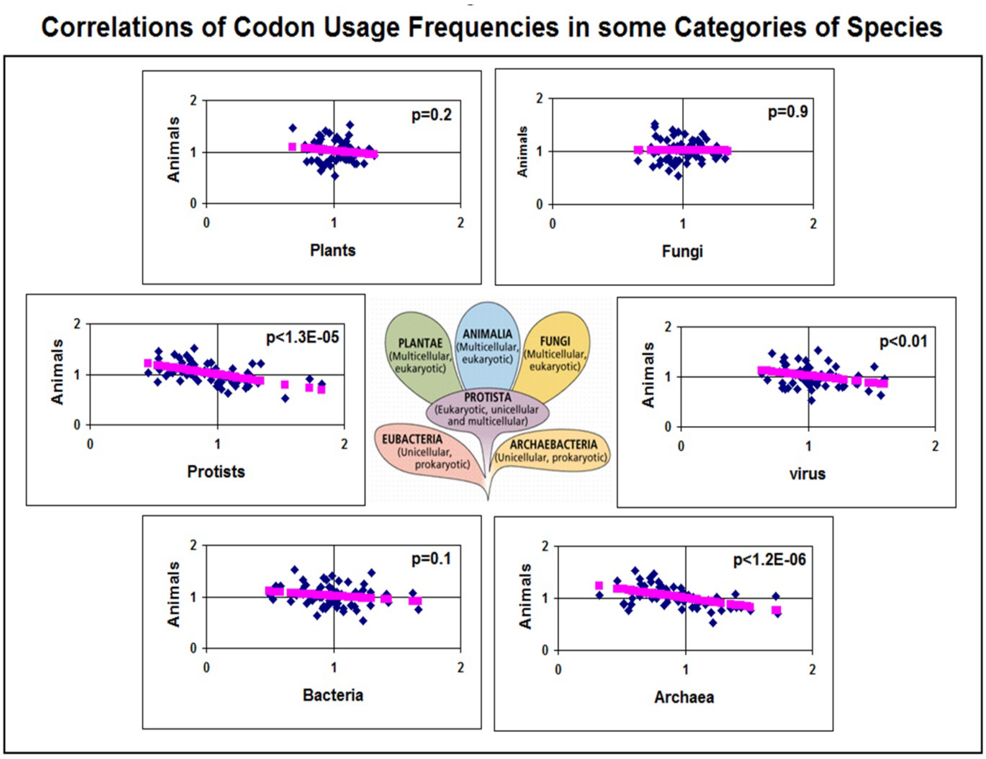


Figure 3: Correlation of Codon Usage Frequencies in some Categories of Species. The relative CUF values were calculated for the 64 codons from the data in Figure 1 and plotted against each other (blue symbols). Red symbols indicate the linear regression lines.

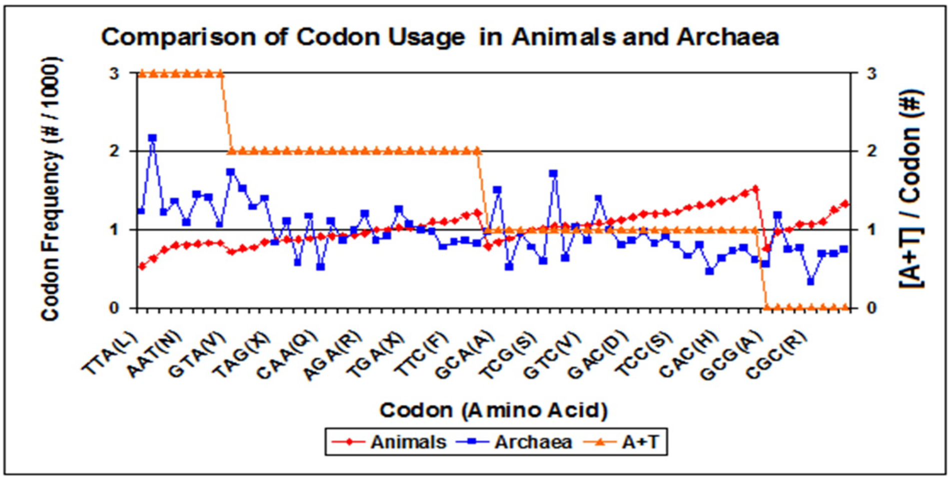


Figure 4: Comparison of Codon Usage in Animals and Archaea. Codons were sorted in descending order of A and T content and the relative CUF data of animals and archaea were plotted. Data are from Figure 3.

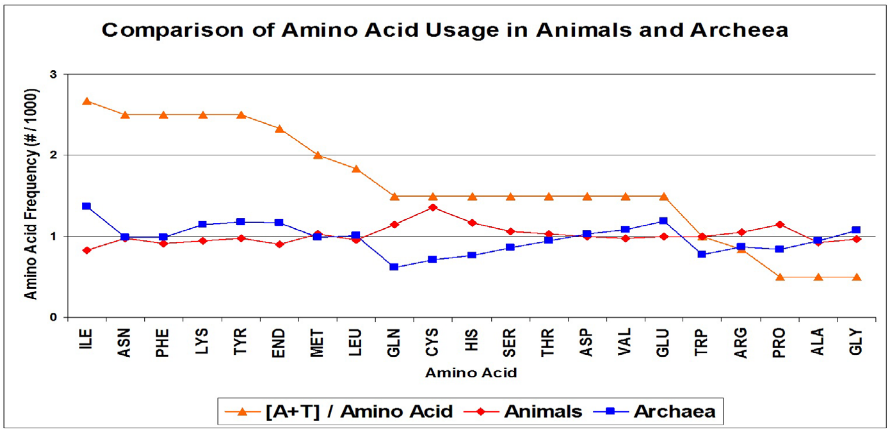


Figure 5: Comparison of Amino Acid Usage in Animals and Archaea. The 64 relative CUF data from Figure 4 were combined into 20 relative amino acid usage frequency values by calculating the means of the corresponding synonymous CUF values.

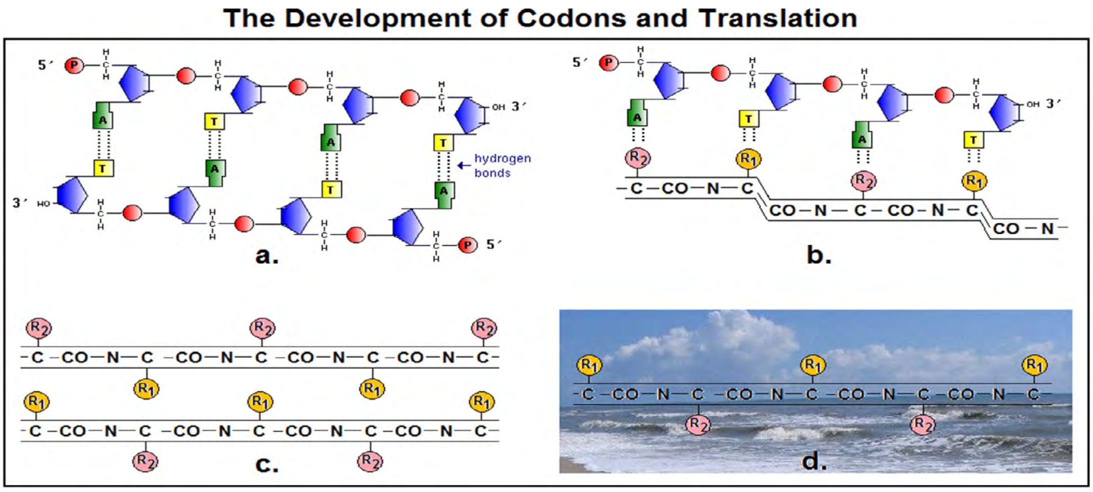


Figure 6: The Development of Codons and Translation. The primitive, ancient nucleic acids were composed only of A and T bases (a), which encode and interact with only two amino acids (R1 and R2, b). The primitive, ancient oligopeptides were physicochemically compatible with each other and may have interacted specifically with each other (c). The peptide and/or nucleic acid complexes would have become located preferentially on hydrophilic/hydrophobic surfaces, forming molecular layers.

and A>Lys [now encoded by AAA]. Thus, the hypothetical primitive $[FK]_n$-type oligopeptides and $[AT]_n$-type oligonucleotides may have formed a nucleo-peptide complex (the positively charged K attracting the negatively charged nucleic acid), which would preferentially have become located on aqueous boundary surfaces, forming a layer (primitive membrane?) (Figure 6).

The next step toward the complexity of the recent Genetic Code might have been recognition of the order (sequence) of bases. Bases in the order ATAATA form a different shape from, for example, TATTAT, and this shape difference might have been utilized to distinguish between two amino acids (regarding their coding as well as their binding). There is strong evidence suggesting that codons and amino acids developed in parallel (co-evolution) [4] and there is a significant connection between the physicochemical properties of the amino acids and their codon structures [11]. Therefore, the development of a triplet code that still utilized only A & T seems a rather "logical" possibility.

This triplet code would have added two more apolar amino acids, Leu [UUA] and Ile [AUU], and two polar amino acids, Tyr [UAU] and Asn [AAU], to the protein building units. A recent stop codon [UAA] might also have developed at this time as codon for Gln (still used in ciliate and flatworm mitochondrial codes [12]).

The development of this triplet code immediately raises the question of translation reading frames: where are the beginning and the end of a triplet codon? We assume that the codon boundaries were not yet defined in the primitive code; AT-only triplet codons were overlappingly translated (Figure 7).

The idea of overlapping translation goes back to the 1950s. In the years immediately following the proposal of the structure for dsDNA, George Gamow suggested a so-called "diamond code" to explain the connection and information transfer between DNA and proteins [13-15]. In his model the nucleic acid bases from 20 different cavities into which the 20 different amino acids fit specifically. The order of cavities determines the order (sequence) of encoded amino acids, which polymerize and form the individual proteins. Gamow's model was the very first model for translation and it turned out to be overlapping, which means that the 2nd and 3rd bases of a triplet codon are identical to the 1st and 2nd bases in the next triplet codon, so amino acid neighbors are interdependently encoded. The attractive feature of the overlapping codon model is that it takes advantage of an interesting structural similarity between amino acids and nucleic acids, namely that the distance between the amino acids and the distance between nucleotide bases is the same, which strongly suggests a connection and 1:1 relationship between these very different residues. In addition, the "frame shift" problem does not arise in overlapping translation. A big disadvantage of this model, which turned out to be "fatal", is that it simply doesn't permit some amino acid neighbors that do exist in real proteins [16]. Gamow's model is still revisited time after time as a way of avoiding frame-shift problems when no other way is apparent. We know today that codon boundaries are physicochemically defined in modern codons by the periodic distribution of GC bases [7], which was not the case in the AT-only model described above.

There is no overlapping translation in recent or modern organisms (as far as we know), but signs that it once existed might have been preserved. When we look for such molecular, phylogenetic "fossils" it is important to bear in mind that biological history is preserved in the form of DNA (not protein) and historical DNA records are located in the non-translated DNA domains (exons, and even in regions called "junk" DNA) [17].

Suppose that overlapping translation did exist in the past, but at a certain point in evolution it was replaced by the now-practiced non-overlapping translation. In that case, some nucleic acid sequences might exist in two different forms with the same translational meaning (protein sequence): one "compact", which was overlappingly translated in the past, and one "*extended*", which is nowadays translated non-overlappingly (*Extended OTS*). A third category of nucleic acid sequences, which developed later, comprises those that cannot be compressed into OTS and are called *Extended non-OTS* (Figure 8).

To test this idea we constructed 64 polycodon frequencies, each corresponding to one codon repeated 10 times. We were looking for the incidence of these simple monotone repeats in the Nucleotide Sequence Databases, provided by the Blast server of NCBI [18] (which contains all GenBank + EMBL + DDBJ + PDB sequences, but no EST, STS,

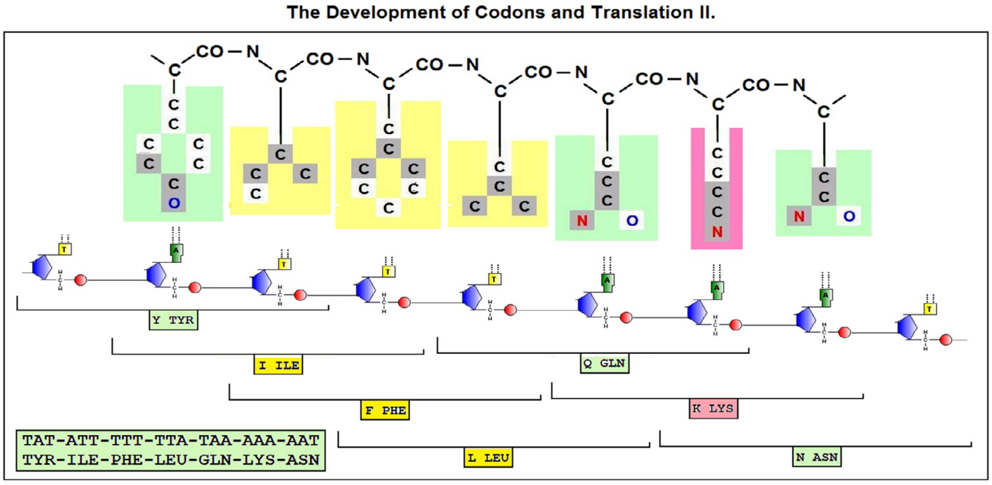
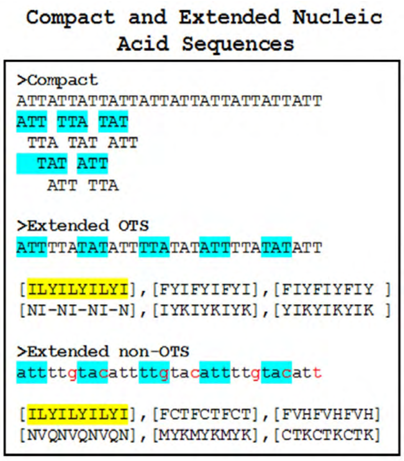

The Development of Codons and Translation II.

Figure 7: The Development of Codons and Translation II. The A- & T-triplet codons might have encoded seven amino acids (indicated) representing hydrophobic (yellow boxes), hydrophilic (green boxes) and positively charged (pink boxes) properties. The codons (and amino acids) have been sorted into a special order to illustrate the principle of overlapping translation.

GSS, environmental samples or phase 0, 1 or 2 HTGS sequences). The database contained 26,409,867,045 letters, corresponding to 8,281,433 sequences, on February 26, 2009. Nucleic acid similarity searches were performed using BLASTN 2.2.19 provided by the NCBI server (default settings) [19]. Surprisingly many sequences were found that were significantly similar to the polycodon-like query oligonucleotides, and their frequencies differed depending on the base composition of the query (Figure 9). This base composition-dependent distribution recalled what we found in the CUF tables, i.e. AT-rich sequences (codons) were once more frequently used than GC-rich codons (sequences) (see figures 2 and 4).

In the next step we assumed - very arbitrarily - that these highly frequent polycodon sequences represent compact, primitive nucleic acids that could be extended and translated overlappingly as well as non-overlappingly (as shown in figure 8). At about 1/3 of the compact sequences were found even in the extended OTS forms, especially those derived from AT-rich sequences (Figure 10). The extended non-OTS forms had the lowest frequency, and this was independent of base composition.


Compact and Extended Nucleic Acid Sequences

```
>Compact
ATTATTATTATTATTATTATTATTATTATT
ATT TTA TAT
 TTA TAT ATT
  TAT ATT
   ATT TTA

>Extended OTS
ATTTTATATATTTTATATATTTTATATATT

[ILYILYILYI],[FYIFYIFYI],[FIYFIYFIY ]
[NI-NI-NI-N],[IYKIYKIYK],[YIKYIKYIK ]

>Extended non-OTS
atttgtacatttgtacatttgtacatt

[ILYILYILYI],[FCTFCTFCT],[FVHFVHFVH]
[NVQNVQNVQN],[MYKMYKMYK],[CTKCTKCTK]
```

Figure 8: Compact and Extended Nucleic Acid Sequences. There are three categories of nucleic acid sequences. The oldest form is the *compact* sequence, in which codon boundaries cannot yet be recognized, so it is translated overlappingly (OTS). *Extended* OTS sequences developed from compact sequences and made it possible to read the OTS non-overlappingly. *Extended non-OTS* sequences are late developments and they cannot be compressed into compact forms. Codons are indicated by blue boxes. The difference between extended OTS and non-OTS is indicated by red letters. Note that the translation of extended OTS and non-OTS sequences into protein sequences is the same (yellow boxes).

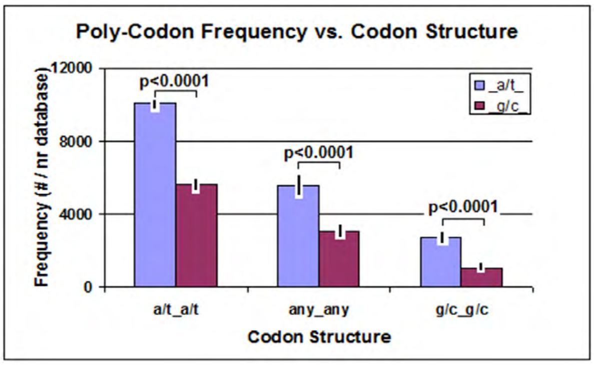


Figure 9: Polycodon Frequency vs. Codon Structure. Nucleotide Sequence Databases (NCBI) were searched by 64 different queries each containing 10x repeats of the 64 possible codons and the BlastN matches were automatically counted (default server settings). The collected data were divided into six groups corresponding to the composition of the 1st and 3rd codon positions (x axis) and the 2nd codon position (colors). Each column represents the mean ± S.D. of a total of n= 64 counting.

These findings suggest that codon-like repeats (especially AT-rich) played a significant role in the genome and were therefore preserved. They represent compact sequences from the early period of codon development, which could be overlappingly translated. However, these sequences successively lost their importance (translation?) with the development of GC-containing codons and the shift to non-overlapping translation.

Our recent and previous experiments provide support for the ideas of Jimenez-Sanchez [8], who suggested that the recent genetic code developed from a simpler, AT-only code, and G and C were added in a second, later step of development. However, an AT-only-containing nucleic acid cannot be dissected into well-defined triplet codons, so frame-shifts have to exist and the overlapping translation of codons is unavoidable. These simple nucleic acids and their translation have, of course, significant limitations for the development of biological functions and life as we know it today. Another concern regarding the AT-only codon is the absence of any experimental evidence. The AT-only code bearers might have been disappeared. More likely, the idea of Jimenez-Sanchez is an extreme and theoretical extrapolation of the biological reality, that older species contains more AT, while younger species more GC bases in their genome.

The situation changed dramatically with the addition of G and C bases. This addition increased the number of possible codons to 64, provided the possibility of high energy signatures along the nucleic acid sequences (physicochemical definition of codon boundaries [7]) and made it possible to shift from overlapping translation to the recent, more permissive, non-overlapping variant.

An alternative evolutionary model was proposed by Ikehara et al. [9], who emphasized the importance of GC at the 1st and 3rd codon positions. We completely agree with Ikehara's statements about the special significance of G and C in defining codon boundaries. However, our recent analyses of CUF tables clearly show the dominance of AT-rich codons in terms of frequency of use. This indicates that AT-rich codons have an evolutionary importance of their own.

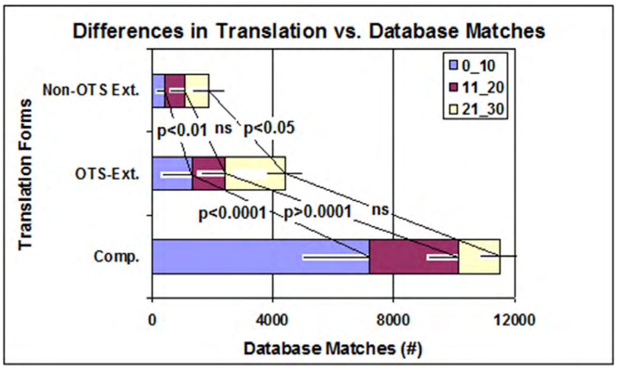


Figure 10: Differences in Translation vs. Database Matches. Nucleic acid databases were searched for matches with specially designed queries (see figure 8): compressed (Comp.), overlappingly translated and extended (OTS-Ext.) and Non-OTS (Ext). All three forms could be translated into the same protein sequence but in different ways. The GC content (#/30) is color-coded. Columns represent the mean ± S.D. The significance of differences between groups is indicated on the connecting lines.

The involvement of overlapping translation in our model of codon evolution solves the obvious problem of frame shifts, but creates new concerns at the same time. It might be difficult to understand how the transition to non-overlapping translation could have happened without serious conflict between the two systems.

Evolutionary questions often have philosophical aspects. Biological sciences recently tend to picture evolution as an uninterrupted, continuous, linear process. However it is not known whether the biological evolution that scientists are able to see and observe on the Earth "is the" biological evolution or it is only a local variant of a greater biological evolution in the Universe. Francis C. Crick, the founding father of molecular biology, lunched the idea of panspermia [20]. It suggested, that life developed somewhere in the universe and spread in the cosmos, even to the Earth, as DNA trapped in cosmic debris. In this case only a short part of biological evolution is available for us; we can see trends, suggest necessary events (like the AT-only nucleic acids [8] and overlapping translation [13-15]) without the possibility to find remains of these evolutionary steps.

It is concluded that the well-known triplet codons and the 64/20 translation is a complex system that is the result of successive development from much simpler systems, like AT-only codons (which were coding only a few amino acids) and overlapping translation. The evolutionary "addition" of GC nucleotides was necessary to define the recent codon-structure that physicochemically marks codon boundaries and makes the more sophisticated, non-overlapping, translation possible.

Acknowledgement

The author wish to thank to Professor Kenji Ikehara for a very stimulating personal discussion (Los Angeles, February 26th, 2009).

References

1. Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A 47: 1588-1602.
2. Leder P, Nirenberg M (1964) RNA Codewords and Protein Synthesis. II. Nucleotide Sequence of a Valine RNA Codeword. Proc Natl Acad Sci U S A 52: 420-427.
3. Bergman J (2001) The Functions of Introns: From Junk DNA to Designed DNA. Perspectives on Science and Christian Faith 53.
4. Woese CR (1967) The Genetic Code: The Molecular Basis for Gene Expression. Harper & Row, New York.
5. Crick FH (1968) The origin of the genetic code. J Mol Biol 38: 367-379.
6. Biro JC, Benyó B, Sansom C, Szlávecz A, Fördös G, et al. (2003) A common periodic table of codons and amino acids. Biochem Biophys Res Commun 306: 408-415.
7. Biro JC (2006) Indications that "codon boundaries" are physico-chemically defined and exons contain even protein-folding information in the redundant Genetic Code. Theor Biol Med Model 3: 28.
8. Jiménez-Sánchez A (1995) On the origin and evolution of the genetic code. J Mol Evol 41: 712-716.
9. Ikehara K, Omori Y, Arai R, Hirose A (2002) A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. J Mol Evol 54: 530-538.
10. Codon Usage Database (2007) NCBI-GenBank Flat File Release 160.0.
11. Biro JC, Biro JM (2004) Frequent occurrence of recognition site-like sequences in the restriction endonucleases. BMC Bioinformatics 5: 30.
12. Andrzej (Anjay) Elzanowski, Jim Ostell (2008) The Genetic Code. National Center for Biotechnology Information (NCBI).
13. Gamow G, A Rich, M Ycas (1955) Advances in Biological and Medical Physics, Vol. 4. Academic Press, New York.
14. Gamow G (1954) Possible Relation between Deoxyribonucleic Acid and Protein Structures. Nature 173: 318.
15. Kgl Danske Videnskab Seltkab (1954) Biol Medd 22.
16. Brenner S (1957) On the Impossibility of All Overlapping Triplet Codes in Information Transfer from Nucleic Acid to Proteins. Proc Natl Acad Sci U S A 43: 687-694.
17. Ohno S (1972) So much "junk" DNA in our genome. Brookhaven Symp Biol 23: 366-370.
18. Basic Local Alignment Search Tool (BLAST) National Center for Biotechnology Information (NCBI), National Library of Medicine National Institutes of Health.
19. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
20. Crick FH, Orgel LE (1973) Directed Panspermia. Icarus 19: 341-348.