Recent emergence of the modern genetic code: a proposal

Michael Syvanen

This article proposes that the genetic code was not fully formed before the divergence of life into three kingdoms. Rather, at least arginine and tryptophan evolved after the diversification of archaea, bacteria and eukaryotes, and were spread by horizontal gene transfer. Evidence for this hypothesis is based on data suggesting that enzymes for biosynthesis of arginine and tryptophan, and for arginine tRNA ligase, have shorter divergence times than the underlying lineages. Also, many of these genes display 'star' phylogenies. This proposal is an extension of the idea that the genetic code was unified because of the evolutionary pressure from horizontal gene transfer. These considerations further undermine the need to postulate the existence of a 'last common ancestor'; a simpler model would be that multiple lineages gave rise to life today.

Phylogenetic trees based on genes for ribosomal RNA and proteins show that archaea and eukaryotes are more closely related to one another than either is to bacteria. Among the more striking observations from the genome-sequencing projects are the numerous genes whose phylogeny appears at odds with the phylogeny based on ribosomal RNA. These genes are often interpreted as examples of horizontal gene transfer [1–5]. For some classes of genes, it appears that most of the trees are incongruent with the universal tree; this is particularly true for those genes responsible for the biosynthesis of amino acids, nucleotides and amino acid tRNA ligases. There was a second unexpected finding: the age of the 'last common ancestor' (LCA) is much younger than had been thought.

This article will illustrate some of these points, focusing on the biosynthetic genes for tryptophan and arginine, and their amino acid tRNA ligases, and relating their divergence times to ideas on the evolution of the genetic code. These considerations raise the possibility that the genetic code continued to evolve after the diversification of the three major kingdoms: Archaea, Bacteria and Eukaryota.

Age of the last common ancestor

In 1996, Doolittle *et al.* [6] suggested that archaea, bacteria and eukaryotes diverged only about 1800–2200 million years (Myr) ago. This was based on examining 57 different genes from the three kingdoms for which calibration points were available, and then averaging the divergence times. The Doolittle study has been confirmed by another study that determined the age of 13 genes and obtained an average divergence time of 2200 Myr [7]. However, there is abundant evidence for the presence of bacterial life dating back 3800 Myr. In fact, on the basis of clock studies of paralogous genes within species, the age of the gene duplication event supports the age of 3000–4000 Myr [7,8]. Feng *et al.* [9], after accounting for horizontal gene transfers, have pushed the divergence time further back to about 3000 Myr.

Nevertheless, there are genes in all three kingdoms that appear to be much younger than these divergence times (Table 1). In this article, a novel interpretation is offered for these numbers, questioning a major assumption that affects all efforts to estimate either the nature or time of the LCA. In the past, it has been assumed that, if all three kingdoms share an orthologous gene, for example tryptophan tRNA ligase, then the LCA also contained that gene. And it was further assumed that the measured differences in age were simple statistical variations of the molecular clock about some common mean. However, the recent appreciation that horizontal gene transfer between the kingdoms has been common renders this assumption questionable. If this assumption were incorrect, then it would be an error to average the times of divergence for different genes. However, this could mean that the measured age of individual genes is more reliable than previously believed.

Star phylogenies

The most parsimonious phylogenetic tree for several of the genes listed in Table 1 takes the shape of a star. A star phylogeny is encountered when eukaryotes, remotely related archaea and remotely related bacteria are all approximately equally related to one another [10]. Figure 1a shows a typical pattern for arginine tRNA ligase. The star in this example arises from the relatively short internal branches compared with the much longer external branches. Furthermore, the arrangement of clades around those internal branches defies any taxonomic scheme. The amino acid tRNA ligases generally display phylogenies indicating that they have been involved in numerous horizontal gene transfers [11–14]. In addition to being involved in horizontal gene transfer, amino acid tRNA ligases appear among the unusually young genes listed in Table 1.

It is relevant to point out the unusual pattern of tryptophan tRNA ligase evolution (Fig. 1b). There are two unusual features in this tree. First, the bacteria are divided into two very remotely related groups that are unrelated to any known phylogenies. Second, there is a very long internal branch separating the genes from bacteria and mitochondria from those of eukaryotes and archaea. Indeed, it has been proposed that this ligase is biphyletic and that it evolved on two separate occasions from tyrosine tRNA: once in the line leading to bacteria and once in the line leading to eukaryotes [15]. This interpretation of the data was questioned by Brown *et al.* [16] who used a different strategy for the multisequence alignments of the tryptophan and tyrosine tRNA ligases. Resolution of this issue requires that the tyrosine tRNA ligase be properly rooted to the tree in Fig. 1b.

Michael Syvanen
Medical Microbiology and Immunology, 3146 Tupper Hall, School of Medicine, University of California, Davis, CA 95616-8645, USA.
e-mail: msyvanen@ucdavis.edu

 0168-9525/02/$ – see front matter PII: S0168-9525(02)02647-1

Table 1. Age of some genes found in Bacteria, Archaea and Eukaryota[a]

	Age ($\times 10^9$ years)	Ref.
Threonine tRNA ligase	2.01	[7]
Valine tRNA ligase	1.25	[7]
Isoleucine tRNA ligase	2.99	[7]
Arginine succinate lyase	1.48	[7]
Ornithine transcarbamylase	1.4	[10]
Arginine tRNA ligase	1.2–2	see Fig. 1

[a]The ages are determined from protein distances and molecular clock assumptions. In no case are constant molecular clocks assumed in the various lineages. The large variation in the age of arginine tRNA ligase depends on whether the plant–metazoan or the fungi–metazoan diversification is used to calibrate the trees.

Besides the tryptophan and arginine tRNA ligases, it appears as though most of the genes involved in the biosynthesis of arginine from ornithine, and the biosynthesis of tryptophan, display star phylogenies [10]. They also appear to be over-represented among the young genes listed in Table 1. (It should be noted that none of the tryptophan biosynthetic genes are in Table 1 because of the absence of good calibration points.)

A star phylogeny can mean one of two things. Either the gene found in these multiple lineages diverged at a single time, or the gene has become so highly diverged in the remotely related lines that phylogenetic information is lost and the deep internal branches cannot be resolved. Although this latter explanation might seem reasonable, paradoxically many of the genes that give rise to a star phylogeny are neither highly diverged, nor do they appear to be unusually functionally constrained, thus it is difficult to see why phylogenetic information would have been lost. A recent paper has documented these points with the genes for ornithine transcarbamylase and tryptophan synthase [10]. This article will consider the explanation that the star phylogeny arises because these genes diverged well after the diversification of the underlying lineages. These genes then spread by horizontal transfer, with the young age recording the time of this spread.

Evolution of the genetic code

In an earlier paper, I suggested that the unity of the genetic code was the outcome of an evolutionary mechanism in which horizontal gene transfer was a major factor [17]. In the early 1980s, there were two competing explanations for the unity of the code. One posited that the unity of the genetic code was the result of functional constraints such that certain amino acids would only fit, at the mechanistic level, with certain codons [18–20]. The second notion was that all life descended from a single interbreeding population that had today's code. This was referred to as the 'frozen

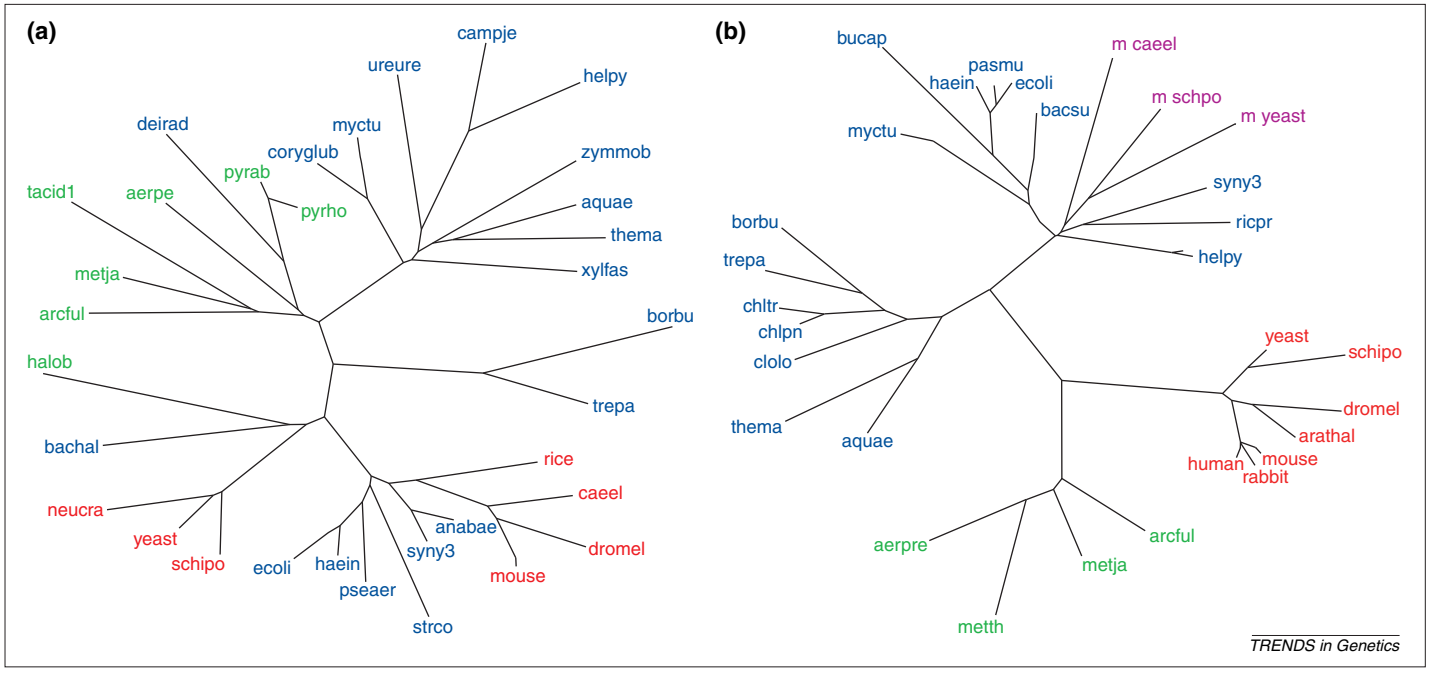

Fig. 1. Two tRNA ligases showing star or near star phyologenies. The age of a selection of genes found in archaea, bacteria and eukaryotes. Green, archaea; blue, bacteria; red, eukaryotes; purple, mitochondrial genes (m). (a) Arginine tRNA ligase distance tree. Distances were determined from multiply aligned sequences according to the Kimura protein distance method and nearest neighbor trees were determined. (b) Tryptophan tRNA ligase distance tree. The ages are determined from protein distances and molecular clock assumptions. In no case are constant molecular clocks assumed in the various lineages. The large variation in the age of arginine tRNA ligase depends on whether the plant–metazoan or the fungi–metazoan diversification are used to calibrate the trees. Abbreviations of species represented are: aerpe, *Aeropyrum pernix*; anabae, *Anabaena sp.* 90; aquae, *Aquifex aeolicus*; arathal, *Arabidopsis thaliana*; arcful, *Archaeoglobus fulgidus*; bachal, *Bacillus halodurans*; bacsu, *Bacillus subtilis*; borbu, *Borrelia burgdorferi*; bucap, *Buchnera aphidicola*; caeel, *Caenorhabditis elegans*; campje, *Campylobacter jejuni*; chltr, *Chlamydia trachomatis*; chlpn, *Chlamydophila pneumoniae*; clolo, *Clostridium longisporum*; coryglub, *Corynebacterium glutamicum*; deirad, *Deinococcus radiodurans*; dromel, *Drosophila melanogaster*; ecoli, *Escherichia coli*; haein, *Haemophilus influenzae*; halob, *Halobacterium* sp. NRC-1; helpy, *Helicobacter pylori*; metth, *Methanobacterium thermoautotrophicum*; metja, *Methanococcus jannaschii*; mouse, *Mus musculus*; myctu, *Mycobacterium tuberculosis*; neucra, *Neurospora crassa*; rabbit, *Oryctolagus cuniculus*; rice, *Oryza sativa*; pasmu, *Pasteurella multocida;* pseaer, *Pseudomonas aeruginosa*; pyrab, *Pyrococcus abyssi*; pyrho, *Pyrococcus horikoshii*; ricpr, *Rickettsia prowazekii*; yeast, *Saccharomyces cerevisiae*; schipo, *Schizosaccharomyces pombe*; strco, *Streptomyces coelicolor*; syny3, *Synechocystis* PCC6803; tacid1, *Thermoplasma acidophilum*; thema, *Thermotoga maritima*; trepa, *Treponema pallidum*; ureure, *Ureaplasma urealyticum*; xylfas, *Xylella fastidiosa*; zymmob, *Zymomonas mobilis*.

Box 1. Informational suppressors and the genetic code

Translation of information on mRNA into protein is not a perfect process and can result in errors. Mutations that enhance the translational error rate, and even change relative specificity are called the informational suppressors (Table I). In all of these cases, a cell containing the suppressor will either recognize a normal codon but will insert an incorrect amino acid in its place, or will recognize a 4-bp or 2-bp sequence as a triplet. In addition, bacteria carrying these mutations have growth advantages over the wild-type bacteria depending on selective media. Of course, in the absence of selection, suppressor mutations can cause clear growth deficiencies when compared with wild-type strains, although continued growth under conditions that select for the suppressed gene often result in additional mutations that compensate for these deficiencies. It seems likely during the evolution of the code that a similar process occurred as new specificities were added. Perhaps current efforts to construct strains of *Escherichia coli* with radically altered genetic codes are mimicking a pathway similar to the emergence of the modern genetic code. For examples of efforts to construct such strains of *E. coli*, see Refs [a–d].

Table I. A partial list of the types and translation gene products affected by informational suppressors[a]

	Informational suppressor	Mutation in or affecting
Nonsense	UGA, UAA, UAG	Numerous tRNAs
	UGA, UAA, UAG	ssu rRNA
	UGA	lsu rRNA
Missense	General	Elongation factor 1
	General	rsp3 (yeast), rpl20 and rpl35 (*E. coli*)
	Q→W,Q→R	tRNA Q
	Start codon	Initiation factor 1, rps18 (yeast)
		Methyl folate accumulation
Frameshift	Plus or minus	tRNAs and ssu rRNA

[a]Abbreviations: lsu, large subunit; rpl, ribosomal protein, large subunit; rps, ribosomal protein, small subunit; ssu, small subunit.

References

a Liu, D.R. and Schultz, P.G. (1999) Progress toward the evolution of an organism with an expanded genetic code. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4780–4785

b Ohno, S. *et al.* (1998) Co-expression of yeast amber suppressor tRNATyr and tyrosyl-tRNA synthetase in *Escherichia coli*: possibility to expand the genetic code. *J. Biochem.* (*Tokyo*) 124, 1065–1068

c Furter, R. (1998) Expansion of the genetic code: site-directed p-fluoro-phenylalanine incorporation in *Escherichia coli*. *Protein Sci.* 7, 419–426

d Yan, W. *et al.* (1996) A tRNA identity switch mediated by the binding interaction between a tRNA anticodon and the accessory domain of a class II aminoacyl-tRNA synthetase. *Biochemistry* 35, 6559–6568

accident'. I proposed an entirely different idea. If horizontal gene transfers were a major factor in evolution, then lineages that could decipher genes from foreign hosts would have a selective advantage, and any lineage that lost the ability to read foreign genes would be at a disadvantage. Applying this principle over all lineages would ensure that all would have the ability to read foreign genes. Thus, the unified genetic code was actively selected. It is interesting to note that, since this proposal was first made, several organisms have been described that have different genetic codes. However, other than mitochondrial codes, all of these exceptional organisms retain the ability to translate foreign genes, even though their genes might be indecipherable to foreign hosts [21]. These findings are consistent with the original argument and this article suggests an extension to this idea.

It is proposed here that the united genetic code is the product of an evolutionary process that has continued since the diversification of the major kingdoms. Specifically, it is proposed that the LCA (which defines the origins of the three kingdoms Archaea, Bacteria and Eukaryota) did not use arginine and tryptophan.

In the 1970s, several papers appeared that attempted to reconstruct the evolutionary pathway which formed the genetic code. These speculative pieces imagined more-primitive codes that had fewer amino acids to which extra amino acids were added [22–27]. Several points emerged from these speculations, two of which are relevant here: it was suggested that, first, tryptophan was one of the last amino acids added and, second, arginine occupied a curious position within the table of the genetic code. The hypothesis concerning tryptophan is the simplest to understand; it is postulated that there was a simpler code that had tyrosine but not tryptophan. A duplication of the tyrosine tRNA ligase gene occurred and one paralog evolved an affinity for tryptophan. Either a primitive stop codon or one of the ancestral codons of tyrosine was then recruited for tryptophan.

On the basis of the position of the arginine codons within the table of the genetic code, Jukes [28,29] argued that arginine was added late and that it replaced a more-primitive amino acid (the 'intruder' hypothesis). In addition, Jukes proposed that the amino acid replaced by arginine was ornithine. Of course, the proposal that arginine and tryptophan were added late was made in the context of a lineage leading to a LCA. However, given that we now accept that horizontal gene transfer occurred often during the early evolution of the Archaea, Bacteria and Eukaryota, there is no reason to assume the genetic code could not have continued to evolve. Indeed, studies of informational suppressors (Box 1) have revealed that the genetic code is quite flexible and the notion of it being 'frozen' is outdated.

We must ask the question: is it even reasonable to suggest that the genetic code changed through the action of horizontal gene transfer and natural selection? A change in the genetic code would affect the expression of all of the genes in a cell. In addition, one could assume that multiple genes would have to be transferred to effect a change in the genetic code. Let us consider each of these problems separately. First, work on informational suppressors has shown that the genetic code can be changed (Box 1), even if expression of hundreds of genes is influenced. This is not a problem unique to my suggestion, but one that confronts any scenario for the sequential evolution of the modern genetic code. Second, the number of genes that need to be transferred is not really that large. If we consider Juke's intruder hypothesis for arginine [28,29], a more-primitive code exists that uses ornithine; thus, in this case, there are preexisting codons for a charged amino acid and a preexisting tRNA gene. The only genes that need be transferred would be an arginine recognizing a tRNA ligase and the biosynthetic genes for converting ornithine to arginine, if arginine was not present in the environment. In addition, we would possibly need

to include a gene encoding an arginine-containing protein that provides a selective advantage to its host. Thus, the transfer of fewer than six genes could initiate the change.

The next problem has to do with natural selection for such a change. The question can be raised of why arginine would be superior to ornithine. There is one possible explanation. Primary amines, such as those in ornithine and lysine, are efficient in hydrolyzing the phosphodiester bond in RNA, whereas the guanidinium group is not. Presumably, dispersal of the change over the multiple nitrogen atoms in arginine attenuates the positive charge that promotes hydrolysis. Thus, regulatory and other binding proteins that interact with RNA might use arginine in place of lysine to attenuate this potential for hydrolysis. The positive charge on arginine could form a favorable salt bridge with the phosphate group without danger of promoting hydrolysis. Thus, according to this scenario, an unknown lineage evolved the biosynthesis of arginine for use in its RNA-binding proteins. This innovation produced an advantage that allowed this lineage to flourish. Horizontal spread of this gene to other lineages was driven by this selective advantage. This selective advantage, plus selection for unity *per se*, could have led to the modern genetic code.

We can envisage a similar scenario for tryptophan; that is, tryptophan serves a function in some enzymes that is unique and cannot be accommodated by tyrosine, phenylalanine or other hydrophobic amino acids. Transfer of a tryptophan tRNA ligase, the tryptophan operon and a unique tryptophan-containing enzyme is all that would be involved.

LCA

One of the obvious conclusions from these considerations is that the existence of a single homologous trait found in the three kingdoms does not necessarily imply that the LCA carried that trait. It is not even necessary to postulate the existence of such an ancestor, as has been suggested by Woese [30] and Doolittle [31]. It is simpler to posit that, since the origin of life, multiple lineages emerged and that multiple lineages are responsible for the so-called 'LCA'. The LCA is an unnecessary addition for any theory of evolution and eliminating the LCA makes for a more parsimonious theory. This point can be illustrated most simply by asking of a theory that contains a single lineage leading to a LCA the following questions. First, lysine tRNA ligase is present in two different, nonhomologous, forms known as class I and class II. How could a lineage that used a class I lysine tRNA ligase ever evolve a class II enzyme that is completely nonhomologous to the first? Second, consider the hydrophobic amino acids leucine, valine and isoleucine. How could a single lineage that used valine and leucine (or just one or other of the two) have the need for the other hydrophobic amino acid(s)? In fact, several puzzles concerning the evolution of the genetic code are rendered simpler if one assumes that multiple lineages with differing genetic codes evolved first and that, through the action of horizontal gene transfer and selection for unity *per se*, the modern genetic code emerged. These changes in the genetic code need not have occurred after the diversification of life into the three kingdoms, although the evidence outlined here suggests that tryptophan and arginine were indeed added after this diversification.

Acknowledgements
I thank Hy Hartman and Steve Daubert for useful suggestions.

References

1 Koonin, E.V. and Galperin, M.Y. (1997) Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* 7, 757–763
2 Woese, C. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6854–6859
3 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129
4 Jain, R. *et al.* (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3801–3806
5 Makarova, K.S. *et al.* (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* 9, 608–628
6 Doolittle, R.F. *et al.* (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271, 470–477
7 Adkins, R.M. (1998) Dating the age of the last common ancestor of all living organisms with a protein clock. In *Horizontal Gene Transfer* (Syvanen, M. and Kado, C.I., eds), pp. 380–391, Chapman & Hall
8 Kollman, J.M. and Doolittle, R.F. (2000) Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. *J. Mol. Evol.* 51, 173–181
9 Feng, D.F. *et al.* (1997) Determining divergence times with a protein clock: update and reevaluation *Proc. Natl. Acad. Sci. U. S. A.* 94, 13028–13033
10 Syvanen, M. (2002) On the occurrence of horizontal gene transfer among an arbitrarily chosen group of 26 genes. *J. Mol. Evol.* 54, 258–266
11 Nagel, G.M. and Doolittle, R.F. (1995) Phylogenetic analysis of the aminoacyl-tRNA synthases. *J. Mol. Evol.* 40, 487–498
12 Doolittle, R.F. and Handy, J. (1998) Evolutionary anomalies among the aminoacyl-tRNA synthases. *Curr. Opin. Genet. Dev.* 8, 630–636
13 Wolf, Y.I. *et al.* (1999) Evolution of aminoacyl-tRNA synthases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* 9, 689–710
14 Weiner, A.M. (1999) Molecular evolution: aminoacyl-tRNA synthases on the loose. *Curr. Biol.* 9, R842–R844
15 Ribas de Pouplana, L. *et al.* (1996) Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. *Proc. Natl. Acad. Sci. U. S. A.* 93, 166–170
16 Brown, J.R. *et al.* (1997) Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthases. *J. Mol. Evol.* 45, 9–16
17 Syvanen, M. (1985) Cross-species gene transfer; implications for a new theory of evolution. *J. Theor. Biol.* 112, 333–343
18 Reuben, J. and Polk, F.E. (1980) Nucleotide-amino acid interactions and their relation to the genetic code. *J. Mol. Evol.* 15, 103–112
19 Sukhodolets, V.V. (1982) Evolutionary changes in the genetic code, predictable on basis of the hypothesis of physical predetermination of the structure of codon bases. *Genetika* 18, 499–502
20 Knight, R.D. and Landweber, L.F. (1999) Is the genetic code really a frozen accident? New evidence from *in vitro* selection. *Ann. New York Acad. Sci.* 870, 408–410
21 Syvanen, M. (1986) Cross-species gene transfer: a major factor in evolution? *Trends Genet.* 2, 1–4
22 Woese, C. (1969) Models for the evolution of codon assignments. *J. Mol. Biol.* 43, 235–240
23 Woese, C.R. (1973) Evolution of the genetic code. *Naturwissenschaften* 60, 447–459
24 Fox, S.W. (1974) Origins of biological information and the genetic code. *Mol. Cell. Biochem.* 3, 129–142
25 Jukes, T.H. (1973) Possibilities for the evolution of the genetic code from a preceding form. *Nature* 246, 22–26
26 Hartman, H. (1975) Speculations on the evolution of the genetic code. *Origins Life* 6, 423–427
27 Hartman, H. (1978) Speculations on the evolution of the genetic code. II. *Origins Life* 9, 133–136
28 Jukes, T.H. (1973) Arginine as an evolutionary intruder into protein synthesis. *Biochem. Biophys. Res. Commun.* 53, 709–714
29 Jukes, T.H. (1974) The 'intruder' hypothesis and selection against arginine. *Biochem. Biophys. Res. Commun.* 58, 80–84
30 Woese, C. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6854–6859
31 Doolittle, W.F. (2000) The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.* 10, 355–358