AMBIGUITY AND THE EVOLUTION OF THE GENETIC CODE

R. R. KOCHERLAKOTA* and N. D. ACLAND
Vincent Massey Collegiate Institute, Winnipeg, Manitoba, Canada

(Received 12 February, 1981; in revised form 1 March, 1982)

Abstract. The evolution of the genetic code is an extremely complex problem. The addition of a new method by which the code could evolve, however, allows much to be explained about the way in which the present codes (Γ_3 and $\Gamma_{\frac{1}{3}}$) originated. The idea that ambiguity would allow the length of the codon to change is very useful, since it predicts the distribution of the 4-blocs and 2-blocs in the code, determines where variations in the code are probable, and presents a scenario for the evolution of the code.

Models for the evolution of the genetic code have been proposed by many authors, using many different approaches. Much has been written about the biochemical aspects of the problem (Crick *et al.*, 1976; Dillon, 1973; Folsome, 1977; Jukes, 1973, 1974; Nagyravy and Fendler, 1974; Wetzel, 1978). The code has been treated as a language (Barricelli, 1979) and from the viewpoints of cybernetics and information theory (Kuhn and Kuhn, 1978; Mackay, 1967; Ratner and Batchinsky, 1976; Walker, 1974). Many attempts have been made to explain the universality of the code (Ishigami and Nagano, 1975; Wong, 1976). The universality of the code has been questioned by the recent work that found a variant form of the code in the mitochondrial DNA (Barrell *et al.*, 1979; Fox, 1979; Martin *et al.*, 1980). In addition, specific problems related to the code ranging from the rarity of arginine in proteins (Jukes, 1973, 1974; Wallis, 1974) to the various strategies employed by different organisms in delegating degenerate bases (Grantham, 1980) have been treated separately.

Another approach is to examine the code and its pattern, not treating its chemical aspects, and to try to find remnants of primordial codes in the present structure. It is this approach that will be used in the present paper. It is important to distinguish between actual remnants of codes and simple coincidences.

The assumption that the code evolved presupposes a method by which the code could change. Hartman (1978) identifies two methods by which this could occur. These are:

(a) The ancient code contained sixty-four codons, but coded for substantially fewer amino acids. As time passed, amino acids were added gradually. Such additions could be called *fine structure alterations.*

(b) The ancient code was similar to the modern code, but only two bases were used (Hartman suggests these were G and C). The other bases were subsequently added. This could be called an *alphabet alteration.*

It should be noted that there is a third variant form of this model that is possible:

(c) The codons are able to change in length. That is, instead of translating a codon of length three (eg. AUA), the ancient code may have translated a codon of length

* Present address: Caltech 1–59, Pasadena, CA 91126, U.S.A.

Origins of Life **12** (1982) 71–80. 0302–1688/82/0121–0071$01.50.

two (eg. AU). Such a model requires a change in codon length called a *word length alteration.*

It has been argued that a word length alteration could not occur, because such a change in the code would necessarily result in the loss of all genetic information, and thus force life to 'start over'. Expansion of the code would allow the production of organisms that could fill niches empty up to that point in time. If a word length alteration occurred early enogh in time, organisms using the altered code would be able to compete effectively with those using the original code. Since the expanded code would produce more diverse forms of life, eventually organisms using this code would be able to eliminate those using the original code.

In any case, the word length alteration requires an extremely complex change in the translation machinery compared to that required by a fine structure or an alphabet alteration. According to Hopfield (1978), the shape of the tRNA molecule did change at some point in time. It is possible to develop a model by which an alteration in the code could occur very rapidly. Moreover, if evolution were assumed to occur by a combination of word length and alphabet alterations, much would be explained about the structure of the code.

It should be noted that what is being proposed is not that the original codon consisted of three positions with fourfold degeneracy in the third position, as was proposed by Jukes (1974) and as was found in valine tRNA by Mitra *et al.* (1977). We believe that it is not necessary to assume that the initial codon contained three positions at all. Rather, the original codon contained only two positions, and it was subsequently expanded to contain three positions.

In order to demonstrate a method by which word length alterations could occur, it is necessary to consider the idea of ambiguity. An ambiguous code is one in which a single codon can code for any of several amino acids. This would seem to be an insurmountably disadvantageous state of affairs. King (1971), however, reports that the slight ambiguity resulting from the presence of ochre and amber suppressors does not necessarily result in the death of an organism. Thus, it is possible for an organism to survive with ambiguity. In fact, in certain circumstances, ambiguity could provide a long term selective advantage to an organism. Any change in the genetic code is bound to be disadvantageous in the very short term, and ambiguity would provide a method by which the time spent in changing the code could be minimized. If the code is required to change very rapidly, ambiguity may be the ideal mechanism. Instead of waiting for a specific mutation, a slow process at best, an ambiguous code could evolve by simply eliminating the translation of one or more amino acids at each codon. For example, assume that the codon $\alpha\beta\gamma$ can produce either Ser, Arg, Asn, or Met. Further, assume that Ser has an evolutionary advantage over the other amino acids for that particular codon. An organism that is more likely to produce Ser would have a selective advantage: the other possibilities would be gradually eliminated. Thus, evolution at the level of the genetic code would occur.

Woese (1965) suggested that ambiguity could be responsible for fine structure alterations. The same method could be used, however, for both alphabet and word

length alterations. The question is, under what conditions could this evolution occur? Clearly, the disadvantage associated with ambiguity would prevent evolution in most cases. If, however, the rate of change of the environment is sufficiently high, the ordinary molecular or organismic evolutionary processes may be too slow to adapt a population to changing environments. In this case, ambiguity would provide a way of altering the code. If such an alteration results in an increase in the vocabulary (the number of amino acids coded for), it has the obvious benefit of increasing the number of possible proteins (the repertoire). This would allow faster molecular evolution and easier adaptation to a changing environment.

Thus, we have two factors that act in opposite directions with varying magnitude: the selective disadvantage associated with ambiguity and the selective advantage associated with the vocabulary expansion that may accompany ambiguity. Since the vocabulary expansion advantage would increase if the environmental rate of change were very high and the ambiguity disadvantage is independent of the environmental rate of change, it should be possible to increase the rate of change to such an extent that the vocabulary advantage would outweigh the ambiguity disadvantage. Thus, if the environmental rate of change were sufficiently high, ambiguity would be an advantageous trait.

So far, we have determined that if the need for a large vocabulary is sufficient, ambiguity may develop. However, we have not dealt with the type of advantage that might be associated with translating a given codon into a specific amino acid. Consider, for instance, the case of conservative mutations. The probability that a mutation is conservative, called the abuttal probability, is a crucial factor in evolution at the level of the protein. If the abuttal probability is very high, the mutation rate will be proportionately reduced in the effects it has on a protein's evolution. The abuttal probability thus serves as a sort of buffer, allowing the effective mutation rate to be more precisely controlled. Moreover, because of the advantages associated with the relative positions of amino acids, the abuttal probability influences the final positions of the amino acids in the code. As far as the initiation of evolution is concerned, however, the abuttal probability is minor.

Thus, we have identified three basic 'forces' involved in shaping the genetic code:

(a) Selective advantages associated with vocabulary expansion.

(b) Selective disadvantages associated with ambiguity.

(c) Advantages associated with the abuttal probability.

The significance of each of these considerations in selection can easily be visualized. If, and only if, the environment is changing fast enough to make a change in the genetic code advantageous, ambiguous codes will have a long term selective advantage. In ordinary circumstances, the disadvantage associated with ambiguity will keep the code nonambiguous. The abuttal probability is of significance in determining the code only in the ambiguous case, although it does play a role in determining the strategy used by an organism in selecting degenerate bases (Grantham, 1980) in the nonambiguous case, allowing the organism to select those bases which optimize the effective mutation rate for that organism.

TABLE I

	Γ_1		
$N = 2$	$W = 2$	$V = 4$	
First	Second		
	G	C	
G	Gly	Ala	
C	Arg	Pro	

Like all evolutionary models, this is a purely probabilistic model. It is possible that an ambiguous code could arise and prosper in a slowly changing or static environment, but it is unlikely. One cannot categorically say that a given event will or will not occur, but only state whether it is likely to occur.

With this knowledge in hand it should not be difficult to construct a specific model for the evolution of the genetic code. There is reason to believe (Hartman, 1978; Kharchenko, 1976) that a code consisting only of C and G may have been a primordial code. The four amino acids whose codons consist exclusively of C and G are structural amino acids, and are suspected of being in the earliest of the codes. Taking the hypothesis one step further than did Hartman, we propose that the primeval code contained only two bases, and that these were arranged in two letter codons. That is, GC and not GCG would have been translated. Let us designate this code by Γ_1 and represent it as in Table I, where N, W, and V represent the alphabet, the codon length, and the vocabulary, respectively. There is one extremely important point that should be noted in this code. All of the amino acids in this code are found in 4-blocs in the modern code. (A K-bloc is a set of K codons of the form $\alpha\beta$N, where α and β are fixed bases and N is a variable base.)

What was the next stage in evolution? Clearly, a fine structure alteration would not change the vocabulary, and the disadvantages associated with ambiguity would virtually preclude such a useless alteration. A word length alteration does not appear to have occurred either. The purpose of any alteration is to increase the vocabulary.

TABLE II

	HPC		
$N = 2$	$W = 3$	$V = 4$	
First	Second	Third	
	G	C	
G	Gly	Ala	G
	Gly	Ala	C
C	Arg	Pro	G
	Arg	Pro	C

That lengthening the codon would not have increased the vocabulary can be deduced from the structure of the modern code. Assuming that the amino acid coded by a given codon has not changed, a word length alteration would give us Hartman's Primeval Code (HPC) as in Table II.

By elimination, the next alteration that occurred was probably an alphabet alteration. It is useful to consider the method by which this would have taken place. Undoubtedly, the oridinary G and C bases were gradually replaced by the new A and U bases at positions in the genetic material. It would take longer for a 'double-new' codon like AU to appear than for a 'single-new' codon like AC to appear. Suppose that p represents the probability that a new base will be introduced into the genome at a given position in a certain interval of time. The expected number of these intervals before the appearance of a given single-new codon is $1/p$. The expected number of intervals before the appearance of a given double-new codon is $1/p^2 > 1/p$. Thus, there would be an expected time lag between the appearance of single-new and double-new codons. Now, looking at the three letter modern codons which have double-new two letter codons as their first two letters, it is evident that there are no 4-blocs present. For instance, UUN consists of two 2-blocs of Phe and Leu. To contrast, in single-new codons, a half of the three letter codons form 4-blocs. As we have seen all of the three letter codons corresponding to codons in the ancient Γ_1 are 4-blocs. This is summarized in Table III. As can be seen the proportion of 4-blocs decreased as the time of appearance increased. It is possible to explain this on the basis of coincidence, but the probability that an arbitrary set of four codons (here the double-new codons) will have no 4-blocs if the 4-blocs are distributed at random is 0.0625. The time lag suggests a more elegant explanation than chance. Suppose that an alphabet alteration had occurred, increasing the vocabulary to V. Furthermore, suppose that a minimum vocabulary of $V^\star$ was necessary to maintain the organism. Prior to this alphabet alteration, the vocabulary had been static at S. Figure 1 is a graph of vocabulary versus time. Slope A represents the time prior to the alphabet alteration with the vocabulary at S. Suddenly, the environmental rate of change increased, favoring ambiguity. This favors an alphabet alteration (B). This alteration, however, has increased the vocabulary to V which is more than $V^\star$. At this stage, the code is still ambiguous, and has an extremely high vocabulary, since more than one amino acid is coded for by each codon. As the ambiguity is gradually eliminated, however, the vocabulary will decrease (C). Eventually, (D), the vocabulary is $V^\star$. If, however, the

TABLE III

Time of appearance versus proportion of 4-blocs

Codon Type	Time of Appearance	Proportion of 4-blocs
Γ_1	t	1.00
Single-new	$t + \Delta t$	0.50
Double-new	$t + (\Delta t)^2$	0.00

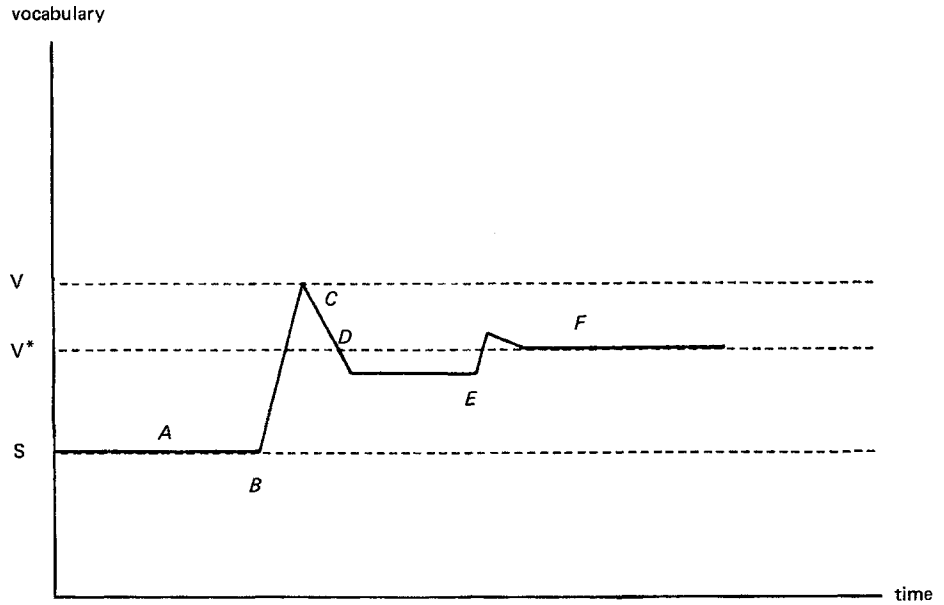

Fig. 1. Vocabulary Versus Time. This is part of a scenario for the evolution of the code, representing the interval of time during which the code Γ_1 was changed to Γ_3. Here, A = a period of static vocabulary before alteration; B = alphabet alteration, marking an increase in ambiguity and vocabulary; C = a period of decline of vocabulary and ambiguity; D = point when the vocabulary is the minimum necessary for survival, but the ambiguity is too high, so the vocabulary and the ambiguity continue to fall; E = a word length alteration to increase the vocabulary; F = a period of static vocabulary.

code is still ambiguous, we are at an impasse. The ambiguity must be eliminated, but the vocabulary cannot be decreased. If the ambiguity continues to decrease, it is probable that another alteration will take place. A word length alteration (E) could occur. Although this would necessarily involve an increase in ambiguity, this increase would quickly be reduced. In this way, ambiguity could be reduced, while vocabulary remains at a high enough value for the environmental rate of change.

This codon is never nonambiguous, so that it is impossible to give a truly explicit table of the code. It is, however, possible to decide which of the codons were ambiguous at the point of minimum ambiguity, just before the word length alteration. These codons were probably those which do not correspond to modern three letter codons that form 4-blocs. The reasoning behind this assertion is as follows: if such a codon were ambiguous, and a third position were added, the resulting codons could become non-ambiguous without reducing the number of amino acids coded for. For instance, AA could be ambiguous, coding for both Asn and Lys. The addition of the third position would allow the amino acid produced by AAU and AAC to be fixed as Asn, and that produced by AAA and AAG to be fixed as Lys. If this occurred at the proper point in time, it would take advantage of the time lag to guarantee that none of the double-new, and some of the single-new codons, formed 4-blocs. The wobble rule of pairing due to Crick specifies which codons would produce the same amino acid when

TABLE IV

Γ_2

$N = 4$	$W = 2$		$V = 21$ (?)	
First	Second			
	U	C	A	G
U	–	Ser	–	–
C	Leu	Pro	–	Arg
A	–	Thr	–	–
G	Val	Ala	–	Gly

the code was fixed. The third position would be subject to wobble since it is the third position that allows degeneracy: several codons coding for the same amino acid. In turn, this is because the third position was added to reduce ambiguity.

The remaining codons, those which correspond to modern 4-blocs, would not be ambiguous. Thus, we can designate this code Γ_2 and represent it as in Table IV, where '---' is an ambiguous codon.

Since it is only in the word length alteration that the codons marked as ambiguous are actually fixed as to which amino acid is produced, we can predict where variations can be expected in the code. The wobble rule allows several stable configurations for the pattern of the code, and one would expect a variety of different codes in existence. In fact, we can predict that the modern codons that would show such variations are probably those which correspond to ambiguous codons in Γ_2. This has been partially confirmed by recent work (Barrell *et al.*, 1979; Fox, 1979; Martin *et al.*, 1980) which suggests that the mitochondrion translates UGA as Trp, not as a terminator codon as is done in the ordinary code. Barrell *et al.* (1979) make the additional suggestion that there is a variation in the code at AUA, which is translated as Met by the mitochondrion. As can be seen, these codons correspond to the ambiguous Γ_2 codons AU and UG. Any other variation in the code would probably occur in codons that correspond to ambiguous Γ_2 codons.

The modern code that is used by most (to our knowledge) prokaryotes and eukaryotes can be designated Γ_3. The variant mitochondrial code can be designated Γ_3^* to distinguish it from the ordinary code. These codes are shown in Table V and Table VI, respectively. The evolution of the genetic code, as expected on the basis of this model, is summarized in a tree diagram in Figure 2. Γ_3^{**} refers to a hypothetical third form of the modern code.

The word length alteration that is a part of this model probably occurred very early in the history of life. If a word length alteration were to occur at a time when complex organisms were already in existence, it is unlikely that the organisms which would bear the new code, having lost their genetic information, would be able to compete effectively with the already present organisms. For this reason, the model supports the endosymbiosis theory of mitochondrial origin. The mitochondrial code Γ_3^* originated early in the history of life. Eukaryotic cells, however, only developed much

TABLE V

Γ_3
$N = 4$ $W = 3$ $V = 21$

First	Second				Third
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

TABLE VI

Γ_3^*
$N = 4$ $W = 3$ $V = 21$

First	Second				Third
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Trp	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Met	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

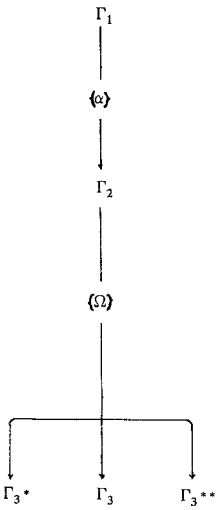

Fig. 2. The Evolution of the Genetic Code. This tree diagram shows the evolution of the genetic code from Γ_1 to Γ_3. Letters in braces indicate the type of alteration that occurred. $\{\alpha\}$ = alphabet alteration, $\{\Omega\}$ = word length alteration.

later, and in the interim, there was no such thing as a mitochondrion. Probably, the mitochondrion is a remnant of an organism that used Γ_3^* that, through endosymbiosis, entered the eukaryotic cell. Thus, although Γ_3 and Γ_3^* originated approximately simultaneously, they did not combine in the eukaryotic cell for many millions of years.

The genetic code is an important component of our conception of the origin of life. An understanding of how the code evolved would give us a better perspective on life as a whole, and how it originated. Models for the evolution of the code are particularly useful for answering questions about the early history of life, such as the validity of the endosymbiotic theory of mitochondrial origin.

Acknowledgement

We would like to thank Dr J. L. King for his encouragement and advice, and for reading an early draft of the manuscript.

References

Barrell, B. G., Bankier, A. T., and Drouin, J.: 1979, *Nature* **282**, 189.
Barricelli, N. A.: 1979, *BioSystems* **11**, 19.
Crick, F. H. C., Brenner, S., Klug, A., and Piecznik, G.: 1976, *Origins of Life* **7**, 389.
Dillon, L. S.: 1973, *Bot. Rev.* **39**, 301.
Folsome, C.: 1977, *Origins of Life* **8**, 391.
Fox, T. D.: 1979, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 6534.
Grantham, R.: 1980, *Trends Biochem. Sci.* **5**, 327.
Hartman, H.: 1978, *Origins of Life* **9**, 133.
Hopfield, J. J.: 1978, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4334.

Ishigami, M. and Nagano, K.: 1975, *Origins of Life* **6**, 551.
Jukes, T. H.: 1973, *Nature* **246**, 22.
Jukes, T. H.: 1974, *Origins of Life* **5**, 331.
Kharchenko, E. P.: 1976, *J. Evo. Biochem. Physiol.* **12**, 255.
King, J. L.: 1971, 'The Influence of the Genetic Code on Protein Evolution', in E. Schoffenields (ed.), *Biochemical Evolution and the Origins of Life* North-Holland, Publ., Amsterdam, pp. 3–13.
Kuhn, H. and Kuhn, C.: 1978, *Origins of Life* **9**, 137–150.
Mackay, A. L.: 1967, *Nature* **216**, 159.
Martin, N. C., Pham, H. D., Underbrink-Lyon, K., Miller, D. L., and Donelson, J. E.: 1980, *Nature* **285**, 579.
Mitra, S. K., Lustig, F., Akesson, B., Lagerkvist, U., and Strid, L.: 1977, *J. Biol. Chem.* **252**, 471.
Nagyravy, J. and Fendler, J. H.: 1974, *Origins of Life* **5**, 357.
Ratner, V. A. and Batchinsky, A. G.: 1976, *Origins of Life* **7**, 225.
Walker, G. W. R.: 1974, *Origins of Life* **5**, 351.
Wallis, M.: 1974, *Biochem. Biophys. Res. Comm.* **56**, 711.
Wetzel, R.: 1978, *Origins of Life* **9**, 39.
Woese, C. R.: 1965, *Proc. Natl. Acad. Sci. U.S.A.* **54**, 1546.
Wong, J. T.-F.: 1976, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2336.