Proc. Nat. Acad. Sci. USA
Vol. 72, No. 5, pp. 1909–1912, May 1975

A Co-Evolution Theory of the Genetic Code

(amino-acid biosynthesis/prebiotic pathways/tRNA)

J. TZE-FEI WONG

Department of Biochemistry, University of Toronto, Toronto M5S 1A8, Canada

Communicated by J. Tuzo Wilson, March 5, 1975

ABSTRACT The theory is proposed that the structure of the genetic code was determined by the sequence of evolutionary emergence of new amino acids within the primordial biochemical system.

The genetic code for protein molecules is a triplet code, consisting of the 64 triplets of the four bases adenine, guanine, cytosine and uracil (1, 2). The cracking of the code was a monumental achievement, but it posed in turn what Monod (3) regards as one of the challenges of biology, namely the "riddle of the code's origin." Crick (4) has discussed two different theories which have been proposed regarding this origin. The Stereochemical Theory postulates that each amino acid became linked to its triplet codons on account of stereochemical reasons, whereas the Frozen Accident Theory postulates that the linkage arose purely by chance. Since neither theory has given a systematic solution to the riddle, the present purpose is to explore a third hypothesis, which postulates that:

> *The structure of the codon system is primarily an imprint of the prebiotic pathways of amino-acid formation, which remain recognizable in the enzymic pathways of amino-acid biosynthesis. Consequently the evolution of the genetic code can be elucidated on the basis of the precursor–product relationships between amino acids in their biosynthesis. The codon domains of most pairs of precursor–product amino acids should be contiguous, i.e., separated by only the minimum separation of a single base change.*

This theory, which may be called a Co-evolution Theory, is readily tested. If many pairs of amino acids which bear a nearest (in terms of the number of enzymic steps) precursor–product relationship to each other in a biosynthetic pathway fail to occupy contiguous codon domains, the theory would be untenable. The known precursor–product conversions between amino acids are (5–7):

Glu → Gln
Glu → Pro
Glu → Arg
Asp → Asn
Asp → Thr
Asp → Lys
Gln → His
Thr → Ile
Thr → Met
Ser → Trp
Ser → Cys
Val → Leu
Phe → Tyr

Of these, only the relationships of Asp to Lys and Thr to Met require some comment. Lys can be synthesized either from Asp via the diaminopimelate pathway (8), or from Glu via the α-aminoadipate pathway (9). Since the former pathway operates in prokaryotes and the latter in eukaryotes, an Asp–Lys pairing has greater prebiotic significance than a Glu-Lys pairing. The biosynthesis of Met can proceed best from Asp, but Thr is nearer to Met in terms of the number of enzymic steps involved (homoserine, which might represent a more primitive form of Thr, is even nearer still to Met). Although Ser and Cys can enter into the Met-biosynthetic pathway subsequent to the entry of Thr, neither Ser nor Cys is a straightforward precursor of Met. Ser is not the only possible contributor of a one-carbon group to Met, and Cys is not the only possible contributor of sulfur (10). α-Transaminations, because of their relative nonspecificity, are not regarded as useful criteria for the tracing of precursor–product relationships. Aside from the above precursor–product relationships, Glu, Asp, and Ala are known to be interconvertible via the tricarboxylate cycle, and Ala, Ser, and Gly via the metabolism of pyruvate, glycerate, and glyoxylate (6).

Evolutionary map of the genetic code

When the codons for various precursor–product amino acids (Table 1) are examined, many of the codon domains of product amino acids are found to be contiguous with those of their respective precursors. The only noncontiguities are those of the Glu–Pro, Glu–Arg, Asp–Thr, and Asp–Lys pairs. If the prebiotic derivations of Gln from Glu, and Asn from Asp, had not occurred at the earliest stages of codon distribution, CAA and CAG could be expected to form part of the early Glu codons, and AAU and AAC part of the early Asp codons. This simple secondary postulate regarding the dicarboxylic amino acids and their amides suffices to remove all noncontiguities between precursors and products. It becomes possible to construct in Fig. 1 a map of the genetic code in which the codon domains of every precursor–product pair of amino acids (connected by single-headed arrows), as well as those of other interconvertible pairs (connected by double-headed arrows) are separated by only a single base change. This confirms the prediction by the Co-evolution Theory that codon distribution is closely related to amino-acid biosynthesis. Furthermore, since the theory suggests that the enzymic pathways of amino-acid biosynthesis largely stemmed from the prebiotic pathways of amino-acid formation, the pathways of this map are regarded as co-evolutionary pathways through which new amino acids were generated within the primordial system, and through which the triplet codons became distributed to finally the 20 amino acids.

Tests for randomness

The correlation between codon distribution and amino-acid biosynthesis indicated in Fig. 1 could arise not only from co-evolution, but also in principle from chance. However, the unlikelihood of the latter explanation can be demonstrated in two different ways. First, consider the widespread contiguities between the codons of precursor and product amino acids. For any precursor codon triplets, there will be *a* other

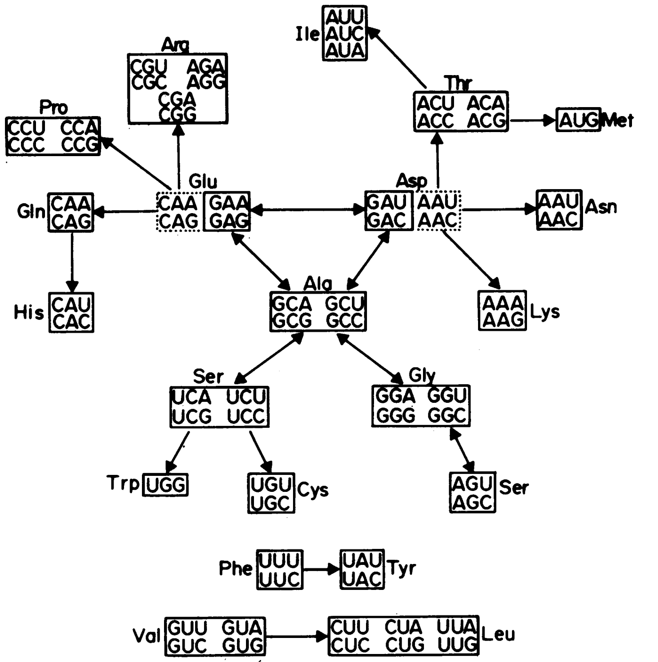

FIG. 1. Evolutionary map of the genetic code: Codons enclosed in solid boxes correspond to the codons used in present day organisms; Glu and Asp codons enclosed in dotted boxes likely belonged to these amino acids at an early primordial stage. Single-headed arrows represent biosynthetic precursor–product relationships between amino acids, and double-headed arrows represent biosynthetic interconversions. All pairs of codon domains connected by either single-headed or double-headed arrows in this map are contiguous, i.e., separated by only a single base change.

TABLE 1. *The genetic code*

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	ter	UGA	ter
UUG	Leu	UCG	Ser	UAG	ter	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

ter = chain termination signal.

triplets in the genetic code which are contiguous with the group, and b other triplets which are noncontiguous. If a product of this amino acid has n codons, the random probability P that as many as x of these n codons turn out to be contiguous with some precursor codon is determined by the hypergeometric distribution:

$$P = \sum_{x}^{n} \frac{a!}{(a-x)!x!} \cdot \frac{b!}{(b-n+x)!(n-x)!} \cdot \frac{(a+b-n)!n!}{(a+b)!}. \quad [1]$$

The calculated values of P for eight precursor–product pairs are shown in Table 2. Using the method of Fisher (11), the eight corresponding $-2\ln P$ values can be summed to give a χ^2 value of 45.01 with 16 degrees of freedom; this indicates an aggregate probability of less than 0.0002 that these eight sets of contiguities could have become so numerous by chance. Amongst the eight amino-acid pairs, either Phe–Tyr or Val–Leu may represent sibling products of a common biosynthetic pathway rather than true precursor and product. Their deletion from calculation leaves a χ^2 value of 27.10 with 12 degrees of freedom, which still points to an aggregate probability of only 0.0075. The potential Glu-Pro, Glu-Arg, Asp-Thr, Asp-Lys, Thr-Met, Ala-Ser-Gly and Glu-Asp-Ala contiguities, plausible but less certain, have not been included in these calculations; their inclusion would lower the aggregate probability even further. Also, there are other ways to perform the statistical analysis, e.g., by taking a pair of codons such as UGU and UGC as one rather than two units in the hypergeometric distribution, but the nonrandom character of the precursor–product contiguities is far too striking to be fundamentally circumventable by statistical methodology.

Secondly, Gln, Pro, and Arg are biosynthetic siblings of the Glu family, and Asn, Thr, and Lys are siblings of the Asp family. Likewise, Cys and Trp are siblings of the Ser family, and Ile and Met are siblings of the Thr family. Of the seven pairs of amino acids in Table 1 that share the first two bases, Ile–Met, Asn–Lys, and Cys–Trp are siblings. His–Gln are precursor–product, and Asp–Glu are either siblings or precursor–product. Only Phe–Leu and Ser–Arg are unrelated pairs. There are 190 possible amino-acid pairs amongst the 20 amino acids, and the four families of siblings generate a total of eight sibling pairs. Accordingly the probability of randomly finding as many as three out of any seven amino-acid pairs to be sibling pairs is only 0.00161 on the basis of Eq. 1 ($a = 8, b = 182, n = 7, x = 3$). If Ile–Met are not regarded as siblings, this probability would be raised to 0.0224, but then there are also grounds to consider Asp–Glu as siblings of the tricarboxylate cycle, whereupon it would be reverted to 0.00161. In any case the enrichment of siblings amongst amino-acid pairs sharing the same first two bases appears strongly nonrandom, and provides further evidence against a chance origin of the correlation between amino-acid biosynthesis and codon distribution.

Unlike codon contiguities between chemically similar amino acids, which might offer the genetic code some protection against the harmful effects of excessive mutations or coding errors, the contiguities between such chemically dissimilar precursor–product or sibling pairs as Ser–Trp, Thr–Ile, Gln–His, Ile–Met, Asn–Lys, and Cys–Trp could offer little advantage, and could not have been accumulated through natural selection. They require interpretation, as attempted by the Co-evolution Theory, as fossilized vestiges of the history of the genetic code. Indeed, the preservation of so many nonrandom and yet nonadvantageous vestiges suggests that the code, once established, did not undergo extensive revision.

Significance of co-evolution

The Co-evolution Theory postulates that amongst the amino acids biosynthetic precursor–products correspond extensively

TABLE 2. *Random probability of precursor–product codon contiguities*

Precursor–product	*a*	*b*	*n*	*x*	*P*
Ser–Trp	34	24	1	1	0.586
Ser–Cys	34	24	2	2	0.339
Val–Leu	24	36	6	6	0.00268
Thr–Ile	24	36	3	3	0.0591
Gln–His	14	48	2	2	0.0481
Phe–Tyr	14	48	2	2	0.0481
Glu–Gln	14	48	2	2	0.0481
Asp–Asn	14	48	2	2	0.0481

The parameters *a*, *b*, *n*, *x*, and *P* are defined by Eq. 1.

TABLE 3. *Stem sequence of the anticodon loop of some pairs of biosynthetically related transfer RNAs from Escherichia coli (19–21)*

tRNA pair	5′-arm........3′-arm
Ser-1, Trp	C-C-G-G-U........A-C-C-G-G
Thr, Ile-1	C-A-C-C-C........G-G-G-U-G
Gln-2, His-1	C-Y-G-G-A........Ψ-C-C-R-G
Val-2, Leu-2	C-Y-A-C-C........G-G-U-R-G
Ala-1, Asp-1	C-C-U-G-C........G-C-A-G-G

R = purine nucleoside; Y = pyrimidine nucleoside.

to prebiotic precursor-products, and the structure of the genetic code was primarily shaped by the prebiotic evolution of amino acids. Therefore, it is a theory of prebiotic amino-acid evolution as much as a theory of the genetic code. Previously Nirenberg *et al.* (12) recognized the contiguities between the codons of some sibling amino acids synthesized from a common precursor, but did not explore any interpretation of their prebiotic significance. The grouping of Ile with Leu and Val as siblings by these workers stemmed from an apparent mistaking of α-ketobutyrate rather than pyruvate as the precursor of Leu and Val, and the grouping of Trp with Phe and Tyr relied on the biosynthetic relationship between shikimates and Trp, which is much more remote than the relationship between Ser and Trp; even Gln, through its participation in the synthesis of anthranilate (13), stands nearer to Trp than Phe and Tyr.

The evolutionary map of Fig. 1 defines one major center of amino acids, consisting of Glu, Asp, Ala, Ser, and Gly, from which 11 other amino acids evolved, and the two minor centers of Phe–Tyr and Val–Leu. This suggestion of the evolutionary primacy of the major center is supported by the finding that its members were formed in greatest abundance when a simulated primitive atmosphere was subjected to electric discharge (14), electron bombardment (15), or thermal treatment (16); by the many metabolic interconversions developed between its members and the intermediates of carbohydrate and lipid metabolism; by the entry of Asp, Gly, and Gln into the evolving pathways of pyrimidine and purine biosynthesis; and by the fact that α-transaminases, usually employing Glu and Ala as amino donors, contribute to the biosynthesis of practically every amino acid, including Phe, Tyr, Val, and Leu in the minor centers. Initially the prebiotic pathways of amino-acid formation probably included reactions uncatalyzed by enzymes. Subsequently these pathways underwent extensive "enzymatization," with or without some modification of their ground plan, when faster reaction rates became an important factor in natural selection. The cyclization of glutamyl-γ-semialdehyde to form Δ^1-pyrroline-5-carboxylate in proline biosynthesis, found by Vogel and Davis (17) to occur spontaneously, may be an interesting example of the survival of a prebiotic, uncatalyzed reaction step into a modern biosynthetic pathway.

Why did prebiotic precursor and product amino acids come to occupy contiguous codons in the genetic code? One of the mechanisms by which codons were assigned to emergent amino acids is suggested by a consideration of the Glu and Gln, and the Asp and Asn, codons. The evolutionary map of Fig. 1 indicates that CAA and CAG were initially Glu codons. Later, as Gln evolved from Glu to join the amino-acid system, these two codons were conceded by Glu to Gln. Similarly, AAU and AAC were initially Asp codons, only later conceded to Asn. Accordingly, at the very early stages, the system of amino acids entering into primordial proteins likely consisted of only a few amino acids, each occupying a continuous domain of the genetic code. As new product amino acids were formed from these on account of the presence of new enzymic or nonenzymic catalysts and co-factors, a precursor might concede some of its codons to a product. The latter also might spill over to occupy any neighboring codons that were not yet firmly assigned to any particular amino acid. Both routes led to contiguity between the codon domains of precursor and product. At least three related mechanisms present themselves by which a precursor could concede some of its codons to a product. First, the product might resemble the precursor sufficiently to compete for attachment to the adaptors of the precursor. Second, the chemical conversion of precursor to product might occur while the precursor was attached to its adaptors. Third, an intermediate in the prebiotic pathway leading from precursor to product might attach to the adaptors of the precursor and then become converted into the product. In each case, the product took over some of the precursor codons through the concession of adaptors of protein synthesis (i.e., the primordial equivalent of tRNAs) from precursor to product. A search for structural similarity amongst the present day prokaryotic tRNAs for biosynthetically related amino acids is inviting, although such a search is subject to four basic limitations: (*i*) the primordial tRNAs conceded by a precursor to a product might be unlike the tRNAs retained by the precursor; (*ii*) the primordial tRNAs for different precursors need not be totally dissimilar; (*iii*) tRNA structures are known to have undergone extensive evolution between the prokaryotic and eukaryotic stages of life, and far greater divergent and convergent changes could have taken place between the primordial and prokaryotic stages; (*iv*) tRNAs accepting different amino acids differ from one another mostly by 40–70% in their base sequences, thus rendering imprecise any attempt to construct an evolutionary tree (18). The last limitation in particular suggests that comparisons between tRNAs should be performed not only on the overall structure, but also on selected regions such as the anticodon loop, which is the primary active center of the molecule. Table 3 shows the similarity in the stem of this loop between some of the *Escherichia coli* tRNAs for pairs of biosynthetically related amino acids. These similarities are not strictly specific, e.g., the common stem sequence for $\mathrm{tRNA}_1^{\mathrm{Ser}}$ and $\mathrm{tRNA}^{\mathrm{Trp}}$ is also shared by $\mathrm{tRNA}_1^{\mathrm{Gln}}$ (22). However, they are also by no means entirely nonspecific, e.g., the stem sequence of $tRNA^{Trp}$ is the same as that of $tRNA_1^{Ser}$ but not the same as that of $tRNA_3^{Ser}$, the stem sequence of which is C-U-C-C-C.......G-G-G-A-G (23, 24), thus correctly reflecting that $tRNA_1^{Ser}$ but not $tRNA_3^{Ser}$ serves any codon that is contiguous to the Trp codon of UGG. Fragmentary as they are, the indicated similarities are consistent with the concept of codon concession by precursor to product, and suggest that at least the anticodon region of the tRNA molecule was possibly an early development, functioning as amino-acid adaptor in protein synthesis throughout the greater part of the co-evolutionary age.

Within this framework of co-evolution of amino acids and their codons, it is expected that additional factors would help to determine the exact allocation of some of the codons. It has been suggested that codon plurality for any amino acid, or codon contiguity between chemically similar amino acids, could minimize the damage due to excessive mutations or coding errors (25, 26). Accordingly, for example when Lys was first synthesized from Asp, different trial versions of the genetic code might have emerged with Lys occupying different domains contiguous to the Asp codons. Eventually the version that allocated AAA and AAG to Lys would be favored over the others if there were survival advantages to having the Lys domain adjacent to that of Arg. Once the code was established, the preservation of precursor–product and sibling contiguities within the code points to a lack of extensive subsequent changes; Crick (27), and Hinegardner and Engelberg (28) have arrived at the same conclusion from a separate consideration of the drastic effects incurred by changes in the code.

As Jukes has suggested (29), early evolutionary arrivals amongst the amino acids would have more opportunity to establish a plurality of codons than late arrivals. Met and Trp were possibly late arrivals which acquired only a single codon each. Even later arrivals, such as hydroxyproline and hydroxylysine residues, could enter into proteins only by way of post-translational modifications. However, since Glu and Asp, which by all other indications were amongst the earliest, are not recognized as early arrivals on the basis of codon plurality alone, earliness could not be the sole determinant of plurality. The Co-evolution Theory proposes that although the acquisition of codon plurality depended upon earliness of arrival, its retention depended upon evolutionary inertness. Thus, Leu and Arg each occupy as many as six coding triplets, but these amino acids never evolved further. In contrast, although Glu was one of the earliest arrivals, it was repeatedly transformed to yield new products. Accordingly, the Glu family occupies a total of sixteen triplets, but Glu itself is left with only two triplets. Likewise, the Asp family occupies 14 triplets, but Asp itself is left with only two. To acquire and retain a high plurality, like Leu and Arg, the amino acid had to be both early in arrival and inert in reactivity.

In conclusion, unanswered questions surrounding the genetic code remain numerous. Nevertheless, the Co-evolution Theory, by offering plausible explanations to many nonrandom characteristics of the code, suggests that the code's origin is by no means a riddle closed to systematic enquiry. The structure of the code begins to appear less haphazard in the light of the likely events of prebiotic evolution.

I wish to thank Mr. L. E. S. Green and Dr. D. B. W. Reid for their valuable advice on statistical methods, and Drs. F. H. C. Crick, C. S. Hanes, J. Mandelstam, and P. J. Piggot for their helpful criticism. This study has been supported by the Medical Research Council of Canada.

1. Nirenberg, M. W. (1963) *Harvey Lectures* **59,** 155–185.
2. Khorana, H. G. (1966) *Harvey Lectures* **62,** 79–105.
3. Monod, J. (1971) in *Chance and Necessity* (A. A. Knopf Inc., New York), p. 143.
4. Crick, F. H. C. (1968) *J. Mol. Biol.* **38,** 367–379.
5. Cohen, G. N. (1968) *The Regulation of Cell Metabolism* (Hermann, Paris), pp. 101–163.
6. Greenberg, D. M. (1969) in *Metabolic Pathways*, ed. Greenberg, D. M. (Academic Press, New York and London), Vol. III, 3rd ed., pp. 237–315.
7. Rodwell, V. W. (1969) in *Metabolic Pathways*, ed. Greenberg, D. M. (Academic Press, New York and London), Vol. III, 3rd ed., pp. 317–373.
8. Gilvarg, C. (1962) *J. Biol. Chem.* **237,** 482–484.
9. Strassman, M. & Weinhouse, S. (1953) *J. Amer. Chem. Soc.* **75,** 1680–1684.
10. Wiebers, J. L. & Garner, H. R. (1967) *J. Biol. Chem.* **242,** 5644–5649.
11. Fisher, R. A. (1950) in *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh and London), 11th ed., p. 99.
12. Nirenberg, M. W., Jones, O. W., Leder, P., Clark, B. F. C., Sly, W. S. & Pestka, S. (1963) *Cold Spring Harbor Symp. Quant. Biol.* **28,** 549–557.
13. Srinivasan, P. R. (1959) *J. Amer. Chem. Soc.* **81,** 1772–1773.
14. Miller, S. L. (1955) *J. Amer. Chem. Soc.* **77,** 2351–2361.
15. Palm, C. & Calvin, M. (1962) *J. Amer. Chem. Soc.* **84,** 2115–2121.
16. Harada, K. & Fox, S. W. (1964) in *The Origin of Prebiological Systems and of Their Molecular Matrices*, ed. Fox, S. W. (Academic Press, New York), pp. 187–194.
17. Vogel, H. J. & Davis, B. D. (1952) *J. Amer. Chem. Soc.* **74,** 109–112.
18. Dayhoff, M. O. & McLaughlin, P. J. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (National Biomedical Research Foundation, Washington, D.C.), Vol. 5, pp. 111–118.
19. Barrell, B. G. & Clark, B. F. C. (1974) in *Handbook of Nucleic Acid Sequences* (Joynson-Bruvvers Ltd., Eynsham, Oxford), pp. 8–59.
20. Williams, R. J., Nagel, W., Roe, B. & Dudock, B. (1974) *Biochem. Biophys. Res. Commun.* **60,** 1215–1221.
21. Clarke, L. & Carbon, J. A. (1974) *J. Biol. Chem.* **249,** 6874–6885.
22. Folk, W. R. & Yaniv, M. (1972) *Nature New Biol.* **237,** 165–166.
23. Yamada, Y. & Ishikura, H. (1973) *FEBS Lett.* **29,** 231–234.
24. Ish-Horowicz, D. & Clark, B. F. C. (1973) *J. Biol. Chem.* **248,** 6663–6673.
25. Sonneborn, T. M. (1965) in *Evolving Genes and Proteins*, ed. Bryson, V. & Vogel, H. J. (Academic Press, New York), pp. 377-397.
26. Woese, C. (1965) *Proc. Nat. Acad. Sci. USA* **54,** 1546–1552.
27. Crick, F. H. C. (1963) *Progr. Nucl. Acid. Res. Mol. Biol.* **1,** 163–217.
28. Hinegardner, R. T. & Engelberg, J. (1963) *Science* **142,** 1083–1085.
29. Jukes, T. H. (1971) in *Prebiotic and Biochemical Evolution*, eds. Kimball, A. P. & Oro, J. (North Holland/Elsevier, Amsterdam, London, New York), pp. 122–147.