Proc. Natl. Acad. Sci. USA
Vol. 75, No. 9, pp. 4334–4338, September 1978
Biophysics

Origin of the genetic code: A testable hypothesis based on tRNA structure, sequence, and kinetic proofreading

(protein synthesis/evolution/primitive cells)

J. J. HOPFIELD

Department of Physics, Princeton University, Princeton, New Jersey 08540; and Bell Laboratories, Murray Hill, New Jersey 07974

Contributed by John J. Hopfield, July 10, 1978

ABSTRACT **We hypothesize that the origin of the genetic code is associated with the structure of the tRNA that existed in primal cells. The sequences of modern tRNA contain correlations which can be understood as "fossil" evidence of the secondary structure of primal tRNA. Kinetic proofreading through diffusion can amplify a low level of intrinsic selectivity of tRNA for its amino acid. Experimental tests of the theory are suggested.**

What is the origin of the particular genetic code existing now? All life known today makes use of a single language for translating mRNA into proteins, in which each three-letter "triplet" of a protein-coding sequence of mRNA corresponds to a particular amino acid or stop instruction. Yet knowing the code tells us little about the reason this particular genetic code exists rather than some other. Speculation about the origin of the genetic code began before the code was deciphered (1). At one extreme of these speculations is the idea that the code has as its basis a chance event plus evolutionary fine tuning (2–4). In such a view, a code obtained from the present one by the permutation G → A, C → U, A → G, U → C (for example) is just as likely as our particular code. At the other extreme lies the theory that codons (or anticodons) select their corresponding amino acids through the stereochemical interactions of free amino acids with nucleic acids (5, 6). Between these viewpoints a virtual continuum of theories (7–10) exists in the literature [and even beyond (11) them!]. Various data have been used to evaluate these speculations: correlations between codon similarities and amino acid similarities (6); statistical analysis of the use frequencies of codons (10); homologies among tRNAs (12); experimental nucleic acid-amino acid affinity studies (6, 13–15), and molecular model building (16, 17).

In this paper, I construct a biochemistry of an early cell. Unlike present-day cells, this primal cell used the structure of primal tRNA to provide its genetic code dictionary. (In modern biology, this dictionary is provided enzymatically by the aminoacyl-tRNA synthetases.) A secondary structure that could be capable of yielding such a dictionary is proposed for primal tRNA. Statistical analysis of the sequences of the modern tRNA of *Escherichia coli* shows that autocorrelations within individual molecules are consistent with those to be expected if modern tRNA evolved from the hypothesized primal tRNA molecules, and provides "fossil" evidence for such a structure. The primal form permits a continuous evolution to modern tRNA. A mechanism of kinetic proofreading by diffusion is shown to produce reasonably accurate proteins in spite of the anticipated low level of intrinsic discrimination of a primal tRNA for its cognate amino acid. Although previous speculations on the evolution of the protein-synthetic machinery cannot be definitively examined in the laboratory, the fundamental hypothesis of this paper leads directly to critical experimental tests.

Primal cells occurred after the era of a primordial prebiotic soup (18, 19), but before highly accurate protein synthesis had yet evolved (20). The cytoplasm of a primal cell would have contained activated amino acids and nucleic acid monomers, RNA or DNA informational polymers, and very simple enzymes.

Relatively accurate reproduction of DNA or RNA informational polymers is possible with nonspecific enzymes or nonbiological catalysts (21). Because of the higher energy of mismatch of noncomplementary base pairs, a simple condensation polymerization (using a nonspecific catalyst) of a mixture of appropriately activated nucleic acid monomers should be capable of yielding a complementary strand to single-stranded DNA or RNA with error rates of 0.1% to 1%. Thus with a 100-fold increase in error rate, nucleic acid replication in modern cells extrapolates smoothly to primal cells.

When one extrapolates present-day protein synthesis back to a primal cell, many present-day aspects remain invariant. mRNA can be produced to contain the relevant sequence information with reasonable accuracy. A crude ribosome will suffice (22–24), because the recognition between tRNA and mRNA is carried chiefly in base pairing. The ribosome need provide only alignment and a general catalyst for covalent bond transfers. It is plausible that very simple proteins and rRNA would permit the production of proper sequences of small proteins from a set of charged (aminoacyl) tRNAs and mRNA.

Charging tRNA presents a paradox to constructing the biochemistry of primal cells. In modern cells, the dictionary by which the cells determine the correspondence between amino acids and codon triplets is provided solely by the approximately 20 aminoacyl-tRNA synthetases. Each type of synthetase selects independently one amino acid and any one of its cognate tRNAs as substrates, rejecting competing substrates to the 0.01% level. If these enzymes were removed from a cell and substituted by a set with a different correspondence between the amino acid and the tRNA binding sites, protein synthesis would be virtually unaltered, but the genetic code would be changed. In modern cells, the synthetases are not general catalysts, (which could be relatively small and simple), but are instead large, highly specific enzymes (25, 26). This presents an evolutionary paradox. Proteins cannot be made accurately without highly specific synthetases to provide the genetic code; but the synthetases are themselves presumably accurately made proteins.

The structure of modern tRNA causes this paradox. The base stacking and covalent backbone of an *E. coli* tRNA is shown in Fig. 1 *upper*, in such a way as to preserve the significant details of the three-dimensional structure seen by x-ray diffraction (27, 28). The acceptor stem is about 60 Å removed from the anticodon loop. The sequence (and presumably the structures) of the acceptor termini of all tRNAs are very similar. Thus there

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

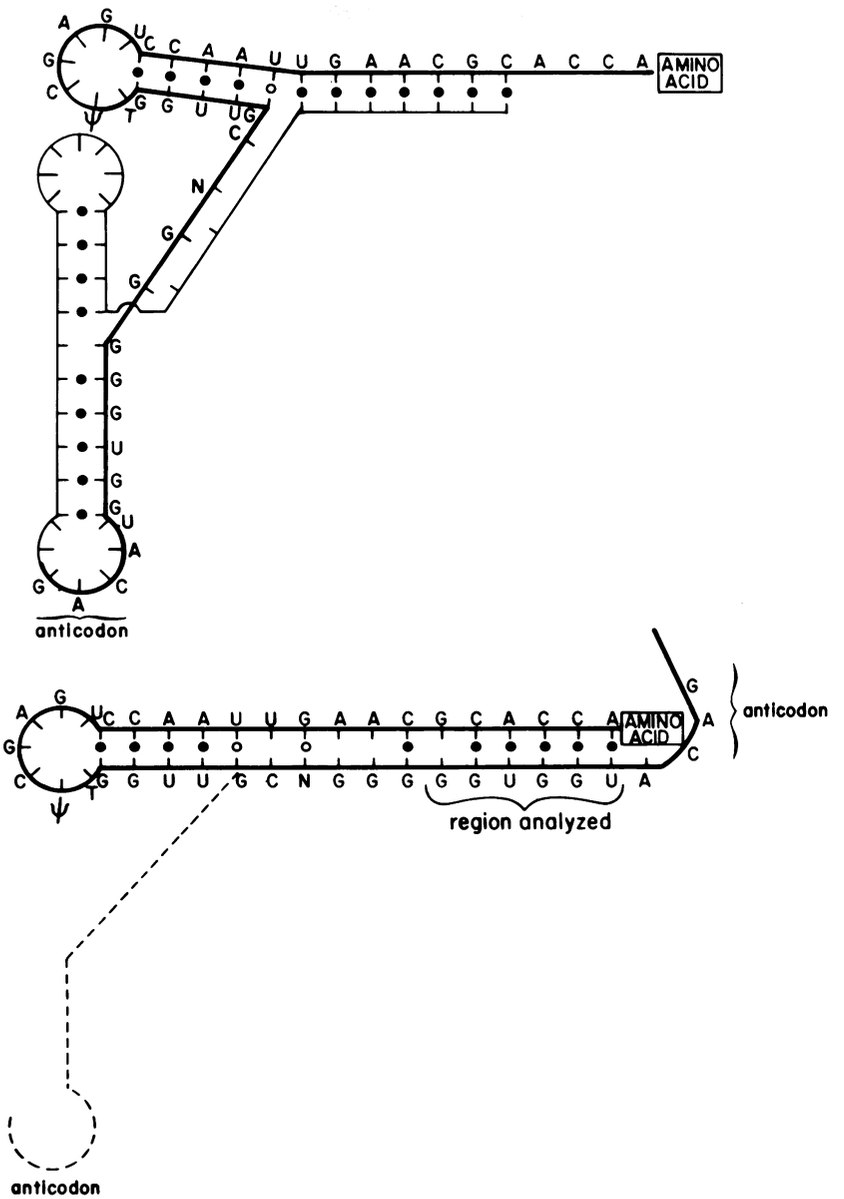

FIG. 1. (*Upper*) The base pairing and stacking of the sequence for *E. coli* $tRNA^{Val}$, positioned in the structure determined for present-day yeast $tRNA^{Phe}$. The part of the backbone used in Fig. 1 *Lower* is indicated by a heavy line. (*Lower*) The same *E. coli* $tRNA^{Val}$ sequence folded into the structure hypothesized for primal tRNA. The structure is drawn extended, but the central region could loop out, depending on sequence in this region. The region for which the pairing was statistically studied is indicated.

is no way by which the amino acid can directly sense which tRNA it is bound to. The selective charging of tRNA takes place only through using an extended enzyme with recognition sites for both an amino acid and the anticodon or other parts of a tRNA molecule. I propose a solution to this "chicken and egg" paradox by a hypothesis on the structure and operation of primal tRNA that obviates the need for synthetases in primitive cells.

Primal tRNA is assumed to differ sufficiently in sequence from modern tRNA in that it had an alternate secondary/tertiary structure, *which was more stable* (for the primal sequences) than the modern folding (Fig. 1 *upper*). The posited alternate secondary structure and base stacking is shown in Fig. 1 *lower*. (Part of the covalent backbone is not drawn because its unknown location has no direct bearing on the discussions.) The stacking in Fig. 1 *lower* is obtained from Fig. 1 *upper* by opening the acceptor stem and the anticodon stem and allowing the strand leading from the TψC loop toward the anticodon to continue back up the acceptor stem. The number of bases in *E. coli* tRNA then places the anticodon triplet very close to an amino acid bound to the 3′ end of the tRNA in the usual fashion. This primal tRNA structure was chosen because it is the simplest refolding pattern that locates the anticodon in immediate proximity to the amino acid.

If modern tRNA evolved from a primal form having a different base-pairing pattern, two bases which were Watson–Crick paired in primal tRNA should tend to have such a correspondence in modern tRNA, even though these two bases are not in opposition in the modern structure. We next test the primal hypothesis by a statistical study of such correspondences in the modern tRNA sequence. These correspondences can be visualized directly by folding modern tRNA into the primal form and seeing what standard base pairings of previous nonopposing regions result. For statistical reasons, we wish to restrict the study to a single primative species with many tRNAs sequenced. Thus the analysis will be restricted to *E. coli*.

The sequences of all 20 *E. coli* tRNAs given in Barrell and Clark (29) were folded into the hypothesized primal shape. (In

$tRNA^{Leu}$, $tRNA^{Ser}$, and $tRNA^{Tyr}$, the "extra loops" were reduced to normal size.) These 20 tRNAs represent 20 different codons and 14 different amino acids. (The 21st *E. coli* tRNA given duplicates a valine codon already included, and was omitted.) Fig. 2 *left* shows a histogram of the number of Watson–Crick G-C or A-U pairs made in the structure shown in Fig. 1 *lower* involving the six base oppositions at the right whose 5′ sequence comes from the right-hand half of the anticodon stem. In the valine example illustrated, five of these six base oppositions make Watson–Crick pairs. The average number of Watson–Crick pairs in this region is $\bar{n} = 2.50 \pm 0.32$ for the 20 tRNAs. The standard deviation of this average number, if all cases are viewed as independent, is $\{(\bar{n}^2 - \bar{n}^2)/19\}^{1/2} = 0.32$. The number $\bar{n}_0 = 6(1/4) = 1.5$ is expected if there is no reason for Watson–Crick pairing. A χ-square comparison of the data in Fig. 2 *left* with a binomial distribution for $\bar{n} = 2.5$ (Fig. 2 *right*) shows a good fit. A *random* binomial distribution of 20 tries will fit the average binomial distribution (Fig. 2 *right*) less well than does the sample (Fig. 2 *left*) 58% of the time.

There are many ways that $\bar{n}$ could exceed that based on random assignments—for example, if the entire tRNA contained only G and C, the *a priori* estimate of a random correct pairing would be greater than $6(1/4)$. To investigate such biases I counted base pairings for a comparison of the six bases (*i*) with the acceptor stem frame-shifted in register one to the right, (*ii*) with it frame-shifted by one to the left, and (*iii*) with a particular anagram of the acceptor stem. For these three cases, values of $\bar{n}$ (± SD) were, respectively, $1.85 \pm .29$, $1.30 \pm .18$; 1.20 ± 0.23, consistent with random pairing.

What is the statistical significance of $\bar{n} = 2.5$? When the 20 tries are taken as independent, the difference between $\bar{n}$ and 1.5 is more than 3σ, an event whose random occurrence has a probability $P < 0.002$. However, one might easily argue that the inter-relations of different tRNA studies imply that less than 20—perhaps 10—of the pairings studied are really independent cases. In that case, the estimate becomes $\bar{n} = 2.50 \pm 0.46$, and $\bar{n}$ differs from 1.5 by 2.2σ with a corresponding $P < 0.05$. An alternate way of evaluating the histogram of Fig. 2 *left* is to examine the improbable events. For random pairing, the probability of finding five or greater numbers of pairs in a single trial is 0.0046 if $\bar{n}_0 = 1.5$. Thus for 20 independent trials, the probability that *at least* one trial be a five or a six is 0.092. If the sequences being tried are *not* independent, the probability of obtaining at least one trial which yields a five or a six will be *reduced*. Thus the occurrence of at least one event in the "five" bin is significant on the $P < 0.1$ level, and is significant on the $P < .05$ level if there were really only 10 independent tRNAs.

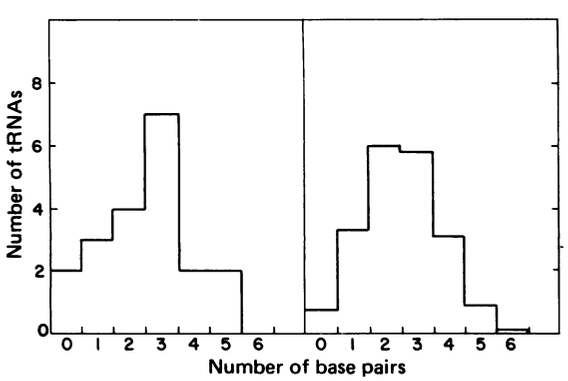

FIG. 2. (*Left*) The distribution of the number of *E. coli* tRNAs having various numbers of Watson–Crick base pairs in the six oppositions indicated nearest the amino acid in the structure of Fig. 1 *Lower*. (*Right*) A comparison average histogram for a binomial distribution with $\bar{n} = 2.5$.

The probability of two or more events in the combined five-six bin in 20 trials is less than 0.005 for random pairing. This $P < 0.005$ level of significance is to be associated with the results if the two matches of five bases ($tRNA^{Leu}$ and $tRNA^{Val}$) are independent events.

How much residual base pairing would be expected to now exist if primal tRNA changed to the modern form in which the primal pairing constraint on sequence is not operative? If τ is the time since synthetases developed, W is the rate of replacement of a base in the population by mutation and evolution, and all mutational changes are equally likely, then the probability of finding now a base pair in a given correspondence (30) is

$$P = 0.25 + 0.75e^{1.33W\tau}.$$

Modern cells have a low mutation rate in base-paired and tertiary-structured rRNA. Values (mean ± SD) of W have been found to be $1.8 \pm 0.5 \times 10^{-10}$ by comparing human and *Xenopus* 5S RNA, and $1.4 \pm 0.2 \times 10^{-10}$ by comparing yeast and rat 5.8S RNA (30). Blue-green algae similar to modern cells are believed to have been in existence more than 2.5×10^9 years (31), and thus correspond to an evolutionary age $W\tau > 0.4$. An $\bar{n}$ of 2.50 ($P = 0.42$) implies an evolutionary age of $W\tau = 1.1$ if all six bases were paired in primal tRNA and slightly less if five of the six were paired. Thus the primal tRNA appears appropriately older on an evolutionary scale, if tRNA substitutions become fixed in a population at the rRNA rate. The possibilities of error rates which are high in early epochs and of frame-shifts in evolution make this estimate only on the upper limit of the chronological age.

Thus in spite of the ambiguity of the effective number of independent trials, the distribution of Fig. 2 *left* is significantly different from an $\bar{n}$ of 1.5 on at least the level of $P < 0.05$, and quite possibly at the level $P < 0.01$. This appropriately small remnant of base pairing when modern tRNA is folded into the posited primal structure provides "fossil" evidence for the primal tRNA structure.

A second normally higher energy structure in which the anticodon is in a spatial position removed from the amino acid binding site is assumed for primal tRNA. This configuration must be stabilized by interactions with the primal ribosome, and it enables an unencumbered anticodon to interact with the message codon. This second structure might more nearly resemble the structure of modern tRNA. Alternate folding patterns for tRNA (32, 33) and 5S RNA (34) are well documented. The structure of free tRNA is believed to differ considerably in the TψC loop region (27, 28, 35) from the structure the tRNA assumes on the ribosome (36, 37). This region of the free molecule is folded inward and inaccessible, whereas on the ribosome it is believed to base-pair with tRNA. Therefore modern tRNA cannot be viewed as having only a single, unique conformation, and primal tRNA may also plausibly have exhibited multiple functional structures.

The structural change between modern and primal tRNA is then only a choice between the relative stabilities of two alternative secondary structures, rather than a discontinuity between two primary structures. The primal tRNA hypothesis fits smoothly into an evolutionary sequence. A hairpin loop of RNA with only a few paired bases and an exposed random sequence which could function as an anticodon could provide an earlier ancestor.

A structure with the anticondon in intimate contact with the amino acid might result in a modest selectivity of a particular anticodon for a particular amino acid in the unknown charging process in primal cells. More important to the following discussion, the binding of the anticodon region to the amino acid

should protect the aminoacyl linkage between tRNA and amino acid against hydrolytic attack. A poorer binding between these two should yield less protection.

A general weakness of the "amino acids select their codons" viewpoint is that the low level of fidelity expected produces poor protein yields. Suppose primal cells used five different amino acids (or classes of amino acids) and needed to make specific elementary proteins 10 amino acids long for their catalysts (or equivalently that 10 sites in a longer sequence must be correct for enzyme function). A nonselective system would have a yield of $(1/5)^{10} \approx 10^{-7}$ for decapeptides of the proper sequence. If the protection of the acyl linkage by the anticodon produced a stabilization of 0.5 kcal (1 cal = 4.184 J), the "off" kinetic constant k for the reaction

$$\text{tRNA·AA} \rightarrow \text{tRNA} + \text{AA}$$

for correct matches will be only one-half the value k_e for erroneous matches. If tRNA is then charged with random amino acids, and the charging and hydrolysis reactions reach a dynamic steady state, the probability of correct charging becomes $(1 + 4k/k_e)^{-1} = (1/3)$. The probability of making the desired decapeptides correctly is then $(1/3)^{10} \approx 1.8 \times 10^{-5}$, an abysmal yield. For that reason, discrimination energies of 0.5 kcal are generally viewed as insignificantly small for producing primal enzymes.

However, the accuracy and yield of this system is *not* limited to such small values, because a time-delay aspect of kinetic proofreading (38) can be used. The dynamic steady state selectivity in homogeneous chemistry is based on the fact that when a tRNA molecule is charged, the probability at time t later that it is still charged decays exponentially, with a mean life that is different for correct and incorrect charging. The ratio of the dynamic steady state populations (if there is no selectivity in charging) equals the ratio of areas under the two solid curves of Fig. 3.

Suppose now that the site of charging tRNA and the site of protein synthesis are physically separated, with a mean diffusion time τ between these two locations. The probability that a tRNA molecule which leaves the charging site at time zero arrives at the primal ribosome before time t is given by the dashed curve in Fig. 3. During the diffusion period the acyl linkage is exposed to hydrolytic attack. The probability of using a particular tRNA, charged at time zero, is proportional to the

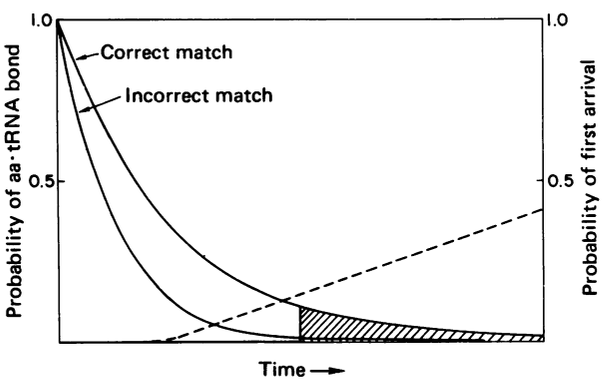

FIG. 3. The solid lines show the probability of the continued existence of correctly and incorrectly charged tRNA at a time t after charging. The dashed curve gives the probability of availability at the ribosome (physically separated from the charging site) at time t after charging. The actual use of aa·tRNA is proportional to the area under the product of solid and dashed curves. The general effect of multiplying by the dashed curve is to reduce the error fraction from the relative areas under the two solid curves to the ratio of the two shaded areas.

area under the product of the exponential solid lines and the dashed diffusion arrival probability. The diffusion arrival probability switches on fairly quickly after an initial lag, and can qualitatively be viewed as switching on the use of an amino acid after a time delay. The fraction of errors can be described qualitatively by the relative areas under the solid curves after a time delay (shaded), *and it is much smaller* than without the time delay. For a value of τ chosen so that 5% of the correctly charged tRNA survives to be used in protein synthesis, a misacylated molecule has only a 0.5% probability of being used. In the same example used before, the probability of correctly putting in an amino acid at a single site becomes (1/1.4), and the probability of making a decapeptide correctly is $(1/1.4)^{10} \approx 0.03$. If a cell were to need a modest number of enzymes, a 3% yield may be quite adequate.

This illustration shows that a *modest* discrimination against hydrolytic attack—too small a discrimination to appear useful at first sight—can, in conjunction with a kinetic proofreading using diffusion and spatial separation, provide the fidelity in tRNA charging necessary to get substantial yield of correctly made oligopeptides. A rudimentary form of the genetic code may plausibly have been already irreversibly fixed when the discrimination energies were $\lesssim 0.5$ kcal. For discriminations in modern experiments, a rate ratio of 2 or less may be significant. (The enhancement of accuracy by using kinetic proofreading by diffusion does not require the particular primal structure, but it does require *some* chemical mechanism of intrinsic recognition of an amino acid by its tRNA.)

The level of discrimination and which bases are important in amino acid–anticodon interactions must be determined experimentally. The possibility of appropriate discriminations in primal tRNA is enhanced by the relatively rigid structure available at the double helix end and by the fixed location of the amino acid. Model building suggests that the NH_2 of an amino acid acylated to the 2′ OH can form a hydrogen-bond with the purine adjacent to the first base of the anticodon. Its residue is then in contact with the first base of the anticodon. Significantly, this contact is made only for L amino acids. Thus the primal structure and model building provides a rationale for which end of the anticodon is most significant; for the invariant purine adjacent to the first position of the anticodon; for the handedness of amino acids used in protein synthesis (given the handedness of the nucleic acids); and for the fact that ribose (rather than deoxyribose) nucleic acids are used for the function of tRNA.

The hypothesized structure provides a conceptual foundation for experiments. tRNA can be modified or RNA synthesized to form a stem having the structure of the proposed primal tRNA. A particular amino acid can be linked to the acceptor terminus, and the effectiveness of different sequences in the single-strand region in slowing deacylation studied. For such *in vitro* studies (as *in vivo*), using an appropriate time delay should amplify the relatively weak discriminations anticipated. The central questions to be asked include: do the experiments generate a genetic code (particularly in conjunction with ideas about the abundances of early amino acids)? If so, how is it related to *the* genetic code? Is there an explanation for the particular 20 amino acids used in protein synthesis? Which bases in the sequence are most important? A determination of the sequences of tRNA from presumed primitive life, such as blue-green algae or methanogenic bacteria, provides a possibly different kind of test of the primal hypothesis. Papers on the origin of the genetic code too often end with a disclaimer that the evidence available—present and probably future—cannot fully resolve the issue raised. There are, in contrast, direct tests possible for these ideas.

I thank T. Yamane and J. Fresco for discussions. The work at Princeton was supported in part by National Science Foundation Grants DMR-75-14264 and DMR-78-05916.

1. Gamow, G. (1954) *Nature (London)* **173,** 318–320.
2. Crick, F. H. C. (1968) *J. Mol. Biol.* **38,** 367–379.
3. Ratner, V. A. & Batchinsky, A. G. (1976) *Origins Life* **7,** 225–228.
4. Eigen, M. (1971) *Naturwissenschaften* **58,** 465–523.
5. Woese, C. R. (1969) *J. Mol. Biol.* **43,** 235–240.
6. Woese, C. R. (1967) *The Genetic Code* (Harper & Row, New York).
7. Orgel, L. E. (1968) *J. Mol. Biol.* **38,** 381–393.
8. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. (1966) *Proc. Natl. Acad. Sci. USA* **55,** 966–974.
9. Jukes, T. H. (1977) *Comprehensive Biochem.* **24,** 235–293.
10. Jukes, T. H. (1973) *Nature (London)* **262,** 22–26.
11. Crick, F. H. C. & Orgel, L. E. (1973) *Icarus* **19,** 341–346.
12. Reanney, D. & Ralph, R. (1967) *J. Theor. Biol.* **15,** 41–48.
13. Woese, C. R. (1968) *Proc. Natl. Acad. Sci. USA* **59,** 110–117.
14. Saxinger, C. & Ponnamperuma, C. (1971) *J. Molec. Evol.* **1,** 63–73.
15. Raszka, M. & Mandel, M. (1972) *J. Molec. Evol.* **2,** 38–43.
16. Pelc, S. R. & Welton, M. G. E. (1966) *Nature (London)* **209,** 868–870.
17. Dunnill, P. (1966) *Nature (London)* **210,** 1267–1268.
18. Oparin, A. I. (1964) *The Chemical Origin of Life*, translator, Synge, A. (Thomas, Springfield, IL).
19. Miller, S. L. & Urey, H. C. (1959) *Science* **130,** 245–251.
20. Rich, A. (1965) *Evolving Genes and Proteins*, ed. Bryson, V. & Vogel, H. G. (Academic, New York), pp. 453–468.
21. Weimann, B. J., Lohrmann, R., Orgel, L. E., Schneider-Bernloehr, H. & Sulston, J. E. (1968) *Science* **161,** 387.
22. Spirin, A. S. (1976) *Origins Life* **7,** 109–118.
23. Crick, F. H. C., Brenner, S., Klug, A. & Pieczenik, G. (1976) *Origins Life* **7,** 389–397.
24. Woese, C. (1970) *Nature (London)* **226,** 817–820.
25. Söll, D. & Schimmel, P. R. (1974) *The Enzymes* **10,** 489–538.
26. Loftfield, R. B. (1972) *Prog. Nucleic Acid Res. Mol. Biol.* **12,** 87–128.
27. Kim, S. H., Sussman, J. L., Suddath, F. L., Quigley, G. S., McPherson, A., Wang, A. M. J., Seeman, N. C. & Rich, A. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 4970–4974.
28. Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C. & Klug, A. (1974) *Nature (London)* **250,** 546–551.
29. Barrell, B. G. & Clark, B. F. C. (1974) *Handbook of Nucleic Acid Sequences*, (Joynson-Bruvvers, Oxford, England).
30. Hori, H., Hijo, K. & Osawa, S. (1977) *J. Molec. Evol.* **9,** 191–201.
31. Schopf, J. W. (1970) *Biol. Rev.* **45,** 319–335.
32. Lindahl, T., Adams, A. & Fresco, J. (1966) *Proc. Natl. Acad. Sci. USA* **55,** 941–948.
33. Ishida, T. & Sueoka, N. (1967) *Proc. Natl. Acad. Sci. USA* **58,** 1080–1087.
34. Richards, E. G., Lecanidou, R. & Geroch, M. E. (1972) *Eur. J. Biochem.* **34,** 262–271.
35. Yoshida, M., Kaziro, Y. & Ukita, T. (1968) *Biochem. Biophys. Acta* **166,** 646–652.
36. Jordan, B. R. (1971) *J. Mol. Biol.* **55,** 423–439.
37. Schwarz, U., Menzel, H. M. & Gassen, H. G. (1976) *Biochemistry* **15,** 2484–2490.
38. Hopfield, J. J. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 4135–4139.