was noted again, and the spheres became complex (Fig. 3). The preparation was then allowed to dry again. We carried out another rehydration with water, and the same phenomena were observed except that the spheres obtained seemed to be even more complex in regard to internal structure. After three or four more cycles of dehydration and rehydration the degree of complexity in the spheres seemed to have reached a maximum. When the foregoing steps were repeated in the presence of methylene blue, the spheres concentrated the stain.

The factors which may cause the development of complex morphological forms in the system are numerous. During the drying process, there are localized high concentrations of proteinoid material. The first rehydration with buffer seems to be an important factor in causing development of complexity. This change is probably due to interaction of the buffer ions with the proteinoid.

Thus simple environmental changes of dehydration and rehydration can indeed cause the development of complexity in an experimental prebiological system; this finding seems to support the suggestion of Hinton and Blum (*2*). We believe that we have developed, from thermal proteinoid, a system which behaves like coacervates (*5, 6*). The large spheres described above have several properties in common with coacervates: complex morphology, selective adsorption as evidenced by concentration of methylene blue, and ability to coalesce. Our experiments may possibly link the seemingly different concepts of the origin of life, the coacervates of Oparin (*6*) and the proteinoids of Fox (*1*).

ADOLPH E. SMITH
FREDERICK T. BELLWARE
Physics Department,
Sir George Williams University,
Montreal 25, Quebec

References and Notes

1. S. W. Fox, in *The Origins of Prebiological Systems*, S. W. Fox, Ed. (Academic Press, New York, 1964), p. 361.
2. H. E. Hinton and M. S. Blum, *New Sci.* **28**, 270 (1965).
3. S. W. Fox and K. Harada, *J. Amer. Chem. Soc.* **82**, 3745 (1960).
4. R. S. Young, in *The Origins of Prebiological Systems*, S. W. Fox, Ed. (Academic Press, New York, 1964), p. 254.
5. H. G. Bungenberg de Jong, in *Colloid Science*, H. R. Kruyt, Ed. (Elsevier, New York, 1949), vol. 2, p. 433.
6. A. I. Oparin, *The Origin of Life on the Earth* (Academic Press, New York, 1957).
We thank M. E. Burns and R. S. J. Manley for helpful discussions. Supported by the National Research Council Grant A-2528.

2 February 1966

Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences

Abstract. *The structure of present-day ferredoxin, with its simple, inorganic active site and its functions basic to photon-energy utilization, suggests the incorporation of its prototype into metabolism very early during biochemical evolution, even before complex proteins and the complete modern genetic code existed. The information in the amino acid sequence of ferredoxin enables us to propose a detailed reconstruction of its evolutionary history. Ferredoxin has evolved by doubling a shorter protein, which may have contained only eight of the simplest amino acids. This shorter ancestor in turn developed from a repeating sequence of the amino acids alanine, aspartic acid or proline, serine, and glycine. We explain the persistence of living relics of this primordial structure by invoking a conservative principle in evolutionary biochemistry: The processes of natural selection severely inhibit any change in a well-adapted system on which several other essential components depend.*

Many of the principles of organic evolution have long been known and are productively used in the organization of biological concepts, but are seldom used to full advantage in biochemistry. In nature, biochemistry is included in biology. An organism is a functioning system composed of the structures, organs, tissues, and organelles of classical biology. These in turn are composed of metabolites, macromolecules, enzyme aggregates, and biochemical feedback systems. Biochemical details concerning these components have only recently become accessible. Potentially, a much greater amount of information relevant to evolution is available in biochemistry than in classical biology.

According to evolutionary theory, each structure or function of an organism is subject to occasional changes or mutations, but the infrequency of these mutations necessitates that they will almost always occur, and be selected for, one at a time. Each change or addition must be an improvement, or at least not too severe a disadvantage, in order that the processes of natural selection permit its survival. This limitation has a very conservative effect. If its ecological niche stays the same, a well-adapted organism strongly resists change. Thus we find familiar-looking fossil shells a third of a billion years old. If its niche changes, new functions evolve, but the most primitive structures tend to remain unchanged, since these older components have already come to be relied upon by several later additions. Any change in a very old component, even though it might be advantageous in some way, would coincidentally disturb so many other things that it would almost always be extremely disadvantageous to the organism. This conservatism is well illustrated in the amino acid sequences of proteins. For example, we can compare the amino acid sequences of cytochrome *c* from yeast (*1*) and from horse (*2*), position by position. In 64 of the 104 positions the amino acids in the two chains are identical. Between horse and human cytochrome *c* (*3*) there are only 12 amino acid differences.

When we consider evolution retrospectively, the constraints are even more severe. One basic evolutionary principle is that every living organism or structure or function had ancestors very similar to itself, but simpler. (This is true even if it had more complex *immediate* ancestors.) In a particular case there are generally only a few plausible slightly simpler ancestors. As we trace the changes in a structure or function back through time, we must bear in mind that *all* of the structures and functions of the cell may be simpler. We are then dealing with primitive components ancestral to those seen today.

The amino acid sequence of ferredoxin from *Clostridium pasteurianum*, a nonphotosynthetic anaerobic bacterium, has been reported (*4*). This protein seems to have arisen at an earlier times than many others which have been studied. We draw this inference from the following considerations.

1) Ferredoxin occurs in primitive anaerobic organisms, both photosynthetic and nonphotosynthetic (*5*). It must have been present in simpler organisms, the extinct common ancestors of these.

2) Ferredoxin contains iron and sulfur, bonded to the protein at its active site (*6*). Ferrous sulfide, FeS, is a widely dispersed mineral, a catalyst which would have been readily available to the most primitive organism.

3) The functions of ferredoxin are basic to cell chemistry. The reduction of ferredoxin is the key photochemical event in photosynthesis by chloroplasts (*5*). All the energy is channeled through this compound to other cellular energy-storage mechanisms. Ferredoxin is the most highly reducing stable compound so far found in the cell, having a reducing potential near that of molecular hydrogen (5). This suggests that its function may have evolved at a very early time when the earth's atmosphere was still strongly reducing. It reduces nicotinamide adenine dinucleotide (NAD) (*5*), a ubiquitous reducing agent in the cell. Therefore, it may be even more primitive than NAD. It catalyzes adenosine triphosphate (ATP) formation by radiation (5). This indicates possible relation to primitive energy transfer processes. It catalyzes the synthesis of pyruvate from carbon dioxide and acetylcoenzyme-A (*5*). This indicates its involvement with one of the simplest, most primitive synthetic processes in intermediary metabolism, the fixation of CO_2. It participates in nitrogen fixation (*7*) and hydrogenase-linked reactions (*5*).

4) Ferredoxin contains an unusually high proportion of the smaller, more thermodynamically stable (*8*) amino acids, such as glycine, alanine, cysteine, serine, and aspartic acid. Furthermore, their synthesis from inorganic substrates by autotrophic organisms requires only a small number of endothermic steps.

5) Ferredoxin is smaller than most other enzymes, having only 55 amino acid units. It appears to have some sort of repeating structure, so that it may once have been still smaller.

Let us now consider the amino acid sequence of ferredoxin. Applying the constraints imposed by the principles of evolution, can we find traces of its ancestry?

Figure 1 shows the reported sequence of ferredoxin (*4*) and some manipulations with it. For the purposes of our study, a single-letter notation (*9*) is much more suitable than the usual three-letter notation (row 1). If links 30 to 55 (row 3) are placed under links 1 to 29 (row 2), it is evident that the number of coincidences (row 4) far exceeds that which would be expected by chance ($P << .001$). It appears that this protein has evolved by doubling of the nucleic acid (gene) which determines it. Smithies, Connell, and Dixon showed how nonhomologous crossing over of chromosomes could produce this effect, and proposed it as the explanation of the apparent doubling which they detected in haptoglobin molecules (*10*). Because of the very high statistical improbability of the chance occurrence of the twelve coincidences, we consider for this study that the two halves of the ferredoxin sequence are in fact homologous, and attempt to decipher some details about their common ancestor. Presumably the ancestral sequence contained all those amino acid units which are common to both parts. Where the units are different, probably one of the two was present originally.

The ancestral sequence of 29 units must itself have had a simpler ancestor. An attempt was made to discover this by the same process, but no further regularities could be found by superimposing quarters of the chain.

Another kind of simplicity would be that the ancestor had fewer kinds of amino acids. During the evolution of the genetic code there must have been a time when the genetic mechanism could discriminate fewer amino acids. Perhaps those which coincide in

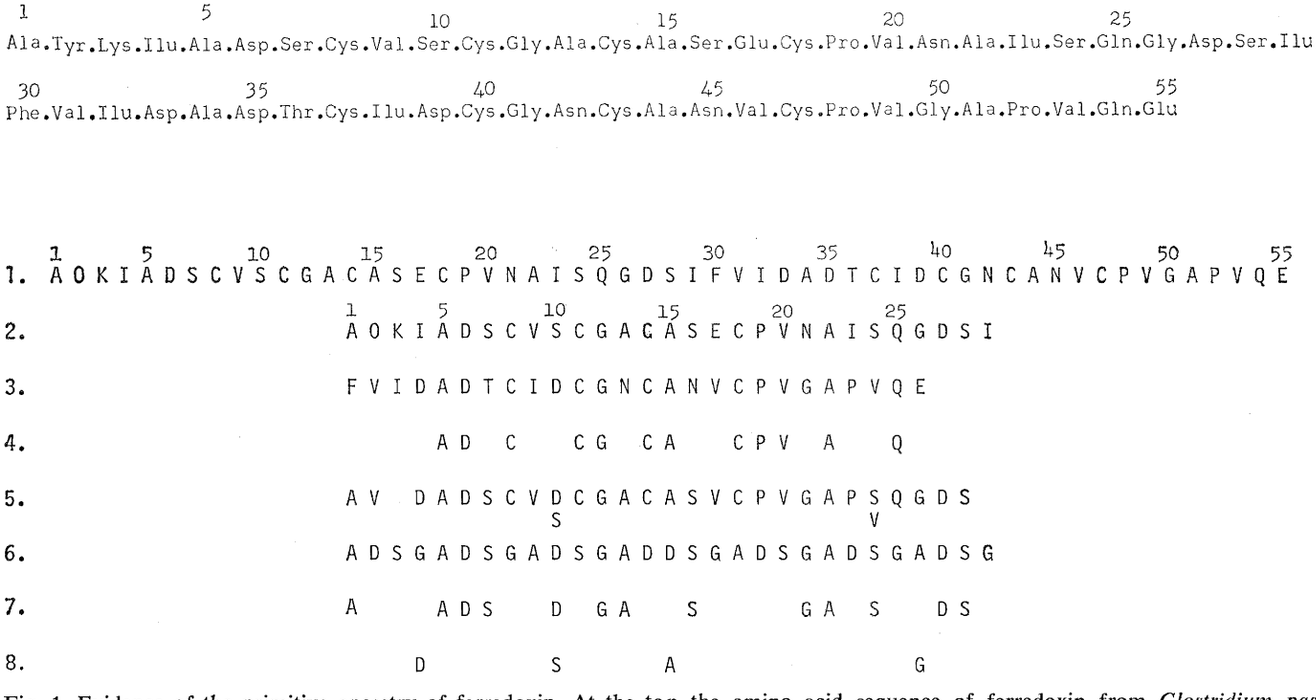

Fig. 1. Evidence of the primitive ancestry of ferredoxin. At the top the amino acid sequence of ferredoxin from *Clostridium pasteurianum* is shown in conventional notation (*4*). Row 1: The sequence is translated into a one-letter code (*9*), a notation more suitable for this type of study. Rows 2 and 3: The two halves of the chain compared by alignment. Row 4: Twelve amino acids which are identical in both halves. If the sequences were unrelated, one would expect only two or three such identities. The probability of finding as many coincidences as this by chance is negligible. Row 5: The same seven simple amino acids found in row 4, plus serine, whenever they occur in *either* chain. Row 6: A simple repeating sequence of four amino acids (with a discontinuity at position 15) from which row 5 appears to be derived. Row 7: Thirteen amino acids from row 5 which conform to row 6. Row 8: The four amino acids which do not conform to the cyclic pattern. The chance probability of as many coincidences as this would be very small if the pattern in row 6 had been independently given. Since it was derived from this study itself, the coincidence is less extreme, but still seems to be good evidence for the validity of the cyclic pattern.

the two halves of the ferredoxin sequence (glycine, cysteine, alanine, proline, valine, aspartic acid, and glutamine) may be survivors of this early stage. And perhaps when any of these occurs in either half, it is also likely to be a survivor of the earlier stage. Therefore, we record the positions and occurrence of these seven amino acids or serine (see below) in the paired ferredoxin sequences (row 5 of Fig. 1). In this arrangement, two regular patterns emerge. Cysteines occur in a cycle of three, and alanines occur in a cycle of four. A discontinuity in both patterns occurs at the midpoint, position 15. None of the other amino acids confirms a cycle of three. However, the cycle of four is confirmed by glycine, proline, aspartic acid, and serine. A repeating sequence of four amino acids (with a break at position 15) has been written in row 6. The combined halves agree with this cycle in 13 positions (row 7); they disagree in only four positions (row 8). Altogether 17 occurrences of these four simple amino acids in the living ferredoxin chain agree with this cycle, and five disagree with it. Figure 2 shows the reported sequence rewritten in groups of four for direct comparison with the repeating pattern.

The probabilities involved in a pattern of this kind are difficult to compute, because the pattern was discovered rather than predicted. Some sort of pattern can always be found if random data are examined in fine detail. In principle, one should modify the computed probability to reflect all the other patterns which would have been considered equally good if they had been discovered. This number is difficult to determine and must be somewhat arbitrary. However, in this case it does not seem to be very large, and the number of coincidences still seems beyond ordinary chance. We consider the observed number of coincidences to be good evidence for the pattern of a cycle of four.

Using a computer program, we have matched the sequence of ferredoxin against itself in all combinations, and against the various possible cycles. The result of this objective method is the same as found by inspection. Only cysteines occur in a cycle of three. Alanine, aspartic acid, serine, and glycine agree with a cycle of four; and proline in three places occupies the position of aspartic acid. This substitution is reasonable stereochemically, since the two side-chains have a very similar

Position				
1	A	*O*	*K*	*I*
30	*F*	*V*	*I*	D
5	A	D	S	*C*
34	A	D	*T*	*C*
9	*V*	S	*C*	G
38	*I*	D	*C*	G
13	A	*C*		
42	*N*	*C*		
15		A	S	*E*
44		A	*N*	*V*
18	*C*	P	*V*	*N*
47	*C*	P	*V*	G
22	A	*I*	S	*Q*
51	A	P	*V*	*Q*
26	G	D	S	*I*
55	*E*			
Cycle	A	D	S	G
		P		

Fig. 2. Evidence for a cycle of four in ferredoxin. The two half-chains, rows 2 and 3 in Fig. 1, are written in successive groups of four, with a break at positions 15 and 44. Alanine (A) occurs mainly in the first column, D and P in the second, S in the third, and G in the fourth. Exceptions are underlined. All other amino acids are shown in italics. This good fit appears unlikely to be due to chance, but a numerical evaluation of the probability would involve several arbitrary assumptions.

conformation, when aspartic acid folds into a hydrogen-bonded intramolecular ring. Aspartic acid is metabolically simpler to synthesize than proline, and therefore seems likely to have evolved earlier.

We will now use these patterns and the rules of evolution to reconstruct the history of the ferredoxin molecule. We first consider the prosthetic group, or inorganic part, in a prebiological environment. Then starting with an extremely primitive living system, we follow the development of the increasing intricacy of the protein.

At chemical equilibrium in an ideal gas mixture of a reducing nature at standard temperature and pressure, H_2S, CH_4, CO_2, H_2O, and N_2 predominate; ammonia and organic compounds occur in small amounts (*11*). At equilibrium, FeS is stable in this gas mixture. Possibly life may have organized about such a stable inorganic catalyst, one that could participate in capturing photons and in directing energy toward the reduction and fixation of CO_2 and toward the synthesis of pyrophosphate. In this photon-activated system, other, less stable catalysts may then have been synthesized, permitting development of more complex systems.

Regardless of how life originated, there was at one time a very primitive organism, far simpler than any known to be living today. It was capable of making and polymerizing some of the simplest amino acids and nucleotides. Perhaps the nucleic acids were made of only two nucleotides, simpler than the ones which occur in the present genetic mechanism. A variety of such polymers could be formed having useful structural or catalytic functions, for which natural selection preserved them.

One sequence of 12 nucleotides doubled and redoubled itself, making a longer, repetitive chain.

At about the same time, the primitive amino-acid-polymerizing mechanism of the cell began to utilize this nucleic acid chain as a template, and it coded for the amino acids alanine, aspartic acid, serine, and glycine. If these events occurred when the nucleotide chain was still only 12 units long, the resulting peptide would be A D S G, as in Fig. 3, row 1. After the nucleotide chain became longer, it coded for a simple repeating protein, A D S G A D S G . . . (row 2). This protein had some advantageous, perhaps structural, function in the cell, unrelated to the present energy-transfer function of ferredoxin. An aberration in the nucleotide sequence produced a break in the cycle (row 2, underlined).

The synthesizing abilities of the organism became more versatile and more efficient. The genetic code became more complex, so that the genetic mechanism was able to incorporate other amino acids. Mutations occurred which modified and complicated the particular amino acid sequence which we are following (row 3).

Cysteine was among these new amino acids added to our sequence. The sulfide bond of the cysteine unit became attached to iron sulfide. The protein thus cooperated in the photon-coupled catalytic function, which it still retains, and became protoferredoxin. On the principle that evolution proceeds one step at a time, we assume that the cell was already using iron sulfide as a catalyst, probably attached to cysteine alone, or to some peptide less suitable than protoferredoxin. This new attachment would merely have increased the efficiency of this function.

Eventually four cysteines were added by mutation, and two identical chains combined to make an intricate protein-iron-sulfide complex of greatly increased efficiency. It still retains essentially this structure.

The nucleic acid doubled in length

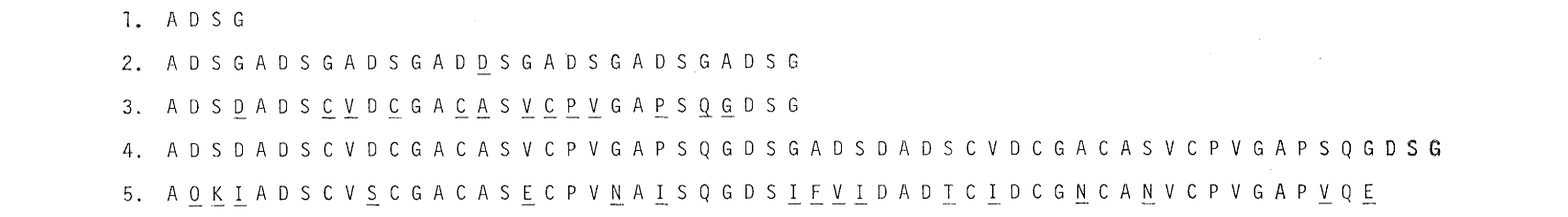

Fig. 3. Proposed origin and evolution of ferredoxin (see text for fuller details). Row 1: Originally, in an extremely primitive organism, a short sequence of four of the simplest amino acids (alanine, aspartic acid, serine, and glycine) could be produced. Row 2: This sequence lengthened by doubling of the genetic material, and one discontinuity occurred (underlined). Row 3: The genetic code becoming more versatile, mutations (underlined) occurred, but only to relatively simple amino acids (the same four, plus cysteine, valine, proline, and glutamine). Iron sulfide was attached to the cysteines, which constituted the "active site" of the respiratory function of this primitive ferredoxin. This configuration still persists. Row 4: By "chromosome" aberration, the whole chain doubled. Row 5: The present more intricate genetic code having evolved, further mutations (underlined) to more complex amino acids occurred. The last three links were deleted. The result was the present sequence of ferredoxin from *C. pasteurianum* (*4*).

by a process that was the prototype of a chromosome aberration, resulting in a protein of 58 units (row 4). In the three-dimensional structure, the effect of this change was to attach the two shorter chains end-to-end. They must already have been in a configuration which was only moderately disturbed by this new constraint. The attachment was an improvement but not a radical change. We predict that when the three-dimensional structure of ferredoxin is worked out, evidence will be found for the previous stage, with its two identical, cooperating, shorter chains. The three end units may have been lost at this time or later, to give the present total length of 55 units.

Many functions of the cell improved in efficiency and complexity, evolving new capabilities. The genetic mechanism also evolved and became capable of incorporating additional amino acids. Mutations occurred and were selected, each of which made a slight improvement in the overall function, until the present sequence was produced (row 5).

By this time there were many lines of descent in the phylogenetic tree, and different species must have produced different mutational variations on the earlier sequences. Comparison of amino acid sequences of many ferredoxins from diverse species should produce clear "living-fossil" evidence of the earliest stages of protein evolution.

At any of the stages, other aberrations may have occurred, resulting in additional duplication and separation of the genetic material, followed by mutations. This would create new genes, producing other proteins, which today may have varying degrees of similarity to ferredoxin.

Genes and their corresponding proteins have not only become more numerous but they have also become longer. Duplication of nucleic acids such as that inferred here in the ferredoxin gene may have been a major means of accomplishing this increase in length. If so, we may expect to see evidences of duplication in other protein sequences, when ways of recognizing distant homologous relationships become more precise than the mere counting of the few identical amino acids remaining. The diheme peptide of *Chromatium* may possibly be such a case (*12*).

Just as the salt composition of our tissue fluids is supposed to represent a stabilized sample of ancient sea water, so the simplest, metabolically most ancient components of cellular metabolism preserve some aspects of their original environment. In modern organisms, primitive reactions, such as those involving glutathione or coenzyme-A, operate under their primordial reducing conditions, isolated from the harsh outer environment by later adaptations. Such ancient systems are extremely conservative, because so many diverse later reactions have become intricately dependent on them that they are no longer "free" to evolve. A mutational change which might be beneficial in one way, in almost every case would be a strong disadvantage in many other ways. When such a mutation occurred, the process of natural selection would therefore reject it. This conservative principle enables us to comprehend why ferredoxin from a living organism could still retain detectable details of its ancient origin.

Thus, in organisms still living there may exist biochemical relics of the era encompassing the origin and evolution of the genetic mechanism. Determination of the sequences of proteins such as ferredoxin and of nucleic acids such as transfer RNA, whose prototypes must have functioned at this early time, should make possible a detailed reconstruction of the biochemical evolutionary events of this era.

RICHARD V. ECK
MARGARET O. DAYHOFF
National Biomedical Research Foundation, 8600 16th Street, Silver Spring, Maryland 20910

References and Notes

1. K. Narita, K. Titani, Y. Yaoi, H. Murakami, *Biochim. Biophys. Acta* **77**, 688 (1963).
2. E. Margoliash, E. Smith, G. Kreil, H. Tuppy, *Nature* **192**, 1121 (1961).
3. H. Matsubara and E. Smith, *J. Biol. Chem.* **237**, 3575 (1962).
4. M. Tanaka, T. Nakashima, A. Benson, H. F. Mower, K. T. Yasunobu, *Biochem. Biophys. Res. Commun.* **16**, 422 (1964).
5. D. I. Arnon, *Science* **149**, 1460 (1965).
6. W. J. Lovenberg, *Biol. Chem.* **238**, 3899 (1963).
7. L. E. Mortenson, *Proc. Nat. Acad. Sci. U.S.* **52**, 272 (1964).
8. All amino acids are highly unstable and decompose into CO_2, H_2O, CH_4, N_2, H_2 (and H_2S), in a reducing atmosphere, given a suitable catalyst. At thermodynamic equilibrium the smaller amino acids have a relatively higher concentration than the more complex ones. For example, in an ideal gas system containing C, H, O, N, and S in the proportions 20 : 50 : 30 : 400 : 1, glycine is present in 10^{-22} mole fraction. Others, in decreasing order of concentration are alanine, cysteine, serine, aspartic acid, and valine. The 14 other coded amino acids have concentrations of less than 10^{-39} mole fraction. These amounts are too low to be significant for the organization of living systems, 10^{-22} representing less than one molecule per droplet. The proportions, however, give some indication of the relative ease with which the amino acids might be made in a very simple, primitive system (M. O. Dayhoff, R. V. Eck, E. R. Lippincott, G. Nagarajan, in preparation).
9. A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; O, tyrosine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan.
10. O. Smithies, G. E. Connell, G. H. Dixon, *Nature* **196**, 232 (1962).
11. M. O. Dayhoff, E. R. Lippincott, R. V. Eck, *Science* **146**, 1461 (1964).
12. K. Dus, R. G. Bartsch, M. D. Kamen, *J. Biol. Chem.* **237**, 3083 (1962); C. J. Epstein, and A. G. Motulsky, *Progr. Med. Genet.* **4**, 85 (1965).
13. This work was supported by NIH grants Nos. GM-08710 and GM-12168, and NASA contract 21-003-002.

21 February 1966