INFERENCES FROM PROTEIN AND NUCLEIC ACID SEQUENCES: EARLY MOLECULAR EVOLUTION, DIVERGENCE OF KINGDOMS AND RATES OF CHANGE

M. O. DAYHOFF, W. C. BARKER, and P. J. McLAUGHLIN
National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, N.W., Washington, D.C. 20007, U.S.A.

Abstract. Presently the sequences of more than 150 different kinds of proteins and nucleic acids are known from the many thousands thought to exist in all living creatures. Some few of these have occupied much the same functional niche within the living cell from near the beginning of life. In three of these latter, sequence evidence pointing to duplications of genetic material in a primitive ancestor is available and in the fourth other evidence suggests it. Such a duplication, shared by the many descendant species, permits us to locate the point of earliest time on an evolutionary tree and to infer the actual order of subsequent evolutionary events. The amounts of change which have occurred in each descendant line can be estimated with good confidence. Some inferences can be made of the structure of the ancestral duplicated sequence, the evolutionary mechanisms which have been operative on it, and the functional capacity of the organism in which it originated. We will describe new, sensitive, objective methods for establishing the probable common ancestry of very distantly related sequences and the quantitative evolutionary change which has taken place. These methods will be applied to the four families, and evolutionary trees will be derived where possible. Of the three families containing duplications of genetic material, two are nucleic acids: transfer RNA and 5S ribosomal RNA. Both of these structures are functional in the synthesis of coded proteins, and prototypes must have been present in the cell at the inception of the fundamental coding process that all living things share. There are many types of tRNA which recognise the various nucleotide triplets and the 20 amino acids. These types are thought to have arisen as a result of many gene duplications. Relationships among these types will be discussed. The 5S ribosomal RNA, presently functional in both eukaryotes and prokaryotes, is very likely descended from an early form incorporating almost a complete duplication of genetic material. The amount of evolution in the various lines can again be compared. The other two families containing duplications are proteins: ferredoxin and cytochrome c. Ferredoxin from photosynthetic and nonphotosynthetic bacteria shows clear evidence of a duplication of genetic material. This duplication is very possibly shared by the ferredoxin from plant plastids and the related adrenodoxin from mammalian mitochondria. If so, a chronology of the details of evolution of these groups can be inferred. From these examples of protein and nucleic acid sequence, we conclude that the amount of change in the bacterial lines is less than that in the eukaryote lines. Even though mutant bacteria are easily produced in the laboratory, though their evolutionary adaptation to new drugs is very rapid, and though new virulent strains often appear spontaneously, nevertheless the sequences of ancient structures in the wild types have changed less than those in the eukaryote lines. Cytochrome c sequences from many eukaryotes and the closely related cytochrome c_2 from *Rhodospirillum rubrum* are known. Other types of cytochrome, such as c_{551} and c_{553}, are probably related to these through gene duplication. Knowledge of enough of these structures to establish an early duplication will provide a time orientation for the cytochrome c evolutionary tree. This quantitative tree now contains sequences from animals, fungi, green plants, protozoa, and bacteria, examples from all five biological kingdoms. (Supported by NIH Grant GM-08710, NASA Contract NASW-2288 and HEW GRS Grant RR-05681.)

At present the sequences of more than 150 different kinds of proteins and nucleic acids are known from the many thousands thought to exist in all living creatures (Dayhoff, 1972, 1973). A few of these have occupied much the same functional niche within the

Origins of Life **5** (1974) 311–330.

living cell almost from the first development of life based on nucleic acid replication. In three of these latter molecules, 5S ribosomal RNA, transfer RNA (tRNA) and ferredoxin, sequence evidence indicating duplications of genetic material in a primitive ancestor is available and in a fourth, the cytochrome c group, additional evidence suggests it. Such duplications, shared by the many descendant species, permit us to locate the point of earliest time on an evolutionary tree derived from sequences of molecules in living organisms and to infer the actual order of the evolutionary divergences. The amounts of change which have occurred in each descendant line can be estimated with good confidence. Some inferences can be made about the structure of the ancestral duplicated sequence, the evolutionary mechanisms which have been operative on it, and the functional capacity of the organism in which it originated.

A useful quantity of sequence data from bacterial groups is just now accumulating. It is already clear that functionally similar sequences in the different orders are so different from each other that sensitive, objective methods are needed to assure ourselves that two sequences show evidence of a common evolutionary origin. Once this relationship is established for a group of sequences, we can refine our understanding of the evolutionary process affecting the early prokaryotic sequences and infer the phylogeny of bacteria and the development of metabolic capacities. In this paper we first describe new objective methods for comparing sequences and then discuss evolutionary inferences from the four groups. Unless otherwise referenced, all protein and nucleic acid sequences mentioned appear in the Data Section of the *Atlas of Protein Sequence and Structure*, Vol. 5 (Dayhoff *et al.*, 1972a) or Supplement I to Vol. 5 (Dayhoff, 1973).

A pair of sequences being compared will fall into one of three general classes: closely related, distantly related or unrelated. In the first group are those sequences which are so similar that sophisticated statistical tools need not be employed to confirm the similarity. It is generally believed that these have been derived from a single ancestral gene either by species divergence or gene duplication and subsequently have accumulated mutations independently. Long regions of such sequences can be aligned without the introduction of any 'gaps' in the sequences. In the third group are sequences whose relationship is either nonexistent or so remote that even the most delicate statistical tools now available cannot distinguish it from chance. If no gaps are permitted in an alignment of typical unrelated protein sequences of equal length, 7% of the residues will match. On the other hand, 23% identities can be obtained by allowing unlimited gaps (Barker and Dayhoff, 1972). For unrelated nucleic acid sequences, 25% match when no gaps are allowed, whereas 48% match if unlimited gaps are permitted. Unrelated sequences may thus appear to be related because most observers focus attention on the number of similarities rather than upon the number of gaps that were required to produce them (see also Cantor, 1968). This illusion has resulted in a number of statistically indefensible reports of relationships in the literature.

The second group of sequences lies between the two extremes. These we call the distantly related sequences. Using statistical methods, the similarity of the sequences can be demonstrated to be greater than that produced by a chance ordering of the residues. These sequences are generally believed to represent genes derived from a

common ancestor which have then undergone divergent evolution. Although similar structures could evolve by chance more than once to perform a particular function, such convergent evolution is not thought to play a major role in sequence similarities. So many sequences could be constructed to perform a particular function that independent evolution of the same sequence based on function would be very improbable.

A number of groups including our own have worked out computer methods for the detection of distant relationships between sequences. Papers describing work by Needleman and Wunsch (1970), Fitch (1966, 1970), McLachlan (1971), Haber and Koshland (1970), Sankoff (1972), Gibbs and McIntyre (1970), and Sackin (1971), as well as our own (Barker and Dayhoff, 1970, 1972, 1973) have appeared recently.

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val	B Asx	Z Glx	X ?
A Ala	1																						
R Arg	-3	9																					
N Asn	1	0	2																				
D Asp	1	-2	1	4																			
C Cys	-1	-6	-2	-3	12																		
Q Gln	0	1	1	1	-3	3																	
E Glu	1	-2	1	3	-3	1	3																
G Gly	1	-4	0	0	-2	0	1	5															
H His	-1	3	1	0	-5	1	-1	-2	8														
I Ile	-1	-3	-1	-2	0	-1	-2	-3	-2	4													
L Leu	-2	-4	-3	-3	-3	-2	-2	-4	-2	3	6												
K Lys	0	3	1	0	-4	0	0	-1	0	-1	-2	4											
M Met	-1	0	-2	-2	1	-1	-2	-3	-2	2	3	-1	4										
F Phe	-4	-4	-5	-6	-6	-5	-6	-7	-2	0	1	-5	0	8									
P Pro	1	-4	0	0	-4	1	1	1	-3	-2	-4	-1	-2	-6	6								
S Ser	1	-2	1	1	0	0	1	1	0	-1	-2	0	-1	-4	0	1							
T Thr	1	-2	1	0	0	0	0	0	-1	0	-1	0	0	-3	0	1	3						
W Trp	-8	-6	-7	-9	-12	-8	-9	-11	2	-4	-5	-8	-4	6	-11	-7	-8	16					
Y Tyr	-7	-5	-7	-9	-10	-7	-8	-10	-1	-2	-3	-5	-2	8	-9	-6	-6	7	11				
V Val	0	-3	-1	-1	0	-1	-1	-1	-2	3	2	-1	2	-2	-1	0	0	-7	-5	3			
B Asx	1	-1	2	2	-2	1	2	0	0	-1	-3	1	-2	-6	0	1	1	-8	-8	-1	2		
Z Glx	0	0	1	2	-3	2	2	0	0	-1	-2	0	-2	-5	1	0	0	-8	-8	-1	1	2	
X ?	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 1. Mutation data matrix. This matrix is derived by taking the logarithm of each element of a relatedness odds matrix for an evolutionary distance of 256 accepted point mutations per 100 amino acid links. Each term of the relatedness odds matrix represents the probability that two amino acids would occur at the same position in two proteins because of relationship, divided by the probability that the combination would occur by chance. The matrices that we are currently using are based on 814 mutations observed in deriving the evolutionary histories of eleven groups of closely related proteins. In addition, the relative mutabilities of each type of amino acid and the relative frequencies of occurrence of each were calculated and used in the derivation of the odds matrix. When a protein is compared to another, position by position, one should multiply the odds for each position to calculate an odds for the whole protein. However, it is more convenient to add the logarithms of the odds terms. Therefore, we use the matrix formed by taking the logarithm of each of the odds terms as our mutation data matrix. The matrix is symmetrical. The value of zero for an unknown amino acid and any known residue is arbitary; such a comparison seldom occurs.

We have found the algorithm of Needleman and Wunsch to be sensitive, rapid, objective, and easy to interpret. The work reported here is based on this algorithm. It determines the highest possible score for any alignment (including gaps) of two sequences. This score is then compared with the highest possible scores obtained by aligning pairs of randomized sequences having the same amino acid or nucleotide composition as the two real sequences. We have made 300 such random comparisons for each pair. The scores form a normal distribution for which the mean and standard deviation are calculated.

The sensitivity of the method depends ultimately upon scores accumulated from comparisons of a residue in one sequence with one in the other sequence. The contribution for each pair is specified in a matrix of pair scores which must be supplied to the program. The simplest such matrix, the 'unitary matrix', counts one for identities and zero for non-identities. We use this simple matrix for nucleic acids because there are not sufficient data for more precise estimates. For proteins, we have derived a matrix from mutation data which has proven to be very sensitive for detecting distant relationships (Barker and Dayhoff, 1970). The mutation data result from the combined effects of several factors, including the nature of the genetic code, the rates of mutation at the nucleotide level and natural selection.

The derivation of this matrix, shown in Figure 1, is described in detail in the *Atlas of Protein Sequence and Structure* (Dayhoff *et al.*, 1969; Barker and Dayhoff, 1972). There are many levels of distinction in this matrix. Amino acids which are often conserved in distantly related proteins, such as cysteine or tryptophan, have high scores on the diagonal and low scores elsewhere. On the other hand, highly mutable amino acids such as serine or alanine are as likely to have changed to any of several other amino acids as to have remained unchanged. The alignment score derived with this matrix is very little affected by exchanges among similar amino acids.

We have used the mutation data scoring matrix for the protein comparisons reported here. For each comparison a statistic is derived whose reliability can be estimated and improved. The difference between the score obtained with the two real sequences, s, and the mean score from the 300 pairs of randomized sequences, m, is divided by the standard deviation, σ, of the random scores, to give a score, A, which we call the alignment score.

$$A = \frac{s - m}{\sigma}.$$

The alignment score is thus expressed in units of standard deviations from the mean of random scores. The probability of obtaining an alignment score more than 2 standard deviation units above that from randomized sequences is less than 3%, and the probability of obtaining a score above 3.1 S.D. units is less than 0.1%.

Implicit in the method is a penalty for increasing the overall length of an alignment by inserting gaps in the sequences. We have set this parameter in such a way that the overall length is increased by only a few percent and the resolution of relationships is

near optimal. This penalty is applied by adding a constant to every term of the scoring matrix. For amino acids we have used +12 and for nucleic acids, +2.

We have been investigating the possibility that this test for the probability of a distant relationship could be made into a quantitative method for ascertaining the evolutionary distance between two sequences. We have constructed a model sequence of 100 links having an average amino acid composition. From the initial sequence a whole family of other sequences with a known number of point mutations was generated by random processes, with the assumptions that each amino acid has a different probability of mutating in a given interval and each has a distinctive probability spectrum for the replacement amino acids. These probabilities were derived from the changes inferred to have taken place in the many evolutionary trees which we presented in the *Atlas of Protein Sequence and Structure 1969,* as described in Chapter 9 (Dayhoff *et al.*, 1969). This model does not fully express the unique mutability pattern of each position in a protein chain and differs from reality in giving fewer parallel mutations in a group of related sequences and in having fewer positions where no mutations are observed at all. In spite of the simplicity of the assumptions, considerable insight can be obtained with the model.

The dependence of alignment scores on total number of mutations in 100 links is shown by the lower curve of Figure 2. Even after 550 changes have occurred, the alignment score is 3. S.D. units above the score for infinite distance; after 1000 changes the score is 1 S.D. unit above and therefore positive scores are obtained 5/6 of the time. The percent difference between aligned sequences shown in the upper curve of Figure 2, although a good measure at short distance, deteriorates rapidly. At 300 changes the score is 3 standard deviations above the score for infinite distance and

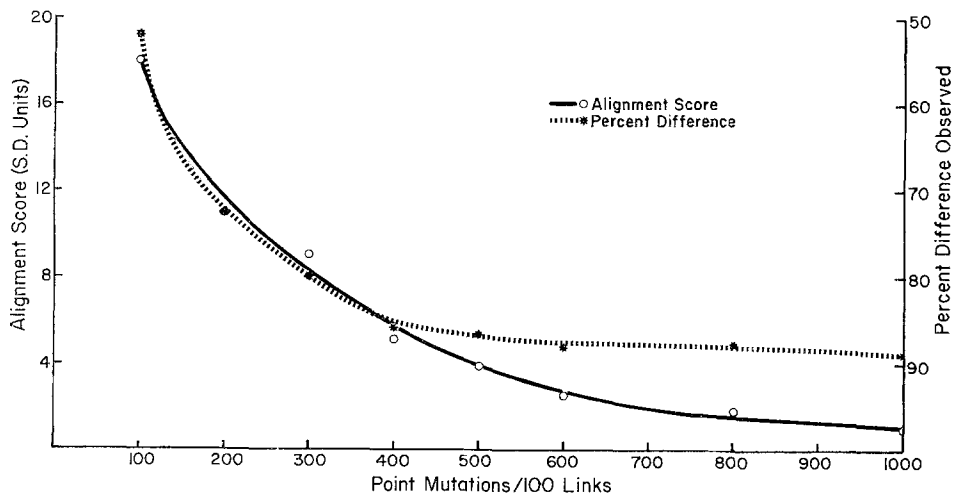

Fig. 2. Dependence of alignment scores on total number of mutations in model protein sequences of 100 links with average amino acid composition. The percent difference between sequences is a good measure of short evolutionary distances; but for sequences that are more than 80% different, the alignment score obtained with our mutation data matrix is a much better measure of relatedness.

after 390 changes it is within 1 standard deviation. For the region with more than 80% difference, the alignment score is a much better measure of relatedness than a count of differences.

It is possible to detect very old relationships using alignment scores. For example, if the related cytochrome sequences behaved like the model sequences, and changed at the same rate that is currently observed in the vertebrate c's, sequences which diverged

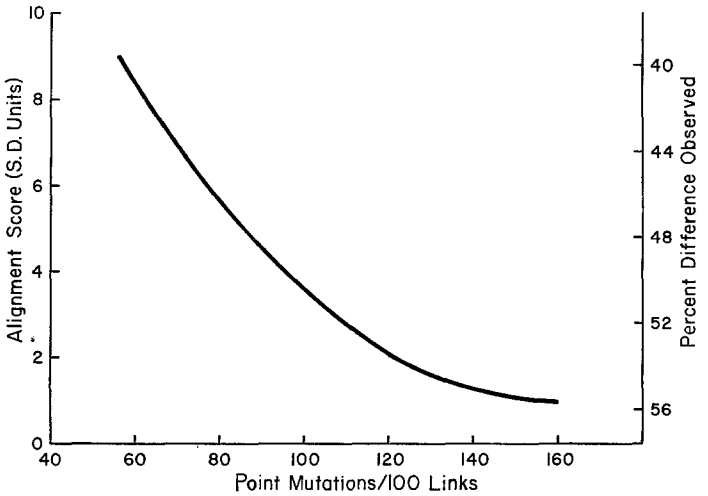

Fig. 3. Dependence of alignment scores and average percent difference on total number of mutations in model nucleic acid sequences of 100 links with equal frequency of occurrence of the four bases. A unitary matrix was used in determining the alignment scores.

3 b.y. ago would give a score of over 12 S.D. units. We do not see such high scores for most of the comparisons. This is in part due to changes in rate and in part because real sequences also suffer changes in length due to the insertion or deletion of genetic material. Such changes degrade the evidence of relationship in the sequence information considerably. In the case of cytochrome c and bacterial c_2 the equivalent number of changes inferred from alignment scores increased by 50% because of the several changes in length.

For nucleic acid sequences, the information degrades much more rapidly than with proteins, as can be seen in Figure 3. In this case we have used a model sequence consisting of 100 residues with equal numbers of each of the four nucleotides and we have assumed an equal probability of change from each nucleotide to each other one. This may not have been the case, but there is as yet insufficient information to estimate the comparison matrix elements more precisely. An alignment score of 3 S.D. units occurs after only 106 changes in a sequence of 100 links and a score of 1 S.D. unit occurs after 160 changes. This poor performance may result from a combination of two factors: that there are only three alternatives for a replacement and that these do not have the chemical distinctiveness of the amino acids.

From Figure 3 it is clear that nucleic acid sequences which are not related at all can still be aligned so that 44% of the nucleotides are identical and there are very few gaps. To protein buffs this may appear to be a good relationship, but in fact it is not meaningful. In spite of the high conservation necessary, we will see that many nucleic acid sequences still preserve recognizable relationships even though they diverged prior to or early in the development of coded proteins.

In the following discussion of the four groups of sequence data, we will use the alignment scores to establish the probability of relationship and the model to calculate approximate amounts of change for the distant sequences.

1. 5S Ribosomal RNA

One of the most interesting molecules, whose structure may antedate the origin of the genetic code, is 5S ribosomal RNA. This molecule, approximately 120 nucleotides in length in the known sequences, has been found in all species which have been examined (Maden, 1971). The genetic code for constructing proteins from nucleic acid templates involves two recognition processes, both involving transfer RNA (tRNA). In the first, each amino acid is recognized by a protein enzyme and bound to an appropriate tRNA. In the second, the anticodon of the charged tRNA binds to the messenger

TABLE I

Alignment scores of 5S RNA sequences (in S.D. units)

	Hu	Ye	*Ps.*	*E.c.*
Human				
Yeast	8.5			
Pseudomonas	2.8	2.2		
E. coli	4.8	4.6	12.2	

RNA. These recognition processes occur either in solution or on the smaller of the two ribosomal subunits. The 5S RNA, although its exact function is not known, is associated with the larger ribosomal unit which is the site of amino acid polymerization. This part of the process of protein synthesis is independent of the kind of amino acid which reacts and could date back to a more ancient synthetic process preceding that using the code.

Five 5S RNA sequences have been completely determined, two from bacteria, one from yeast (Hindley and Page, 1972) and two very similar ones from the vertebrates. For simplicity in the following discussion of distant relationships, we will use only the human sequence from the vertebrates. The four sequences are recognizably related to each other, as shown by the alignment scores of the complete sequences in Table I. All pairs have alignment scores greater than 2 S.D. units and all scores involving the *E. coli* sequence are above 4.6. The corresponding probability that the *E. coli* sequence and any of the others could be this similar by chance is less than 10^{-5}. It is immediate-

ly evident that the two bacterial sequences are most similar to each other and the human and yeast sequences are most similar to each other. An alignment incorporating all of the sequences can be made, requiring only a few gaps which reflect changes in length.

The *E. coli* sequence was the first to be published, by Brownlee *et al.* in 1968. These authors noted that there were some repeating patterns in the sequence. The nine or ten residues at either end were very similar to each other, being identical at 7 positions. The rest of the molecule seemed to show a doubling. They aligned residues 10 to 60 and 61 to 110 and showed regions where 10 and 8 consecutive residues, respectively, were identical. In addition there were some shorter regions of matching. This amount of similarity is very unlikely to have occurred by chance.

We cut the alignment of the four sequences at the places suggested by the *E. coli* doubled regions, discarded the short terminal sequences, and investigated the relationships of these 'half chains' using alignment scores. The resulting scores are shown in Table II. The first halves show a definite relationship from chain to chain and the relationship among the second halves is also clear. In comparing the set of first halves

TABLE II

Alignment scores of 5S RNA half chains
(in S.D. units)

	Hu 1	Ye 1	*Ps.* 1	*E.c.* 1	Hu 2	Ye 2	*Ps.* 2	*E.c.* 2	Anc.
Human 1									
Yeast 1	7.4								
Pseudo. 1	3.2	2.5							
E. coli 1	5.0	3.5	8.3						
Human 2	−0.8	−0.7	0.5	0.0					
Yeast 2	1.8	0.3	−0.3	−0.9	4.1				
Pseudo. 2	0.7	0.4	0.7	1.1	1.2	2.6			
E. coli 2	2.4	2.0	2.5	3.1	1.5	3.2	8.6		
Ancestor	3.3	3.1	4.3	7.4	1.9	3.4	7.0	11.6	

Note: The regions of the sequences used are shown in Figure 4. Where two alternatives appear for the ancestral sequence, we have used the upper character.

with the set of second halves, the *E. coli* halves are definitely related, and many of the other comparisons give positive scores. Some of the scores are low or even negative, indicating that there has been so much change that the common origin is no longer detectable. The degradation of information has involved either point mutations or insertions and deletions of nucleotides. Although the overall length is closely maintained, the alignment of the halves shows evidence of accumulated shifts of portions of the chain.

In order to establish the order of events independently of the somewhat subjective alignment of eight chains, we have used Figure 3 to determine for all pairs the evolutionary distances in terms of point mutations equivalent to the alignment scores. For every pair of organisms, the average evolutionary distance of the corresponding half

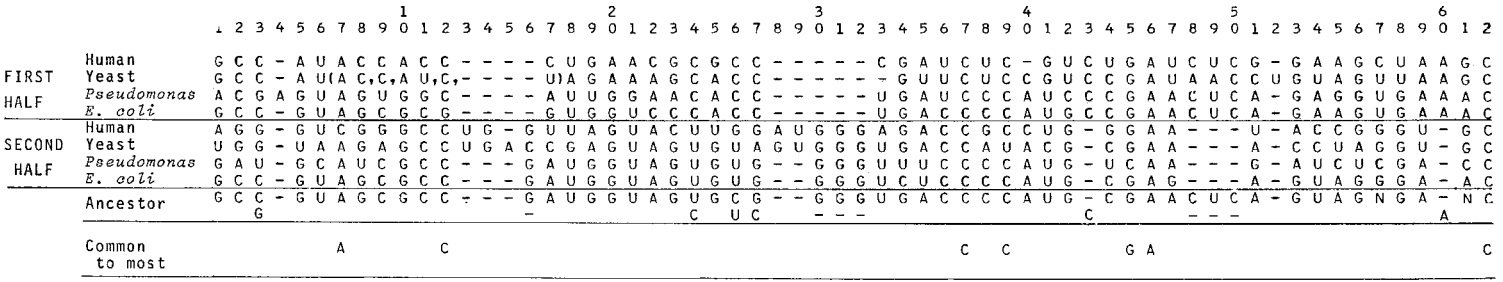

Fig. 4. Alignment of 5S ribosomal RNA half chains. There are two regions in the *E. coli* sequences where remarkable similarity is preserved; the first 10 residues above are identical and residues 36 to 46 are identical except for the insertion of one residue in the first half chains. It is very probable that these similar regions were produced by a duplication of genetic material. Many mutations are evident in other regions of the alignment. The half-chain sequences from the other organisms, which have changed more than those from *E. coli*, have been aligned to conserve the match to the total *E. coli* sequence. The amino-terminal 9 residues of *E. coli* are almost identical to the carboxy-terminal 10 residues and must have been produced by an event separate from those producing the elongation of the main portion. We have omitted these terminal residues and the homologous regions in the other sequences from the alignment.

chains is always less than the average distances between the first halves and the second halves, which indicates early gene elongation preceding the divergence of species. This result is not closely dependent on the model and is also true if alignment scores themselves are used.

An alignment of the half sequences, shown in Figure 4, follows the main features of the original alignment of Brownlee *et al.* derived from the *E. coli* sequence. It is easy to derive an ancestral sequence from this set, as shown below the actual sequences. In each case of a single entry, the base shown occurs at least once in each set of half chains and is the most frequent, usually occurring in more than half of the sequences. In the cases where two alternatives are shown, each is conserved in at least 3 chains of one half, but does not occur in the other half. Where a decision could not be made, we placed an N. In a number of places insertions and deletions must have occurred. We have postulated a minimum number of these.

The alignment scores of the ancestral sequence with the half chains are shown in Table II. It is clear that the scores with all of the four bacterial half chains are higher than those with the four eukaryote half chains. This indicates that there has been more change in the eukaryote lines since the primordial duplication.

Finally, we have calculated the evolutionary tree using the ancestral sequence method (Dayhoff *et al.*, 1972b; McLaughlin *et al.*, 1972). The approximately 100 residues of each of the four sequences shown in the alignment and the corresponding portion of the toad sequence, as well as the duplicated ancestral sequence, were used. The best estimate of the tree is shown in Figure 5. From the relationships among the half chains,

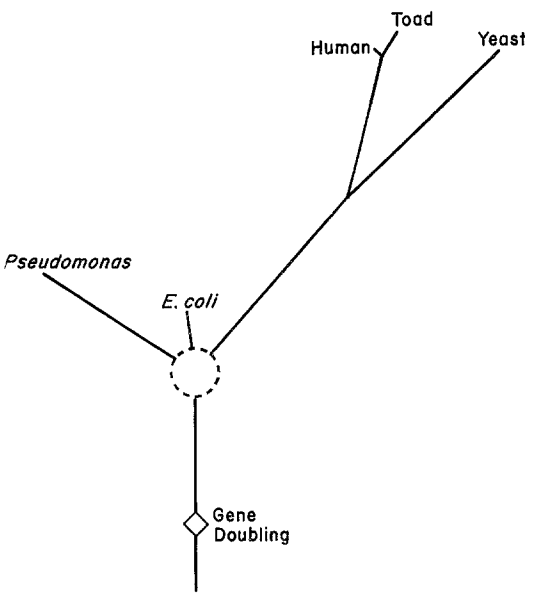

Fig. 5. 5S ribosomal RNA evolutionary tree. The shortened sequences corresponding to the length of the two halves in the alignment of Figure 4, with the addition of a doubled ancestral sequence, were used. The line lengths are suggestive of the number of mutations which are calculated to have occurred.

the time orientation in this tree was established: the chain elongation came first and was succeeded by the divergence of the eukaryote line from the bacterial lines leading to *Pseudomonas* and *E. coli.* Unfortunately, the order of divergence of these three lines cannot be unequivocally placed with this information. Among eukaryotes, first the fungal and vertebrate lines diverged and finally the two vertebrate lines to the mammal and the amphibian separated. The lengths of the lines in the figure are suggestive of the number of point mutations which are calculated to have occurred.

The relative importance of the various mechanisms producing genetic variability may have changed over the time interval represented by this tree. Between the time of the doubling and the divergence of the three lines, the introduction or deletion of genetic material appears to have predominated. During more recent evolution in all the lines, the point mutations far outnumber the changes in amount of genetic material.

This 5S RNA tree shows very different amounts of change in the various lines, the bacteria having changed much less than the eukaryotes. To check possible bias in the method, we also used two alternative lines of reasoning, which omitted all consideration of the doubling, for the upper portions of the tree. A tree made by matrix methods (Margoliash and Fitch, 1967) from the changes equivalent to the alignment scores for complete sequences (Table 1) is in agreement with the corresponding parts of the tree shown. Further, alignments of the five sequences were made under the assumption that all sequences were equally subject to change, and trees were derived from these alignments by the common ancestor method. The *E. coli* arm is still shortest and the *Pseudomonas* arm is shorter than those to the eukaryotes.

TABLE III

Comparisons between tRNA types in eukaryotes and in prokaryotes
(complete sequences without modifications)

tRNA types	Alignment scores		% Identity	
	Yeast–Yeast	*E. coli*–*E. coli*	Yeast–Yeast	*E. coli*–*E. coli*
Asp–Arg	4.2	6.6	50	61
Asp–Gly	3.6	7.1	48	64
Asp–Phe	2.3	5.5	44	53
Asp–Trp	3.1	7.2	47	62
Arg–Gly	2.5	5.3	46	55
Arg–Phe	2.2	4.9	43	55
Arg–Trp	6.6	6.5	61	57
Gly–Phe	4.7	7.7	53	62
Gly–Trp	4.9	8.1	48	64
Phe–Trp	4.1	4.7	53	53
Average	3.82	6.36	49.3	58.6

Note: The sequences which have not been published in the *Atlas of Protein Sequence and Structure* were taken from Harada *et al.*, 1972 (Asp tRNA, *E. coli*), Weissenbach *et al.*, 1972 (Arg tRNA, brewer's yeast) and Yoshida, 1973 (Gly tRNA, yeast).

2. Transfer RNA

The next question which arises is whether great disparity of evolutionary change is usual among bacteria. There is another large body of data for which the time orientation is known – the tRNA sequences. It now seems likely that the divergence of types for the various amino acids occurred very early, long before the divergence of the lines to *E. coli* and the eukaryotes, and that the types have evolved independently since then (Dayhoff and McLaughlin, 1972). Presently 5 types of similar length with homologous codons from both *Saccharomyces* yeast and *E. coli* are known. Alignment scores were derived from all pairs of these within each organism, giving a total of 10 comparisons. We have also derived a measure of percent identity from a single optimized alignment in which the codons and paired regions were matched. Table III presents the results. In both cases, the evidence overwhelmingly supports the view that there is less difference between the pairs of *E. coli* sequences than between the pairs of yeast sequences. Translating the average alignment scores into equivalent mutations, we see that the accumulated change in yeast evolution is 1.4 times larger than that in *E. coli* evolution.

3. The Ferredoxins and Adrenodoxin

The ferredoxins are small iron-containing enzymes which participate in various organisms in such fundamental biochemical processes as photosynthesis, nitrogen fixation, sulfate reduction and other oxidation-reduction reactions. Sequences are known from three main groups: plant plastids, anaerobic photosynthetic bacteria and anaerobic non-photosynthetic bacteria. The sequences of ferredoxins from plant plastids are nearly twice as long as those from non-photosynthetic bacteria, whereas that from *Chromatium*, a photosynthetic bacterium, is intermediate in length. There are also differences between the three types of ferredoxin in content of iron, sulfur, and cysteine and in the functional role of the protein. Nevertheless, portions of the sequences do exhibit similarities when they are aligned.

Adrenodoxin (Tanaka *et al.*, 1973) is an electron transport protein which has been isolated from mammalian adrenal mitochondria, where it acts in the intermediate steps of steroid hydroxylation. It is similar to the plant ferredoxins in length and in

TABLE IV

Alignment scores of mitochondrial adrenodoxin and plastid ferredoxin complete sequences (in S.D. units)

	Bovine adrenodoxin	Algal ferredoxin
Ferredoxin-algal (*Scenedesmus*)	2.8	
Ferredoxin-higher plant (alfalfa)	2.5	19.4

the numbers of cysteines and of bound iron and sulfur atoms which it contains. Alignment scores for adrenodoxin compared to the plant ferredoxins are shown in Table IV. The probabilities of obtaining such scores using randomly scrambled sequences of the same composition are less than 0.01, suggesting a remote relationship. Together with the similarities in overall length, function, and numbers and correspondence of functionally important amino acid residues and prosthetic groups, it seems likely that adrenodoxin shares a distant common ancestor with the plant ferredoxins (Barker *et al.*, 1972).

It has long been recognized that the bacterial ferredoxin sequences display evidence of a gene doubling (Eck and Dayhoff, 1966). Matsubara *et al.* (1967) suggested that the plant and bacterial sequences were recognizably related. Matsubara *et al.* (1968) have advanced a detailed hypothesis regarding the many evolutionary steps linking the two types of sequences. They suggested that the evolution included an early doubling in the ancestral form and the appearance of additional material including smaller duplications and deletions in the plant forms. In what follows we have examined for relationships the portions of the sequences which align with the duplicated bacterial sequences (Dayhoff *et al.*, 1972a, p. D–39); we have omitted the additional terminal material from the plant, *Chromatium*, and adrenodoxin sequences.

Table V gives the alignment scores of the half chains of sequences from *Micrococcus aerogenes* (a non-photosynthetic anaerobic bacterium), *Chromatium* (a photosynthetic anaerobic baterium), plastids of alfalfa (a higher green plant), and from bovine mitochondrial adrenodoxin. The bacterial sequences are definitely related and they share the duplication. Comparison of the sequences from bacteria with those from eukaryote

TABLE V

Alignment scores of ferredoxin half-chains from non-photosynthetic and photosynthetic bacteria, green plant plastids and bovine mitochondria

		First half				Second half			
		Micro-coccus	*Chrom-atium*	Alfalfa	Adreno-doxin	*Micro-coccus*	*Chrom-atium*	Alfalfa	Adreno-doxin
First Half	*Micrococcus aerogenes* (1–26)								
	Chromatium (1–28)	6.4							
	Alfalfa (32–61)	2.3	4.4						
	Adrenodoxin (37–75)	1.5	2.8	3.4					
Second Half	*Micrococcus* (27–54)	5.6	4.5	1.6	1.8				
	Chromatium (29–65)	3.6	1.3	1.0	0.0	3.4			
	Alfalfa (62–88)	0.8	1.0	0.5	1.0	1.5	1.7		
	Adrenodoxin (76–104)	0.3	0.6	1.5	1.9	0.5	1.0	2.0	

organelles indicates that the relatedness of the first halves is very probable, while that of the second halves is quite distant. The cross terms between the two halves for the bacterial sequences are at a greater distance than the corresponding halves, indicating that the duplication preceded the divergence of these species. While the evidence from the plant ferredoxin and adrenodoxin is not inconsistent with a duplication shared by all the types, neither does it show it conclusively.

A number of bacterial ferredoxin sequences are known and from an alignment of the halves an ancestral sequence can be readily constructed (cf. Dayhoff *et al.*, 1972a, p. D-40). We used such a sequence to obtain the alignment scores of Table VI. If the mitochondrial adrenodoxin and the plastid ferredoxin shared the ancient duplication, it is quite clear that the adrenodoxin sequence has changed most, followed by the plastid sequences, then by that of *Chromatium* and finally by those of the other bacteria.

TABLE VI

Alignment scores of the ancestral ferredoxin half chain with other half chains (in S.D. units)

	First half	Second half
Clostridium butyricum	8.48	6.55
Micrococcus aerogenes	6.80	6.80
Chromatium	7.92	3.16
Scenedesmus	3.95	0.80
Alfalfa	4.36	1.43
Bovine adrenodoxin	1.89	0.22

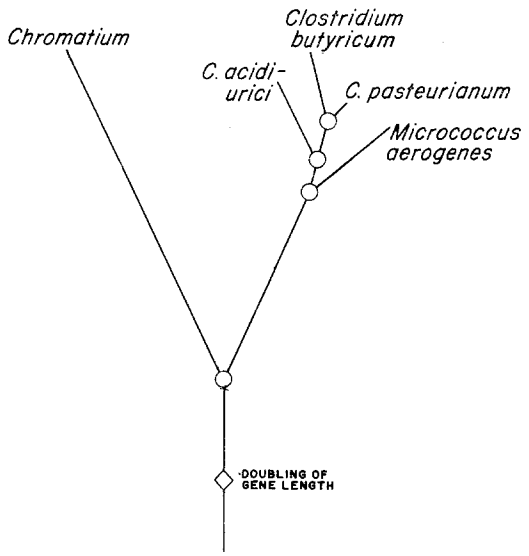

Fig. 6. Evolutionary tree of bacterial ferredoxins. The earliest event was the gene duplication. This was followed by the divergence of the photosynthetic and non-photosynthetic lines. It is impossible to decide from present evidence whether the ancestral organism was photosynthetic or not.

Figure 6 shows the order of events in the bacterial lines. It should be noted that all of the non-photosynthetic forms are on one branch and *Chromatium* is on the other. It is not possible to decide from this whether or not their common ancestor was photosynthetic. The possibility that sequence information could provide an earliest point on the phylogenetic tree of all life and an answer to the question of the primacy of the photoautotroph or the heterotroph is tantalizing.

4. C-Type Cytochromes

Complete sequences of five diverse c-type cytochromes, demonstrably related in sequence to the eukaryote c, are now available. A great deal of information is known about the structure and function of the eukaryote cytochrome c. It is coded in the nucleus and is functional in the respiratory chain of the mitochondrion. Many sequences representing all four eukaryote kingdoms have been educed. Among the bacterial sequences known, by far the most similar to the eukaryote c is cytochrome c_2 from *Rhodospirillum rubrum*. At considerable evolutionary distance from this pair and from each other are three other sequences: c_{551} (Ambler and Wynn, 1973) and c_5 (Ambler and Taylor, 1973) from pseudomonads, functional in respiratory chains; and c_{553} from the plastid of a chrysophycean or golden alga (Laycock and Craigie, 1971; Laycock, 1972), possibly functional in electron transport between photosystems I and II. These cytochromes all have standard reducing potential in the range +0.25 to +0.39 (Mahler and Cordes, 1966; Laycock and Craigie, 1971). Most distant from all of the others is cytochrome c_{553} from *Desulfovibrio vulgaris* (Bruschi and Le Gall, 1972); this form functions in anaerobic metabolism and has a negative standard reducing potential. All six kinds of protein have a heme group covalently attached to cysteine in the amino terminal portion of the chain, and in addition, the c_5 has a second heme similarly attached near the carboxyl end. The alignment scores of these sequences are shown in Table VII. The four remote sequences show >99.9% probability of relationship to either c or c_2.

TABLE VII

Alignment scores of c-type cytochromes

	Horse c	*R.* c_2	Algal plastid c_{553}	*Ps.* c_{551}	*Ps.* c_5	*D.* c_{553}
c Horse						
c_2 *Rhodospirillum*	11.6					
c_{553} Algal Plastid	5.2	3.1				
c_{551} *Pseudomonas*	2.5	3.4	4.3			
c_5 *Pseudomonas*	2.9	4.5	5.4	3.8		
c_{553} *Desulfovibrio*	3.3	2.5	1.9	2.7	2.0	

Note: Because the molecular architecture of the c_5 sequence is somewhat different from the others, with a longer sequence preceding the first cysteines, we have deleted the first 5 residues for these comparisons. The species from which the sequences were derived include *Rhodospirillum rubrum* (c_2), *Monochrysis lutheri* (c_{553}), *Pseudomonas aeruginosa* (c_{551}), *Pseudomonas mendocina* (c_5), *and Desulfovibrio vulgaris* (c_{553}).

The c and c_2 sequences show marked similarity along their entire lengths, with identical residues at 42 of their 100 aligned residues. The pairs with scores higher than 3 S.D. units show unusual similarity at both ends of the chains, whereas those with scores below 3 S.D. units are similar only in the first third of their sequences where the heme-binding sites are located.

Using the relationship derived from model sequences, we have translated the alignment scores into the equivalent number of point mutations per 100 residues. These range from a total distance of 200 for the c and c_2 sequences to an average of 600 for the *Desulfovibrio* comparisons. From these distances we have estimated a tree, as shown in Figure 7. *Desulfovibrio vulgaris* is an obligatory anaerobe which is generally

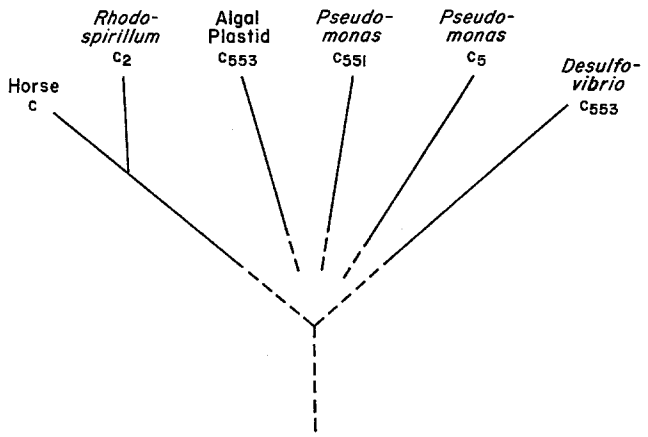

Fig. 7. Estimated evolutionary tree of various distantly related c-type cytochromes. The alignment scores shown in Table VII were translated into the equivalent number of point mutations per 100 residues using the curve in Figure 2. The tree was derived from these estimated distances. This procedure circumvents the necessity of aligning the very distantly related sequences, which is difficult to do. It is evident that c and c_2 are strikingly related whereas all of the other types are very distant from these two and from each other.

conceded to have developed from an early branching during evolution (Margulis, 1970), before the development in other lines of proteins with high positive reducing potentials. We have assumed that an early event on the tree was this species divergence. The great evolutionary distance of the sequences does not conflict with this assumption. The exact order of divergence of the early branches cannot be decided. However, it is clear that c and c_2 are relatively very close together. A great many branches are missing on this tree, so that it is not always possible to distinguish branch points produced by gene duplications within one organism from those produced by divergences of bacterial groups.

Comparison of the evolutionary distances of c and c_2 from each of the lines diverging earlier shows no significant differences in the amount of change in the two lines. We have drawn the tree as though the amounts of change on the two lines were equal.

The complete set of cytochromes which might be recognizably related to this family has not been characterized from any of the groups of organisms mentioned. In eukaryotes there is a c_1 chain which also functions in the mitochondrion, and there is a c_6 from the plastids of higher green plants (Mahler and Cordes, 1966), believed to be an analogue to the algal c_{553} (Laycock, 1972). Ambler has intensively investigated the sequences in pseudomonads. There appears to be at least one additional type, termed c_4 (Ambler and Murray, 1973), for which the amino-terminal sequence was determined and found to be similar to that of c_2 and c.

There are other cytochromes designated broadly as c-type whose sequences do not give significant alignment scores with the c–c_2 group (Dayhoff *et al.*, 1972a; Dayhoff, 1973). These include cytochrome c_3 from *Desulfovibrio*, cytochrome $c_{551.5}$ of undetermined origin and cytochrome cc′ from pseudomonads. Cytochrome b_5 and b_{562} sequences also do not yield significant alignment scores with the c group.

From the c-type cytochromes we can develop an outline of the differentiation of eukaryotes from prokaryotes. This outline is shown in Figure 8, which is an evolutionary tree produced by the ancestral sequence technique. This tree is adapted from a more detailed study of cytochrome c evolution (McLaughlin and Dayhoff, 1973). From this evidence and from that in the previous tree, it is seen that the earliest divergence here is of two lines of prokaryotes. One line leads to *Rhodospirillum* as the present descendant; when other comparable bacterial sequences are known, undoubtedly there will be branches from this line leading to other bacterial groups. The other line evolved into the common ancestor of the eukaryotes. The scheme of evolution

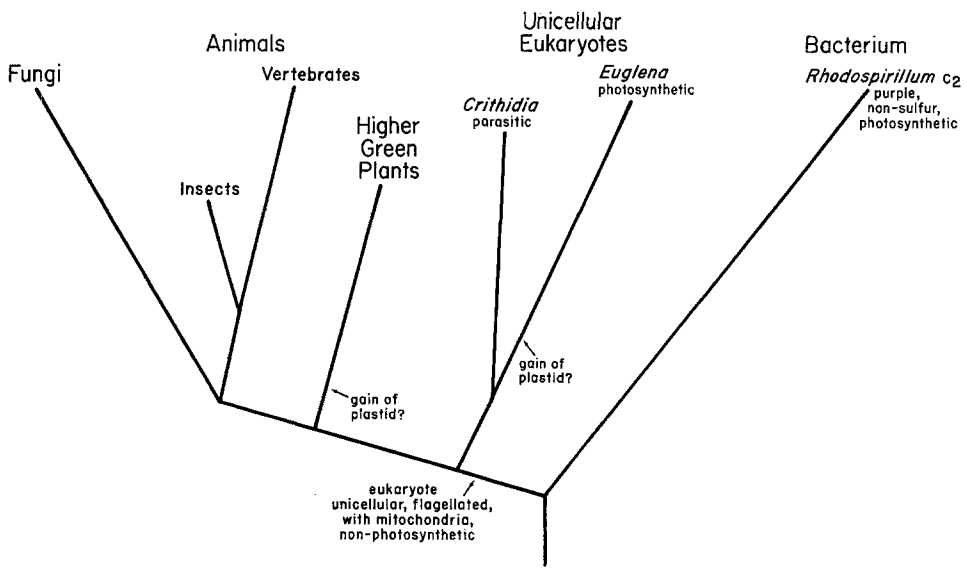

Fig. 8. Outline evolutionary tree of cytochromes c and c_2. The evidence in cytochrome sequences indicates the order of divergence of eukaryote kingdoms shown here. The lengths of the branches are proportional to the number of mutations estimated to have occurred on these branches. The 'stem' of the tree is placed by reference to the preceding tree. The *Euglena* sequence is from Pettigrew, 1973. (Adapted from McLaughlin and Dayhoff, 1973.)

shown by cytochromes slightly favors the symbiotic theory on the origin of plastids in eukaryote cells (Margulis, 1970) and we show the tree for this hypothesis. The common ancestor of the eukaryotes would have been a heterotrophic, flagellated unicellular organism. This ancestor would have had mitochondria which were functioning in electron transport and were already advanced in the process of becoming integrated into the host cell's metabolism. Probably, like *Rhodospirillum* and the eukaryote groups which do not have plastids, this ancestor did not have photosystem II. Separate mechanisms for ingestive nutrition evidently evolved in the diverging protozoa and metazoa. Photosynthetic nutrition would have arisen by the association of blue-green algal symbionts, which became plastids, with heterotrophic hosts. In this tree we can suppose that the acquisition of algal symbionts occurred twice – once on the line leading to the unicellular *Euglena* and another time in an early ancestor of the multicellular plant line.

The mitochondrion, because it is found in all eukaryote species in this tree but not in any prokaryotes, must have appeared in the short evolutionary interval between the divergence of the *Rhodospirillum* line and that of the flagellates. There are two basic theories for this. Either a symbiont was acquired which has since lost or transferred to the nucleus many of its genes and has retained only a few genes and a coding apparatus within itself, or alternatively the eukaryote nucleus and mitochondrion developed from genetic material already in the bacterial ancestor through duplications and reorganization. The correlation of evolutionary trees worked out in the future for various related sequences of prokaryotes and eukaryotes should distinguish between these two theories, or, should some other series of events actually have occurred, provide insight into them.

The answer to the problem of the connection of eukaryote and prokaryote organisms seems imminent. In phylogenetic terms, the question of the symbiotic *vs* nonsymbiotic origin of eukaryote cells becomes a question of whether there are or are not free-living forms on the branches leading to nuclear, mitochondrial and plastid groups of genes. So far, the c-type cytochrome tree has branches leading to a plastid and to the nucleus, with a free-living form on the nuclear branch very close to the beginning of eukaryotic cells. The blue-green algae have long been suggested as free living descendants of the lines which became plastids, but alas, no cytochrome c sequences have been determined from this group.

From these examples, we feel confident that it will be possible to establish the time orientation and the main features of bacterial evolution, including the lines of origin of the eukaryote nucleus, the plastid, and the mitochondrion. We are encouraged in this hope because the main features of eukaryote evolution can, in large measure, be worked out with only one protein, cytochrome c (McLaughlin and Dayhoff, 1973), and because many bacterial sequences show evidence of relationship to each other and the trees derived from them are consistent with other evidence. Further, the bacterial sequences are often more strongly conserved than the eukaryote types, which compensates for the increase in the time span over which evolution has occurred. Thus many 'living fossil sequences' of bacteria await our inquiry.

Acknowledgements

This investigation was supported by Contract NASW-2288 from the National Aeronautics and Space Administration, NIH Grant GM-08710 from the Institute of General Medical Sciences and by NIH Grant RR-05681 from the Division of Research Resources.

We wish to thank Dr Lois T. Hunt for helpful discussions and M. J. Gantt and A. Y. Kulkarni for technical assistance.

References

Ambler, R. P. and Murray, S.: 1973, *Biochem. Soc. Trans.* **1**, 162.
Ambler, R. P. and Taylor, E.: 1973, *Biochem. Soc. Trans.* **1**, 166.
Ambler, R. P. and Wynn, W.: 1973, *Biochem. J.* **131**, 485.
Barker, W. C. and Dayhoff, M. O.: 1970, *Biophys. Soc. Abs.* **10** 152a.
Barker, W. C. and Dayhoff, M. O.: 1972, in M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, Vol. **5**, National Biomedical Research Foundation, Washington, D.C.
Barker, W. C. and Dayhoff, M. O.: 1973, *Biophys. Soc. Abs.* **13**, 205a.
Barker, W. C., McLaughlin, P. J., and Dayhoff, M. O.: 1972, *Fed. Proc.* **31**, 837Abs.
Brownlee, G. G., Sanger, F., and Barrell, B. G.: 1968, *J. Mol. Biol.* **34**, 379.
Bruschi, M. and Le Gall, J.: 1972, *Biochim. Biophys. Acta* **271**, 48.
Cantor, C. R.: 1968, *Biochem. Biophys. Res. Commun.* **31**, 410.
Dayhoff, M. O., (ed.): 1972, *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, D. C.
Dayhoff, M. O. (ed.): 1973, *Atlas of Protein Sequence and Structure*, Vol. 5, *Supplement I*, National Biomedical Research Foundation, Washington, D. C.
Dayhoff, M. O. and McLaughlin, P. J.: 1972, in M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Washington, D.C.
Dayhoff, M. O., Eck, R. V., and Park, C. M.: 1969, in M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, vol. 4, National Biomedical Research Foundation, Silver Spring, Md.
Dayhoff, M. O., Hunt, L. T., McLaughlin, P. J., and Barker, W. C.: 1972a, in M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Washington, D.C.
Dayhoff, M. O., Park, C. M., and McLaughlin, P. J.: 1972b, in M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Washington, D.C.
Eck, R. V. and Dayhoff, M. O.: 1966, *Science* **152**, 363.
Fitch, W. M.: 1966, *J. Mol. Biol.* **16**, 9.
Fitch, W. M.: 1970, *J. Mol. Biol.* **49**, 1.
Gibbs, A. J. and McIntyre, G. A.: 1970, *Eur. J. Biochem.* **16**, 1.
Haber, J. E. and Koshland, D. E. Jr.: 1970, *J. Mol. Biol.* **50**, 617.
Harada, F., Yamaizumi, K., and Nishimura, S.: 1972, *Biochem. Biophys. Res. Commun.* **49**, 1605.
Hindley, J. and Page, S. M.: 1972, *FEBS Letters* **26**, 157.
Laycock, M. V. and Craigie, J. S.: 1971, *Can. J. Biochem.* **49**, 641.
Laycock, M. V.: 1972, *Can. J. Biochem.* **50**, 1311.
Maden, B. E. H.: 1971, *Prog. Biophys. Mol. Biol.* **22**, 127.
Mahler, H. R. and Cordes, E. H.: 1966, *Biological Chemistry*, Harper and Row, New York.
Margoliash, E. and Fitch, W. M.: 1967, *Science* **155**, 279.
Margulis, L.: 1970, *Origin of Eukaryotic Cells*, Yale Univ. Press, New Haven, Conn.
Matsubara, H., Jukes, T. H., and Cantor, C. R.: 1968, *Brookhaven Symp. Biol.* **21**, 201.
Matsubara, H., Sasaki, R. M., and Chain, R. K.: 1967, *Proc. Nat. Acad. Sci. U.S.* **57**, 439.
McLachlan, A. D.: 1971, *J. Mol. Biol.* **61**, 409.
McLaughlin, P. J. and Dayhoff, M. O.: 1973, *J. Mol. Evol.* **2**, 99.
McLaughlin, P. J., Hunt, L. T., and Dayhoff, M. O.: 1972, *J. Human Evol.* **1**, 565.
Needleman, S. B. and Wunsch, C. D.: 1970, *J. Mol. Biol.* **48**, 443.

Pettigrew, G. W.: 1973, *Nature* **241**, 531.
Sackin, M. J.: 1971, *Biochem. Genet.* **5**, 287.
Sankoff, D.: 1972, *Proc. Nat. Acad. Sci. U.S.* **69**, 4.
Tanaka, M., Haniu, M., Yasunobu, K. T., and Kimura, T.: 1973, *J. Biol. Chem.* **248**, 1141.
Weissenbach, J., Martin, R., and Dirheimer, G.: 1972, *FEBS Letters* **28**, 553.
Yoshida, M.: 1973, *Biochem. Biophys. Res. Commun.* **50**, 779.