(1981); A. J. Bard, *Science* **207**, 139 (1980); M. S. Wrighton, *Acc. Chem. Res.* **12**, 303 (1979); A. J. Nozik, *Annu. Rev. Phys. Chem.* **29**, 189 (1978).
3. J. R. Bolton, *Science* **202**, 705 (1978).
4. P. R. Ryason, *Sol. Energy* **19**, 445 (1977).
5. E. Collinson, F. S. Dainton, M. A. Malati, *Trans. Faraday Soc.* **55**, 2096 (1959).
6. L. J. Heidt, M. G. Mullin, W. B. Martin, Jr., A. M. J. Beatty, *J. Phys. Chem.* **66**, 336 (1962).
7. L. J. Heidt and A. F. McMillan, *J. Am. Chem. Soc.* **76**, 2135 (1954).
8. P. K. Eidem, A. W. Maverick, H. B. Gray, *Inorg. Chim. Acta* **50**, 59 (1981).
9. V. Balzani, F. Bolletta, M. T. Gondolfi, M. Maestri, *Top. Curr. Chem.* **75**, 1 (1978).
10. H. D. Gafney and A. W. Adamson, *J. Am. Chem. Soc.* **94**, 8238 (1972).
11. P.-A. Brugger and M. Grätzel, *ibid.* **102**, 2461 (1980).
12. Y. Tsutsui, K. Takuma, T. Nishijima, T. Matsuo, *Chem. Lett.* (1979), p. 617.
13. For a discussion of methods for optimizing the performance of these heterogeneous catalysts, see J.-M. Lehn, J.-P. Sauvage, R. Ziessel, *Nouv. J. Chim.* **5**, 291 (1981); P.-A. Brugger, P. Cuendt, M. Grätzel, *J. Am. Chem. Soc.* **103**, 2923 (1981).
14. K. Kalyanasundaram and M. Grätzel, *Angew. Chem. Int. Ed. Engl.* **18**, 701 (1979); J. Kiwi, E. Borgarello, E. Pelizzetti, M. Visca, M. Grätzel, *ibid.* **19**, 646 (1980).
15. G. M. Brown, B. S. Brunschwig, C. Creutz, J. F. Endicott, N. Sutin, *J. Am. Chem. Soc.* **101**, 1298 (1979).
16. N. Serpone, M. A. Jamieson, M. S. Henry, M. Z. Hoffman, F. Boletta, M. Maestri, *ibid.*, p. 2907.
17. D. Miller and G. M. McLendon, personal communication.
18. W. C. Trogler, G. L. Geoffroy, D. K. Erwin, H. B. Gray, *J. Am. Chem. Soc.* **100**, 1160 (1978).
19. D. K. Erwin, G. L. Geoffroy, H. B. Gray, G. S. Hammond, E. I. Solomon, W. C. Trogler, A. A. Zagars, *ibid.* **99**, 3620 (1977).
20. P. K. Eidem, thesis, California Institute of Technology (1981).
21. D. R. Tyler and H. B. Gray, *J. Am. Chem. Soc.* **103**, 1683 (1981).
22. K. R. Mann, N. S. Lewis, V. M. Miskowski, D. K. Erwin, G. S. Hammond, H. B. Gray, *ibid.* **99**, 5525 (1977).
23. H. B. Gray *et al.*, in *Fundamental Research in Homogeneous Catalysis*, M. Tsutsui, Ed. (Plenum, New York, 1979), vol. 3, p. 819.
24. V. M. Miskowski, I. S. Sigal, K. R. Mann, H. B. Gray, S. J. Milder, G. S. Hammond, P. R. Ryason, *J. Am. Chem. Soc.* **101**, 4384 (1979).
25. I. S. Sigal, K. R. Mann, H. B. Gray, *ibid.* **102**, 7252 (1980).
26. K. R. Mann, M. DiPierro, T. P. Gill, *ibid.*, p. 3965.
27. I. S. Sigal and H. B. Gray, *ibid.* **103**, 3330 (1981).
28. T. P. Smith and H. B. Gray, paper presented at the 180th National Meeting of the American Chemical Society, Las Vegas, Nev., 24 to 29 August 1980.
29. V. M. Miskowski, G. L. Nobinger, D. S. Kliger, G. S. Hammond, N. S. Lewis, K. R. Mann, H. B. Gray, *J. Am. Chem. Soc.* **100**, 485 (1978).
30. S. J. Milder, R. A. Goldbeck, D. S. Kliger, H. B. Gray, *ibid.* **102**, 6761 (1980).
31. S. F. Rice and H. B. Gray, *ibid.* **103**, 1593 (1981).
32. R. F. Dallinger, V. M. Miskowski, H. B. Gray, W. H. Woodruff, *ibid.*, p. 1595.
33. J. S. Najdzionek, unpublished results.
34. A. W. Maverick and H. B. Gray, *J. Am. Chem. Soc.* **103**, 1298 (1981).
35. V. M. Miskowski, R. A. Goldbeck, D. S. Kliger, H. B. Gray, *Inorg. Chem.* **18**, 86 (1979).
36. V. M. Miskowski, A. J. Twarowski, R. H. Fleming, G. S. Hammond, D. S. Kliger, *ibid.* **17**, 1056 (1978).
37. D. G. Nocera and H. B. Gray, *J. Am. Chem. Soc.*, in press.
38. P. B. Fleming and R. E. McCarley, *Inorg. Chem.* **9**, 1347 (1970).
39. J. A. Ferguson and T. J. Meyer, *J. Am. Chem. Soc.* **94**, 3409 (1972).
40. H.-D. Scharf, J. Fleischhauer, H. Leismann, I. Ressler, W. Schleker, R. Weitz, *Angew. Chem. Int. Ed. Engl.* **18**, 652 (1979).
41. For several years our research in inorganic and organometallic photochemistry has been supported by grants from the National Science Foundation (Chemical Dynamics Program). Collaboration with researchers at the Jet Propulsion Laboratory and with Dr. D. S. Kliger and his group at Santa Cruz has been aided by grants from the Continental Group Foundation, the Jet Propulsion Laboratory Director's Discretionary Fund, and the U.S. Department of Energy. Instrumentation was obtained with grants from the National Science Foundation (CHE78-10530) and from the Union Oil Company of California Foundation. Certain rhodium and iridium salts used in our research were lent to us by Johnson Matthey, Inc. A.W.M. acknowledges the National Science Foundation (1977 to 1980) and the Standard Oil Co. (Ohio) (1980 and 1981) for graduate fellowships. This is contribution No. 6415 from the Arthur Amos Noyes Laboratory.

Determination of Nucleotide Sequences in DNA

Frederick Sanger

In spite of the important role played by DNA sequences in living matter, it is only relatively recently that general methods for their determination have been developed. This is mainly because of the very large size of DNA molecules, the smallest being those of the simple bacteriophages such as ϕX174 (which contains about 5000 base pairs). It was therefore difficult to develop methods with such complicated systems. There are, however, some relatively small RNA molecules—notably the transfer RNA's of about 75 nucleotides, and these were used for the early studies on nucleic acid sequences (*1*).

Following my work on amino acid sequences in proteins (*2*) I turned my attention to RNA and, with G. G. Brownlee and B. G. Barrell, developed a relatively rapid small-scale method for the fractionation of ^{32}P-labeled oligonucleotides (*3*). This became the basis for most subsequent studies of RNA sequences. The general approach used in these studies, and in those on proteins, depended on the principle of partial degradation. The large molecules were broken down, usually by suitable enzymes, to give smaller products which were then separated from each other, and their sequence was determined. When sufficient results had been obtained they were fitted together by a process of deduction to give the complete sequence. This approach was necessarily rather slow and tedious, often involving successive digestions and fractionations, and it was not easy to apply it to the larger DNA molecules. When we first studied DNA some significant sequences of about 50 nucleotides in length were obtained with this method (*4, 5*), but it seemed that to be able to sequence genetic material a new approach was desirable and we turned our attention to the use of copying procedures.

Copying Procedures

In the RNA field these procedures had been pioneered by C. Weissmann and his colleagues (*6*) in their studies on the RNA sequence of the bacteriophage Qβ. Phage Qβ contains a replicase that will synthesize a complementary copy of the single-stranded RNA chain, starting from its 3′ end. These workers devised elegant procedures involving pulse-labeling with radioactively labeled nucleotides, from which sequences could be deduced.

For DNA sequences we have used the enzyme DNA polymerase, which copies single-stranded DNA as shown in Fig. 1. The enzyme requires a primer, which is a single-stranded oligonucleotide having a sequence that is complementary to, and therefore able to hybridize with, a region on the DNA being sequenced (the template). Mononucleotide residues are add-

The author is head of the Division of Protein and Nucleic Acid Chemistry at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England. This article is the lecture he delivered in Stockholm, Sweden, on 8 December 1980, when he received the Nobel Prize in Chemistry, a prize he shared with Walter Gilbert and Paul Berg. Minor corrections and additions have been made by the author. The article is published here with the permission of the Nobel Foundation and will also be included in the complete volume of *Les Prix Nobel en 1980* as well as in the series *Nobel Lectures* (in English) published by the Elsevier Publishing Company, Amsterdam and New York. Dr. Berg's lecture appeared in the 17 July issue, page 296. Dr. Gilbert's lecture will be published in a subsequent issue.

0036-8075/81/1211-1205$01.00/0

ed sequentially to the 3′ end of the primer from the corresponding deoxynucleoside triphosphates, making a complementary copy of the template DNA. By using triphosphates containing ^{32}P in the α position, the newly synthesized DNA can be labeled. In the early experiments synthetic oligonucleotides were used as primers, but after the discovery of restriction enzymes it was more convenient to use fragments resulting from their action as they were much more easily obtained.

The copying procedure was used initially to prepare a short, specific region of labeled DNA which could then be subjected to partial digestion procedures. One of the difficulties of sequencing DNA was to find specific methods for breaking it down into small fragments. No suitable enzymes were known that would recognize only one nucleotide. However, Berg, Fancher, and Chamberlin (*7*) had shown earlier that under certain conditions it was possible to incorporate ribonucleotides, in place of the normal deoxyribonucleotides, into DNA chains with DNA polymerase. Thus, for instance, if copying were carried out with ribo CTP (*7a*) and the other three deoxynucleoside triphosphates, a chain could be built up in which the C residues were in the ribo form. Bonds involving ribonucleotides could be broken by alkali under conditions where those involving the deoxynucleotides were not, so that a specific splitting at C residues could be obtained. Using this method we were able to extend our sequencing studies to some extent (*8*). However extensive fractionations and analyses were still required.

The "Plus and Minus" Method

In the course of these experiments we needed to prepare DNA copies of high specific radioactivity, and in order to do this the highly labeled substrates had to be present in low concentrations. Thus if [α-^{32}P]dATP was used for labeling, its concentration was much lower than that of the other three triphosphates and frequently when we analyzed the newly synthesized DNA chains we found that they terminated at a position immediately before that at which an A should have been incorporated. Consequently a mixture of products was produced all having the same 5′ end (the 5′ end of the primer) and terminating at the 3′ end at the position of the A residues. If these products could be fractionated on a system that separated only on the basis of chain length, the pattern of their distribution on fractionation would be proportional to the distribution of the A's along the DNA chain. And this, together with the distribution of the other three mononucleotides, is the information required for sequence determination. Initial experiments carried out with J. E. Donelson suggested that this approach could be the basis for a more rapid method, and it was found that good fractionations according to size could be obtained by ionophoresis on acrylamide gels.

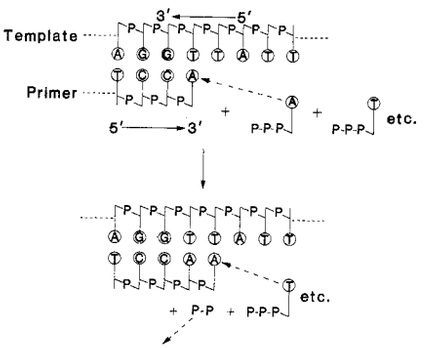

Fig. 1. Specificity requirements for DNA polymerase.

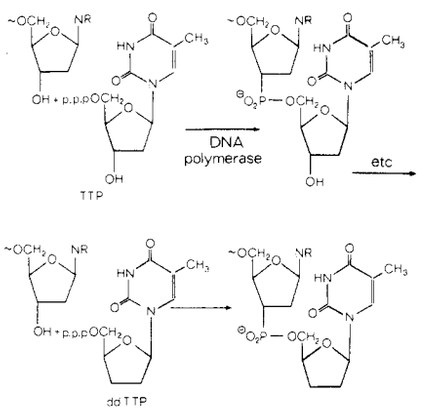

Fig. 2. Diagram showing chain termination with dideoxythymidine triphosphate (ddTTP). The top line shows the DNA polymerase-catalyzed reaction of the normal deoxynucleoside triphosphate (TTP) with the 3′ terminal nucleotide of the primer: the bottom line the corresponding reaction with ddTTP.

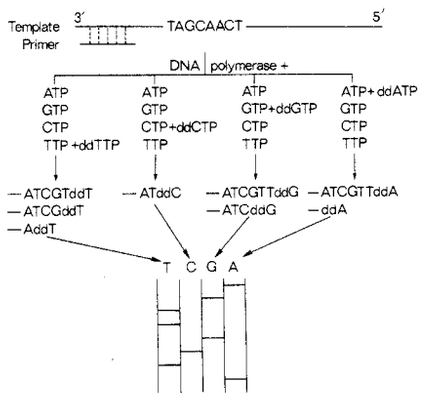

Fig. 3. Principle of the chain-terminating method.

The method described above met with only limited success but we were able to develop two modified techniques that depended on the same general principle, and these provided a simpler and much more rapid method of DNA sequence determination than anything we had used before (*9*). This, which is known as the "plus and minus" technique, was used to determine the almost complete sequence of the DNA of bacteriophage φX174, which contains 5386 nucleotides (*10*).

The "Dideoxy" Method

More recently we have developed another similar method that uses specific chain-terminating analogs of the normal deoxynucleoside triphosphates (*11*). This method is both quicker and more accurate than the plus and minus technique. It was used to complete the sequence of φX174 (*12*), to determine the sequence of a related bacteriophage, G4 (*13*), and has now been applied to mammalian mitochondrial DNA (mtDNA).

The analogs most widely used are the dideoxynucleoside triphosphates (Fig. 2). They are the same as the normal deoxynucleoside triphosphates but lack the 3′ hydroxyl group. They can be incorporated into a growing DNA chain by DNA polymerase but act as terminators because, once they are incorporated, the chain contains no 3′ hydroxyl group and so no other nucleotide can be added.

The principle of the method is summarized in Fig. 3. Primer and template are denatured to separate the two strands of the primer, which is usually a restriction enzyme fragment, and then annealed to form the primer-template complex. The mixture is then divided into four samples. One (the T sample) is incubated with DNA polymerase in the presence of a mixture of ddTTP (dideoxythymidine triphosphate) and a low concentration of TTP, together with the other three deoxynucleoside triphosphates (one of which is labeled with ^{32}P) at normal concentration. As the DNA chains are built up on the 3′ end of the primer, the position of the T's will be filled, in most cases by the normal substrate T, and extended further, but occasionally by ddT and terminated. Thus at the end of incubation there remains a mixture of chains terminating with ddT at their 3′ end but all having the same 5′ end (the 5′ end of the primer). Similar incubations are carried out in the presence of each of the other three dideoxy derivatives, giv-

ing mixtures terminating at the positions of C, A, and G, respectively, and the four mixtures are fractionated in parallel by electrophoresis on acrylamide gel under denaturing conditions. This system separates the chains according to size, the small ones moving quickly and the large ones slowly. As all the chains in the T mixture end at T, the relative position of the T's in the chain will define the relative sizes of the chains, and therefore their relative positions on the gel after fractionation. The actual sequence can then simply be read off from an autoradiograph of the gel (Fig. 4). The method is comparatively rapid and accurate, and sequences of up to about 300 nucleotides from the 3′ end of the primer can usually be determined. Considerably longer sequences have been read off but these are usually less reliable.

One problem with the method is that it requires single-stranded DNA as template. This is the natural form of the DNA in the bacteriophages ϕX174 and G4, but most DNA is double-stranded and it is frequently difficult to separate the two strands. One way of overcoming this was devised by A. J. H. Smith (*14*). If the double-stranded linear DNA is treated with exonuclease III (a double-strand specific 3′ exonuclease) each chain is degraded from its 3′ end, as shown in Fig. 5, giving rise to a structure that is largely single-stranded and can be used as template for DNA polymerase with suitable small primers. This method is particularly suitable for use with fragments cloned in plasmid vectors and has been used extensively in the work on human mtDNA.

Cloning in Single-Stranded Bacteriophage

Another method of preparing suitable template DNA that is being more widely used is to clone fragments in a single-stranded bacteriophage vector (*15–17*). This approach is summarized in Fig. 6. Various vectors have been described. We have used a derivative of bacteriophage M13 developed by Gronenborn and Messing (*16*), which contains an insert of the β-galactosidase gene with an Eco RI restriction enzyme site in it. The presence of β-galactosidase in a plaque can be readily detected by using a suitable color-forming substrate (X-gal). The presence of an insert in the Eco RI site destroys the β-galactosidase gene, giving rise to a colorless plaque.

Besides being a simple and general method of preparing single-stranded DNA, this approach has other advantages. One is that it is possible to use the same primer on all clones. Heidecker *et al.* (*18*) prepared a 96-nucleotide-long restriction fragment derived from a position in the M13 vector flanking the Eco RI site (Fig. 6). This can be used to prime into, and thus determine, a sequence of about 200 nucleotides in the inserted DNA. Smaller synthetic primers have now been prepared (*19, 20*) which allow longer sequences to be determined. The approach that we have used is to prepare clones at random from restriction enzyme digests and determine the sequence with the flanking primer. Computer programs (*21*) are then used to store, overlap, and arrange the data.

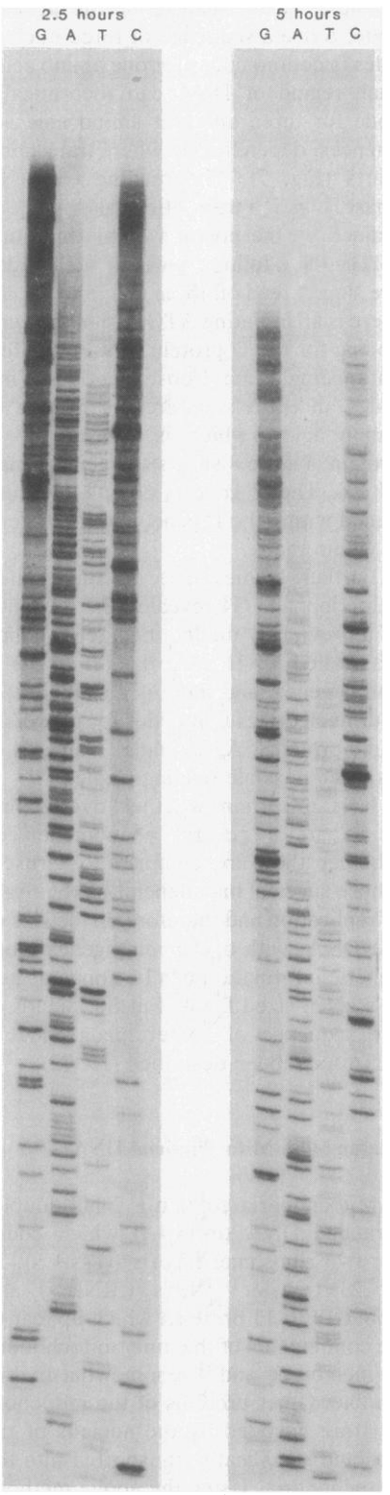

Fig. 4. Autoradiograph of a DNA sequencing gel. The origin is at the top and migration of the DNA chains, according to size, is downward. The gel on the left has been run for 2.5 hours and that on the right for 5 hours with the same polymerization mixtures.

Another important advantage of the cloning technique is that it is a very efficient and rapid method of fractionating fragments of DNA. In all sequencing techniques, both for proteins and nucleic acids, fractionation has been an important step, and major progress has usually been dependent on the development of new fractionation methods. With the new rapid methods for DNA sequencing, fractionation is still important; and, as the sequencing procedure itself is becoming more rapid, more of the work has involved fractionation of the restriction enzyme fragments by electrophoresis on acrylamide. This becomes increasingly difficult as larger DNA molecules are studied and may involve several successive fractionations before pure fragments are obtained. In the new method these fractionations are replaced by a cloning procedure. The mixture is spread on an agar plate and grown. Each clone represents the progeny of a single molecule and is therefore pure, irrespective of how complex the original mixture was. It is particularly suitable for studying large DNA's. In fact, there is no theoretical limit to the size of DNA that could be sequenced by this method.

We have applied the method to fragments from mtDNA (*22, 23*) and to bacteriophage λ DNA. Initially new data can be accumulated very quickly (under ideal conditions at about 500 to 1000 nucleotides a day). However, at later stages much of the data produced will be in regions that have already been sequenced, and progress then appears to be much slower. Nevertheless, we find that most new clones studied give some useful data, either for correcting or confirming old sequences. Thus, in the work with bacteriophage λ DNA, we have about 90 percent of the molecule identified in sequences and most of the new clones we study contribute some new information. In most studies on DNA one is concerned with identifying the reading frames for protein genes, and to do this the sequence must be correct. Errors can readily occur in such extensive sequences and confirmation is always necessary. We usually consider it necessary to determine the sequence of

each region on both strands of the DNA.

Although in theory it would be possible to complete a sequence determination solely by the random approach, it is probably better to use a more specific method to determine the final remaining nucleotides in a sequencing study. Various methods are possible (*22, 24*), but all are slow compared with the random cloning approach.

Bacteriophage ϕX174 DNA

The first DNA to be completely sequenced by the copying procedures was from bacteriophage ϕX174 (*10, 12*)—a single-stranded circular DNA, 5386 nucleotides long, which codes for ten genes. The most unexpected finding from this work was the presence of "overlapping" genes. Previous genetic studies had suggested that genes were arranged in a linear order along the DNA chains, each gene being encoded by a unique region of the DNA. The sequencing studies indicated, however, that there were regions of the ϕX DNA that were coding for two genes. This is made possible by the nature of the genetic code. Since a sequence of three nucleotides (a codon) codes for one amino acid, each region of DNA can theoretically code for three different amino acid sequences, depending on where translation starts (Fig. 7). The reading frame or phase in which translation takes place is defined by the position of the initiating ATG codon, following which nucleotides are simply read off three at a time. In ϕX there is an initiating ATG within the gene coding for the D protein, but in a different reading frame. Consequently this initiates an entirely different sequence of amino acids, which is that of the E protein. Figure 8 shows the genetic map of ϕX. The E gene is completely contained within the D gene, and the B gene is within the A.

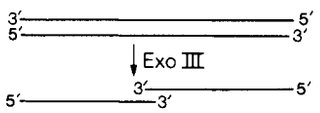

Fig. 5. Degradation of double-stranded DNA with exonuclease III.

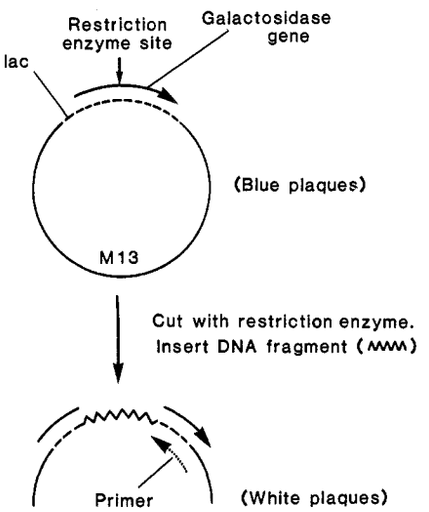

Fig. 6. Diagram illustrating the cloning of DNA fragments in the single-stranded bacteriophage vector M13mp2 (*16*) and sequencing the insert with a flanking primer.

Table 1. Coding changes in mitochondria.

Codon	Amino acid coded		
	Normal	Mitochondria	
		Mammalian	Yeast
TGA	Term*	Trp	Trp
AUA	Ile	Met	Ile
CTN	Leu	Leu	Thr
AGA, AGG	Arg	Term?	Arg
CGN	Arg	Arg	Term?

*Terminating.

Further studies (*25*) on the related bacteriophage G4 revealed the presence of a previously unidentified gene, which was called K. This overlaps both the A and C genes, and there is a sequence of four nucleotides that codes for part of all three proteins, A, C, and K, using all of its three possible reading frames.

It is uncertain whether overlapping genes are a general phenomenon or whether they are confined to viruses, whose survival may depend on their rate of replication and therefore on the size of the DNA: with overlapping genes more genetic information can be concentrated in a given-sized DNA. Further details of the sequence of bacteriophage ϕX174 DNA have been described (*10, 12*).

Mammalian Mitochondrial DNA

Mitochondria contain a small double-stranded DNA (mtDNA) which codes for two ribosomal RNA's (rRNA's), 22 or 23 transfer RNA's (tRNA's) and about 10 to 13 proteins which appear to be components of the inner mitochondrial membrane and are somewhat hydrophobic. Other proteins of the mitochondria are encoded by the nucleus of the cell and specifically transported into the mitochondria. Using the above methods we have determined the nucleotide sequence of human mtDNA (*23*) and almost the complete sequence of bovine mtDNA. The sequence revealed a number of unexpected features which indicated that the transcription and translation machinery of mitochondria is rather different from that of other biological systems.

The genetic code in mitochondria. Hitherto it has been believed that the genetic code was universal. No differences were found in the *Escherichia coli*, yeast, or mammalian systems that had been studied. Our initial sequence studies were on human mtDNA. No amino acid sequences of the proteins that were encoded by human mtDNA were known. However Steffans and Buse (*26*) had determined the sequence of subunit II of cytochrome oxidase (COII) from bovine mitochondria, and Barrell, Bankier, and Drouin (*27*) found that a region of the human mtDNA that they were studying had a sequence that would code for a protein homologous to this amino acid sequence—indicating that it most probably was the gene for the human COII. Surprisingly the DNA sequence con-

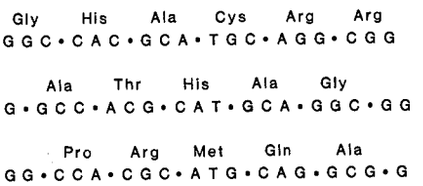

Fig. 7. Diagram illustrating how one DNA sequence can code for three different amino acid sequences. The dots indicate the positions of triplet codons coding for the amino acids.

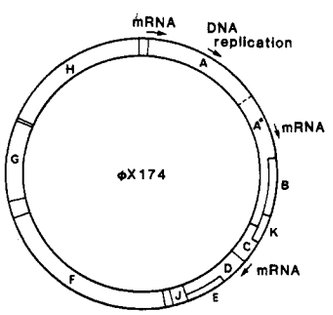

Fig. 8. Gene map of ϕX174 DNA.

Table 2. Normal coding properties of alanine tRNA's. The first position of the codon pairs with the third position of the anticodon and vice versa, for example:
$^{5'}GCU^{3'}$ (codon)
$_{3'}CGG_{5'}$ (anticodon)

Codon	Anticodon	
	Wobble	Mitochondria
GCU	GGC	UGC
GCC		
GCA	UGC	
GCG		

tained TGA triplets in the reading frame of the homologous protein. According to the normal genetic code TGA is a termination codon, and if it occurs in the reading frame of a protein the polypeptide chain is terminated at that position. It was noted that in the positions where TGA occurred in the human mtDNA sequence, tryptophan was found in the bovine protein sequence. The only possible conclusion seemed to be that in mitochondria TGA was not a termination codon but was coding for tryptophan. It was similarly concluded that ATA, which normally codes for isoleucine, was coding for methionine. As these studies were based on a comparison of a human DNA with a bovine protein, the possibility that the differences were due to some species variation, although unlikely, could not be completely excluded. For a conclusive determination of the mitochondrial code it was necessary to compare the DNA sequence of a gene with the amino acid sequence of the protein it was coding for. This was done by Young and Anderson (*28*) who isolated bovine mtDNA, determined the sequence of its COII gene, and confirmed the above differences.

Figure 9 shows the human and bovine mitochondrial genetic code and the frequency of use of the different codons in human mitochondria. All codons are used with the exception of UUA and UAG, which are terminators, and AGA and AGG, which normally code for arginine. This suggests that AGA and AGG are probably also termination codons in mitochondria. Further support for this is that no tRNA recognizing the codons has been found (see below) and that some of the unidentified reading frames found in the DNA sequence appear to end with these codons.

In parallel with our studies on mammalian mtDNA, Tzagoloff and his colleagues (*29, 30*) were studying yeast mtDNA. They also found changes in the genetic code, but surprisingly they are not all the same as those found in mammalian mitochondria (Table 1).

Transfer RNA's. Transfer RNA's have a characteristic base-pairing structure which can be drawn in the form of the "cloverleaf" model. By examining the DNA sequence for cloverleaf structures and using a computer program (*31*), it was possible to identify genes coding for the mitochondrial tRNA's.

Besides the cloverleaf structure, normal cytoplasmic tRNA's have a number of so-called "invariable" features that are believed to be important to their biological function. Most of the mammalian mitochondrial tRNA's are anomalous in that some of these invariable features are missing. The most bizarre is one of the serine tRNA's in which a complete loop of the cloverleaf structure is missing (*32, 33*). Nevertheless, it functions as a tRNA.

In normal cytoplasmic systems at least 32 tRNA's are required to code for all the amino acids. This is related to the "wobble" effect. Codon-anticodon relationships in the first and second positions of the codons are defined by the normal base-pairing rules, but in the third position G can pair with U. The result of this is that one tRNA can recognize two codons. There are many cases in the genetic code where all four codons starting with the same two nucleotides code for the same amino acid. These are known as "family boxes." The situation

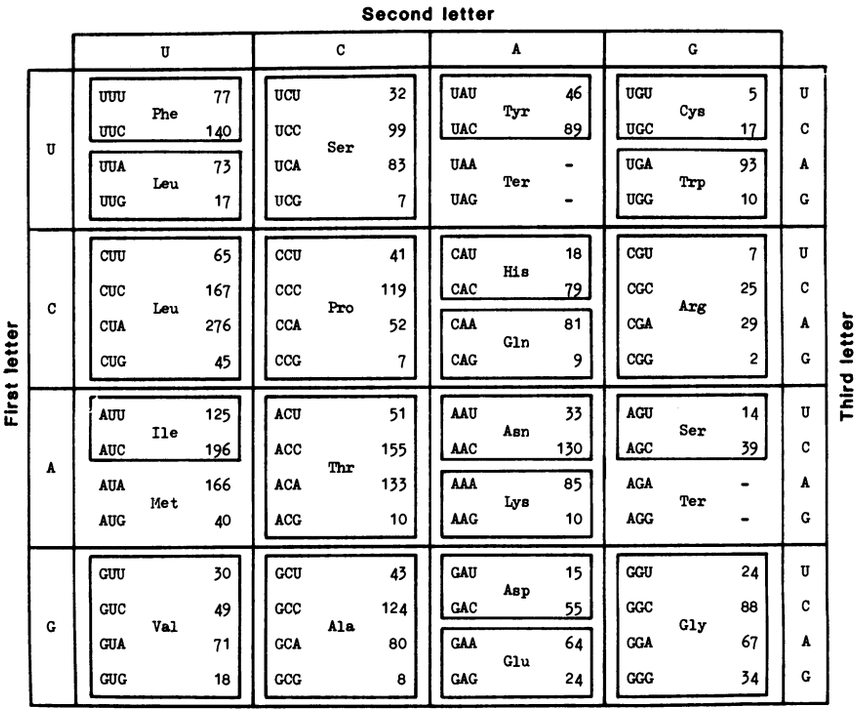

Fig. 9. The human mitochondrial genetic code, showing the coding properties of the tRNA's (boxed codons) and the total number of codons used in the whole genome shown in Fig. 10. (One methionine tRNA has been detected, but since there is some uncertainty about the number present and their coding properties, these codons are not boxed.)

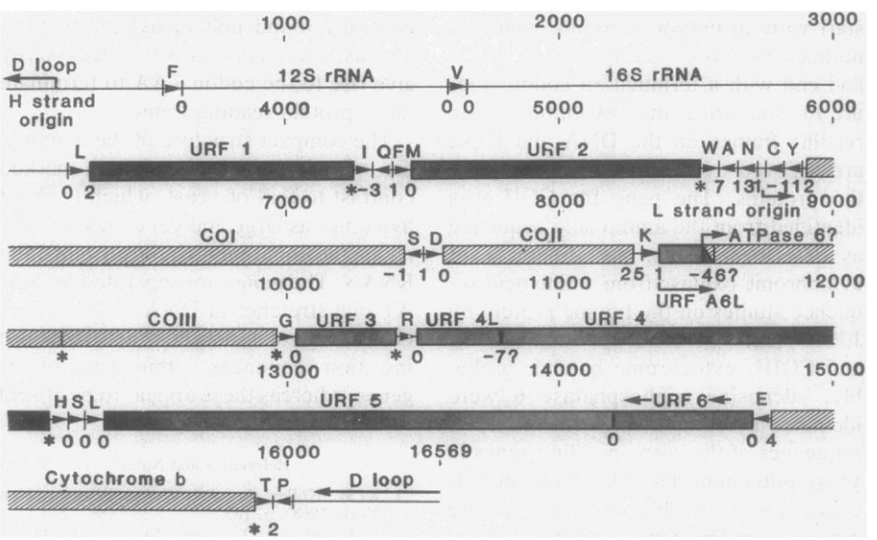

Fig. 10. Gene map of human mtDNA deduced from the DNA sequence. Boxed regions are the predicted reading frames for the proteins. URF, unidentified reading frame. The tRNA genes are denoted by the one-letter amino acid code and are either L strand coded (▶) or H strand coded (◀). Numbers above the genes show the scale in nucleotides and below the predicted number between genes. * Indicates that termination codons are created by polyadenylation of the mRNA.

for the alanine family box is shown in Table 2, indicating that with the normal wobble system two tRNA's are required to code for the four alanine codons.

Only 22 tRNA genes could be found in mammalian mtDNA, and for all the family boxes there was only one, which had a T in the position corresponding to the third position of the codon (*34*). It seems very unlikely that none of the other predicted tRNA's would have been detected, and the most feasible explanation is that in mitochondria one tRNA can recognize all four codons in a family box and that a U in the first position of the anticodon can pair with U, C, A, or G in the third position of the codon. Clearly in boxes in which two of the codons code for one amino acid and two for a different one, there must be two different tRNA's and the wobble effect still applies. Such tRNA's are found, as expected, in the mitochondrial genes. The coding properties of the mitochondrial tRNA's are shown in Fig. 9. Similar conclusions have been reached by Heckman *et al.* (*35*) and by Bonitz *et al.* (*36*), working respectively on neurospora and yeast mitochondria.

Distribution of protein genes. Mitochondrial DNA was known to code for three of the subunits of cytochrome oxidase, probably three subunits of the adenosine triphosphatase complex, cytochrome b, and a number of other unidentified proteins. In order to identify the protein-coding genes, the DNA was searched for reading frames; that is, long stretches of DNA containing no termination codons in one of the phases and thus being capable of coding for long polypeptide chains. Such reading frames should start with an initiation codon, which in normal systems is nearly always ATG, and end with a termination codon. Figure 10 summarizes the distribution of the reading frames on the DNA and these are believed to be the genes coding for the proteins. The gene for COII was identified from the amino acid sequence as described above, for subunit I of cytochrome oxidase from amino acid sequence studies on the bovine protein by J. E. Walker (personal communication), and COIII, cytochrome b, and, probably, adenosine triphosphatase 6 were identified by comparison with the DNA sequences of the corresponding genes in yeast mitochondria. Tzagoloff and his colleagues were able to identify these genes in yeast by genetic methods. It has not yet been possible to assign proteins to the other reading frames.

One unusual feature of the mtDNA is that it has a very compact structure. The reading frames for coding for the proteins and the rRNA genes appear to be flanked by the tRNA genes with no, or very few, intervening nucleotides. This suggests a relatively simple model for transcription, in which a large RNA is copied from the DNA and the tRNA's are cut out by a processing enzyme, and this same processing leads to the production of the rRNA's and the messenger RNA's (mRNA's), most of which will be monocistronic. Strong support for this model comes from the work of Attardi (*37, 38*) who has identified the RNA sequences at the 5′ and 3′ ends of the mRNA's, thus locating them on the DNA sequence. One consequence of this arrangement is that the initiation codon is at, or very near, the 5′ end of the mRNA's. This suggests that there must be a different mechanism of initiation from that found in other systems. In bacteria there is usually a ribosomal binding site before the initiating ATG codon, whereas in eukaryotic systems the "cap" structure on the 5′ end of the mRNA appears to have a similar function, and the first ATG following the cap acts as initiator. It seems that mitochondria may have a more simple, and perhaps more primitive, system with the translation machinery recognizing simply the 5′ end of the mRNA. Another unique feature of mitochondria is that ATA and possibly ATT can act as initiator codons as well as ATG.

On the basis of the above model, some of the mRNA's will not contain termination codons at their 3′ ends after the tRNA's are cut out. However they have T or TA at the 3′ end. The mRNA's are generally found polyadenylated at their 3′ ends, and this process will necessarily give rise to the codon TAA to terminate those protein reading frames.

The compact structure of the mammalian mitochondrial genome is in marked contrast to that of yeast, which is about five times as large and yet codes for only about the same number of proteins and RNA's. The genes are separated by long AT-rich stretches of DNA with no obvious biological function. There are also insertion sequences within some of the genes, whereas these appear to be absent from mammalian mtDNA.

References and Notes

1. R. W. Holley, *Les Prix Nobel* (Elsevier, New York, 1968), p. 183.
2. F. Sanger, *Les Prix Nobel* (Elsevier, New York, 1958), p. 134.
3. _____, G. G. Brownlee, B. G. Barrell, *J. Mol. Biol.* **13**, 373 (1965).
4. H. D. Robertson *et al.*, *Nature (London) New Biol.* **241**, 38 (1973).
5. E. B. Ziff, J. W. Sedat, F. Galibert, *ibid.*, p. 34.
6. M. A. Billeter, J. E. Dahlberg, H. M. Goodman, J. Hindley, C. Weissmann, *Nature (London)* **224**, 1083 (1969).
7. P. Berg, H. Fancher, M. Chamberlin, *Informational Macromolecules* (Academic Press, New York, 1963), pp. 467–483.

7a. The abbreviations C, A, G, U, and T are used to describe either the ribonucleotides or the deoxyribonucleotides, according to context. Other abbreviations used are CTP, cytidine triphosphate; dATP, deoxyadenosine triphosphate; TTP, thymidine triphosphate; N, any nucleotide. Single-letter and three-letter abbreviations of the amino acid residues are as follows: A, Ala, alanine; R, Arg, arginine; N, Asn, asparagine; D, Asp, aspartic acid; C, Cys, cysteine; Q, Gln, glutamine; E, Glu, glutamic acid; G, Gly, glycine; H, His, histidine; I, Iso, isoleucine; L, Leu, leucine; L, Lys, lysine; M, Met, methionine; F, Phe, phenylalanine; P, Pro, proline; S, Ser, serine; T, Thr, threonine; W, Trp, tryptophan; T, Tyr, tyrosine; V, Val, valine.

8. F. Sanger, J. E. Donelson, A. R. Coulson, H. Kössel, D. Fischer, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 1209 (1973).
9. F. Sanger and A. R. Coulson, *J. Mol. Biol.* **94**, 441 (1975).
10. F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, M. Smith, *Nature (London)* **265**, 687 (1977).
11. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
12. F. Sanger, A. R. Coulson, T. Friedmann, G. M. Air, B. G. Barrell, N. L. Brown, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, M. Smith, *J. Mol. Biol.* **125**, 225 (1978).
13. G. N. Godson, B. G. Barrell, R. Staden, J. C. Fiddes, *Nature (London)* **276**, 236 (1978).
14. A. J. H. Smith, *Nucleic Acids Res.* **6**, 831 (1979).
15. W. M. Barnes, *Gene* **5**, 127 (1979).
16. B. Gronenborn and J. Messing, *Nature (London)* **272**, 375 (1978).
17. R. Herrmann, K. Neugebauer, H. Schaller, H. Zentgraf, in *The Single-Stranded DNA Phages*, D. T. Denhardt, D. Dressler, D. S. Ray, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1978), pp. 473–476.
18. G. Heidecker, J. Messing, B. Gronenborn, *Gene* **10**, 69 (1980).
19. S. Anderson, M. J. Gait, L. Mayol, I. G. Young, *Nucleic Acids Res.* **8**, 1731 (1980).
20. M. J. Gait, unpublished results.
21. R. Staden, *Nucleic Acids Res.* **8**, 3673 (1980).
22. F. Sanger, A. R. Coulson, B. G. Barrell, A. J. H. Smith, B. A. Roe, *J. Mol. Biol.* **143**, 161 (1980).
23. B. G. Barrell, S. Anderson, A. T. Bankier, M. H. L. de Bruijn, E. Chen, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, I. G. Young, *Nature (London)* **290**, 457 (1981).
24. G. Winter and S. Fields, *Nucleic Acids Res.* **8**, 1965 (1980).
25. D. C. Shaw, J. E. Walker, F. D. Northrop, B. G. Barrell, G. N. Godson, J. C. Fiddes, *Nature (London)* **272**, 510 (1978).
26. G. J. Steffans and G. Buse, *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 613 (1979).
27. B. G. Barrell, A. T. Bankier, J. Drouin, *Nature (London)* **282**, 189 (1979).
28. I. G. Young and S. Anderson, *Gene*, in press.
29. G. Macino, G. Coruzzi, F. G. Nobrega, M. Li, A. Tzagoloff, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 3784 (1979).
30. G. Macino and A. Tzagoloff, *ibid.*, p. 131.
31. R. Staden, *Nucleic Acids Res.* **8**, 817 (1980).
32. M. H. L. de Bruijn, P. H. Schreier, I. C. Eperon, B. G. Barrell, E. Y. Chen, P. W. Armstrong, J. F. H. Wong, B. A. Roe, *ibid.*, p. 5213.
33. P. Arcari and G. G. Brownlee, *ibid.*, in press.
34. B. G. Barrell, S. Anderson, A. T. Bankier, M. H. L. de Bruijn, E. Chen, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, I. G. Young, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 3164 (1980).
35. J. E. Heckman, J. Sarnoff, B. Alzner-DeWeerd, S. Yin, U. L. RajBhandary, *ibid.*, p. 3159.
36. S. G. Bonitz, R. Berlini, G. Coruzzi, M. Li, G. Macino, F. G. Nobrega, M. P. Nobrega, B. E. Thalenfeld, A. Tzagoloff, *ibid.*, p. 3167.
37. J. Montoya, D. Ojala, G. Attardi, *Nature (London)*, in press.
38. D. Ojala, J. Montoya, G. Attardi, *ibid.*, **287**, 79 (1980).