centrifugation, he dissolved them and tested the solution. They showed intense urease activity. Subsequent work revealed the crystals to be a protein with a molecular weight of 483,000. Not only had Sumner shown that the enzyme was a protein, he had actually crystallized it. The substances that promote the fundamental processes of life were compounds that crystallize in the manner of common chemicals.[1]

from W. Day Genesis On Planet Earth, 2nd ed. Yale, 1984

8.

The Molecular Basis of Life

#12

The astonishing feature of living things is that they can reproduce facsimiles of themselves generation after generation. The conclusion is inescapable. Organisms must possess within their makeup some means of retaining and passing on a store of information that is the inheritance from the preceding generations. This information contains the instructions for synthesizing the organism and all its components.

Since prehistoric times man has realized that heredity was an influence in the physical characteristics of plants and animals, but the mechanism remained obscure and mystifying. Not until the middle of the last century was a systematic study of inheritance carried out. In 1856, an Augustinian monk named Gregor Mendel, growing varieties of the common pea in the cloister gardens in Brünn (Brno), Moravia (now a part of Czechoslovakia), began experiments in cross-breeding them and observing the transmission of various traits in the offspring in the first, second, third, and following generations. As early as 1866, Mendel published statistical rules regarding inheritance in the Proceedings of the Naturforschender Verein in Brünn. The paper received little attention at the time and was forgotten, only to be rediscovered in 1900 when three European botanists, Carl Erich Correns (Berlin), Erich Tschermak von Seysenegg (Vienna), and Hugo De Vries (Leiden), simultaneously and independently reported results similar to Mendel's, only to find that the experimental data and theory had been published 34 years previously.

The great controversies over evolution raged throughout the latter half of the nineteenth century and into the twentieth. In order to explain Darwin's theory, biologists formulated the concept that biological characteristics are

inherited by physical factors that are passed on through successive generations. The English biologist William Bateson gave this branch of biology the name *genetics* (from genesis) in 1906, and in a peculiar retrogressive derivation, the inheritance factors came to be known as genes. But what is the actual chemical nature of the gene?

The genetic substance was isolated from the nuclei of cells nearly 70 years before its true biological significance was realized. In 1868, Friedrich Miescher, a young medically trained Swiss chemist from Basel on the borders of France and Germany, set out for Tübingen. At twenty-four he had just completed his doctoral examination and was going to work for Ernst Hoppe-Seyler, the great German physiological chemist. It was autumn before Miescher began his postdoctoral research, but by February of 1869 he wrote to his former professor in Switzerland of isolating a new substance from cell nuclei.

Little was known of the nucleus of the cell at this time, and the function of cellular material was almost completely obscure. Miescher had initially planned on carrying out his investigation on lymph cells, but their limited availability compelled him to use pus cells that he extracted from surgical bandages. Contaminated with grease and carbolic acid, the cells were first washed with sodium sulfate solution, filtered, and treated with alkali. Miescher then shook the cell fragments vigorously for a long time in mixture of ether and extremely dilute hydrochloric acid. The fats, decomposition products, and detritus either dissolved in the ether or went to the interface of the immiscible liquids: the slightly denser nuclei slowly settled and fell to the bottom of the water layer as a fine, whitish sediment.[1]

Miescher's substance from the nuclei of pus, which he called nuclein, contained a substantial percentage of phosphorus. Until this time, lecithin was the only phosphorus-containing natural product known. The nuclein was a complex of protein and nucleic acid; but subsequent purification procedures led to the nucleic acid being separated as a mass of long, fibrous, threadlike material that could be collected from the precipitated matter by entwining it on the end of a glass rod. Apparently Miescher, without knowing the structural nature of his nuclein, realized it had some connection with the genetic function. It was not until 1944, however, that an experiment by O. T. Avery[2] verified that the young Swiss had isolated the substance that is the chemical basis for the hereditary features of all living things—the nucleic acid.

Miescher's preparation of nuclein came ten years after Darwin published *On the Origin of Species*. By 1880 the mitotic process was established, and biologists, working with the light microscope, discovered that all cells contain nuclear material in a definite number of rodlike units called chromosomes. Chromosomes are the carriers of specific hereditary factors (genes), and elegant microscopic studies revealed their mode of replication in cellular reproduction. When isolated and analyzed, chromosomes were found to contain protein and nucleic acid in nearly equal proportions.

Nucleic acid is the phosphorus-containing polymeric substance found in the nuclein that Miescher had isolated from pus cells in 1868. The German chemist Albrecht Kossel, working at Heidelberg with nucleic acid from the thymus glands of calves, discovered that it contained nitrogenous bases called purines and pyrimidines, and he isolated and identified two different derivatives of each base. Being a polymer with only four kinds of subunits, nucleic acid, like starch, seemed to be a long, monotonous chain molecule.

For the first few decades of this century, while biologists expanded genetics by clarifying the mathematical relation of biological inheritance, biochemists achieved considerable success in showing the role of enzymes in controlling the life processes. Neither had any clear concept of the chemical nature of a gene or how enzymes were made. Then in the 1930s, the American geneticists George Beadle and Edward Tatum,[3] working with *Neurospora crassa*, the common red bread mold, linked Darwinism to chemistry by showing that enzymes control structure and genes control enzymes. They recognized that genes are somehow coded for enzymes, and they postulated that for each enzyme there is a specific gene.

It was an impressive feat for cells to reproduce proteins containing hundreds of amino acids repeatedly in the exact sequence. For a protein to be synthesized time and again in identical composition, it was reasoned, the amino acids must be polymerized in their precise order on some type of template. Such templates would contain in their composition the information of heredity.

Because of the immense amount of information an organism would need in its heredity material, biologists generally accepted that for any substance to be the genetic factor it would have to have a large number of subunits to act as letters in the informational code. Since proteins contain over twenty kinds of amino acids, these macromolecules appeared ideally suitable for the role; and throughout the 1930s, most biologists and biochemists believed the protein found in chromosomes was the genetic material and regarded the nucleic acid to be of little significance.

During that same decade, in 1935, Wendell Stanley[4] at the Rockefeller Institute in New York did an astonishing thing. He crystallized a virus. Viruses are biological entities that occupy a zone between living cells and inanimate chemicals. They consist of protein and nucleic acid without the cellular machinery for reproduction and metabolism. Lacking the means of regeneration, viruses are perpetuated by inserting their genetic factor into cells, and the machinery of the infected cell is taken over to produce copies of the virus. For a virus to be crystallized like so much salt impressed upon scientists that genes might eventually be isolated and studied as chemical compounds. The basis of life moved one step closer to being purely a matter of chemistry.

Then in 1944, Erwin Schrödinger,[5] the renowned Austrian-born physicist living in Dublin as an émigré from Nazi Germany, published a small book, *What Is Life?* In it he stressed upon his fellow physicists that biology was on the threshold of the crucial question of the basis of life, and they must not be discouraged by

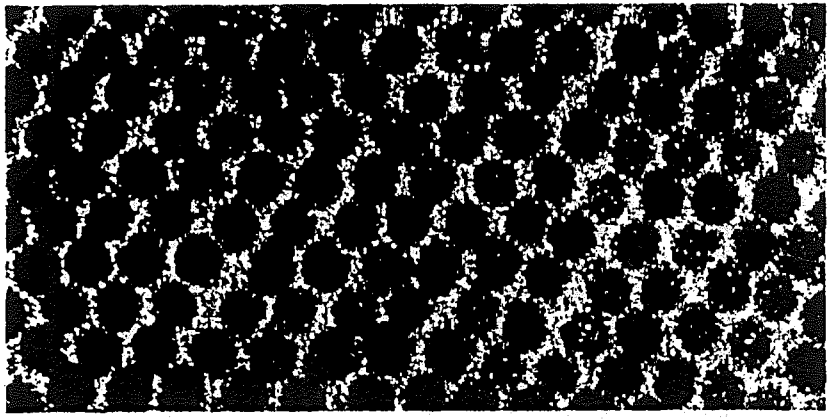


Figure 8.1. Electron photomicrographs of human wart virus. The virus is an icosahedral shell containing DNA. This particular virus has no envelope. Magnification is 120,000 X.

the difficulty in interpreting life by the ordinary laws of physics. He emphasized that they should consider finding how biology can be explained on the molecular level.

The key to the puzzle lay in viruses that infect bacteria and are called bacteriophages. In 1952, Alfred Hershey and Martha Chase, using radioactive phosphorus and sulfur as labels to follow the respective biochemicals, demonstrated that the DNA of a bacteriophage entered the bacterium and it alone was responsible for the reproduction of new viruses. This compelling evidence that the chemical form of genes was the nucleic acid astounded biologists, who regarded protein to be the material of inheritance. The discovery marked the beginning of molecular biology.

When Kossell analyzed nucleic acid in the last century, he found two purines he called adenine and guanine, and two pyrimidines, cytosine and thymine. Later research with plant nucleic acid led to the discovery of a third type of pyrimidine called uracil (fig. 8.2).

In 1910, Levine at the Rockefeller Institute found that nucleic acids also contained a five-carbon sugar. The nucleic acid from plants had the sugar ribose, whereas animal nucleic acid had the same sugar minus one oxygen, and hence was known as deoxyribose (fig. 8.3).

The structural units of nucleic acid consist of the purine and pyrimidine base bonded to the terminal carbon atom (no. 1) of the sugar, and the sugar portion has a phosphate group attached. These three constituents—base, sugar, phosphate—together form a nucleotide. When nucleotides join through a phosphate diester by a 3′, 5′-linkage of their sugars, they create the long chains known as nucleic acids (fig. 8.4).

It soon became apparent that there were two kinds of nucleic acids: one contained adenine, guanine, cytosine, and thymine with deoxyribose and was

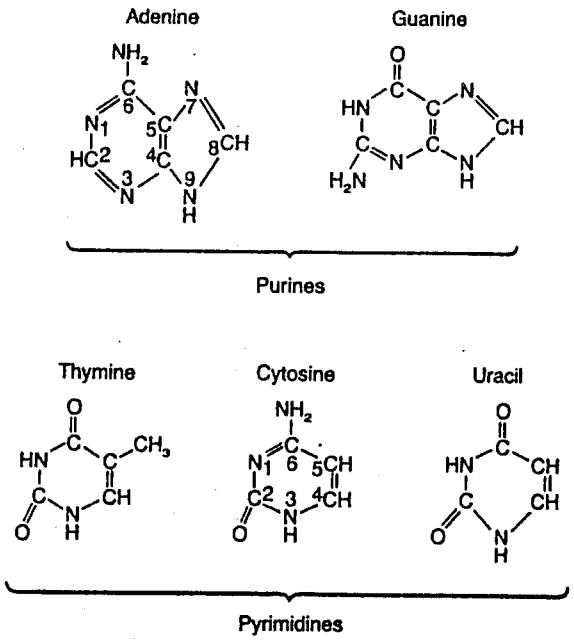


Figure 8.2. Structures of adenine, guanine, thymine, cytosine, and uracil.

called deoxyribonucleic acid (DNA) (fig. 8.5); the other nucleic acid also consisted of four bases but contained uracil in place of thymine, and, since the sugar of uracil was ribose, it was called ribonucleic acid (RNA).

Eventually, biologists realized that the two kinds of nucleic acid did not distinguish plants and animals, but that all living things contained both types, DNA and RNA. Beneath life's immense diversity there was an astonishing unity. The principles of genetics were found to be coextensive to all forms of life. The

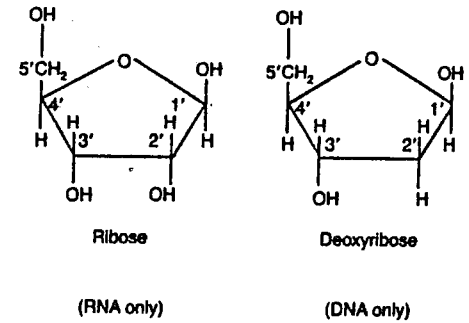


Figure 8.3. Structures of ribose and deoxyribose.

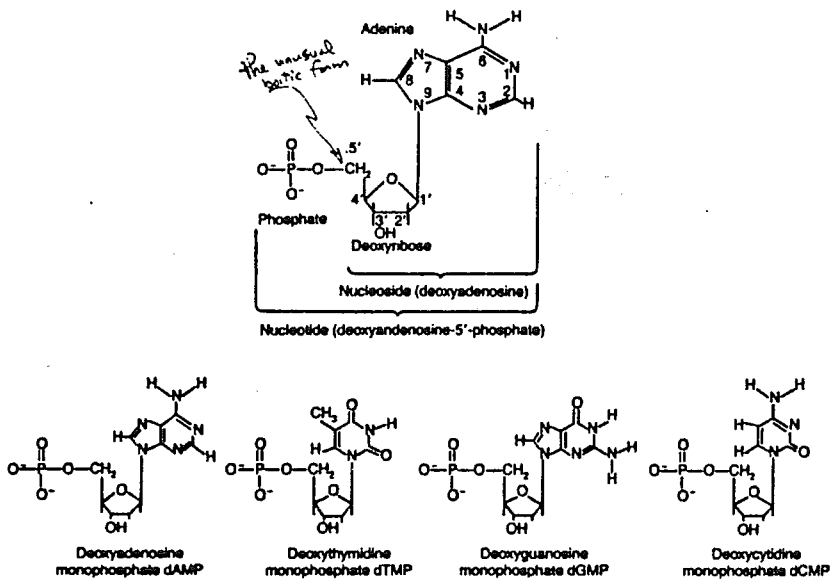


Figure 8.4. Structures of deoxynucleotides.

continuity of living substances through reproduction is based on the multiplicity of genes which are copied and passed on from generation to generation—the material of inheritance being the same for viruses and man alike. That substance is DNA.

Only four basic chemical units are used to create the blueprint in a coded sequence of units in an informational molecule where the molecular weight can run into billions even for bacteria. A section of the DNA constitutes a gene that carries the information for the amino acid sequence for a particular protein. In bacteria the number of genes can be in the thousands, but in mammals it runs as high as 100,000.

Cellular reproduction is ultimately molecular reproduction, and the unique nature of the chemical structure of nucleic acids allows these biopolymers to be copied faithfully. In procaryotes, with their circular DNA, cellular reproduction is replication of the nucleic acid, followed by binary fission into daughter cells with each carrying a full complement of cellular material. The chemistry is simple and direct and must represent the mode of replication assumed soon after the origin of the first living systems. Eucaryotic cells, on the other hand, have reproduction procedures so elaborate that they must have taken an extremely long time to evolve.

For DNA to be the informational reservoir of the cell, there had to be a chemical procedure for making duplicates of the molecule. Before understanding

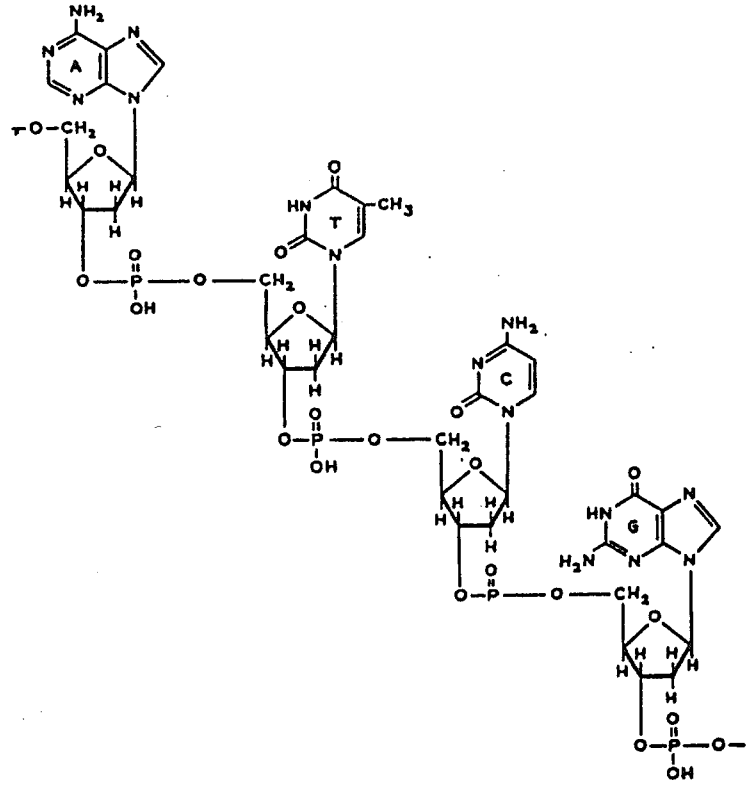


Figure 8.5. A tetranucleotide portion of one strand of DNA composed of adenine (A), thymine (T), cytosine (C), and guanine deoxynucleotides.

how this was accomplished, it was necessary to determine the complete structural arrangement. Nucleic acids are polynucleotides that exist as long, unbranched chains, but only the precise three-dimensional configuration would reveal the mode of action of their biological function. And thus was launched the search for the secret to DNA's architecture.

In 1950, Erwin Chargaff and his students[6] at Columbia University made an odd discovery. After carefully analyzing the purine and pyrimidine composition of various DNAs, they noticed that the number of adenine bases nearly always equaled the number of thymines, and the number of guanine and cytosine bases were also nearly equal. To put it more conveniently, they found A = T and G = C. This was despite large variations in the amounts of A = T to G = C in different DNAs.

While Chargaff's laboratory was analyzing the base composition of DNA, Rosalind Franklin and Maurice Wilkins at King's College, London, were using X-ray crystallography to obtain some precise measurements of DNA. And in

California, Linus Pauling and his associates[7] at the California Institute of Technology attacked the problem of DNA's structure by using the bond lengths and angles of the quantum theorists as guides to construct atomic models. Both the London group and the Caltech chemists were inclined to believe that there were three polynucleotide chains in the DNA molecule.

The principal force of attraction that would hold the polynucleotide strands together was hydrogen bonding. Hydrogen bonds are weak compared to covalent bonds. To break a covalent bond requires between 12 and 24 KJ/mole,* whereas the energy to break hydrogen bonds ranges from 1 to 3 KJ/mole. Nonetheless, hydrogen bonds are of enormous importance in biology by being primarily responsible for the specificity of interactions between macromolecules.

A critical question concerning the structure of DNA was whether the bases pointed toward the outside or toward the center of the molecule. Pauling suggested that the bases were on the outside, but Franklin felt that she had evidence that the phosphates were on the outside and the bases were toward the center.

This was the situation in 1951 when James Watson, a twenty-two-year-old American postdoctoral fellow, arrived in Cambridge and met Francis Crick, a physicist working on his Ph.D. in biophysics. Although ostensibly Crick was doing research on protein and Watson was interested in the structure of the Tobacco Mosaic Virus, they both had an abiding interest in DNA and decided to collaborate on solving the riddle of its structure.

Franklin's X-ray diffraction patterns indicated a regular and compact configuration. The helical structure had been shown by Pauling to be a favored configuration of macromolecules, and Crick became enamored with helices. The question of prime importance was the chemical basis of the procedure organisms used to copy their DNA. Presumably the purines and pyrimidines were important. Reasoning that nature has a penchant for doing things in pairs, Watson discarded the three-strand concept and felt it more plausible to build a model with two strands twisting around each other with the bases directed toward the center. Using cardboard cutouts of the four bases, he attempted to build a model on hydrogen bonding with each base facing a like base as it could be imagined duplication of a strand to occur. But the model was not compact as shown by the X-ray studies. And as Jerry Donahue, an American crystallographer working in the same laboratory, pointed out, Watson, like everyone else, was using the wrong tautomeric form of the bases. Donahue felt that the bases existed in the keto rather than the enol form.

With new cutouts, Watson again attempted to combine base pairs for his model. At this point he discovered that the cutouts of adenine with thymine were the same size and shape as guanine with cytosine. The significance of Chargaff's A = T and G = C ratios suddenly became obvious.

* Kilojoules. One kilojoule = 0.239 kilocalorie.

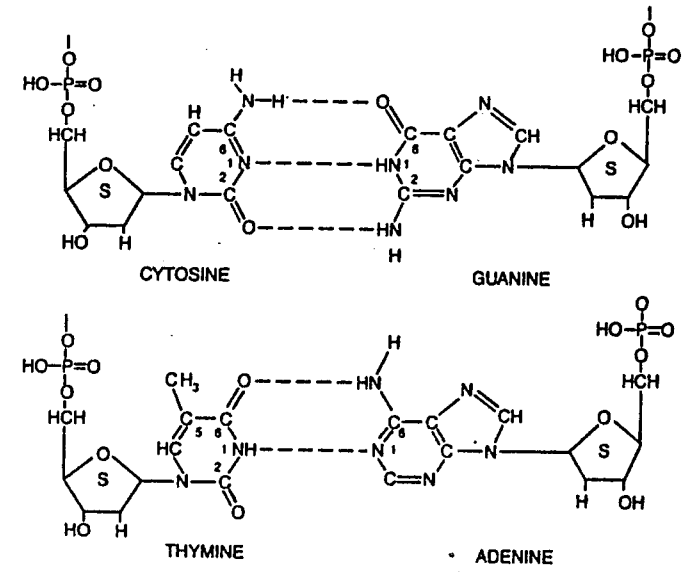


Figure 8.6. The arrangement of nucleotide pairing in DNA.

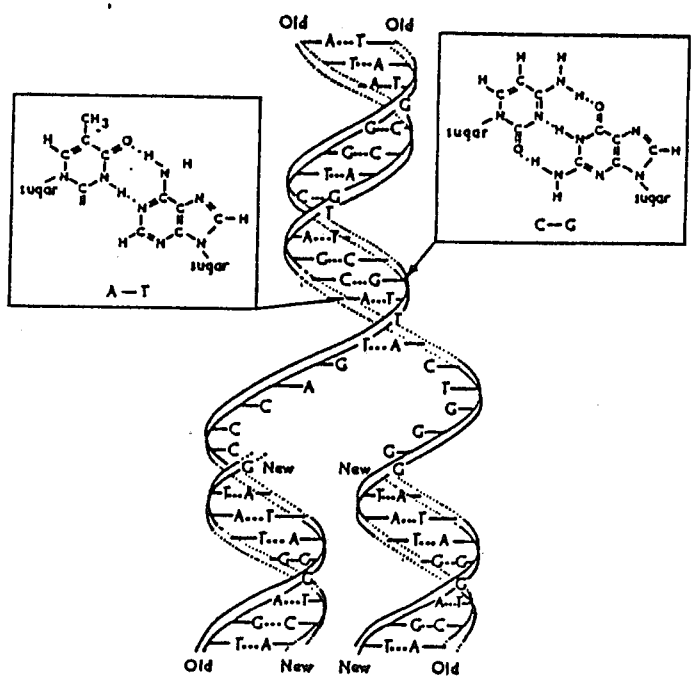


Figure 8.7. Illustration of unwinding and replication of the DNA molecule.

9
From Blueprint to Organism

Miller's experimental simulation of atmospheric conditions on primordial earth opened the research to how life began by showing the manner in which the building blocks were formed before the first cells. Among the building blocks are the nucleotides that join to form nucleic acids, and it is the nucleic acid DNA that is the molecular basis of life's reproductive capacity. But DNA is only an informational molecule—like a computer tape. The proteins, in their immense variety and roles, are the chemicals most directly responsible for the shape, the composition, and the functionality of an organism. Only after the information coded in the chemical structure of DNA is retrieved and translated to protein structure does it complete its biological purpose. How then are proteins created from the DNA structure?

There are two kinds of nucleic acids: DNA and RNA. DNA is the molecule that is the store of hereditary information and is found in the nucleus of the cell. RNA, on the other hand, is found both in the nucleus and in the cytoplasm. Protein synthesis takes place in the cytoplasm. Even in the 1940s, before DNA was known to be the material of genes, research by Torbjörn Caspersson[1] in Stockholm and Jean Bracht[2] in Brussels indicated that RNA somehow was involved in protein synthesis.

In 1950, Henry Borzook and his colleagues[3] at the California Institute of Technology and Tore Hultin of the Wenner-Gren Institute in Stockholm independently identified the microsomes, later known as ribosomes, in the cytoplasm as the site of protein synthesis. The DNA does not act directly as the template for the synthesis of protein, but instead, the DNA sequence is transcribed to an RNA molecule which is copied from the DNA in the same manner that

DNA is replicated. The RNA copy of DNA is a messenger RNA (mRNA) that carries the hereditary information in its sequence to the ribosome to be translated into the amino acid sequence of a polypeptide.

$$\text{DNA} \xrightarrow{\text{(transcription)}} \text{mRNA} \xrightarrow[\text{ribosome}]{\text{(translation)}} \text{protein}$$

In the mid-1950s biologists thought ribosomal RNA was literally a template for protein synthesis which contained structural cavities that were complementary to the shape of amino acids from which the protein was constructed. But no one could construct a model of RNA with specified holes that could conceivably act as a template. It was Francis Crick who recognized that amino acids do not fit directly in the template molecule but require an adapter molecule that is specific for the amino acid that recognizes the particular site on the template.

Mahlon Hoagland of Harvard University had discovered a type of RNA that could not be settled by centrifugation but remained in solution. Hence he called them soluble RNAs. It was not long before molecular biologists realized that Hoagland's soluble RNAs were Crick's adapter molecules. Within a short time a specific soluble RNA was found for each of the twenty amino acids, and more in keeping with their role, they became known as transfer RNAs (tRNAs).

The sequencing of tRNA molecules showed that all of them had CCA at the end of the polynucleotide that has the sugar with a free dydroxyl group on the 3′–carbon—the end to which the amino acid is attached. Also, since the A/U and G/C ratio approached equality, there was the possibility of considerable base pairing within the molecule. Robert Holley and his coworkers[4] at Cornell University observed that tRNAs contain a number of unusual bases that differ from adenine, guanine, cytosine, and uracil by having one or more methyl groups substituted at various positions in their structures.

A complete analysis of a tRNA for alanine was accomplished by Robert Holley and others in 1945, and the nucleotide sequence of a yeast tRNA for tyrosine was reported by Madison and coworkers[5] the following year. The structures proposed for these two tRNAs was a cloverleaf configuration held together by hydrogen bonding and with the CCA bases at the 3′ position and sticking out. Since then, sequences of some 75 different tRNA molecules have been determined, and they all can be organized in the same general cloverleaf folding.

The middle loop of the cloverleaf on the opposite end from the CCA contains a triplet of unpaired bases that form the anticodon, these three bases being complementary to the triplet of bases in the mRNA that represent a codon, or one word in the information in the mRNA that translates to one amino acid. The tRNAs range in number of nucleotides from 74 and 91, but of the various tRNA molecules, the overall distance from the CCA at one end to the anticodon at the other end appears to be constant, the difference in the nucleotide number being compensated for by the size of the little loop located between the right hand and

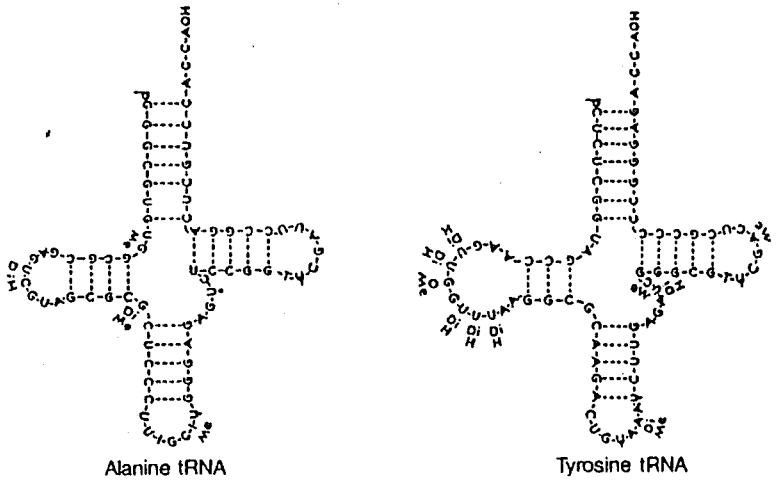
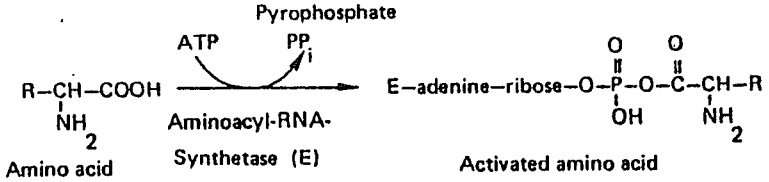


Figure 9.1. Proposed structure for alanine tRNA and tyrosine tRNA.

bottom limbs. Also, the unusual bases are located in regions not forming the hydrogen bonds. Transfer RNAs, unlike the other forms of nucleic acids, must have a highly specific three-dimensional structure, and in this respect they actually resemble proteins.

Amino acids are prepared for use in protein synthesis by being activated in a reaction with adenosine triphosphate (ATP), which requires the enzyme aminoacetyl RNA synthetase. The result is an activated complex of enzyme-AMP-amino acid in which the 5′-phosphate group of adenosine monophosphate (AMP) is linked as a mixed anhydride to the carboxyl group of the amino acid. Pyrophosphate is split from the activating ATP in the process.



In the next step, the activated amino acid molecules are transferred to the corresponding tRNA by another enzyme. There is a particular tRNA for each amino acid, and a specific enzyme for each amino acid to be attached to the tRNA. The tRNA complex containing the amino acid then moves to a ribosome.

The mRNA, having been formed from the DNA and moved to the cytoplasm, has ribosomes attached to it at the chain initiator codon (AUG), which binds, for example, N-formylmethionyl-tRNA. The ribosome moves down the chain of the

mRNA one codon at a time, accepting the respective tRNA complex that corresponds to the codon and connecting the amino acid to the growing peptide chain. As the ribosome moves along the mRNA, the polypetide chain grows. Other ribosomes attach themselves to the mRNA as the initiator codon is vacated. In this way a single mRNA can serve as many as six or eight ribosomes at the same time. There is only one specific site of the ribosome where peptide bond formation can occur, but from each ribosome moving along the mRNA, a polypeptide chain grows, each on an average between 125 and 400 amino acids long.[6]

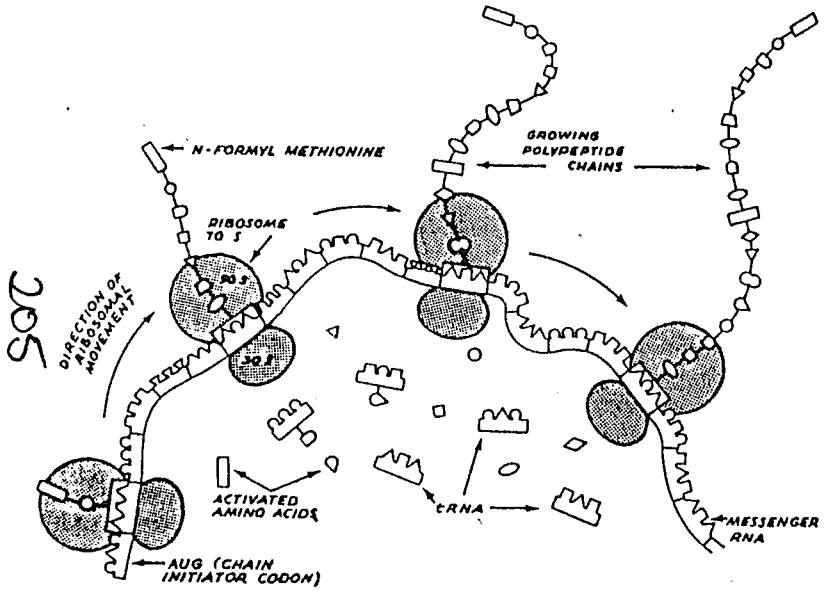


Figure 9.2. Illustration of ribosomes on a mRNA synthesizing polypeptides.

The sequence of purine and pyrimidine bases on the mRNA strand, originally transcribed from the nuclear DNA, directs the arrangement of amino acids in the synthesis of proteins. This information carried on mRNA is not in single nucleotides but resides in sequences of three nucleotides, or triplets. The "genetic code," therefore, consists of nonoverlapping triplets in the mRNA chain, each triplet being a code word, or codon.

After the resolution of the question of how hereditary information was stored in DNA, then transcribed to mRNA, and translated to protein structure, it remained to decipher the code. What nucleotides correspond to what amino acids?

M. W. Nirenberg and J. H. Matthael[7] in 1961 were the first to decipher a codon. They discovered that a synthetic polynucleotide containing only uridylic acid (polyuridylic acid) could serve as a template for ribosomes isolated from the bacterium *E. coli*. When poly U was mixed with ribosome preparations and

tRNA molecules were added, each bearing its own specific amino acid, a polypeptide was synthesized that contained only phenylalanine. From this they were able to conclude that the codon for phenylalanine is UUU.

In similar experiments, GUU directed the synthesis for valine-containing polypeptides; UGU was the code word for cysteine, and UUG for leucine. Eventually, by the use of "block copolymers"—polynucleotides of repeating sequences of two or three bases—the entire genetic code was deciphered.

SECOND LETTER

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

* UAA (ochre), UAG (amber), and UGA (umber) are chain terminating codons. AUG is used as a chain initiating codon standing for formylmethionine in E. coli. In the middle of a protein chain it stands for methionine.

Figure 9.3. The genetic code.

As a result of many additional studies, not only were the triplets for all the amino acids assigned, but codons for special instructions were also found. When biologists discovered that all proteins synthesized by *E. coli* have N-formylmethionine as the N-terminal amino acid residue, they recognized that AUG was the initiator codon. All polypeptides begin with N-formylmethionine, which may or may not be removed later. Three amino acids have six codons, five have four codons, and ten have two codons. UAA, UAG, and UGA are chain-terminating triplets that signal for the peptide to end at that point.

Unlike proteins, nucleic acids are robust molecules. They are stable in mild

acids and alkalies and can be heated to almost 100°C, conditions that wreck the delicate protein structures. Since the other biopolymers can be hydrolyzed without disrupting the nucleic acids, this property facilitates the isolation of DNA and the RNAs from the various cellular components. Another feature that simplifies the purification of DNA is that, in contrast to protein, there is only one kind of DNA in an organism.

Both DNA and proteins are biopolymers, built of subunits that evolved for different purposes. Proteins require shape and different types of functional groups to achieve specificity and chemical activities. To obtain these they needed a variety of subunits, and they found it in the twenty or so amino acids. DNA, on the other hand, did not need many shapes—it needed only to store information. For this purpose it had a choice. It could use a large alphabet (subunits) and have a large vocabulary, or it could use just a few letters but make the informational molecule extremely long. In the end, biological systems adopted a DNA of just four subunits and used three-letter words with a vocabulary of only 64 words. But the litany of life's message seems endless, for the number of units in DNA molecules often runs into the billions.

It can be argued that this is all that DNA needed to accommodate twenty amino acids. But there are many more possible amino acids than twenty that could have been adopted for proteins. Certainly, a DNA constructed of four kinds of nucleotides seems adequate and could have been the type that proved superior to other more complex arrangements by being simpler, more stable, and compact. And there may be no appreciable advantage in using more kinds of amino acids if twenty is enough.

The answer, however, may be for another reason. The emerging cell would have constructed its proteins and nucleic acids from the selection of amino acids and nucleotides in the primordial environment. Once committed to an efficient biological system, however, the primitive cells would have been unable to incorporate any additional building blocks. The amino acids and nucleotides adopted may have been the only ones available in the prebiotic environment of primordial earth.

10
A Thread Unbroken

All living things on earth are tied to an invisible evolutionary thread that stretches back to the beginning of life. The thread is the DNA molecule that is in each of our cells and carries the genetic information for the construction of our very being. That molecule has existed, has been altered, lengthened, and copied generation after generation from the moment the first living cells formed on earth over 3.4 billion years ago. The infrequent changes that occurred in the molecule were retained and passed down to succeeding generations; and each change in DNA became reflected in a change in a protein translated from its structure.

There are many kinds of proteins in a functional cell performing or monitoring essentially all biochemical reactions; a single cell can have 5,000 to 10,000, and man may have as many as 100,000. The metabolic processes, the matrix for bone and shell, the transport of reactants, the synthesis of constituents, and even the shape and mechanical properties of biological systems are all governed by proteins, each protein delicately adapted to perform its specific role by its chemical structure, which in turn is a direct expression of the arrangement of different amino acids in its chain. When, therefore, a change in the amino acid composition of a protein occurs, it often has a profound effect on the entire cell or organism. Nevertheless, this is the basis of evolution, and without it, life could not have developed.

It is the mutability of the genetic apparatus that results in alteration of proteins. A gene is a segment of the DNA molecule and carries the coded message for the synthesis of a single polypeptide—a polypeptide that may itself be a protein or may be combined with other polypeptides to form a protein. If, then, by mutation a nucleotide is changed in the

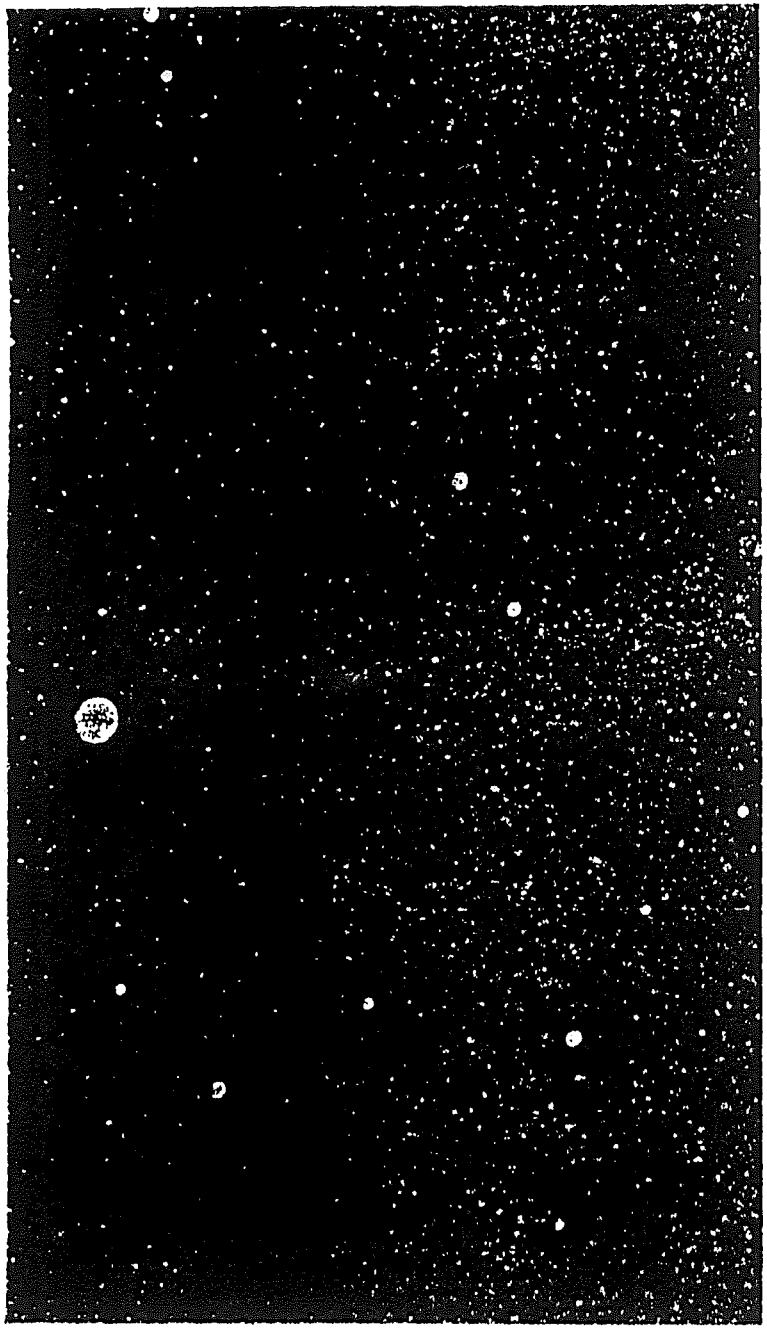


Figure 10.1. A molecule of linear double-stranded DNA. The bar represents 1,000 nm; magnification is 20,000 X.

gene, one of the coded messages is altered and is read for a different amino acid, so that when the protein is synthesized a substitution occurs in the peptide chain. Mutations generally result from the low background level of natural radioactivity, but they can also be caused by certain chemicals.

This is mutation on the simplest level and for single polypeptides. When advanced plants and animals evolved, their DNA molecules became clustered in chromosomes and reproduction became more complex. This allowed for greater variation of the genetic material. The diversity of genotypes can be attributed to different allele combinations, chromosomal interchange, inversion, recombination, or polyploidy. As a result, except in the case of twins, no two individuals of advanced species are identical. This complexity of reproduction introduced a new set of mutations by errors that occur in the procedure, but the mutation of the DNA molecule resulting in an amino acid substitution in a single protein is the subject of interest to molecular evolutionists. Chemically, it is simpler and can be studied analytically and mathematically.

The rate of mutation, like that of radioactive disintegration, is a statistical factor that is impressively constant. E. Zuckerkandl and L. Pauling[1] estimated that each amino acid in hemoglobin undergoes substitution by genetic mutation at an average rate of once every 800 million years. Since hemoglobin has 140 amino acids, this averages to one substitution for the molecule every 5.07 million years.

The constancy of the rate of substitution was confirmed by Motoo Kimura[2] by comparing the number of amino acid substitutions that have occurred in hemoglobin chains of man and the carp. Hemoglobin consists of two sequences of polypeptides called alpha and beta chains that evolved from an ancient hemoglobin of only one chain. Man and carp, with their two-chain hemoglobins, had a common ancestor with the primitive globin that lived during the Devonian period 350 to 400 million years ago, and divergence of the separate evolutionary lines is taken to have occurred around 375 million years ago. By comparing the alpha and beta chains of human hemoglobin which come from separate genes, Kimura found that they differed in the total number of amino acid substitutions by 75. When he then compared the number of substitutions of the human beta chain and the carp alpha chain, he found the difference to be 77, or essentially the same. In other words, after diverging from a common ancestor 375 million years ago, man and carp underwent virtually the same number of mutations in the alpha chain of their hemoglobins.

Various species have many of the same proteins in common which perform the same function, but differ slightly from species to species by amino acid composition. Such proteins are called homologs, and the more closely related species are, the more similar are the homologous proteins in composition. Vernon Ingram[3] of the Massachusetts Institute of Technology suggested in 1961 that the substitution rate in these homologous proteins by mutation could be used as a molecular evolutionary clock.

The degree of divergence in the composition of blood proteins in different

animals is particularly evident in the study of blood serum. When human blood is injected into a rabbit, the rabbit's immune response produces antibodies to the human proteins. This antihuman serum in return, when mixed with human blood, causes clumping of 100 percent of the blood protein. The same antihuman serum mixed with blood of other species gives the following percent precipitation: gorilla, 64; orangutan, 42; baboon, 29; ox, 10; deer, 7; horse, 2; and kangaroo, 0. This is an extremely critical measure of chemical resemblance and affinity, and the order of relationship correlates with evidence from comparative anatomy, embryology, and paleontology.

Molecular evolution can now compare this divergence of amino acid composition of homologous proteins directly. Since 1965, the known sequences of proteins have been published annually by Margaret Dayhoff and her colleagues of the National Biomedical Research Foundation in the *Atlas of Protein Sequence and Structure*. The number of entries has grown from 50 protein sequences in the 1965 edition to 409 in 1972.[4] By the known sequences of homologous proteins from a number of species, a comparison can be made to assess to what extent each species has undergone amino acid substitution in the protein by mutation.

One of the earliest proteins to be studied was cytochrome c by Emanuel Margoliash of Abbott Laboratories and Walter M. Fitch of the University of Wisconsin Medical School.[5] This heme protein is easily extracted and purified and is found in every mitochondrion of eucaryotic cells, which is to say, of all living things above bacteria and blue-green algae. It therefore is the expression of a distinct, recognizable gene that has existed for some 1.3 billion years.

Cytochrome c in the vertebrates has 104 amino acids, a few more in others. Because the structural shape of cytochrome c is closely associated with its ability to perform by interacting with cytochrome oxidase and cytochrome reductase, the surface of the protein is conserved. Any mutation whose amino acid substitution changes the protein's shape is deleterious, and consequently, does not survive. As a result, 50 percent of the molecule remains invariant. Within the variable amino acids, there are 19 particular ones that play a role in the conformation of the polypeptides, and 6 others that have remained the same for undetermined reasons. There is also a constancy of the hydrophobic and basic segments of the protein, and some classes of amino acids are mutably interchangeable.

Thus, in studying the mutation distance between cytochrome c's of 20 species ranging from yeast to man, Fitch and Margoliash found a connection between the taxonomic kinship of species and the number of residue differences. Proteins of closely related species showed few or no differences, whereas the greatest differences in amino acids was with the most diverse species: the monkey and human lines are distinguished by a single mutation in the human line that resulted in the substitution of isoleucine for threonine at one position; and the horse differs from the tuna fish by 20 amino acids, from Baker's yeast by 45. The average difference between primates and the other mammals is 10.1 residues. With the amino acid substitutions from mutation through the history of cyto-

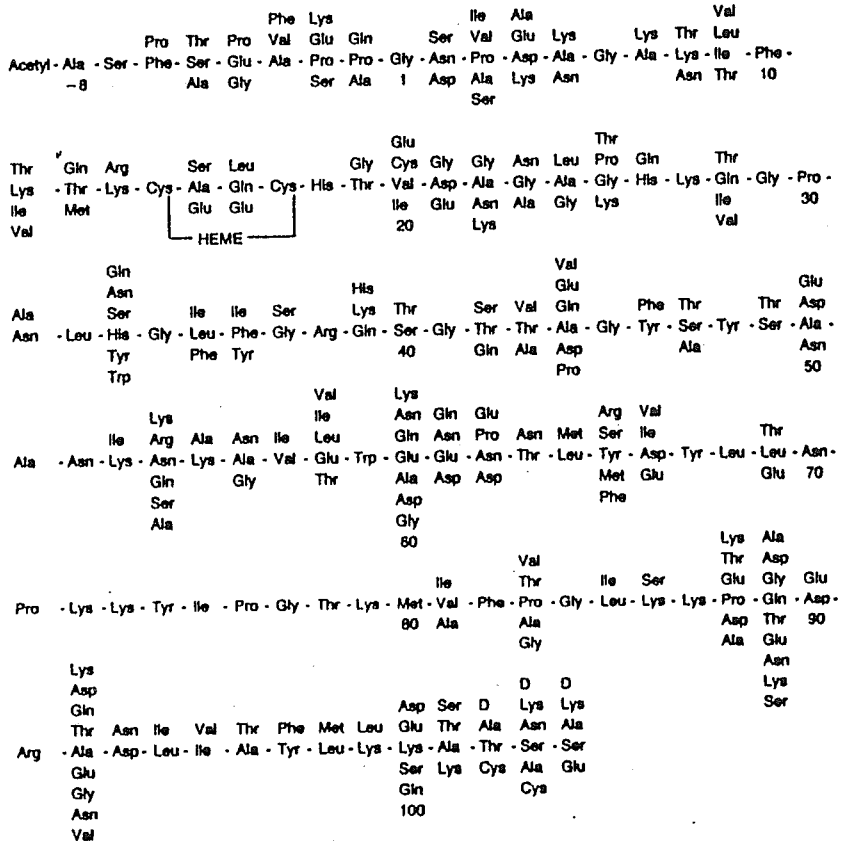


Figure 10.2. Composite of amino acid sequence of 30 eucaryotic cytochrome c's. The continuous sequence of 112 residues is that of wheat germ protein. All the other cytochrome c's are shorter with their various amino acids shown above and below the linear sequence.

chrome c calculated, Fitch and Margoliash were able to reconstruct the phylogenetic tree. Although substitutions in cytochrome c may not have been the mutation most responsible for a particular divergence, its evolution coincides remarkably with the taxonomic development of the species.

To derive the topology of a phylogenetic tree, the amino acid sequences of homologous proteins are programmed in a computer. The computation is based on the premise that the fewest number of changes occurred in the ancestral organism. The computer then considers each amino acid along the polypeptide chain, one at a time. Amino acids that have remained unchanged have been conserved and are considered invariant. For the others, the method involves a mathematical formula for each possible combination to take into account possible sequences for the ancestral structures. The detailed procedure is described in chapter 2 of the *Atlas of Protein Sequence and Structure* (1969).

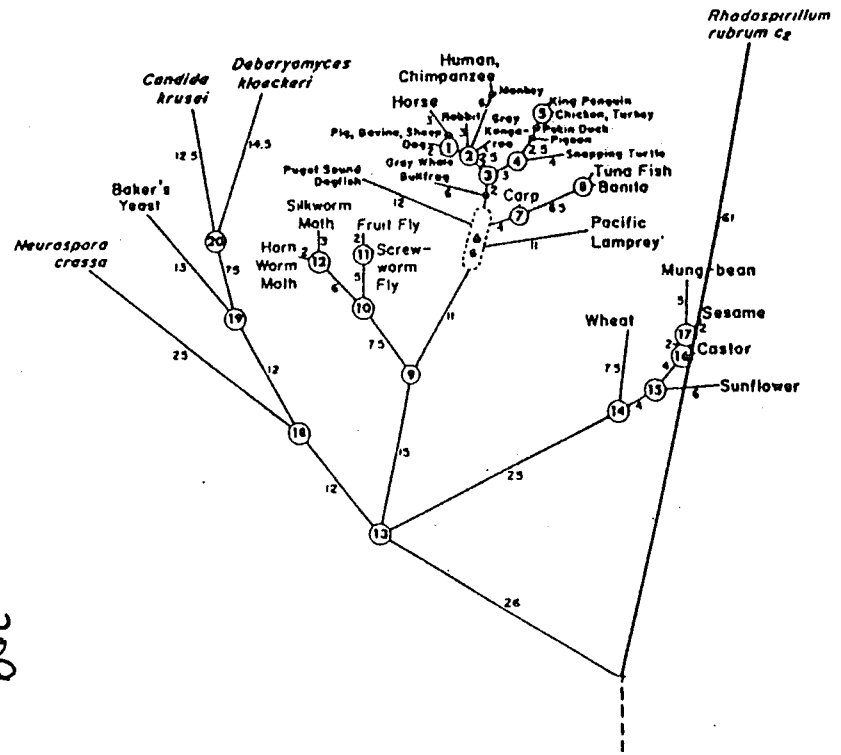


Figure 10.3. Phylogenetic tree based on cytochrome c. The numbers are the inferred amino acid changes per 100 links. (*Atlas of Protein Sequence and Structure*, 1972).

Motoo Kimura and Tomoko Ohta[6] have tabulated some of the principles governing molecular evolution. The rate of amino acid substitution is approximately constant if the change does not affect the function or tertiary structure of the protein. Therefore, those parts of the protein that are functionally least important undergo the fastest change. And when a gene having a new function is evolved, it is preceded by gene duplication. In this way the organism has one copy to accumulate mutations and eventually emerge as a new gene, while the other copy retains the old function required for survival.

A study of the globins reveals species divergence coincidental with duplication of genetic material in the last billion years. The alpha and beta chains of hemoglobin, presently formed from separate genes on different chromosomes, were derived from a single ancestral gene by an ancient duplication of the genetic structure. The hemoglobin of the lamprey, a primitive fish, still retains a single chain. And too, myoglobin, a related heme protein found in muscle, evidently is the result of a divergence from hemoglobin by an ancient duplication that took place in the pre-Cambrian, approximately a billion years ago. In this way, as

organisms evolved to more advanced levels, their genetic material and associated proteins expanded in number and complexity.

A gene contains an average of 1,000 nucleotides; and in man the total number of nucleotides making up a genome (the set of chromosomes and all their genes) is about 4 billion. This is roughly the same for all mammals. In a study of seven proteins, Jack King and Thomas Jukes[7] found in the evolutionary history of mammals the number of amino acid substitutions averaged 2 to 2.5 per year in a species. If the average generation time in the line to man was 10 years, there would have been 20 substitutions in each generation. This is much too rapid for a stable species; variants would be lost before natural selection would have time to derive potential advantages that they might present. For this reason, the majority of changes in proteins have been neutral or almost neutral in natural selection.

The rate at which different proteins evolve depends upon the number of amino acids that are invariant for the protein to be functional. To study the rate of molecular evolution Russell Doolittle and Birger Blombäck[8] chose the fibrinopeptides. Unlike hemoglobin and cytochrome c, the fibrinopeptides do not have to conserve particular amino acids for their biological role. They are the peptide segments, 13 to 21 amino acids long, that are cut out of fibrinogen when it is converted to fibrin in a blood clot. Since fibrinopeptides can consist of virtually any amino acid, they are excellent for studying the unrestricted rate of mutation.

Application of the fibrinopeptides is essentially limited to the mammals, but their information is an outstanding contribution for hominoid evolution. Humans, chimpanzees, and gorillas all have identical fibrinopeptides, whereas there are differences in the fibrinopeptides of such closely related pairs as the cat–lion, dog–fox, donkey–horse, and the water buffalo–Cape buffalo. In studying the molecular evolution of the six hominoids by comparing their fibrinopeptides, Doolittle and his coworkers[9] discovered that, although the gorilla is identical to man and chimpanzee, he differs from the Asian apes, the orangutan and siamang. It appears that this chemical evidence of man's relationship to the African apes supports Leakey's thesis that our early progenitors lived in Africa.

Paleontologists tend to place the emergence of man distinct from the other primates as early as 14 million years ago.[10] Molecular evolutionists, on the other hand, believe that man and the African apes diverged as late as 4 to 5 million years ago.[11] For better or for worse, the chemical analysis and comparisons are less susceptible to subjectivity than taxonomy. Consequently, we learn that the molecular relationship between man and the other primates is much closer than is generally believed—and actually closer than between members of some other recognized species. Of 50 different proteins isolated from man and chimpanzee, the difference in amino acid sequences has been found to be less than 1 percent.

There are two forms of chromosomal organization: the single chromosome of bacteria and blue-green algae as a double-stranded DNA in a closed loop, and the beadlike chain of chromosomes of the higher organisms. Mutations can occur

by errors in a number of ways in replication of the chromosome arrangement, but these chromosomal aberrations are almost always lethal. On the other hand, the effect of a mutation of the genes resulting in amino acid substitution depends upon the protein. If the substitution has no effect on the function of the protein, it is neutral; but if the substitution alters the shape and efficiency of the protein, it will be deleterious to the organism, and the mutation will not survive by loss of the individual from reproduction.

A species is an interbreeding pool, so any surviving mutation can spread throughout a breeding range. New species arise when mutations accumulate in reproductively isolated groups. As the changes from mutations become prevalent, interbreeding with populations of the ancestral species is first rendered unusual, then eventually it becomes impossible.

Whereas substitutions of invariant amino acids in existing proteins is a mutation that is not retained, it is the means by which new proteins achieved their optimal amino acid sequence for functionality. In the early stages of life on earth the initial proteins were probably of low efficiency, but were continually improved upon by amino acid substitution until they attained maximum effectiveness. Since then, the sequence responsible for their function has been rigorously conserved. But without mutation, life would have remained on the level of the first primitive cells on primordial earth.

It appears that life began on earth with a few basic reactions to synthesize components, and hence perpetuated itself by reproduction. The proteins for these fundamental reactions were conserved, and as organisms evolved, they developed improvements by introducing new reactions to the old ones and refining the efficiency of biochemical conversions. Because of this conservative manner in which evolution proceeds, we all have within us "relics" of our progenitors, even of the ancient microscopic organisms that floated in Archean seas. Life does not discard what the survival of the species has depended upon, but retains the old alongside any new developments. For this reason, many of the basic biochemical reactions derived from extremely remote ancestors, little changed in billions of years, can be seen in the biochemical architecture of contemporary plants and animals.

One such vestige of our origins in an anaerobic environment is a metabolic pathway for glycolysis. A main source of nutrition now and probably from the beginning is the metabolism of carbohydrates. In the simple process of fermentation, glucose is broken down to pyruvic acid, yielding 11.3 kilojoules of chemical energy per mole of sugar. Contemporary anaerobic organisms carry out this conversion in 10 reactions. Since the development of this basic series of reactions, it has been extended by some organisms to metabolizing pyruvic acid farther to lactate, propionate, acetate, ethanol, acetone, butyric acid, and higher fatty acids. Despite the fact that the decomposition of carbohydrates to pyruvic acid by fermentation has been overshadowed by more evolved metabolic pathways in all except the anaerobic organisms, it is preserved in the biochemistry

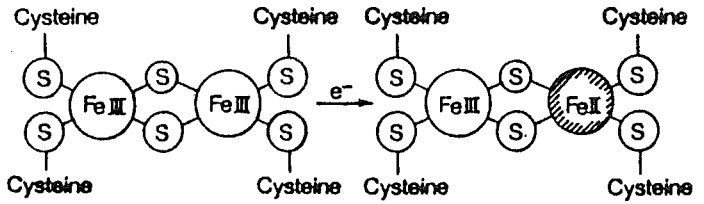


Figure 10.4. Model of the active site of ferredoxin.

of higher plants and animals as a relic from an age when it was the sole means our primitive anaerobic ancestors had of extracting chemical energy.

But the development of the whole biochemical structure has branched out from a few fundamental reactions. The result of this is that the original reactions can no longer be changed. Too much of the overall biochemistry of the organisms stems from them and depends upon their products. For example, acetate is involved in energy transfer systems, but it is also the essential starting material of such diverse components as carbohydrates, amino acids, and fats. Even if some other substance could be found that worked better than acetate, it could not be adopted without devastating consequences because so many reactions require it. A single change would pyramid into changing much of the metabolism of the organism.

This conservation of biochemical processes accounts for the close unity of all living things in the very fundamental steps that make life possible. This is why the determination of the amino acid sequences of many proteins and the application of computers have allowed biochemists to trace back the evolutionary development of today's plants and animals on a molecular level. Whereas the genetic complement of contemporary mammals is estimated to be 100,000 genes, the number of metabolic reactions that the various phyla have in common is quite small. From this it appears that all animals and man descended from an ancestral primordial cell that had only about 200 genes.[12]

The gene that transcribes cytochrome c dates back as early as the development of the eucaryotic cell 1.3 billion years ago. But a universal protein that may extend as early as the beginning of life on earth is ferredoxin, an iron-containing protein that is vital in photochemical reactions for electron transport to cellular energy storage. Ferredoxin has a reducing potential near that of molecular hydrogen, making it the most highly reducing stable compound in a cell and suggesting that it evolved at a time when the earth's atmosphere was still strongly reducing.

The various roles of ferredoxin in the cell are fundamental: it assists in the ATP formation by radiation;[13] it participates in the reduction of carbon dioxide to pyruvate; and it is used in the fixation of nitrogen.[14] Apparently ferredoxin is more ancient than nicotinamide adenine dinucleotide (NAD), a ubiquitous

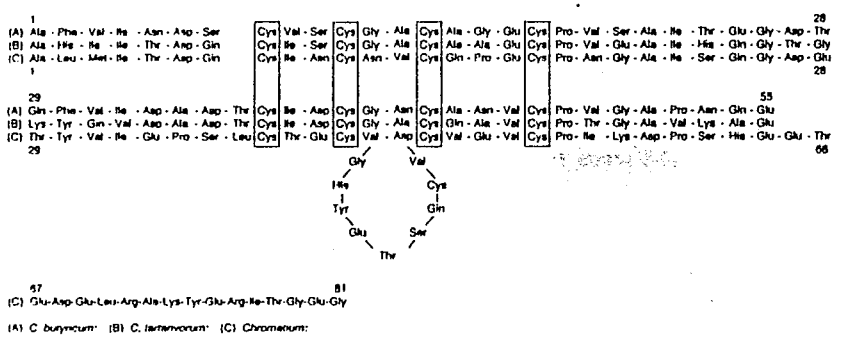


Figure 10.5. Comparison of sequences of *Clostridium* and *Chromatium* ferredoxins.

reducing agent in cells. In the primitive microbe *Clostridium pasteurianum* and in the photosynthetic bacterium *Chromatium thiosulfatophilium* ferredoxin participates directly as a reductant of carbon dioxide with acetyl coenzyme A, instead of through NAD.[15]

Ferredoxin is an iron–sulfur protein of only 55 amino acids, consisting of an unusually high proportion of the smaller and thermodynamically stable amino acids: glycine, alanine, serine, aspartic acid, and cysteine. From a study of the sequence of amino acids of ferredoxin, Richard Eck and Margaret Dayhoff[16] concluded that the protein had an ancestral sequence of 29 units and that the original molecule was based on a repeating sequence of alanine, serine, aspartic acid, and glycine. It appears that the original genetic mechanism was a sequence of 12 nucleotides that doubled, then doubled again, making a long, repetitive chain. Later, as the genetic code became more complex, other amino acids, including cysteine, were added and the sulfide bond of the cysteine became attached to the iron. Eventually, four cysteines were added by mutation and two identical chains combined to make an intricate protein–iron–sulfide complex of greatly increased efficiency.

Ferredoxin occurs in the primitive anaerobic organisms, both photosynthetic and nonphotosynthetic, and is basic to cell chemistry. The simplicity and evolutionary development of ferredoxin suggest that it may have been one of the earliest proteins formed by life on earth.[17]

Life extends uninterruptedly from the first living cell on earth down through the ages to all its living descendants. The genetic information coded in the thread of DNA nucleotides has been passed on from generation to generation, slowly being improved upon, becoming more efficient, adding new genes, and progressively moving toward the higher forms of plants and animals.

All of today's descendants are genetically the same age. But all have not climbed the evolutionary ladder to the same heights. There are the "living fossils"—plants and animals that exist little changed from their fossil ancestors.

The shark has remained essentially the same for the last 70 million years, the horseshoe crab for 180, and the cockroach, scorpion, millipede, and nautilus have changed little in several hundred million years. The ginkgo tree of China flourished during the age of the dinosaurs, and in 1958 off the coast of southwest Africa some fishermen caught in their nets a strange fish they had never seen before. And for good reason: the fish was a coelacanth, a primitive species thought by paleontologists to have become extinct around 150 million years ago.

The living fossils extend even to the most ancient forms of life. About 15 years ago, Sanford Siegel, a University of Hawaii botanist who was on a chance visit to Harlech Castle in Wales, observed tourists honoring a time-seasoned practice passed down from medieval days. In the manner of the knights of old, or at least of those of the soldiers who stood guard duty, they were urinating at the base of the castle walls. Siegel, supported by the National Aeronautics and Space Administration, had worked for several years investigating microorganisms living under harsh environments likely to be encountered in space travel. Because ammonia is a major component of the atmosphere of Jupiter and may have been common on primitive earth, it was one of the environments studied. A natural condition high in ammonia would be soil saturated with urine.

Siegel returned home with soil samples he had collected from around the castle walls at Harlech and attempted to culture them in concentrated ammonium hydroxide. Most organisms would have been killed or greatly inhibited by the medium, but Siegel observed one growing in microscopic clusters of star-shaped bodies attached to slender stalks. It did not fit a description of any known living organism, although it did closely resemble a Precambrian microfossil Barghoorn had discovered in a 2-billion-year-old Gunflint chert at Kakabeka, Ontario, Canada, which he had named *Kakabekia umbrellata*.[18]

Siegel hypothesized that he had found a living microfossil that was an obligate ammonophile, an organism that requires ammonia to grow. Since he collected the first specimens at Harlech Castle, studies by Siegel and his wife, Barbara,[19] have shown that the organism's need for ammonia is not absolute. The organism has also been found in soils from Alaska, Iceland, and various alpine regions that are low in ammonia but high in alkalinity. It does not need oxygen, but, unlike most anaerobic bacteria, is not killed by it. Siegel's discovery, which may be a living relative of Barghoorn's fossil from the Middle Proterozoic period, was named *Kakabekia barghoorniana*.

The cyanobacteria, although widely dispersed, have probably remained essentially the same for at least a billion years. And the anaerobic microbes that were dominant for the first 2 billion years of life before free oxygen existed in any significant amounts have managed to remain with us in sheltered niches and are represented today by the clostridia that survive to give us tetanus, botulism, and gas gangrene. The clostridia, which lack even cytochrome c, are listed as one of the most primitive forms of life. Even the photosynthetic bacteria, those primitive microbes that have the ability to synthesize organic matter from carbon

Figure 10.6. The gatehouse at Harlech Castle.

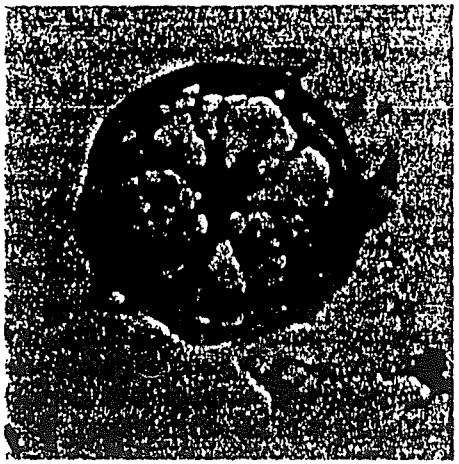


Figure 10.7. *Kakabekia barghoorniana* from Wales.

dioxide in an anaerobic environment using hydrogen sulfide as their source of hydrogen and requiring only acetic acid, light, and some minerals, survive as one of the earliest life forms on earth.

But, although the living fossils appear to be frozen in time while other plants and animals are undergoing dynamic evolution, this is not the case. Even the genes of these relics of the past have undergone mutation at the same rate as all other living descendants. In order to maintain an unchanged morphology they apparently have been under an incessant action of natural selection for hundreds of millions and even billions of years, while a steady stream of almost neutral genetic variations has flowed through, transforming their informational molecules tremendously.

All living organisms have within their biochemical structures traces of the events that have led to the advancement of life for the last 3.5 billion years. Unlike the fossil record, which has left many branches that became deadends, the chemical relics within all living things are derived from direct ancestry. Application of molecular evolution is an immensely valuable technique. With it, scientists can conceivably fulfill an old dream—trace back the evolution of the species to the very beginning of life itself.