11

LIFE IN THE UNIVERSE

A Man that is of Copernicus' Opinion, that this Earth of ours is a Planet, carry'd round and enlighten'd by the Sun, like the rest of the Planets, cannot but sometimes think, that it's not improbable that the rest of the Planets have their Dress and Furniture, and perhaps their Inhabitants too as well as this Earth of ours.

. . . But perhaps they'll say, it does not become us to be so curious and inquisitive in these Things which the Supreme Creator seems to have kept for his own Knowledge: For since he has not been pleased to make any farther Discovery or Revelation of them, it seems little better than presumption to make any inquiry into that which he has thought fit to hide. But these Gentlemen must be told, that they take too much upon themselves when they pretend to appoint how far and no farther Men shall go in their Searches, and to set bounds to other Mens Industry; as if they knew the Marks that God has placed to Knowledge: or as if Men were able to pass those Marks. If our Forefathers had been at this rate scrupulous, we might have been ignorant still of the Magnitude and Figure of the Earth, or that there was such a place as America.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

14

On the definition of life

. . . For any living thing that has reached its normal development and which is unmutilated, and whose mode of generation is not spontaneous, the most natural act is the production of another like itself, an animal producing an animal, a plant a plant, in order that, as far as its nature allows, it may partake in the eternal and divine. That is the goal towards which all things strive, that for the sake of which they do whatsoever their nature renders possible . . . Since then no living thing is able to partake in what is eternal and divine by uninterrupted continuance (for nothing perishable can forever remain one and the same), it tries to achieve that end in the only way possible to it, and success is possible in varying degrees; so it remains not indeed as the selfsame individual but continues its existence in something *like* itself—not numerically but specifically one . . .

Aristotle, *De Anima*

. . . Ultimately life can be unequivocally explained in physicochemical terms. . . . We eat, drink, and reproduce not because mankind has reached an agreement that this is desirable, but because, machine-like, we are compelled to do so.

Jacques Loeb, *The Mechanistic Conception of Life,* 1912

. . . if "dead" matter has reared up this curious landscape of fiddling crickets, song sparrows, and wondering men, it must be plain even to the most devoted materialist that the matter of which he speaks contains amazing, if not dreadful powers, and may not impossibly be, as Hardy has suggested, "but one mask of many worn by the Great Face behind."

Loren Eiseley, *The Immense Journey,* 1946

▽ The problem of his own beginnings has intrigued man since remotest antiquity. Of more recent origin—and of perhaps even greater fascination—is the question of life on other worlds beyond the Earth. It is our immense good fortune to be alive at the first moment in history when these tantalizing issues can be approached with rigor and in detail. To hold in our hands the keys to these ancient riddles is a triumph of the highest order; it heralds an age of exploration and discovery unsurpassed in the history of mankind.

▽ The questions of extraterrestrial life and the origin of life are intertwined. But before we can approach either question, we must have some general understanding of the nature of living systems. Here, we face a fundamental handicap: our knowledge of biology is restricted to essentially one example. The inner workings of terrestrial organisms—from microbes to men—are so similar in their biochemical details as to make it highly likely that all organisms on the Earth have evolved from a single instance of the origin of life. This hypothesis is supported by a variety of observations. All organisms are composed of one or more cells. The organization and functioning of these cells show enormous similarities. In the biochemistry of the photosynthetic and respiratory apparatus; in the detailed reproductive behavior of cells; in the ubiquity of the molecule DNA as the genetic material; in the details of the breaking down of foodstuffs to extract energy; in the asymmetry of the constituent molecules; even in the microstructure of membranes and flagellae, and in the molecular basis of animal colors, the same materials, the same methods have been used over and over again in the extremely diverse aggregate of plants and animals which we describe collectively as "life on Earth."

▽ Thus, we see immediately one reason why the discovery and characterization of extraterrestrial life has a pervasive appeal to the biologist. He might then be able to separate what is necessary from what is contingent. He might begin to learn which characteristics terrestrial living systems have because *any* living system must have them; and which characteristics are historical accidents, the results of arbitrary and random concatenations of events which might elsewhere have developed along different lines and produced different kinds of living systems. In other sciences, it is possible to test Earthbound insights elsewhere in the universe. Indeed, this is one reason why physics and chemistry are, in a sense, "universal" sciences. But for all we know, biology is literally mundane and provincial, and we may be familiar with but one special case in a universe of diverse biologies.

▽ Because we have only one example, the question of the definition of life is beset with difficulties. Remembering that our conclusions may lack the generality

we desire, let us try to find what life on Earth is all about. Is life merely a particularly subtle organization of matter, or is there something more to it? Any child can tell the difference between a live puppy, a dead puppy, and a toy puppy. Exactly what is this distinction?

▽ In primitive times, when very little was understood about the nature of living systems, the most routine biological activities, such as the germination of a seed or the flowering of a plant, were attributed to divine intervention. In the early years of the Industrial Revolution, when advances in celestial mechanics gave something close to a complete understanding of the positions and motions of the heavenly bodies, the concept arose that living systems may be nothing more than a particularly intricate kind of clockwork. But when early investigations failed to unveil the clockwork, a kind of ghostly mainspring was invented—the "vital force." The vital force was a rebellion from mechanistic biology, an explanation of all that mechanism could not explain, or for which mechanisms could not be found. It also appealed to those who felt debased by the implication that they were "nothing more" than a collection of atoms, that their urges and supposed free wills arose merely from the interaction of an enormously large number of molecules, in a way which, although too complex to use predictively, was, in principle, determined.

▽ But today, we find no evidence for a vital force; indeed, the concept is very poorly defined, a kind of universal catch-all for anything we cannot explain. The opposite tack—that all living systems are made of atoms and nothing else—has proved a particularly useful idea. An entire new science of molecular biology has made startling progress and achieved fundamental insights starting from this assumption. And there is nothing debasing in the thought that we are made of atoms alone. We are thereby related to the rest of the universe; and if we are made of the same stuff, more or less, as everything else, then elsewhere there may be things rather like us. We are a tribute to the subtlety of matter. As the distinguished American physicist Richard P. Feynman of the California Institute of Technology has put it to a lecture audience:

> If a piece of steel or a piece of salt, consisting of atoms one next to the other, can have such interesting properties; if water—which is nothing but these little blobs, mile upon mile of the same thing over the Earth—can form waves and foam, make rushing noises and strange patterns as it runs over cement; if all of this, all the life of a stream of water, can be nothing but a pile of atoms, *how much more is possible?* If instead of arranging the atoms in some definite pattern, again and again repeated, on and on, or even forming little lumps of complexity like the odor of violets, we make an arrangement which is *always* different from place to place, with different kinds of atoms arranged in many ways, continually changing, not repeating, how much more marvelously is it possible that [matter] might behave? Is it possible that that "thing" walking back and forth in front of you, talking to you, is a great glob of these atoms in a very complex arrangement . . . ? When we say we are a pile of atoms, we do not mean we are *merely* a pile of atoms, because a pile of atoms which is not repeated from one to the other might well have the possibilities which you see before you in the mirror.

▽ Let us consider an animal—any animal—in as general a way as we can. (We could just as well consider a plant, but animals are more fun.) What does it do? Mostly, it does nothing. It waits in a hole, lies in the grass, floats downstream, or rocks on the porch. Occasionally it does something interesting: it acquires a fragment of organic matter produced by some other organism, and incorporates it —not quite all of it, because some passes right through the animal virtually unchanged. But the ingested material per se is not utilized; it is broken down further, chemically rearranged, built up into molecules that the animal needs, and then used.

▽ Eating is so common that we tend to forget what an extraordinary process it is. We eat corn flakes, and yet, we do not *become* corn flakes. The corn flakes are changed into *us*. In Feynman's splendid phrase, "Today's brains are yesterday's mashed potatoes."

▽ How did such an effortless ability for molecular transmutation ever develop? And if the animal is an adult, why must it eat at all? It has reached its optimum mass; why must it go to all that trouble to seek, process, and utilize food? There seem to be two general reasons. First, the animal eats in order to acquire energy, so it can carry out other biological processes—e.g., moving, breathing, or further eating. Second, it eats so that it has a source of materials for repair. The chemical bonds which hold together the molecules composing the animal tend to break. If unchecked, molecules would dissociate, cellular systems would dissipate, and the organism would die. The relatively short periods of time which most familiar animals can survive without eating testify to the essential instability of higher living systems. Some organisms, when faced with a food shortage, or other environmental embarrassments, can simply shut up shop and wait until conditions improve; but most higher organisms need constant alimentary reassurance.

▽ Of the other obvious activities of animals, most are either accessory to eating or unessential. Respiration is a device to extract the maximum amount of energy from food; motility tends to insure the acquisition of food; excretion is a method of removing unmetabolizable food; irritability—the response of the organism to external stimuli—and the ability to learn increase the probability that the animal does not end its days in the gullet of some other animal: they are also properties of some electronic computing machines which no one is willing to call "alive," quite yet.

▽ The remaining characteristic, which does seem essential for living systems and which all animals share, is reproduction. It is true that a narrow definition of a living system as a self-reproducing one would seem to exclude mules; but despite superficial appearances, even they reproduce at a colossal rate, or at least their constituent cells do.

▽ If we observe the reproductive habits of animals, we are struck by several facts. Reproduction of the whole animal occurs relatively rarely. The animals are exceptionally strongly motivated to reproduce themselves, even though there is no obvious material benefit which thereby accrues to them. Organisms tend to reproduce their own kind, so that reproduction occurs only within a species.

Reproduction is far from identical, especially in higher animals, where fertilization is accompanied by a random reassortment of the parental genetic factors in establishing the characteristics of the new animal. There is generally an age when most animals are no longer capable of reproduction. Finally, we observe that most animals die a natural death soon after this age is reached.

▽ Can it be that reproduction is in some sense the "point" of biological activity? We can imagine an organism which carries out metabolism and all the other functions ordinarily ascribed to living systems in elementary biology textbooks; which has very efficient repair mechanisms, so that it easily survives the vicissitudes of its environment; and which has no reproductive organs and never reproduces. We can imagine such an organism, but we never find one. Why not? Because there is no way for such an organism to arise. The only mechanism which we know for the production of biological complexity is evolution by natural selection, the differential survival of the organisms which, by chance, are best adapted to their environments. But natural selection can occur only if the well-adapted organisms reproduce themselves. Thus, the development of complexity in living systems is intimately connected with their self-replication.

▽ The paleontological record shows clearly the gradual development of biological complexity during the history of the Earth. A billion years ago, there was, apparently, nothing more complex than one-celled protozoa and their colonies. Yet today you are a walking collection of about 10^{14} cells, coordinated in origin and function. Each of your cells bears strong family resemblances in size, function, and chemistry to contemporary protozoa. The principle of evolution by natural selection permits us to understand how this increase in complexity has occurred. It is not that complexity per se has survival value, but rather that a solution to an environmental crisis which involves many molecules is often qualitatively superior to a solution which involves only a few molecules. For example, the image-forming eye of the vertebrates is a qualitatively superior light receptor than the pigmented eye-spots of the protozoa; but a protozoan cannot construct an eye, because an eye is constructed of more molecules than there are in the entire protozoan. In an environment where the ability to detect swiftly-moving predators or prey is at a premium, organisms with efficient visual receptors will preferentially survive. Efficiency and complexity are here coupled. We expect that in time and through natural selection, large living systems of extraordinary complexity will develop, adapted in detail to their environments.

▽ The enormous complexity of even a simple, single-celled organism is illustrated in Figure 14–1, a photograph of part of an algal cell, taken with an electron microscope. The magnification is 25,000 times. At the periphery can be seen the cell wall, which separates the alga from its environment. Towards the upper left-hand corner is the cell nucleus, which contains the genetic material. Between the nucleus and the cell wall is the cytoplasm of the cell, which contains elaborate apparatus for photosynthesis, respiration, and enzyme production and operation, among many other functions. (The enzymes are largely proteins, composed of sequences of amino acids.) The cell is no sack of enzymes and other

FIGURE 14–1. *An electron micrograph of a portion of an algal cell, a simple plant. The magnification is 25,000 times. (Courtesy of Dr. G. E. Palade, Cytology Laboratory, The Rockefeller Institute for Medical Research, New York.)*

chemicals. It has a detailed, functional architecture of great sophistication and complexity. Such a "simple" cell is clearly well along in the evolutionary process, and when we speak of the origin of life, we must be speaking of the origin of a much simpler entity.

▽ If we now follow the evolution of biological complexity backwards in time, we can imagine self-replicating and mutating entities even simpler than a cell, yet still capable of subsequent steps up the evolutionary ladder. Since the universe is primarily composed of very simple non-self-replicating molecules, we must eventually come face to face with the problem of the origin of the first self-replicating system, the subject of the next chapters. However, for the moment, let us probe a little deeper into our hypothetical animal.

▽ It reproduces. How does it reproduce? Think of the enormous number of characteristics which familiar animals have. There is the gross anatomy, the overall architecture of the organism. Then, there is the physiology, the dynamic functioning and articulation of the different parts of the organism in carrying out its functions. It has inherited behavior patterns—how to build a nest, how to bury a bone. It has ten trillion or so cells, each one of which is itself an extraordinarily complex structure. At the present time, we are making only the first fumbling steps towards assembling a cell from scratch. Yet the information to construct the entire organism is somehow contained in the genetic material, because, with striking regularity, animals look like their parents.

▽ The problem of heredity is really two problems: How is the genetic information transmitted from generation to generation? And how is this information translated into action, in the development of the new organism? These two questions can be phrased another way: What is the genetic "code"? And how, in the developing organism, is the code "read"?

▽ The most significant aspects of life are often not the most obvious. The most abundant organic molecule on Earth is cellulose; yet we are not constructed of cellulose, and we have great difficulty metabolizing it. Understandably, we might feel that its significance has been overrated. A tree, if its opinion could be polled, would probably disagree. The phylum with the largest number of identified species is the arthropods; in this sense, life on Earth is mostly beetles. Yet we would feel, with justified annoyance, that a biological survey party from some other planet that spent all its time on Earth studying the beetles had overlooked some items of importance. Similarly, it is possible to be misled when we examine the chemical composition of a typical cell—say, a bacterium—whose molecular census might be as follows: lipids, 30 million molecules; phospholipids, 20 million molecules; proteins, 5 million molecules; polysaccharides, 1 million molecules; ribonucleic acid (RNA), 40 thousand molecules; deoxyribonucleic acid (DNA), 1 molecule. If biochemists were to concentrate all their attention on the lipids, they would also be missing the point.

▽ In experiments with peas and other flowering plants, the Moravian monk Gregor Mendel, in 1865, derived certain empirical rules for the transmission of hereditary characteristics, subsequently known as Mendel's laws. At about the

same time, the nucleic acids were isolated, and the chromosomes discovered. Yet the connection among Mendel's empirical laws, the microscopic behavior of the chromosomes, and the chemistry of the nucleic acids was not demonstrated until the 1950's; indeed, some aspects of the connection—especially between chromosomes and nucleic acids—remain obscure today.

▽ The chromosomes are small threadlike bodies in the cell nucleus which undergo an intricate ritual of duplication and segregation during the reproduction of a cell. In the first decade of the twentieth century, it was realized that the chromosomal choreography was exactly the process required by Mendelian genetics to account for the transmission of heredity characteristics. Thus, the original basis for the belief that the chromosomes are the genetic material was their behavior during reproduction and had nothing to do with their chemical composition. The fact that chromosomes are composed, to a significant degree, of nucleic acids has in turn supported the view that the nucleic acids are involved in heredity. Many more powerful demonstrations of this thesis are now available, including the observation that the DNA of a virus, injected into a bacterial cell, can change the entire function of the cell, converting it from a factory for making more bacteria to a factory for making more viruses.

▽ The genetic material of all known organisms on Earth is composed largely of DNA and RNA. These nucleic acids have coded into their structure the information which is reproductively transmitted from generation to generation. In addition, they have the capability for self-replication and mutation. DNA serves as a kind of molecular blueprint which controls metabolism, produces a replica of itself for the next generation to follow, and, through the centuries, gradually changes, or mutates, occasioning new forms of life.

▽ The structure and function of DNA have been elucidated chiefly by the American molecular biologist James D. Watson of Harvard University, and the British molecular biologist Francis H. C. Crick of Cambridge University. DNA is a long molecule, comprising two molecular strands wound about each other in a coil or helix; a short section of a DNA molecule is shown schematically in Figure 14–2 and as a molecular model in Figure 14–3. During cell division, the strands separate, and each synthesizes a copy of the other, yielding two molecules of DNA where originally there was only one. This is the primary molecular reproductive event. The building blocks for this synthesis are called nucleoside phosphates. Much of the activity of the cell is devoted to constructing these building blocks from yet simpler molecules acquired from ingested food and joining them together to form nucleic acids. The nucleoside phosphates are each composed of a sugar, a base, and some phosphates. A given nucleic acid molecule is generally composed of four kinds of nucleoside phosphates. Their sequencing along the chain is a kind of four-letter code that determines which sequences of amino acids, and therefore which proteins, a cell will make.

▽ In Figure 14–2 the two helical strands can be seen running vertically in opposite directions on the right- and left-hand sides of the Figure. As the detailed inset shows, the strands are connected by pairs of bases chosen from the four bases,

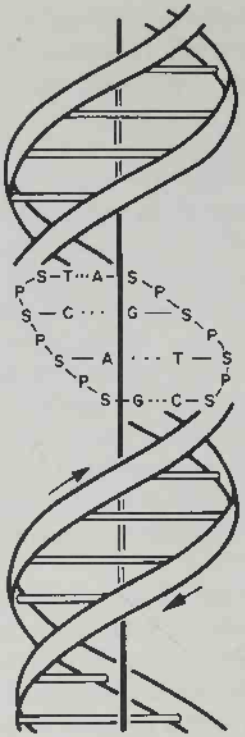


FIGURE 14–2. *Schematic diagram of a short segment of a DNA molecule. The two complementary strands are composed of alternating sugars and phosphates. The strands are joined by bases, either in the combination adenine-thymine or in the combination guanine-cytosine. The vertical axis in this drawing is for orientation only, and does not correspond to any feature of the DNA molecule.*

adenine (A), cytosine (C), guanine (G), and thymine (T). The strands themselves are made of sugars (S) and phosphates (P). Thus, a nucleoside is a combination of a base and a sugar, such as AS, while ASP is an example of a nucleoside phosphate. (A nucleoside phosphate with only one phosphorous group is called a nucleotide.) The sequence of bases—for example, the bases TCAG along the left-hand strand of the inset—specifies the genetic code by determining which proteins the cell will construct. Proteins, in turn, are long chains of amino acids. Recent evidence indicates that three nucleoside phosphates in the nucleic acid are required to specify each amino acid in the protein. The transcription sequence is this: DNA makes RNA; several kinds of RNA together make proteins, in particular, enzymes; and enzymes, by controlling the rates and varieties of chemical reactions in the cell, govern its metabolism. In this way, the nucleic acids actively control the form and functions of all cells.

▽ Exact replication—the production of two identical DNA molecules from one—occurs because only certain combinations of bases can fit across the two strands. During DNA replication, the strands of the helix separate. The bases

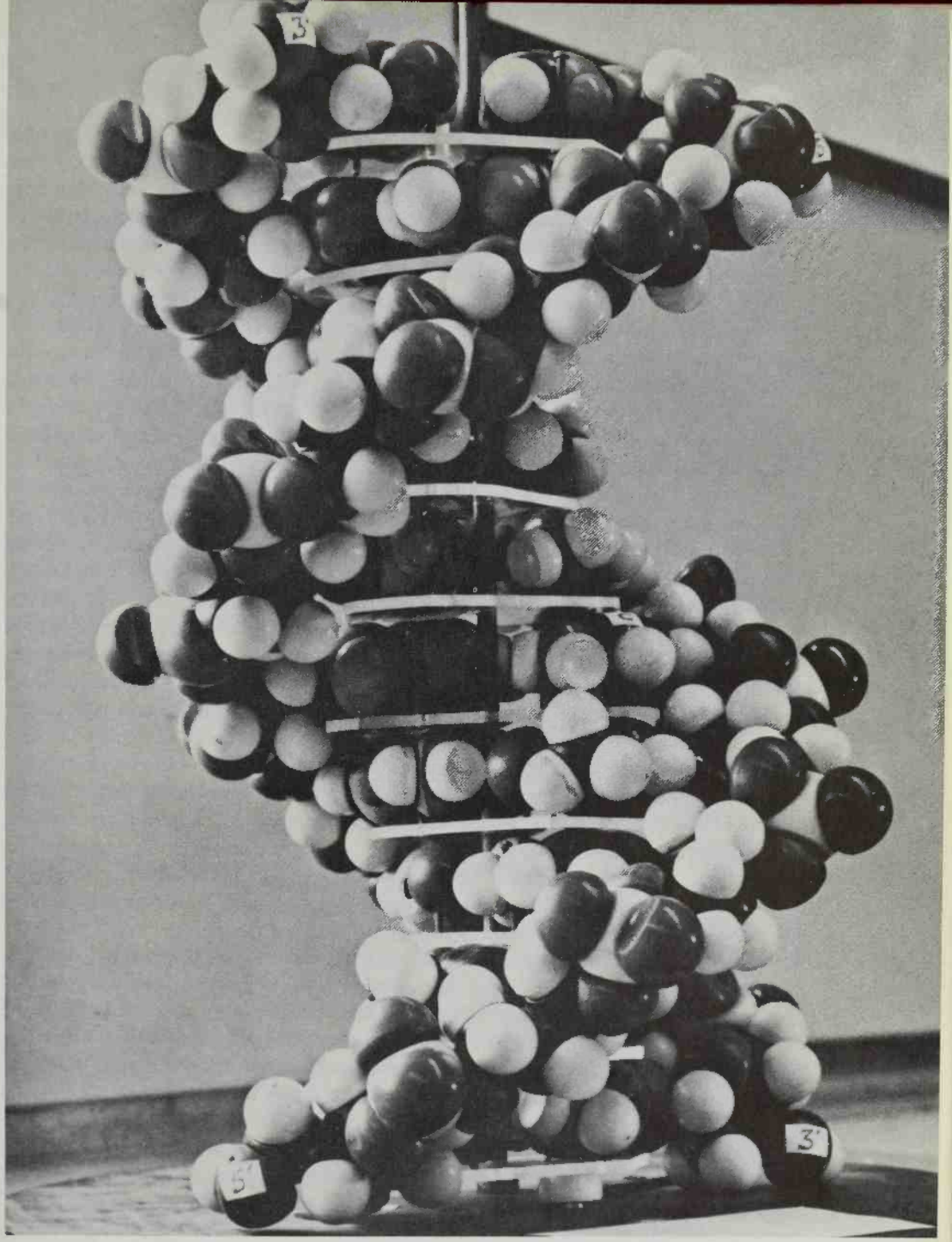


FIGURE 14–3. *A model of a short segment of a DNA molecule, in which atoms are represented by spheres and sections of spheres. Different varieties of atoms are represented by different colors. In Figure 14–2, letters such as S or A stood for molecules containing a dozen or more atoms. In this Figure, each atom is represented directly. Actual DNA molecules may be thousands of times longer than the short segments displayed in Figures 14–2 and 14–3. The complexity of the nucleic acid molecules is evident. (Courtesy of Professor Paul Doty, Department of Biological Sciences, Harvard University.)*

exposed to the cell medium determine which nucleoside phosphates from the medium can combine with the separated strands. For example, suppose an adenine-containing nucleoside phosphate is sitting bound in one strand. A variety of other nucleoside phosphates are available in the medium and occasionally come close enough for chemical bonding to occur. If a guanine-containing nucleoside phosphate is added, DNA synthesis will not proceed, because the guanine-adenine combination will be too large for the space available between the strands. An adenine-cytosine combination will not combine properly, nor will an adenine-adenine combination. Only an adenine-thymine combination fits between the strands. Elsewhere, a thymine-cytosine combination will be too small for the DNA double helix, and not reach between strands. The replication of DNA occurs in large part because thymine (T) will bind only with adenine (A), and guanine (G) only with cytosine (C). Thus, once the sequence of bases along one strand is specified, the sequence along the other strand is determined [see Figure 14–2]. For example, if a section of one strand of a DNA molecule had the sequence of bases T C A G A G T G A C C G A T A T T C, we could immediately decide that the sequence of bases along the other strand must be A G T C T C A C T G G C T A T A A G.

▽ The replication of nucleic acids is usually identical, but not always. Under the influence of external factors such as radiation, or by purely random molecular motions, changes in the structure of the nucleic acids may occur. For example, one base in a strand may be deleted, or may be substituted by another base; or the sequence of bases may be inverted in a short linear sequence. Since the sequence of bases is now changed, the proteins which this nucleic acid molecule now constructs will be different. In general, the changed sequence of bases will yield a "nonsense" portion of the protein it has coded; that is, the altered portion of the protein will serve no useful function.

▽ Some aspects of evolution and the genetic code can, perhaps, be illustrated by a parable. Once upon a time there was an ancient and stable empire, whose capital was connected with its many provinces by a system of royal roads. Day-to-day government in the provinces was the responsibility of the satraps, usually imperial appointees of local origin. The satraps excelled in executing the imperial commands; but their personal initiative was limited. They had a certain repertoire of responses to crisis situations; these responses survived repeated testing in previous generations, and generally functioned well. From time to time, a message from the emperor would arrive concerning some major issue, such as next year's agricultural quotas, or the retraining of industrial workers. These messages were obeyed instantly and literally; their wisdom was legendary. But in those days, the imperial messages were invariably phrased in a foreign tongue. The parlance of the capital was not the dialect of the provinces. The messenger, billeted at the imperial court, knew only the language of the capital. The burden of translation therefore fell to the satrap, who maintained for this purpose a group of bilingual viziers.

▽ The emperor, a member of a venerable and illustrious family, was permanently sequestered in the remote and inaccessible capital, insulated from the

stresses of provincial life. The emperor was a conservative, in the best and worst senses. It was his firm belief that, with very minor changes, the imperial methods and mandates of his predecessor emperors were precisely valid in his day. Therefore, it was his practice to consult the ancient almanac and read the hoary mandates of his imperial forebears. At propitious moments, identical mandates were dispatched to the provinces. The times were stable, and external threats were few. The empire prospered.

▽ Yet there was a certain anxiety which the emperor's courtiers shared. It concerned a dark secret which had, from time to time, concerned imperial households through ages past, back to the founding of that august family: the emperor was afflicted with a hereditary mumble. Not all the time, you understand; mostly, the emperor was lucid and assured, his commands confidently transformed into provincial action. But on rare occasions, the mumble would come upon him. At these times, he could not be understood at all. The emperor would say, "Mumble, mumble!" The courtiers would repeat, "Mumble, mumble," and nod their heads sagely. The imperial courier, his message in hand, would mount his steed and race along the royal highway to the provinces. "Mumble, mumble!" he would say to the satraps. The satraps would turn to their translators. "Mumble, mumble," they translated, in the dialect of the provinces. "Mumble, mumble," the satrap would say to the workers and soldiers, his duty discharged. But the workers and soldiers knew nothing of "mumble, mumble." They continued to await an intelligible imperial command. And soon a lucid message was on its way with another courier. All would be put right, and only a little time lost.

▽ But the bad moments, the courtiers knew, were when the emperor's mumble *seemed* lucid. Oh, sometimes he would say, "Doubled be must quota potato the," and the workers might either reconstruct his meaning, or be idle. However, he did other things. He might insert a "not" into a decree, where none was intended, or substitute one noun for another. This invariably led to disaster. No courtier, no courier, no satrap ever questioned the imperial command. The word of an absolute autocrat is law, as the case of Lieutenant Kijé, the subject of a composition by Prokofiev, attests. So occasionally there was serious trouble in the provinces.

▽ One day it came to pass that a desperate crisis of external origin arose in the provinces. Its nature need not concern us here; the emperor never heard of it. He was very quick with mandates for the provinces and very loath to receive news from the provinces. The crisis was beyond the capabilities of the satraps; it was not encompassed in their repertoire of responses. At this moment, by chance—for the emperor knew nothing of the crisis—an imperial mumble was dispatched to the provinces. It was of the apparently lucid variety, and led to activities by the workers. Fortuitously, miraculously, the misapprehended mumble solved the crisis. Of all the possible mumbles—and of these there were many—the emperor had spoken by chance the right mumble at the right moment. The empire was saved.

▽ This is no way to run an empire, you are saying. Yet, in a sense, this is how living systems function. Very crudely—for the analogy is inexact—the capital is the nucleus, and the provinces the cytoplasm of a cell. The cloistered emperor with

his almanac is the nuclear DNA; the courier and his message, RNA, coded in the nucleus and passing into the cytoplasm. The satraps and their viziers are the ribosomes and adapter RNA, which serve as a kind of molecular scaffolding, organizing cytoplasmic amino acids into the sequence specified by the messenger RNA. Translation is necessary, because the messenger RNA carries in the base sequence of its nucleotides the information on the amino acid sequence of the proteins to be constructed. The workers and soldiers are the enzymes. The feedback from cytoplasm to nucleus is apparently insignificant. An accidental disruption of the DNA base sequence almost invariably has a deleterious effect on the functioning of the assembled proteins. But very rarely, a mutation has a salutary effect. Biological evolution is based upon the fortuitous emergence of such random beneficial mutations. Clearly for each organism which is better adapted because of a beneficial mutation there are millions which perish because of a deleterious mutation. Natural selection works only because (1) enormous numbers of organisms are involved, and (2) the beneficial mutations are preferentially reproduced. But evolution by hereditary mumble is a slow process.

▽ From breeding experiments, primarily with the common fruit fly, *Drosophila melanogaster* (literally, the "black-bellied dew-lover"), it was found by Thomas Hunt Morgan and his students at Columbia University, in the second decade of the twentieth century, that the hereditary characteristics, or genes, were arranged in linear order along the chromosomes, with each gene controlling one or more traits of the organism. The common organisms all have more than one chromosome, and in sexual reproduction the chromosomes of the parents are randomly reassorted, thereby giving the offspring a chance for a previously untried physical constitution. The number of possible reassortments is so large that it provides a natural explanation for the fact that, except for identical twins, no two individuals are alike.

▽ *D. melanogaster* and other insects have, fortunately for us, a set of giant chromosomes in their salivary glands. These chromosomes [see Figure 14–4] are naturally banded, and the bands have a 1-to-1 correlation with the genes deduced from breeding studies. Since all cells in the fly arise from the same fertilized egg, we expect the chromosomes of the salivary glands to be structurally identical with the chromosomes of the reproductive cells. When a gene is absent, a band is absent; when it appears doubled, we see a doubled band, etc. This has provided a strong observational confirmation of the fact that the genes are strung in linear order along the chromosome, and that each gene does control at least one hereditary characteristic.

▽ Some salivary gland chromosomes show an occasional enlarged, bulbous region [see Figure 14–4]. It now appears that these puffs are the sites of active genes, where the genetic material is coding a particular sequence of nucleoside phosphates in messenger RNA. The puffs are found to be associated with high concentrations of RNA. The messenger RNA then presumably detaches itself from the DNA of the chromosome puff and migrates to the cytoplasm of the cell, where it directs protein synthesis. While such conclusions have been deduced for only a

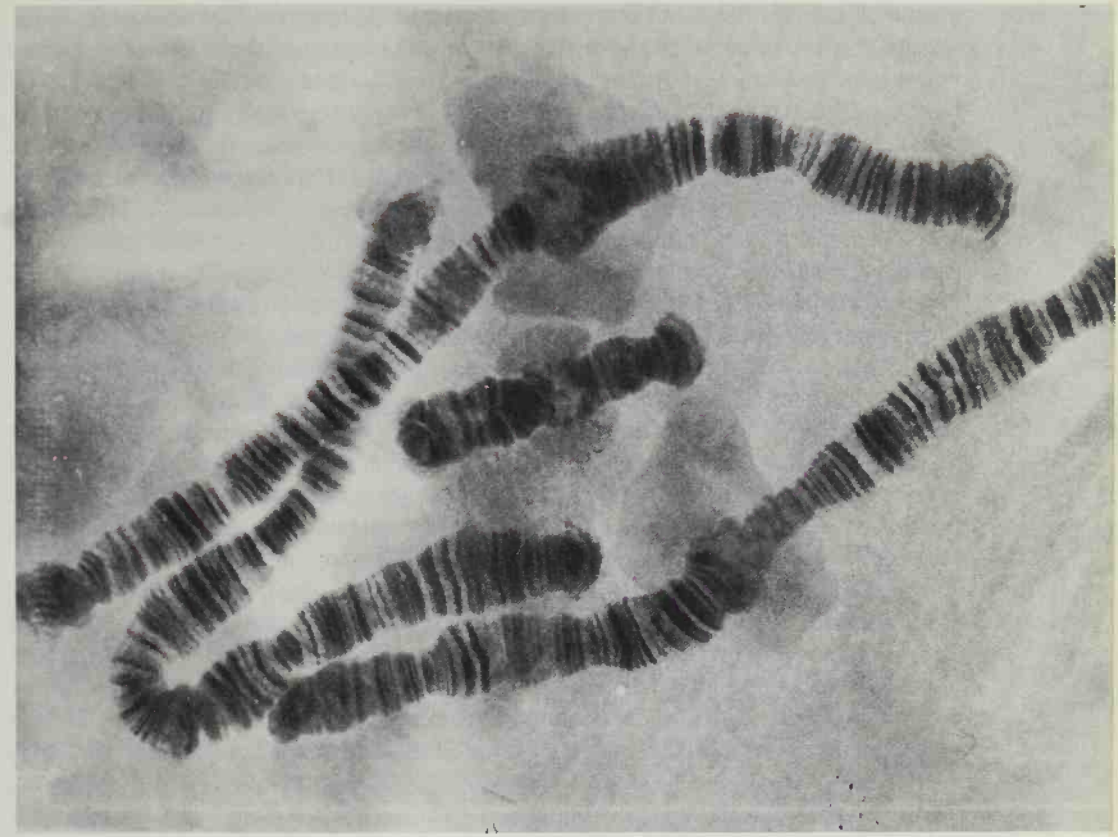


FIGURE 14–4. *A photograph, magnified some 700 times, of the four chromosomes in the salivary gland of the midge* Chironoumus. *The sequence of bands corresponds closely to the sequence of genes in the genetic material of the organism. The enlarged regions may be sites of active production of messenger RNA. (Courtesy of Professor W. Beermann, Max Planck Institut für Biologie, Tübingen.)*

small variety of organisms, there is every reason to believe that the same process occurs for all the organisms on the Earth.

▽ While there is a very vague correlation between the number of chromosomes and our estimate of the evolutionary complexity of the organism (for example, *Drosophila* has four chromosome pairs; human beings have 23), this correlation is far from a general rule. The particular plants which contain the largest number of chromosomes per organism are far from being the dominant species on the planet. What counts is not the number of chromosomes or the amount of genetic material, but rather its information content.

▽ A typical *Drosophila melanogaster* chromosome may be about 1 micron ($= 10^{-4}$ cm) long. From *Drosophila* breeding experiments, it has been possible to determine, without ever looking at its chemistry, that a given chromosome of *D. melanogaster* contains at least 1000 genes. If we divide one micron into a thousand

equal parts, we find that each gene is about 10 Å long (1 Å = 10^{-8} cm). Since there may be many genes which control traits not readily discernible to the geneticist, we would expect the size of the gene to be somewhat less than 10 Å.

▽ The size of a gene so deduced from breeding studies is confirmed in the DNA structure, where we find that the distance between adjacent nucleoside phosphates along the same polynucleotide strand is 3.4 Å. A substitution in one of these nucleoside phosphates changes the meaning of the triplet code of which it is a part and will thereby alter one amino acid in the protein which this nucleic acid codes. For example, some hereditary diseases in human beings are known to result from just such a substitution of one amino acid. Therefore, the smallest genetically significant unit of the hereditary material is about 3.4 Å long.

▽ Is it really possible that the sequence of nucleoside phosphates along the DNA chain can contain enough information to construct an entire organism? A human being? The mass of DNA in one set of human chromosomes is about 4×10^{-12} grams. One pair of nucleoside phosphates (on opposite strands of the double helix) has a mass of the order of 10^{-21} grams. Therefore, there seem to be some 4×10^{9} pairs of nucleoside phosphates per chromosome set, according to this calculation by the American geneticist, H. J. Muller, of Indiana University.

▽ In each of the 4×10^{9} positions available for a nucleoside phosphate, only four combinations of nucleoside phosphates are possible. On a given strand, the base may be T, C, A, or G. The base on the complementary strand must then be A, G, T, or C, respectively. The number of possibilities for one slot along the DNA molecule is four. The number of possible combinations for two consecutive slots is $4 \times 4 = 16$; for three slots, $4^3 = 64$, etc. Thus, for 4×10^{9} slots, there are $4^{4\times10^9}$, or $10^{2.4\times10^{10}}$ possible varieties of human chromosomes. This is, of courses, an enormous number vastly larger than the number of elementary particles in the detectable universe, which as we saw in Chapter 11, is about 10^{80}. It is vastly larger, also, than a googol; but much smaller than a googolplex.

▽ This number is a measure of our improbability. If you threw 4×10^{9} nucleoside phosphate pairs up in the air, and they came down by twos in random order, there would be only one chance in $10^{2.4\times10^9}$ of reassembling one of your chromosomes. You could perform this imaginary exercise of reassembling DNA molecules randomly, once a second, for the lifetime of the Galaxy, and come nowhere near constructing one of your own or anyone else's chromosomes.

▽ But if our DNA is *this* improbable, how could it have come into being at all? Our improbability is extracted from the environment by natural selection. Our nucleic acids are *not* constructed randomly. The vast majority of base combinations has never been tried. Each combination is built upon pre-existing ones. The sequences of nucleoside phosphates which work are carried over from generation to generation, unchanged for millions of years. In fact, the similarity of base sequences between the DNAs of two different organisms can now be used as an index of evolutionary kinship. In this way, it has been possible to show, for those who have wondered, that man and monkey are more closely related than man and mouse.

▽ The information contained in a single human sperm cell is equivalent to that of 133 volumes, each of the size and fineness of print of *Webster's Unabridged Dictionary*. Yet we now can understand how this information could have come into being through natural selection. We understand how natural selection can extract order from chaos, if there are self-replicating and mutating systems in a non-static environment. But we are again faced with the question of the origin of the first such system.

▽ The origin of life on Earth seems intimately tied to the prebiological origin of proteins and nucleic acids. We do not know that proteins and nucleic acids must be intimately involved in living systems on other planets, although some evidence supporting this view will be presented in Chapter 18. But if we are to design for the detection of extraterrestrial life equipment which is not hopelessly parochial in outlook, we must have some general approach to living systems. △ We may well encounter phenomena on other planets which, while possessing all the essential attributes of life on Earth—perhaps even intelligence—exist in different forms, and function according to different principles. It would be useful to have a purely operational definition of life which is not confined to familiar terrestrial chemistry. ▽ We conclude the chapter by returning to this question. △

Some interesting preliminary ideas on the subject have been formulated, in terms of cybernetics, by the Soviet mathematician A. A. Liapunov. Cybernetics, ▽ a term coined by the American mathematician Norbert Weiner, △ is concerned with the study of control processes and the construction of control systems. ▽ Cybernetics developed at the same time that the first large electronic computing machines were constructed. △ Liapunov believes that control, in its broadest sense, is the most universal property of life, independent of form.

▽ Because of the necessity for evolution by natural selection in order to develop living systems of any complexity, one possibly useful definition of life is this: a living system is any self-reproducing and mutating system which reproduces its mutations, and which exercises some degree of environmental control. This definition is much narrower than that of Liapunov. △

Another definition of life, supported by the Soviet biochemist A. I. Oparin of the A. N. Bach Institute, is expressed in terms of complexity and a highly regulated metabolic system for material exchange with the environment. Metabolism, of course, is an essential attribute of life. But in the context of origins, does life lead to metabolism, or does metabolism lead to life? No entirely satisfactory answer has been found to date. It should be noted that simple forms of material interchange (which are not highly regulated, and thus not metabolic processes) can be observed in non-living systems—for example, in liquid solutions.

In Liapunov's view, living systems have the following special characteristic: through definitely prescribed channels, the transmission of small quantities of energy or material containing a large volume of information is responsible for the subsequent control of vast amounts of energy and materials. (An obvious example is the control by the genetic material in man of the form, development, and chemical processes of the much larger individual.) Liapunov points out that

heredity, irritability, and so forth can be described in cybernetic terms as information storage, feedback, communication channeling, etc.

All biological materials are dependent upon their mass, chemical composition, energy state, electric and magnetic properties, and so forth. Generally speaking, these properties will change over a period of time. But a small fraction of the materials will remain relatively stable. These substances maintain their stability despite changes which occur in the external environment. Liapunov calls such reactions, in which the substance survives changes in the external environment, *maintaining reactions*. Maintaining reactions underlie all biological processes. Indeed, life is characterized by its adaptation to the external environment.

In the language of cybernetics, maintaining reactions can be outlined as follows: the sensing material receives information about the external environment in the form of coded signals. This information is reprocessed and sent in the form of new signals through defined channels, or networks. This new information brings about an internal reorganization of the system which contributes to the preservation of its integrity. The mechanism which reprocesses the information is called the control system. It consists of a vast number of input and output elements, connected by channels through which the signals are transmitted. The information can be stored in a recall or memory system, which may consist of separate elements, each of which can be in one of several stable states. The particular state of the element varies, under the influence of the input signals. When a number of such elements are in certain specified states, information is, in effect, recorded in the form of a text of finite length, using an alphabet with a finite number of characters. ▽ These processes underlie contemporary electronic computing machines and are, in a number of respects, strongly analogous to biological memory systems. △

The control system directs the maintaining reactions of the organism or computing machine and their response to the external environment. The response occurs by collecting information about the external stimuli, analyzing it into its component parts, and comparing it with information which is already recorded in the memory. The greater the amount of information previously stored, the more adaptable the control system. An important property of the maintaining reaction is its speed of response. If the reactions are slow, the survival of the system is jeopardized. Thus, a large information storage capacity is required for the memory bank, and the information must be stored in an exact and stable manner.

Liapunov suggests that individual molecules, consisting of sufficiently large numbers of atoms, can conceivably act as stable physical information carriers. Such molecules are quantum systems. In order to reach another information state, they must be raised to another energy level, sufficiently distant from the original state that few transitions will occur due to random thermal motion.

The energy supply for the maintaining reactions must not run down. Nevertheless, such systems will constantly lose heat and energy because of their activity. According to the laws of thermodynamics, the energy levels in a closed system—one completely isolated from its environment—must eventually reach

equilibrium. If a living system were closed, any loss of energy would endanger its stability. Thus, a stable state will not be maintained unless energy is obtained from the external environment, and the living system becomes an open system.

An important thermodynamic characteristic of any such system is its entropy. Entropy may be defined as a measure of the unavailability of energy in a thermodynamic system; ▽ or, alternatively, as a measure of the disorder of a system. In any closed system, no process can occur in which the entropy decreases; that is, the disorder of any closed system will increase as time goes on. In an infinite amount of time, disorder should be complete, and the atoms randomly distributed, in the absence of other influences. If a living system is represented as a closed system, its entropy would continuously increase. The burgeoning disorder would, in time, bring all biological processes to a halt. Consequently, a living organism must systematically lose entropy, to maintain internal order. This is possible only at the expense of the external environment. The organism must extract energy from, and increase the entropy of, the external environment, so that its own entropy may be continuously decreased, and its structural and functional integrity maintained. As we pointed out earlier, this is one reason why cells metabolize. △

Older definitions of life, which identified life with metabolism, were inadequate. Such definitions, in our opinion, are entirely worthless. Liapunov characterizes life as a highly stable material system which uses information coded by molecular states for the production of maintaining reactions.

The actual organization of living systems into subcellular organelles, cells, organs, organisms, populations, species, and so forth is analogous to a hierarchy of control systems. Each structural unit is controlled by its own semiautonomous control system, which acts upon all those units subordinate to it, and in turn is acted upon by those control systems which are above it in the hierarchy.

There is a distinction between the control systems within an individual organism and those acting upon an ensemble of organisms (for example, populations, species, and so forth). In the former case, the control system consists of units acting directly down through the hierarchy. Liapunov calls this the structural method of control. In the latter case, we have a large number of more or less independent statistically equal systems which interact by chance meetings. Liapunov calls this the statistical method of control. Those systems of higher rank order—for example, the species—are significantly more stable than any given individual constituent (here, an individual organism). But this greater stability of the higher system is possible only if the constituent parts are replaceable; that is, if reproduction occurs.

For the newly synthesized constituent part to partake of its measure of stability, it must contain a pre-formed supply of information, stored in its memory bank, which guarantees its maintaining reactions. It is quite inconceivable that this information supply could arise spontaneously within the constituent itself. Thus, a new constituent must obtain this store of information necessary for its functions

from other constituents—most reasonably from other similar constituents which we may refer to as the previous generation. Thus, reproduction is seen to be in large measure information replication.

The transmission of information from generation to generation occurs in a background of interference which can partially alter its character. ▽ If such an alteration of the hereditary information cache is itself replicated identically—that is, if the altered information is transmitted to succeeding generations—then such an alteration can be called a "mutation." △ Such mutations change the control system, modifying the maintaining reactions, and thereby changing the character of the interaction of the system with its environment; they can radically alter the efficiency with which a given individual copes with its environment.

▽ It is possible therefore to describe living systems from a cybernetic point of view. At the moment, this is perhaps no more than a useful analogy. It has provided an insight, but not yet any new information. △ It is possible that in the future a synthesis of such a cybernetic approach with molecular biology will lead to a complete understanding of the nature of life, an understanding which we do not yet have, as Liapunov himself is fully aware. These ideas, and the related viewpoint of the Soviet physicist Kolmogorov—discussed in Chapter 35—may ultimately prove to be of great significance in the analysis of the problem of the origin of life on Earth and the probable widespread distribution of life in the universe.

15

The origin of life: Historical views and panspermia

But now to carry the Search farther, let us see by what Steps we must rise to the attaining some knowledge in the deeper Secrets concerning the State and Furniture of these new Earths. And, first, how likely is it that they may be stock'd with Plants and Animals as well as we? I suppose no Body will deny but that there's somewhat more of Contrivance, somewhat more wonderful in the Production and Growth of Plants and Animals, than in Lifeless Heaps of inanimate Bodies, be they never so much larger; as Mountains, Rocks, or Seas are. For the Finger of God, and the Wisdom of Divine Providence, is in them much more clearly manifested than in the other. One of Democritus's or [Des] Carte's Scholars may venture perhaps to give some tolerable Explication of the Appearances in Heaven and Earth, allow him but his Atoms and Motion; but when he comes to Plants and Animals, he'll find himself non-plus'd, and give you no likely account of their Production. For every Thing in them is so exactly adapted to some Design, every part of them so fitted to its proper Use, that they manifest an Infinite Wisdom, and exquisite Knowledge in the Laws of Nature and Geometry, as, to omit those Wonders in Generation, we shall by and by show; and make it an Absurdity even to think of their being thus happily jumbled together by a chance Motion of I don't know what little Particles.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

[May one] doubt whether, in cheese and timber, worms are generated, or, if beetles and wasps, in cow-dung, or if butterflies, locusts, shellfish, snails, eels, and such life be procreated of putrefied matter, which is to receive the form of that creature to which it is by formative power disposed[?] To question this is to question reason, sense, and experience. If he doubts this, let him go to Egypt, and there he will find the fields swarming with mice begot of the mud of the Nylus, to the great calamity of the inhabitants.

A seventeenth century opinion quoted by L. L. Woodruff, *The Evolution of Earth and Man* (1929)

Nothing seems now more contrary to reason, than that chance and nastiness should give a being to uniformity, regularity, and beauty . . . and create living animals . . . This, however, was the opinion not only of the ignorant and illiterate, but of the most learned grave philosophers of preceding ages; and would probably still have been taught and believed had not microscopes discovered the manner how all these things are generated . . .

Henry Baker, *The Microscope Made Easy* (1742)

▽ In an earlier and simpler age, life was believed to arise spontaneously, from nothing. It was a commonplace observation. Not quite from nothing, perhaps; but mice from the mud of the Nile, maggots from putrefying meat, lice from sweat, and fireflies from conflagrations—as the most elementary observation seemed to show. The question of the origin of life was trivial—life was arising all the time, at least for lower animals.

▽ Since higher animals arose from the reproduction of their own kind, the question of *their* ultimate origin was more difficult. The dominant view, found in Genesis 1, in the Hesiodic *Theogony,* and in the Sumerian creation myths, invokes a separate creation of each species by divine fiat. There were early murmurings against these beliefs, for if lower organisms arise spontaneously, might not the earliest self-reproducing higher organisms have developed in some way from simpler predecessors? The pre-Socratic philosopher Anaximander postulated that life arose in the sea, and that man developed from something like a fish. A view similar to Darwinian natural selection was expressed by Empedocles. Is it possible that higher organisms, so marvelously adapted to their environments, arose through some natural process from simpler organisms? In Book II of his *Physics,* Aristotle restated Empedocles' opinion in the following characteristic phrases:

> Wherever, then, all the parts came to be just what they would have been if they had come to be for an end, such things survived, being organized spontaneously in a fitting way; whereas those which grew otherwise perished, and continued to perish . . .

This passage is quoted by Darwin on the first page of *The Origin of Species,* apparently without his being aware that Aristotle then went on to criticize the Empedoclean hypothesis:

> Yet it is impossible that this should be the true view, for . . . all . . . natural things either invariably or normally come about in a given way: but of not one of the results of chance or spontaneity is this true. . . .

What is bothering Aristotle? He seems to be saying that a random but useful change in the characteristics of an organism cannot be maintained, because there is no way for the characteristic to become established among many organisms. The children of one-armed men are generally born with two arms. The possibility is overlooked that random variations in the genetic material may be reproduced, and adaptive characteristics thereby become established in an entire population.

▽ After the time of Lucretius, who echoed Empedocles' views, the combined authority of Aristotle and the medieval Church was so great that spontaneous origins of lower animals were accepted as tenets of faith. But when the Italian

Renaissance flowered across Europe, the confidence in hoary antiquity and traditional explanations became eroded: "Anyone who in discussion relies upon authority uses, not his understanding, but rather his memory . . . ," wrote Leonardo da Vinci. Experimental verification of hypothesis became widely accepted. Thus, in 1665, an Italian physician, Francesco Redi, put the hypothesis of spontaneous generation to an experimental test. When putrefying meat is covered with fine gauze, Redi found, maggots never develop. He discovered that maggots were the larval forms of flies, which deposited their eggs on the meat. When the meat was covered by gauze, the flies were unable to lay their eggs, and no maggots developed.

▽ About a decade after Redi was disproving spontaneous generation at the level of the house fly, a Dutchman, Antony van Leeuwenhoek, was discovering microorganisms, and thereby, through no fault of his own, extending the debate on spontaneous generation for another two centuries. Leeuwenhoek found that apparently pure water, but especially water which contained organic impurities such as hay infusions, abounded with microorganisms. His charming account of these discoveries follows:

> On April 24th, 1676, observing this water by chance, I saw therein with great wonder unbelievably very many small animalcules of various sorts; among others, some that were three to four times as long as broad. Their entire thickness was, in my judgment, not much thicker than one of the little hairs that cover the body of a louse. These creatures had very short, thin legs in front of the head (although I can recognize no head, I speak of the head for the reason that this part always went forward during movement). . . Close to the hindmost part lay a clear globule; and I judged that the very hindmost part was slightly cleft. These animalcules are very cute while moving about, oftentimes tumbling all over.

▽ From where did these "animalcules" come? Leeuwenhoek himself believed that tiny seeds, or germs, of the animalcules were present everywhere, and, upon gaining access to nutrient media such as hay infusions, proceeded to grow. Since the germs can arise from the microorganisms themselves, there is no necessity to invoke spontaneous generation. Yet many learned men were unable to accept the extraneous origin of microorganisms, especially when a variety of experiments seemed to show that organic solutions, whether covered or uncovered, boiled or not boiled, always developed "animalcules." It was not until 1861, two years after the publication of Darwin's *The Origin of Species,* that this problem was finally put to rest. Louis Pasteur, in his *Memoir on the Organized Bodies which exist in the Atmosphere,* demonstrated rigorously that the air does contain "germs," as Leeuwenhoek thought; that introduction of these germs into a sterile medium invariably leads to the appearance of microorganisms; and that sterile organic media exposed to air, but not to germs, never develop microbial cultures.

▽ (It is a curious fact that shortly after utilizing sterile techniques in solving one problem in the origin of life, Pasteur applied them to an experimental question on extraterrestrial life. In 1864 a large meteorite of a type now known as a carbonaceous chondrite fell near Orgeuil, France. Pasteur caused a special drill to

be constructed, which, he hoped, would remove samples from the interior of the meteorite without contaminating them with microorganisms from outside. Using sterile techniques, Pasteur inoculated an organic medium to search for growth of any indigenous microorganisms which the meteorite interior might contain. The results were negative, and have relevance today: Pasteur extracted his sample shortly after the fall of the meteorite, and was, of course, a very careful experimentalist. In Chapter 23, we will return to more recent studies of possible living forms in the Orgeuil and other meteorites.)

▽ Thus, by the 1860's, it was no longer possible to hold that contemporary organisms, no matter how simple, spontaneously arise from non-living precursors. By this time, Darwin had provided an intellectual framework in which the development of complex organisms from simpler ones by natural selection could be understood. Yet the problem of the origin of the first organism remained. No one felt the difficulty of this problem more keenly than Darwin himself: "It is mere rubbish, thinking at present of the origin of life," Darwin wrote in a letter to Hooker in 1863. "One might as well think of the origin of matter." In this, Darwin was correct. As we have seen in Chapters 7 and 8, we *are* today thinking of the origin of matter, with some success; and serious scientific studies of the origin of matter and of the origin of life have occurred contemporaneously. But at the end of the nineteenth century there was no experimental approach to the origin of a living organism from inanimate matter. The problem seemed impossibly difficult.

▽ In this intellectual climate, the Swedish chemist Svante Arrhenius in 1907 proposed the panspermia hypothesis. Arrhenius suggested that terrestrial life did not originate on Earth. He imagined that simple living forms may have drifted from world to world, propelled by radiation pressure through interstellar space. An extraterrestrial origin of life at least postponed the difficulties inherent in the origin of life, much as studies of stellar nucleogenesis ignore the problem of the origin of hydrogen. There may be no problem of the ultimate origin of matter; the universe may be infinitely old, and may always have had matter in it. By the same token, if panspermia is even a relatively inefficient method of populating a planet, in sufficient time the descendants of one organism might populate a static universe. △

Let us first consider the philosophical tractability of the panspermia hypothesis. What objection, in principle, could there be to spores making this magnificent cosmic journey from planet to planet, and from star system to star system, and then, by chance falling on a planet where conditions were suitable, reviving and initiating life? The idea is not inconsistent with materialist philosophy. Indeed, is it *necessary* to assume that life on Earth should arise locally, from non-living matter? And starting from the assumption that there are a multiplicity of populated worlds, is it not completely logical to investigate the possibility that organisms are exchanged between planets? Only by an interdisciplinary approach, through astronomy, biology, and allied sciences, is it possible either to confirm or forever lay to rest the panspermia hypothesis.

Sagan recently attempted a careful analysis of this problem. ▽ Arrhenius supposed that terrestrial microorganisms were sometimes wafted into the strato-

sphere by winds in the terrestrial atmosphere. There is some evidence, from balloon studies, that microorganisms can be found at great heights, well into the stratosphere. Arrhenius postulated that occasionally such organisms will be entirely ejected from the Earth by electrical forces. Such a mechanism works in principle, but in practice we do not know with what efficiency microorganisms are ejected—if they are ejected at all. This is one motivation for obtaining high-altitude microbiological profiles of the terrestrial atmosphere and exosphere.

▽ Let us assume, with Arrhenius, that such electrostatic ejection of microorganisms from the upper terrestrial atmosphere occasionally occurs. What will be the fate of such a microorganism? For notational convenience, let us call this microorganism a "bug," in full awareness that it will be much smaller than any ordinary insect usually called a bug. The fate of an ejected bug depends upon the ratio p/g, where p is the magnitude of the force due to radiation pressure, which tends to drive the microorganism away from the sun; and g is the magnitude of the gravitational force due to the Sun, which tends to drag the bug into the Sun. In the absence of other forces, if $p/g = 1$, the bug just sits in interplanetary space; if p/g is less than 1, it falls into the Sun; and if p/g is greater than 1, it leaves the solar system. Since p and g are both inversely proportional to the square of the distance, r, of the microorganism from the Sun, p/g is independent of r. But the value of the *net* force, $p - g$, varies inversely as the square of r.

▽ For a model bug, something like terrestrial microorganisms, only a narrow size range has p/g greater than 1 and can escape. These bugs must be approximately 0.2μ to 0.6μ in radius, assuming them spherical; their diameter must therefore be comparable to the wavelength of visible light [1 micron (μ) = 10^4 Å = 10^{-4} cm]. Any bugs seeding the Earth—to initiate life in primitive times, for example—would have to be outside this size range. Bugs leaving the solar system would have to fall within this size range. A characteristic dimension for an ordinary terrestrial microorganism is several tens of microns; but bacterial and fungal spores and many viruses have dimensions between 0.2 and 0.6μ.

▽ Since the force due to radiation pressure continues to act as an organism recedes from the sun, its velocity continues to increase, and it soon reaches very considerable speeds Thus, a bug in this size range starting in the vicinity of the Earth would reach the orbit of Mars in weeks, the orbit of Jupiter in months, the orbit of Neptune in years, and the distance to the nearest star in a few tens of thousands of years. If the bug made no collisions in its interstellar peregrination, it could transit the Galaxy in a few hundred million years. But Shklovskii has pointed out that over these distances, the bug will almost certainly not move in a straight line. △ The bug would tend to move in the same manner as a particle of interstellar dust (which it resembles in size, mass, and composition), that is, in an irregular, random manner. Having traveled the distance of several tens of light years, it may suddenly change direction upon collision, or even merge with another interstellar dust particle. Thus, the bug would tend to make a random "walk" through the Galaxy, similar to the Brownian motion of small particles in solution. ▽ Because of the resulting erratic path, which crosses back upon itself and retraces

steps already taken, the transit time of a bug between any two places in the Galaxy is correspondingly much longer than if the bug traveled in a straight line and made no collisions. △ In order to traverse a distance of about 1000 light years (approximately $\frac{1}{30}$ our distance from the Galactic center), the bugs would require several hundred million years. To cross the entire Galaxy, they would require 10^{11} years, a time interval some 10 times greater than the estimated age of the Galaxy. ▽ Thus, if the Earth were seeded several billions of years ago—as would be required, to match the evolutionary timescale—that initial bug must have been ejected from a star no more than about 6000 light years away, *provided* that deflecting particles of interstellar dust were as common then as now. However, since, as we have seen in Chapter 5, the interstellar dust particles arise by collisions among themselves and with the interstellar gas, it is possible that some billions of years ago the interstellar dust density was much less than it is today; the Earth could then have been seeded by a bug arising on a planetary system more distant than 6000 light years from the Earth.

▽ But such discussions of transit times neglect a very serious question: Do the bugs survive the environmental hazards of the trip? First, the microorganism would be at a very low temperature and high vacuum for most of the trip. It was known even in Arrhenius' day that at least some spores can be immersed in liquid air (temperature −196°C) for long periods of time, without its affecting their ability to germinate subsequently; and we know today that some microorganisms survive extended laboratory exposures to high vacuum. In such experiments, the vacuums do not nearly approach those found in interstellar space, where the density of atoms is about 1 atom cm^{-3}; nor, for obvious reasons, do they approach the 10^9 or 10^{10} year transit times of which we have been speaking. Although a slow boiling-away of molecules that comprise the bug might occur over such immense journeys, let us assume, for the purpose of argument, that the bugs can survive tolerably well the high vacuum and low temperatures on interplanetary and interstellar space. △

Another hazard for wandering panspermia is the H II regions of hot, ionized interstellar gas surrounding early type stars. These regions encompass hundreds of light years and are extremely hot. ▽ But there is some question whether the densities of H II regions are great enough for the temperatures to affect the bugs. △

What about radiation? The bugs are exposed, among other varieties, to solar ultraviolet radiation and to cosmic rays. ▽ If we assume the bug to have the radiation sensitivity of the most resistant known microorganism, solar ultraviolet radiation at wavelengths short of 3000 Å would kill the putative panspermia at the moment of their departure—within a day of their ejection from Earth into interplanetary space. In the extremely unlikely case that the ejected microorganism has an infinite tolerance to ultraviolet radiation, then x-rays and protons of solar origin would kill the bugs before the orbit of Neptune is reached.

▽ We should emphasize that, at least for ejection from this solar system, the radiation hazards cannot be avoided by providing a protective shielding for the bug. With a shielding thick enough to be useful for radiation protection, the bug would be too large to be ejected by solar radiation pressure. Similarly, we cannot

save the panspermia hypothesis by imagining interstitial spores locked within the fissures of some interplanetary dust particles or meteors and thereby shielded from the harmful radiation.

▽ The same arguments apply for an unprotected spore smaller than 0.2μ entering our solar system, instead of leaving it. It would accumulate a lethal dose of radiation while entering the solar system. But bugs ejected from planets at great distances from a star—for example, at the position of Uranus or Neptune in our own solar system—would run negligible radiation risks. Therefore, the possibility of ejection from or arrival on such worlds cannot be dismissed on grounds of radiation sensitivity.

▽ If panspermia tarry too long, they will be killed by another sort of radiation. We saw in Chapter 7 that the primary cosmic ray flux in the vicinity of the Earth was about 0.04 roentgens (R) per year; in interstellar space, the cosmic ray flux is essentially the same. The most radiation-resistant microorganisms known are not destroyed until they have accumulated doses of 10^6 or 10^7 R. On this basis, it would seem that in $4 \times 10^6/4 \times 10^{-2} = 10^8$ years, the laggard panspermia would be killed by cosmic rays. Actually, the problem is somewhat more complex. When a cosmic ray primary (usually a high energy proton) enters a microorganism, it produces its damaging effect partially by direct ionization of the internal structure of the organism. However, another part of the damage done is by cosmic ray secondaries—less energetic particles that are created in the deceleration of the primary. For a single microorganism floating in space most of the cosmic ray secondaries should emerge harmlessly from the organism into the surrounding space. Because of their small sizes microorganisms should accordingly be more resistant to cosmic rays than larger organisms whose tissues absorb the cosmic ray secondaries. If cosmic rays have always been as intense as they are in the vicinity of the earth today, they may restrict interstellar travel by panspermia over distances more than a few thousand light years, but the exact restriction depends on the largely unknown contribution of secondaries.

▽ Let us now turn to stars other than the Sun. In general, the hotter a star is, the greater is the value of p/g, but also the shorter is its main sequence lifetime—the period during which life can reasonably be expected to develop on the star's planets. Let us assume that the donor planet (the planet ejecting panspermia) was populated for at least a few hundred million years, either for the indigenous origin of life on that planet (in which case the estimate is very generous), or for the proliferation of a spore arriving earlier, from another donor world. We then find that only main sequence stars between spectral types A0 and G5 can eject panspermia. Most of the stars in the Galaxy are cooler than the Sun. Only a few percent lie between spectral types A0 and G5; therefore, only a few percent could hold donor planets for the panspermia hypothesis. For these spectral types, the hotter stars can eject a larger range of sizes of microorganisms; but at the same time, they almost certainly present a much more serious radiation hazard. Stars much cooler than the Sun can eject no panspermia at all. We can conclude that the only reasonable donors are the outer planets of stars in the range of spectral types A0 to

G5; these stars are capable of ejecting microorganisms in a size range between 0.1μ and 3.0μ.

▽ Acceptor planets must clearly be different from donor planets. In our solar system, bugs in the size range between 0.2 and 0.6μ are ejected; accordingly, bugs in this size range can never enter the solar system. The hotter the star, the greater the size range of bugs which are deflected away from that solar system by radiation pressure. Thus, the most likely acceptor planets are those circling cool M dwarfs and the outer planets of G and K stars. The most likely locales in the solar system to search for interstellar panspermia, then, are the moons of the outer planets, especially Triton, the inner satellite of Neptune. Larger panspermia—say, in the 1.0μ size range—might be found elsewhere, if they could survive the radiation of the voyage. It has been suggested by the American geneticist Joshua Lederberg, of Stanford University, that the Moon, which is unlikely to have any indigenous life forms, might be a useful place to search for interstellar panspermia.

▽ Another area of difficulty for the panspermia hypothesis, one which we have not yet touched upon, is geometry. The spaces between the stars are immense. A bug randomly "walking" its way across the Galaxy actually has a very small chance of accidental encounter with a possible acceptor planet. In order for the Earth to have received one microorganism from a stellar source in the first billion years of Earth history, each one of, say, 10^{11} assumed planets in the Galaxy must have ejected about one ton of microorganisms into interstellar space during that period. △ Of course these values could be modified. For example, if there are only 10^8 populated planets, then each of them would have to release 1000 tons of spores every billion years. ▽ Because we do not know the present rate of ejection of microorganisms from planets, especially by electrostatic mechanisms, we cannot assess the plausibility of these values; however, they seem quite high. Studies of the microorganism population of the upper atmosphere of the Earth would be very useful in such considerations; but the over-all prognosis for the panspermia hypothesis is not favorable. The restriction on possible donor and acceptor planets, the radiation hazards during transit, and the geometrical difficulty provide almost insuperable obstacles.

▽ Another sort of planetary seeding should be mentioned, at least in passing. We have so far discussed radiation pressure as a conceivable mechanism for interstellar transport of living things. Thomas Gold of Cornell University has pointed out another possibility. Let us assume, for the moment, that the Galaxy is populated here and there by advanced technical civilizations. Since many such civilizations probably are far in advance of our own, interstellar space flight may have been discovered and exploited [see Chapters 32 and 33]. Suppose, Gold says, that an expedition from such a civilization—a survey party, for example—lands on a previously uninhabited but nevertheless clement planet. Unless the most rigorous precautions are taken, they will contaminate the planet. Their ship, their air, and they themselves are populated with diverse microorganisms. In fact, the prevention of accidental contamination of Mars by the first unmanned space ships destined to land there is a very serious problem [see Chapter 19]. Gold, however, imagines

such planetary contamination in a more vivid way. He pictures the visitors having a picnic on the virgin planet, and leaving their refuse behind. In this view, some microbial resident of a primordial cookie crumb may be the ancestor of us all.

▽ While this garbage theory of the origin of life understandably lacks appeal, we should not exclude it altogether. Perhaps a race of advanced extraterrestrials would be scrupulously careful not to contaminate a previously unpopulated planet; but perhaps not. There is also the complementary possibility that such an advanced civilization may intentionally initiate life on uninhabited planets, for any of a variety of reasons: to prepare the planet for subsequent colonization, with, of course, a very long timescale in mind; to distribute the genetic material of the home planet, so that in case of a disaster, the evolutionary patrimony is not irretrievably lost; or perhaps merely as an experiment in laboratory biology, with a somewhat grander laboratory than those to which we are accustomed. If there is intelligent life in the universe, then it is difficult to exclude such possibilities; but it is also difficult to say very much more about them. If we put away such last resorts—at best, they temporize with the real issues—we must finally come to grips with the problem of an indigenous origin of life. This is the topic of our next two chapters. △

16

The physical setting for the origin of life

Looking back through the prodigious vista of the past, I find no record of the commencement of life, and therefore I am devoid of any means of forming a definite conclusion as to the conditions of its appearance. Belief, in the scientific sense of the word, is a serious matter, and needs strong foundations. To say, therefore, in the admitted absence of evidence, that I have any belief as to the mode in which existing forms of life originated, would be using words in a wrong sense. But expectation is permissible where belief is not; and if it were given to me to look beyond the abyss of geologically recorded time to the still more remote period when the Earth was passing through physical and chemical conditions which it can no more see again than a man can recall his infancy, I should expect to be a witness of the evolution of living protoplasm from not-living matter.

T. H. Huxley, *Biogenesis and Abiogenesis* (1870)

▽ Imagine the solar system viewed from afar: four planets, as the American science writer Isaac Asimov has said, and debris. The four are, of course, Jupiter, Saturn, Uranus, and Neptune. They are large bodies, far from the Sun—easy objects for a small telescope somewhat outside our solar system. The spectra of these outer, or Jovian, planets shows hydrogen, methane, and ammonia in their atmospheres; helium and water are expected. Such spectra are very common in the universe, because of the high cosmic abundance of hydrogen [Chapter 4]. But as we move closer to the sun, some of the debris becomes discernible. The apparent surface features and atmospheric composition of Venus, Earth, and Mars are soon detected. To the best of our present knowledge, the atmospheres of Venus, Mars, and perhaps of Mercury as well, are composed primarily of nitrogen and carbon dioxide. But there is something strange about the Earth. There is oxygen in our atmosphere.

▽ As rust and fire attest, oxygen is a reactive gas. It combines with other molecules rapidly at high temperatures, more slowly at low temperatures, to form new chemical compounds. Sometimes energy must be supplied in order for oxygen to react. Sometimes a catalyst, perhaps water, speeds the rate of reaction. But reaction with oxygen is inexorable, and, in the inorganic world, irreversible—a one-way street—as long as the oxygen lasts. When a material combines with oxygen, it is said to be "oxidized." Thus, water is an oxidized form of hydrogen. Materials with a large hydrogen content are said to be "reduced." Thus, water can also be described as a reduced form of oxygen. The atmospheres of the Jovian planets are reduc*ing;* the atmosphere of the Earth is oxidizing.

▽ Organic matter—matter of either biological or abiological origin which contains carbon—has a high hydrogen content. It and we who are made of it are characteristically reduced. Yet we live in an oxygen atmosphere. The complete oxidation of organic substances produces carbon dioxide, water, and nitrogen. Such indiscriminate oxidation is clearly debilitating, destroying the material of which we are made. Consequently living organisms on the Earth use a variety of mechanisms, some of them very sophisticated, to avoid contact with oxygen altogether, to shunt the oxidation to non-injurious molecular reactions, or to repair the oxidation damage which has occurred. In a very real sense, we Earthly organisms are living in a poison gas. But more startling yet, some of us are breathing it. Our ancestors have evolved systems for *utilizing* oxygen to the point where our accommodation to the poison has given us a great subsidiary advantage. Combination of metabolic products derived from the breakdown of food with molecular oxygen permits our food to be oxidized completely, to carbon dioxide and water.

▽ It appears that a great selective advantage was conferred upon organisms

which evolved mechanisms to cope with the presence of free molecular oxygen in the atmosphere. Not only did they avoid indiscriminate oxidation of their own material and consequent degeneration; they also evolved the capability of *selective* oxidation of foodstuffs, which enables much more energy to be extracted from the food. For example, two organisms, an anaerobe which does not utilize molecular oxygen, and an aerobe, which does, may each ingest the same quantity of sugar, but the aerobe may extract ten times more energy from it.

▽ Because of the metabolic efficiency of aerobes, it has been suggested that organisms on planets which lack oxygen may not be very advanced. But this is an unimaginative conclusion. There may be more energetic foodstuffs available elsewhere; or the organisms there may eat at a faster rate than do organisms here; or their metabolic processes may be correspondingly slower. It is premature to infer that every planet populated with higher organisms must have an oxygen atmosphere.

▽ If oxygen utilization provides a significant metabolic advantage, why are there on Earth today the obligate anaerobes, organisms which are poisoned by molecular oxygen? These organisms, none of them more advanced than worms, live in the relatively few environments on Earth where molecular oxygen is absent—some soils and oceanic mud, for example. It is possible that many obligate anaerobes are degenerate, evolved from predecessor organisms which were capable of utilizing molecular oxygen. A species living many generations in an oxygen-depleted environment would have no selection pressure to improve, or even maintain, its oxygen-utilization apparatus. A mutation which cripples this apparatus would not be deleterious in an oxygen-poor environment. Given enough time, such a mutation is bound to occur, and the facultative anaerobe—one which can take its oxygen or leave it—would have evolved into an obligate anaerobe. But might some of the obligate anaerobes be descended entirely from anaerobic ancestors? Might they be relics of an earlier epoch when anaerobic conditions were more common than they are today?

▽ Interestingly, the metabolism of sugar by many anaerobes is identical in its early phases with the sugar metabolism of aerobes. (Both organisms take hexose —any 6-carbon sugar, like glucose—and convert it to a hexose phosphate. The hexose phosphate is transformed into hexose diphosphate; the diphosphate is split into two molecules of glyceraldehyde phosphate, etc.) Each step is catalyzed by at least one enzyme. In animal tissue, in yeast, and in many bacteria, 10 of the first 14 steps in the metabolism of hexose sugars are identical. This is another illustration of the essential biochemical similarities of diverse terrestrial organisms; it can most simply be explained by the supposition that all organisms now living on Earth had a common ancestor. After these common first steps in the breakdown of sugar, the metabolic pathways of various organisms diverge. The anaerobes carry the energy extraction process only a few steps further and are content. The energy is locked in the phosphorous bonds of a molecule called adenosine triphosphate (ATP), a ubiquitous molecule in terrestrial living systems which serves as a kind of common energy currency. When energy is extracted from food, it is put into the

energy-rich bonds of ATP; when energy is needed to drive a reaction, the energy is extracted from the energy-rich bonds of ATP. (The letter T here stands for "tri," and not for thymidine, as it did in our discussion of DNA in Chapter 14.)

▽ On the other hand, aerobic organisms carry the metabolic pathway many steps further, utilizing molecular oxygen to extract all the chemical energy available in sugar, and to store it in ATP for future use. How curious that the anaerobic metabolic steps are common to both aerobes and anaerobes! Up to the point at which the molecule pyruvic acid is made, the pathways are common; after that, they are strikingly divergent. This circumstance suggested to the versatile Anglo-Indian scientist J. B. S. Haldane, in 1927, that the common ancestor of contemporary terrestrial organisms was an anaerobe, and that aerobic metabolism was a more recent elaboration. Haldane went further: he proposed that early organisms were anaerobic because the primitive atmosphere of the Earth lacked molecular oxygen, and was instead rich in reduced compounds. Haldane felt that the pre-biological synthesis of hydrogen-rich organic molecules would be much easier to understand in a reducing environment.

▽ At the time, it was a radical suggestion. Prevailing views on the chemical composition of the primitive terrestrial atmosphere leaned heavily towards N_2 and CO_2, neither of which are reduced. The fact that the universe is mostly hydrogen was not known until 1929, and the existence of methane and ammonia in the atmospheres of the Jovian planets was discovered only in 1934. A few years later, a book entitled *The Origin of Life* was published in the Soviet Union by the Russian biochemist A. I. Oparin. Drawing heavily upon the astronomical evidence then available, Oparin independently suggested that the primitive atmosphere of the Earth had been reducing. More than a decade earlier, Oparin had drawn the same conclusion, in an article for a Communist Party periodical, from the supposed abiological synthesis of petroleum. Incidentally, whether petroleum is produced entirely biologically, partly biologically, or entirely abiologically is still undecided.

▽ The idea, promulgated by Haldane and Oparin, of a reducing primordial environment of the Earth is the touchstone to later experiments on the origin of life, which form the subject of our next chapter. We have seen previously, in Chapter 4, that from cosmic abundances, a "typical" planetary atmosphere should be composed of hydrogen, helium, methane, ammonia, and water. The atmospheres of the Jovian planets—Jupiter, Saturn, Uranus, and Neptune—have, we believe, just this composition. The Earth, and the other planets of terrestrial type—Mercury, Venus, and Mars—should have begun their careers with similar atmospheres; at least in the case of the Earth, there is independent supporting evidence. Why did the terrestrial planets lose their primitive atmospheres, while the Jovian planets retained theirs?

▽ The outermost region of any planetary atmosphere is called its exosphere. It is from the exosphere that molecules escape into interplanetary space. Suppose we take an object—any object, of any mass—a great distance away from the Earth and then let it fall. In the absence of air resistance, there is a certain velocity with which it impacts the Earth. Reversing the process, if a mass—again, any mass—is

projected upwards from the surface of the Earth with this same velocity, it will reach very great distances. There is a critical velocity, called the escape velocity, beyond which the ejected mass will continue to travel indefinitely—that is, when it is moving so rapidly that terrestrial gravity cannot quite drag it back. For the Earth, it is 11.2 km sec^{-1}, or about 7 miles per second, and is the velocity with which a space vehicle must be ejected, if it is to escape from the Earth and go someplace else.

▽ The same ideas which apply to Gemini capsules and Voskhods apply to atoms and molecules. If they are traveling fast enough in an upward direction, they can escape from the Earth—unless some other molecule is in the way. An oxygen molecule in the air just in front of you, moving upward with a velocity of 11.2 km sec^{-1}, will not escape from the Earth. It is in fact quite unlikely it will even escape from the room. Even if you direct that rapidly-moving oxygen molecule out into the open air, the situation is not much improved. As soon as it gets started on its hopeful journey outward, it collides with another molecule—probably a stolid, slow-moving, type—jostles its neighboring molecules some, and then slows down to a more pedestrian molecular velocity. It is only in the exosphere that a molecule moving outward with escape velocity has a good chance of escaping; there the density of the atmosphere is so low that the probability of collision with another molecule on the outward voyage is small. But since the molecular population of the exosphere is small (by definition, as we have just seen), the amount of mass escaping from a planetary exosphere tends to be relatively small.

▽ Low-mass molecules escape much more easily than high-mass molecules. Why? Because at a given exosphere temperature, all molecules tend to have the same energy. (This is not quite true, because the fastest-moving molecules have already escaped, but their place is soon taken by molecules from below.) Now a massive molecule moving slowly may have the same energy as a faster-moving molecule with a smaller mass. There is a distribution of molecular velocities. Most are moving with some average speed; a few are moving very slowly; a very few are moving very rapidly. Of those few moving very rapidly, some, by chance, are also moving outward. They escape.

▽ Thus, in any planetary atmosphere, hydrogen, the lightest molecule, will escape preferentially. The loss of hydrogen is replenished, to some extent, by outgassing from the interior, especially in earlier times; and by the solar proton wind—ionized hydrogen atoms blown outwards from the solar atmosphere. In the case of the terrestrial planets, the exosphere temperatures are relatively high, and the force of gravity is relatively low. Both circumstances tend to enhance the escape of gases from their exospheres. The rate of escape of hydrogen is today much larger than the rate of replenishment from outgassing or from the solar wind.

▽ The Jovian planets, by contrast, have such large gravitational forces and such low exosphere temperatures (since they are far from the sun) that even hydrogen, the lightest gas, never escapes. A characteristic time required for a significant fraction of the hydrogen in the terrestrial exosphere to escape is

something like 1000 years. The corresponding number for the exosphere of Jupiter is a googol years or so. Heavier gases, of course, move more sluggishly, and have greater difficulty escaping. Significant amounts of helium also escape from the terrestrial exosphere, but molecules as massive as atomic oxygen are too heavy and do not escape from the Earth. Mars, with its lower gravity, may conceivably have permitted the escape of substantial amounts of atomic oxygen during the age of the solar system, a fact of some significance in assessing the possibility that conditions on Mars were more Earthlike in earlier times. In the case of Mercury, its low gravity and high exosphere temperature (because of its proximity to the Sun) suggest that all molecules less massive than argon (atomic weight 40) have escaped during the last 5×10^9 years.

▽ We should emphasize that it is the exosphere temperature, and not the surface temperature, which determines the rate of escape. On the Earth, surface temperatures are about 300°K, which is much too low to permit escape of substantial amounts of hydrogen. But since the exosphere temperature is characteristically 1600°K, and sometimes, during solar activity, goes above 2000°K, hydrogen escapes. Thus, we see that if the terrestrial planets and the Jovian planets started out their careers with extensive reducing atmospheres, the terrestrial planets would have lost their hydrogen by escape into interplanetary space, while the Jovian planets would have retained theirs, in good agreement with observation.

▽ But there is another wrinkle to this problem. As first noted by D. H. Menzel, of Harvard University, and Henry Norris Russell, of Princeton University, the Earth is deficient in such noble gases as neon, argon, krypton, and xenon. From astronomical spectroscopy and from analyses of meteorites—the only samples of extraterrestrial matter that we can get our hands on at present—we know that the noble gases are generally more abundant—relative, say, to silicon—almost everywhere else in the universe. Thus, if the Earth started out with cosmic composition, some process has depleted the noble gases. The depletion has been greatest for the lighter noble gases, such as neon and argon and less marked for the heavier noble gases, krypton and xenon. The noble gases are particularly important in studies of the evolution of a cosmic object, because with only a few exceptions they do not form chemical compounds; also, they remain gaseous down to very low temperatures. Since they do not combine chemically and do not freeze out, they must have been removed when they were gases. The preferential removal of the low atomic weight noble gases might suggest exospheric escape; but we have just seen that the escape of significant quantities of any gas heavier than helium would not occur, given the present gravity and exosphere temperature of the Earth.

▽ If we want to explain the noble gas depletion by molecular escape, we must therefore assume that the temperature of the primitive exosphere was greater, or that the force of gravity on the primitive Earth was less than at present. In Chapter 6 [see Figure 6–3], we discussed the early evolution of the sun and saw that as it was contracting approximately vertically, towards the main sequence, in the spectrum-luminosity diagram, its brightness was much greater than it is today.

▽ With present values of the acceleration of gravity on Earth, exosphere temperatures as high as 100,000°K were required for escape of noble gases, temperatures some scores of times larger than contemporary values. The same theory of exospheric escape shows that the escape timescale must have been extremely short, a few thousand years. The exosphere temperature is determined by the absorption in the upper atmosphere of ultraviolet photons from the Sun through the conversion of photon energy into the motion of the absorbing molecules, that is, into heat. While the exosphere temperature does depend on the particular molecular species in the upper atmosphere, very crudely the exosphere temperature is proportional to the intensity of absorbed ultraviolet sunlight. Thus, the depletion of noble gases can be understood if for a period of some thousands of years the Sun had a luminosity some scores of times greater than its present luminosity.

▽ From Figure 6–3, we see that such luminosities and such timescales are consistent with present models of early solar evolution. This rough numerical coincidence lends some credence to the view that the present underabundance of noble gases on the Earth is due to atmospheric escape from a hot primitive exosphere. The approximate numerical agreement makes the argument more persuasive than a purely qualitative argument would have been. However, there are other factors which must be examined—such as the rate of arrival of material from the lower atmosphere at the escape level—before this theory can be accepted.

▽ Alternatively, suppose that the Earth was still in the process of formation at the time the Sun reached the main sequence. The gravity at the surface of an object of the mass of the Earth, but greatly distended, can be much less than the present value—in fact, sufficiently less to explain the escape of the noble gases without invoking exosphere temperatures in excess of present values.

▽ But now, if a large fraction of the Earth's initial complement of a gas as heavy as xenon or krypton escaped into space, then essentially all of the lighter gases—methane, ammonia, water, hydrogen, and helium—must also have been wiped off the Earth. Yet we have an atmosphere today. Where did it come from? The solar wind is an entirely inadequate source. Our present atmosphere must then have arisen from outgassing—from volcanoes and fumeroles, bubbles of gas, penetrating to the surface and unable to escape from the now completed Earth.

▽ There is other geological evidence—from the composition of the Earth's surface and from the rate of outgassing observed today—to support this conclusion: the present atmosphere of the Earth is not the same as the gaseous envelope which surrounded the Earth during its formation. The initial atmosphere was lost; our present atmosphere is of secondary origin. What was the chemical composition of this secondary atmosphere, formed after the Earth reached its present size and the Sun reached the main sequence?

▽ The materials outgassed must have been trapped in the interior of the Earth during formation. Thus, their composition would be typical of the solar nebula in the vicinity of the Earth. Hydrogen was in excess; the other atoms were reduced. Material could have been trapped in two ways: by occlusion and by precipitation.

In occlusion, a gas bubble is physically trapped—for example, inside a rock; in precipitation, a chemical compound is formed which drops out of the atmosphere as a solid or a liquid, and cannot escape to space. Then, during the end of the formation process, the temperature of the Earth began to increase. The aggregation process itself provided one source of heat. Material of all sizes, from dust grains to asteroids, was plummeting down, colliding, and providing the material for the accreting Earth. These collisions released heat—possibly enough to melt the surface. Another source of heat was radioactivity. There are reasons, both from the study of meteorites and from theories of the origins of the elements, for believing that radioactive isotopes now long extinct (due to radioactive decay) were then present in great numbers. Their decay produced heat, and this again contributed to the temperatures required to cause outgassing. Whether the Earth was ever completely molten is today largely unknown, despite intensive geophysical and geochemical attacks on the problem.

▽ If the Earth's secondary atmosphere was also reducing, how do we explain the transition from a reducing to the present oxidizing atmosphere? Here, we recall the discussion of preferential molecular escape of lighter gases, especially hydrogen. In the upper atmosphere of the primitive Earth, hydrogen-rich molecules—in particular, water, methane, and ammonia—were being photodissociated by ultraviolet light. Just as a sufficiently energetic photon can ionize an atom, separating an electron from the nucleus, so can a less energetic photon break a molecule into pieces—water, for example, into the components OH and H. After absorption of another ultraviolet photon, the OH may be further separated into O and H. These hydrogen atoms escape to space; the oxygen atoms cannot. The net result is the preferential escape of hydrogen, and the oxidation of the atmosphere remaining behind. Methane tends to be converted to carbon dioxide, ammonia to molecular nitrogen. If the process continues long enough, free oxygen will form. We do not know whether water vapor photodissociation and the subsequent escape of hydrogen is adequate to account for the oxygen now in our atmosphere and chemically combined in the crust. Some calculations suggest that the process is adequate to account for the present oxygen; others suggest that it is not. Certainly, today, the oxygen content of our atmosphere is determined by green plant photosynthesis—a sophisticated sort of photodissociation of water, where visible photons are employed by the plants, instead of the ultraviolet photons which are effective in the upper atmosphere. Conceivably, there was an oxygen atmosphere before green plant photosynthesis was rampant on Earth; but it is possible also that no free oxygen was produced until green plants flourished.

▽ The date of the transition between the secondary reducing atmosphere of the Earth and the present oxidizing atmosphere is therefore difficult to establish. From Figure 6–3 of Chapter 6, we note that after reaching the main sequence, the sun had a distinctly smaller luminosity several billion years ago, than it has today. Thus, the temperature of the Earth should have been lower. The geological record shows signs that algae inhabited the Earth some 2.7 billion years ago [Figure 16–1]. More recently a number of workers have found phytane (a fragment of the

Dentate structure Collenia structure Fragmental limestone

FIGURE *16–1. Algal limestones found in the Rhodesian Shield of Africa. These structures exhibit the characteristic features of limestone secreted by calcareous algae. It is unlikely, though still possible, that these patterns were produced abiologically. If they are of biological origin, as many scientists believe, they are among the oldest signs of life on the Earth. These limestones were found embedded in rocks dated about 2.7 billion years old.* [*Reproduced from a paper by A. M. MacGregor, in the* Trans. Geol. Soc. South Africa *43:9 (1940); by permission of the Geological Society of South Africa.*]

chlorophyll molecule) and other signs of ancient biological activity in Minnesota sediments dated again at about 2.7 billion years. The temperatures on the surface of the Earth 2.7 billion years ago, must therefore have been above the freezing point of water. Yet at that time, the solar luminosity was so low that the average temperature on the Earth's surface should have been about 20 or 30 Centigrade degrees below present values—that is, well below the freezing point of water—unless there is some other factor which we have not yet taken into account.

▽ If we were to calculate the temperature of the Earth today from the albedo (or reflectivity) of the Earth, the solar luminosity, and the distance of the Earth from the sun, we would conclude that temperatures on the Earth were about 20° below zero Centigrade. Such a calculation neglects the effect of the atmosphere. Carbon dioxide and water in the terrestrial atmosphere are transparent in the visible, as everyday experience attests; however, they tend to be quite opaque at infrared wavelengths. Thus, sunlight passes unimpeded through water vapor and carbon dioxide in our atmosphere and heats the ground. But when the ground tries to radiate back to space in the infrared, it finds its efforts hampered by atmospheric absorption by CO_2 and H_2O. These molecules play the same role as the glass in a greenhouse, which is also transparent in the visible and opaque in the infrared. The effect is known as the "greenhouse effect." The temperatures inside a greenhouse and on the Earth are larger than we would expect if infrared radiation escaped from them unimpeded. It seems reasonable, then, to explain the additional 20 or 30 degree temperature increase 2.7 billion years ago by a slightly extended atmospheric greenhouse effect. But what molecules must we use?

▽ The carbon dioxide abundance in the Earth's atmosphere is believed, for reasons of chemical equilibrium, to be roughly constant throughout geological time. The water vapor abundance depends on the surface temperature. To increase the amount of water vapor, we must increase the surface temperature. But it is exactly an increase in surface temperature that we are trying to explain.

Therefore, some other molecule is needed—one which is not present in significant amounts in the present atmosphere. The reduced gases methane and ammonia are ideal for this purpose. They are such efficient infrared absorbers that even relatively small quantities could explain the required temperature increase. But if this argument is correct, it must follow that the atmosphere was at least slightly reducing as recently as 2 or 3 billion years ago. Is there any geological evidence to support this conclusion?

▽ There are minerals known which change their chemical composition when the oxygen or hydrogen concentration of their environment varies. The appearance of the minerals limonite, hematite and calcite in geological sediments between 2 and 2.5 billion years old shows that large amounts of free hydrogen were not available in that epoch. In very old South African sediments, an extremely rare mineral, uraninite, can be found. With a few exceptions, it is not found in recent sediments, because uraninite is a reduced uranium mineral, UO_2. In the presence of even small amounts of molecular oxygen, it is oxidized to U_3O_8, the form of uranium oxide found, for example, in pitchblende. The geological evidence suggests that the uraninite minerals were formed between 2 and 3 billion years ago.

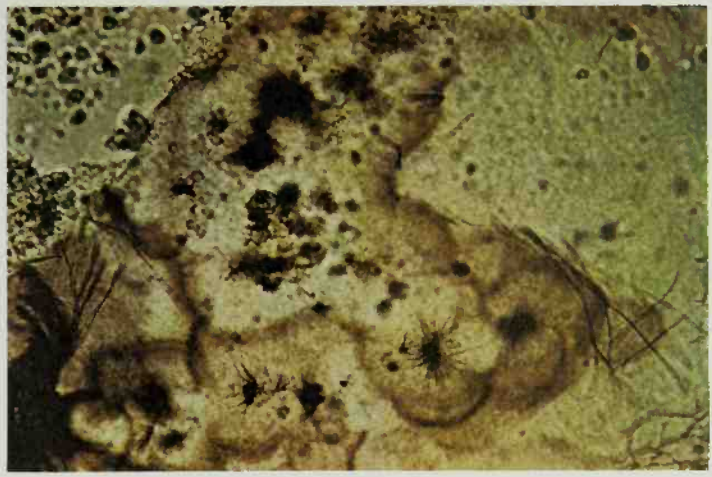
▽ Thus it appears that between two and three billion years ago, the atmosphere was not very reducing, nor was it even slightly oxidizing. To avoid very low temperatures, we must postulate small amounts of methane and ammonia. This is still consistent with neutral or slightly reducing conditions. From several lines of evidence, it appears that green plant photosynthesis and oxygen reduction were fairly extensive. To avoid the accumulation of large amounts of free oxygen in that epoch, we must postulate that the photosynthetic oxygen produced was used in oxidizing the reduced constituents of the atmosphere and of the Earth's surface. Only later could free molecular oxygen form. Unfortunately, there is not yet enough geological evidence to determine when substantial quantities of free oxygen were first formed. The large insects, evolved in Carboniferous times some 500 million years ago, may have demanded large amounts of oxygen for their metabolism. Thus, the transition to an oxidizing atmosphere probably occurred between 2 and 0.5 billion years ago. As we go backward to earlier epochs, the atmosphere was progressively more and more reducing.

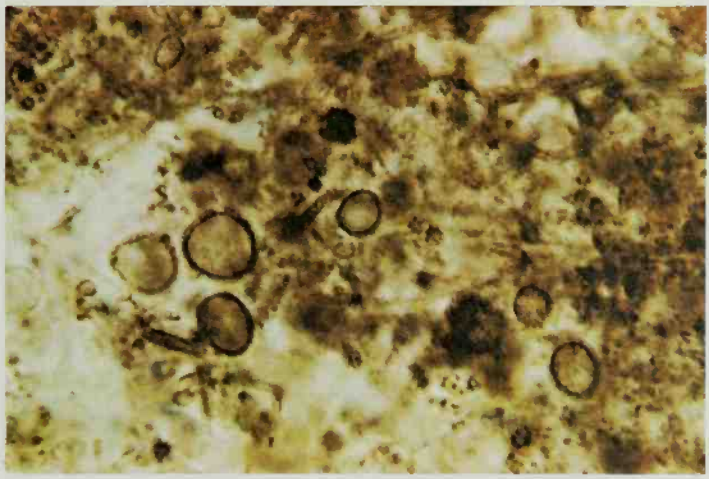
▽ We are now in a position to reconstruct something of the timescale of the origin of life. The Earth was formed some 4.5×10^9 years ago. While the Earth may never have been completely molten, its surface seems to have been quite hot during the process of formation. The origin of the Earth antedates the origin of life on the Earth. It is likely that several hundred million years after the Earth's formation surface temperatures were largely below the boiling point of water, and some atmosphere and ocean had been outgassed from the interior. The atmosphere was reducing, and the stage was set for the origin of life. By 2.7 billion years ago, the atmosphere was not yet oxidizing, but life had come into being and evolved to the stage of complexity of algae. Figures 16–2 through 16–7 give some idea of the kind of small plant and microorganism remains which are found in 2-billion-year-old sediments. They come from the Canadian Shield and were discovered by the

American paleobotanist Elso Barghoorn, of Harvard University, through whose courtesy they appear here in color. Not surprisingly, some of these microfossils correspond to no known variety of organism. More recently Barghoorn, has found fossil evidence of 3.1-billion-year-old bacteria.

▽ Since bacteria and especially algae are very complex living systems, enormously more advanced than the first living systems, the origin of life must be dated considerably before 3.1 billion years ago. As a guess, we may put the origin of life at $4.0 \pm 0.5 \times 10^9$ years ago. If the origin of life was easy, in a sense which we will define in the next chapter, then a date for it of about 4 billion years ago may be close to the truth. If we imagine one billion years from that date spent in the evolution of the first cell from simpler living systems, and two billion years for the elaboration of this single-cell design, we will come to a point about one billion years ago, shortly before the fossil record begins in earnest, when great varieties of multicellular organisms were emerging. Perhaps the great diversification of animals at the beginning of the Precambrian Epoch, some 600 million years ago, was due to the evolution of biological mechanisms for utilizing the large amounts of free oxygen newly available through plant photosynthesis, as the American geophysicist Lloyd V. Berkner has suggested.

▽ The times required for establishing the basic chemical patterns and structural organization of terrestrial life were probably much longer than the time required for the elaboration of such variations on the same theme as microbes, maple trees, mantises, and men.





FIGURES *16–2 and 16–3. Typical photomicrographs of sediments in the Gunflint Chert of the Canadian Shield, which are rich in organic matter. The tangled filaments and the roughly spherical objects are both microfossils of biological origin. The spheres are known as* Huroniospora microreticulata *Barghoorn, which roughly means that it is a spore found in the vicinity of Lake Huron, that it has a network of fine lines in its interior, and that it was discovered by Barghoorn. The thick wall and the network of fine lines both suggest that the subject is the spore of some other organism. A typical* Huroniospora *is a few microns in diameter. The filaments are multicellular, another significant biological advance achieved at the time that these organisms died, some 2.7 billion years ago. Compared with similar contemporary organisms, however, like the blue-green algae, these filaments show much more irregularity from cell to cell—a circumstance not altogether unreasonable for a primitive organism. (Courtesy of Prof. Elso Barghoorn, Harvard University.)*

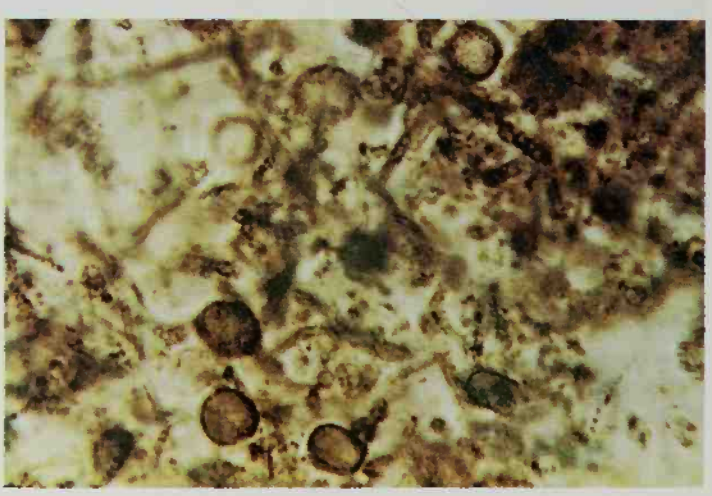


FIGURE *16–4. Photomicrograph of the Gunflint Chert showing an organism, assigned by Barghoorn the genus name* Eoastrion, *the dawn star. From the central body small dark filaments are seen to radiate irregularly. The entire object is shown here encapsulated in an outer sheath some microns across. (Courtesy of Prof. Elso Barghoorn, Harvard University.)*

FIGURE *16–5. Photomicrograph of some other specimens of* Eoastrion *taken from the Gunflint Chert. (Courtesy of Prof. Elso Barghoorn, Harvard University.)*

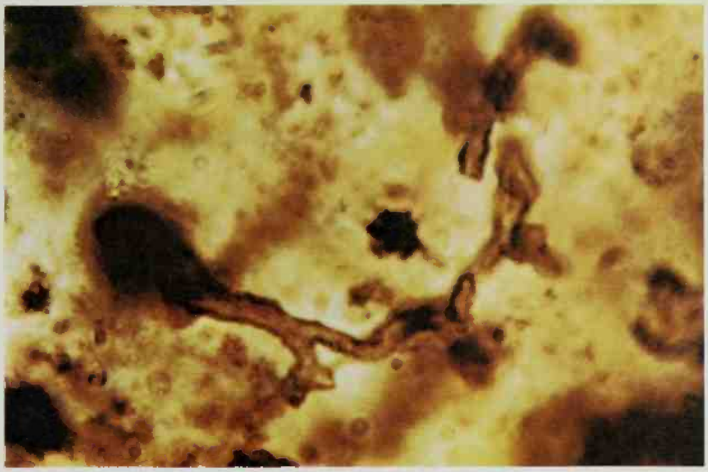


FIGURE *16–6. Photomicrograph of* Archaeorestis schreiberensis *Barghoorn, which means approximately, an old ropey filament found in Schreiber, Ontario, by Barghoorn. The structure is branched and tubular, with occasional bulbous structures. While it shows vague similarities to some contemporary organisms,* Archaeorestis, *like all the other organisms in this group, is probably an extinct genus. The filaments are a few microns in diameter, but may be hundreds of microns in length. (Courtesy of Prof. Elso Barghoorn, Harvard University.)*

FIGURE *16–7. Photomicrograph of an organism found in the Gunflint Chert. It is called* Eosphaera tyleri *Barghoorn, which means approximately the dawn sphere, found by the American geologist, Dr. Stanley Tyler of the University of Wisconsin, and by Prof. Barghoorn. We are here looking at a section of* Eosphaera. *Three-dimensional examination shows that* Eosphaera *is really two concentric spheres. The inner and outer walls are connected by smaller spheroids, six of which are visible in this picture in which the outer sphere has been partially ruptured. A typical diameter from outer wall to outer wall is about 20 microns. (Courtesy of Prof. Elso Barghoorn, Harvard University.)*

17

Chemical syntheses and early evolution of life

. . . It is often said that all the conditions for the first production of a living organism are now present, which could ever have been present. But if (and oh! what a big if!) we could conceive in some warm little pond, with all sorts of ammonia and phosphoric salts, light, heat, electricity, etc., present, that a proteine compound was chemically formed ready to undergo still more complex changes, at the present day such matter would be instantly devoured or absorbed, which would not have been the case before living creatures were formed.

Charles Darwin (1871)

A hen is only an egg's way of making another egg.

Samuel Butler, *Life and Habit* (1877)

▽ Four and one half billion years ago the Earth was lifeless. Nowhere—not in the primitive atmosphere, not in the early oceans, nor in the newly forming crust—could even the simplest, most unassuming microorganism be found. Two billion years later, the Earth was fairly teeming with one-celled organisms of appreciable complexity. As we have seen in the previous chapter, the origin of the first living systems must have occurred within about a billion years of the formation of the Earth. How? Was it a vastly improbable event which, to our good fortune, occurred by chance in this small corner of the universe, and not elsewhere? Or, starting from the physics and chemistry of the primitive terrestrial environment, was the origin of life a likely event, given only a billion years of random molecular interactions?

▽ It is true that in enough time, almost any random concatenation of molecules, no matter how complex, will occur. But would the appropriate concatenation of molecules—nucleic acids and proteins, for example—occur in the time available? For example, suppose the probability for the origin of the first self-replicating system in the primitive environment in any given year was 10^{-6}. Then the probability that the origin of life occurs in any given century in that era would have been $10^{-6} \times 10^{2} = 10^{-4}$, a small number. But in 10^{9} years, the probability becomes very close to one, and we may talk of the origin of life as a "forced event"—that is, as a highly probable outcome of the chemical interactions on the primeval Earth.

▽ Suppose, instead, that the probability of the spontaneous origin of life in any given year on the primitive Earth were 10^{-12}. Then the probability of living systems arising, even over 10^{9} years, would be $10^{-12} \times 10^{9} = 10^{-3}$, a very small number. In this case, we would conclude that the origin of life was a highly improbable event in the time available in the early history of the Earth, and that life is here at all only through an extraordinary stroke of luck. We do not know what these probabilities really are, but they may be determinable by experimental investigation. If the probability of the origin of life turns out to be high, we can conclude that the origin of life is a likely event on many planets; if the probability turns out to be low, we must conclude that, except for such possibilities as panspermia or intentional colonization, the universe is sparsely populated.

▽ Several decades ago, it was fashionable to think that the probabilities were very low. In his book *Human Destiny,* le Compte de Noüy computed the probability that a sequence of amino acids arranged in random order would duplicate any given protein. We have performed a similar calculation in computing our own improbability, in Chapter 14. If the protein under consideration is 100 amino acids long, and any amino acid slot may be filled by any one of 20

biologically common amino acids, then the chance of random assembly of the given protein is one in 20^{100}, about 10^{130}, or more than a googol. Le Compte de Noüy concluded that such an event is so vastly improbable that it could not have occurred at all. He concluded that the origin of life required divine intervention.

▽ But there are other possibilities. We have seen in Chapter 14 that natural selection serves as a kind of probability sieve, extracting those structures and functions which improve the adaptability of the organism to the environment. But what of the origin of the first proteins, or the first nucleic acids? Must they have been *random* assemblies of their respective building blocks, the amino acids and the nucleoside phosphates? Or might it be that the molecules which spontaneously arose in the primitive environment are the ones which were later utilized in the origin of life?

▽ Contemporary organisms tend to be about 90 percent water, and human beings are no exception. Water is by far the most common molecule on the surface of the Earth. Evidently, life has used the building blocks available. However, sand (SiO_2) is also very abundant, yet relatively few organisms use it; those which do, use it structurally, and not biochemically or metabolically. The nucleic acids are made exclusively of carbon, nitrogen, oxygen, hydrogen, and phosphorus. The first four atoms are among the most abundant in the universe, as we have seen in Chapter 4; but phosphorus is rather rare. The proteins are made of carbon, nitrogen, oxygen, hydrogen, and sulphur. Again, sulphur is relatively rare. We can conclude that living systems have utilized, wherever possible, those atoms and simple molecules in greatest abundance. However, what is most available and what is most useful are often not identical. Thus, some common atoms or molecules have not been incorporated into living systems, while other uncommon ones have been selectively extracted from the environment.

▽ By chance, certain of the abundant molecules, notably water, seem peculiarly well suited for incorporation into living systems. In a book called *The Fitness of the Environment,* published in 1913, Lawrence J. Henderson, a biochemist at Harvard University, discussed at some length the salutary properties of water. For the origin and development of life, we need a liquid medium (or at the very worst, a very dense gaseous medium), in which molecular interactions can take place. For biological stability, the medium should remain liquid over a wide temperature range, and its temperature should vary only sluggishly when it is heated or cooled. In addition, Henderson felt that it would be useful if the liquid could dissolve salts, and participate in an acid-base chemistry. All these properties are shared by water; in several cases, water possesses them to a greater degree than any other common molecule. Henderson was struck by the apparent preadaptation of water for its biological role, and he felt that the coincidence was worth exploring. Others have used Henderson's arguments to conclude that the origin of life occurred by design. Henderson also discussed the "fitness" of carbon, oxygen, and other atoms and molecules for the fundamental roles they play in contemporary biochemistry.

▽ This question of the fitness of the environment provides some perspective

on the problems inherent in extrapolation from a single example. We are talking and walking aggregates of carbon-based organic compounds in a liquid water solvent system. Might we be biased in our judgment that living systems must be carbon-based and aqueous? The American chemist George Pimentel, of the University of California at Berkeley, has argued that the fitness of water and carbon may be illusory, the product of our limited biochemical imaginations and the historical uniformity of terrestrial biochemistry. Some hydrocarbon solutions have wide liquid ranges and adequate temperature stabilities. The ability to dissolve salts or to participate in acid-base chemistry is not a prerequisite for molecular complexity, and many other alternatives can be imagined. At low temperatures, there are silicon compounds which have high stability, and can generate as much complexity as carbon compounds. At room temperature, however, they are not nearly so stable as comparable carbon compounds. In the presence of liquid water, many of them tend to dissociate. On the other hand, they are much more stable in ultraviolet light than many carbon compounds. Thus, silicon-based biochemistries may be appropriate in low-temperature, non-aqueous environments with high ultraviolet fluxes. Pimentel has pointed out that there are many known chemical reactions which proceed at biologically respectable rates at very low temperatures. At room temperature, however, the reactions occur so rapidly that we see only the products of the reaction, and tend to lose sight of the reactions themselves. Thus, while chemical reactions proceed much more slowly as the temperature declines, there do exist chemical reactions which proceed reasonably fast at low temperatures. Analogous statements may be made about high temperatures. We are just beginning to explore alternative biochemistries, and it is quite premature to conclude that ours is the only, or even the best of all possible, biochemistries.

▽ Our high water content has suggested to many biologists that life on Earth arose in the oceans. In fact, there is a rough correspondence between the content of such elements as calcium and potassium in sea water and in blood and tissues. This is our first hint that living systems tend to incorporate the primitive environment, so that their *milieu intérieur* would tend to resemble the familiar surroundings of the early history of life, a possibility first glimpsed by the nineteenth-century French physiologist Claude Bernard.

▽ Prior to 1953, several attempts were made to simulate the primitive environment of the Earth and synthesize organic molecules. The results were generally discouraging. In many of the earlier experiments, the over-all conditions were not reducing. For example, mixtures of H_2O, CO_2, and N_2 might be used and then irradiated with high-energy electrons. Only very simple organic molecules, such as formaldehyde, were produced, and even these, in very low yield.

▽ But in 1953, a major advance was made at the University of Chicago. Having convinced himself that the solar system arose under reducing conditions, the American chemist Harold C. Urey turned his attention to the problem of the origin of life. Urey's collaborator, Stanley L. Miller, prepared a mixture of methane, ammonia, water, and hydrogen as a simulated primitive atmosphere. The idea was to supply energy to such a mixture, and determine whether organic molecules

were produced in detectable yield. Of the energy sources which seemed to be available in primitive times, and which are capable of driving organic synthetic reactions, solar ultraviolet radiation was the obvious choice. However, ultraviolet light is rather difficult to work with, as ordinary laboratory glassware is opaque to it. (This is why it is hard to acquire a sunburn through a window.) Accordingly, Urey and Miller used an electric discharge: high-energy electrons were passed between two electrodes through the simulated primitive atmosphere. Such a flow of electrons is an adequate lightning simulation. If there was water on the primitive Earth, clouds can be expected; and if there were clouds, electrical discharges between the clouds and the ground—that is, lightning—must have occurred. In the experiment, the gas was circulated so that after being sparked it was carried through a water bath, and the organic products produced in the gas dissolved in the liquid, where further reactions were possible. After a week of sparking, the liquid turned a deep brown. Clearly, new molecules were being produced from methane, ammonia, water, and hydrogen. But which ones? Were they organic?

▽ To analyze the composition of their mixture, Miller and Urey used an analytic technique called paper chromatography, which, with minor variations, has been used extensively in subsequent experimental work on the origin of life. If you dip a piece of white blotting paper in a bottle of black ink, the ink will move up the blotter a certain distance, and then stop, because of the molecular interactions between the ink and the paper. The ink is drawn up through the capillaries of the porous paper a certain distance and no further. If you perform such an experiment, you will observe that the black ink has separated into its constituent green and magenta pigments, and that the green and magenta pigments, being attracted in different degrees by the paper, have moved different distances up the blotter. This is a simple example of paper chromatography. In ordinary laboratory use, an unknown sample is spotted at the corner of a large piece of chromatography paper, which is similar to ordinary filter paper. The paper is then placed in an organic solvent, which seeps along the paper, carrying the unknown sample with it a certain characteristic distance. The paper is then turned at right angles and dipped into another solvent, which carries the unknown in a direction perpendicular to the first. The procedure is usually adequate to separate a large number of unknown samples into discrete spots on the chromatogram. Using the same paper and the same solvents, different organic molecules in the unknown will be carried to certain characteristic positions on the page. If these spots are colored, either intrinsically, as for ink, or extrinsically, by a spray, the position of the spots may be determined, and the composition of the unknown ascertained. Photographic emulsions are very sensitive to electrons emitted during radioactive decay. Thus, when the expected yield is very low, atoms of one of the original reactants—say, methane—may be labeled with a radioactive isotope—for example, radioactive carbon^{14} in place of the ordinary C^{12}. Carbon^{14} is subject to radioactive decay, giving off an electron from its nucleus (a neutron has been transformed into a proton and an electron), and the electron escapes from the molecule. This electron is capable of exposing a grain on a photographic emulsion. Consequently, newly

synthesized compounds, whose atoms are so labeled, can be detected even in very small amounts by placing the paper chromatogram against an x-ray film and then developing the film. The newly-formed labeled molecules, in effect, take pictures of themselves.

▽ Figure 17–1 is an example of a developed autoradiogram negative. The darkest spot in the middle of the other spots is due to the molecule adenine, labeled with C^{14}. The adenine had been mixed with the sugar ribose and phosphoric acid, and irradiated with ultraviolet light for eighteen hours. The products were then run in two solvent systems, and the resulting two-dimensional chromatogram was placed against film with the result shown in the figure. Each of the spots, other than adenine, corresponds to an organic molecule synthesized in this experiment. Many of the molecules shown in this figure have not been identified as yet. This chromatogram was obtained with the assistance of Mrs. Elinore Green in my laboratory at the Smithsonian Astrophysical Observatory.

▽ Using ordinary paper chromatography, with a color stain, Miller and Urey found that they had produced, in high yield, large numbers of amino acids, the building blocks of the proteins. In addition, other organic molecules were produced, most of which are also involved in contemporary biological processes, although some, like urea, are involved mostly as end products. About 85 percent of the products produced in this experiment remain unidentified to this day. Some of these unknown products are believed to be sugars; others were long, tarlike polymers, responsible for the deep brown color which the aqueous solution acquired after a week of sparking.

▽ The original work of Miller and Urey has also been confirmed by the American geochemist Philip H. Abelson, of the Carnegie Institution of Washington. Using electric discharges in a wide variety of gas mixtures, Abelson found that as long as the net conditions were reducing, it was possible to replace CH_4 by CO_2, or NH_3 by N_2, and still produce the amino acids and the other products obtained by Miller and Urey. However, as soon as the net conditions become oxidizing, the organic synthesis effectively turns off. This is strong confirmatory evidence that for the large-scale prebiological synthesis of organic molecules, reducing conditions are required.

▽ Subsequently, in 1959 the German chemists W. Groth and H. von Weyssenhoff, at the University of Bonn, showed that ultraviolet irradiation at wavelengths where the gas mixture absorbed gave results similar to those of Miller and Urey. Ultraviolet light is fairly efficient in producing organic molecules from a mixture of ethane, ammonia, water, and hydrogen. The quantum yield is a quantity which expresses the number of organic molecules of a given type produced for every photon of ultraviolet light absorbed by the gas. Groth and von Weyssenhoff found characteristic ultraviolet quantum yields of about 10^{-5} to 10^{-6}; that is, it took between 100,000 and 1,000,000 photons to produce one organic molecule of a given type—amino acids, for example. Similar quantum yields have been found in later work, which we will discuss presently, on the ultraviolet synthesis of nucleic acid precursors.

▽ If we know the quantum yield, we can, in principle, compute the total

FIGURE 17–1. *An example of a two-dimensional autoradiographic chromatogram. The darkest spot corresponds to a labeled molecule used as a starting point in the experiment; the other spots, many of them unidentified, correspond to new organic molecules synthesized in the experiment.*

amount of organic matter formed while the Earth retained its reducing atmosphere. A typical mass for a simple molecule synthesized under primitive conditions is 10^{-22} gm. With a quantum yield of 10^{-5}, we have $10^{-5} \times 10^{-22} = 10^{-27}$ gm of organic matter produced per photon absorbed. A typical value of the ultraviolet photon flux at the top of the Earth's atmosphere in primitive times is 3×10^{14} photons cm^{-2} sec^{-1}. That is, every square centimeter of the Earth received 3×10^{14} ultraviolet photons per second. Since each photon produced 10^{-27} gm of organic matter, the total solar ultraviolet flux each second produced $10^{-27} \times 3 \times 10^{14} = 3 \times 10^{-13}$ gm over each square centimeter. If the primitive reducing atmosphere lasted 3×10^{8} years (about 10^{16} seconds), then $3 \times 10^{-13} \times 10^{16} = 3 \times 10^{3}$ gm of organic matter must have been produced by ultraviolet radiation during that time over each square centimeter of the Earth's surface. This is three kilograms per square centimeter, a sizable amount. The average depth of the present oceans is about 3 km, or 3×10^{5} cm. Since water has a density of 1 gm cm^{-3}, there are 3×10^{5} gm of water in a column one square centimeter in cross-section and 3 km high. Thus, if all the primitive Earth's organic matter were dissolved in the present oceans, we would have an aqueous solution which is $3 \times 10^{3}/3 \times 10^{5} = 10^{-2}$, or a one percent solution of organic matter. This is just about the consistency of a thin consommé, and confirms the expectation expressed by J. B. S. Haldane, in his earliest papers on the origin of life, that living systems arose in a "hot, dilute soup."

▽ Since the trail-blazing experiments of Urey and Miller, a large number of other, more complex organic molecules have been produced in a similar way. So far, the soup has been garnished according to the following recipe: start with methane, ammonia, water, and hydrogen, and see which simple molecules are produced in liquid water (for example, amino acids), or in the gas when liquid water is absent (for example, aldehydes and hydrogen cyanide). Then take these molecules, mix them together, and supply more energy. Take the products of this second step and use them as the reactants for the third step. Continue until the molecule you are looking for is made. A variety of energy sources have been used: high-energy protons and electrons, ultraviolet light, x-rays, gamma rays, and heat. Some of these energy sources—particularly electrons and ultraviolet radiation—are reasonable simulations for the primitive environment of the Earth; others are not. In some experiments, unrealistically high concentrations of organic matter have been used, as if the primitive seas had been composed of 50 percent organic matter, instead of perhaps one percent or less.

▽ The simulations are at best inexact. For example, pure reactants are used, whereas the primitive environment was not chemically pure. It is, of course, impossible for us to know the detailed chemical and physical conditions over the entire surface of the Earth, some 4×10^{9} years ago. Thus, the chemical reactions which occur in these studies cannot exactly simulate those which occurred on the early Earth.

▽ As another example, the energy source used to make the molecule of interest also tends to destroy it. Short wavelength ultraviolet light, for instance, will

dissociate amino acids. To avoid such destruction, the experimenters have removed the product molecules from the energy source. This is one reason for circulating the gas through a liquid medium, in the Miller–Urey experiment. In experiments with ultraviolet light or heat, the energy source may simply be turned off after the desired products are synthesized. The primitive environment of the Earth was not as accommodating as the organic chemists, who, understandably enough, want to analyze their product before it is destroyed.

▽ A flash of lightning occurs and is finished; molecules synthesized by the flash are unlikely to be struck by a later lightning bolt. But with ultraviolet light, the synthesized molecules are generally even more liable to dissociation by ultraviolet light than are the precursor molecules. Was ultraviolet light, then, a useless energy source in primitive times, because the synthesized molecules were destroyed before they had any chance to react further and form molecules of biological interest? Not if the origin of life occurred in the oceans. A few tens of meters of pure liquid water will absorb essentially all the ultraviolet light incident on the surface of the waters in primitive times. As the content of organic matter in the waters increased, the organic molecules on the top of the ocean shielded organic molecules a few centimeters below from the dissociating effects of ultraviolet light. Because the primitive atmosphere was reducing, no ultraviolet-absorbing ozone could form. For this reason, I believe that light in the approximate wavelength range 2400–2900 Å penetrated to the surface of the waters in primitive times. As the transition to an oxygen atmosphere occurred, ozone was slowly produced by interactions of oxygen atoms and molecules, and eventually enough ozone was formed to establish a kind of molecular blanket at an altitude of about 40 km, which today protects us from the harmful effects of ultraviolet light. It seems likely that in primitive times the ultraviolet absorbing blanket was the top of the ocean, not the top of the atmosphere.

▽ We can now picture the primitive Earth between 4 and 4.5 billion years ago. There is a reducing atmosphere and bodies of water, both outgassed from the interior. Major tectonic changes are occurring; the continents are being formed; due to gravitational accretion and radioactivity, the Earth may have been, at least at certain times and places, much warmer than it is today. During storms, lightning bolts traverse the atmosphere; during the day, some ultraviolet light from the sun penetrates through the atmosphere and is absorbed in the ocean. The atmosphere is composed of methane, ammonia, water, and very small amounts of hydrogen. Soon, the ammonia will become dissolved in the oceans, where it forms ammonium hydroxide (NH_4OH), which will tend to make the oceans alkaline. The atmosphere may have also contained fairly large amounts of the relatively unreactive gases nitrogen and helium. Due to chemical interactions, the atmosphere has small amounts of aldehydes and hydrogen cyanide in it; dissolved in the oceans are amino acids. In such a setting, what else must occur?

▽ The next stages in prebiological organic synthesis, beyond the Miller–Urey experiments, have been performed in the United States in the 1960's, primarily in the laboratories of the Spanish-American chemist, John Oró, at the University of

Houston; the Ceylonese-American chemist, Cyril Ponnamperuma, at the Ames Research Center of the National Aeronautics and Space Administration; the American chemist, Melvin Calvin, at the University of California, Berkeley; and the American chemist, Sidney W. Fox, of the University of Miami.

▽ In experiments designed to simulate primitive conditions, these workers have succeeded in producing the 5-carbon sugars ribose and deoxyribose; the 6-carbon sugar glucose; the bases of the nucleic acids, adenine, guanine, and uracil; and polypeptides—long chains of amino acids which, at least in some gross chemical properties, resemble proteins.

▽ More recently, Ponnamperuma and I have produced nucleoside phosphates, the building blocks of the nucleic acids, under simulated primitive conditions. The rationale for these experiments is as follows: In the absence of ozone on the primitive Earth, it appears that ultraviolet light was penetrating the atmosphere and reaching the surface of the waters in the 2400–2900 Å region. This is a wavelength range in which ultraviolet light is deleterious to contemporary organisms. (Germicidal lamps emit ultraviolet light at these wavelengths.) Such ultraviolet damage occurs because certain molecular groups of contemporary organisms preferentially absorb at these wavelengths. The chief absorbers are the nucleic acid bases which preferentially absorb at about 2600 Å. In the environment of the primitive Earth, 2600 Å is just in the middle of the ultraviolet wavelength "window" transmitted by the atmosphere. Thus, by a curious coincidence, ultraviolet light was available at just the wavelengths at which the bases absorb. After ascertaining that the bases and the sugars ribose and deoxyribose are produced in simulated primitive Earth environments, we wondered what would happen in the presence of phosphorus. The early oceans should have had phosphates and other phosphorus compounds dissolved in them in small amounts. Thus, in one set of experiments, we prepared a dilute solution of the base adenine, the sugar ribose, and a phosphorous compound such as phosphoric acid in some experiments, or the more complex ethyl metaphosphate in others. The adenine was labelled with C^{14}, and the products were analyzed by autoradiographic paper chromatography. One of the compounds produced in highest yield was adenosine triphosphate, ATP. A molecular model of this molecule appears as Figure 17–2. It is a combination of adenine, ribose, and three phosphates, and can be written, in the notation of Chapter 14, as A-S-P-P-P. The possible primordial synthesis of ATP is significant in two respects.

▽ We have mentioned [Chapter 14] that ATP is ubiquitous in contemporary cells, where it serves as a kind of common energy currency. Today, ATP is made directly by plants in photosynthesis and is synthesized by animals and many microbes from food. But this experiment suggests that in primitive times, ATP may have been made "free," produced abiologically and raining down on primitive organisms like manna from heaven. The quantum yields in ATP production are so high that, if primitive synthesis was about as efficient as in our experiments, every square centimeter of the primitive oceans could have supported, essentially indefinitely, a population of 20,000 bacteria, each with contemporary generation

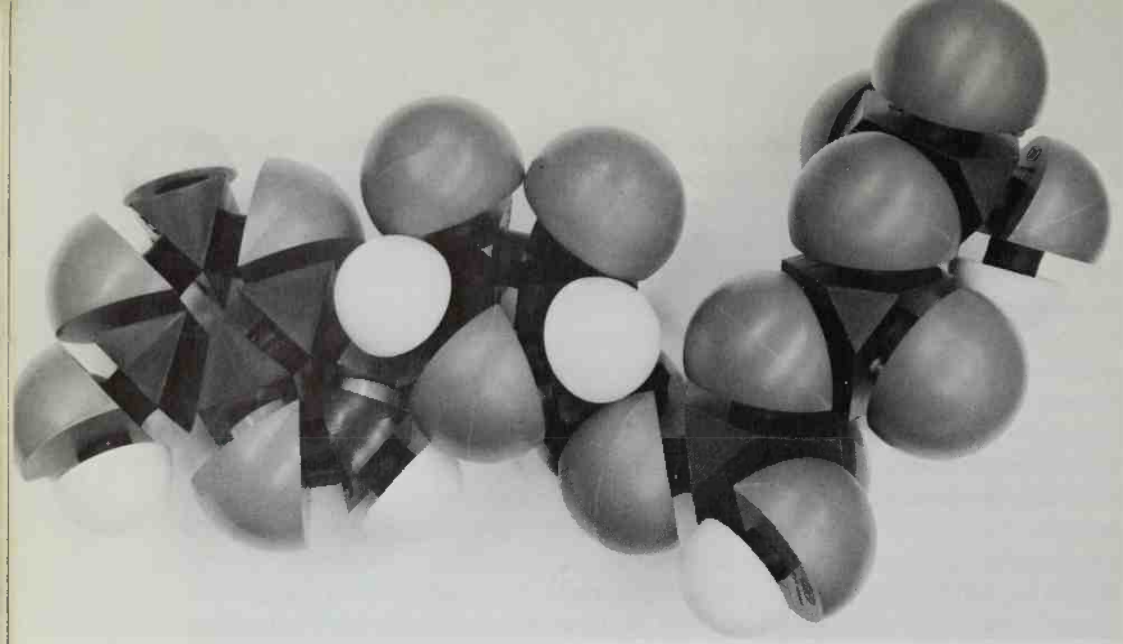


FIGURE 17-2. *A molecular model of the molecule, adenosine triphosphate (ATP). Different atoms are represented by different shaped and colored components. The ring of six wedge-shaped atoms on the left represents the adenine molecule. The "tail" of dark molecules on the right is the triphosphate group, and between them is the ribose sugar.*

times and energy requirements, using as sole energy source the ultraviolet-synthesized ATP. Thus, the elaborate metabolic machinery which today is devoted to generating ATP may have been unnecessary at the dawn of life.

▽ The other significance of ATP is that it is a precursor in the synthesis of nucleic acids in contemporary cells. One of the most exciting recent developments in biochemistry has been the laboratory synthesis of RNA by the Spanish-American biochemist Severo Ochoa, at New York University, and of a kind of DNA by the American biochemist Arthur Kornberg, now at Stanford University Medical School, with their respective collaborators. In each case, the experimenters took nucleoside phosphates, some inorganic substances such as magnesium, and an enzyme of biological origin. For the RNA synthesis, the enzyme is called polynucleotide phosphorylase; for the DNA synthesis, it is called DNA polymerase. Imagine a dilute solution of nucleoside triphosphates, magnesium, and, say, DNA polymerase. Drop into it a small amount of preexisting DNA. In a short time, Kornberg found, there will be synthesized larger quantities of a molecule which, in a great many respects, resembles the original "primer" DNA. A similar situation holds for the RNA. Now, imagine the experiment done again, but without the primer. What happens? Nothing happens, for a while. But soon, it is evident that there has been a slow synthesis of a nucleic acid, even in the absence of a primer. Such spontaneous polymerizations can occur both for RNA and for DNA. The synthesized DNA may be of a very simple and repetitive type (e.g., A T A T A T A T . . .), but it appears to be DNA nevertheless.

▽ To better simulate primitive conditions, imagine that we perform the same experiment, this time with neither the primer nor the enzyme. Now nothing per-

ceptible happens. Why not? What is the function of the enzyme? These enzymes are catalysts which increase the rate, but not the direction, of a chemical reaction. This means that in the absence of DNA polymerase, appropriate nucleoside triphosphates will spontaneously join together, or polymerize, but on a much longer timescale than occurs in the presence of the enzyme. We do not now know what the rate of spontaneous polymerization of nucleoside triphosphates is, in the absence of the appropriate enzyme. Suppose it takes 1000 years. Clearly, such an experiment is not practicable to perform in the laboratory without the enzyme. In fact, the enzyme provides us with the laboratory tool we need. We can trade the enzyme for geological time.

▽ If the spontaneous polymerization of nucleoside triphosphates into nucleic acids takes much more than, say, 10^8 years, we can conclude that primitive nucleic acids did not arise by spontaneous polymerization of nucleoside phosphates. But if nucleic acids do polymerize spontaneously—in intervals short compared with geological time, if long compared with a human lifetime—we can circumvent an otherwise embarrassing problem: As we have mentioned in Chapter 14, proteins are made in contemporary biological systems only by nucleic acids. An enzyme such as DNA polymerase is a protein. So we need nucleic acids to make proteins, and proteins to make nucleic acids. One way onto this biological treadmill is the spontaneous synthesis of nucleic acids in the absence of proteins.

▽ Both some DNA and some RNA precursor nucleoside triphosphates have now been produced under simulated primitive conditions. We can imagine the origin of primitive nucleic acids by the spontaneous polymerization of ultraviolet-synthesized nucleoside triphosphates in a primeval body of water which contained some mineral catalyst such as magnesium. Once the first nucleic acid molecule is synthesized, subsequent syntheses would use it as a primer. A step towards exact nucleic acid self-replication—as occurs in contemporary biological systems—would have occurred on the ancient Earth. After the production of the first polynucleotide, subsequent generations of polynucleotides will mutate, either through interaction with light and with other molecules or "spontaneously." Some nucleoside triphosphates will be removed; others will be substituted for the deletions; in still other cases, short sequences of nucleoside triphosphates will be inverted. Eventually, we can anticipate that the primeval sea would be fairly full of a variety of self-replicating nucleic acids.

▽ If we now have understood in a general way the origin of the first self-replicating, mutating system, have we not also understood the origin of life? No, not quite. There is no way for these primitive nucleic acids to control their immediate environment in a way which enhances their continued replication. In contemporary cells, as we have seen in Chapter 14, there is an elaborate apparatus involving messenger RNA, adapter RNA, ribosomes, and a diversity of specialized enzymes all required for the nucleic acids to control the chemistry of the cell. We cannot imagine these complex and specific accessory molecules to have arisen spontaneously in the primitive environment. The apparatus for the transcription of the genetic code must itself have evolved slowly, through billions of years of evolution.

The major remaining problem in laboratory investigation of the origin of life is the origin of the genetic code. Perhaps nucleic acids are themselves weakly catalytic. Perhaps polynucleotides have a weak ability to order amino acids in a singlet code, rather than the contemporary triplet code. Since primitive nucleic acids would have been composed of four or so kinds of nucleoside triphosphates, this would mean that primitive proteins contained only about four amino acids. Yet the active site—the place on contemporary proteins where most of the catalytic effect occurs —often contains no more than four different kinds of amino acids. We should note that even very weak catalytic abilities will, through many generations, confer significant and perhaps decisive selective advantages on their owners.

▽ Much work has been done on the production of polypeptides from amino acids in simulated primitive environments. A wide range of combinations of amino acids are produced, more recently in aqueous solution. It is possible that the most prevalent varieties of these polypeptides have weak catalytic abilities useful for promoting further syntheses. But since, to the best of our knowledge, polypeptides are non-self-replicating, their spontaneous synthesis in the primitive environment cannot provide the answer to the fundamental questions of the origin of life.

▽ It is conceivable that the first self-replicating molecular system which was capable of evolution was not a nucleic acid, not RNA or DNA, but rather some molecule now biologically extinct, long since replaced by the more efficient self-replicating system involving the nucleic acids. But no evidence for such a molecule has been suggested, and few people support this view. On other planets, other molecules besides nucleic acids may be fundamental for self-replication, but terrestrial life, our only present example of life, seems oriented around nucleic acids and proteins.

▽ Once the problem of the interaction between primitive nucleic acids and primitive polypeptides is solved, it will be possible to state fairly that life has been synthesized in the laboratory. Not, of course, anything familiar, like an aardvark or an axolotl; merely a molecular system capable of self-replication, mutation, replication of its mutations, and some degree of environmental control. If we understand how such a molecular system came into being, we will have begun to understand the long evolutionary chain from the gases and waters of the primitive Earth to the origin of man.

▽ The laboratory synthesis of life, at least in the sense of a molecular system *capable* of evolution by natural selection, may be proved in a decade; some say it has already been accomplished. But if this is all there is to the origin of life, some may object, would not a race of self-replicating robots be alive? Certainly. If we can imagine an environment littered with mechanical arms and legs, transistors, cryogenic apparatus, and whatever other parts we require for a versatile robot, even today a robot can be developed which will use these parts to construct another identical robot. It would need an inheritable set of instructions on how to construct other robots. To be capable of evolution, a random or accidental change in the instructions would have to be incorporated in the next generation. This is, of course, analogous to the biological modus operandi, and such an analog could be

constructed mechanically. As time went on, provided the supply of spare parts did not run out, we would have a great increase in the number and diversity of robots.

▽ Now imagine that the supply of a particular building block—say, an arm—becomes exhausted. All the arms littering the landscape have already been used for robots. What happens? Robot reproduction will then grind to a halt, unless there is a mutation to rework, say, mechanical legs, which are still available, into the much-needed arms. In time, even the supply of legs will be exhausted. If a mutation were developed to make legs from another commodity in abundant supply—say, scrap automobile engines—and then the legs converted into arms, the robots with such an adaptation would, of course, preferentially reproduce. In time, perhaps, we would have an adaptation in which the robots would directly mine iron ore, convert the ore into scrap automobile engines, the automobile engines into legs, and the legs into arms. The most efficient sequence would be directly from ore to arm, but the robots are trapped by their history. Steps can be added only one at a time.

▽ The robots would then have developed a reaction chain analogous to the enzyme-intermediated reaction chains of contemporary biological systems. It was first suggested by the American geneticist Norman Horowitz, of the California Institute of Technology, that the origin of biochemical reaction chains occurred in a way similar to our fanciful robot analogy—that chemical building blocks, organic molecules, were made essentially "free," in the primitive environment, and then utilized by the first living systems. As the number of these living systems increased, the demand for certain critical molecular building blocks exceeded the supply. Those organisms which could use another common and previously untapped molecule and convert it into the desired building block clearly had a distinct selective advantage over their neighbors unable to effect such a transformation. As each required molecule becomes progressively depleted, another step in a long reaction chain must be added. The Horowitz hypothesis provides a neat and elegant understanding of the origin of complex biological reaction chains.

▽ If we imagine an ocean full of a variety of nucleic acids, each organizing its own short but useful reaction chain, utilizing ATP made by solar ultraviolet light at no cost to the organisms, we can see that we are well along towards the development of biological complexity. If, by chance, an aggregation of such nucleic acids were localized together, the interaction between them might be advantageous to both. Several methods are known by which such physical association might be managed. Figure 17–3 shows a number of "coacervates." In experiments which may possibly simulate conditions in the primitive oceans, the Dutch chemist H. G. Bungenburg de Jong found, in the 1930's, that there often is a spontaneous synthesis of objects in the 1–100μ size range which are rich in colloidal organic matter on the inside, and clearly separated from the external environment on the outside. In some of these experiments, nucleic acids were concentrated in the interior of the co-acervate. △

A. I. Oparin believes that these coacervate droplets were, in essence, the first

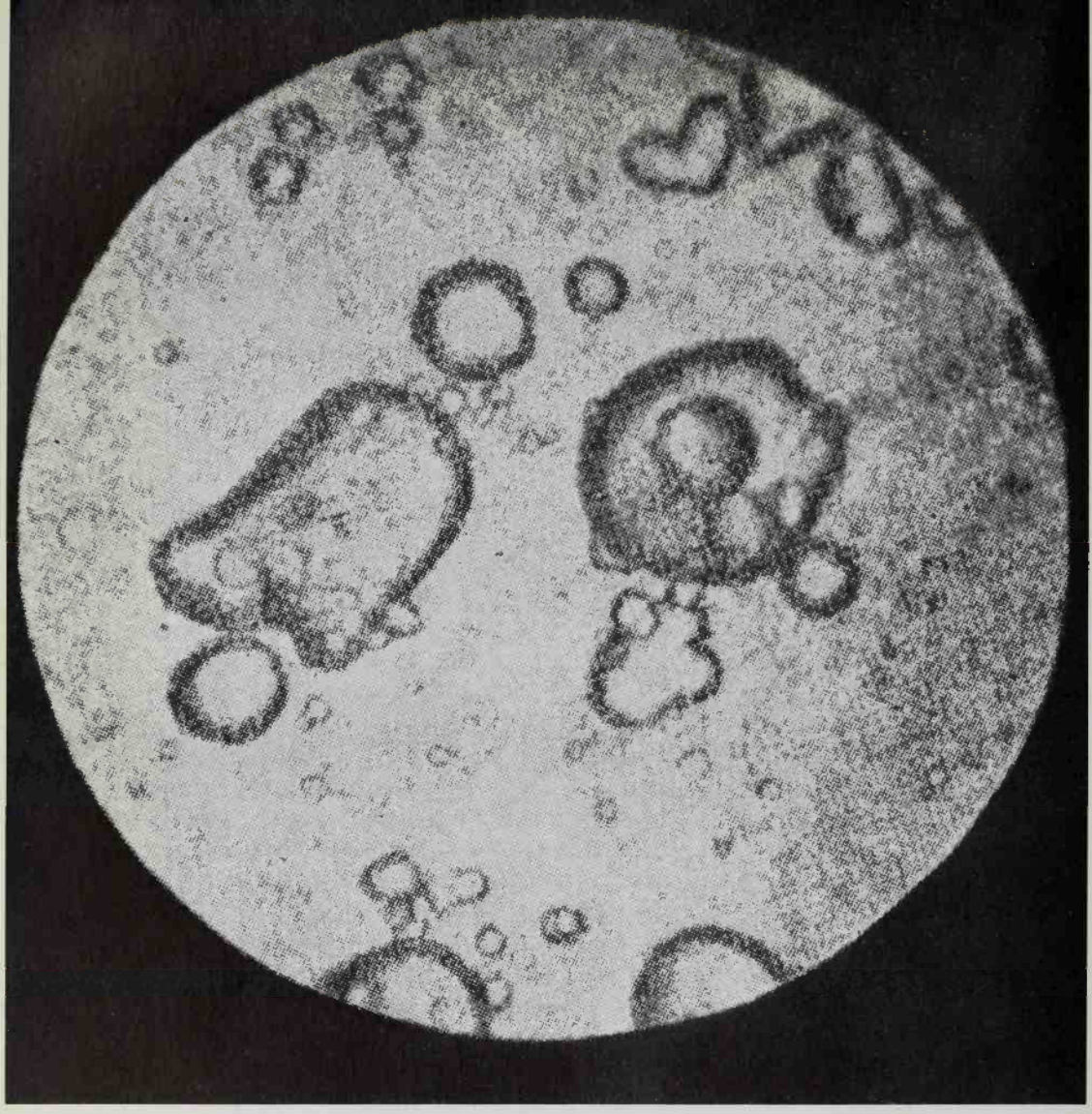


FIGURE *17–3. Examples of coacervate systems with three components: gelatin, gum arabic, and RNA. (Taken from a publication by Prof. A. I. Oparin.)*

forms of life on Earth. They have a number of interesting properties. In particular, they can adsorb into their structure various simple organic molecules from the surrounding medium. Oparin believes that this is an elementary form of metabolism, an all-important property of life. He also states that the coacervate undergoes a process analogous to natural selection, which he describes in the following words:

> The coacervate droplets which were formed in the waters of the Earth found themselves immersed in a solution containing various organic materials and inorganic salts. These substances were adsorbed by the droplet, and entered into

chemical interactions with it. Thus, organic synthesis occurred. But in parallel was the process of decomposition. The rate of each of these processes, and of other processes, depended on the internal organization of the individual coacervate droplet. For a relatively long period of time, only those droplets could survive which possessed sufficient dynamic stability so that their rate of synthesis exceeded their rate of decomposition. If the rate of decomposition exceeded the rate of synthesis, those particles disappeared; such "poorly" organized particles played no role in the further development of living material.

We find it difficult to agree with Oparin that these coacervate droplets were the first forms of life on Earth. While the analogy between material exchange and metabolism is interesting, it hardly proves that coacervates were primitive living organisms. A fundamental property of living systems is self-replication, including the presence of a genetic code which transfers properties from generation to generation. Coacervates have no mechanism of inheritance. Oparin's hypothesis does not explain the transition from non-living to living systems.

▽ A possibly more relevant model of pre-cellular molecular enclaves are the microspheres of Sidney W. Fox [Figure 17–4]. The microspheres are made by heating and cooling synthetic polypeptides. The microspheres are far less complex than the superbly architectured bacterial cells [see Figure 14–1] which they resemble superficially; however, microspheres have considerably more stability than do the coacervates. They therefore provide some reason to believe that local enclaves of organic matter in the primitive oceans may have been common. We do not know in detail how such an enclave may have evolved into the contemporary cell, with its elaborate reproductive choreography. One promising possibility is that simple free-living organisms aggregated together into a loose cooperative arrangement, which slowly evolved into a smoothly interacting whole. It has recently been found that such cytoplasmic organelles as the chloroplasts (controlling photosynthesis), the mitochondria (controlling respiration), and the small bodies at the base of flagellae [see Chapter 14] all have their own DNA, different from that of the cell nucleus—suggesting their independent origins. The evolution of the cell clearly requires a long period of natural selection. But as we have seen in Chapter 16, the geological and paleontological evidence suggests about a billion years between the origin of life and the origin of the first cells.

▽ A characteristic property of molecules used in biological systems is that they tend to be asymmetric. What does molecular asymmetry mean? Suppose we were constructing gloves. We have material for the palm and the back of the hand, for the thumb, and for the four fingers. There are two ways of putting this material together. We can make a right-handed glove or a left-handed glove; the two kinds are not equivalent, as is evident when you try to put a right-hand glove on your left hand. Similarly, organic molecules, which fill three dimensions, are constructed in a variety of asymmetric and non-equivalent ways. Molecular asymmetry can be detected by optical rotation. If a beam of plane-polarized light is passed through a solution containing an asymmetric molecule, the plane of polarization will be rotated. If the plane of polarization is rotated towards the right, the molecule is

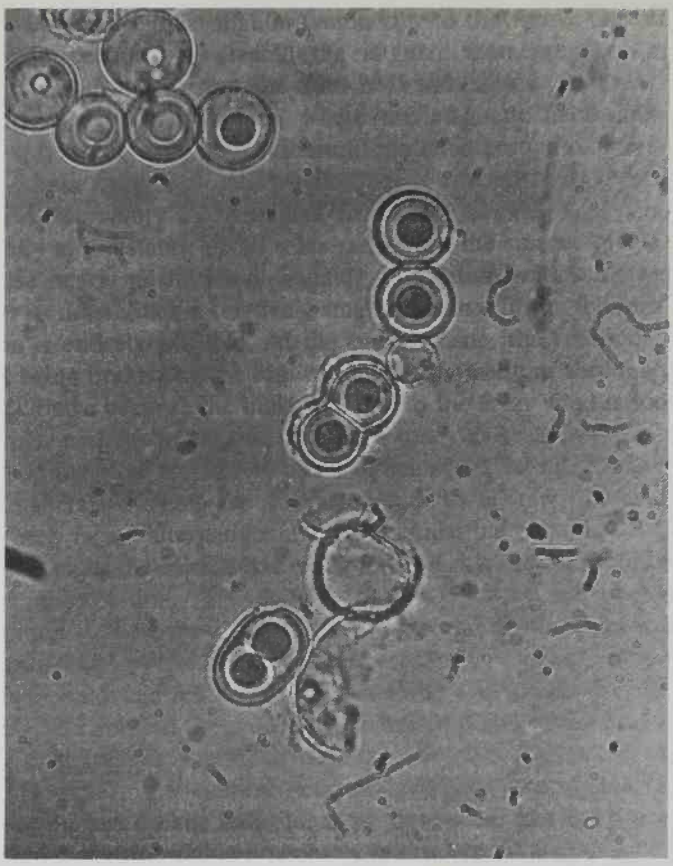


FIGURE 17–4. *Examples of microspheres of synthetic polypeptides produced by Prof. S. W. Fox and Dr. S. Yuyama, University of Miami. The twin aspect of some of these microspheres resembles cells in the process of division; however the microspheres do not always tend spontaneously to division. These twin forms were generally produced by pressing down on single microspheres with a microscope slide cover glass. (Courtesy of Prof. Fox.)*

said to have right-hand asymmetry, or to be "dextro-rotary" (abbreviated D). If the plane of polarization is rotated towards the left, the molecule is said to have left-hand asymmetry, or to be "levo-rotary" (abbreviated L). In either case, we say that the molecule is optically "active." Optical activity is therefore a measure of molecular asymmetry. No optical rotation at all can occur if the molecules in the solution are symmetric, or if there are equal numbers of levo-rotary and dextro-rotary molecules. A pair of molecules composed of just the same atoms but having opposite symmetries, like a left-hand and a right-hand glove, are known as "stereoisomers." A mixture of equal numbers of levo-rotary and dextro-rotary stereoisomers is known as a "racemic" mixture.

▽ In 1848, Louis Pasteur investigated the difference between tartaric acid, a dextro-rotary molecule, and racemic acid, a molecule which has the same apparent

structure as tartaric acid—that is, the same atoms put together in the same way—but which is not optically active. In a brilliant series of experiments, Pasteur found that racemic acid was a mixture of the dextro-rotary tartaric acid and a previously unknown levo-rotary tartaric acid, which rotated the plane of polarization towards the left. Pasteur was able to separate microscopically the levo-rotary from the dextro-rotary crystals of racemic acid, and show that the two parts rotated the planes of polarization in opposite directions.

▽ We might expect that molecules used in living systems are racemic, that there are as many dextro-rotary as levo-rotary isomers. This is not the case. The molecules of living systems are characteristically optically active. In nature, we find only the D form of glucose. The cell walls of bacteria contain only the D isomer of amino acids, and the enzymes inside bacteria are made up only of the L amino acids. Why is optical activity such a common property of biological molecules? The enzymes in intermediate metabolism have a high degree of specificity; they catalyze only one set of reactants and no other. In fact, they often are able to distinguish between two stereoisomers, metabolizing, for example, D tartaric acid and not L tartaric acid. This has suggested that the enzymes work by forming a three-dimensional structure with the reactants in a kind of lock and key arrangement that brings the reactants together; then the enzyme goes in search of more reactants. Substantial evidence in support of this lock-and-key model of enzyme activity has appeared in recent years.

▽ For an organism to have enzymes which make one set of reactions work and not another, it must have a high degree of steric selectivity—that is, the enzyme must somehow be capable of distinguishing among various three-dimensional configurations of similar molecules. The American biochemist Lubert Stryor, of Stanford University, has pointed out that such steric selectivity for enzyme action will, in time, demand that a distinction be made among stereoisomers. We expect that a biochemistry without stereoisomerism is a primitive one, at best. Thus, the search for optical activity will be an important tool in any explorations for extraterrestrial life. If on Mars, for example, we find racemic mixtures of organic compounds, we will be tempted to conclude that biological evolution has not proceeded very far. If we find optical activity, it will be much more exciting, particularly if stereoisomers are found there which are different from the ones found here.

▽ But how did optical activity arise? In the experiments performed by Miller, Urey, and their successors, the synthesized organic molecules form a racemic mixture. Are there any mechanisms for abiological generation of asymmetric molecules? It is known that photochemical reactions involving polarized light can yield optically active products from racemic precursors; or a catalyst which is optically active, such as a quartz crystal, may yield optically active products; or finally, there may be a spontaneous reaction in the absence of optically active factors which nevertheless produces an optically active product. But none of these mechanisms can explain the origin of optical activity in biochemistry, because in the large each stereoisomer of a pair should be produced in equal amounts. The

amount of left-hand polarized light striking the surface of the Earth is balanced by right-hand polarized light; the amount of left-asymmetric quartz equals the amount of right-asymmetric quartz; and the extent of spontaneous synthesis of levo-rotary compounds should be exactly balanced by the rate of synthesis of dextro-rotary compounds. It seems highly probable that the organic molecules synthesized on the primitive Earth at the time of the origin of life were not on the average optically active.

▽ Conceivably, optical activity in biochemistry is the result of natural selection. As enzyme systems and biochemical reaction chains developed, the three-dimensional specificity of the lock-and-key arrangement of enzymes and reactants must have improved. After a time, enzymes must have been capable of distinguishing levo-rotary from dextro-rotary stereoisomers. Let us imagine two organisms, one of which synthesizes L amino acids, and one of which synthesizes D amino acids, both from simpler precursors. Suppose, now, that for a reason entirely unconnected with amino acid stereoisomerism, the L amino acid synthesizer was slightly better adapted to its environment than the D amino acid synthesizer. After some generations have passed, the L amino acid synthesizers should dominate the biological landscape; after a while, the D amino acid synthesizers may become extinct. The descendants of this organism will continue to synthesize L amino acids, not because L amino acids have any intrinsic merit over D amino acids, but rather because L amino acids have been woven early into the fabric of life. If the prevalence of L amino acids in contemporary enzyme systems is such an historical accident, then the chance of finding D amino acids in any extraterrestrial enzymes should be 1 in 2. If, on the other hand, the investigations of extraterrestrial life forms show that L amino acids are predominant everywhere else, we will be forced to revise our beliefs on the origin of optical activity.

▽ We have, in this chapter, been discussing some of the chemical problems in the origin of life. Much of what we have said is speculative, for the fundamental reason that none of us was on the Earth at the time life arose. The best we can do, for the moment, is to invent a likely story and maximize its plausibility by laboratory investigation. But only by investigating living systems elsewhere will we be able to check our story with some rigor. Are extraterrestrial forms composed primarily of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulphur? How do they reproduce? Is the genetic material composed of nucleic acids? Are proteins their molecular catalysts? Are they arranged into cells? Are their molecules dextro-rotary or levo-rotary? Or racemic? We are on the threshold of these discoveries.

▽ The later evolution of life, by natural selection in response to the challenges of the environment, is increasingly well-documented in the fossil record. There is a general trend, as time progresses, towards enhanced complexity. Yet there is no reason to suspect any urge or desire towards complexity by the evolving organisms. We have seen that those mutants in the primitive environment which led to improved survival and replication of nucleic acids prospered; other molecular systems perished

in enormous numbers. There is, accordingly, a sense in which the evolution of the cell, and all subsequent evolution up to man, may be viewed as a device for maintaining the continued survival of the nucleic acids. There is a sense in which our instincts and desires, our loves and hates, our breathing, eating, sleeping and dying exist because they help ensure the continued existence of the molecules of our genetic material; a sense in which we are fundamentally ambulatory repositories for our nucleic acids. Whether we like it or not this is at least in part what human beings are for; because of our intelligence we are more than this—but it is an open question how much more. △

18

Is there life on Earth?

We approach now the only planet in which man is certainly known to exist, and which ought to have an interest for us superior to any which we have yet seen, for it is our own. We are voyagers on it through space, it has been said, as passengers on a ship, and many of us have never thought of any part of the vessel but the cabin where we are quartered. Some curious passengers (these are the geographers) have visited the steerage, and some (the geologists) have looked under the hatches, and yet it remains true that those in one part of our vessel know little, even now, of their fellow-voyagers in another. How much less, then, do most of us know of the ship itself, for we were all born on it, and have never once been off it to view it from the outside!

Samuel Pierpont Langley, *The New Astronomy* (1891)

▽ The origin of life seems to be an incidental adjunct to the early development of a planetary surface. We have seen that only very general conditions—a reducing atmosphere of approximately cosmic abundance, and bodies of liquid water—are required for the large-scale production of complex organic molecules. Judging from the history of the Earth, if such conditions prevail for only a few hundred million years, the origin of life seems to be probable. For all we know, much shorter periods of time are adequate. △

It is believed that during the early days of the formation of our solar system, many of the physical and chemical conditions on the terrestrial planets (Mercury, Venus, Earth, and Mars) were similar. These planets, it is assumed, were all formed from the same solar nebula of gas and dust, and their early chemical compositions were very nearly identical. Thus, we would expect that the conditions which brought about the origin of life on Earth would also have been present on the other terrestrial planets.

▽ However, there are other factors to consider besides a common initial environment. If the surface temperatures are too high, common organic molecules will be thermally dissociated as fast as they are produced; and no liquid medium will be available as a solvent for early chemical interactions, and as a shield against primitive ultraviolet radiation. If surface temperatures are too low, familiar chemical reactions will proceed at insignificant rates; and any liquid medium will freeze and hence be unavailable. The freezing point of water can be lowered by adding salts to the solution. Thus, a reasonable range for liquid water as the medium for a living system, and for a reasonable stability and rate of reaction of familiar organic chemicals is between $-50°C$ and $+100°C$. Since the temperatures on the bright side of Mercury, for example, are much higher than this, there is some basis for a provisional exclusion of life there.

▽ Another factor is the temperature of the exosphere, the level from which molecules escape to space. If the exosphere temperature is very high, then the rate of escape of a planetary atmosphere would be very high. Its reducing atmosphere would be retained for a very short period of time, and there would be an inadequate period available for the origin of life. We have already discussed this point in Chapter 15, in connection with panspermia developing on possible planets of other stars. Mercury, again, has such a low mass and high exosphere temperature that any primitive reducing atmosphere which it may once have had departed into space very early in its history. In low-temperature environments, more exotic biochemistries may be possible. Liquid ammonia or hydrocarbon solutions may replace water as a solvent sysem, and silicon compounds might replace carbon compounds as structural biochemicals.

▽ Our knowledge of planetary environments permits us provisionally to exclude Mercury and the surface of the Moon as possible abodes of life; probably, also, Venus, the asteroids, and most of the other moons in the solar system. But in applying such a priori negative judgments, we must be very careful that we are not deceived by terrestrial analogy. On other worlds there may be chemistries and living systems which we cannot even imagine. The best approach is observational, not deductive.

▽ How easy would it be to detect, from a remote observation platform, living systems on the Earth? The mass of the Earth is 6×10^{27} gm; the mass of the atmosphere is 5×10^{21} gm. Yet the mass of biological material on the surface of the Earth is only a few times 10^{17} gm, according to the best recent estimates; less than 0.0001 percent of the mass of the *air,* and some 10^{-8} percent of the mass of the Earth. Thus, for all our feelings of self-importance, we are only a kind of biological rust, clinging to the surface of our small planet, and weighing far less than the invisible air which surrounds us. Yet we have tamed and reworked the surface of our planet, altered its character, and are in the process of leaving it for distant parts. Are our activities, obvious to us, noticeable from a distant vantage point? Would our presence be detected? △

To appreciate the situation in which the Earthbound astronomer finds himself, let us imagine that we are Martian astronomers in a Martian observatory. Our equipment includes the most modern astronomical instruments currently available on Earth. From our splendidly equipped observatory, let us ask the question: Is there life on Earth?

The planet Earth, as seen in the Martian sky, would appear as a very bright star, only slightly less brilliant than Venus appears from the Earth. Just as we can see the planet Venus going through phases like those of the Moon, so could the hypothetical Martian astronomer observe the phases of the Earth. ▽ Since the Earth would appear in the Martian sky, at a greater angle from the Sun than Venus does in ours, it would be easier to observe the Earth from Mars than Venus from the Earth. The Earth would appear as a morning or evening "star," low in the Martian sky. Because of the phases, it would be impossible to see a place on the Earth near the middle of the day, local time, except when the Earth was a great distance from Mars, on the other side of the Sun. △

Could the engineering works of men be observed from a Martian observatory? Dams, reservoirs, cities—would they be detectable? ▽ Because of the turbulence in the Earth's atmosphere, even the largest of our telescopes, the 200-inch Hale reflector at Mt. Palomar, California, is capable of photographing detail no smaller than about 300 km across on Mars. The Martian atmosphere is much thinner than Earth's, and it is possible that resolution by an observatory on Mars would be less limited by the Martian atmosphere. The smallest feature on Earth visible from Mars might be only kilometers across.

▽ The Earth has by now been photographed many times from space. A systematic program for photography of the Earth has been undertaken in the United States in the Tiros and Nimbus series of satellites, to chart cloud formation,

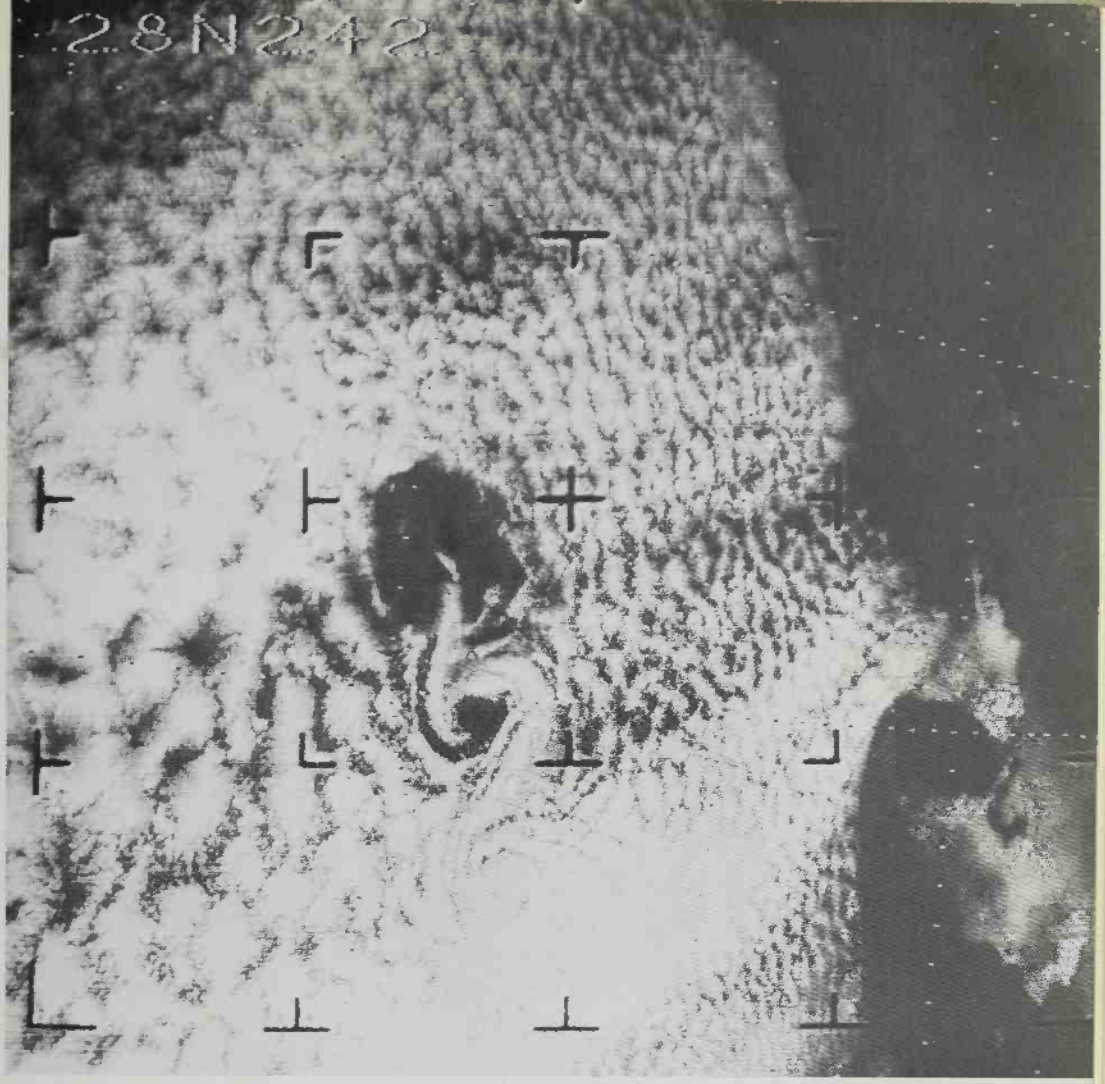


FIGURE *18–1. Nimbus 1 photograph of a cloud eddy pattern over Guadalupe and Baja California, 14 September, 1964. (Courtesy of Goddard Space Flight Center, NASA.)*

movement, and dissipation, in an attempt to improve weather prediction. While the cameras on such satellites are not designed for life-detection, we may use them for such purposes. Photographs of the Earth taken from Tiros and Nimbus satellites sometimes show cloud covers as patterned and provocative as that of Figure 18–1. When there are breaks in the clouds, views of the Earth's surface, such as those in Figures 18–2 and 18–3, can be obtained. Figure 18–2 shows the eastern seaboard of the United States from Chesapeake Bay to Cape Cod. Figure 18–3 shows the southern tip of India and the island of Ceylon. The regions depicted in these photographs are among the most heavily populated and densely vegetated areas of the Earth; yet even close inspection shows no sign of life at all. New York appears deserted; India and Ceylon appear barren. These conclusions

FIGURE 18–2. *Tiros 7 photograph of the eastern seaboard of the United States, 23 June, 1963. (Courtesy of Goddard Space Flight Center, NASA.)*

have been repeated hundreds of times in close investigation of Tiros photographs of populated regions of the Earth: when the resolution is no better than a few kilometers, there is no sign of life on Earth.

▽ Altogether, there have been several hundred thousand photographs taken and examined in the Tiros series. On some of these photographs, objects as small as 2000 feet across are discernible. Yet in all of these photographs, only one—Figure 18–4—shows any clear sign of life on Earth. This is a photograph taken by Tiros 2 of the forest near the Canadian logging town of Cochrane, Ontario, on 4 April, 1961. In the upper left part of the picture, several wide parallel stripes can be seen; and at right angles to them, another set. Swaths one mile across had been cut through the Canadian forest by the loggers. The swaths were separated by about two miles. After the swaths had been made, snow fell, enhancing the contrast between the trees still standing and the treeless swaths. But even here, on this one-in-a-million photograph, do we have unambiguous signs of life, from the vantage point of Mars? Can the Martians imagine no geological process which could make such a pattern? Even here, with resolutions better than

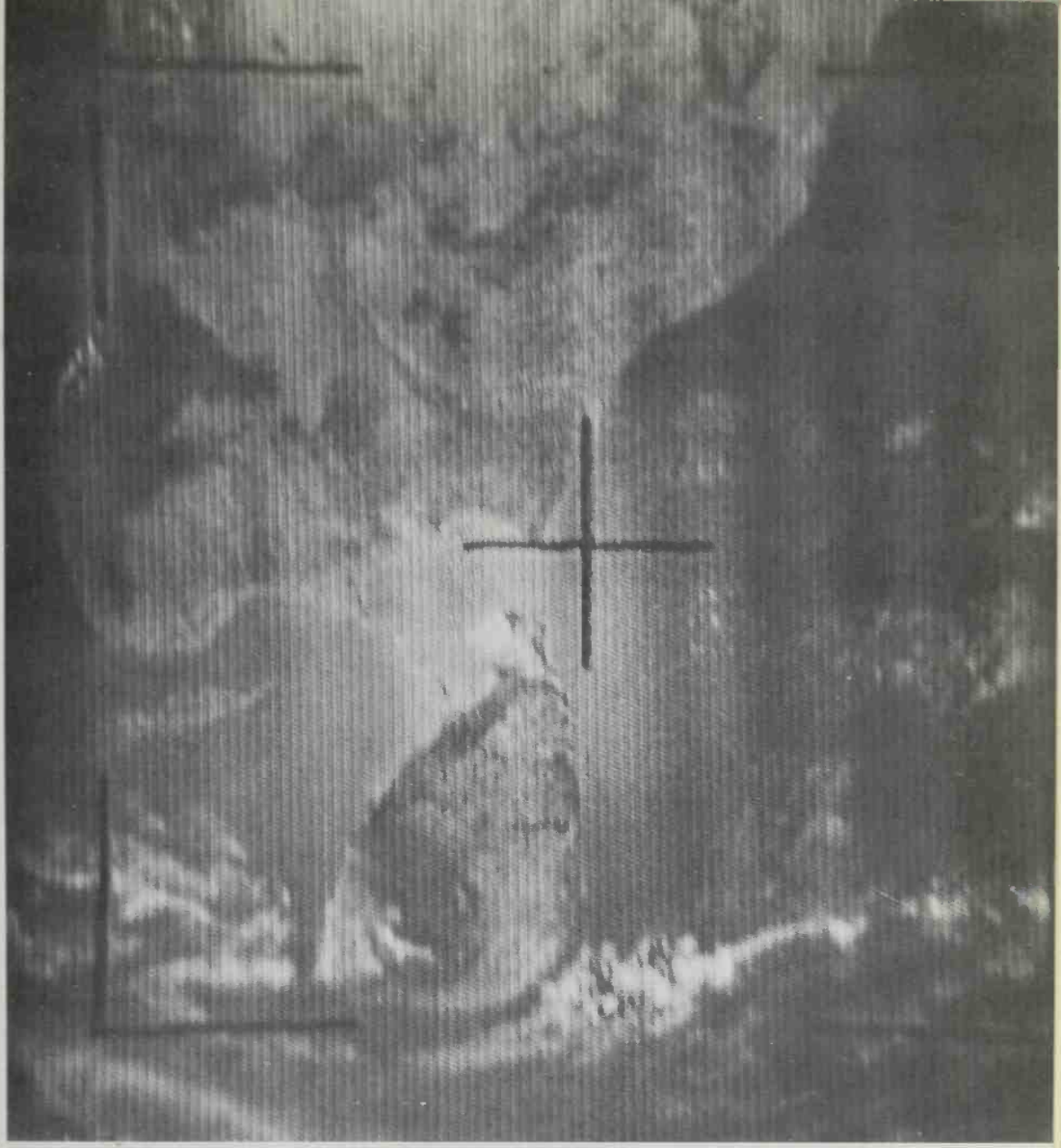


FIGURE 18-3. *Tiros 5 photograph of southern India and Ceylon, 6 March, 1963. (Courtesy of Goddard Space Flight Center, NASA.)*

the Martians could reasonably have, there is no rigorous proof of life on Earth.

▽ My colleagues and I have made a study of the higher resolution photographs available from the Nimbus satellite. With resolutions of a few tenths of a kilometer, we have discovered a recently completed highway in Tennessee, perhaps a jet contrail in the Davis Straits, the wake of a ship in the Red Sea, but also a very straight feature off the northern coast of Morocco which had all the apparent signs of intelligent design, but was, in fact, a natural peninsula. At a few tenths of a

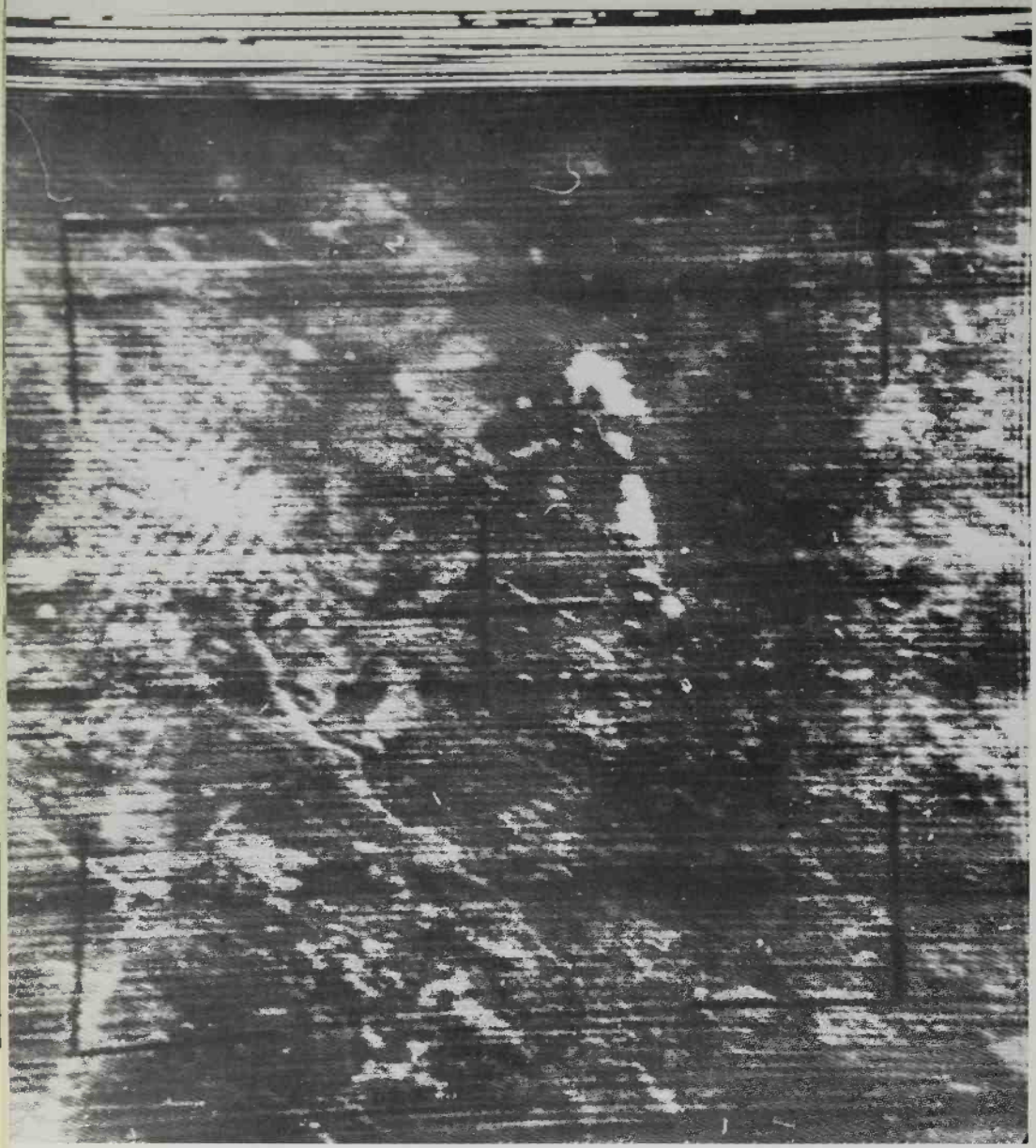


FIGURE 18-4. *Tiros 2 photograph of the region of Cochrane, Ontario, Canada, taken on 4 April, 1961. (Courtesy of Goddard Space Flight Center, NASA.)*

kilometer resolution the signs of intelligent life on Earth can be detected but not unambiguously. Convincing photographic evidence of intelligent life on Earth requires resolution of 10 meters or better. △

Could the night-time illumination of the largest cities of the Earth—New York, Moscow, Tokyo, Paris, London, Chicago—be detected? Let us assume that the artificial illumination from one of our largest cities is, on the average, ten times

greater than the illumination it receives from the full moon, and is confined to a region 10 km square. Then there is a bare possibility that the Martian astronomer observing the night hemisphere of the Earth could see a tiny speck of light, of about 16th magnitude. In reality, however, due to the scattering of sunlight from the illuminated hemisphere of the Earth, the Martian astronomer would be at best only marginally able to detect such a weak signal. ▽ Another factor which tends to make our largest cities invisible from above is smog. It seems that whenever a city becomes large enough so that its night-time illumination might be observable from Mars, the city also generates enough industrial pollution, even at night, so that the city becomes invisible. The American astronaut M. Scott Carpenter was able to observe mountain trails and smoke from chimneys when over Tibet; but when orbiting over Southern California, he could find no sign of the city of Los Angeles. △

Nuclear explosions which unfortunately sometimes occur on the planet Earth could be seen from Mars as short-lived, very bright flashes of light. Nevertheless, since nuclear weapons tests occur only infrequently, and since the resulting flash is itself visible for only a brief moment, it is highly improbable that such explosions would be detected from Mars. If a special program were initiated for synoptic observations of the Earth, perhaps nuclear explosions could be observed reliably. However, it seems unlikely that the civilized Martian astronomer could deduce from these short-lived flashes of light that life—to say nothing of intelligent life—existed on Earth. Even we who live on Earth can hardly consider these barbarous experiments, which could lead to the destruction of life on our beautiful world, as manifestations of intelligence!

With an optical telescope, an astronomer on Mars might be able to detect seasonal color variations over a vast area of the Earth's surface. ▽ There are major seasonal color and brightness changes in deciduous forests, and in areas with cultivated crops, such as the Ukraine, or the American Midwest. △ With these observations in hand, a wide range of explanations could be imagined. ▽ Perhaps there are crystals in some regions of the Earth whose color depends on temperature, or whose darkness depends upon humidity. Or perhaps they are due to life of some sort on Earth. △ But it seems unlikely that the Martian astronomer could *reliably* conclude that the seasonal color variations are of biological origin.

If the Earth were regularly observed over a period of several decades, major transformations of the surface might be noticed—for example, the systematic destruction of the forests. But could the Martian astronomer draw definitive conclusions from these observations? ▽ Similar major and systematic "secular" variations can be observed on the surface of Mars, as we shall see in Chapter 20. △ In themselves, such variations are very interesting, but certainly cannot be considered irrefutable proof of the presence of life. A number of such variations have been reported for the moon (although they were on a smaller scale), but the surface of the moon is almost certainly devoid of life.

▽ Extensive spectroscopic measurements of the Earth could be made from a Martian observatory. In a search for life, spectral bands in the infrared, which are

due to absorption by surface organic matter, might be sought. But unfortunately, such bands at 3.5μ and longer wavelengths must, for existing equipment, be seen in reflected light to be detected. The light transmitted from the Earth to Mars in wavelengths longer than 3.5μ is mostly infrared energy radiated from the Earth, and not sunlight reflected from it. The Martian astronomer would find it difficult to detect spectroscopic signs of surface organic matter.

▽ Some minor atmospheric constituents which are of biological origin, such as CH_4 and N_2O, might be identified. Methane is a highly reduced gas, and must be continually produced in the Earth's atmosphere, so that the total amount is not depleted by oxidation. The methane in the Earth's atmosphere is produced primarily by methane bacteria, which convert organic compounds into CO_2 and CH_4. The methane bacteria live in mud at the bottoms of ponds, where there is much organic matter and the conditions are anaerobic. Hence, CH_4 is often called "marsh gas." Similar bacteria also live in the rumens of cows and other ungulates. Accordingly, one of the major sources of methane in the terrestrial atmosphere is bovine flatulence. The identification of methane in the Earth's atmosphere from a Martian vantage point would thus be a very significant observation, if only the Martian astronomer knew how to interpret it. But it seems unlikely that the appropriate interpretation would be forthcoming.

▽ What about methane in the atmospheres of the Jovian planets? The Soviet astronomer G. A. Tikhov proposed that methane on Jupiter has the same source as methane on Earth, leading to the conclusion that there must be at least Jovian bacteria, if not Jovian cows. Since we have seen [Chapters 4 and 16] that methane is a constituent of primordial planetary atmospheres, we need not take this suggestion very seriously. It does, however, emphasize the difficulties in connecting the presence of a simple molecule with biological activity. △

Nearly all of the free oxygen in the Earth's atmosphere is a product of plant photosynthesis. The main source of oxygen is not the higher plants, but rather marine plankton, which fill the oceans. ▽ The crust of the Earth is underoxidized —that is, it is capable of further chemical reaction with atmospheric oxygen. △ If it were not for the continuous production of oxygen by biological activity, it would vanish from the atmosphere within a relatively brief span of years. If the amount of free oxygen in a planetary atmosphere is so small that it can be detected only at the very limits of instrumental sensitivity, then its presence might be explained by abiological hypotheses. But such a vast amount of oxygen as is present in the Earth's atmosphere can be explained only in terms of extensive biological activity. ▽ Yet we can make two reservations here. It is conceivable that an extensive oxygen atmosphere could be produced by photodissociation of water [see Chapter 16]. In addition, I wonder whether an intelligent anaerobic organism, who finds oxygen a poison gas, would conclude very readily that an extensive oxygen atmosphere can only be the product of biological activity. △

▽ If Martian astronomers had an instrument which permitted very sensitive examination of the visible spectrum of the Earth at one wavelength of light, they

would have observed an apparent major increase in the abundance of such gases as neon, argon, mercury, and sodium in the spectrum of the night sky of the Earth over the past few decades. Whether they would attribute this change to instrumental error, improvements in terrestrial illumination engineering, or an imminent catastrophe is a matter of conjecture.

▽ Routine spectroscopic measurements of the Earth would reveal the presence of quantities of oxygen and water on the Earth that are enormous, especially when compared with Mars. Our temperatures would seem uncomfortably high, and the absence of ultraviolet light from the surface would be noted. It is quite possible that the Martian scientists, arguing from Martian analogy, would conclude that with no convincing evidence for life on Earth, and with such an unpromising environmental inventory, further searches for life on Earth should be abandoned. △

However, there is another method which might be used to detect life on Earth. Let us assume that there are Martian observatories equipped with a modern radiotelescope—an instrument which permits us to detect, measure, and record the radio emission from various celestial objects. The Martian astronomer, like his Earthly counterpart, would investigate the radio emissions of the planets. ▽ He would find that Venus is a radio source, probably because its surface is hot; that Jupiter is a radio source, because the electrons in its magnetic field are emitting synchrotron radiation; and so forth. △ Turning his radiotelescope to Earth, however, he would make an amazing discovery: at meter wavelengths, the otherwise unassuming planet Earth emits almost the same power of radio radiation as does the sun in a period of low sunspot activity! ▽ A planet as bright as a star! △ In the meter wave band, the Earth radiates a million times more radiation than Venus or Mercury. This discovery could be made on Mars by using only a modest radiotelescope.

Further investigation would show that different regions on the surface of our planet radiate unequally; a periodic relation would be found between the intensity of radio emission and the rotation of the Earth on its axis. For example, when Africa or South and Central Asia were facing Mars, the radio intensity would fall sharply; when Europe and North America were facing Mars, the power emitted would sharply increase. If observations had been continued over a long period of time, the Martian astronomer could make an even more astounding discovery—today the Earth is emitting radio radiation which is 10^6 times more intense than what it emitted a few decades ago. ▽ The Martian scientists would perhaps attempt a "natural" explanation for the phenomenon; such attempts would eventually prove unsuccessful. △ The clever Martian astronomers would come to realize that the radio emission could not be explained by the action of natural forces, but could only be produced by artificial means. They would conclude that intelligent life exists on Earth—a remarkable discovery indeed.

Several thousand television transmitters exist on the Earth. If we take into account the average power of each transmitter (approximately 20 kilowatts), the frequency bandwidth emitted, the average operating period of each transmitter

(say, 6 hours out of every 24), and the fact that all wavelengths in television transmission (1.5–6.0 meters) pass unobstructed through the atmospheres of the Earth and Mars, we can calculate the power transmitted from Earth to Mars.

Radio astronomers may be interested to know that the so-called "brightness temperature" of the Earth at television wavelengths is some hundreds of millions of degrees. This is 100 times greater than the radio brightness of the sun at comparable wavelengths, during a period of low sunspot activity. In addition to television transmitters, there are a large number of radio stations, and other installations which emit radiation strongly in the ultra-high frequency wavelength range. ▽ The United States' Ballistic Missile Early Warning System (BMEWS) would, some years ago, every now and then, detect the moon on its radar screens.

▽ We have presented this fantasy of a Martian observatory investigating the Earth because it illustrates the actual difficulties and potential triumphs of remote investigations of planetary biology. If the hypothetical Martians could not find any sign of life on Earth except at radio frequencies, it should not surprise us that unambiguous, indisputable, rigorous evidence for life on Mars is not yet forthcoming. Such searches as have been made for intelligible radio transmission from Mars have yielded entirely negative results. The radio emission from Mars is the random noise of thermal emission. △

In the example of radio emission from Earth, we have encountered for the first time, the cosmic implications of the biological activity of intelligent beings. Due to the presence of a technical civilization on our planet, there has been a drastic modification of an important feature of the Earth as seen from afar—the nature and power of its radio emission. ▽ Arduous work by an extraterrestrial astronomer could probably convince him that the signals have intelligible content (despite the quality of many television programs). △ The Earth has become strikingly different from all the other planets in our solar system. An essential attribute of intelligent life is that sooner or later its activity will attain a cosmic character. In Part III of this book, we shall elaborate this possibility.

Does the absence of strong, nonthermal, and intelligible radio emission from Mars imply, in itself, that there are no highly developed life forms on that planet? Generally speaking, no. Much of the radiation associated with television transmission is dissipated into space. Perhaps the radio emission from Mars is not sufficiently powerful to reach the Earth. It is natural to assume that as technical civilizations become more advanced, less wasteful methods of utilizing electromagnetic energy will be devised. Electromagnetic waves will probably be focused into tight, discrete beams, and scattering of this energy away from the intended source will be minimized. ▽ Thus, if a Martian civilization is somewhat more advanced than our own, it may have devised economical means of electromagnetic communication which do not permit eavesdropping from Earth. Nevertheless, if a civilization exists on Mars which is substantially in advance of our own, it is surprising that we have no sign of its existence (although if they have monitored *our* television transmission, perhaps we have some clue to their absence!).

▽ The American radio astronomer Frank Drake of Cornell University has

pointed out that no serious search for narrow band radio transmission from Mars has actually been carried out. Mars has been observed with broad-band receivers, to measure its subsurface temperatures, but searches for intelligible signals have been carried out, at best, unofficially and unsystematically. On the other hand, Drake argues, the likelihood for success of such a program is probably small. If the Martians were as much as 50 years ahead of us, we should (with the reservations noted above) have had some other signs of their existence. If they are as much as 50 years behind us, they are incapable of radio transmission. These estimates are based on terrestrial analogy, and assume that the almost discontinuous recent advances in our technical civilization are characteristic of civilizations elsewhere. We certainly do not know that this is the case; on the other hand, it is the most reasonable assumption we can make—there are no known counterexamples! Since both the Earth and Mars have existed for 5×10^9 years, the probability of a successful search for intelligible Martian radio emission is $50/(5 \times 10^9) = 10^{-8}$, or a millionth of a percent. Thus, in budgeting no time for such investigations of Mars, the directors of radio observatories have probably chosen wisely. Yet because of the interest of the search, it will not be surprising if occasional moments are stolen informally, between observing programs, to peer, with an amalgam of wistfulness and hope, at distant Mars. △

19

The planet Mars

. . . I am apt to believe that the Land in Mars is of a blacker Colour than that of Jupiter or the Moon, which is the reason of his appearing of a Copper Colour, and his reflecting a weaker Light than is proportionable to his distance from the Sun. . . . His Light and Heat is twice, and sometimes three times less than ours, to which I suppose the Constitution of his Inhabitants is answerable.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

▽ We now come to discuss our enigmatic planetary neighbor Mars, which seems to provide the best opportunity, in the immediate future, for the study of extraterrestrial life. In this chapter we will discuss the physical environment of Mars and the possibility that life could have come into being in the ancient Martian past and survived to the present day. In the following chapter we will discuss the observational evidence which has suggested more directly that Mars may harbor life, and the experiments which have been planned for future programs of Martian exploration.

▽ Seen for the first time through a telescope, Mars is a disappointing sight. You see an orange-buff or ochre colored disk of varying brightness and flickering visibility, swimming erratically across the telescope's field of view. The evanescent, will-o'-the-wisp appearance of Mars is due to "seeing," erratic atmospheric motions near the base of the Earth's atmosphere which change the directions of photons journeying from Mars to Earth and thereby distort the image of Mars which we see in reflected sunlight. To obtain a better image of the Martian surface, we should construct our observatory at higher altitudes, leaving the bulk of the atmospheric turbulence below us. The best visual and photographic observations of Mars have been made from telescopes situated on high, usually isolated, mountain peaks. Among the best observatories for studying Mars are those in the American Southwest, and at the Pic du Midi, in the French Pyrenees. There we find that the image of Mars is much more steady, and we can make out surface details quite clearly. Yet Mars never appears so large that it fills the entire field of view of the telescope. Generally, it appears as a small orange disk, with an angular diameter no larger than a modest-sized lunar crater.

▽ If we carry out observations over a period of time, we find that some of the surface features are disappearing over the west edge, or "limb," of Mars, and that others are appearing over the east limb. We are observing the planet's rotation. In time, we see familiar features reappearing at the east limb; we have observed a complete rotation. Mars rotates once about its axis every 24 hours and 37 minutes, just 41 minutes longer than the period of the Earth's rotation. The rotation of Mars can be seen in parts A, B, and C of Figure 19–1. The feature in the center of the disk in view A is Sinus Meridiani, which has disappeared over the eastern limb of the planet in view C. By observing the rotation of Mars, we can demarcate the Martian equator, and therefore, the axis of rotation. The Martian axis of rotation is inclined about 24 ° to the perpendicular to the plane of its orbit. The inclination of the Earth's axis of rotation is 23½ °. Because the Martian period of rotation is similar to ours, Mars has a familiar day-night cycle. Since the axes of rotation of the two planets have such similar inclinations, the seasons on Earth and Mars have quite

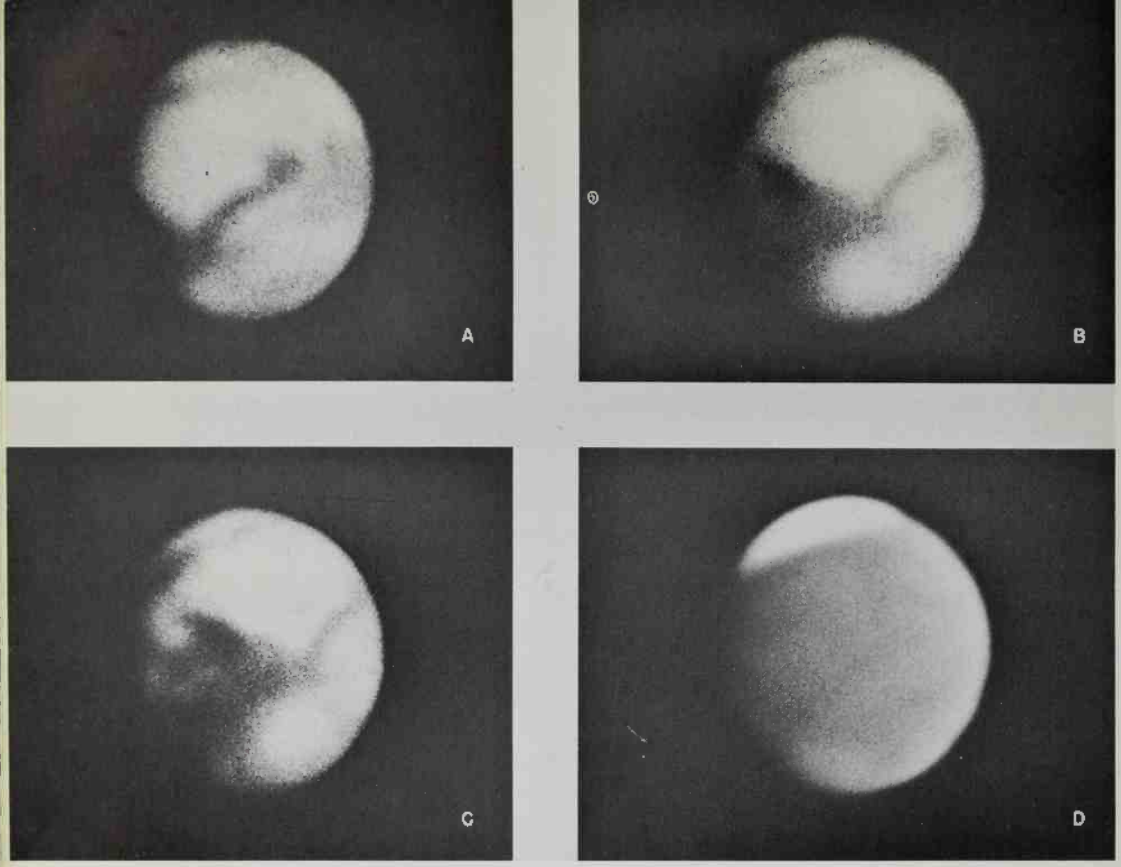


FIGURE 19-1. *Four photographs of Mars. Contrary to the usual astronomical convention, south is here at the bottom. Figures A, B, and C are taken in red light and illustrate the rotation of the planet. Figure D is taken in blue light, and illustrates the absence of surface detail in the blue and violet, a phenomenon known as the blue haze. (Courtesy of Mt. Wilson and Palomar Observatories.)*

similar forms. No one knows why the periods of rotation and the axial inclinations of Earth and Mars are so nearly identical. It may be merely a coincidence, or it may indicate some deeper connection between Earth and Mars, dating to the time of the origin of the solar system.

▽ Because Mars is about 50 percent further from the Sun, on the average, than is the Earth, its year is much longer—about 687 of our days. Thus, while winter lasts for the same fraction of the year on Mars as it does on Earth, its actual duration is almost 200 days—a long and, as we shall see, very cold winter indeed.

▽ Observing Mars under superior seeing conditions [Figure 19–2], we can make out three general regions: brilliant white polar caps; generally neutral gray dark areas, often concentrated in the equatorial regions; and the buff or orange ochre-colored bright areas, which give Mars its ruddy hue. The polar caps, which wax and wane with the seasons, are a form of thin and loosely packed water ice known as hoarfrost. This identification is based on a variety of independent lines of argument. Just as spectral absorption lines are formed when light passes

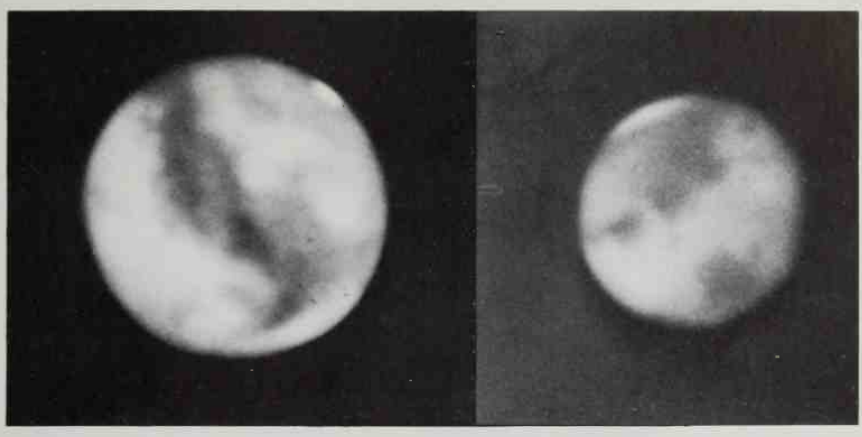


FIGURE 19-2. *Two photographs of Mars taken at the Observatoire du Pic du Midi in the 1940's. These pictures illustrate the wealth of surface detail which can be photographed under excellent seeing conditions. The photograph at left, taken during late spring on Mars, shows a very small southern polar cap. The triangular shaped region in the lower right is Syrtis Major. Above it and to its left are Mare Tyrrhenum and Mare Cimmerium. At the right, a photograph taken during late fall, we see a fairly extensive southern polar cap. The region to the bottom right is Mare Acidalium. Directly above it at the top of the picture is Solis Lacus. (Courtesy of Dr. A. Dollfus and the Meudon Documentation Center of the International Astronomical Union.)*

through an absorbing gas [see Chapter 4], so are absorption bands formed when light is reflected off a solid. In reflection, the light actually penetrates a small distance into the solid, and those wavelengths which correspond to the characteristic absorption wavelengths of the material are subtracted from the reflected light. Ice has a characteristic reflection spectrum in the near infrared which is matched by the near infrared reflection spectrum of the Martian polar caps.

▽ When sunlight is reflected off a solid, it tends to acquire a characteristic polarization, not as pronounced as in synchrotron emission [see Chapter 7], but nevertheless detectable. The polarization of sunlight reflected off the Martian polar caps is exactly matched by the reflection of sunlight off hoarfrost in the laboratory. Finally, the brightness of the polar caps is consistent with the brightness of hoarfrost. Thus, even a quick glance at Mars through a modest telescope under good seeing conditions shows the polar ice cap, and the polar ice cap means there is water on Mars. Since we have found water to be intimately involved with terrestrial life processes, this simple observation justifies some first hope that Mars may also have its own biology.

▽ Two stages in the regression of the northern polar cap of Mars can be seen in Figure 19–3, two drawings re-sketched as if we were looking down on the polar cap. The two views are about two Earth months apart, during northern Martian spring. We see that as the cap retreats towards the pole, small islands of hoarfrost are left behind. It has been suggested that these regions are elevations, which hold the hoarfrost longer because they are at higher and colder altitudes. One such locale,

where the hoarfrost is regularly left behind, is called the Mountains of Mitchell. However, elevations are not necessarily colder, on Mars, and these areas may be colder simply because they reflect more sunlight.

▽ From the rate of regression of the Martian polar caps in local spring, we can compute the thickness of the caps. We know how much sunlight is striking the caps, and how much of it is absorbed by the ice. This sunlight tends to heat the ice and make it evaporate. A very thick ice cap would appear to shrink very slowly; a thin ice cap, rapidly. For the amount of sunlight available, the ice cap regresses at a rather rapid rate; this permits us to conclude that the thickness of the cap is generally about a centimeter or less. All the water in the ice cap, if melted, would perhaps fill the Great Lakes of North America. Thus, while the caps indicate the existence of water on Mars, they do not point to very large amounts.

▽ Actually, there seems little chance that lakes of pure water exist on Mars.

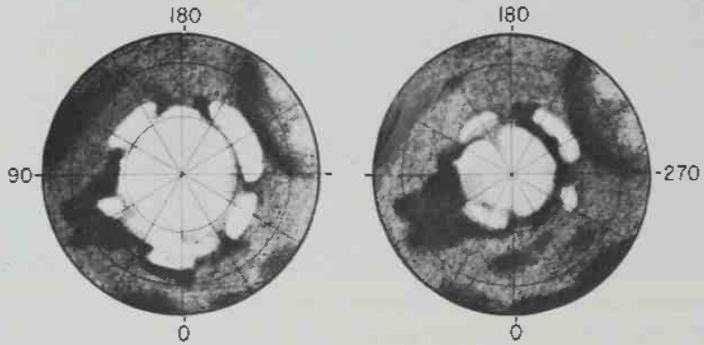


FIGURE *19–3. Two phases in the average seasonal development of the northern polar ice cap. These drawings are based on observations in the 1946, 1948, 1950, and 1952 oppositions. (Courtesy of Dr. A. Dollfus.)*

In order for a liquid to form at a given temperature, the atmospheric pressure must exceed a certain value. The atmosphere provides a kind of lid over the body of water. In a vacuum, the water would vaporize almost instantly. The total atmospheric pressure on Mars, discussed below, is so small that no effective lid is available to keep a pool of water liquid. Instead, if ice on Mars is heated, it will turn directly into water vapor; just as on Earth, under atmospheric pressure, we can observe that dry ice (frozen carbon dioxide), when heated, is converted into gaseous CO_2. But we never observe liquid CO_2 at 1 atm pressure.

▽ There is other evidence for the absence of open bodies of water on Mars. At certain angles of observation, we should see a bright image of the Sun reflected off the mirror-like surfaces of the hypothetical Martian lakes; despite much searching, no such image has ever been seen. We can conclude with some confidence, then, that no open pools of pure water exist on Mars.

▽ While the polar caps wax and wane with the seasons, the bright and dark areas of Mars generally maintain their relative configurations. Figure 19–4 is a

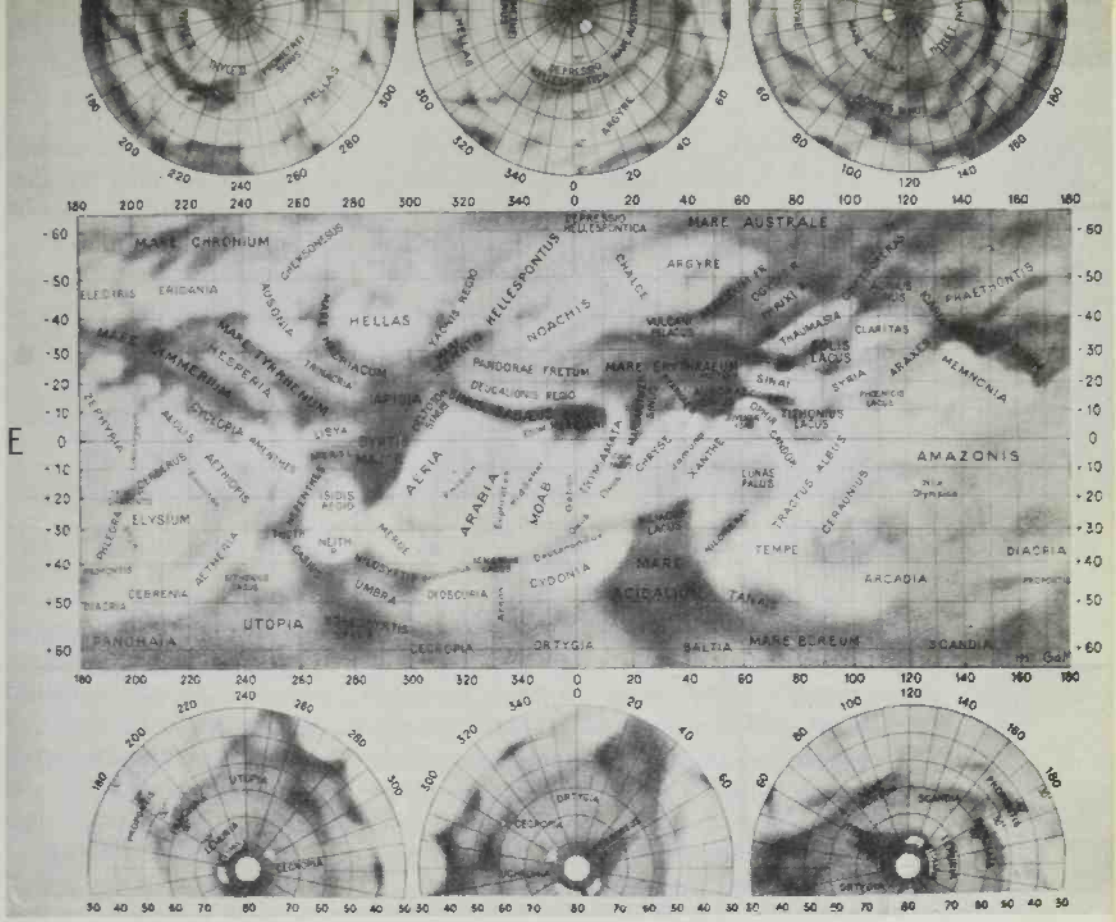


FIGURE 19-4. *The International Astronomical Union Mars cartography. Illustrated in this map are only those features which have been photographed during several recent oppositions. The smallest features depicted are some hundreds of kilometers across.*

map of Mars in mercator projection, like many maps of the Earth. The vertical scale shows latitude; the horizontal scale shows longitude. Because astronomical telescopes invert images, astronomers tend to think of south as being "up," a convention which is followed in this map. The features shown and named are those which have been repeatedly photographed, year after year. Under the best seeing conditions, observers of Mars have noted that the dark areas seem actually to be composed of many very dark spots, now referred to as the dark nuclei [Figure 19–5]. When the seeing conditions are not excellent, the dark nuclei appear smeared together, and we see Mars roughly as it appears in the map in Figure 19–4. There are other more elusive features, which have not been photographed but which can be seen visually through a large telescope under good seeing conditions. These we shall describe later.

▽ Around the turn of the century, the Martian bright and dark areas were given names, usually of Latin or Greek origin and sometimes with copious allusions to classical antiquity. Thus, Mare Erythraeum is the "Red Sea," although it is neither red nor a sea; Hellespontus is "The Greek Bridge," though neither in fact;

FIGURE 19-5. *A map of Mars based upon visual observations of Dr. J. H. Focas. The resolution of the dark areas into dark nuclei is evident. Cf. Figure 19–4. (Courtesy of Dr. J. H. Focas.)*

Solis Lacus is "The Lake of the Sun"; Sinus Meridiani is "Meridian Bay" because the zero meridian of Mars, corresponding to the meridian of Greenwich on Earth, passes through it. This is, of course, a terrestrial convention; Martian cartographers, if any, will have other conventions. And as a hopeful augury for the future, at 250° longitude, +55° north latitude is a locale that many have been seeking. It is an arresting thought that these unearthly places with strange names—Trivium Charontis, Tithonius Lacus, Thoth-Nepenthes—places until now observed only with large telescopes over immense distances, will, one day in the lives of most of us, be trod by men.

▽ While the general configurations of the bright and dark areas have, with a few exceptions, remained fixed for decades, we sometimes see a transient incursion of a bright area into a dark. It is as if a small piece of dark area becomes eaten away, replaced with buff-colored material typical of the bright areas. The incursion may continue, obscuring large fractions of the dark areas. There is every reason to believe that such incursions are great dust storms. The polarization of the light reflected from the bright areas indicates that they are covered with innumerable small, opaque particles. The incursions of bright areas into dark have just the same polarization properties as the bright areas. We conclude that the bright areas of Mars are vast deserts, and that occasionally the winds carry large amounts of desert material over the dark areas, temporarily obscuring them from our view. In 1956, a dust storm of planet-wide proportions was observed; for about a month it obscured almost all detail on the planet. Yet even after the most striking of these dust storms, the winds subside and the dark areas reappear. If the dark areas are at higher elevations than the bright areas, we can understand how they can be obscured by a dust storm, and later, scoured by the winds, reappear to the view of the astronomer on Earth. Recently the American astronomer James Pollack and I have found radar evidence suggesting that the dark areas are at systematically higher elevations than the bright deserts.

▽ The polarization of the light reflected from the bright areas not only tells us

that the bright areas are covered with dust; it also tells us something of the composition of this dust. Of hundreds of terrestrial minerals examined in the laboratory, only one shows the same polarization properties as the Martian bright areas—a mineral called limonite. Each molecule of limonite is an iron oxide polyhydrate—that is, a compound of iron and oxygen, Fe_2O_3, which has several molecules of water loosely bound to it. Limonite explains both the polarization properties of the Martian deserts and their color and brightness.

▽ The dark areas of Mars have also been examined polarimetrically and spectroscopically. The polarization of light reflected from the dark areas shows them also to be covered by or composed of small, opaque particles, which are even more opaque than the already very dark particles in the Martian deserts. Pollack and I believe that larger particles of limonite, several tenths of a millimeter across, can also account for the polarizing properties and the darkness of the dark areas. In general, larger particles are darker. The nature of the Martian dark areas carries us directly to the question of the possible existence of life on Mars, but we wish first to describe some other features of the Martian environment.

▽ Just as we saw, in Chapters 3 and 4, how the physical conditions in the stars could be determined from the light they emitted, so the physical conditions of planetary atmospheres and surfaces can be partially deduced from an examination of the sunlight that they reflect and the infrared and radio radiation that they emit. Sunlight which is reflected off the Martian surface passes twice, at a slant angle, through the Martian atmosphere. The molecules in the Martian atmosphere preferentially remove the sunlight at their respective absorption wavelengths, and the reflection spectrum of Mars contains lines and bands, just as does the emission spectrum of a star [Chapter 4]. In this way, the gases CO_2 and H_2O have been identified in the Martian atmosphere. The amount of water vapor in the atmosphere is roughly equal to the amount of water locked as hoarfrost in the polar caps. It is about 0.1 percent of the amount of water vapor in our own atmosphere. On the other hand, the amount of carbon dioxide in the Martian atmosphere is much larger than in the Earth's atmosphere—perhaps thirty times more, although the exact figure is still in question. Oxygen has been searched for unsuccessfully. If it is present at all, it exists as a trace constituent. The absence of oxygen, of course, does not necessarily preclude the existence of life on Mars, even in fairly advanced forms [see Chapter 14].

▽ If there is no oxygen in the Martian atmosphere, then we expect no ozone. In the Earth's atmosphere, ozone absorbs ultraviolet light between 2000 and 3000 Å, light which would otherwise be lethal to most terrestrial organisms. Does the absence of ozone mean that the Martian surface is bathed in ultraviolet radiation? If some other atmospheric absorber were present, either as a gas or a solid aerosol, the ultraviolet light might not reach the surface. When Mars is observed in ordinary visible light, it appears as in Figure 19–1 A, B, and C. On the other hand, when we observe Mars in ultraviolet light, we see something like Figure 19–1D. The polar region shows up very prominently, but almost all surface detail in the bright and dark regions has vanished. The source of this mysterious phenomenon is called the

"blue" or "violet haze," although it may not be a haze, and it certainly is not blue or violet. If it is a haze, it absorbs blue, violet, and ultraviolet light, and transmits light of longer wavelengths. If we had some of it—whatever it is—in a bottle, it would appear red. If the Martian blue haze is some unidentified atmospheric absorber, then the intensity of ultraviolet sunlight at the surface of Mars may still be small. However, a recent rocket observation of Mars in ultraviolet light implies the absence of atmospheric absorbers effective at these wavelengths on Mars. What, then, causes the blue haze? We have suggested that limonite particles cover both bright and dark areas, but that bigger particles exist in the dark areas. With the particle sizes necessary to explain both brightness and polarization, Pollack and I find that dark and bright areas reflect light equally well at violet and ultraviolet wavelengths. On this basis we expect the contrast between bright and dark areas to disappear in the violet and ultraviolet. Thus, the "blue haze" may be purely a surface effect; if this view is correct, there should be substantial penetration by ultraviolet light to the surface, posing, perhaps, an additional hazard for organisms on Mars.

▽ Many other gases besides oxygen have been searched for in the Martian atmosphere with negative results. From a radio experiment on the U.S. Mariner IV space vehicle and from the shape of absorption lines in the Martian atmosphere, it is possible to determine the total atmospheric pressure at the surface of Mars. This figure is about 10^{-2} atm—that is, about 1.0 percent of the total atmospheric pressure at the Earth. If we add up the carbon dioxide and water vapor which has been identified in the Martian atmosphere, we find that some of the atmosphere of Mars is unaccounted for. Some other gas which we have not yet identified is present. Certain molecules, such as N_2 and the noble gases, have their absorption lines in ultraviolet wavelengths which cannot be observed from the surface of the Earth because of absorption in the terrestrial atmosphere. Nitrogen is present in large quantities in our own atmosphere (78%), and its cosmic abundance is high. For these reasons, astronomers have concluded that some of the Martian atmosphere is composed of N_2. But this is an argument by default, and direct ultraviolet observations of Mars should be performed—for example, from Orbiting Astronomical Observatories above the atmosphere—to check the presence and abundance of nitrogen on Mars.

▽ Mention should be made of another gas, one which has been used in arguments about the presence of life on Mars. In 1956, the American astronomer C. C. Kiess, of Georgetown University, and his collaborators obtained a spectrum of Mars which seemed to show certain weak absorption features in the blue, green, and yellow parts of the spectrum. They attempted to match these features with a variety of gases and concluded that only nitrogen dioxide, NO_2, could explain the observations. Kiess and collaborators maintained that the quantities of this poison gas on Mars were so large as to exclude the possibility of any indigenous life on that planet. This discussion was taken up and extended by others; it was used to argue that future biological exploration of Mars by space vehicles was unnecessary, because life on Mars was impossible.

▽ This conclusion must certainly be considered premature. Kiess and his

collaborators never actually computed how much nitrogen dioxide their observations implied. If this computation is performed, the quantity of NO_2 in the Martian atmosphere is shown to be about 0.001 percent.

▽ But might this not still be large enough to constitute at least a chemical embarrassment to the Martians? The amount of NO_2 in our own atmosphere can be measured in quite analogous ways. For example, a spectrometer may be pointed at the Sun, and the absorption by NO_2 in the overlying atmosphere recorded on a photographic plate. Since NO_2 is a primary constituent of smog and other urban pollution, such studies are performed almost routinely in cities such as Los Angeles. There, the NO_2 content of the atmosphere varies with time. A typical pattern has maximum NO_2 abundances at 8:00 A.M. and at 5:00 P.M., and subsidiary peaks at 7:00 P.M. and 11:00 P.M. (all Pacific local time). These peaks correspond to the morning and evening rush hours, and probably to evening car-borne social activities of the inhabitants of that exotic city. Such studies open up entire new fields of research, such as spectrochemical sociology. But the important point is that the average amount of NO_2 above the city of Los Angeles is greater than the average amount of NO_2 in the atmosphere of Mars. Conditions in Los Angeles, while inclement by some standards, do not quite preclude life there; the same conclusion applies to Mars.

▽ By observing the infrared and radio emission which Mars radiates to space, it is possible to obtain an approximate picture of what the local surface temperatures are. At an average locale in the Martian desert, at the equator, in summer, near noon, a typical temperature might be as high as 20°C (68°F), or warmer than room temperature in Great Britain. Yet that very night, the temperature will plunge to 70 or 80° below zero (−94 to −112°F). Mars has been described as having an extreme continental climate. As we move closer to the poles, the average temperatures become less, and the diurnal fluctuation becomes smaller. A temperature average over latitude, longitude, season, and time of day, might be −30 or −40°C (−22 to −40°F). There is no spot on Mars which stays above the freezing point of water during any 24-hour day. Still, some locales tend to be much warmer than others. Despite out terrestrial tendency to consider deserts as being hotter than other places, the Martian bright areas tend to be cooler than the dark areas, in part because being brighter, they absorb less sunlight during the day. The dark nuclei of the dark areas are very dark indeed; therefore, they absorb significantly more sunlight than the deserts. While even in the dark nuclei, the night-time temperatures tend to be low, during local spring and summer the average daytime temperature of a dark nucleus tends to stay above the freezing point of water.

▽ Mars therefore appears to be generally cold, arid, and oxygen-poor. A man transplanted to Mars with no protective equipment would asphyxiate before he would freeze. Otherwise, he might die of thirst or be scorched by ultraviolet light. But men are not the only organisms on Earth. The most ubiquitous terrestrial life forms are the microorganisms. What happens if we inoculate a simulated Martian environment with terrestrial microorganisms? Such experiments have in fact been

performed. A chamber is prepared which may contain dried limonite powder, an oxygen-free atmosphere composed mostly of CO_2 and N_2 is introduced under reduced pressure; an ultraviolet lamp shines on the limonite; and the entire chamber is taken through a daily freeze-thaw cycling. Such chambers are, of course, known as "Mars jars." Perhaps remarkably, when samples of terrestrial soil rich in microorganisms are introduced into such an environment, while some of the microorganisms die, mostly after the first freeze-thaw cycle, a significant fraction survive indefinitely. The survivors include a wide variety of microorganisms, including types which form spores and types which do not. The ultraviolet light kills the microorganisms unlucky enough to be exposed. Those microorganisms which hide under rocks survive. They do not need oxygen, the temperatures do not bother them, and the very low water content is adequate for their needs. If the water content is increased, corresponding, say, to what may happen in local Martian spring, the surviving organisms are found to grow and reproduce. △

It is important to remember that people in the Antarctic live at temperatures which are within the ranges found in the Martian polar regions. The lowest temperature ever recorded on Earth, in the Antarctic, was −82°C (−116°F). Man living in the Antarctic creates his own artificial biosphere, but organisms have a great capacity for evolutionary adaptation to severe environmental conditions. The inclemency of the Martian climatic conditions does not in itself exclude the possibility of life.

▽ Such simulation experiments are clearly relevant to the question of life on Mars. They indicate that perfectly adequate biological mechanisms exist for survival under average Martian conditions, and for growth when the conditions are relatively favorable. Since there are few natural terrestrial environments quite as rigorous as those on Mars, it is remarkable that terrestrial organisms have Martian survival capability; but they do. These experiments, a kind of natural selection on a laboratory scale, have only been carried out for periods of months. If we imagine them carried out over very long periods, we can see that through mutation and selection, an evolutionary process will occur in which the survivors are increasingly better adapted to the Mars jars. In just the same way, we can imagine Martian organisms evolving, in the Martian environment, into forms well adapted to it. For all we know, life forms much more advanced than microorganisms might have evolved under such circumstances. While these experiments increase the plausibility of indigenous life on Mars, they, of course, do not prove that life exists on Mars.

▽ But such experiments do have relevance to the question of biological contamination of Mars. Suppose that in future attempts at Martian exploration, a space vehicle were to impact the Martian surface. On its journey from Earth to Mars, such a spacecraft would have been externally sterilized by solar ultraviolet radiation. This radiation does not penetrate into the spacecraft interior, and any organisms which were there at the launch of the spacecraft would survive the impact of the spacecraft on Mars. If the spacecraft fractures on impact, terrestrial microorganisms will be deposited on the Martian surface. If precautions are not

taken before launch, all interior materials of the spacecraft will have a large complement of terrestrial microorganisms of many varieties. Martian winds and dust storms are, as we have seen, prevalent; the microorganisms would be distributed over the entire surface of Mars. It is still possible that little ultraviolet radiation reaches the surface of Mars, and that the microorganisms—or "bugs," as microbiologists affectionately, but inaccurately, call them—would not be killed by germicidal sunlight. Even if the ultraviolet flux is large, bugs adhering to particles of dust might survive nicely.

▽ There is a certain "compound interest" to microbial reproduction. In the absence of predators or competitors, the bugs reproduce exponentially. As a simple example, consider a bug which is deposited in an environment in which it grows very slowly. Some terrestrial microorganisms reproduce once an hour; imagine the microorganism in question reproducing on Mars once every thirty days. Thus, at the end of thirty days, we have two organisms; at the end of sixty days, we have $2 \times 2 = 4$ organisms; after ninety days, $2 \times 2 \times 2 = 8$ organisms; and after $30 \times n$ days, 2^n organisms. After 300 days ($n = 10$), we would have 2^{10}, or approximately 10^3 microorganisms. After 1500 days ($n = 50$), or slightly longer than two Martian years, we would have $2^{50} = (10^3)^5 = 10^{15}$ microorganisms. After eight Earth years (about 3,000 days; $n = 100$), we would have $2^{100} = (10^3)^{10} = 10^{30}$ microorganisms, a number which is larger than the entire microbial population of the planet Earth. This example illustrates the seriousness of biological contamination of Mars.

▽ We suspect that there are already at least some microorganisms on Mars, and we wish to examine them in detail. What do they look like? How are they constructed? How do they function? Are they composed of cells? Is the basic hereditary material made of nucleic acids? Are proteins used as catalysts? There is a long list of basic biological questions to be asked.

▽ Now imagine that despite the danger of biological contamination, we send unsterilized spacecraft to Mars—for example, to learn more about its physical environment. In later missions, we send instruments designed to search out and characterize indigenous Martian organisms, if any. We find microorganisms on Mars—in fact, microorganisms which are very similar to some terrestrial bugs. What do we conclude? That similar forms have developed independently on the two planets? That Mars and Earth have had some common biological contact in the distant past? Or that a spacecraft from Earth inadvertently deposited organisms on a previous mission? Biological contamination of Mars would be a major scientific disaster. For this reason, a program of space vehicle decontamination and sterilization has been announced by the National Aeronautics and Space Administration of the United States. But the United States is not the only spacefaring nation. The Soviet Union has an imminent capability for Martian landings, and other nations may, in the not-too-distant future, also participate in the search for life on Mars. It matters little if contamination of Mars is effected by a Russian or an American bug; the microorganisms know no nationalities. Without some effort on our part, they may not even respect interplanetary boundaries. For this reason,

it is cheering that the Soviet Union has shown signs of willingness to sterilize its spacecraft. Efforts were made to sterilize the Soviet lunar rocket Luna II, and a resolution calling for rigorous sterilization of space vehicles launched to Mars was approved in May, 1964, by Soviet representatives at the meeting of the Committee on Space Research of the International Council of Scientific Unions. In this area of space exploration, the peoples of the planet Earth appear to have a common purpose singularly apposite for our first venture of another world.

▽ We have assayed the known environment of Mars. We find it rigorous, but probably not too rigorous for indigenous organisms. Yet it is clear that organisms could not have originated and evolved on a planet similar to contemporary Mars [see Chapter 16]. Might the conditions have been more clement on primitive Mars? As we have seen in Chapters 11–13, there is good evidence to support the belief that all the planets in the solar system were formed in an analogous manner, out of the same cloud of gas and dust, which had a common reducing chemistry. There is no reason to doubt that the primitive atmosphere of Mars was reducing; that due to an atmospheric greenhouse effect, the temperatures were warmer; and that some open bodies of water may have been in existence—although these matters are not rigorously demonstrated. The change from the primitive to the contemporary Martian environment must have had the same cause as the transition from the primitive to the contemporary terrestrial environment—namely, atmospheric escape. Mars has a lower mass and therefore provides a greater opportunity for escape of a given molecule from its gravitational field. As the atmosphere of Mars slowly boiled off to space, during eons of geological time, the atmospheric conditions became less reducing, the surface temperatures declined, and eventually most of the water evaporated to space or was frozen subsurface. These changes were gradual, and we can easily imagine the adaptation through natural selection of Martian organisms to the changing conditions.

▽ Some independent support for this picture is provided by the composition of the Martian deserts. As we saw earlier in this chapter, the polarization of light reflected from the deserts strongly suggests that they are composed of limonite, $Fe_2O_3 \cdot nH_2O$. Limonite appears on Earth primarily in equatorial climates, mixed with hematite and bauxite as lateritic soil. Limonite and lateritic soil are both highly oxidized, and have large water content. Limonite is generally five to ten percent water by mass. Terrestrial limonite and lateritic soils are formed, geologists believe, only in the presence of oxygen in hot, humid environments. Contemporary Mars has no oxygen; it is cold and arid. We can understand the presence of large quantities of limonite on Mars only if we postulate the existence of earlier conditions much like those in the tropical zone on the Earth. If limonite requires molecular oxygen for its formation, we then have evidence that at one epoch in its history Mars held an oxidizing atmosphere, an atmosphere which presumably has by now escaped to space or has reacted chemically with the Martian surface.

▽ The oxygen in the terrestrial atmosphere, as we saw in Chapter 16, is probably produced by plant photosynthesis. Might that also have been the case for Mars?

Was Mars once lush and verdant? While oxygen may not be necessary for advanced life forms, the only examples which we have here on Earth point to its usefulness in extracting energy from foodstuffs. Could advanced life forms have developed at some epoch in the distant Martian past, only to be destroyed at a later time by the escape of oxygen to space and its reaction with the crust? Or might organisms on Mars have continued to adapt to the changing Martian environment, in ways at which we now can only dimly guess? △

20

The quest for life on Mars

Where stars are shining in the mist,
In measured steps the Martian treads.
On hillocks of monastic hue
No grass, no trees, no, none of these . . .

I. Smelianov

Brothers . . . stoop not to renounce the quest
Of what may in the sun's path be essayed,
The world that never mankind hath possessed.

Ulysses, in Dante, *Inferno,* XXVI

▽ The origin of life on primitive Mars seems not unlikely. The present physical environment of Mars does not exclude life. We have seen, in Chapter 18, the difficulties in remote detection of life on Earth from a Martian vantage point. How can we say anything more about life on Mars? Remarkably enough, there are a variety of observations which have been interpreted, with varying degrees of success, as indicating life on Mars. Some of the early arguments we now know to be almost certainly erroneous, but even the most recent pieces of evidence do not unambiguously demonstrate the existence of life on Mars. A coherent picture can be gained only by considering all the facets of this enigmatic subject.

▽ Some decades ago, it was commonly reported that the Martian dark areas were green. If the dark areas were green, what could they be made of? The most common greenish materials on Earth are plants, and it was readily concluded that vegetation was thriving on Mars. But the detection of colors by astronomical observation is a thorny problem. It is possible to be deceived for physical reasons and for psychophysiological reasons. Around the turn of the century, it was common to use refracting telescopes, which employ lenses to collect light. Today, almost no refracting telescopes are being constructed for professional astronomical work; instead, reflectors, using large mirrors, are the rule. One of the advantages of reflectors is that they are not plagued by chromatic aberration, as are refractors. Chromatic aberration occurs because light of different colors is brought to different focuses in transmission through a lens. Thus, if the yellow wavelengths of sunlight, reflected off Mars into the telescope, are in focus, many other colors will be out of focus. In particular, the extrafocal blue and green light is smeared over the image of Mars. Mixed with the red-orange coloration of the bright areas, there is little apparent change in color; but mixed with the neutral gray of the dark areas, a distinct blue-green coloration appears. The use of reflecting telescopes largely removes such problems of chromatic aberration.

▽ There remain, however, psychophysiological problems. When a neutrally colored area is placed alongside a brightly colored area, it tends to acquire a complementary color. This corresponds to no real coloration in the neutral area, but is simply a quirk of human color vision. The colors complementary to the red-orange Martian bright areas are greens and blues; and again, neutrally colored dark areas on Mars are invested with a spurious blue-green coloration. Confusion due to extrafocal light and to color contrast effects can both be removed by using a diaphragm with a large reflecting telescope. The diaphragm isolates a dark area, so that the adjacent ruddy bright areas are not seen; the reflecting telescope removes the extrafocal blue light. Under these circumstances, the dark areas appear an almost neutral gray. There is some tendency for the dark areas to appear slightly

reddish; this is not surprising, because some of the dusty material from the bright areas must also be present in the dark areas. Occasional subtle and delicate colors have been observed in recent years, but they are a far cry from earlier days, when "chocolate," "carmine," "kelly green," and "dragon's blood" could be found in scientific descriptions of Mars.

▽ Color changes which have been reported on Mars are probably also largely illusory. As we have mentioned, given a brightly colored area with a red or orange hue adjacent to a neutral gray area, the eye invests the dark area with some blue-green color. Now, if the dark area changes its darkness—that is, varies its contrast with the bright area—it will appear to vary in color. The eye's interpretation of contrast variations as color variations is, in fact, one of the principles of the Land process of color photography. Thus, if the Martian dark areas change their brightness, it should not surprise us that they also seem to change their colors.

▽ If, then, the Martian dark areas are neutrally colored and not green, is the possibility of vegetation in the dark areas excluded? It is true that the most easily seen terrestrial plants are colored green. The color arises from a highly specific and ubiquitous molecule known as chlorophyll. Chlorophyll, a photon-acceptor, is involved in the first step in the long photosynthetic chain which converts the energy of sunlight into the energy-rich bonds of the ATP molecule [see Chapters 14 and 17]. Chlorophyll appears green because it absorbs in the red and the blue; the middle of the spectrum is reflected back from the plant, to give it its greenish hue.

▽ The absorption properties of chlorophyll are critically dependent on its molecular structure. A slight change in molecular side groups can produce a major change in the absorption properties of the molecule. A large part of the solar spectrum is in the yellow and green wavelengths which chlorophyll tends to reject. To utilize these yellow and green photons, plants on Earth have made many special adaptations. Many plants use a wide range of accessory pigments, molecules quite different from chlorophyll, such as the carotenoids, which give carrots their distinctive color. Here, it is the orange and red parts of the spectrum which are mostly not utilized, and the shorter wavelengths, including the green and yellow light, which are absorbed. Higher plants concentrate large amounts of chlorophyll, so that the relatively weak absorption in the yellow and green is compensated by the large number of absorbers. There seems to be no particular adaptive advantage to greenish coloration in plants. Most likely, it is an historical accident; that is, at the time of the origin of plants, chlorophyll molecules absorbing primarily in the red and the blue were evolved, and all of subsequent plant evolution has been built upon these early adaptations. Instead of changing the fundamental ground plans, plants have made accessory adaptations to correct the grosser photosynthetic deficiencies.

▽ On another planet, it is entirely possible that other pigments evolved early in the origin of life; there really seems no reason to expect extraterrestrial vegetation to be green. In fact, a neutral color such as brown, gray, or black has a subsidiary advantage on chilly Mars: no part of the spectrum is rejected; all photons are used,

either for photosynthesis or for simple heating of the plant. Neutral colorations may make much more sense for Martian plants than greens. This point has been stressed by G. A. Tikhov, the former director of the world's first and only Institute of Astrobotany, at Alma Ata, Kazakhstan, U.S.S.R. This institute, now defunct, was an early focus of Soviet enthusiasm for extraterrestrial life, at a time when the subject was lacking both solid observational techniques and the support of the scientific community.

▽ While some Martian surface features, such as Syrtis Major, were probably observed as early as the eighteenth century, the first systematic mapping of Mars, with the aid of adequate telescopes, did not occur until the last half of the nineteenth century. Many of our present names for Martian surface features derive from that time. A leader in early Martian cartography was an Italian astronomer, Giovanni Schiaparelli. In 1877, during relatively routine observations of Mars, under conditions of relatively good seeing, Schiaparelli was surprised to find long, dark, rectilinear features which seemed to connect dark area with dark area, traversing thousands of kilometers of the Martian deserts. Schiaparelli called these features "canali," which, in Italian, denotes channels, or grooves. The word was, however, translated into English as "canals," a word embracing the distinct implication of intelligent design.

▽ The existence and significance of the canals were most eloquently championed, some decades later, both in the scientific and the popular literature, by Percival Lowell, an American diplomat turned astronomer, who established an observatory in Flagstaff, Arizona, for the express purpose of studying Mars. The seeing conditions in Arizona were superior to those of most other observatories at the time. The further observations which Lowell and his associates recorded were developed into a coherent picture of Mars, which went something like this: long, rectilinear features are observed crossing the Martian deserts. They apparently undergo seasonal brightness and color changes. Occasionally, one such line apparently germinates into two. The lines never stop at some desert locale, but always continue from dark area to dark area. Straight lines, Lowell argued, are not natural features; therefore, they must be artificial. If the canals are in fact artifices, what is their function? Even at the turn of the twentieth century it was known that the gravitational field of Mars was not likely to hold an extensive atmosphere, and that liquid water was not in great abundance on the Martian surface. Lowell therefore proposed that the canals were canals—carrying liquid water from the polar ice caps to the thirsty Martians residing in the dark equatorial regions. Small, dark nuclei which were observed at the interconnections of several canals were appropriately called "oases."

▽ Two immediate scientific objections to the canal theory were disposed of by Lowell and collaborators with arguments which retain their validity today. First, it was suggested that the reported widths of the canals were too small to be detected with the resolving power of the telescopes used. Lowell showed that long, rectilinear features against high-contrast backgrounds can be seen even if their

widths are far below the theoretical resolving power. He performed an experiment with overhead transmission wires in Flagstaff, Arizona, showing that he could station himself far from the wires and still detect their presence.

▽ The second objection was that the canals were too wide—many kilometers across, much wider than necessary to carry water from the polar ice caps. Lowell countered as follows:

> The fundamental fact in the matter is the dearth of water. If we keep this in mind, we shall see that many of the objections that spontaneously arise answer themselves. The supposed herculean task of constructing such canals disappears at once; for, if the canals be dug for irrigation purposes, it is evident that what we see, and call by ellipsis the canal, is not really the canal at all, but the strip of fertilized land bordering it,—the thread of water in the midst of it, the canal itself, being far too small to be perceptible. In the case of an irrigation canal seen at a distance, it is always the strip of verdure, not the canal, that is visible, as we see in looking from afar upon irrigated country on the Earth.

▽ With these successes in the scientific dialogue, Lowell and his followers constructed an inverted pyramid of deductions upon the apex of the canal observations. The canals were a massive engineering work; therefore, the Martians are in substantial technological advance of contemporary human society. The canals obviously cross what we would term international boundaries; hence, a world government exists on Mars. One of Lowell's followers went so far as to place the capital in Solis Lacus (latitude −30°, longitude 90°, in Figure 19–4). The hydraulic engineering required was discussed, and Lowell painted moving verbal portraits of a race of superior beings, engaged in heroic attempts to maintain their civilization on a dying planet. Lowell's ideas were incorporated into fictional form by Edgar Rice Burroughs, in a series of books about John Carter, a terrestrial adventurer cavorting on Mars, which introduced Lowell's ideas to an even larger public.

▽ Lowell developed a long chain of argument, ultimately based on the reality of the canals as a genuine Martian surface feature. But were the canals really on Mars, or, as with other beauty, in the eye of the beholder? An acrimonious scientific debate ensued, stretching over decades in time. Although largely resolved, it can still be heard echoing occasionally in contemporary scientific literature. Had Lowell been less articulate, had he not directed his eloquence to the general public, the debate would probably have terminated much earlier. It became so bitter, and seemed to many scientists so profitless, that it led to a general exodus from planetary to stellar astronomy, abetted in large part by the great scientific opportunities then developing in the application of modern physics to stellar problems. The present shortage of planetary astronomers can be largely attributed to these two factors. To savor the spirit of this debate, consider the following two scientific reports from eminent and experienced planetary observers:

> I have been watching and drawing the surface of Mars. It is wonderfully full of detail. There is certainly no question about there being mountains and large greatly elevated plateaus. To save my soul I can't believe in the canals as Schiaparelli draws

them. I see details where he has drawn none. I see details where some of his canals are, but they are not straight lines *at all*. When best seen these details are very irregular and broken up—that is, some of the regions of his canals; I verily believe—for all the verifications—that the canals as depicted by Schiaparelli are a fallacy and that they will so be proved before many favorable oppositions are past. (E. E. Barnard, 1894)

At the first glance through the 32½-inch on 1909, September 20, I thought I was dreaming and scanning Mars from his outer satellite. The planet revealed a prodigious and bewildering amount of sharp or diffused natural, irregular detail, all held steadily; and it was at once obvious that the geometrical network of single and double canals discovered by Schiaparelli was a gross illusion. Such detail could not be drawn; hence only its coarser markings were recorded in the notebook. (E. -M. Antoniadi, 1916)

Other observers were similarly unable to record the existence of canals, even under very good seeing conditions. W. H. Pickering, a supporter of Lowell, countered one such report as follows:

Dr. Van Biesbroeck records that he saw no canals or lakes on that evening in spite of the excellent seeing. The main reason for this is that on that evening, at that time, there were no canals or lakes visible.

But Lowell's rejoinders were more eloquent:

The straightness of the lines is unhesitatingly attributed to the draughtsman. Now this is a very telling point. For it is a case of the double-edged sword. Accusation of design, if it prove not to be due to the draughtsman, devolves *ipso facto* upon the canals . . . Let us not cheat ourselves with words. Conservatism sounds finely, and covers any amount of ignorance and fear.

▽ Under fine seeing conditions, some observers saw the canals; others did not. Did one group have superior visual acuity, or did the other have superior powers of imagination? The answer seems to be this: when the atmospheric seeing conditions are moderately good, the polar cap, bright areas, and dark areas can be discerned, as in Figures 19–1 and 19–2, but no canals appear. When the seeing conditions improve, many observers suddenly glimpse the canals standing out, as Percival Lowell put it, "like the lines in a fine steel etching," in a complex and interlacing pattern encompassing the whole planet. But at the best observing sites, such as Pic du Midi, under superb seeing conditions, experienced observers have found that the canals can be resolved into disconnected fine features, the "bewildering amount of sharp or diffused natural, irregular detail" of Antoniadi. Then, as the seeing conditions become worse, and the image shimmers, the canals reappear. The eye has a compulsive need for order, and in the few moments of superior seeing, in which we must glimpse, remember, and record the surface of Mars to sketch in our notebooks, it is far easier to remember a few straight lines than a multitude of fine detail. Similar results have been demonstrated in laboratory experiments. When observers are left to glimpse and record discon-

nected, mottled features under poor seeing conditions, they tend to construct straight lines where none exist.

▽ A comparison of a given region of Mars under conditions of good and of superb seeing, where the canals are, respectively, glimpsed and resolved, is shown in Figure 20–1, taken from Antoniadi's book *La Planète Mars,* 1929. This result has been confirmed by the French astronomer Audouin Dollfus and by other workers in recent years. Thus, the problem of the Martian canals, as with so many other apparent observables on Mars, appears to be largely psychophysiological rather than astronomical.

▽ There have been no photographs of the thin, straight canals of Mars, the nub of the Lowellian controversy, although there are photographs of broader features, sometimes called canals, such as Thoth-Nepenthes (longitude 260°, latitude

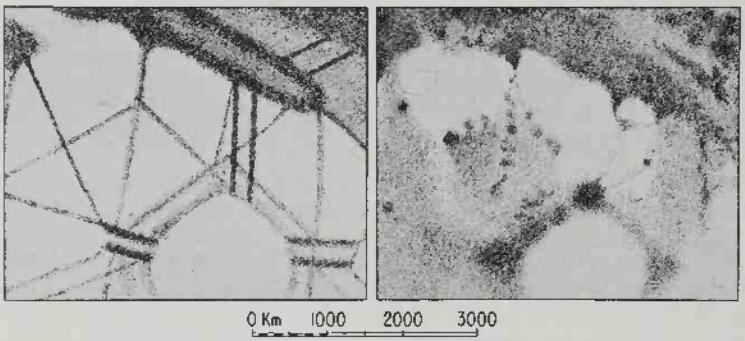


FIGURE 20-1. *Example of the resolution of the "canals" of Mars into fine detail. On the left are observations of the region of Elysium as observed by Schiaparelli between 1877 and 1890. There is a network of fine linear canals, both single and double. On the right is a drawing of the same area, as observed by E.-M. Antoniadi, between 1909 and 1926. By squinting and alternately looking at the left- and right-hand illustrations, the reader may test the hypothesis that the Martian surface features are actually as depicted on the right and that, when the atmospheric seeing conditions become poor, are glimpsed as in the illustration on the left. (Reproduced from* La Planète Mars *by E. M. Antoniadi, Hermann et Cie, 1930.)*

+20° in Figure 19–4, left-hand photograph in Figure 19–2). The photographic plate has the advantage of objectivity. Rarely will wishful thinking introduce a canal in a photograph where none was originally. But photography has the disadvantage that Mars must be viewed in a time exposure, encompassing moments when the seeing ranges from poor to superb. The photographic plate records an average, while the eye can remember the one moment of superb seeing. To record on photographs the details which Schiaparelli and Lowell interpreted as canals requires resolution which cannot be obtained from the surface of the Earth. We are like fish in the ocean depths, longing to view the flights of eagles.

▽ Telescopes lofted by balloon to the stratosphere, or taken on spacecraft nearer to Mars, should, in the not-too-distant future, obtain for us an accurate photographic representation of the features formerly interpreted as canals. The features must have some significance; although there is disconnected detail on the Moon,

which is observed through the same telescopes as Mars, no one has ever reported canals on the Moon. There are characteristic features on Mars, unlike the canals of Lowell, but which are the basis of the canal reports. One recently suggested possibility is that they are strings of sand dunes. Thus, while the canals are almost certainly not the massive engineering works of an advanced Martian civilization, their study may yet give us some further insights into the Martian environment. Meanwhile, the controversy has served at least the useful purpose of emphasizing the dangers of too many conclusions from too little data. As the Swedish chemist, Svante Arrhenius, put it (1918):

> The theory that intelligent men exist on Mars is very popular. With its help everything can be explained, particularly if we attribute an intelligence vastly superior to our own to these beings, so that we not always are able to fathom the wisdom with which their canals are constructed . . . The trouble with these "explanations" is that they explain anything, and therefore in fact nothing.

▽ Most of the present evidence suggesting life on Mars is of a different character. Each year, as the Martian ice caps recede towards the poles, sizable quantities of water vapor are released into the atmosphere. The Martian atmospheric circulation is apparently adequate to transport this water vapor across the equator, so that the water released by the retreat of one polar cap is available for the reformation of the polar cap in the opposite hemisphere. The radius of Mars, R, is 3380 km. The circumference of Mars is $2\pi R$, so the distance from pole to pole is πR. It takes half a Martian year for the water vapor to travel from pole to pole, or about $687/2 = 344$ days. The average rate at which the water vapor travels from pole to pole is therefore $\pi R/344$, or about 30 km per day.

▽ At the same time that the water vapor is being transported through the atmosphere, a remarkable phenomenon, known as the wave of darkening, occurs on the surface. The dark areas become progressively darker, and their contrast with the unchanging bright areas increases. This occurs in a wavelike movement, the front of the darkening wave progressing from the vaporizing polar cap towards and across the equator, and into the opposite hemisphere. Half a Martian year later, the wave of darkening proceeds in the opposite direction. The wave of darkening is not subject to the uncertainties of eyeball astronomy; it has been repeatedly photographed, and quantitatively measured on telescopes equipped for photometry. The wave of darkening proceeds according to recent measurements by the Greek astronomer J. H. Focas of the Athens Observatory, at an average rate of 35 km per day, close enough to the presumed rate of transport of water vapor in the atmosphere to suggest that the two phenomena are connected. It is this seasonal contrast enhancement which is the source of reports of seasonal color changes on Mars.

▽ Now what is the origin of the wave of darkening? Svante Arrhenius, whom we have just encountered attacking the Lowellian dogma, proposed an inorganic explanation of the darkening wave. Arrhenius suggested that salts exist in the dark areas of Mars (but not in the bright areas) which change their darkness and color with the humidity. Materials of this general type, such as cobalt chloride, are known on Earth, and in fact, are used to measure humidity changes. The amount

of water released by the polar cap, if distributed over the entire planet, is very small, about 10^{-3} gm over each square centimeter of the planet, some thousand times less than the water vapor content of the Earth's atmosphere. At the moving front of the wave of darkening, the water vapor content may be ten times larger, or 10^{-2} gm over each square centimeter. No materials on Earth are known which change their darkness (or color) in the manner observed on Mars due to such a small increase in the absolute quantity of moisture. Also, those materials whose absorption properties are responsive to humidity, the so-called hygroscopic salts, polarize the light reflected from them in a manner inconsistent with the observed polarization of sunlight reflected from Mars. The bright areas of Mars, we recall, are composed of limonite, a very dark, very strongly absorbing material. The dark areas of Mars are darker still, and cannot be composed of a semitransparent salt.

▽ There is an alternative explanation of the wave of darkening. Mars appears to be an arid world. If there are organisms there, we might expect them to be very responsive to the availability of water. Mars is further from the Sun than Earth is, and we might expect any photosynthetic plants on Mars to be more hungry for photons than are plants on Earth. We observe that when the local humidity increases, the Martian dark areas become darker. Are we in fact observing the seasonal growth and proliferation of Martian vegetation? The suggestion is a natural one, and was proposed as long ago as 1884 by the French astronomer E. L. Trouvelot, who mused:

> Judging from the changes that I have seen to occur from year to year in these spots one could believe that these changing grayish areas are due to Martian vegetation undergoing seasonal changes.

▽ Visual observation of the wave of darkening indicates that the darkening changes in the individual dark nuclei occur in periods characteristically as short as a week. The changes cover vast areas of Mars. The sudden flourishing of plant life over sizable areas of the Earth is a fairly common occurrence. Algal blooms are one example. A possibly more relevant example is the rapid growth of vegetation during the annual rainy season in many terrestrial deserts. Figures 20–2 and 20–3 illustrate the dramatic changes in a landscape which occur within about a month of a significant increase in the amount of available moisture. If the wave of darkening is a biological phenomenon, it follows that life on Mars is widespread and, furthermore, responds very rapidly to slight increases in the local moisture content. What an extraordinary conclusion to draw, from fifty million miles away! Yet we must recall the demise of the canals. Have we considered all alternatives? Is Martian biological activity the only reasonable explanation for the wave of darkening, or is there another, inorganic explanation, closer to the truth, which has eluded us up to now?

▽ Probably related to the wave of darkening is the dark collar which surrounds the retreating polar ice cap on its journey towards the pole. The collar has variously been described as black, brown, or blue. It is a real Martian phenomenon, not a contrast effect, as can be demonstrated by blotting out the polar

FIGURE *20–2. Scrub vegetation in a semi-desert area at Jebel Sileitat, Khartoum, Sudan, before the onset of the annual rains.* [*Courtesy of Dr. M. J. Chadwick, University of Cambridge.* (*See also* Life in Deserts, *by J. L. Cloudsley-Thompson and M. J. Chadwick, Dufour, Philadelphia, 1964.*)]

FIGURE *20–3. The same area of the Jebel Sileitat shown in Figure 20–2, but now in the rainy season. Because of the vagaries of color reproduction the grasslands here do not appear so green, nor the sky so blue as they do in Khartoum.* [*Courtesy of Dr. M. J. Chadwick, University of Cambridge.* (*See also* Life in Deserts, *by J. L. Cloudsley-Thompson and M. J. Chadwick, Dufour, Philadelphia, 1964.*)]

cap at the telescope and noting that the collar is still much darker than the surrounding regions. The polarization of light reflected from the polar collar shows that it is not simply due to dampening of the Martian soil. Something else is going on.

▽ At the edge of the receding polar ice cap, there is probably a greater supply of atmospheric water vapor than in any other region on Mars. There is even the bare possibility of temporary and shallow pools of liquid water, although this has never been confirmed polarimetrically. It seems to make some sense for organisms on an arid planet to proliferate at the edge of the polar cap. Despite the low average temperatures on Mars, daytime summer temperatures in the dark nuclei near the polar cap tend to be mild, even by terrestrial standards. The edge of the summer polar cap seems an ideal place to search for life on Mars.

▽ In addition to these seasonal variations, Mars exhibits striking secular changes. While the relative configurations of the bright and dark areas generally retain their integrity for many decades, some areas of Mars characteristically undergo marked, rapid, erratic changes. In Figure 20–4, we see four drawings, three of them made by the Greek astronomer Antoniadi. The upper drawings were made in 1877 and 1911; the lower drawings, in 1924 and 1926. The region is Solis Lacus, erstwhile capital of Mars in the Lowellian *Weltansicht*. As an example of the superiority of visual to photographic observations, compare Antoniadi's drawings of Solis Lacus with the representation of Solis Lacus in Figure 19–4 (longitude 90°, latitude −30°), taken from photographic plates alone. The dotted lines in the lower two pictures indicate areas covered by clouds at the time of observation. Something extraordinary happened in Solis Lacus between 1877 and 1911, and in the even shorter period of time between 1924 and 1926. From the scales of these figures, we see that the changes have been major. Areas 1000 km on a side are involved. Great variations in fine detail have occurred. Some dark areas have appeared in a desert; elsewhere, deserts have encroached into the dark areas. What is happening? Perhaps the Martian secular changes represent ecological successions on Mars. On the Earth, due to changing geological and climatic conditions, often a species of organism will arrive in a previously uninhabited area and there proliferate mightily in a relatively brief period of time. At other times, the climatic conditions may prove so severe that the species is locally destroyed.

▽ Possibly related to the secular changes is the reappearance of the dark areas after a dust storm. We observe material from the Martian deserts blown by winds over the Martian dark areas, which are consequently obscured. The dust storm does not move on; yet, after a short period of time, characteristically about a week, the dark area reappears. Where has the dust gone? Has Martian vegetation grown through the dust in the short period of a week, as the Estonian-Irish astronomer Ernst Öpik has suggested? Do the Martian plants shake themselves clean? Or is there an inorganic explanation, related to elevation differences?

▽ In Chapter 19, we alluded to the identification of limonite in the Martian bright areas, from analysis of the polarization of the sunlight which they reflect.

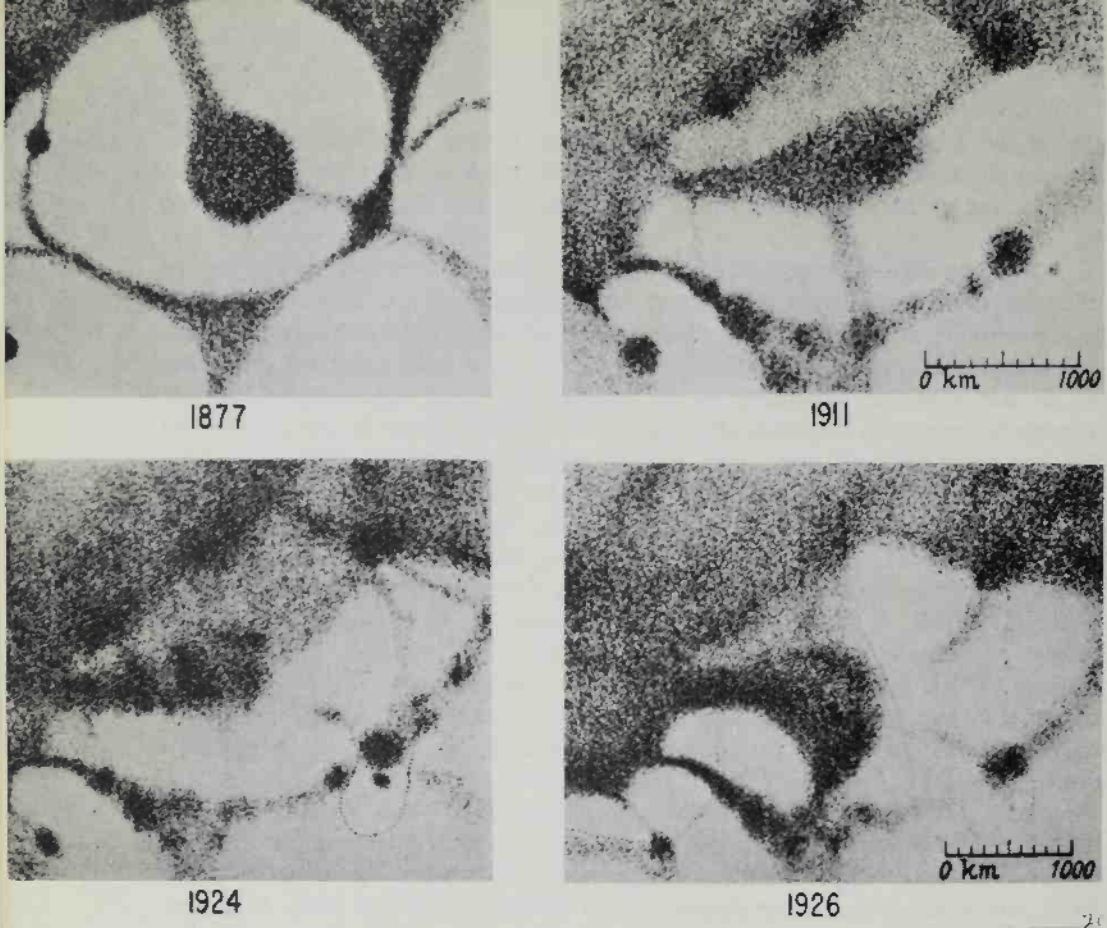


FIGURE 20-4. *Examples of secular changes on Mars. These are four drawings of the same area of Mars, the region of Solis Lacus. The drawing in upper left was made by Schiaparelli. The remaining three drawings were made in 1911, 1924, and 1926 by E.-M. Antoniadi. Even allowing for the difficulties in visual observations and the differences in styles of drawing between Schiaparelli and Antoniadi, it is clear that there have been substantial topographical changes on the Martian surface.* (*Reproduced from* La Planète Mars *by E.-M. Antoniadi, Hermann et Cie, 1930.*)

The polarization of the Martian deserts is independent of the Martian seasons; the polarization curves have the same character in Martian summer as in Martian winter. However, in the dark areas, the polarization depends very much on the season of the year, being much more striking in local Martian spring and early summer than in late fall and winter. To reproduce anything like these seasonal polarization changes in the laboratory, Dollfus was forced to conclude that the particles in the Martian dark areas change their darkness or their size—probably both—periodically with the seasons. We have already encountered seasonal changes in the wave of darkening and the reports of color changes in the Martian dark areas. But changes in the particle size are something else.

▽ To match the polarimetric observations, in the Martian spring we need some particles which are larger than the average winter particle size, which is perhaps

0.1 mm. Such a redistribution of particle sizes is exactly what we might expect from biological activity. If there is growth and proliferation in the Martian spring, we can understand why larger particles should then exist.

▽ Yet a non-biological explanation of these seasonal and secular changes appears to be possible. We have mentioned (p. 265) that the dark areas seem to have systematically higher elevations than the bright areas, and that both may be covered with a finely granulated material resembling limonite. The existence of dust storms shows that winds can distribute dust over substantial distances on Mars. The seasonal changes in brightness and in polarization of the Martian dark areas can be understood in terms of seasonal changes in particle sizes. If, in local spring, the dark highlands are scoured by the winds and the small bright particles carried down to the deserts below, the dark areas will appear to darken and the average particle size in them will appear to increase. If, in the Martian autumn, the small particles are carried back up to the highlands by local turbulent winds, the contrast between the dark areas and bright areas will diminish, and the average particle size in the dark areas will decrease. Pollack and I have shown that the particle sizes and wind velocities needed for such an inorganic explanation of the wave of darkening are plausible for Mars.

▽ In this view, the reappearance of a dark area after being covered in a dust storm can be understood: the winds soon sweep clean the high lying dark area. In areas where the elevations are not greatly different, perhaps the shifting wind-blown sands sometimes uncover underlying dark material, sometimes cover over pre-existing dark material, thereby giving rise to that other Martian enigma, the secular changes. All of these changes can be understood in terms of the scattering properties of small grains of limonite with diameters roughly equal to the period at the end of this sentence.

▽ The possibility that all the seasonal and secular changes on Mars—previously attributed to biological activity on that planet—can be understood by winds and dust and elevation differences does not, of course, disprove life on Mars. In our studies of the earth with the Tiros and Nimbus meteorological satellites, we were unable to detect seasonal changes in cultivated crops or in forests. Earth-based observations of Mars can detect much finer contrast gradations than the meteorological satellites can, but at the same time their ability to resolve fine details is much worse than for the meteorological satellites. For such gross changes as have been observed on Mars to be due to biological activity there, life would have to be much more extensive on Mars than on earth. For all we know, the planet may have a complex variety of organisms, yet there may be no means of ascertaining this fact over interplanetary distances.

▽ A final category of modern evidence relating to life on Mars is provided by infrared spectroscopy. Organic molecules have characteristic absorption features at a wavelength near 3.5μ in the infrared. The reflection spectra of most terrestrial organic materials show such absorption features. As the photons penetrate a small distance into the sample before being reflected, they are absorbed by vibration of carbon-hydrogen groups in the sample. The American astronomer William

Sinton has observed similar features in the spectrum of sunlight reflected from Mars. Sinton finds three features that can be understood if material on Mars contains large quantities of the methyl group, CH_3, the methylene group, CH_2, and the aldehyde group, CHO, as constituents of larger molecules.

▽ Sinton's observations were very difficult to perform, because of the small amounts of infrared radiation reflected from Mars and the limitations in terrestrial infrared detector systems. Using the 200-inch Hale telescope at Mt. Palomar, California, Sinton seemed to show that the absorption features preferentially appear in the Martian dark areas. The existence of the Sinton bands has been confirmed by the Soviet astronomer V. I. Moroz, at the Crimean Astrophysical Observatory. But more recently, Sinton and the Canadian-American spectroscopist D. G. Rea, of the University of California, Berkeley, have suggested that two of the three Sinton bands may, nevertheless, be due to inorganic molecular contaminants in our own atmosphere, and not due to organic matter on Mars. The remaining Sinton band is apparently a true Martian feature. For the moment, we can only suspend judgment on whether organic molecules have been detected in the Martian dark areas. The problems in obtaining and identifying the Sinton bands illustrate the great difficulties in spectroscopic observations of faint features over interplanetary distances and the desirability of observations from closer range.

▽ The observations we have described are tantalizing. They are considered evidence for life on Mars by many. △ Although the information taken as a whole may indicate the presence of life on Mars, we will have a rigorous answer only when new observational techniques are developed.

Our success to date in mastering space enables us to plan experiments designed to solve this ancient mystery. ▽ We may observe Mars in many ways: from a fly-by vehicle which spends only an hour or less within close range of Mars; from an orbiter, a spacecraft in orbit about Mars which performs observations, perhaps over many months; from an atmospheric entry probe which determines information regarding atmospheric physics before it reaches the Martian surface, but which performs no experiments on the surface; from a small soft-lander which performs a few experiments of biological relevance over a short period of time; and from a larger, automated biological laboratory which performs a wide range of interconnected biological investigations over a long period of time on Mars. Any strategy of Martian exploration will involve some mixture of these systems. △ A fly-by carrying an automatic camera can scan the Martian landscape from a distance of several thousand kilometers. The photographic images can be transmitted, ▽ slowly, point-by-point, △ to Earth by television. The experiment of the Soviet lunar rocket Luna III, which photographed the far side of the Moon [Chapter 21, Figures 21–7 and 21–8], indicates the complete feasibility of such an experiment.

▽ A typical contemporary American vehicle is Mariner IV, shown in Figure 20–5. It was designed to take 22 pictures of Mars from a distance of a few thousand kilometers. Planetary photography was the only Martian experiment directed to the surface of Mars from Mariner IV. The unsuccessful Soviet fly-by spacecraft

Mars-1 is shown in Figures 20–6 and 20–7. In addition to television equipment, Mars-1 was equipped with infrared and ultraviolet spectrometers, and a system to detect long wavelength radio emission. △ Fly-by photography will allow us to study details as small as one kilometer across on the Martian surface. We will then be able to study the surface of Mars in the same detail with which astronomers can, using Earth-based facilities, study the surface of the Moon. After such investiga-

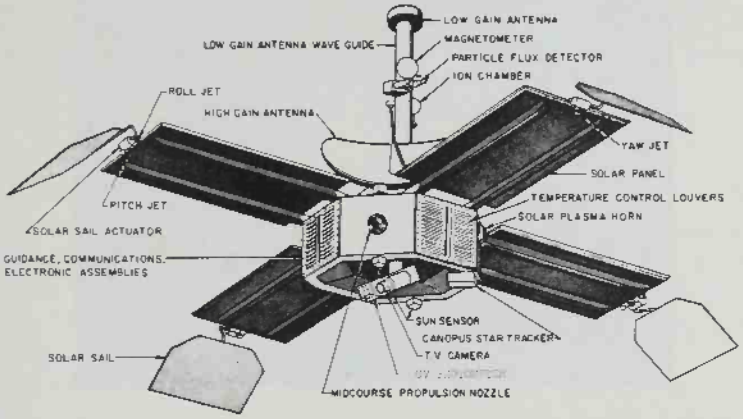
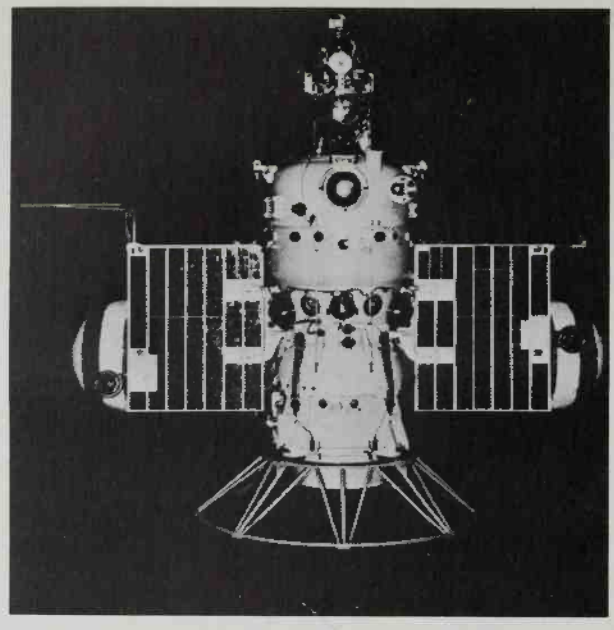


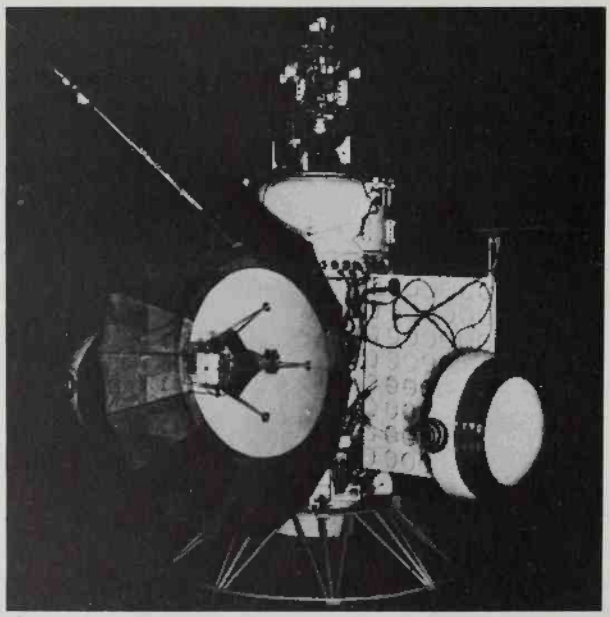
FIGURE *20–5. Diagram of the United States spacecraft Mariner 4 which flew by Mars on 14 July, 1965. Most of the experiments depicted were designed to measure interplanetary particles and magnetic fields. The ultraviolet photometer was an experiment removed from the spacecraft before launch. (Courtesy of NASA.)*

tions are carried out, many of the problems concerning the nature of the Martian dark areas and the controversial canals will be resolved.

▽ On 14 July, 1965 the United States spacecraft Mariner IV successfully flew by Mars, performing a variety of scientific observations. In one elegantly simple experiment, the spacecraft flew behind Mars, and its radio signal to Earth was gradually eclipsed by the planet's atmosphere. From the rate of fading of the radio signals, information on the temperature and pressure variation of the Martian atmosphere was obtained. Experiments directed at finding a planetary magnetic field and the associated Van Allen radiation belts gave negative results. The absence of a magnetic field on Mars is of some interest. The earth's magnetic field is thought to arise from its liquid iron core, formed through geological time by the migration of iron downward through the surface and mantle. The absence of a magnetic field on Mars suggests that the iron on Mars has not made a similar migration, and therefore that substantial quantities of iron may still exist near the Martian surface. This may be the explanation of the limonite, an iron oxide, which seems to exist on Mars.

▽ Mariner IV successfully acquired some fourteen or fifteen photographs of the Martian surface in the day-lit hemisphere. No usable information was acquired from photographs of the night side. In Figure 20–8 we see three renditions of the





FIGURES 20-6 *and* 20-7. *Two views of the Soviet spacecraft Mars 1, which was launched on an unsuccessful voyage towards Mars on 1 November, 1963. Mars 1 had a much larger and more sophisticated instrumental payload than did Mariner 4, as is evident by comparing these photographs with Figure 20–3. The Soviet spacecraft Zond 2 may have been similar in design to Mars 1. (Courtesy of Sovfoto, Moscow.)*

FIGURE *20–8. Three transcriptions of Frame 1 of the Mariner IV picture sequence of the Martian surface. A possible cloud may be visible above the planetary horizon in the last two renditions. The center of the picture corresponds to appoximately 35°N. latitude, 172°E. longitude, primarily a desert area. (Courtesy of NASA.)*

same first frame of the Mariner photographic sequence. The radio signals from Mars gave, as in the transmission of newspaper wirephotos, information on the darkness of each bright or dark point comprising the picture. The numbers can then be rendered into a picture and the apparent contrast of the Martian surface features can be increased or decreased on Earth at will. Three different choices of surface contrast are shown here. Among the bright and dark markings which we see, there is a curious dark line paralleling the horizon. Its nature is unknown. A bright patch may also be seen above the horizon in the sky in one of these pictures. Whether this is a dust cloud sitting in the Martian atmosphere, or an optical defect in the lens system is still an open question.

▽ The area covered by the Mariner IV photographs comprises primarily Martian deserts west of Amazonis and eventually regions of the dark area Mare Sirenum [cf. Figure 19–4]. In the early photographs of the bright areas, the Sun was approximately overhead; shadows were short, and details were difficult to see. In Figure 20–9, a rendition of Frame 7, several circular markings can be seen. Figure 20–10 is a rendition of Frame 11, taken near Mare Sirenum, when the Sun was at a low angle, permitting longer shadows. It is now quite clear that the circular markings previously seen are craters like those on the Moon.

▽ The large lunar craters [see Chapter 21] are almost certainly produced by the impact of objects many kilometers across onto the surface of the Moon. Many of these impacting objects are believed to be fragments of asteroids wandering more or less erraticly in the inner part of the solar system. Since Mars is much closer to the asteroid belt than is the moon, it should be subject to many more impacts—perhaps 25 times as many. Yet the number of craters of a given size in a given area on the Martian surface is no larger than the comparable number on the moon. This must mean that processes exist on Mars which efficiently erode even large impact craters. In Figure 20–10 we see a very large crater, over 100 kilometers across, whose ramparts have been seriously breached. Major erosion of its walls has occurred. Is the erosion due to windblown dust, or perhaps, as on earth, to running water? We find that the craters on the Martian bright areas are shallower, more eroded, and have their bottoms more filled in than the craters in the Martian dark areas. This is an explicable circumstance when we recall that the bright areas prob-

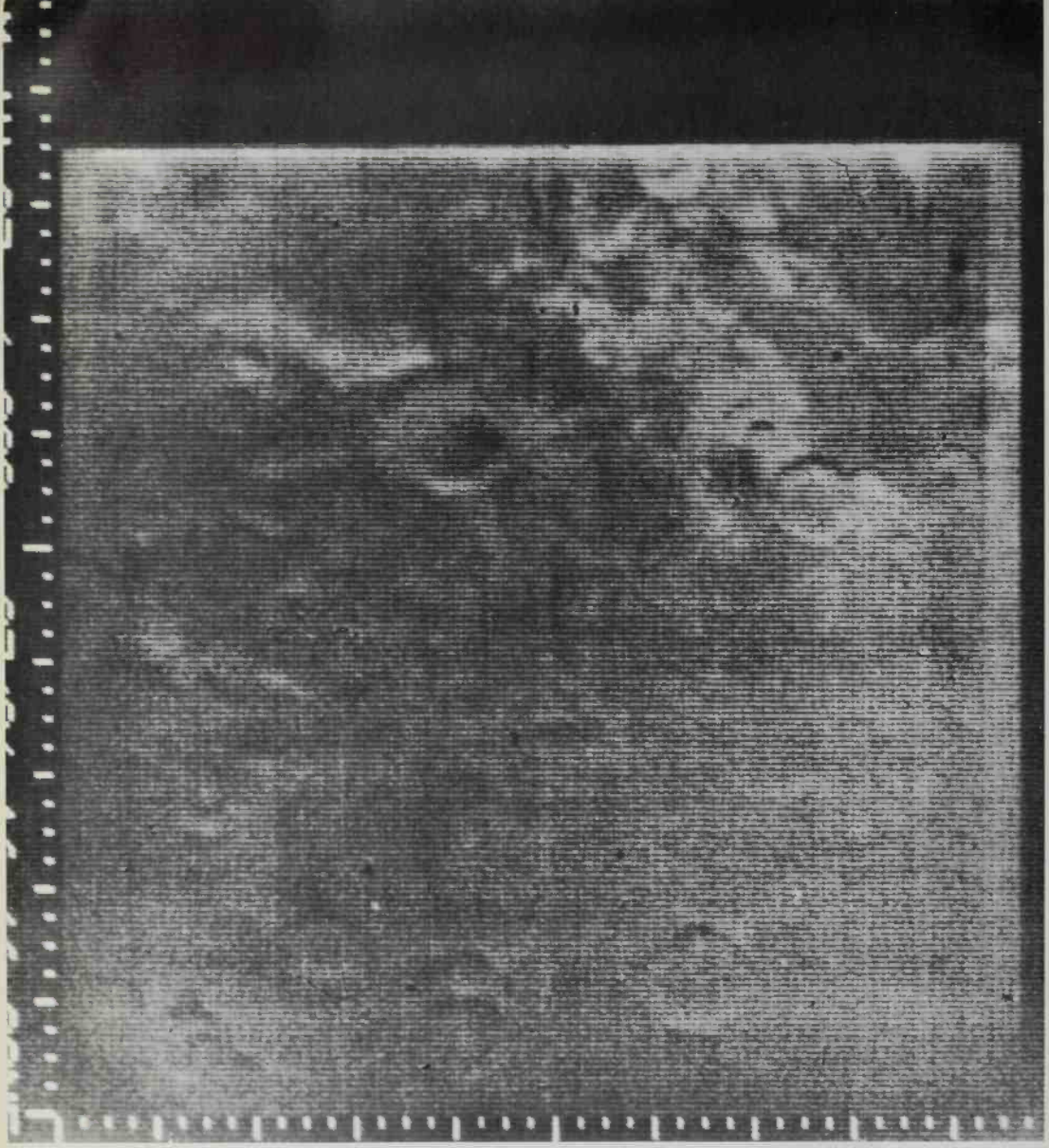


FIGURE 20–9. *A transcription of Frame 7 of the Mariner IV photographic sequence. The Sun is now 29° from the zenith, and circular markings are beginning to become visible. The center of the picture corresponds to approximately 13°S. latitude, 186°E. longitude, primarily a desert area. (Courtesy of NASA.)*

ably have substantial amounts of drifting dust, which tend to fill in and erode craters formed there.

▽ There is certainly no extensive liquid water on the Martian surface today; however, it is not out of the question that water erosion was important hundreds of

millions of years ago. Because of the efficiency of Martian crater erosion—regardless of the mechanism—the surface we see is not that of a very ancient Mars, and it is therefore impossible to judge from the Martian geology viewed in these photographs whether there were extensive bodies of liquid water in the early history of Mars. Some curving depressions which look very much like fluid flows can be seen in some of the photographs, e.g. in Frame 11, Figure 20–10; this need not necessarily be due to running water. Although it is not easily visible in the reproduction of Figure 20–10 shown here, a straight line extending from the lower left to the middle right, traversing the large eroded crater, can be seen on the original. Whether this feature has anything to do with the classical Martian canals, or whether it is an easily understood feature such as a fault, remain open questions.

▽ The best ground resolution obtained in the Mariner IV photographs was a few miles. We have already seen in Chapter 18 that photographs of the planet Earth with comparable resolution give no sign of life, intelligent or otherwise. The Mariner IV television experimenters, Robert B. Leighton, Bruce C. Murray, and their colleagues, have been careful to emphasize that the Mariner IV photographic experiment was not designed to search for life on Mars, and that it neither demonstrated nor precluded the existence of life on that planet. Neither has it resolved the canal controversy. It has demonstrated the utility of remote planetary photography and pointed out the need for greatly improved resolution in future space missions.

▽ A Mars orbiter permits us to garner information about even smaller scale features on Mars; but, more important, it permits us to gather this information over many months. Since the bulk of the astronomical evidence suggesting life on Mars is seasonal in character, it would be of enormous interest to examine the Martian dark areas, for example, during the passage of the wave of darkening. The best opportunities for such investigations lie in the years 1969 and 1971. Perhaps we can gain information on the distribution of organic matter over the surface of Mars. Some organic molecules may characteristically be present in one dark area; other organic molecules, in other dark areas. From an orbiter, observations with a resolution of 10 meters or better should be possible. Observations of the Earth with similar resolution show unambiguous signs of life, although most of these are signs of human habitation. Due to the fact that Mars is an exterior planet, lying further from the Sun than the Earth, no one has ever observed any region of Mars in the middle of local night. If life on Mars is not distributed over the whole planet, but is localized in a few favored high-temperature, high moisture environments, an orbiter may be the ideal vehicle for finding such "hot spots." In infrared night-time observations, a large Martian hot spot would be easily detectable. The potentialities of an orbiter for Martian exploration are many; there seems little doubt that orbiters will be used in the early stages of the biological investigation of Mars.

▽ But before very long, we will want to land scientific instruments on the Martian surface. In addition to the simple question, "Is there life on Mars?" biologists are interested in the anatomy, physiology, genetics, biochemistry, ecology, and behavior of Martian organisms, to name only a few sub-disciplines. Such information can only be acquired on the spot. In the design of instrumentation for

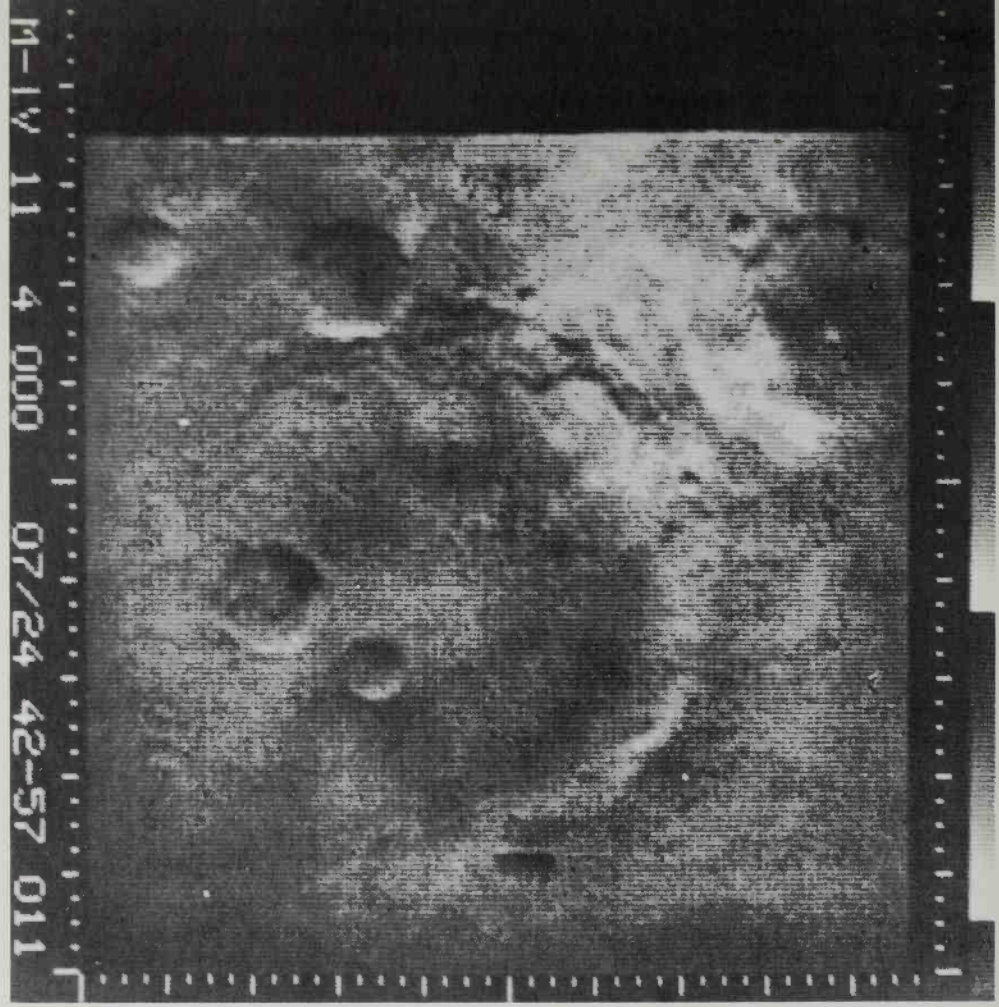


FIGURE *20–10. A transcription of Frame 11 of the Mariner IV photographic sequence. The Sun is now 47° from the zenith, and impact craters are clearly visible. At the Northern ramparts of the large central eroded crater, a sinuous flow pattern may be seen; a straight line exists traversing the Southern ramparts of this crater diagonally. The center of the picture corresponds to approximately 31°S. latitude, 197°E. longitude, a region localized in a dark area. The contrast of features in this frame is difficult to compare with the contrasts of other frames. (Courtesy of NASA.)*

the detection and characterization of life on Mars—a major occupation of some biologists today—there are two fundamental questions: First, is life on Mars ubiquitous, or localized in only a few areas of Mars? Second, how close are the Martian life forms to those on Earth? Can an automated biological laboratory land anywhere on Mars, or are some places vastly preferable to others? We have presented arguments that some locales, such as the edge of the retreating polar ice cap during Martian spring, are favorable habitats for Martian organisms. Other areas, such as Syrtis Major and Solis Lacus, show striking seasonal or secular

changes. Yet the Martian winds should distribute small organisms essentially uniformly over the planet, and we find on Earth today that microorganisms can be found in essentially every locale, from the Sahara and Gobi deserts to the Mindanao Deep; from the top of Mt. Everest to the top of the Empire State Building. Human beings are not so uniformly distributed, and an extraterrestrial expedition to Earth, seeking indigenous life, would be well advised to look for microorganisms: there are more microorganisms, and they are more readily caught. Yet any biological inventory of Earth should, we know, consider organisms larger than microbes.

▽ How, in fact, will an automated biological laboratory detect life on Mars? One set of possible experiments, which are actively being pursued at the time of writing, involves landing a nutrient medium on Mars, inoculating it with samples of Martian soil, and looking for signs of growth and reproduction. As the microorganisms grow, they may increase the turbidity of the nutrient medium or change its acidity. Alternatively, the Martian microorganisms, like many terrestrial organisms, including people, may give off carbon dioxide in the course of metabolizing the food brought from Earth. The Martian biological experiment, Gulliver, shown in Figure 20–11, is such a CO_2 monitor. But what if the Martians fail to find the food sent from Earth palatable? What if their tastes are more exotic? The nutrient medium will be inoculated, but the Martian organisms present will not grow; there will be no turbidity changes, no acidity changes, no CO_2 given off. The experiment will send negative results back to Earth. Shall we conclude that there is no life on Mars?

▽ An alternative experiment is to look for particular categories of enzymes on Mars. We have seen that phosphorus compounds play a fundamental role in metabolic energy transfer and other terrestrial metabolic activities. If phosphorus is also important in Martian metabolism, we might expect enzymes known as phosphatases, which transfer phosphorus groups in metabolism, to be present in the Martian soil, as they are in terrestrial soil. A device called Multivator, designed to search for phosphatase and other signs of Martian metabolism, is shown in Figure 20–12.

▽ But what if the Martian organisms do not contain phosphatases? Phosphorus is present in terrestrial organisms in an abundance far out of proportion to the cosmic abundance. Perhaps some other atom takes the place of phosphorus on Mars. The fluorescence technique used to search for phosphatases will give spurious positive results for some relatively rare minerals. Perhaps such minerals are present on Mars, and the positive result will not actually indicate the presence of phosphatases on Mars. It is for such reasons that a combination of different experimental approaches in an automated biological laboratory is necessary for a thorough search for and characterization of life on Mars. There is no single "life detector."

▽ As an example of the advantages which a combination of experimental techniques provides, consider the following device, which is under preliminary study in the United States: Individual particles of the Martian soil are made to

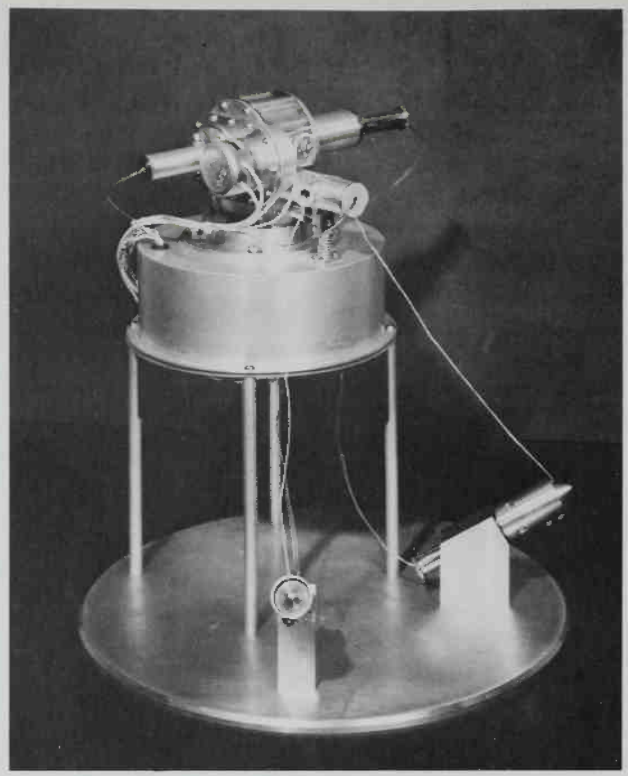


FIGURE 20-11. *Model of the Mark III Gulliver. The two projectiles on the disk are the sample collectors. When fired they would drag a length of sticky string over the Martian surface, which, when reeled back into Gulliver, would have samples of the surface adhering. The strings are then drawn into a nutrient broth containing radioactively labelled carbon compounds. If Martian microorganisms adhering to the string metabolize the labelled broth and release carbon dioxide, this event would be recorded and radioed back to Earth. (Courtesy of Dr. Gilbert V. Levin, Hazelton Laboratories, Inc., and Prof. Norman H. Horowitz, California Institute of Technology.)*

adhere electrostatically to a moving belt, which is carried through an infrared spectrometer. Individual particles 0.1 mm in diameter and smaller are automatically analyzed by infrared spectroscopy. If their infrared spectrum is characteristic of minerals, as the spectra of most of the particles will be, the belt moves on to the next particle. But when a particle is scanned which has the infrared spectrum of organic matter, it is also photographed through a microscope. Such a device can, in principle, sift through large numbers of uninteresting particles to determine something about the chemistry and morphology of what we consider the interesting particles. Only the spectra and photographs of the interesting particles would be transmitted by radio back to Earth. If we now imagine this procedure amplified, with many devices examining samples of Martian soil for their physical and chemical properties, their possible metabolic activities, and their responses to new

FIGURE 20-12. *Cutaway view of a test model of Multivator. Martian dust would be drawn up by a small vacuum cleaner, and deposited into a number of different reaction chambers, each containing its own enzyme or metabolite. Each chamber is then observed by a photometer for changes in turbidity due to growth of the acquired Martian organisms, for fluorescence due to the presence of enzymes, etc. (Courtesy of Prof. Joshua Lederberg and Dr. Elliot Levinthal, Stanford University Medical School.) A somewhat similar device, not pictured here, called Wolf Trap, was designed by Prof. W. Vishniac of the University of Rochester.*

stimuli, we see that an automated biological laboratory controlled by a computer can be a very powerful tool in the quest for life on Mars.

▽ In order that high-risk, high-return possibilities are not completely overlooked, such a device should also obtain periodic television scans of the Martian landscape. A laboratory of this type should be mobile—we may think of it as a small tank. It is heavy, but well within the range of possible payloads deliverable to Mars by the booster rockets being planned by the United States and the Soviet Union for manned exploration of the Moon. Perhaps the television pictures will be unspectacular—rocks, lava flows, sand dunes. An occasional scraggly plant would not be unexpected. But there are other possibilities—fossils, footprints, minarets. . . . We will know only when we drop our instruments on the surface of Mars. △

21

The Moon

The Surface of the Moon then is found, by the least Telescopes of about three or four Foot, to be diversified with long Tracts of Mountains, and again with broad Valleys. For in those Parts opposite to the Sun you may see the Shadows of the Mountains, and often discover the little round Valleys between them, with a Hillock or two perhaps rising out of them. Kepler from the exact roundness of them would prove that they are some vast work of the rational Inhabitants. But I can't be of his mind, both for their incredible Largeness, and that they might easily be occasioned by natural Causes. Nor can I find anything like Sea there . . .
What then, is it credible that this great Ball was made for nothing but to give us a little Light in the Night-time, or to raise our Tides in the Sea?

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

Though I am old with wandering
Through hollow lands and hilly lands,
I will find out where she has gone,
And kiss her lips and take her hands;
And walk among long dappled grass,
And pluck till time and times are done
The silver apples of the moon,
The golden apples of the sun.

William Butler Yeats, *The Song of the Wandering Aengus*

▽ Each of the nine planets, their thirty-one satellites, and the innumerable smaller objects of our solar system has its own individuality. While certain broad relationships exist—for example, within the terrestrial and Jovian planetary groups—the differences are more striking. To the interplanetary explorer of the next century, the differences among these objects will be far more vivid than, for example, the differences among the major ports of call of the last century. In this and the following two chapters, we will briefly describe the environments of these diverse worlds—within the rather restricted limits of present knowledge—and examine them for the possible presence of life.

▽ Our nearest neighbor is our familiar satellite, the Moon. It revolves about the Earth, of course, about once a month. Since, from our terrestrial vantage point, we always see approximately the same face of the Moon, we conclude that it must be rotating about its axis at the same rate that it revolves about the Earth, thus always keeping the same face to the Earth. This reluctance of the Moon to expose her backside is a coyness probably shared by the other thirty satellites in our solar system. Such synchronous rotation is due to tidal friction—that is, to the tides introduced in the body of the satellite by the more massive planet. Operating over the four and a half billion year lifetime of our solar system, such body tides are very efficient in producing synchronous rotation. Once the periods of the satellite rotation and revolution are equal, however, the tidal forces are usually unable to slow down the satellite's rotation any further.

▽ While the Earth has slowed the Moon's rotation, tides produced by the Moon in the body and oceans of the Earth have slowed the rate of rotation of the Earth. This tidal "braking" increases the day by about 0.002 second per century, an amount which, while incredibly small, is still within the range of detectability of astronomical techniques. The tidal retardation of the Earth has produced no perceptible lengthening of the day during the span of Man's sojourn on this planet; but one billion years ago, if the rate of retardation is constant, the day must have been 2×10^5 sec $yr^{-1} \times 10^9$ yr $= 2 \times 10^4$ seconds, or roughly six hours shorter. An apparent confirmation of these astronomical deductions on the tidal braking of the Earth has come from an unlikely source.

▽ In the Bahamas, there is a reef coral called *Acropora palmata,* which has ringed ridges on its skeletal structure. The ridges are formed by the growth of the coral, and each ridge corresponds to one year's growth. When the rings are examined in closer detail, it is found that they are composed of large numbers of much finer ringed markings—roughly 360 to the annual band. The American geologist John W. Wells, of Cornell University, has postulated that these finer rings represent a daily growth of the coral.

▽ Now consider a sample of the coral from much earlier times—say, the Middle Devonian, a time about 350 million years ago. The length of the year should not change with time. But if the length of the day was shorter in the Middle Devonian than it is now, we should expect to see more fine lines per yearly band in a Middle Devonian fossil than in a contemporary fossil. Wells has in fact examined Middle Devonian fossils, and finds that they have about 400 ridges per year. Thus, 350 million years ago, there were some 400 days per year, and each day was about (365/400) × (24 hrs) = about 21.9 hours long. The astronomical data give essentially the same result for the Middle Devonian. This is one of many examples of the connection between astronomy and biology. Indeed, this book is devoted to the examination of such a connection.

▽ Everyone knows that through a small telescope the Moon looks something like the view in Figure 21–1. There are bright areas and dark areas, the so-called *continentes* and *maria,* designations dating from the time of their discovery by Galileo, who thought that the dark areas were indeed lunar bodies of water. We now know that there is no liquid water at all on the lunar surface, and that the *maria* are dark, relatively flat, depressions.

▽ In order to study the bright and dark areas more accurately, new photographic methods are being used. A photograph of the Moon is projected onto a blank globe and then rephotographed from any desired angle. Such a procedure, called rectification, removes the effects of foreshortening near the edges, or limbs, of the Moon. The rectified photograph in Figure 21–1 is therefore a view of the Moon that no human being has yet seen. At the center of the figure appears the great lunar rayed crater Tycho, which ordinarily appears far to the south, in naked eye observations or astronomical photographs. The bright rays may be seen emanating from Tycho and traversing substantial distances across the lunar surface. Figure 21–1 is based on a photograph taken at full moon—that is, when the Moon, the observer, and the Sun are approximately in a straight line, with the observer in the middle. At full moon, the lunar surface takes on the high-contrast guise seen in Figure 21–1. The rays of craters like Tycho then become extremely prominent.

▽ At other times, such as half moon, however, when the sunlight is striking the center of the lunar disk from an angle, a given area of the Moon seems much less bright, the contrast between *maria* and *continentes* declines, and the rays all but disappear. In other photographs of the Moon, Tycho, apart from its rays, is in fact a very modest and unspectacular crater.

▽ A more usual photograph of a smaller region of the Moon is shown in Figure 21–2, a non-rectified photograph of the region of Mare Imbrium, a large circular *mare* in the northwest quadrant of the Moon. We see that the *mare* floors are rugged, occasionally pockmarked with craters of all sizes. Towards the bottom edge of this picture is the lunar terminator, the range of locales on the lunar surface at which the Sun has just set. In the late lunar afternoon, shadows of mountains become long, as in the lower left-hand corner of the photograph, and it is possible to compute from the length of these shadows the height of lunar

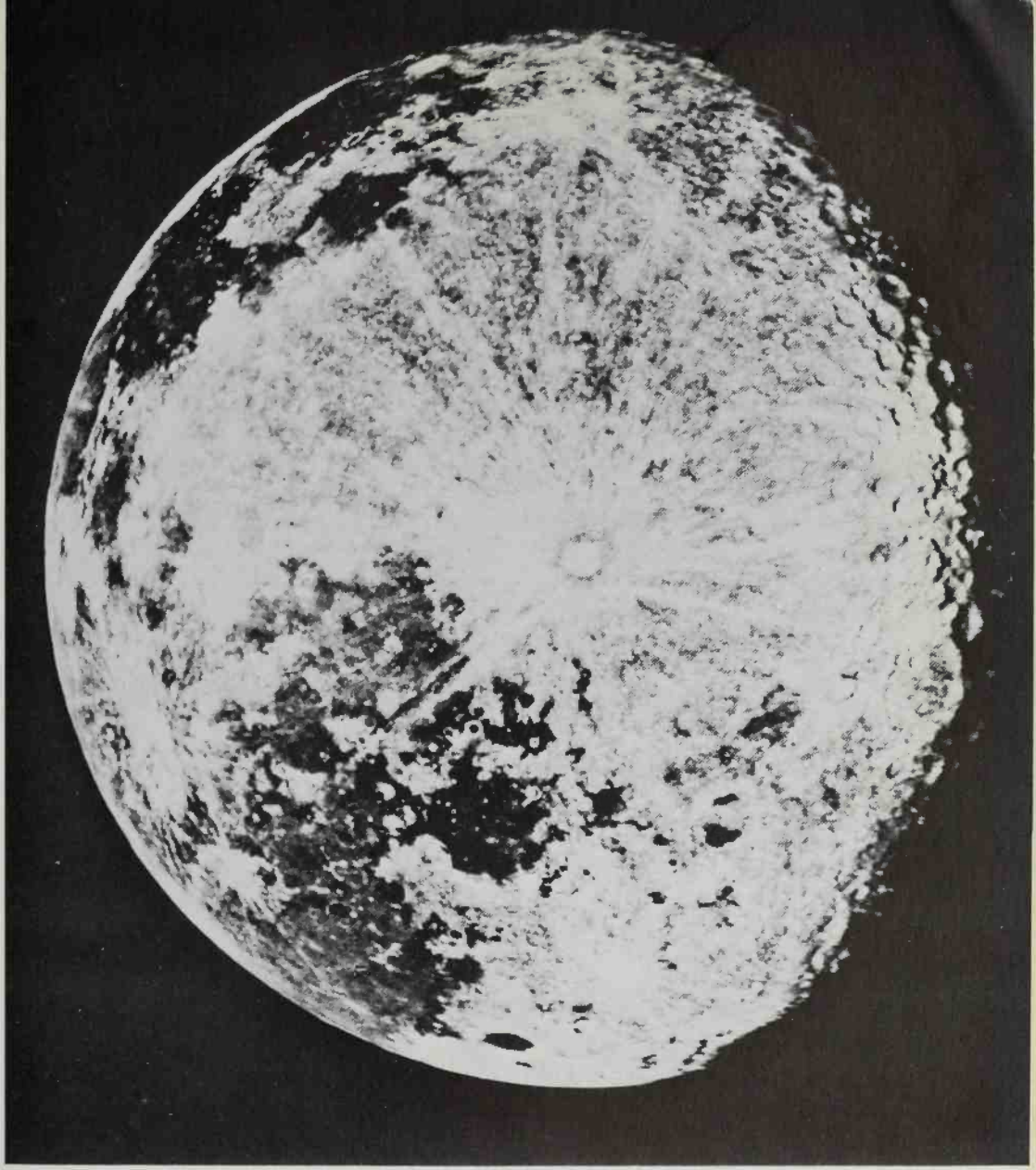


FIGURE 21–1. *Rectification of a full moon photograph showing the rayed crater, Tycho, at the center of the disk. (Courtesy of Dr. Ewan Whittaker, and Dr. G. P. Kuiper, Lunar and Planetary Laboratory, University of Arizona.)*

features. In this way, it has been possible to determine that there are mighty mountain ranges on the Moon, some of them approaching, if not surpassing, the altitude of the Himalayas. The mountains in the upper left-hand corner of the picture are called the Alps, after their terrestrial counterparts; the slash through the Alps, going towards the extreme left-hand corner of the picture, is called the

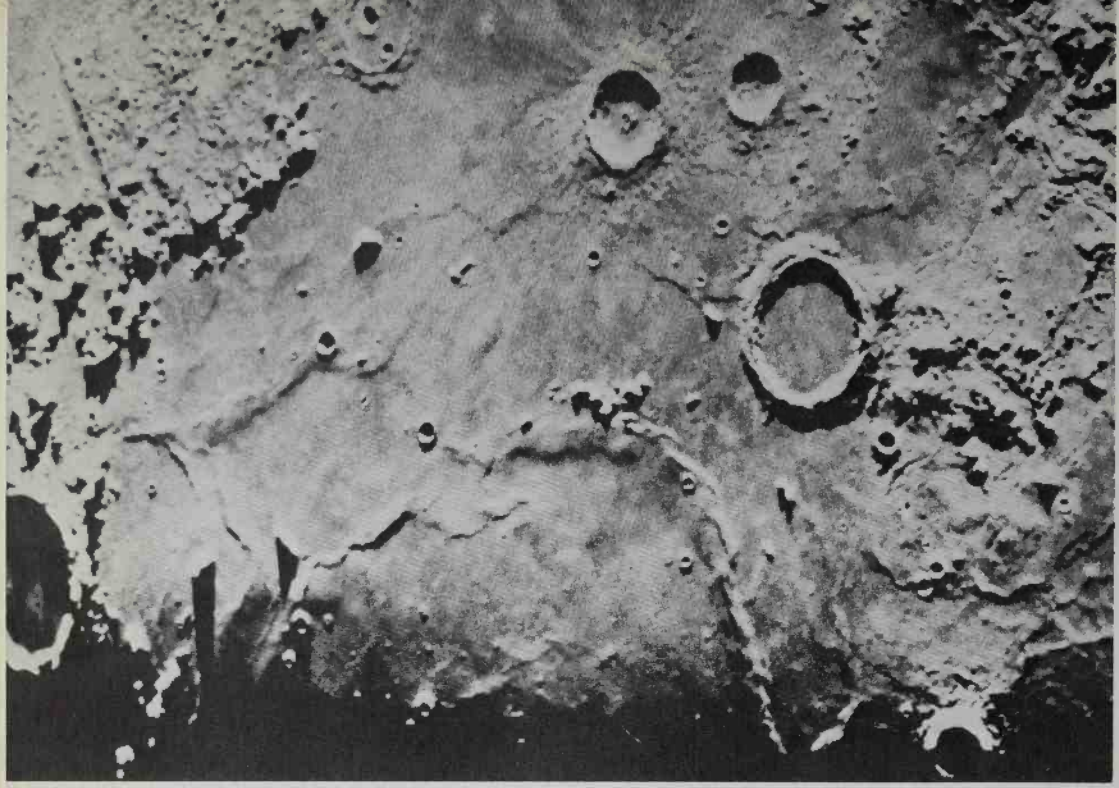


FIGURE *21–2. Unrectified photograph of the region of Mare Imbrium.* (*Yerkes Observatory photograph courtesy of Dr. G. P. Kuiper, Lunar and Planetary Laboratory, University of Arizona.*)

Alpine Valley. Close inspection of the Alpine Valley reveals it to be filled with a set of closely spaced craters, unlike typical valleys known on Earth.

▽ Far superior resolution of the Moon has been obtained from the U.S. Ranger spacecraft. A typical photograph of the Moon obtained by Ranger VII appears as Figure 21–3. The dense clustering of craters in this photograph occurs along the path of a lunar ray, much as the craters of the Alpine Valley are arranged along the cut through the Alps. Many of the lunar craters, such as Archimedes, the large crater towards the center of Figure 21–2, are more appropriately described as ring walls, or walled plains. Because the radius of the Moon is so small, the lunar horizons are much closer to the observer than on Earth. If one stood at the center of a large lunar crater, or ringed plain, the walls would be beyond the horizon and out of sight.

▽ The relation between the depth and width of the lunar craters follows the same mathematical law which impact craters of all sizes follow on the Earth. This and other evidence have convinced the majority of astronomers that the larger lunar craters have been formed by the impact of some objects from interplanetary space.

▽ The alternative view is that the lunar craters are of volcanic origin. A study of the close-up photographs of the Moon by the Ranger series of space vehicles has recently convinced many students of the Moon that the smaller lunar craters—

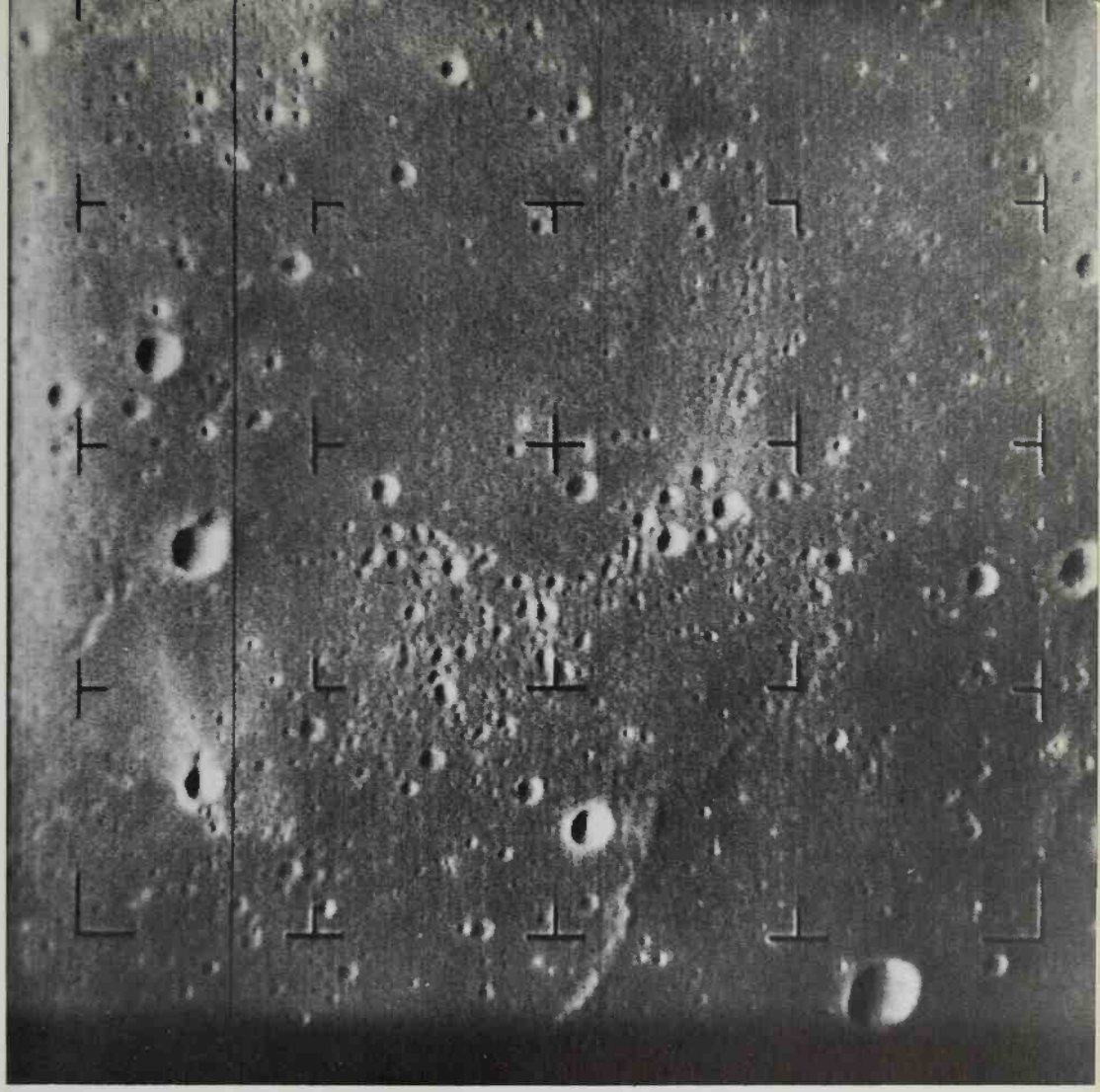


FIGURE *21–3. Ranger VII photograph of a cluster of craters in the floor of the newly named Mare Cognitum. (Courtesy of NASA.)*

particularly those which are too small to be visible from the Earth—may be in part of volcanic origin. One line of evidence supporting this view can be seen in Figure 21–4, a close-up of the lunar crater Alphonsus. The dark region to the center and below is the crater floor. In the crater floor we see a set of rills—slightly meandering cracks in the lunar surface. Oriented along one of these rills is a sequence of at least six fair-sized craters. This correlation between the rills and the craters is too striking to be attributed to mere chance. Either the craters caused the rills, the rills caused the craters, or both have a common cause. It has been argued that the most likely of these alternatives is that both rills and volcanic craters are formed along faults by stresses in the lunar surface. In some cases long chains of craters have been observed which are far too orderly to be an accidental configura-

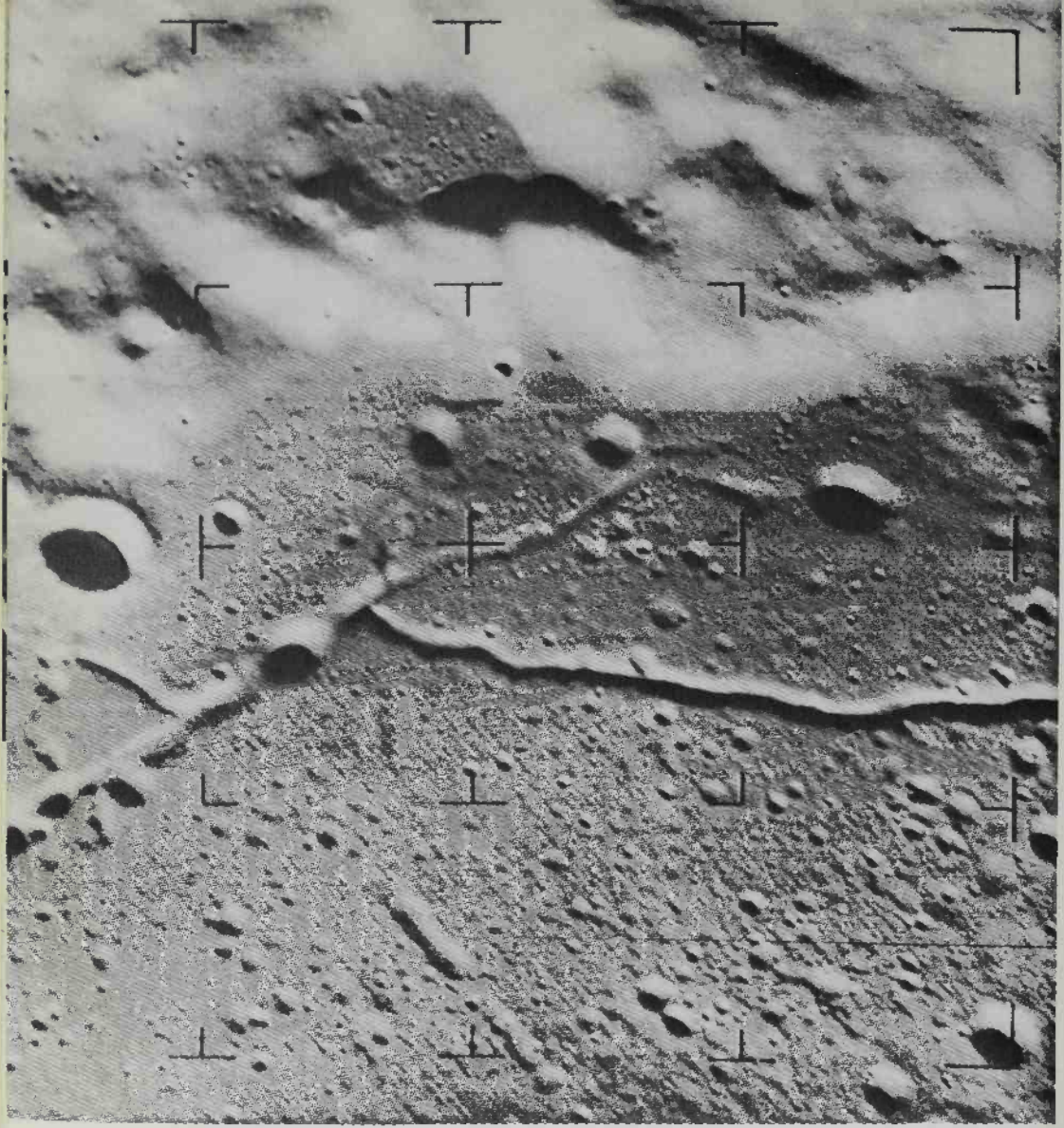


FIGURE 21-4. *Ranger IX photograph of the eastern edge of the floor of the lunar crater Alphonsus. The photograph was taken from an altitude of 115 miles above the lunar surface, at a time 1 minute and 17 seconds before impact. (Courtesy of NASA.)*

tion produced by impacting projectiles from interplanetary space. (In the vicinity of ray craters, however, the clustering of craters along the ray can be understood as the result of the debris flung out by the explosion forming the crater.)

▽ Some—perhaps all—of the *maria* show a roughly circular shape. Figure 21–5 is a rectified photograph of Mare Humorum, in the southwest quadrant of the lunar hemisphere which faces the Earth. The *maria* may in fact be nothing more than a group of very large craters, produced by just the same impact mechanism

FIGURE 21-5. *Rectification of a Meudon Observatory photograph of Mare Humorum. Note the partially destroyed circular craters around the periphery of Mare Humorum. (Courtesy of Dr. G. P. Kuiper, Lunar and Planetary Laboratory, University of Arizona.)*

that produced the craters. The Alpine Valley would then be exactly analogous to a lunar ray, a gash carved out of the lunar Alps by debris from the giant impact which formed Mare Imbrium.

▽ If the lunar craters are formed by impact of large meteorites and asteroidal-sized objects, should we not see an occasional crater being formed? Should astronomers invest time in the comparison of photographs of the same region of the Moon, taken years apart, to look for the appearance of new craters? We can easily convince ourselves that this is a profitless enterprise. There are about 10^5 craters visible in the lunar hemisphere which faces the Earth, on photographs obtained with large astronomical instruments. If we assume that these craters were formed uniformly throughout the last 5 billion years of lunar history, then we see that new craters should be formed about once each $(5 \times 10^9)/10^5 = 5 \times 10^4 = 50{,}000$ years. We should wait, on the average, some 50,000 years before seeing the next crater formed; even then, it is unlikely to be very spectacular.

▽ Similar impacts must have occurred on the Earth during the same 5×10^9 year period. Yet the Earth is not prominently marked with large circular craters. This discrepancy can be understood entirely in terms of erosion by wind and water

on the Earth's surface. Structures become eroded on the Earth in periods of time very short compared with the history of the planet. In fact, in recent years, such structures as the Riss Kessel, in Germany, have been identified as fossil meteor impact craters. The number of such recent craters on the Earth, extrapolated through its history, gives a total number of terrestrial craters entirely consistent with the number on the lunar *maria*.

▽ Why are some craters rayed, like Tycho, and others, like Archimedes, not, if both types of craters are formed by the same process—the collision with the Moon of some interplanetary derelict? There is now reason to believe that erosion occurs on the lunar surface, both from the impact of innumerable micrometeorites on the Moon each day, and by the re-impact on the Moon of lunar material ejected after collisions of larger objects. Some sign of this erosion can be seen in Figure 21–5. In and around Mare Humorum are several "ghost craters," circular features which evidently were once craters, but which now have been partly eroded away. It is also possible that lava flows attended at least some of the earlier and larger lunar impacts, thereby obscuring preexisting features.

▽ Another eroding influence is solar radiation and the solar wind. When materials are bombarded by solar ultraviolet radiation and x-rays, and the ejected charged particles of the solar proton wind, their crystalline structure tends to break down; the recombination of these molecular fragments forms chemical groups which impart color to the objects; and we are left with a very dark, fine powder. Spectroscopic searches have failed to find any sign of a lunar atmosphere. When a cosmic radio source such as the Crab Nebula passes behind the Moon, it instantaneously "winks out," rather than slowly fading, as it would if the Moon possessed even a modest atmosphere. Especially in the absence of an atmosphere, small particles tend to sinter; that is, they become welded to each other, not solidly but rather with only one or two points of contact. The result is a low-density fluff of unimaginable complexity, which has been called a "fairy castle structure." Such a material explains many of the spectroscopic, polarimetric, and radio properties of the lunar surface.

While the lunar surface is known to be covered by such low-density material, there is still some debate on the depth of this layer. Most astronomers believe that its thickness is of the order of a few centimeters or less. Thomas Gold believes that the depth may run to kilometers. Such a material, even if very deep, has an appreciable bearing strength, and would probably feel "crunchy" to the first astronaut who treads it.

▽ When a small asteroid, let us say, impacts the Moon, it blasts out a fragment of lunar surface material, some of which escapes to space; the remainder is distributed over the lunar surface. The larger chunks produce secondary craters when they impact. The material of the asteroid and the impact site are both ground into a fine rock flour, which, when distributed over the lunar landscape, gives the appearance of the rays. As time progresses, the ravages of solar radiation and the solar proton wind will cause a gradual deterioration in the brightness of the rays. In about 10^6 years, the highly reflecting character of the ray material will be

destroyed, and except for the secondary craters there will be no sign of their previous existence. It may be that ray craters such as Tycho are not very many millions of years old.

▽ We have already alluded to laboratory experiments, in which fine powders were irradiated in vacuum by protons simulating the solar proton wind. In experiments of this type, performed by the American astronomer Bruce Hapke of Cornell Universty, it is found that, almost independent of the composition of the irradiated material, the powder becomes as dark as the lunar surface is today after the equivalent of 10^6 years' solar proton irradiation. Continued irradiation tends to make the materials darker yet. Thus, to explain the fact that the Moon is not even darker than it is, we must postulate that underlying brightly colored material is stirred up, in timescales of a million years or less. This is clearly related to the erosion processes which we have already mentioned.

▽ We then have an interesting model of the superficial layers of the lunar surface. Underneath perhaps only a few centimeters of the dark, irradiated, sintered, pulverized material, is a layer of brighter sintered material which has not been recently exposed to the solar proton wind. It may be that the first astronaut to walk the surface of the moon will leave on the dark *mare* surface a trail of brighter white footprints, which will survive a hundred thousand years before irradiation and erosion inexorably extinguish their outlines.

▽ The darkening effect of solar radiation is, of course, not restricted to the Moon. It should apply to all bodies in the solar system which have little or no atmospheres and small magnetic fields (so that solar protons reach their surfaces). In this category are probably most of the satellites in the solar system, perhaps the planet Mercury, and the dust and debris which fill the spaces between the planets. The vast lanes of dust which fill the plane of the solar system, and which we see as the zodiacal light, are composed of very dark particles—undoubtedly darkened by solar protons.

▽ Erosion of the lunar surface tends to wipe away small craters in times short compared with the age of the Moon, but is unable to destroy the larger craters. Since there are more impacts of small particles than of large ones, there must be more smaller craters formed than larger ones. The combined result of impact and erosion must produce a certain distribution of crater sizes on the lunar surface. Studies of crater counts in the Ranger photographs are being used to reconstruct the history of lunar crater formation and destruction.

▽ One of the most striking conclusions of the Ranger photographs is the general uniformity of the lunar surface. Figure 21–6 is a Ranger IX photograph in the crater Alphonsus. The white circle is the impact area of the Ranger IX spacecraft on the lip of a small crater. Parts 1, 2, and 3 of Figure 21–6 were taken with progressively improved resolution. Thus the crater near which impact occurred, as seen in Part 1 is barely detectable, but in Part 3, it is clearly visible. What is striking is the similarity in form of craters of very different sizes and the general uniformity of the lunar landscape at different resolutions. The largest craters in Part 1 and Part 3 (Figure 21–6) are practically indistinguishable, yet one

FIGURE *21–6. The last three frames of the P1 camera of the Ranger IX spacecraft. Frame three was taken 0.45 seconds before impact from an altitude of three quarters of a mile above the lunar surface. (Courtesy of NASA.)*

is about ten times larger than the other. The best Ranger IX photographs were able to see detail as small as a few inches across. Unlike the Earth there is no qualitative novelty which appears on the lunar surface when we increase our ability to see fine detail. The Soviet spacecraft, Luna IX, soft-landed on the Moon, revealed a lunar landscape remarkably similar in general character to that seen with the incomparably poorer resolution obtainable with Earth-based telescopes.

▽ By noting which craters have partially destroyed other craters, and by counting the number of asteroidal fragments in the vicinity of the Moon, astronomers have been able to reconstruct something of the history of lunar crater formation. The craters within the *maria* can be accounted for by asteroidal fragment impact during the last few billion years. The lunar *maria* must have been formed earlier, probably three to four billion years ago. There are too many craters in the lunar *continentes* to be accounted for by contemporary impact rates. Also, enormous objects—many hundreds of kilometers across—are necessary to produce such major and lasting scars on the face of the Moon as the *maria*. Thus, many of the lunar surface features must be due to the impact of debris present in the vicinity of the Moon shortly after the time of its origin. It may be that this debris was the final fragments of the swarm of bodies which gravitationally condensed to form the Moon. △

▽ Figure 21–7 is a montage, prepared by the Anglo-American selenologist E. A. Whittaker, of the Lunar and Planetary Laboratory, University of Arizona. It is made from photographs taken primarily of the far side of the Moon by the Soviet cosmic rocket Luna III, when that side of the Moon was in direct sunlight. (The near side was therefore dark, at the phase that we on Earth call new moon. There is an unfortunate tendency to call the far side of the Moon the "dark" side of the Moon. This is obviously an error; it is probably related to the unconscious feeling of some people that the Earth, if not at the center of the universe, is at least the source of all light.)

▽ The large, circular *mare* at left center is Mare Crisium, and is identical, when rectified, with this *mare* as seen from the Earth. Except around the edges, where there is overlap with what is seen from the Earth, all the features are new, discovered by Luna III. A rectified and recentered photograph of the overlapping region—that photographed by Luna III and by ground based telescopes both—can be seen in Figure 21–8. Comparisons of Figures 21–7 and 21–8 show the quality of the Luna III photographs and give a good idea of the relation of the newly discovered features on the lunar surface to those known earlier. There seems to be

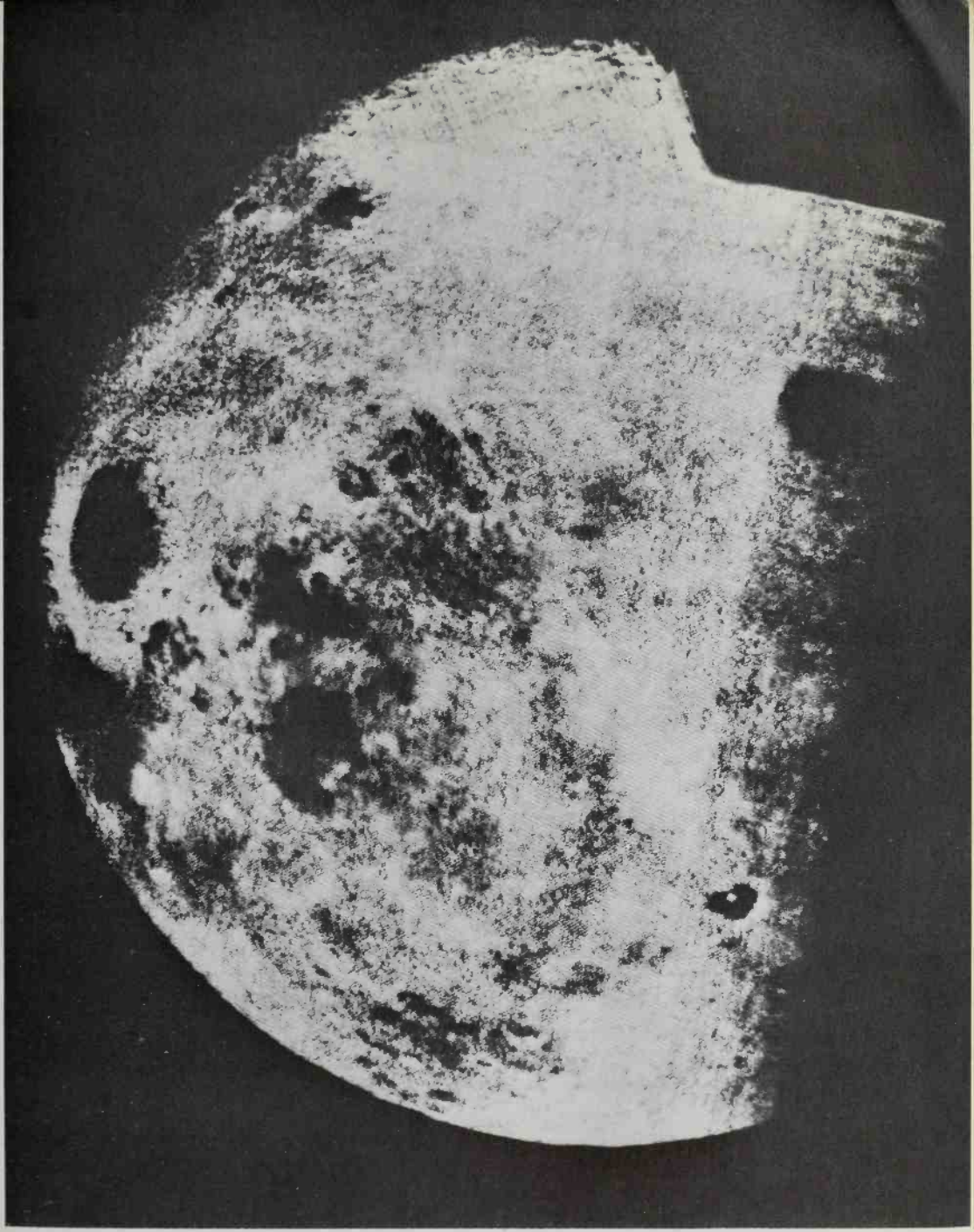


FIGURE 21–7. *A montage of Luna III photographs of the far side of the Moon. (Courtesy of Dr. Ewen A. Whittaker, Lunar and Planetary Laboratory, University of Arizona.)*

a puzzling lack of *maria* on the far side of the Moon, a fact which is connected with unanswered questions of origin. Some 5 to 10 percent of the Moon has not been observed by Luna III, by the later Soviet spacecraft Zond III, or from Earth, and remains to be discovered and named by future lunar spacecraft.

▽ The Moon's scarred and pitted face is a treasure trove of clues to the early history of the solar system. Because there is no erosion by wind or water, subsurface features have remained relatively undisturbed. What erosion there is, due to micrometeorites, secondary meteorite impacts, and the solar proton wind, operates very slowly, and may not disturb the lunar subsurface to great depths. There is consequently a possibility that as we dig beneath the lunar surface, we will find intact objects of earlier and earlier ages, until we reach material essentially as formed at the time of the origin of the solar system.

▽ In Chapters 16 and 17, we argued that the origin of life resulted from the production of organic molecules in a secondary reducing atmosphere, outgassed from the primitive Earth. A similar outgassed atmosphere must have accompanied the early Moon. The Moon today has essentially no atmosphere; its mass is so low that it is unable to gravitationally bind even the heaviest gases to it. Instead, they dissipate to space by the process of gravitational escape. In primitive times, atmospheric escape must have been equally efficient on the Moon. Thus, the Moon could have retained an atmosphere for, say, 10^9 years only if the gases escaping to space were continually replenished by outgassing from the lunar interior. There is some reason to suspect that such outgassing occurred, and that the Moon may have retained an atmosphere, and even a hydrosphere, in its early history. If so, organic molecules must have been abiologically produced in appreciable quantities. If life never developed on the Moon, such material may now be sequestered below the layer of micrometeoritic material which gently rained down through the primitive lunar atmosphere.

▽ We do not know how long the Moon retained an atmosphere and hydrosphere. It seems possible, although unlikely, that a primitive form of life arose in the early history of the Moon. It certainly could not survive on the lunar surface today. Any terrestrial organism placed on the surface of the Moon would be killed by conditions far more rigorous than those of Mars. During the lunar day (one of our months), the surface temperatures range from the normal boiling point of water to −180°C. There is no atmosphere, and no liquid water. The solar ultraviolet radiation alone is adequate to destroy in a period of hours the most radiation-resistant microorganism known. Along with solar x-rays and the solar proton wind, even dead organic matter would be charred in a few years.

▽ The possibility of life on the surface of the Moon has been proposed from time to time, on the basis of supposed color and other changes which have been reported. For example, the American astronomer William H. Pickering, observing in Mandeville, Jamaica, in the 1920's and 1930's, reported many observations of periodic color changes in the floors of craters, in step with the local time of day on the Moon. He reported, for example, that the floor of the crater Stevinus became reddish-brown as the Sun rose high in the lunar sky, while the floor of the large

FIGURE 21–8. *Rectified and recentered photograph of the region of the Moon observable from the surface of the Earth and also observed by the Soviet spacecraft Luna III. Figure 21–8 and the left-hand portion of Figure 21–7 show the same region of the Moon photographed under comparable lighting conditions. (Courtesy of Dr. Ewen A. Whittaker, Lunar and Planetary Laboratory, University of Arizona.)*

crater Grimaldi became increasingly green towards local noon. Pickering also reported seeing moving spots in the crater floors, some bright and some dark. The moving bright spots he attributed to clouds on the almost airless Moon; the moving dark spots, to the rapid growth of plants across the crater floor, or even, in some cases, to the movements of large migratory insects!

▽ Similar observations have been made up to the present day, although the interpretations have become less facile. Many of these reports can be attributed to the seeing conditions in the Earth's atmosphere, the dependence of the reflecting power of the lunar surface on the Sun's angle, and the motion of shadows across crater floors when the Sun is low in the lunar sky. There is no good evidence for life on the surface of the Moon; and indeed, the physical conditions there provide a strong independent argument against a lunar surface biology. But may not conditions be less inclement below the lunar surface?

▽ The Moon, like every other object in the solar system, is a source of radio radiation. The more intense the radio radiation, the hotter the Moon must be. At short radio wavelengths, radiotelescopes "see" close to the lunar surface and derive a temperature close to that determined by infrared and other techniques, figures which we have already quoted. But as we go to radiotelescopes tuned to longer and longer wavelengths, we are in effect observing the Moon at greater and greater depths. The Soviet astronomer V. I. Troitskii of Gorkii University finds a systematic increase of this derived "brightness temperature" with wavelength,

corresponding to an increase in temperature by about 1.6 C° per meter. The source of this temperature increase must be the heat of the lunar interior. The temperature beneath the Earth's surface similarly increases, as temperature measurements in borings and mines have shown.

▽ In addition to the average temperature increasing with depth, the daily temperature fluctuations are greatly reduced as we go below the lunar surface, because of the excellent insulating properties of the lunar surface material. The particles in the "fairy castle" structure have few points of contact with each other; therefore it is difficult for heat from the Sun to propagate downwards. At a depth of roughly 50 meters below the lunar surface, comfortable temperatures by our standards can be expected, temperatures which remain constant through the lunar day and night. At similar depths, several astronomers and geologists have suggested, there should be subsurface water which is prevented from escaping to the surface by an overlying layer of ice, in exact analogy to the permafrost of the terrestrial antarctic regions. If there is a region below the surface of the Moon which has warm temperatures, liquid water, and the possibility of primitive organic matter, then it seems premature to exclude the possibility of life on the Moon. Indigenous life seems highly unlikely because there is no energy source besides the chemical energy locked in the organic matter and other subsurface materials. This energy is limited at best; if life once arose at such depths, it would, in a brief period of time on the astronomical timescale, have died of malnutrition. But the lunar subsurface does seem able to support terrestrial microorganisms, a fact which is the basis for concern about biological contamination of the Moon.

▽ By the beginning of 1966, a dozen vehicles have impacted the Moon: the Soviet spacecraft Luna II, IV–IX, and the United States spacecraft Rangers IV, VI, VII, VIII, and IX. While attempts were made to sterilize Luna II and Ranger IV, there does not seem to have been a thorough sterilization of any of these vehicles. The difficulties in the sterilization procedure would have impaired the progress of the national space programs involved; as a result, the sterilization requirements were relaxed. Fortunately, the risk is not nearly as grave as that which would attend the biological contamination of Mars. We have no independent evidence for life on the Moon; but there is, as we saw in Chapter 20, some suggestion of life on Mars. If terrestrial microorganisms are to replicate on the Moon, they must find their way down some tens of meters below the lunar surface. In addition, while there is a mechanism for widespread atmospheric distribution of replicating contaminants on Mars, no comparable mechanism exists on the Moon.

▽ There may, however, be direct evidence for subsurface lunar organic matter. From time to time, there have been reports of gas clouds, mists, and reddish glows observed on the lunar surface. The earliest such report appears to be that of Sir William Herschel, the discoverer of the planet Uranus. His accounts follow:

> May 4, 1783. I perceived in the dark part of the Moon a luminous spot. It had the appearance of a red star of about the fourth magnitude. It was situated in the place of Hevelli Mons Porphyrites [a feature which we call today the crater Aristarchus].

> The instrument with which I saw it was a 10 feet Newtonian Reflector of 9 inches aperture. Dr Lind's lady who looked in the telescope immediately saw it, tho' no person had mentioned it, and compared it to a star . . .
>
> ***
>
> Last night I had an opportunity to view the Moon in a favorable situation and found that the volcano of which I saw the eruption last month was still considerably luminous. The crater seemed to glow with a degree of brightness, which I should not have been able to account for, if I had not seen the eruption last month. It appeared to me as if the crater was nearly doubled in its dimension since last month . . .
>
> ***
>
> This is all I can give you at present. Believe me, Sir, I have not the least desire of keeping such observation to myself, but have so many subjects (which I think of greater consequence to astronomy) in hand at present that I had postponed giving an account of them to some other opportunity.

Quite analogous glows were seen in Aristarchus in 1963 by the American astronomers, Edward Barr and James Greenacre at the Lowell Observatory, Flagstaff, Arizona. Their observations were confirmed by several other observers at another telescope.

▽ The lunar crater Alphonsus had already been the subject of reports of low-lying gas clouds, or ground hazes, which obscured surface detail, when the Soviet astronomer N. A. Kozyrev performed a remarkable observation of the Moon on 3 November, 1958, at the Crimean Astrophysical Observatory. While photographing the spectrum of the sunlight reflected from Alphonsus, Kozyrev noticed a reddish cloud enveloping the central peak of the crater. (A Ranger IX photograph of Alphonsus is seen in Figure 21–9. It is the largest crater and is in the left-hand portion of the photograph. The featureless central mountain peak may also be viewed. A close-up of the crater floor of Alphonsus is seen in Figure 21–4.) He immediately obtained another spectrum, and after about thirty minutes, observed the cloud to have dissipated. He then obtained a final spectrum. Kozyrev's first and third spectra showed the usual spectrum of the Moon—namely, no lunar features at all; merely the solar spectrum superimposed on the absorption features of the Earth's atmosphere (since we are looking at sunlight reflected off the Moon and passing through the atmosphere of the Earth). The second spectrum, however, showed a broad spectral feature confined to the region of the central peak of Alphonsus that was absent on the other two spectra. Kozyrev identified this feature as due to the molecule C_2. This identification has stood the test of time and a number of critical analyses of the observation.

▽ C_2 is not a molecule encountered in everyday life on the Earth, because it is highly reactive, and combines, for example, with oxygen to form CO_2. C_2 is, however, a constituent of comet tails, where it apparently is the breakdown product of a more complex organic molecule. Similarly, the presence of C_2 on the Moon must be attributed to some larger molecule which contains two carbon atoms. The simplest such molecule is acetylene, C_2H_2, although more complex organic

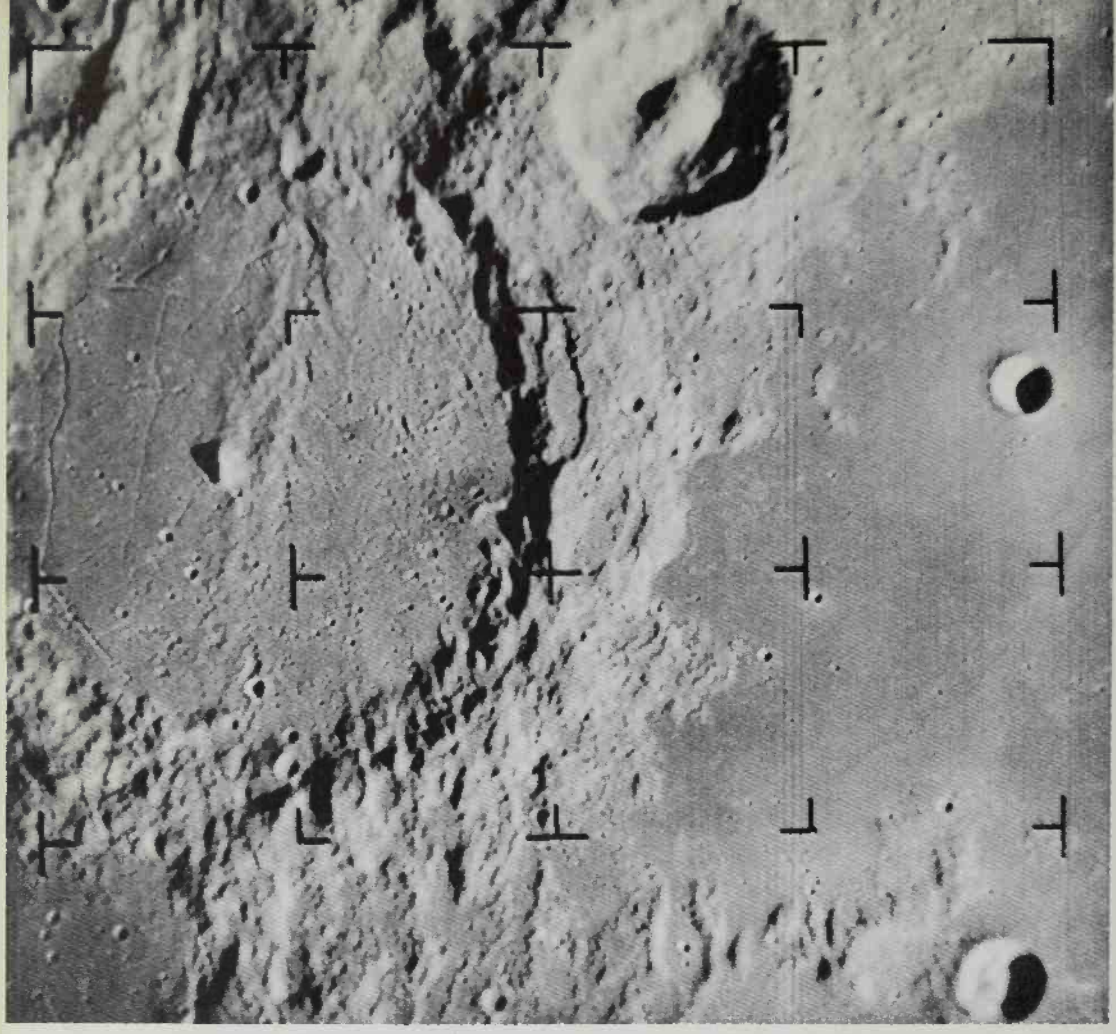


FIGURE 21-9. *Ranger IX photograph of the lunar crater, Alphonsus, at left. The central peak in this crater is the apparent source of the gas emission seen by Kozyrev and others. The peak shows no central opening. (Courtesy of NASA.)*

molecules are also possible. In the reconstruction of events now accepted, acetylene or a more complex molecule was outgassed from the lunar interior, escaping in the vicinity of the central peak of Alphonsus. The molecule was then bombarded by solar radiation, which broke it down into simpler fragments, including C_2. The C_2 molecule was then excited by sunlight; the absorption and emission of light by C_2 account for Kozyrev's spectrum.

▽ Although Kozyrev, like Herschel, attributed what he saw to lunar volcanism, there seems little evidence for active volcanism of the terrestrial type on the Moon. But as Kozyrev's observations showed, outgassing from the lunar interior may be a contemporary process. Also in the crater Alphonsus, there are craters which are surrounded by dark halos. A general discoloration can be seen in the region of the large crater at the bottom center of Figure 21–4, in the floor of

Alphonsus. It has been suggested that such dark halo craters are also the result of the outgassing of material from the lunar interior. One possibility for such a material is acetylene. Kozyrev's observations, the halo craters, and the more frequent reports of glows and gas clouds do suggest the presence of organic matter beneath the lunar surface and the occasional penetration of the surface by escaping organic materials.

▽ The United States and the Soviet Union have extensive programs for the scientific exploration of the Moon, first by unmanned orbiters and landing vehicles, and followed by scientific exploration parties. Such explorations have the promise not only of determining the physical conditions at the time of the origin of the solar system, but also, through organic chemical analysis of subsurface material, of illuminating the processes which led to the origin of life on Earth. △

22

Mercury and Venus: environments and biology

. . . I have often wonder'd that when I have view'd Venus . . . she always appeared to me all over equally lucid, that I can't say I observed so much as one Spot in her . . . is not all that Light we see reflected from an Atmosphere surrounding Venus?

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

I. Mercury

▽ Until recently, the planet Mercury was described as both the hottest and the coldest place in the solar system. Because it is the planet closest to the Sun and because it absorbs almost all the sunlight reaching it, its illuminated hemisphere should be very hot. But since the planet was thought to rotate synchronously, always keeping the same face towards the Sun, it seemed that the dark side would be heated primarily by the heat flow from within and by starlight—energy sources so feeble that the temperatures on the dark side were estimated at 20 or 30 degrees above absolute zero (about −240°C).

▽ Measurement of the infrared and radio emission of the bright side confirmed the theoretical expectation of high temperatures; values around 350°C were obtained. Recently, the first accurate measurements of the temperature on the dark side of Mercury have been made by the American astronomer Kenneth Kellerman at the Parkes radiotelescope near Sidney, Australia. Kellerman found that temperatures on the dark side were in the vicinity of 0°C, the normal melting point of ice. Thus, since the bright side of Venus is hotter, and the clouds of the Jovian planets colder, Mercury is neither the hottest nor the coldest place in the solar system.

▽ How is the temperature on the sunless side maintained? One possibility is that Mercury is not in synchronous rotation. The observations which led to deductions of synchronous rotation, by Schiaparelli, Antoniadi, and Dollfus, are very difficult to perform. Figure 22–1 shows three drawings of the Mercurian surface made by Dollfus at the Pic du Midi Observatory in the French Pyrenees, among the finest observing locales on Earth. Each drawing is based on observations made on a different night. Except for the wobble in the position of the axis of rotation of Mercury—an effect known as libration—they show approximately the same regions. While the general configuration of the markings is very similar in each of the three drawings, the differences from drawing to drawing illustrate the difficulties in observing so small an object as close to the sun as Mercury. The remarkable feature of the maps of all observers of Mercury is that they only show one hemisphere, while if the planet were in nonsynchronous rotation, it should be possible to see, at various times, aspects of both hemispheres.

▽ More recently radar techniques have been used to observe the rotation rate of Mercury. This technique for determining the rate of rotation is exactly analogous to the Doppler effect methods discussed in Chapter 13 for determining stellar rotational velocities. When a radar pulse is reflected from a planetary surface, the edge approaching the Earth changes the frequency of the reflected radar

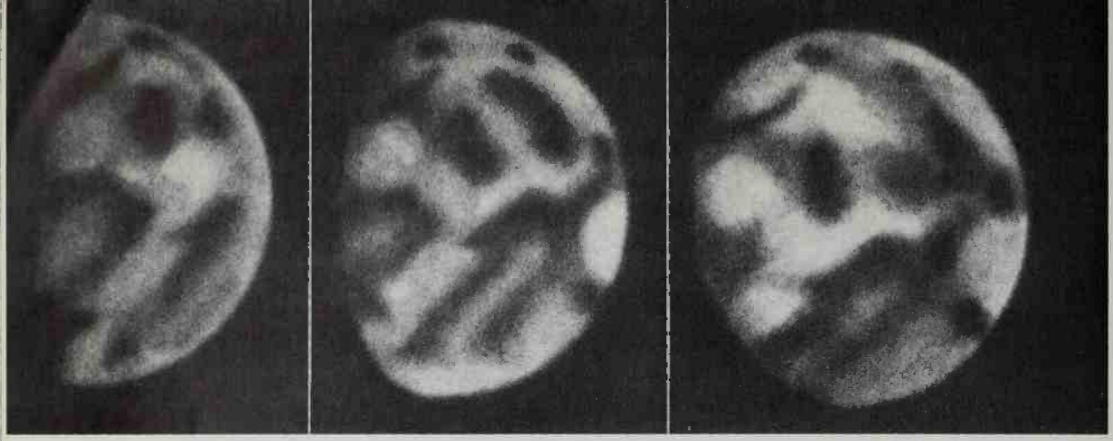


FIGURE 22–1. *Three drawings of the illuminated hemisphere of the planet Mercury made by Dr. Audouin Dollfus in October 1950. Each drawing represents the collation of several observations. Mercury evidently exhibits phases like the Moon. (Courtesy of Dr. Audouin Dollfus.)*

beam to smaller wavelengths, while the edge receding from the Earth Doppler shifts to longer wavelengths. A radar pulse which has a very small range of frequencies will, because of planetary rotation, have a much broader range of frequencies when received back on Earth after reflection. Using the world's largest radiotelescope, a dish of 1,000 ft diameter in Arecibo, Puerto Rico, the American radio astronomers Gordon H. Pettengill and Rolf Dyce of Cornell University have shown that Mercury is apparently not rotating synchronously. Rather than rotating once every eighty-eight days—a period of rotation equal to its period of revolution—these radar results suggest that Mercury is rotating once about every fifty-six. If this is the case, then the problem of maintaining the relatively high temperatures on the dark side of Mercury is solved. Even the coldest place on Mercury would have been in sunlight a few weeks previously, and there would not be sufficient time for it to cool to lower temperatures than are observed. Nevertheless, the conflict between the visual and the radar observations, both of which are difficult to perform, remains nettling, and at the present time we cannot say that the problem is solved.

▽ Polarimetric observations by Audouin Dollfus, in France, and spectroscopic observations by V. I. Moroz, in the Soviet Union, have each indicated that Mercury has a definite, although very thin, atmosphere. The illuminated hemisphere of Mercury has permanent features; these have been drawn in Figure 22–1. Antoniadi, in the 1920's, reported seeing atmospheric "veils" which temporarily obscured the dark features, an apparent analogy to the situation on Mars. The veils also suggest the presence of some atmosphere.

▽ That Mercury has any atmosphere at all is extraordinary. Because of its nearness to the Sun, its exosphere temperature should be very high. The low gravitational field of the planet and the high exosphere temperature together suggest that any but the most massive molecules would have escaped during the history of the solar system. The gas identified by Moroz is carbon dioxide, which is in fact a fairly heavy gas. But even CO_2 would have escaped during the lifetime of Mercury. The presence of an atmosphere on Mercury is probably due to an equilibrium between outgassing and escape. CO_2 and other gases are exhaled from

the Mercurian interior, spend a fairly brief period of time in the atmosphere proper, and escape from the exosphere. The gas which we observe at any time is the gas which happens to be in transit between the Mercurian interior and interplanetary space.

▽ If Mercury were in synchronous rotation, its dark side might be heated by hot gases which circulate from the illuminated hemisphere to the dark side, carrying their heat with them. The wind speeds required are enormous—hundreds of miles per hour. If we imagine ourselves standing near twilight on Mercury, the Sun will appear two and a half times its size as seen from Earth, and will be low in the black Mercurian sky. The landscape before us is even more parched, withered, and seared than on our own airless Moon. There is a thin but violent wind blowing towards us. The temperatures on the bright side of Mercury are hotter than the highest temperatures in the average oven; it is difficult for us to imagine any life thriving under that cruel and blazing sky.

▽ Behind us is the dark side. We think the temperatures there are equable, but we know nothing else about it. An extensive atmosphere is probably absent on the dark side, because it would recirculate to the bright side and boil off into interplanetary space. Liquid water may be present there temporarily, and we may therefore begin thinking about the possibility of life on the night side of Mercury. In the absence of sunlight, we cannot expect photosynthesizing plants. There are other energy sources, but because of the paucity of our knowledge, it does not seem profitable to speculate upon them. Yet in any inventory of biologically interesting planets Mercury must be included. Hopes have been expressed by the United States' National Aeronautics and Space Administration that unmanned exploration of the Mercurian surface may begin in the late 1970's.

II. Venus

▽ Seen through a large telescope, Venus is an even more disappointing sight than Mars. When the planet is full, we see a completely featureless disk. In the course of months, Venus exhibits crescent phases like those of the Moon, since it, too, passes between us and the Sun, and we are often presented with some combination of the bright and dark hemispheres. The dark hemisphere is invisible against the blackness of space beyond, and the crescent shape of the illuminated hemisphere is all we see [Figure 22–2]. When the planet is photographed in ultraviolet light, faint, evanescent markings can be discerned [Figure 22–3]. Venus is surrounded by an extensive, unbroken cloud deck whose composition was, until very recently, unknown. Both in the visible and in the ultraviolet, we are seeing only the clouds of Venus.

▽ Before we can proceed with the discussion of Venus, there is a semantic problem which must first be solved. Except for the Earth, the names of the planets are derived from the gods of Roman mythology. Corresponding to each god, there

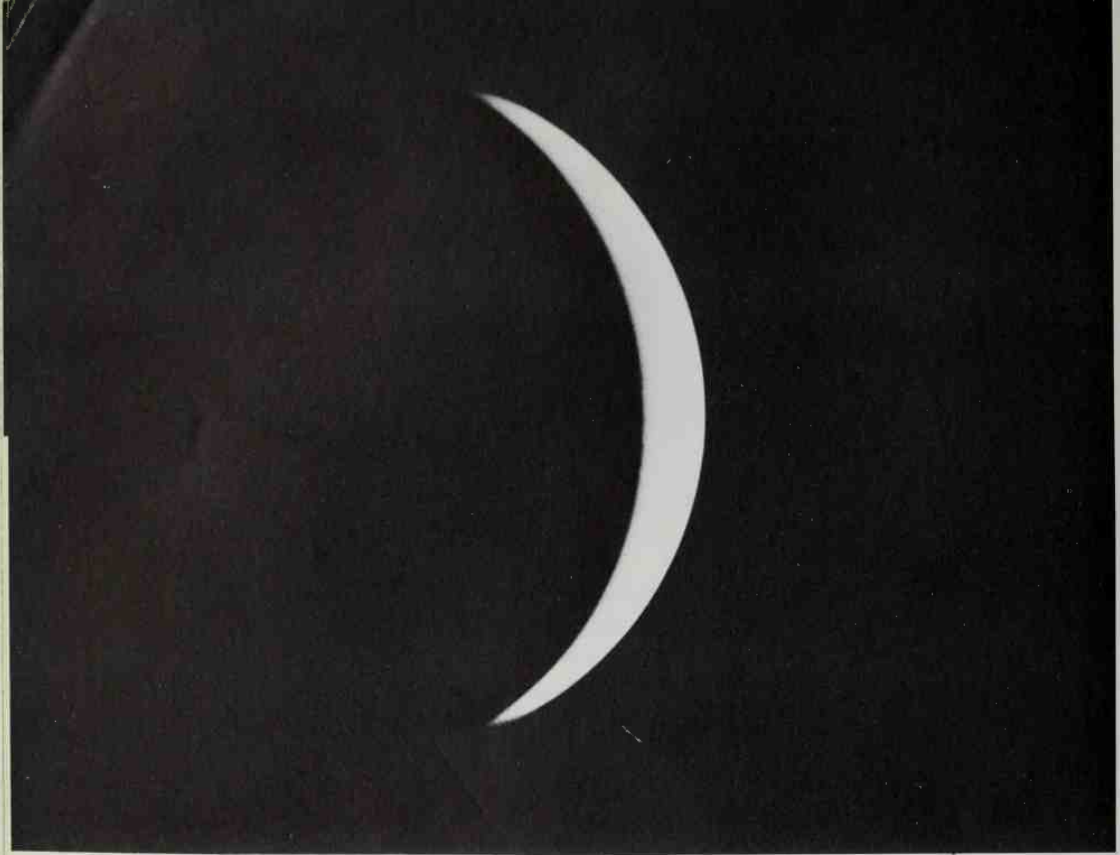


FIGURE 22–2. *Venus, photographed in blue light in crescent phase. At all visible frequencies photographs of Venus such as this show no discernible markings on the disk. (Courtesy of Mt. Wilson and Palomar Observatories.)*

is a certain character, personality trait, or geological term which is commemorated in adjectival form: Mercury, mercurial; Mars, martial; Jupiter (or Jove), jovial; Saturn, saturnine; Neptune, neptunian; and Pluto, plutonic. In the case of Venus, the corresponding adjective is "venereal." Long after most of these words were ensconced in the English language, astronomers discovered the need for planetary adjectival forms which would not be confused with similar adjectives of different meaning. For some planets there was no problem. Thus, Mercury, Mercurian; Mars, Martian; Jupiter, Jovian; Saturn, Saturnian; and Uranus, Uranian. Since almost no astronomical work has been done on the physical environments of Neptune and Pluto, there is little confusion in the use of Neptunian and Plutonic.

▽ But what to do for Venus? The proper word, by analogy, is "Venerean," but many astronomers felt it to be too closely associated with its cognate, and preempted by other areas of human activity. The Italian-American astronomer Luigi Jacchia, of the Smithsonian Astrophysical Observatory, has suggested that "Venerean" be used, nevertheless, and *honi soit qui mal y pense:* but "Venerean" has not proved at all popular in the scientific literature. One sometimes finds "Venusian" as an alternative; but this is a barbarism, comparable to "Marsian,"

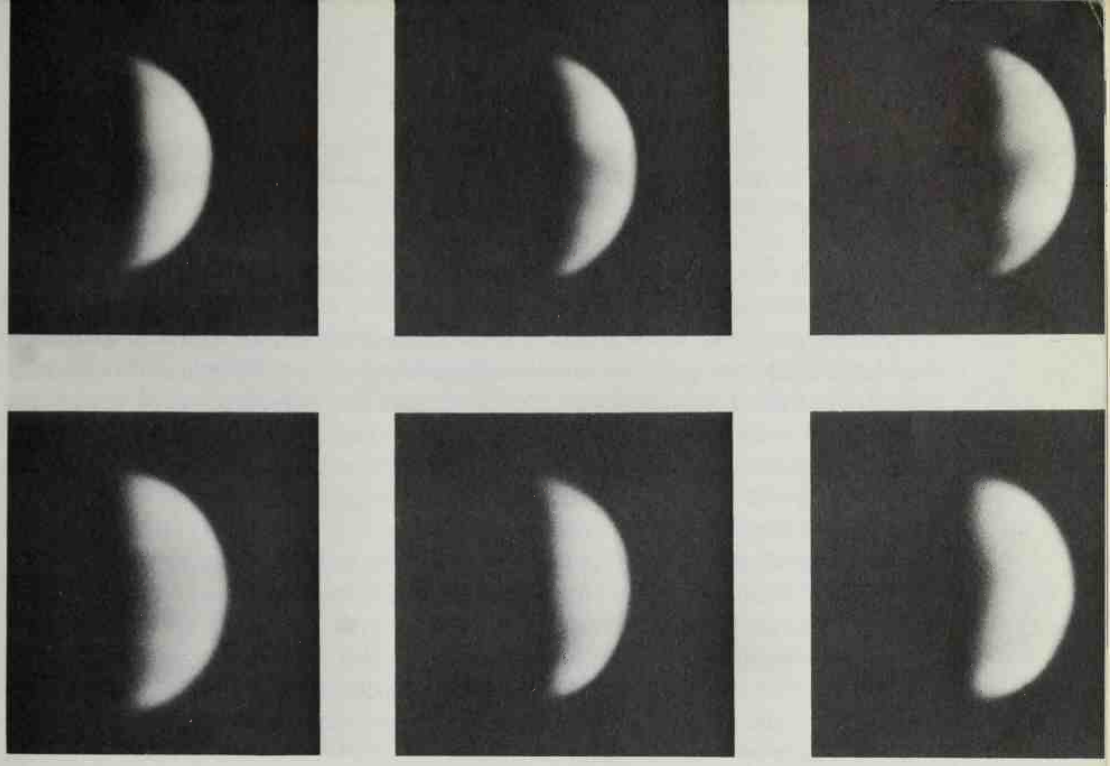


FIGURE 22–3. *Six photographs of Venus taken in ultraviolet light. At these frequencies, markings can sometimes be discerned, especially near the terminator—the line dividing the bright and dark hemispheres. The ultraviolet features vary from day to day—perhaps from hour to hour—as these photographs show. In ultraviolet, as in visible light, we are looking at the Venus cloud level, and the variations in ultraviolet indicate, at most, violent movements of the Venus clouds. (Courtesy of Mt. Wilson and Palomar Observatories.)*

"Jupiterian," or "Earthian." The Greek goddess corresponding to Venus is Aphrodite. The appropriate adjective here, "Aphrodisian," or "Aphrodisial," again has other connotations, which some astronomers, in the interests of clarity and decorousness, prefer to avoid. The currently accepted alternative is "Cytherean," from the Ionian island of Cythera, onto which Aphrodite is said to have emerged from the sea. It is portrayed in the right foreground of Botticelli's celebrated painting, "Birth of Venus."

▽ Having clarified these monumental issues, let us consider further the Cytherean environment. Because Venus is enshrouded by clouds, direct telescopic examination of its surface was beyond the ability of the early planetary observers. In the absence of direct observations, they adduced a variety of differing and mutually inconsistent environments. Since only water clouds were familiar, the apparent thickness of the Cytherean cloud layer seemed to argue for a great abundance of water. From there, it was only a step to the assertion, seriously put forth in 1918 by Svante Arrhenius, that

> everything on Venus is dripping wet . . . a very great part of the surface of Venus is no doubt covered with swamps. . . . The constantly uniform climatic conditions which exist everywhere result in an entire absence of adaptation to changing exterior

> conditions. Only low forms of life are therefore represented, mostly no doubt, belonging to the vegetable kingdom; and the organisms are nearly of the same kind all over the planet.

Arrhenius, it will be remembered, had criticized Lowell for deducing too much about Mars from too little data.

▽ Spectroscopic observations of such a wet world should easily demonstrate, one would think, the presence of atmospheric water vapor. Thus, it was with some surprise that observers in the 1920's found that they were unable to detect any water vapor above the clouds of Venus at all. Thus, the Carboniferous swamp model was generally abandoned, and replaced by the arid, planetary desert model. The clouds could not then be water; they were instead attributed to a permanent pall of dust, raised from the windswept surface.

▽ Unsatisfied with such an explanation of the brilliant white clouds of Venus as dust, the American astronomers Donald H. Menzel and Fred L. Whipple, of Harvard University, pointed out in 1955 that the absence of spectroscopically detectable water vapor was not a good argument against water clouds. The situation can be demonstrated by the simple analogy of a pan of water whose temperature can be controlled. At a given moment, some fast-moving H_2O molecules have broken the weak chemical bonds which bind them to their neighbors and are escaping from the pan. At the same moment, some other H_2O molecules are reentering the pan from the overlying atmosphere. Just as in the atmosphere of Mercury, the amount of water vapor above the pan depends on the equilibrium between two processes. As we reduce the temperature of the pan, there are far fewer fast-moving molecules in the liquid and therefore far fewer water vapor molecules in the atmosphere above. If the temperature of the water is sufficiently low—say, many tens of degrees below 0°C, so that the water has frozen to ice—then the amount of water vapor above the pan will be very small indeed.

▽ From the infrared emission of Venus, it was determined that the temperature of the clouds of Venus is about −40°C (by coincidence, this is also −40°F). If the clouds of Venus were made of ice crystals at a temperature of −40°C, the amount of water vapor above them would be undetectable, and no contradiction with the spectroscopic results would be implied. Menzel and Whipple then went on to argue that if large amounts of water existed in the clouds, even larger amounts must exist on the surface. In the previous unsuccessful searches for water vapor, it had been found, quite by accident, that great quantities of carbon dioxide existed in the atmosphere of Venus. Menzel and Whipple proposed, in effect, that the surface of Venus was largely covered by carbonated oceans—seltzer water.

▽ As a final example of the variety of descriptions of Venus which could be derived from the very limited data then available, let us consider the model proposed, also in 1955, by Fred Hoyle. In the early history of any planet, there will be a certain amount of water and other materials outgassed from the planetary interior, as we saw in Chapter 16. In the upper atmosphere of the planet, the water vapor tends to be photodissociated by solar ultraviolet radiation; the hydrogen

escapes to space, and the oxygen remains behind to oxidize the atmosphere [see Chapter 16]. If the planet initially has much more water than hydrocarbons, all the hydrocarbons will eventually be oxidized, and an aqueous, oxidizing environment will result as on Earth. But if the initial complement of hydrocarbons greatly exceeds the amount of water, all the water will be used up in partially oxidizing the hydrocarbons to CO_2, and a CO_2 atmosphere with a large residue of surface hydrocarbons will result. While the atmosphere of Venus is thought to be largely composed of N_2, by the same argument from default that we encountered for Mars (Chapter 19), the proportion of CO_2 is perhaps a hundred times greater than in the Earth's atmosphere. Hoyle therefore proposed that the surface of Venus was covered with oil, or other hydrocarbons, and that the cloud layer was smog.

▽ The state of our knowledge of Venus in 1956 is amply illustrated by the fact that the Carboniferous swamp, the windswept desert, the planetary oilfield, and the global seltzer ocean each had its serious proponents. Those optimists planning, in 1956, eventual manned missions to Venus must have had considerable difficulties in deciding whether to send along a paleobotanist, a mineralogist, a petroleum geologist, or a deep-sea diver. We now know that none of these models is correct, and that a proper description of Venus incorporates features from several of the early models.

▽ In 1956, a team of American radioastronomers at the U.S. Naval Research Laboratory, headed by Cornell H. Mayer, first turned a large radiotelescope towards Venus. The observations were made near inferior conjunction, the time when Venus is nearest the Earth, and when, also, we are looking almost exclusively at the dark hemisphere of the planet. Mayer and his colleagues were astounded to find that Venus radiated as if it were a hot object at a temperature of about 300°C. Subsequent observations at a variety of wavelengths have confirmed these observations and have shown that the deduced temperature of Venus increases away from inferior conjunction—that is, as we see more and more of the illuminated hemisphere. The most natural explanation of these observations is that the surface of Venus is hot—far hotter than anyone had previously imagined. Venus is about 0.7 A.U. from the Sun. By the inverse square law it should therefore receive $1/(0.7)^2$, or about twice as much solar energy as does the Earth. On the other hand, its clouds are very highly reflecting. When both effects are considered, it turns out that despite its smaller distance from the Sun, Venus absorbs less sunlight than the Earth. Ordinarily, it should not even be as hot as the Earth; yet it was 300° warmer.

▽ Some early difficulties in providing a detailed explanation of the high surface temperatures led to an alternative explanation of the intense radio radiation from Venus. Douglas E. Jones, an American physicist at the Jet Propulsion Laboratory of the National Aeronautics and Space Administration, proposed that the high temperatures apply not to the surface of Venus, but to a dense ionized layer, or ionosphere, high in the Cytherean atmosphere. The difference between the hot surface and hot ionosphere models can be understood by reference to Figures 22–5 and 22–6. The radio spectrum of Venus at inferior conjunction is

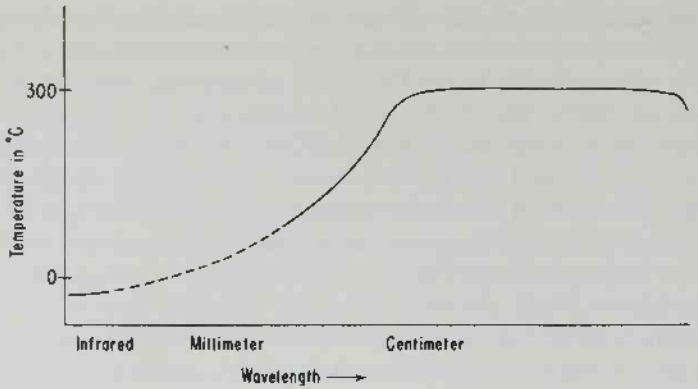


FIGURE 22–4. *Schematic representation of the long wavelength spectrum of Venus. The temperature observed rises from about −40° C, at infrared wavelengths, to over 300° C at centimeter wavelengths. Any model of the Venus environment must be able to explain this spectrum.*

given roughly by Figure 22–4. At centimeter wavelengths, the same temperature—about 300°C—is deduced at all wavelengths. But at millimeter wavelengths, there is a sharp decline in temperature, as we would expect, since the spectrum should join smoothly to the temperature of −40°C deduced in the infrared.

▽ In the hot surface model [Figure 22–5], the centimeter wavelength radiation arises from the surface and is transmitted by the atmosphere and clouds, which must be transparent at those wavelengths. At millimeter wavelengths, however, the atmosphere and clouds must absorb the radiation, so that in fact the shorter wavelength radiation arises from higher, and therefore cooler, levels of the atmosphere. In the infrared, we are observing the cold clouds.

▽ In the hot ionosphere model, however, at centimeter wavelengths we are observing emission by the ionosphere. A dense ionosphere becomes transparent towards shorter wavelengths. In the ionospheric model, at millimeter wavelengths

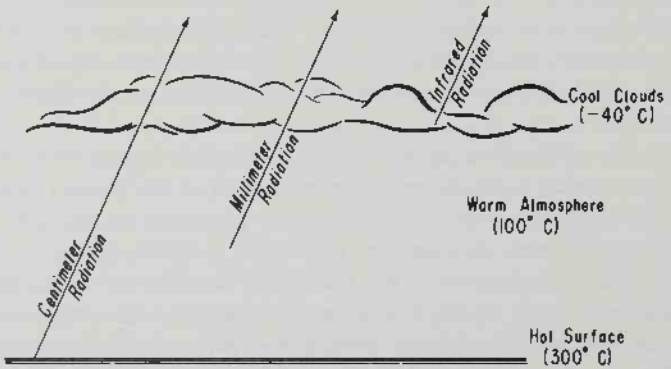


FIGURE 22–5. *Schematic representation of the hot surface model of Venus.*

we are seeing radiation emitted directly by the surface [Figure 22–6]. Note that the ionospheric model preserves the possibility of relatively low surface temperatures and therefore of the possibility of life on Venus. This is its great appeal.

▽ A distinction between the hot ionosphere and hot surface models can be gained if we imagine a radiotelescope scanning across the disk of Venus, tuned to a wavelength of about one centimeter. In the hot surface model, the atmosphere and clouds are slightly absorbing at 1 cm wavelength. Thus, when the radiotelescope looks towards the edge of the disk, there is more absorbing material in the light path than when the radiotelescope points to the center of the disk [Figure 22–7]. Thus, in the hot surface model, there should be less radiation coming from the edges, or limbs, of Venus than from the center, a circumstance known as limb-darkening.

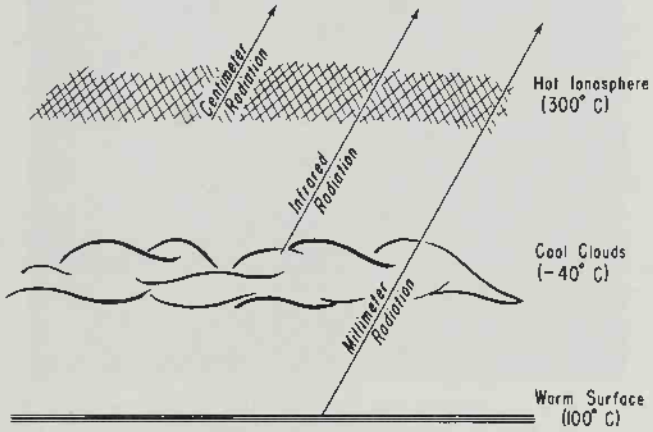


FIGURE 22–6. *Schematic representation of the hot dense ionosphere model of Venus.*

▽ In contrast, consider the hot ionosphere model [Fig. 22–8]. Here, the semitransparent ionosphere is the primary source of emission at 1 cm wavelength. At the center of the disk, the radiotelescope sees a smaller thickness of the emitting ionosphere than at the limbs. Where there is more emitting material, there should be more emission. Thus, the hot ionosphere model predicts limb-brightening. Unfortunately, the available radiotelescopes on Earth are unable to resolve, or scan across, Venus. At 1 cm wavelength, they could only determine the average emission over the entire disk. A relatively small radiotelescope, flown to the vicinity of Venus, could distinguish between limb-brightening and limb-darkening by scanning the Cytherean disk; this was a primary mission of the United States spacecraft Mariner II.

▽ A photograph of Mariner II may be seen in Figure 22–9. The extended horizontal panels are solar cells for the conversion of sunlight into electricity. At the very bottom is a directional antenna for radioing scientific results back to Earth. The radiotelescope used to scan the disk of Venus is the small disk sitting just above the main hexagonal electronics housing.

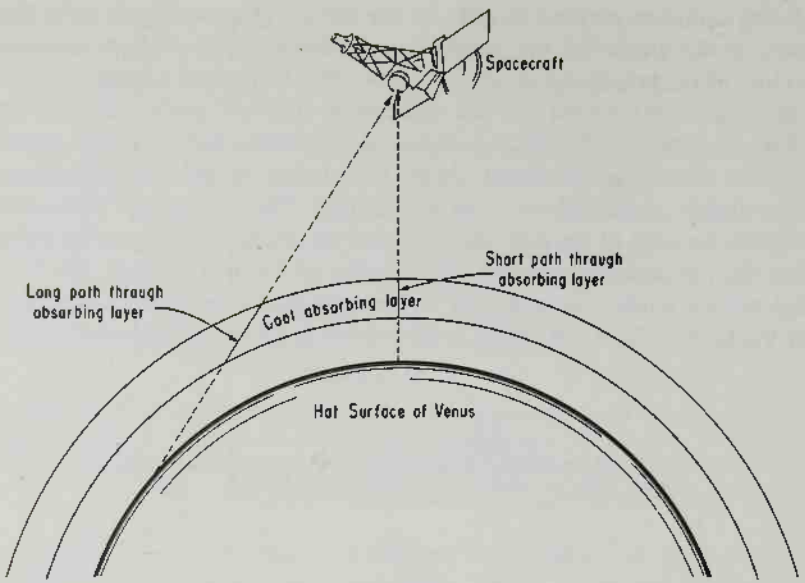


FIGURE 22–7. *Illustration of limb-darkening expected at 1 centimeter wavelength if Venus has a hot surface and a cool absorbing layer. When the spacecraft looks towards the limb of the planet, it is looking through a greater quantity of absorbing material and therefore sees a lower effective temperature.*

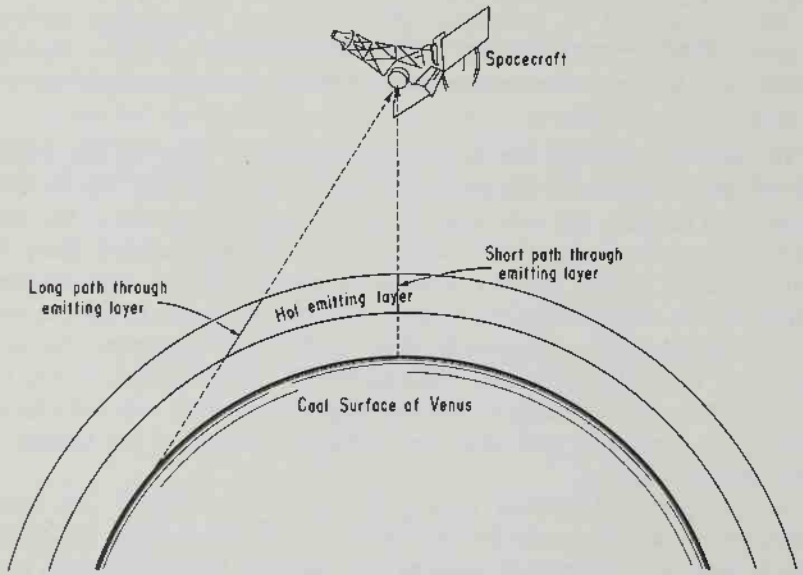


FIGURE 22–8. *Representation of expected limb-brightening on the ionospheric model of Venus. Here, when the spacecraft looks towards the limb, it sees a greater path of emitting material and therefore enhanced emission.*

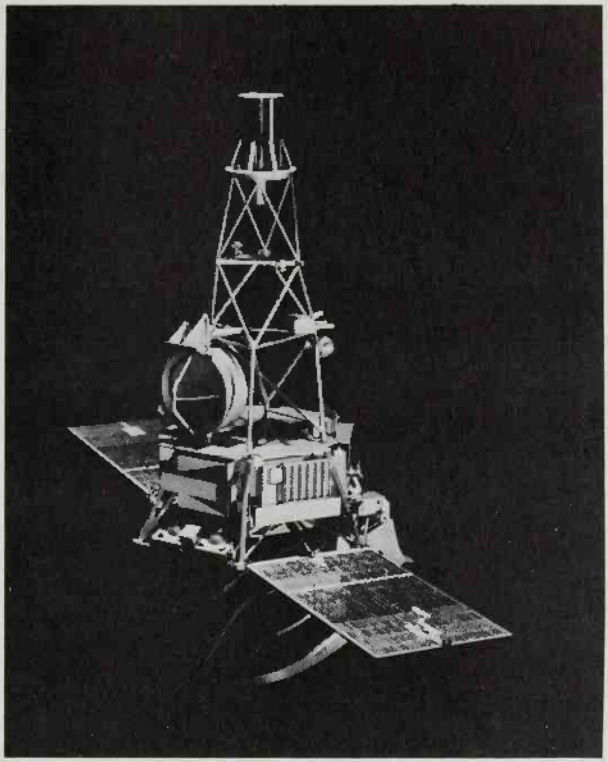


FIGURE 22–9. *A photograph of the Mariner II space vehicle as it may have looked when flying by Venus on 14 December, 1962. The microwave and infrared radiometers are mounted on the disk set in the wire superstructure just above the middle of the spacecraft. (Courtesy of NASA.)*

▽ On 14 December, 1962, Mariner II passed within 35,000 to 40,000 km of Venus, and scanned across the disk at two wavelengths near 1 cm. Mariner II found no limb-brightening. Instead, a distinct limb-darkening was observed. These results contradict the ionospheric model and provide support for the hot surface model.

▽ This Mariner II experiment is an excellent example of the role of space vehicles in the investigation of planetary environments. A specific model of Venus had been proposed which explained most of the observations then available. The model had predictable consequences, which were different from the consequences of other models, but which could not be tested from the vicinity of the Earth. A space vehicle was needed. The spacecraft and the radiotelescope were designed and built together. Despite the fact that some expectations for the mission were not fulfilled, both the spacecraft and the radiotelescope worked well enough to provide the critical tests of the theoretical model.

▽ If, then, the ionospheric model is invalid, what makes Venus hot? From a

variety of observations at visual, infrared, and radio frequencies, it has recently been established that the clouds of Venus are indeed made of water: ice crystals in the colder cloudtops, which are seen in ordinary photographs [Figures 22–2 and 22–3]; and water droplets in the bottom of the clouds, which are "seen" at long wavelengths. The CO_2 and H_2O in the Cytherean atmosphere, plus the water in the clouds, combine to produce a very efficient greenhouse effect. The atmosphere is in convective motion. Sunlight is deposited either in the clouds or directly on the surface. The sunlight which is deposited on the surface heats it immediately; the sunlight which is deposited in the clouds or atmosphere is transported by the downward convective motions, to heat the surface. The hot surface attempts to radiate in the infrared, but the absorption by the atmospheric CO_2 and H_2O and the water clouds is so great that very little heat from the surface or lower atmosphere escapes directly to space. The surface temperature must then be sufficiently high so that the small fraction of radiation which does escape to space equals the intensity of the sunlight which is absorbed by Venus.

▽ The American astronomer James B. Pollack of the Smithsonian Astrophysical Observatory and I have explored the role which water clouds can play in determining the characteristic features of the Venus environment. We find that a fairly thick layer of ice crystal clouds with water droplets below can explain in detail the spectrum of infrared and microwave radiation emitted from the planet, the limb-darkening at microwave frequencies observed by Mariner II, the variation of the centimeter wavelength temperatures with the phase of Venus, the limb-darkening observed in the infrared, and the polarization properties of the Venus clouds at optical frequencies. In addition, these clouds can explain, through the greenhouse effect, the high surface temperatures deduced from radio observations. While there are still a number of unsolved problems about Venus, the hypothesis that the clouds are water explains, in a straightforward way, a wide variety of observations.

▽ When a radar pulse is sent to Venus at centimeter wavelengths, it is transmitted by the atmosphere and clouds and strikes the surface, where it is partially absorbed and partially reflected. The part which is reflected is then returned to Earth, where it can be detected with a large radiotelescope. The ability of Venus to reflect radar gives a clue to its surface composition, just as the brightness and color of an object in reflected visible light can be used for estimating its composition. For example, extensive oceans of water or hydrocarbons can be excluded. In addition, the rotation of Venus causes a Doppler broadening of monochromatic radar pulses reflected from the planet, and the rate of rotation of Venus can be deduced.

▽ When the passive radio observations and these active radar observations are combined, some interesting conclusions about the body of Venus emerge. Venus is rotating very slowly, approximately once every 250 days; but more remarkable yet, it is rotating backwards. Except for Uranus, which is a marginal case, all the other planets in the solar system have direct rotation; that is, they are rotating in the same direction that they are revolving about the Sun. If we stand above the Earth's

north pole and observe the Earth rotating from West to East beneath us, we will find that it is rotating in a counterclockwise direction. From the same vantage point, we would see the Earth revolve about the Sun, also counterclockwise. This is called direct rotation. But if we were able to make the same observation at Venus, we would find that while it revolves about the Sun in a counterclockwise sense, it rotates clockwise about its axis. This is called retrograde rotation. The cause of the retrograde rotation of Venus is unknown, but it and the slow rotation period are both probably related to tidal friction. The rotation and revolution of Venus together imply that the time from local sunrise to sunrise—the "day" on Venus—is about 116 of our days. The nights on Venus are long and hot.

▽ The surface of Venus is not covered with liquid water or pools of hydrocarbons. But any one of a large number of pulverized common terrestrial minerals could account for the properties of the Cytherean surface as determined by radio measurements. The coldest temperature on Venus is about 200°C; the warmest, about 700°C. At these temperatures, any familiar terrestrial organisms would be scorched. It is perhaps premature to exclude the possibility of completely novel organisms, based on exotic chemistry, but, from our present vantage point, the prospects for life on the surface of Venus appear very bleak indeed.

▽ The Cytherean water clouds are, perhaps, another story. They are at moderate temperatures, bathed in sunlight, abundantly supplied with water, and must contain small amounts of minerals convectively transported from the underlying surface. It is possible to imagine organisms carrying out their entire life cycle in such an environment. The clouds of Venus appear to be a possible habitat for microorganisms from Earth, if not indigenous Cytherean organisms. We will discuss this possibility further in Chapter 34.

▽ As for the surface of Venus, it is appallingly hot; because of the thick clouds, it is overcast and gloomy even in the daytime. The temperatures are so high that in some places the surface should glow with the deep ruby red of its own heat. Venus, the bright morning star, has for millennia been called and identified with Lucifer. The identification is curiously appropriate. Venus is very much like hell. △

23

The solar system beyond Mars: environments and biology

The inhabitants of Jupiter must . . . it would seem, be cartilaginous and glutinous masses. If life be there, it does not seem in any way likely that the living things can be anything higher in the scale of being than such boneless, watery, pulpy creatures . . .

William Whewell, 1854

I. Jupiter

▽ Moving outwards from the Sun, we glimpse the familiar Earth and ruddy Mars, which we have already discussed. If we pass those mountains of erratically drifting rubble and debris, the asteroids, we arrive at mighty Jupiter, eleven times larger than the Earth, three hundred times more massive, where the day is ten of our hours and the year is twelve of our years. The nearest and best-known of the Jovian planets, Jupiter is still far less understood than Venus or Mars.

▽ When we look at Jupiter, we see a whirling, turbulent mass of clouds and gases. The atmosphere of Jupiter is composed primarily of hydrogen and helium, with smaller amounts of ammonia, methane, and probably water. The clouds of Jupiter [Figure 23–1] are thought to be composed of frozen crystals of ammonia, but this is not certain. The temperature at the clouds is about $-100°C$. In this environment of unfamiliar substances and low temperatures, spots are observed suddenly to appear in the Jovian clouds. Due to the differential rotation of Jupiter (it rotates faster at the equator than near the poles), the spots are stretched out into the conspicuous, brightly colored bands which are one hallmark of the Jovian planets. The Great Red Spot of Jupiter, seen in the upper left central portion of Figure 23–1, is a generally brick-red feature observed probably for the last three centuries. It is of unknown composition and unknown origin. Jupiter is a powerful source of radio emission, but unlike that of Venus this emission does not come from any underlying surface; instead, it is probably synchrotron radiation such as that which characterizes supernova remnants [Chapter 7]. It is believed that Jupiter has an intense magnetic field which traps charged particles from the solar wind and produces the Jovian analog of the terrestrial Van Allen belts. These charged particles are accelerated by the magnetic field of Jupiter, and are thereby induced to emit synchrotron radiation.

▽ No one knows what lies far beneath the clouds of Jupiter. As in all planetary atmospheres, the density of the air must increase with the depth below the clouds. From the motions of the satellites of Jupiter, it can be concluded that there is not a solid surface a short distance below the clouds. The atmosphere is extensive, and very high densities (for a gas) will be reached only a few hundred kilometers beneath the clouds. The atmosphere of Jupiter must have densities which approach the densities of ordinary solids. Under these enormous pressures, materials take on unfamiliar properties; walking through the lower reaches of the Jovian atmosphere would be very similar to swimming. In a sense Jupiter is a vast planetary ocean, not of water, but of hydrogen and helium, with smaller amounts of methane, ammonia, and water. Far below, in the innermost recesses of the Jovian

FIGURE 23–1. *Jupiter in blue light showing bands and belts parallel to the equator, and, in the upper left-hand corner, the Great Red Spot. (Courtesy of Mt. Wilson and Palomar Observatories.)*

interior, according to the German-American astronomer Rupert Wildt of Yale University, metallic hydrogen abounds, a form of hydrogen uncommon to everyday experience, produced only under enormous pressures.

▽ It has been customary to dismiss instantly the possibility of life on Jupiter, with a reference to poisonous gases and freezing temperatures. But the gases of the Jovian atmosphere, let us recall, are far from unambiguously poisonous; indeed, they are just the components of the primitive atmosphere in which life arose on Earth [Chapter 17]. And while the temperatures at the visible cloudtops are very low, temperatures approaching room temperature will almost certainly be found a few tens of kilometers further down. Ultraviolet light supplies energy to the upper atmosphere, and lightning discharges must be common in the clouds. With an atmosphere of hydrogen, methane, ammonia, and water, an abundance of energy sources, and equable temperatures, we have exactly the conditions used in experiments on the origin of life on Earth [see Chapter 17]. Theoretical models of the Jovian atmosphere below the visible clouds, constructed by the French

astronomer Roger Gallet, at the U.S. National Bureau of Standards, even predict the existence of a thick liquid water cloud. It therefore seems inescapable that large quantities of organic molecules are being produced abiologically in the atmosphere of Jupiter today, and that such conditions have been maintained for the past 4.5×10^9 years. Jupiter is in fact an immense planetary laboratory in prebiological organic synthesis.

▽ It is much more difficult to say anything about the possibility of the origin and present existence of life on Jupiter. For example, we can imagine organisms in the form of ballasted gas bags, floating from level to level in the Jovian atmosphere, and incorporating pre-formed organic matter, much like plankton-eating whales of the terrestrial oceans. However, such speculations are profitless, except as an encouragement to future studies. But when the preliminary detailed reconnaissance of our solar system is completed a century hence, it may well turn out that the greatest surprises and the most striking advances for biology attended the exploration of Jupiter.

II. Saturn, Uranus, Neptune, and Pluto

▽ The other Jovian planets, Saturn, Uranus, and Neptune, are believed to be similar in their general structure and composition to Jupiter; but since they are further from the Sun, their cloudtops are colder, and since they lie further from us, they are more difficult to study. Saturn [Figure 23–2] has spots and atmospheric bands similar to those of Jupiter, but in addition it possesses one extraordinary feature: its rings. The rings, rather than being a thin sheet, as some early astronomers naïvely supposed, are an immense swarm of small particles orbiting Saturn, much as the planets and asteroids orbit the Sun. The innermost constituents of the rings revolve about Saturn in significantly less time than the outermost. The thickness of the rings has been determined by the American astronomer Fred Franklin, of the Smithsonian Astrophysical Observatory, as at most a few centimeters. The spectra of the rings and other theoretical considerations indicate that they may be composed of ordinary water ice; if not, they are at least covered by ice. Thus, the rings of Saturn are, more or less, made of snowballs.

▽ A photograph of the outermost planet, Pluto, is shown in Figure 23–3, compared with standard photographs of Mars, Jupiter, and Saturn. Pluto is 40 astronomical units from the Sun, and even in this photograph by the world's largest optical telescope it is indistinguishable from the background stars.

III. Satellites of the Jovian Planets

▽ The 31 natural satellites of planets in the solar system vary greatly in size, appearance, and over-all density. What they have in common is our ignorance

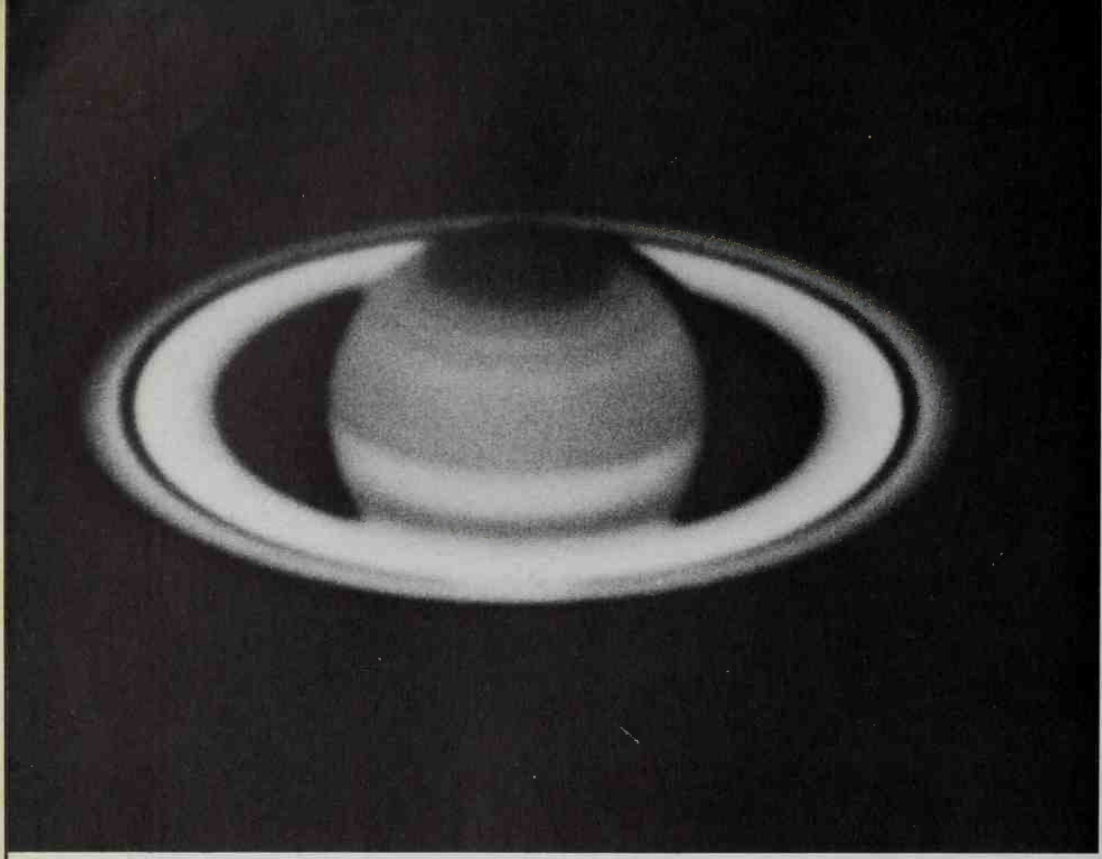


FIGURE 23–2. *Saturn and its ring system. A system of bands and belts on the body of Saturn can be seen as can one of several divisions in the rings. (Courtesy of Mt. Wilson and Palomar Observatories.)*

about them. The distribution of satellites in the solar system is as follows: the Earth, of course, has one; Mars, 2; Jupiter, 12; Saturn, 9; Uranus, 5; and Neptune, 2. By the late 1970's or the early 1980's investigation of the satellites of the Jovian planet by space vehicles may begin. Until then we must be content with our very limited knowledge and the names which astronomers have granted to the satellites. Particularly exotic are the names of the satellites of Saturn, which are, in order of distance from Saturn, Mimas, Enceladus, Tethys, Dione, Rhea, Titan, Hyperion, Iapetus, and Phoebe.

▽ The satellites vary in size from the satellites of Mars, Phobos, and Deimos, which are only several kilometers in radius, to Ganymede, the giant satellite of Jupiter which is half again as large as our moon. Roughly the same size as our moon are the satellites Triton of Neptune, Titan of Saturn, and the satellites of Jupiter, Io, Europa, and Callisto. Galileo discovered Io, Europa, Callisto, and Ganymede, and for this reason they are called the Galilean satellites of Jupiter. The average densities of these satellites vary from the density of a typical rock—our moon, for example—to densities apparently below the density of water—as in the case of some of the satellites of Saturn. Such low density satellites may be nothing more than large balls of icy fluff.

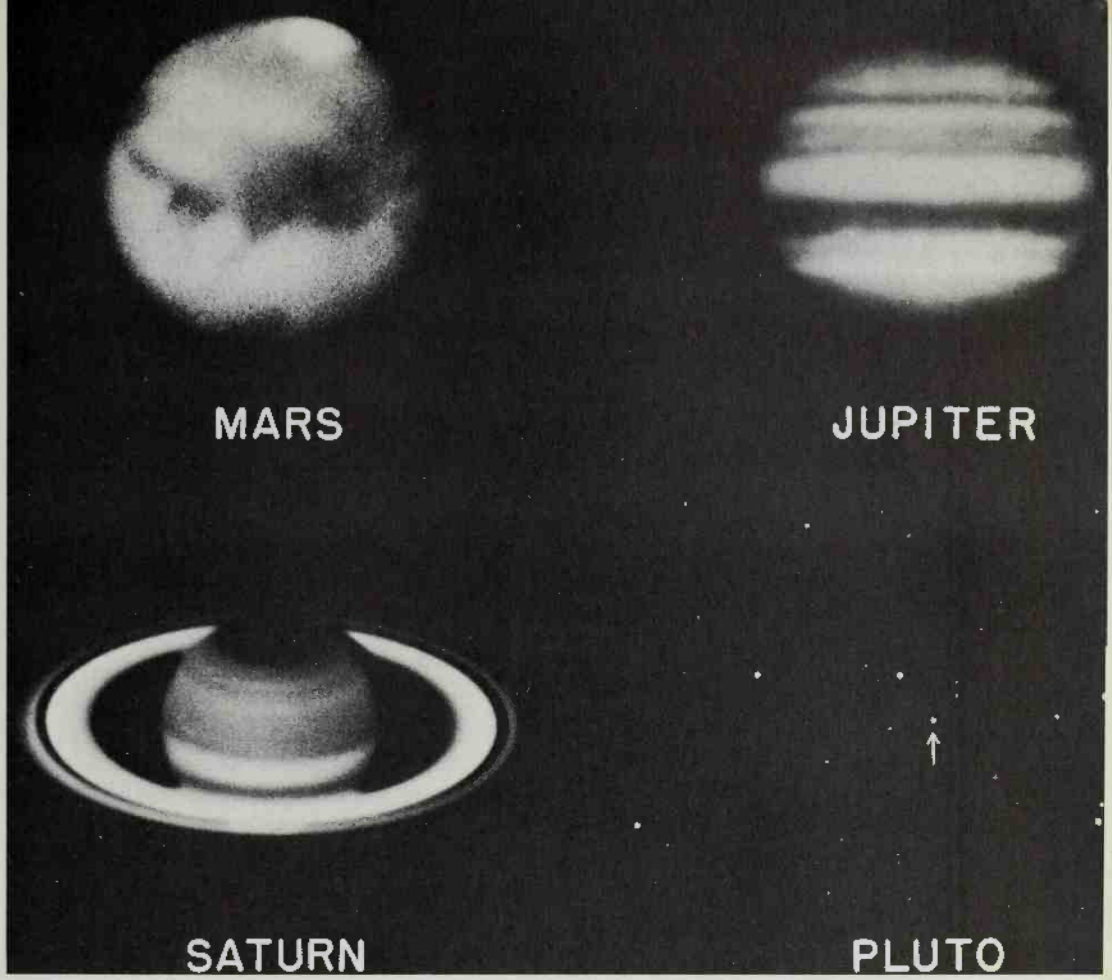


FIGURE 23–3. *Comparison of typical photographs of Mars, Jupiter, Saturn, and Pluto. (Courtesy of Mt. Wilson and Palomar Observatories.)*

▽ Aside from our own satellite, we know most about the Galilean satellites of Jupiter, Io, Europa, and Callisto. Crude maps of them have been prepared from visual observations and are displayed in Figure 23–4. Because they are so small and so far away, the Galilean satellites are even more difficult to observe than Mercury. All four, like Mars, show irregular patterns of dark and bright features, with some concentration of dark features towards the equator. The nature of the dark areas is entirely unknown. Similar dark features on Mars are thought by some to be connected with biological activity [Chapter 20]. The surface temperatures on these bodies are very low, $-100°C$ or less in the bright areas; somewhat warmer in the dark areas. We are largely ignorant of possible biological processes indigenous to such low temperatures.

▽ Figure 23–4 does not give a completely adequate picture of the relative brightness of the Galilean satellites. Io is about three times brighter in the visible than Callisto, for example (and Titan reflects even less light than Callisto.) The colors of Io and Europa vary with position on their surfaces to a much greater

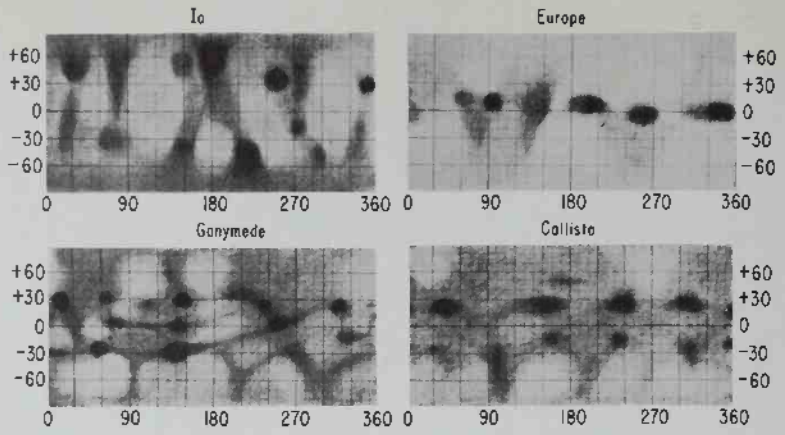


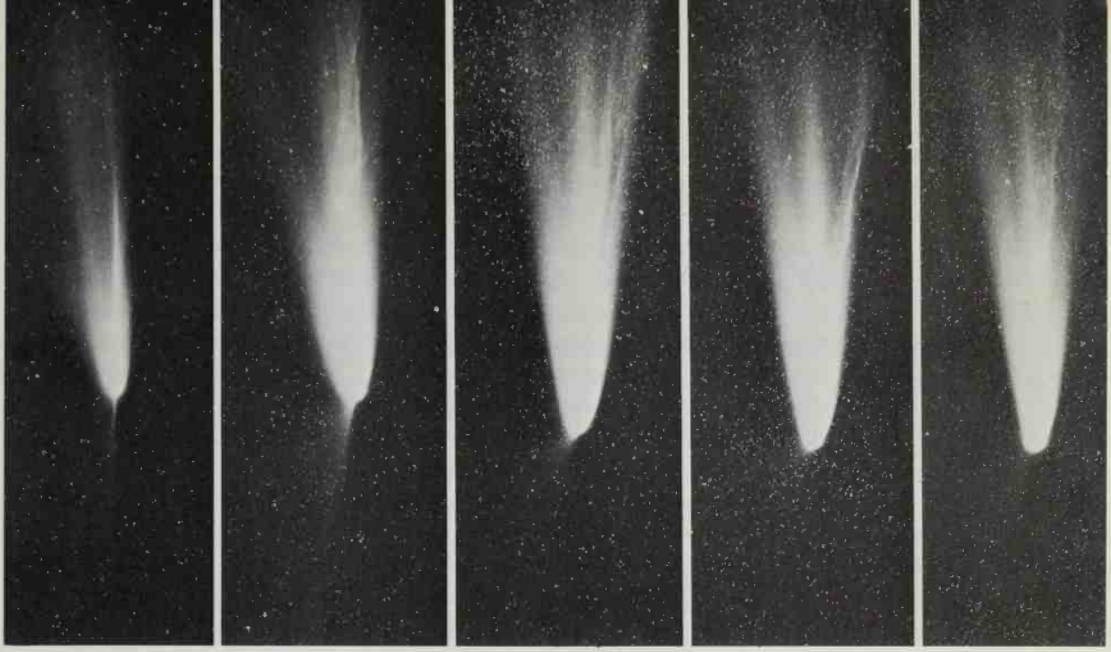
FIGURE 23–4. *Maps in Mercator projection of the four Galilean satellites of Jupiter, Io, Europa, Ganymede, and Callisto. These maps are based on visual observations, although very recently some photographs of the Galilean satellites showing surface detail have been obtained. (Courtesy of Dr. Audouin Dollfus, Meudon Observatory of Paris.)*

extent than do the colors of, for example, the other Galilean satellites. Both Io and Saturn's satellite Titan are extremely red; they reflect much less light at short visible wavelengths than at long. This fact may be connected with the presence of an atmosphere on these two satellites. Titan is known, on spectroscopic grounds, to have an atmosphere containing methane and there is some indirect evidence for an atmosphere on Io. The bright areas of the Galilean satellites are very likely to be snow, but whether this is snow of H_2O, NH_3, or CH_4, we do not yet know.

IV. Comets

▽ Comets have been a subject of fear, awe, and reverence from the beginnings of recorded history. Comets are seen as brilliant streaks against the familiar backgrounds of stars, as in Figure 23–5. Generally they do not perceptibly move against this stellar background during a single night's observing. A period of months characteristically elapses between the time when they are first seen and the time when they are too distant to be seen. New comets are discovered each year, but only rarely are they visible to the naked eye. One naked eye comet is Arend-Roland, shown in Figure 23–5, and named after the two amateur astronomers who discovered it.

▽ When a bright comet appears that is visible to the naked eye, interesting public reaction sometimes follows. When Halley's Comet last appeared, in 1910, the Earth passed through the comet's tail, which was known to be composed of poison gases. Many people expected the asphyxiation of everyone on Earth, an expectation which led to several celebrated sybaritic parties saluting the end of the



April 26 April 27 April 29 April 30 May 1
1957

FIGURE 23–5. *Five views of the comet Arend-Roland, all but one taken on consecutive nights. (Courtesy of Mt. Wilson and Palomar Observatories.)*

world. Their outcomes were somewhat anticlimactic: the density of matter in comet tails is so extraordinarily small that there were no detectable consequences on Earth, except the aftermaths of the celebrations.

▽ The spectral analysis of sunlight reflected from comets (which do not shine by their own light) has indicated the presence of the molecules C_2, C_3, CN, CH, NH, NH_2, OH, CO, CO_2, and N_2, either in neutral or in ionized forms. Many of these molecules—C_2 and C_3, for example—are unfamiliar because they are chemically highly reactive. They exist in comet heads and comet tails only because the density there must be very low, so that the probability of one of these molecules colliding and interacting with another molecule is small. The molecules therefore have lifetimes long enough to absorb incident sunlight and be detected by Earth-based astronomical spectroscopy. The low density of the comet tails is attested to by the fact that we can see stars, and in some cases, the Sun, through the tail and even, in some cases, through the head of the comet. The only dense portion of the comet is the nucleus, which is about 10 km across. The diffuse tail, on the other hand, may be 10^7 to 10^8 km long.

▽ The most widely accepted theory of the nature of comets, due to the American astronomer Fred L. Whipple, of the Smithsonian Astrophysical Observatory, holds them to be conglomerates of methane, ammonia, and water ices, with an admixture of impurities. Many comets are in extremely elongated orbits about the Sun. They are not observed until they are at about the distance of the orbit of Mars. At that point, the intensity of sunlight and the solar proton wind is sufficient to excite the molecules in the nucleus, force them outward from the Sun by

radiation and particle pressure, and produce the prominent cometary tail. The radiation pressures involved are physically just the same as the pressures discussed in the context of the panspermia hypothesis [Chapter 15].

▽ When radiation falls upon such an orbiting snowbank, chemical interactions among H_2O, CH_4, and NH_3 will produce organic molecules, as has been demonstrated in laboratory experiments on simulated comets. The dissociation of these organic molecules by solar radiation leads to the molecular fragments, like C_2 and C_3, which are observed spectroscopically. The tails so produced are sometimes multiple, as in Figure 23–5, showing a great complexity in the fine structural details which may vary from day to day. When the comets are receding from the Sun, their tails precede them in their flight.

▽ Many comets are thought to arise from a region several hundred thousand Astronomical Units from the Sun, essentially in interstellar space. The Dutch astronomer Jan Oort, of the University of Leiden, believes that there is a vast population of cometary nuclei in orbit around the Sun at these enormous distances; these are occasionally perturbed by passing stars into orbits which enter the inner solar system and are detected by astronomers on Earth. Subsequent perturbations by the Sun, Jupiter, and other planets may drive the comets into orbits of much smaller size, so that their periodic returns may occur frequently enough to be noted on Earth. Halley's Comet has a period of about 76 years, and has been observed in historical records some 29 times. The next return of Halley's Comet will be in 1986.

▽ If comets are ordinarily denizens of interstellar space (although in orbit about the Sun), then their detailed examination would give significant clues on the still unexplored regions between the stars. In addition, many theories of the origin of the solar system hold that the comets are similar to the original material from which the solar system was formed. A detailed chemical analysis of cometary nuclei might be useful in studies of the early history of the solar system, including primitive organic chemistry. Comets approach sufficiently closely to the Earth that a rendezvous between a comet and a spacecraft is almost within the range of present technology. Interests in a cometary probe have been expressed by the European Space Research Organization, a joint establishment of many Western European countries for the scientific exploration of space.

V. The Asteroids

▽ Between Mars and Jupiter a vast horde of particles, ranging from Ceres, 350 km in radius, down to pea-sized grains and smaller, are orbiting the Sun. This is the asteroid belt. Repeated collisions among asteroids over the history of the solar system have produced a great number of small particles. The asteroid belt has been called a "cosmic grist-mill." The collisions frequently inject material into elongated orbits; some of them intersect the orbit of the Earth, and a small fraction, quite by accident, collide with the Earth in its travels about the Sun. These

asteroidal fragments are the meteorites. They are divided into two general varieties, the irons and the stones, whose composition is approximately indicated by their names.

▽ The Earth can be considered divided into two regions: the core, which is composed, we think, primarily of iron; and the mantle and crust, which are composed primarily of silicates. Were the Earth shattered by some unimaginably powerful explosion, we could imagine interplanetary space littered with debris similar to the meteoritic irons and stones. There are, indeed, serious scientists who hold that the asteroids are the fragments of a destroyed planet. The total mass of the present asteroid belt is roughly equivalent to a sphere with the same density as the Earth, but with a diameter of about 1000 km. This is equivalent to the mass of a small Jovian satellite, and not to a planet. But for all we know, much of the material may have escaped during the explosion.

▽ If the asteroids are a shattered planet, we may wonder if some previous technical civilization blew its world apart. We may point out that the destruction of a planet by a technical civilization requires a state of advance—if that is the word—far beyond our present capabilities. For example, the great meteor crater in Arizona was produced by a very minor asteroidal fragment, but the energy expended in carving out the meteor crater by explosive impact is comparable to an explosion of a 20-megaton nuclear weapon, close to the present technological limits on thermonuclear devices.

▽ The bulk of astronomers who have studied the problem believe that the asteroids are not the result of a titanic explosion, but rather the remnants of a planet which never formed, perhaps because of the tidal perturbations from the nearby massive planet Jupiter.

▽ The major fraction of the stony meteorites are known as chondrites, because of the chondrules, small, glassy inclusions in the stones. Of the chondrites, a small fraction are called carbonaceous, because they contain significant quantities of organic matter. About 2% of the known meteorites are carbonaceous chondrites, and about 0.5% of the carbonaceous chondrites by mass are made of organic matter. Thus, about 10^{-2} percent of all meteoritic material which has fallen on the Earth is organic matter. For comparison, the mass of the Earth is 6×10^{27} gm; the mass of the biosphere—all the living and non-living organic matter on Earth—is a few times 10^{17} gm. Thus, the Earth is composed of something like 10^{-8} percent organic matter, and most of this is of biological origin. Why is there a million times more organic matter in the asteroid belt than on the Earth?

VI. The Carbonaceous Chondrites

▽ The carbonaceous chondrites have been used in three different ways to argue for the presence of extraterrestrial life. First, there is the organic matter itself. In 1864, a meteorite that fell near Orgueil, in southern France, was analyzed by Jon Jacob Berzelius, of Sweden, and several other of the famous chemists of the day.

They were astonished to find a large content of organic matter. The possibility of contamination by terrestrial organic matter—for example, in the soil on which the meteorite fell—was shown to be insubstantial, and the possibility of living organisms on the parent body of the Orgueil meteorite was seriously raised in the scientific literature.

▽ In recent times, Orgueil and other carbonaceous chondrites have been subjected to a rigorous and multifaceted chemical analysis. There seems little doubt of the existence of indigenous high molecular weight paraffins, long-chain aromatic hydrocarbons like tar, fatty acids, and porphyrins, the building blocks of chlorophyll. We know today, as Berzelius did not, that very complex organic molecules may be produced in the absence of life, under reducing conditions [see Chapter 17]. Thus, in itself, the demonstration that organic matter exists in the meteorites does not prove that life also exists on the meteorite parent body. It has been argued that the relative abundance of organic molecules in the carbonaceous chondrites is similar to that in undisputed samples of biological origin—for example, recent sediments, or even butter. But not enough is known about the relative distribution of organic molecules in prebiological synthetic reactions to give much weight to this argument.

▽ An even more intriguing discovery has been made by the Hungarian-American geochemist Bartholomew Nagy, of the University of California, and his colleagues. We recall from Chapter 14 that the optical activity of organic molecules is one of the hallmarks of their biological origin. With a few insignificant exceptions (insignificant because the conditions used are unlikely to recur in nature), all organic molecules produced under simulated prebiological conditions are racemic mixtures of approximately equal numbers of dextrorotatory and levorotatory stereoisomers [see Chapter 14]. Nagy and his colleagues extracted a certain fraction of the organic matter from the Orgueil meteorite and tested it for optical rotation. They found that the Orgueil organic matter was distinctly levorotatory. As controls for possible sources of contamination, they used dust and wax from the museums in which the Orgueil meteorite has been stored, and pollen, soil samples, and other organic matter. All the samples of terrestrial origin similarly prepared showed dextrorotatory optical activity.

▽ Let us consider the significance of these results. The Orgueil meteorite, like all chondrites, is porous. It had been sitting in a French museum for a century. Sizable opportunity for contamination existed. Yet it appears that all possible contaminants are dextrorotatory, while the meteorite organic matter is levorotatory. Must we then conclude that the meteorite organic matter was initially levorotatory, and that there was biological activity on the meteorite parent body? Not necessarily, for we can imagine how the optical activity may have been generated, even after the meteorite's arrival on Earth.

▽ Suppose that the Orgueil meteorite originally had only a racemic mixture of organic molecules, but that these molecules were palatable to terrestrial microorganisms. Life forms here preferentially metabolize one of the two stereoisomers.

Since most terrestrial organic matter of the type extracted by Nagy and co-workers is dextrorotatory, the dextrorotatory organic matter in Orgueil may have been digested and metabolized by terrestrial microorganisms. The levorotatory fraction would have remained untouched. In time, the meteorite would be left with only levorotatory organic matter as a result of terrestrial biological activity. It seems a pity that, because of the possibility of contamination, we cannot unambiguously deduce biological origins from optical activity. Whether the extracted fraction of Orgueil were levorotatory, dextrorotatory, or racemic, we would be unable to draw any significant conclusions about biological origins. If optical activity is to be used for the detection of extraterrestrial life, it is clear that very stringent sterilization techniques must be used; for biological contamination can destroy the entire utility of the method.

▽ Contamination is also a problem in the second argument from carbonaceous chondrites—the discovery of organized elements. In the course of their investigations of carbonaceous chrondrites, Nagy and the Hungarian-American microbiologist George Claus, of New York University Medical School, discovered that these meteorites seemed full of highly structured forms, roughly 10 microns across, which seemed to them obvious remnants of living microorganisms. Some of them were fairly amorphous, some nondescript spheres; but others had highly provocative features, such as the organized element of Type 5, seen in Figure 23–6. Here was a structure which, when stained, appeared to be of indisputable biological origin, because of the complexity of its form. It was apparently embedded in the meteorite, and resembled no known terrestrial microorganism. Is the organized element of Type 5 the first known example of extraterrestrial life?

▽ We have mentioned that carbonaceous chondrites are porous. In the course of a meteorite's entry into the Earth's atmosphere, it "breathes," and large volumes of air, containing microorganisms, are drawn through its structure. Some of them may easily become embedded in the interior of the chondrite. Although the organized element of Type 5 resembled no known terrestrial microorganism, the Latvian-American geochemist Edward Anders and the American pathologist Frank Fitch, both of the University of Chicago, found that when ordinary ragweed pollen is prepared and stained by the same procedure used by Nagy and Claus, a structure is produced which looks extraordinarily like the organized element of Type 5 [see Figure 23–7]. Two alternatives remain: We may assume that ragweed flourishes in the asteroid belt, as in the illustrations of Antoine de St. Exupery's book *The Little Prince;* or we may assume that the Orgueil meteorite was contaminated by ragweed pollen. Somewhat reluctantly we must choose the latter alternative.

▽ This does not dismiss the wide variety of other identified and named organized elements. But consider the problems inherent in their identification. We would like to be sure that an organized element was present in the meteorite when it fell. In many cases, this is impossible, because of the porosity of the meteorite. If the organized element has an exotic morphology, we should like to be sure that no terrestrial microorganisms can give a similar appearance. This is not always easy.

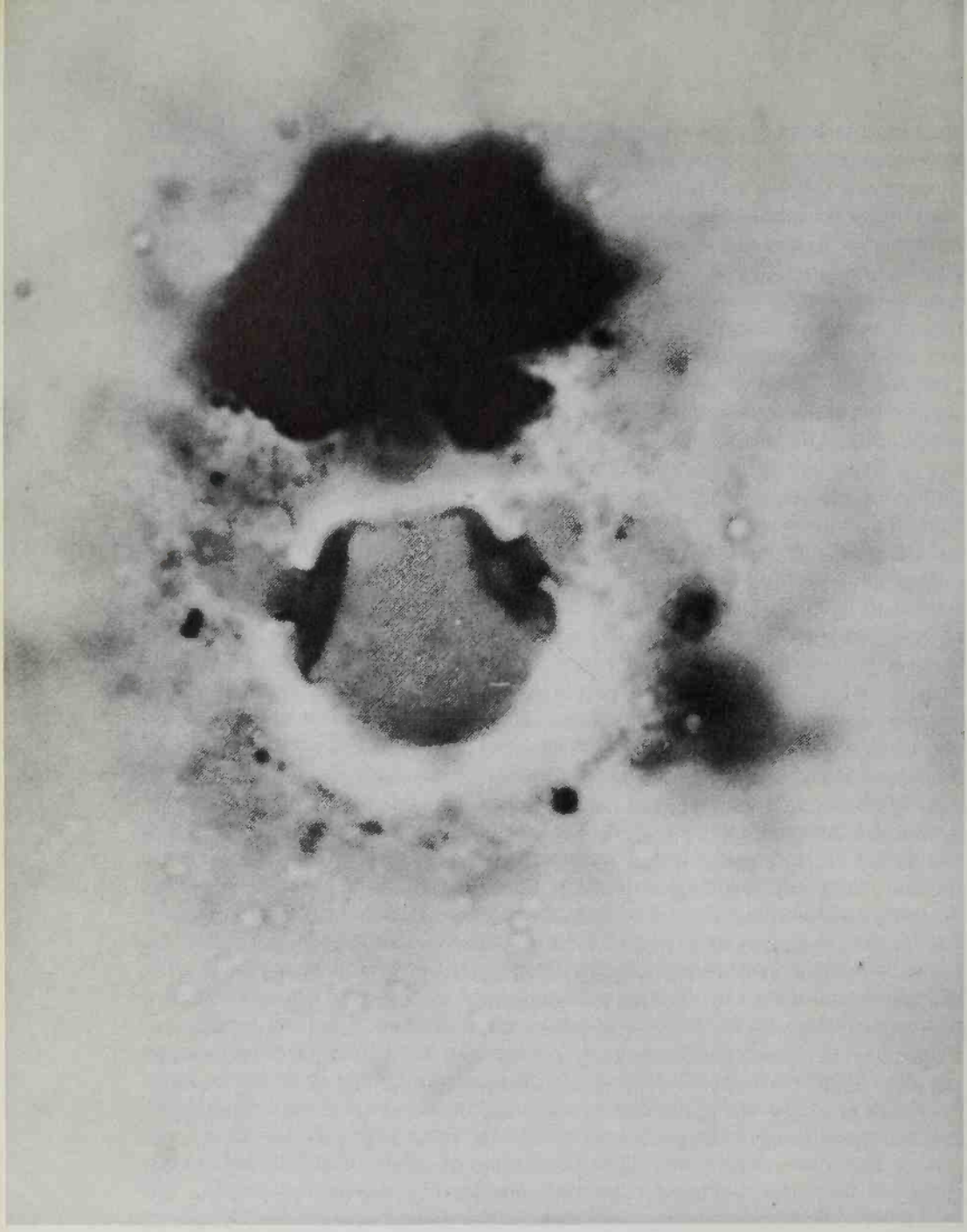


FIGURE 23–6. *The structured object below the black smudge is an organized element of Type 5, in the designation of Claus and Nagy, with an applied Gridley stain. (Courtesy of Prof. Edward Anders and Prof. Frank Fitch, University of Chicago, and Prof. Bartholomew Nagy, University of California.)*

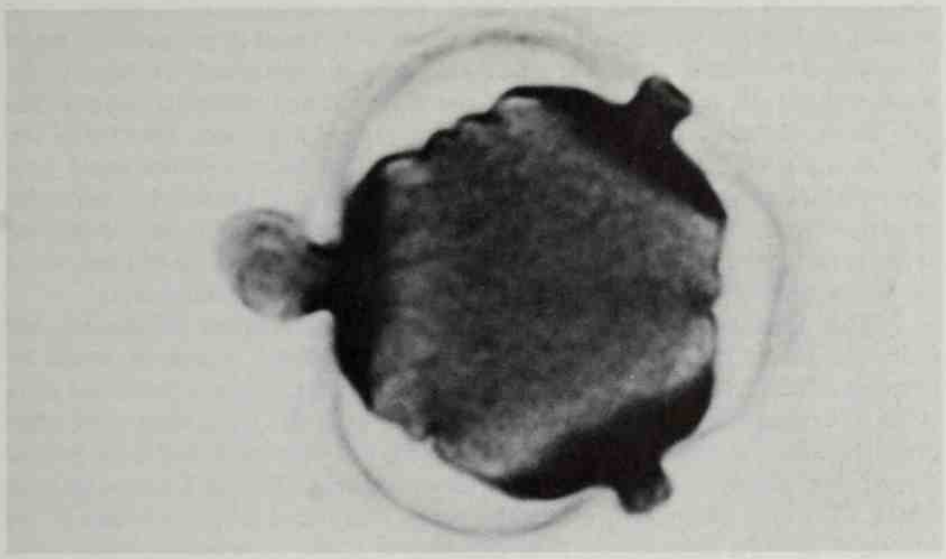


FIGURE 23–7. *Gridley stained* Ambrosia elatior *pollen grain. (Courtesy of Prof. Edward Anders and Prof. Frank Fitch, University of Chicago.)*

We should like to demonstrate that the organized elements are in fact themselves composed of organic matter, or a likely fossil replacement; but because they are so small, it is difficult to perform such microchemical analysis.

▽ Finally, even if the organized elements could be proved to be morphologically unique, composed of organic matter, and indigenous to the meteorite, we still have not proved that life exists on the meteorite parent body. As we saw in Chapter 17, experiments related to the origin of life have shown that highly structured forms of organic matter can be produced in the absence of life. These problems are extremely difficult, and a simple solution with a single experiment is unfortunately unrealistic.

▽ In a third category of experiments, some microbiologists have attempted to extract living microorganisms from the interiors of carbonaceous chondrites. They have tried to use extremely careful techniques to extract uncontaminated cores from the interiors of the meteorites, and to perform the microbiological cultures under sterile conditions. But as we have seen, the meteorites are porous, and contamination is virtually unavoidable. The Soviet microbiologist A. A. Imshenetskii, of the Soviet Academy of Sciences, has shown that completely sterilized meteorites become contaminated with microbes even in their deep interiors, simply after sitting on a shelf for a short period of time. △ In recent years, the Soviet scientists Bairiev and Mamedov announced to the press that they had "discovered" a special variety of bacteria in the *iron* Sichotz-Alinscii meteorite. However, it soon became apparent that this "discovery" was invalid because of the crude nature of the investigations. ▽ Similarly, in the United States, Frederick D. Sisler, of the U.S. Geological Survey, cultured samples from the interiors of carbonaceous chondrites

and found that after a long period under sterile conditions, his nutrient broth clouded, and several varieties of microorganisms were found to be present. One of them was a facultative anaerobe; that is, although it was capable of living in the absence of oxygen, it also showed a preference for utilizing molecular oxygen. The only planet on which significant quantities of oxygen have been detected is our own. It is out of the question that an extraterrestrial microorganism could have developed the complex electron transfer apparatus required for utilizing molecular oxygen without a long period of evolution in an oxygen environment. Despite the apparent unfamiliarity of Sisler's microorganisms, the fact that one of them was a facultative anaerobe is strong evidence that they are in fact contaminants. △

What conclusions, then, can we draw from the organic substances and inclusions which have been found in the meteorites? Of course, it would be tempting to say that the carbonaceous chondrites constitute definite proof that there is life on other planets. In the history of science, however, there have been many instances where a desired answer was accepted, not because it had been proved correct, but merely because it was the answer sought. An old Chinese proverb states: "The man who eagerly awaits the arrival of a friend should not mistake the beating of his own heart for the thumping hooves of the approaching horse." The true nature of the carbonaceous meteorites—whether they actually contain the remnants of extraterrestrial life, or whether there is some other explanation for the presence of organic matter and organized elements—remains an unresolved question. ▽ The difficulties, frustrations, and scientific debates which this question has engendered may presage the consequences of the first search for life on Mars by an unmanned lander. But the experience gained in the analysis of the carbonaceous chondrites will be invaluable in other searches for extraterrestrial life. △

24

Life in other solar systems

. . . And yet 'tis not improbable that those great and noble Bodies have somewhat or other growing and living upon them, though very different from what we see and enjoy here. Perhaps their Plants and Animals may have another sort of Nourishment there.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

Let us . . . consider a giant man sixty feet high—about the height of Giant Pope and Giant Pagan in the illustrated *Pilgrim's Progress* of my childhood. These monsters were not only ten times as high as Christian, but ten times as wide and ten times as thick, so that their total weight was a thousand times his, or about 80 to 90 tons. Unfortunately, the cross sections of their bones were only a hundred times those of Christian, so that every square inch of giant bone had to support ten times the weight borne by a square inch of human bone. As the human thigh bone breaks under about ten times the human weight, Pope and Pagan would have broken their thighs every time they took a step. This was doubtless why they were sitting down in the picture I remember. But it lessens one's respect for Christian and Jack the Giant Killer.

J. B. S. Haldane, *On Being the Right Size* (1932)

▽ In previous chapters, we have considered the properties of the stars, the likelihood that other stars have planetary systems accompanying them through space, the requirements for the origin and early evolution of life, and the range of planetary environments which may conceivably support biological processes. Let us now try to collect these results, attempt to specify which stellar types are likely to have inhabited planets, and perhaps even indicate which of the nearer stars are the most likely candidates. While much of what we say will concern life in general, bearing in mind the objectives of Part III of this book, we will also consider at least some minimal criteria for the development of intelligent life.

▽ On our own planet, it has taken at least several hundreds of millions of years for the evolution of simple one-celled organisms from the materials of the primitive atmosphere and oceans. If we consider this a typical figure, it follows that planets of stars of spectral type earlier than A0 have not resided on the main sequence sufficiently long for the evolution of protozoa [see Table V].

▽ Again from terrestrial analogy, an even larger time interval seems to be required △ for the simplest forms of life to evolve into intelligent beings ▽ capable of constructing a technical civilization. △ The driving force behind this evolution is the natural selection of randomly produced mutations. A vast number of mutations must take place before one of them, purely by chance, aids in the development of a more advanced life form. ▽ On our planet, this process has taken about three billion years. If we believe this figure to be typical, then △ technical civilizations will probably be found only on planets associated with stars which have resided on the main sequence for at least several billions of years, that is, stars of spectral type later than F0 [see Table II, Chapter 6]. The argument from stellar rotation [see Chapter 13] has suggested that only stars of spectral type later than F2 are accompanied by planetary systems. Hereafter we shall assume, with one exception, that all main sequence stars of spectral type later than F2 have associated planetary systems.

The one exception is the first-generation stars—the sub-dwarfs [see Chapter 6]. These stars contain only negligible amounts of the heavier elements and are unlikely to have Earthlike planets. ▽ They may, however, have planets of the Jovian type; as we have seen [Chapter 23], the production of complex organic molecules, and perhaps even living systems, seems to be possible on such worlds. But we again emphasize that the general existence of planetary systems of other stars is not yet rigorously demonstrated. The existence of dark companions of the nearest stars, the argument from stellar rotation, and contemporary theories of the origin of solar systems together strongly point to a plurality of habitable worlds.

But only future developments in astronomy can demonstrate beyond the shadow of a doubt the existence of large numbers of such planets. △

All planets and satellites are not habitable. We do not expect to find life on the lunar surface because the Moon is devoid of an atmosphere and lacks surface water, ▽ or other liquids which might perform the biological role of water. △ The chemical reactions necessary for the origin of life can occur only under a certain range of temperatures. If the local climate is too hot or too cold, ▽ the larger synthesized molecules will immediately be broken down by the heat, or the rate of reaction will be so slow that relatively few chemical reactions will have occurred in the lifetime of the planet. △ In either case, life will not arise.

We can imagine each star surrounded by a spherical shell, throughout which planetary temperatures are equable, and the origin and development of life are possible. ▽ We can call this region the "zone of habitability," or the "ecosphere." The science of ecology is concerned with the relation between organism and environment. △ If a planet lies too close to its sun, the high local temperatures may preclude life. A possible example is the planet Mercury, where the temperatures on the surface facing the Sun are higher than the melting point of lead. Similarly, if a planet lies too far from its star, the local temperatures will be too cold for the origin and development of life. ▽ It is difficult, for example, to conceive of life on the planet Pluto, where the temperatures are approximately −200°C. △ However, very high temperatures are more hazardous for living processes than very low temperatures. We know that simple forms of life, such as viruses and bacteria, can withstand extremely low temperatures for extended periods of time.

The temperature of a planet is determined first of all by the amount of radiation that it receives from its sun on each unit of area of its surface in a given period of time. For this reason, the size of the ecosphere varies from star to star, depending on the stellar luminosity. If the local sun is of early spectral type, and hence has high luminosity, the dimensions of its ecosphere will be large. On the other hand, if the local sun is a red dwarf (of spectral types between late K stars and M's), the luminosity will be low, and the external radius of the ecosphere will be very small—smaller, in fact, than the radius of the orbit of the planet Mercury.

▽ Let us assume that planetary systems everywhere have the same relative dimensions as our own, the innermost planet lying at about 0.4 A.U. from the Sun, the outermost, at about 40 A.U., and the spacing of the planets also more or less as in our solar system. Then the probability that red dwarfs have even one planet within the local ecosphere is very small. But we do not in fact have any good theory on the expected dimensions of extraterrestrial solar systems. For all we know, the stars of late spectral type have planets which in effect huddle together very close to their star, while stars of early spectral type may have planetary systems which are very distended. △ We cannot exclude the possibility that red dwarfs, in general, have planets within their ecospheres.

If we exclude red dwarfs, and assume that habitable planets occur only around main sequence stars lying between spectral types F2 and K5, then only one to two percent of the stars in our Galaxy simultaneously meet the requirements ▽ for long

lifetimes and at least one planet within the stellar ecosphere. △ However, since there are approximately 150 billion stars in our Galaxy, we still arrive at the comforting conclusion that at least a billion may have habitable planets.

▽ The American engineer Stephen Dole, of The RAND Corporation, believes that there is an additional reason that stars of later type than K1 are unsuitable. The luminosity of such stars is so low that planets must be very near their stars, for equable temperatures to be maintained. But the distance to the star must then be so small, Dole believes, that the planets would soon be locked into synchronous rotation, always keeping the same face to their sun. On a planet with no atmosphere, such trapped rotation would lead to very high temperatures in the illuminated hemisphere, and to very low temperatures in the unilluminated hemisphere; therefore life would probably be excluded from either hemisphere. However, as we have seen in the case of the planet Mercury, even a small atmosphere may be sufficient to carry enough heat by atmospheric convection to maintain a fairly balmy climate on the dark side; also, recent studies of Mercury's rotation suggest that synchronous rotation is not the only possibility for a close planet in an elliptical orbit. △

Let us consider another circumstance which might limit our estimate of the number of habitable planets in the Galaxy. Nearly half of the stars belong to multiple systems. The orbit of a possible planet associated with a binary stellar system would, in general, follow a very complex and unclosed path ▽ —that is, an orbit which does not, like familiar circular or elliptical orbits, close back upon itself. △ The calculation of such an orbit is an extremely complex mathematical undertaking, the solution of the so-called restricted three-body problem. Despite its difficulties, this problem is simpler than the general three-body problem, because here the mass of the planet is negligible compared with the mass of the two stars. The planet has no influence on the stellar motion, although its motion is influenced by the more massive stars. Traveling in such a complex orbit in or about a binary system, the planet would at times approach one of the stars very closely; at other times, it would recede to great distances. Thus, its surface temperature would vary greatly ▽ and more or less erratically, posing great difficulties for life, if not excluding it altogether. △

Until recently, it was generally accepted that habitable planets could not be associated with multiple star systems. However, this problem has now been reexamined by the Chinese-American astronomer Su-Shu Huang, of Northwestern University, who has shown that in certain special cases there do exist possible periodic planetary orbits in multiple star systems.

The temperature variation on such planets is within the allowable limits for the development of life. In these special cases, the relative orbits of the stars are approximately circular. Figure 24–1 shows a cross section of the so-called "critical surfaces" in the restricted three-body problem. Clement temperatures are associated with periodic orbits which lie either inside the surface passing through the point L_1, or outside the surface passing through the point L_2. Let us consider the simpler case that the masses of both stars are identical. We express the distance between the stars, measured in Astro-

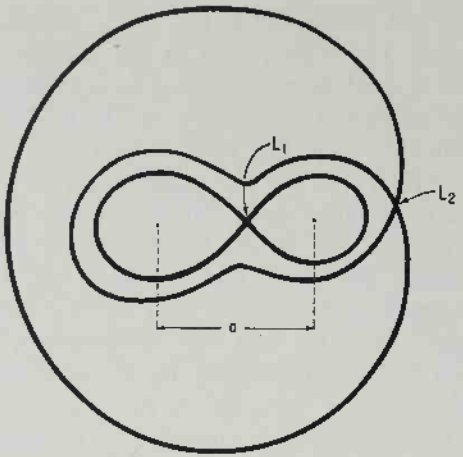


FIGURE 24–1. *Diagramatic representation of the critical surfaces in the restricted three-body problem. Periodic orbits are permitted about each of the stars represented by points, provided these orbits lie within the figure-eight curve. Alternatively periodic orbits are permissible outside the outer curve. Such orbits would circle both stars.*

nomical Units, as a, and the luminosity of each star, in units of the luminosity of our Sun, as l. ▽ Thus, if $a = 5$, the stars are 5 A.U. apart; if $l = 0.8$, then each star has $\frac{8}{5}$ the luminosity of our Sun. △ Huang has shown that inside the surface passing through the point L_1, orbits suitable for the development of life exist, provided that a is greater than 2 $l^{1/2}$. When a is greater than 13 $l^{1/2}$, it is possible to consider each component of the binary star system as a single star. ▽ Thus, for stars of solar luminosity, appropriate orbits exist if they are separated by more than 2 A.U.; and we may ignore the fact that it is a binary star system if the components are separated by more than 13 A.U. △

In many binary systems, the distance between the components exceeds the critical value of 2 $l^{1/2}$. When the components of the double star system are close to one another, suitable periodic orbits must be outside the surface which passes through the point L_2 [again, compare with Figure 24–1]. The calculations of Huang show that when the components of a binary system have equal mass, orbits suitable for the origin and development of life can exist, provided that a is less than 0.4 $l^{1/2}$. Thus, habitable planets are excluded if the distance between the stars is between 0.4 $l^{1/2}$ and 2 $l^{1/2}$. Analogous results can be obtained for the more common case that the masses of the two components are unequal.

Conceivably, multiple star systems are formed in such a way that the simultaneous formation of planetary systems is excluded. ▽ But this is, at the moment, only a conjecture. △ Although it appears much more likely that habitable planets orbit single stars, in principle they can also exist in association with multiple star systems.

The stellar system closest to our own, Alpha Centauri, is in fact a triple system. ▽ The two major components, Alpha Centauri *A* and Alpha Centauri *B*, have spectral types G2 and K4, respectively. Thus, Alpha Centauri *A* has a spectral type very similar to that of our Sun. The third component, Alpha Centauri

C, also called Proxima Centauri, because at times it is the closest star to our Sun, is a red dwarf of spectral type M5e. △ The relative orbits of the two larger components are elliptical. The semi-major axis* of these two components of Alpha Centauri is 23.4 A.U. The distance between these two major components is therefore great enough to permit each of them to have planets with stable periodic orbits of biological interest. However, the extreme eccentricity of the paths of these stars with respect to each other requires special consideration, since the expressions above were deduced for binary systems having circular orbits. Further, Alpha Centauri seems to be a comparatively young system. It is possible that the component stars have not as yet entered the main sequence. ▽ In this case, there would not have been sufficient time for the origin and early evolution of life on planets of the Alpha Centauri system. △

For a star to have a habitable planetary system, the radiation emitted by it must remain approximately constant for perhaps billions of years. ▽ Other factors being equal, a few percent change in the solar luminosity would have drastic effects on the temperature of the Earth [Chapter 16]. △ The overwhelming majority of stars on the main sequence are remarkably constant in their radiation output. Geological studies indicate that our own Sun has varied its luminosity no more than a few tenths of one percent ▽ over the last few hundreds of millions of years △. There is, however, a large class of variable stars where luminosities fluctuate greatly. ▽ Such stars are unlikely to have habitable planetary systems. △

Among other conditions which must be satisfied, if a planet is to harbor indigenous life, are the mass of the planet and the chemical composition of its atmosphere. As we discussed in Chapter 16, these two characteristics are apparently not independent of one another. ▽ We saw there that at a given exosphere temperature and planetary gravitational field, the lighter atoms will preferentially escape to space. As the exosphere temperature becomes higher (because, for example, the planet is nearer its sun), or as the force of gravity declines (because, for example, we are considering a planet of low mass), the rate of escape of all atoms is enhanced.

▽ With the mass, radius, and exosphere temperature of the Earth, hydrogen should escape in geologically brief periods of time, while the rate of escape of oxygen over all of geological time is insignificant. The mass of the Moon, however, is so much less than the mass of the Earth that even if its exosphere temperature were the same as the Earth's, very heavy gases should escape during the lifetime of the solar system. Thus, any residual hydrogen in the Earth's atmosphere, and any residual atmosphere at all on the Moon, must be due to a continuous supply of atmosphere, probably from the interior of these bodies. In the absence of an atmosphere, a planet cannot maintain an ocean of water or of any other kind of liquid; but liquids or very dense gases seem required for molecular interaction in the origin and evolution of life. Thus, except for the fairly remote possibility of the

* ▽ The longest straight line that can be drawn through an ellipse is called the major axis; half the major axis is called the semi-major axis. △

subsurface origin and evolution of life, a planet must have an atmosphere to be habitable. △

On the other hand, a very large planetary mass can also be a limiting factor. For example, the giant planets Jupiter and Saturn have almost completely retained their original atmospheres, rich in hydrogen and helium. If a planet preserves the original composition of the medium from which it was formed, its hydrogen-helium atmosphere must be very dense. The possibility of such a planet forming a hard surface is problematical. ▽ Whether Jupiter, Saturn, Uranus, and Neptune are completely gaseous spheres, with their density rapidly increasing towards their centers, or whether they in fact have a crust of rock or of more exotic composition far below the visible atmosphere, is at the present time unknown. △ We have already pointed out that if the mass of a planet were 5 or 10 times greater than that of Jupiter, it would not differ appreciably from a dwarf star. ▽ We have mentioned that the chemical composition of a massive planet of the Jovian type does not necessarily exclude the origin and development of life, although it does imply that the character of that life would be very different from life on Earth. △

In order for life to arise and develop on a planet, the mass of the planet must lie between certain limits. ▽ For a planet with an exosphere temperature similar to the Earth's (about 1500°K), a mass about 10 times smaller than the Earth's would result in a significant escape of the atmosphere into interplanetary space during geological time. If the mass were several tens of times greater than the Earth's, substantial quantities of hydrogen would be retained, and the planetary chemistry would be highly reducing. If the planetary mass were some 2000 times larger than the Earth's, the "planet" would in fact be a small star. But these mass limits for habitability—between 0.1 and 2000 Earth masses—are so broad that the bulk of expected planets will be included. △

The fact that the terrestrial planets lie in the inner and warmer part of the solar system, while the larger Jovian planets, of reducing composition, lie in the outer and colder portions is probably not a simple coincidence. ▽ It may be that planets of terrestrial type always form in the inner regions of the solar system, where the early dissipation of the lighter gases, hydrogen and helium, is facilitated. Jovian-type planets tend to contain the fully reduced gases methane, ammonia, and water vapor, which are very efficient absorbers of infrared radiation. Thus, Jovian planets in general tend to have very efficient greenhouse effects, so that relatively warm temperatures can be expected, at least at some level in their atmosphere or clouds, even if they are very far from their sun.

▽ With the preceding criteria we could make some estimate of the probable habitability of a given extraterrestrial planet, if only we had some information about that planet. Unfortunately, as we discussed in Chapter 11, the identification and characterization of extraterrestrial planets are just beyond our present capabilities. Such information may be supplied by new astronomical techniques within the next decade. All we can do, therefore, is to survey the nearest stars and estimate what fraction of them have appropriate ecospheres.

▽ If we assume that other planetary systems always have the same distribution

TABLE V

THE TWENTY NEAREST STARS OF SPECTRAL TYPE BETWEEN F2 AND K5

Star	*Spectral type*	*Distance in light years*
Alpha Centauri A	G2	4.3
Alpha Centauri B	K4	4.3
Epsilon Eridani	K2	10.8
61 Cygni A	K5	11.1
Epsilon Indi	K5	11.3
Tau Ceti	G8	12.2
70 Ophiuchi A	K1	17.3
70 Ophiuchi B	K5	17.3
Eta Cassiopeiae A	F9	18.0
Sigma Draconis	G9	18.2
36 Ophiuchi A	K2	18.2
36 Ophiuchi B	K1	18.2
HR 7703 A	K2	18.6
HR 5568 A	K4	18.8
Delta Pavonis	G7	19.2
82 Eridani	G5	20.9
Beta Hydri	G1	21.3
HR 8832	K3	21.4
p Eridani A	K2	22.0
p Eridani B	K2	22.0

of planets as our own system, we will probably obtain a lower limit on the number of nearby stars likely to have habitable planets. We then find that the nearest stars likely to possess habitable planets are, in order of their distance from the Sun, those of Table V. We have presented the nearest twenty such stars. Designations beginning "HR" refer to the Harvard Revised Catalogue. We emphasize that the list would be very much longer had we included late K and M dwarfs. Since the number of nearby stars increases as the cube of the distance from the Sun, the bulk of these stars lie between 17 and 22 light years from the Sun. We have included components of multiple star systems where they have matched the range in spectral types, although there is a suspicion that the process of formation of multiple star systems may preclude the formation of planets.

▽ If we exclude multiple star systems, we find that the three nearest stars of potential biological interest are Epsilon Eridani, Epsilon Indi, and Tau Ceti. It is reasonable that any search for life beyond our solar system should begin with these stars. If we restrict our attention to single stars of spectral type quite close to that of the Sun—let us say, between F5 and G5—we find that the four closest such stars are Tau Ceti, Sigma Draconis, 82 Eridani, and Beta Hydri. If, despite the implications of present evidence, the formation of planetary systems and the origin

of life are rare events, none of the stars listed in Table V would hold inhabited planets. Instead, the nearest life forms would lie at much greater distances beyond the nearest stars.

▽ The range of planetary environments, even in our own solar system, is startling. The airless, waterless surface of our Moon is alternately very hot and very cold. The night side of Mercury, with a very thin atmosphere and no sun at all, nevertheless is at a moderate temperature. Venus, with a massive atmosphere, is at temperatures approaching red heat; Jupiter, with its dense, reducing atmosphere, is very cold at the clouds, and probably very warm beneath them. The force of gravity at the clouds of Jupiter is more than 10 times that at the surface of the Moon. The environments are diverse; each planet and satellite are unique. In other solar systems, we can expect an even greater diversity, although general patterns should also emerge: The distinction between Jovian and terrestrial planets is probably a universal one.

▽ What can we say about the forms of life evolving on these other worlds? We have argued that the early chemical processes leading to the origin of life may be similar on many diverse worlds, although this is far from proved. But it is clear that subsequent evolution by natural selection would lead to an immense variety of organisms; compared to them, all organisms on Earth, from molds to men, are very close relations.

▽ There are limiting sizes to organisms on any given planet. An organism must be large enough to carry out a minimum of metabolic functions required for its continued replication. The smallest organism known on Earth capable of independent replication is called PPLO, for pleuropneumonia-like organism. It is about 10^{-5} cm across. The upper limit to the size of land-dwelling animals derives from several factors. As the Haldane epigraph to this chapter underlines, if an organism is too large, it will be unable to support its own weight. A second limitation concerns the rate of propagation of signals through the animal's nervous system. If the animal is too large, a signal—for example, from the light receptors, saying, "Stop; a crevasse lies immediately ahead"—will be received by the distant legs too late for useful action. As a partial solution to this problem, dinosaurs had extensive neural networks in their posteriors. Larger animals can be maintained if there is a buoyant medium for support, as, for example, in the oceans, or in a very dense atmosphere.

▽ Most familiar organisms have two, four, or six legs, but adaptations to none or to many, as in snakes or millipedes, have occurred. There seems no reason for extraterrestrial organisms to have any particular number of legs—or, for that matter, any legs at all. In other environments, other specialized motility schemes may have been derived. Certainly, at the level of protozoa this is the case, where flagellae, cilia, and even a kind of ramjet, are commonly used for biological propulsion.

▽ The greater the gravity of the planet, the smaller will be the largest animals. On planets with low gravity, there may be organisms which, from our point of view, would be long and spindly. The same, incidentally, applies to architecture in

advanced extraterrestrial civilizations. High-gravity worlds should have short and squat structures; low-gravity worlds are at least permitted more delicate forms.

▽ We have mentioned in Chapter 16 that on other worlds respiration may not be required, and that fairly advanced forms may be found even in reducing environments. The size of respiring organisms is also limited by the method of respiration. There are no insects larger than about a foot across, because insects introduce oxygen into the interior parts of their bodies by diffusion, a much slower and less efficient process than the circulation of blood.

▽ The number of possible sensory receptors in extraterrestrial organisms is apparently limited. On planets with fairly extensive atmospheres or oceans, sensory receptors for direct chemical analysis of molecules in the atmosphere or ocean would clearly be useful. While a variety of bioanalytic techniques would be possible, these senses would be roughly equivalent to our senses of taste and smell. The usefulness of a sense of hearing depends on the composition, and temperature, of the atmosphere, which determines the velocity of sound. Pressure receptors, such as our sense of touch, seem useful in almost any environment.

▽ The most efficient means of sensing distant objects is the reception of electromagnetic radiation. Since the velocity of light is so large, the propagation time on a planetary surface is negligible. Almost all stars of interest emit the bulk of their radiation in what we call the visible part of the spectrum. In general, we should expect more reflected visible light to be available than light of any other frequency. Furthermore, the visible part of the spectrum is the wavelength range least likely to be absorbed by atmospheric constituents. Transitions of the electrons in atmospheric gases result in light absorption in the ultraviolet. The vibrations of molecules cause absorption in the infrared. The rotation of molecules causes absorption at infrared and short radio wavelengths. Thus, for fundamental physical reasons, the visible wavelength interval is a "window" region of transparency in all planetary atmospheres. Another window should, in general, be found at long radio wavelengths, beyond 3 cm. However, there is a primary difficulty in imagining organisms which "see" with radio waves: In order to have any useful resolution—that is, detection of fine visual detail—the effective collecting area must be enormous. To have the same resolving power at 5 cm wavelength that the eye has at 5000 Å wavelength, an extraterrestrial microwave "eyeball" would have to be roughly half a mile in diameter. This seems awkward.

▽ In terrestrial organisms, optical senses are used primarily for observations in reflected sunlight. There are occasional instances of animals which emit visible light, such as certain marine animals and fireflies. The female of the latter species winks seductively at the male. Elsewhere—particularly on worlds where sound propagation is not utilized (for example, because of a very thin atmosphere)—we can imagine the more elaborate development of communication by electromagnetic propagation, probably of visible light, but not necessarily. If such a species communicated by radio waves, despite the attendant poor resolution, we would probably attribute extrasensory perception to them. But it should be emphasized that this is "extrasensory" only in that it is a sense we lack. Such an adaptation can

be based on perfectly sound physical principles. There is some evidence that human beings can sense high-intensity radar, although the mechanism is at the present time unknown.

▽ One eye gives two-dimensional resolution; a second eye, through binocular vision, gives three-dimensional resolution. Three eyes represent not nearly the same improvement over two that two represent over one. But if placed in the back of the head, for example, a third eye might serve some useful purpose. Some animals in the Mesozoic seem to have had three eyes, all in the front of the head, and some physiologists believe that the human pineal gland is a vestigial remnant of a third eye in the middle of the forehead. Representations of the Buddha sometimes show such a third eye.

▽ Devices for acquiring, reprocessing, and excreting food would probably vary widely from world to world, depending on the nature of the food chain and the relationship of the various organisms. There seems no reason to expect elsewhere the same combination of functions that we find on the Earth, where the vocalizing, breathing, and eating organs have been combined, to a certain degree, as have the organs of excretion and reproduction. Elsewhere, different combinations of functions may prevail.

▽ Even such a brief and tentative excursion into extraterrestrial ecology cannot be tested until we have obtained samples of extraterrestrial organisms. Yet such simple considerations are useful, because they shed some light on the selective advantages of the forms and functions of terrestrial organisms. △

12/31/56

"Want To Know How It Ends?"

"I'm sorry, sonny. We've run out of candy."

III

INTELLIGENT LIFE IN THE UNIVERSE

Lights come and go in the night sky. Men, troubled at last by the things they build, may toss in their sleep and dream bad dreams, or lie awake while the meteors whisper greenly overhead. But nowhere in all space or on a thousand worlds will there be men to share our loneliness. There may be wisdom; there may be power; somewhere across space great instruments, handled by strange, manipulative organs, may stare vainly at our floating cloud wrack, their owners yearning as we yearn. Nevertheless, in the nature of life and in the principles of evolution we have had our answer. Of men elsewhere, and beyond, there will be none forever.

Loren Eiseley, *The Immense Journey* (1957)

25

The assumption of mediocrity

. . . That which makes me of this Opinion, that those Worlds are not without such a Creature endued with Reason, is that otherwise our Earth would have too much the Advantage of them, in being the only part of the Universe that could boast of such a Creature . . .

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

. . . the intelligent part of creation is thrust into the compass of a few years, in the course of myriads of ages; why not then into the compass of a few miles, in the expanse of systems?

William Whewell, *Plurality of Worlds* (1854)

Life, even cellular life, may exist out yonder in the dark. But high or low in nature, it will not wear the shape of man. That shape is the evolutionary product of a strange, long wandering through the attics of the forest roof, and so great are the chances of failure, that nothing precisely and identically human is likely ever to come that way again.

Loren Eiseley, *The Immense Journey* (1957)

▽ Are there other intelligences in the universe? Is the Galaxy filled with civilized worlds, diverse and unimaginable, each flourishing with its own commerce and culture, befitting its separate circumstances? Or can it be that we are alone in the universe, that by some poignant and unfathomable joke, ours is the only civilization extant?

▽ The idea that we are not unique has proved to be one of the most fruitful of modern science. The atoms on Earth are the same in kind as those in a galaxy some 5 or 10 billion light years distant. The same interactions occur, the same laws of nature govern their motions. The formal statement summarizing the attraction of massive lead spheres in a terrestrial laboratory can also be used to predict accurately the motions of binary stars, or the orbit of the Moon. One of the major intellectual revolutions of the Renaissance, one for which Copernicus and Galileo fought, and for which Giordano Bruno lost his life, was the idea that the Earth was but one of many planets in our solar system and beyond. △

The power of this idea, ▽ the assumption of our own mediocrity, the thought that our surroundings are more or less typical of any other region of the universe △ has been emphasized by the German astronomer Sebastian von Hoerner at the National Radio Astronomy Observatory in the United States. The ancient Greeks were unaware of the true nature of the stars or the scale of the universe. ▽ Some primitive peoples believe the stars to be lanterns hung from the vault of heaven, or holes cut in the celestial panoply, showing the fires which burn beyond. That the stars are distant suns is a subtler and more powerful idea. △ But in principle, the Greeks could have determined the dimensions of the solar system and the distances to the stars.

Let us assume that the Earth is an average planet, and that the Sun is an average star. Then the diameter of the Earth, its distance from the Sun, and its albedo, or reflectivity, should be characteristic of planets in general. Since the Greeks already knew the approximate dimensions of the Earth (Eratosthenes had performed an essentially correct calculation), a comparison of the apparent brightness of the five planets then known with the apparent brightness of the Sun permits a calculation of the distance from the Earth to the Sun. The value obtained in this way is about twice the true value.

If we were to assume that the ten brightest stars in the sky were also suns like our own, and if we knew how much brighter the Sun appears than these stars, it would be possible to compute *their* distance from the Earth in terms of the distance of the Sun from the Earth. And with the value of the Astronomical Unit obtained from the apparent brightness of the planets, the ancient Greeks could have estimated the average distance between the stars with an error of a mere 10%.

▽ In the seventeenth century, Christian Huygens in fact attempted such a calculation. He constructed a thin metallic plate which could artificially eclipse the image of the Sun. A series of small holes were punctured in the plate, until a hole was made which was so small that the sunlight passing through it seemed no brighter than the star Sirius. Huygens found that his hole had an angular diameter about 1/28,000 that of the Sun. Assuming that Sirius and the Sun have the same intrinsic brightness, Huygens deduced that Sirius was 2.8×10^4 further from the Earth than the Sun. Since one Astronomical Unit is about 1.5×10^{13} cm, Huygens concluded, in effect, that Sirius was 4.2×10^{17} cm, or about 0.45 light years distant. The correct figure is almost a factor 20 larger. The discrepancy is due primarily to Huygens' adoption of the assumption of mediocrity: Sirius, a dwarf star of spectral type A1, has an intrinsic luminosity some 60 times greater than the Sun. △

Thus, although such estimates have only probabilistic character, the assumption of mediocrity will, in many cases, give a valid rough answer, when a detailed scientific justification lies beyond the present capabilities of science.

▽ Nevertheless, the application of this method to areas where we have little knowledge is essentially an act of faith. For example, one exercise which we shall later carry through is to estimate the likelihoods of the origin of life in a suitable planetary system, the origin of intelligence, the origin of technical civilization, etc. Such estimates are, either implicitly or explicitly, based upon terrestrial experience. But it is dangerous to extrapolate from one example. This is why, for example, the discovery of life on one other planet—e.g., Mars—can, in the words of the American physicist Philip Morrison, of the Massachusetts Institute of Technology, "transform the origin of life from a miracle to a statistic."

▽ For the origin of intelligence and of technical civilizations, finding another example may be even more difficult than the detection and characterization of life on Mars. We must recognize the possibility that even with as many as 10^{22} planets in the accessible universe, the probability that one of them possesses a technical civilization may be 10^{-22} or less. We may feel that the probability must be higher, but we do not *know*. Indeed, the determination of such probabilities is one of the major motivations of a search for intelligent extraterrestrial life.

▽ Another question of some relevance to our own time, and one whose interest is not restricted to the scientist alone, is this: Do technical civilizations tend to destroy themselves shortly after they become capable of interstellar radio communication? The establishment of interstellar radio contact may permit an estimate of such probabilities.

▽ As an example of the difficulties inherent in establishing a priori probabilities, consider the question of the origin of intelligent life on Earth. We have emphasized that evolution is opportunistic, not foresighted. We have five digits on each hand and foot not, we think, because there is any intrinsic advantage to five, versus four or six, but because we have evolved from a Devonian predecessor, an amphibian with five bones homologous to our present phlanges. This example is trivial for the question of the origin of intelligence; but suppose that we had some

evolutionary patrimony which was not irrelevant, but rather *detrimental* to the development of intelligence—some characteristic so deep-seated, so intimately woven in the fabric of life that the development of intelligence would be unlikely. Surely, all conceivable adaptations are not achieved, even when they may have high selective value. There must be tractable evolutionary pathways from here to there. For example, there are no organisms on the Earth which have developed tractor treads for locomotion, despite the usefulness of tractor treads in some environments. The improbability of achieving such an adaptation through slow evolutionary process must outweigh its potential adaptive advantage.

▽ So we may ask whether the development of human intelligence was a fortuitous occurrence. Intelligence itself arose early, and the development of tool-using capabilities evolved with birds and non-human primates. But the ecological circumstances surrounding the evolution of contemporary human intelligence are essentially unknown. Some anthropologists believe that human communities developed in response to the inclemency of late Pliocene and Pleistocene times, perhaps because of the recession of the forests in which pre-human communities had lived, or perhaps because the cold new climate placed a premium on new habits of dress, food, and habitat. But if there had been no Pleistocene ice ages, would intelligence have developed on Earth?

▽ Some scientists have been especially impressed by the number of individually unlikely events which are together responsible for the development of men and human intelligence. They have emphasized that even if the Earth were started out again from scratch, and only random factors allowed to operate, the development of anything like a human being would be highly unlikely. But others have been impressed with the high selective value of intelligence. Provided that we do not use our intelligence to destroy ourselves, it and the civilization which now accompanies it are among the most significant developments in the history of life on Earth. We have occupied all habitats, tamed or destroyed all competitors and predators, and some of us are about to leave these Earthly confines for other places. Even though the development of humans—or their rough extraterrestrial anatomical equivalents, humanoids—is unlikely, might not the development of their intellectual equivalents be a pervasive evolutionary event?

▽ The development of intelligence and technical civilization has occurred about midway during the main sequence residence time of our Sun. If we were to extrapolate from one example, using the assumption of mediocrity, we would conclude that all planets on which life has flourished for several billions of years have a high probability of the development of intelligence and technical civilization. But this is at best a plausibility argument; we do not know the detailed factors involved in the development of intelligence and technical civilization. △

It would seem, then, that this book is concerned with an unsolved—if not unsolvable—problem. Is it in fact possible to call a book dealing with intelligent life in the universe "scientific"? ▽ We are deeply convinced that the problem can be approached responsibly only if the assumptions involved are stated explicitly, and if the most efficient use of the scientific method is made. Even then, we shall not

come to many final answers, but the formulation of the problems has itself significance and excitement.

▽ One conceivable approach is to assume that civilizations in various states of historical development exist throughout our Galaxy, and then to see what observational consequences this assumption implies. Humanity is relatively young; our civilization is in its infancy. Hominids have inhabited the Earth for about 0.1% of its history; our civilization has so far endured only for one-millionth the lifetime of the Earth; and technical civilization, in the sense of the capability for interstellar radio communication, has been present for about one-billionth of geological time. It is then immediately obvious that if there are civilizations on planets of other stars, they should, in general, be much more highly developed than our own. Whether this development includes social, scientific, artistic, or technical aspects, or other aspects which we cannot even imagine, is difficult to foretell. But establishing contact with an extraterrestrial civilization evokes, in exaggerated form, some of the same problems as would face the crew of an Algonquin war canoe, miraculously transported to contemporary Upper New York Bay. △ It would seem an almost impossible task to forecast the development of society for thousands or more years into our future. Historians tend to avoid such problems; ▽ they have difficulty enough understanding the past, without foretelling the future. △ Nevertheless, we believe that some regularities and general tendencies about the evolution of civilizations can be stated.

Judging from our only example, there is an important peculiarity of advanced forms of intelligent life: they strive for active control of the universe. Man is already venturing beyond the Earth and taking his first timid steps toward remaking the solar system. Possible influences of intelligent life in the Galaxy, but on a much grander scale, will be discussed in Chapter 34. For billions of years, the Earth has had only one satellite; ▽ now, there are thousands. △ The artificial satellites are, of course, small; yet they are larger than the tiny satellites of Saturn which form its remarkable rings. Our civilization could establish an artificial ring about this planet, an engineering feat which seems well within the reach of contemporary technology. There appears to be no use for such a ring at present; but if there were, we could create one within a few decades. ▽ In fact, the orbiting of a belt of small needles by the United States Air Force some years ago, in an operation called "Project Westford," demonstrated the feasibility of such an enterprise. △

In Chapter 18, we mentioned that because of the activities of man, the brightness temperature of the Earth in the meter wavelength range has increased a millionfold during the last two or three decades. Intelligent life has made our small planet the second most powerful radio source in the solar system. It is entirely possible that in future decades our planet will—at least at some times and some frequencies—become as powerful a source of radio radiation as the sun.

We shall show in Chapter 28 that an analogous situation can, in principle, be created at optical frequencies. The development of quantum generators of optical radiation—the lasers—opens the possibility of sending narrow beams of almost monochromatic light over vast interstellar distances. At a given frequency and

direction, the intensity of optical emission from the Earth may greatly exceed that of the Sun.

These are only a few examples of those cosmic manifestations of intelligent life that can be predicted by modest extrapolations of existing technology. But what will come after? The specific course of the active influence of intelligent life on the universe is not easy to forecast; but the trends in development are entirely obvious. ▽ If there are many technical civilizations in the universe, only a small fraction need have the same urge to expand and control that our species possesses, for there to be a wholesale remaking of the universe.

▽ When we attempt to make a prognosis of the most general aspects of intelligent society in the distant future—say, millions of years hence—modest extrapolations from existing technology do not suffice. We might restrict ourselves only to what is physically possible, even though it may be technically far beyond our present imaginings. But for million-year timescales, even this procedure is hopelessly modest. New scientific principles will of course be discovered, and it is impossible for us to forecast their nature or even their direction. Perhaps a sign of a very advanced civilization will be the abandonment of the urge to expand and control. Perhaps a sign of a truly advanced civilization will be the voluntary abandonment of technical pursuits for activities of another kind. In the Russian edition of the present work, Shklovskii expresses his hope that "Marxist philosophers will become interested in the problem of large-scale predictions of the future of humanity," and apologizes for setting forth his own ideas, because he is not a specialist in scientific augury. But there are no specialists in this subject; there may not even be such a subject. △ But at least, the mistakes made here will encourage more fruitful discussions in the future.

In the remainder of Part III, we will touch upon a large number of problems. First, we shall consider an analysis of some modes of possible reconstruction of the cosmos by intelligent beings. As examples (perhaps not entirely hypothetical), we shall consider the problems of the moons of Mars and of the Dyson shell hypothesis. Then we consider a wide range of possible modes by which contact may be established with intelligent extraterrestrial life. Some of the chapters of Part III contain mathematical calculations which may present some difficulty for the general reader. However, they are required to substantiate some of the conclusions derived. ▽ Since this phase of the subject is so new, it is not possible to refer to standard references. We have tried to construct the analysis in such a way that the detailed calculations can be omitted without substantially impairing comprehension of the major points. To this end, unessential mathematical details are given in small print. △ The material in these chapters is new, and to a certain degree, not previously published.

26

Are the moons of Mars artificial satellites?

. . . Nor has [Mars] any Moon to wait upon him, and in that . . . he must be acknowledged inferiour to the Earth.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

Round the decay
Of that colossal wreck, boundless and bare
The lone and level sands stretch far away.

Percy Bysshe Shelley, *Ozymandias*

The two moons of Mars are among the most intriguing objects in the solar system. Their existence was first suggested by the English satirist Jonathan Swift. In his work *Gulliver's Travels*, published in 1726, more than 150 years before the Martian satellites were discovered, Swift has Gulliver describe the scientific activities of the kingdom of Laputa, an inhabited island in the sky. In Lemuel Gulliver's account, the following curious sentence appears:

> They [the Laputan astronomers] have likewise discovered two lesser stars, or 'satellites,' which revolve about Mars, whereof the innermost is distant from the centre of the primary planet exactly three of his diameters, and the outermost five; the former revolves in the space of ten hours, and the latter in twenty-one and an half; so that the squares of their periodical times are very near in the same proportion with the cubes of their distance from the centre of Mars, which evidently shows them to be governed by the same law of gravitation, that influences the other heavenly bodies . . .

▽ This Laputan discovery was echoed by Voltaire in his interplanetary romance, *Micromegas,* published in 1752. △

Swift's characterization of the number of Martian moons, their periods of revolution, and their distances from the planet are uncannily close to the truth. There has been some speculation on how Swift arrived at his prognosis. It seems probable that he was not operating on fantasy alone; in a sense, his description was based on the prevalent astronomical ideas of the day. It was known, of course, that the Earth had one moon, and believed that Jupiter had four. (Only the four Galilean satellites had been detected in Swift's time; now we know that the giant planet has 12 moons, many of which can be observed only through the largest telescopes.) Since Mars is situated between Earth and Jupiter, the assumption that there was a geometric progression of the number of satellites for the more distant planets may have led Swift to the deduction that Mars had two moons. We should recall that Pythagorean ideas about the harmony of numbers were widely accepted in those days.

Swift probably believed that the moons were small in size because non-Laputan astronomers had not yet detected them. He may have reasoned that the moons were relatively close to Mars, because even very small satellites could be detected if they were sufficiently distant from their primary. Close proximity, however, would have hidden them in the scattered light of the planet. ▽ The proportionality between the period of revolution of these satellites about Mars and the 3/2 power of their distance from the center of Mars is simply an expression of Kepler's Third Law, which, together with its derivation from the theory of gravitation by Newton, was well known in Swift's time. △

The moons of Mars were actually discovered in 1877 by the American astronomer Asaph Hall, shortly after the completion of a large refracting telescope at the United States Naval Observatory. Since 1877, they have been observed repeatedly, mostly when Mars is at opposition. The outermost of the two moons is called Deimos, and is approximately 23,000 km from the center of the planet. The inner moon is Phobos, some 9,300 km distant. ▽ Phobos and Deimos are Greek for, respectively, "fear" and "panic," the chariot horses of the god of war. △

The period of revolution of Deimos about Mars is 30 hours, 18 minutes; Phobos revolves every 7 hours, 39 minutes. Thus, Swift predicted the period of revolution of Phobos to within 25%, and of Deimos to within about 40% of the true values—a rather remarkable guess. The period of rotation of Mars about its axis is 24 hours, 37 minutes, and 23 seconds. Thus, if we neglect the artificial satellites of Earth, Phobos is the only known moon in the solar system with a period of revolution about its planet which is less than the period of rotation of the planet itself. For this reason, future explorers of the planet Mars will be able to see Phobos rise in the western Martian sky and set in the east; its apparent period of revolution is 11 hours.

▽ In the early decades of this century, the English author Edgar Rice Burroughs published a series of Martian romances based on the adventures of one John Carter, a Virginian miraculously transported to Barsoom (as Mars was called by its inhabitants). The Martian scenarios were based in large part on Percival Lowell's views of Mars, and served to fix Lowell's ideas of the Martian environment in the minds of an entire generation.

▽ Among other preconceptions which are, even today, sometimes difficult to shake, Burroughs populated Mars with an intelligent race of indigenous human beings, breathing an oxygen atmosphere, living in the dry bottoms of oceans long ago evaporated, and drinking water pumped by Lowell's elaborate planetary canal system. One of Burroughs' phrases, "beneath the hurtling moons of Barsoom," has given the impression that Deimos and Phobos, viewed by an observer on Mars, move rapidly and perceptibly through the nighttime sky. In fact, it would take some 5 1/2 hours for Phobos to rise in the West, follow a meridian through the zenith, and set in the East. Those who have watched artificial Earth satellites (whose comparable periods are about 50 minutes) can testify that such motion would be almost imperceptible. △

If the period of revolution of a satellite were exactly equal to the planet's period of rotation, the moon would appear fixed in the sky from one hemisphere of the planet, and never seen from the other. (▽ Such synchronous orbits are the basis of the SYNCOM system of communication satellites under development in the United States; here, three satellites launched into synchronous orbits could, in principle, revolve just as fast as the Earth rotates, so that every locale on the Earth's surface would be in direct line-of-sight communication with one of the satellites, and each satellite would always be in direct view of the other two.) △ The period of revolution of Deimos about Mars is fairly close to the period of planetary rotation. The length of the Deimos "month" from new Deimos to new Deimos is

about 132 hours. ▽ The word "month" is, of course, derived from the word "moon." It might be tempting to invent new names for the "months" of satellites of other planets; but since the word "moon" can also be used in a generic sense, we will refrain from wrestling with such splendid barbarisms as "donth" and "phonth." △ The orbits of both Martian satellites lie in the equatorial plane of the planet, and are very close to being circular. The eccentricity of the orbit of Phobos is 0.017, and of Deimos, 0.003. ▽ A perfect circle has an eccentricity of zero; the larger the eccentricity, the more elongated the orbit. △ The angles of inclination—that is, the angles between the satellites' orbital planes and the Martian equatorial plane—are 1′75″ and about 1′, respectively.

During an average opposition of Mars, the apparent "stellar" magnitudes of Phobos and Deimos are +11.5 and +13, respectively. Thus, if it were not for their proximity to the planet, they could easily be seen from Earth with a telescope of moderate size. ▽ This difficulty in detection is the same one which we encountered in Chapter 11, when we considered the prospects for photographic detection of planets of nearby stars. △

At present, it is impossible to measure the angular dimensions of the two moons—and therefore, their true sizes—by direct observation from Earth, because their diameters are so small. But there is an indirect method which enables us to obtain approximate values of their dimensions. ▽ We know to sufficient accuracy how far the moons are from us, and from the Sun. We can measure their brightness and derive their apparent magnitudes, as given above. Why are they as bright as they are? Because of their reflectivity, or albedo, and because of their size. The higher the albedo and the larger the size, the brighter they should be. Thus, if we make an assumption about the albedo of the satellites, we can deduce their size. △ If we assume that the albedo of the satellites is about the same as that of Mars (some 15%, in the visible), then it can be calculated that Phobos has a diameter of about 16 km, and Deimos, of perhaps 8 km. ▽ If the satellites have albedos comparable to the moon's or Mercury's (values about half that of Mars), then the diameters will be somewhat larger. △ Phobos and Deimos are, then, the smallest known moons in the solar system. We should note, however, that if satellites of similar dimensions exist very close to Jupiter, Saturn, or the other Jovian planets, they would not be detectable at the present time.

To an explorer on Mars, Phobos would be a brilliant celestial object with a discernible disk. Its angular diameter would approach 10 minutes of arc—that is, one-third the size of the lunar disk seen from Earth—and its luminosity would be about 4% that of our own moon, quite sufficient to cast shadows during the Martian night. Deimos, more distant from its primary, would look very much like a bright star, perhaps ten times brighter than Venus appears from the Earth.

Lowell noted that neither of these moons has the characteristic red color of Mars itself, an observation confirmed by later investigators. ▽ We have seen in Chapter 19 that the red color of Mars is probably due to large amounts of the mineral limonite, which also accounts for the planetary albedo. Thus, it seems clear that the chemical composition of the surfaces of Deimos and Phobos differs

from that of Mars. The albedo of the satellites therefore need not necessarily be the same as that of Mars. This underlines the uncertainties involved in the calculation of the diameters of Phobos and Deimos. If the surfaces of these moons initially had compositions similar to that of Mars, subsequent differentiation would have occurred. For example, the Martian atmosphere is thick enough to absorb protons incident upon it from the solar wind. The satellites have no such atmosphere, and the incident protons will strike their surfaces and discolor them as they have discolored our moon [see Chapter 21]. There are also other possible causes for the composition difference, as we shall see. △

In 1945, the American astronomer B. P. Sharpless, ▽ working at the same Naval Observatory at which Hall discovered the satellites of Mars, △ detected a remarkable peculiarity in the motion of Phobos. Comparing a series of old observations made by Hermann Struve with more recent observations, he noted that the orbital velocity of Phobos is increasing. The magnitude of this acceleration is small, but, from his data, apparently real [see Figure 26–1]. ▽ In the case of Deimos, there was less clear evidence for such an acceleration. If we call the angular velocity of Phobos around Mars ω, and $\Delta\omega$ the change in this velocity in some time interval, then $\Delta\omega/\omega$ will be the relative change in angular velocity in the same time period. △ According to Sharpless, this relative change was

$$\Delta\omega/\omega = +(7.98 \pm 0.73) \times 10^{-12}.$$

▽ The first plus sign denotes an acceleration, as opposed to a deceleration; the sign $\pm$ indicates Sharpless' estimate of the uncertainty in the measurements.

▽ Since the period of revolution of Phobos about Mars is 7 hours, 39 minutes,

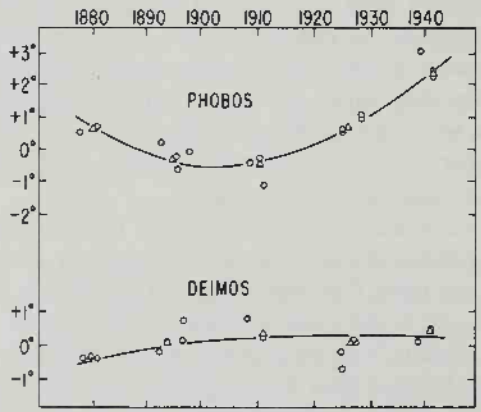


FIGURE 26–1. *Evidence on secular accelerations of Phobos and Deimos gathered by B. P. Sharpless. A satellite with no secular acceleration would show a horizontal line. Within the errors of observation the data points for Deimos are very close to a horizontal line. On the other hand, it is difficult to draw a horizontal line through the points for Phobos, and it is on this basis that Sharpless deduced a secular acceleration of Phobos.* (*Courtesy of the* Astronomical Journal.)

or about 28,000 seconds, Phobos travels at roughly $(28{,}000)^{-1} \simeq 2.1 \times 10^{-5}$ revolutions per second. Due to its acceleration, Phobos moves $8 \times 10^{-12} \times 2.1 \times 10^{-5} = 1.7 \times 10^{-16}$ of a revolution faster each second. Since the distance of Phobos from the center of Mars is 9300 km, the acceleration causes a *decrease* in the radius of the orbit of $9.3 \times 10^{8} \times 1.7 \times 10^{-16} = 1.6 \times 10^{-7}$ cm each second. In 60 years the orbital radius has shortened some three meters, an extremely small amount; but the consequent changes in the period of Phobos have been detectable from Earth. At this rate, Phobos will impact the surface of Mars, now some 5,900 km below, in 5.9×10^{8} cm/1.6×10^{-7} cm sec$^{-1} \simeq 3.7 \times 10^{15}$ sec, or about one hundred million years. △

In celestial mechanics, a continuous, non-periodic change in one of the components, or characteristics, of an orbit is called a "secular change." ▽ This use of the words "secular change" is similar to that introduced in Chapter 20, when we spoke of secular changes in the configuration of dark areas on the surface of Mars. A periodic, and hence more easily predictable, change is called a "canonical" change. The words hark back to the Middle Ages, when the church had calendrical (hence, periodic) responsibilities. Those occurrences not conforming to the ecclesiastic calendar were, by default, secular. △ Thus, Sharpless detected a secular acceleration in the motion of Phobos about Mars. No unequivocal evidence was adduced for a secular acceleration of Deimos.

Since the Martian moons are very difficult to observe, even with the best equipment, it is possible that Sharpless' compilation of the data may contain substantial errors. However, the large magnitude of the secular acceleration, $\Delta\omega/\omega$, suggests that the effect is real. Suppose a hypothetical observer on Mars were to predict the position of Phobos in the sky over a 50-year period, neglecting the effect of secular acceleration. At the end of this time, the actual position of Phobos would be 2° from the predicted position—in celestial mechanics, a very large deviation.

Let us assume for the moment that the secular acceleration is a real effect. We will suggest several conceivable causes, ▽ and then examine their consequences △:

1. Atmospheric drag. If Phobos were traversing a sufficiently dense gas during its motion about Mars, the gas would "drag" the satellite, causing its orbit to shrink, and resulting in a net acceleration. This effect strongly influences the motion of artificial satellites of Earth, and is a main factor in determining their lifetime in orbit.
2. Tidal friction, an effect which probably played an important role in the evolution of the Earth-Moon system.
3. Electromagnetic braking of the motion of Phobos by the magnetic field of Mars.
4. The effects of radiation pressure.
5. Classical celestial mechanical perturbations.

Let us now consider each of these possibilities in turn.

1. *Atmospheric drag.* In 1954, the astronomers Frank J. Kerr and Fred L. Whipple, working in the United States, concluded that a gaseous, resisting medium could not explain the secular acceleration of Phobos. They computed the density of the resisting medium required to produce the observed effect. By a variety of assumptions, they arrived at figures for the density of the resisting medium ranging between 3×10^{-16} and 5×10^{-16} gm cm^{-3}. ▽ A more recent estimate of the required gas density in the vicinity of Phobos by the Austrian-American scientist Gerhard Schilling of the Rand Corporation puts it at $\rho \simeq 5 \times 10^{-16}\,\delta$ gm cm^{-3}, where δ is the bulk density of Phobos itself. Thus, if Phobos were made of ordinary terrestrial surface rocks, or the material which comprises the moon, δ would equal 3.3 gm cm^{-3}, and ρ would then be some 2×10^{-15} gm cm^{-3}. If Phobos were made of ice, δ would be 1 gm cm^{-3}, and ρ would be 5×10^{-16} gm cm^{-3}. Thus, the required density of the resisting medium must be approximately 10^{-15} gm cm^{-3}, if Phobos is a solid object composed of ordinary materials, as, in a preliminary analysis, we would certainly expect it to be. △

Kerr and Whipple then assumed that the resisting force was due to drag of the *interplanetary* gas and dust. The interplanetary medium is about equally dense at the orbits of Deimos and Phobos. Thus, if the secular acceleration of Phobos is caused by a resisting interplanetary medium, Deimos should have a similar secular acceleration. Since this was not the case, Kerr and Whipple concluded that the existence of a damping medium could not explain the secular acceleration of Phobos.

However, if we assume that the opposing medium is the Martian atmosphere, then at a distance of some 23,000 km from the center of the planet, at the orbit of Deimos, the atmospheric density would be much less than in the vicinity of Phobos. Therefore, we must estimate the density of the Martian atmosphere at the various altitudes before we can exclude the resisting medium explanation.

▽ Before the successful completion of the Mariner IV mission to Mars, a calculation of the expected densities of the Martian atmosphere at the distances of Phobos and Demos were beset with many difficulties. The uncertainties included the number density and altitude of the base of the Martian exosphere, the exosphere temperature and its mean molecular weight. As a result of the occultation experiment on Mariner IV (see Chapter 20) some of these parameters are now known to better accuracy. The base of the exosphere—the region from which gravitational escape can occur—now appears to be at an altitude of less than 200 km, as compared with the approximately 1,500 km that had been previously computed. The exosphere temperature appears to be a few hundred °K, much colder than had previously been thought, and the mean molecular weight at this level appears to be close to 44, the molecular weight of carbon dioxide. However, at greater altitudes, the major constituent should become atomic oxygen, and at still greater altitudes, atomic hydrogen, arising from the photo-dissociation of water vapor. With these figures, the number density at the base of the Martian exosphere is approximately 2×10^{9} cm^{-3}, a factor of 200 larger than the best previous theoreti-

cal estimates. At the altitude of Phobos, the number density will, of course, be much less.

▽ Let n_P be the density of the Martian exosphere at the distance of Phobos, and n_b be the density at the base of the exosphere. If the distances of Phobos and the base of the exosphere from the surface of Mars are, respectively z_P and z_b, then

$$n_P = n_b \exp\{-R^2/H\,[1/(R + z_b) - 1/(R + z_P)]\}$$

Here, R is the radius of Mars, and $H = kT/mg$ is the scale height of the atmosphere, a measure of how rapidly the atmospheric density declines with altitude. k is Boltzmann's constant, T is the absolute temperature, m is the mass of the atmospheric constituent in question, and g is the acceleration due to gravity at the surface of Mars. The equation takes into account the variation of g with altitude; the dependence is as the inverse square of the distance from the center of the planet. This equation does not hold exactly because we are considering the exosphere, where collisions are infrequent. However, it should apply to first approximation.

▽ Putting numbers into this equation, and noting that the Martian exosphere will be composed over most of its extent of atomic hydrogen, we find that the number density at the distance of Phobos, some 6,000 km, is about $2 \times 10^5 f_H$, where f_H is the fractional abundance of atomic hydrogen at the base of the exosphere. We do not know the value of f_H, but we suspect it to be very small by analogy with the small amount of hydrogen (about 0.3%) at the base of the terrestrial exosphere; in particular, the source of hydrogen in the Martian atmosphere must be the photo-dissociation of water, and there is about 1/1,000 as much in the Martian as in the terrestrial atmosphere. It would be very surprising if f_H were larger than, say, 10^{-3}. We then find that an upper limit to n_P is about 2×10^3 cm^{-3}. For comparison, we found that the required mass density of the exosphere to explain the secular acceleration of Phobos by atmospheric drag was $5 \times 10^{-16}\ \delta$ gm cm^{-3}. For a hydrogen exosphere, this is the same as $3 \times 10^8\ \delta$ cm^{-3}. For ordinary values of δ, we see that the atmosphere is more than 100,000 times too diffuse at the altitude of Phobos for this explanation to be viable. Calculations based on the numbers thought to apply prior to Mariner IV gave a smaller discrepancy of about a factor of 1,000.

▽ Thus, we have reached the conclusion that the density of the Martian exosphere in the vicinity of the satellite Phobos is probably 100,000 times too diffuse to explain its secular acceleration, if Phobos has a density characteristic of ordinary solid materials.

▽ In the discussion immediately following, we analyze the alternative explanations of the secular acceleration of Phobos. The over-all conclusion is that explanations 2 through 5 cannot account for the secular acceleration of Phobos. The reader who does not wish to trouble himself with these technical details may take up the discussion again on page 373. △

2. *Tidal friction.* Another possible explanation of the secular acceleration of Phobos is tidal friction, a problem investigated by the British geophysicist Sir Harold Jeffreys of Cambridge University. Since there are no large bodies of liquid on the surface of Mars, tidal friction can arise only in the solid body of the planet. Jeffreys assumed that

the viscoelastic properties of Mars are the same as for the solid body of the Earth. His calculations indicate that tidal friction could account for only 10^{-4} of the observed secular acceleration of Phobos.

However, the question of the viscoelastic properties of the solid material of a planet is quite controversial. Recently, the Soviet geophysicist N. N. Pariiskii ▽ and, independently, the American geophysicist G. J. F. MacDonald of the University of California at Los Angeles, △ concluded that body tides in the Earth (and, by analogy, in Mars) are significantly greater than Jeffreys anticipated. According to the calculations of Pariiskii, body tides are necessary in order to explain the secular motion of our moon and might also account for the secular acceleration of Phobos.

However, we now have evidence, from entirely different considerations, that the secular acceleration of Phobos cannot be explained by tidal friction. According to Jeffreys, the theoretical value of the secular acceleration of a satellite by viscoelastic tidal friction in the solid body of a planet can be represented by the expression

$$\frac{d\zeta}{dt} = \frac{9}{4}\frac{m\omega}{M}\left(\frac{R}{r}\right)^5 \psi \sin 2\theta,$$

where m is the mass of the moon, M is the mass of the planet, R is the radius of the planet, r is the radius of the orbit of the satellite at any time, ω is the average angular velocity of the moon at any time, and θ is the angle of lag of the tidal bulge. The quantity ψ depends only on the viscoelastic properties of the planet. Furthermore, according to Jeffreys,

$$\sin 2\theta = \frac{\Phi}{2(\Omega - \omega)},$$

where Ω is angular velocity of rotation of the planet, and Φ depends only on the viscoelastic properties of the planet. From the above equations and Kepler's Third Law, we can determine the time it takes for the radius of the circular orbit of the moon to pass from r to r_0 because of tidal effects:

$$t(r) = t_0(1 - \omega/\Omega)^{-1}\{[(r/r_0)^{13/2} - 1] - (13\omega/10\Omega)[(r/r_0)^5 - 1]\};$$

$$t_0 = 3\omega_0/13(d\omega/dt)_0.$$

Here, ω_0 and $(d\omega/dt)_0$ are the current values of the average angular velocity of the moon and its variation with time, ▽ and $r_0 \simeq 2.8\ R$ is the current value of r. △ We note that in the case of Phobos, $r < 2.17\ r_0$ (the distance from the center of Mars for which $\omega_0 = \Omega$); otherwise, this moon would not approach Mars, but would recede from it. Having carried out these calculations, we obtain

$$t(r < 2.15\ r_0) < 5 \times 10^8 \text{ yrs.}$$

But 500 million years as an upper limit on the time which has passed since the formation of Phobos is an inadmissibly small value. Five hundred million years ago, conditions on Mars (which has itself existed for 4 to 5 billion years) were not significantly different from contemporary conditions. Accordingly, it is inconceivable that in such a recent

epoch a moon could have been formed having an almost circular orbit and lying practically in the plane of the equator of the planet.

There is another possibility: let us suppose that Phobos was formed at a distance of $2.15\, r_0 < r < 2.17\, r_0$, and that its period of revolution was almost exactly equal to the period of rotation of Mars. We must also suppose that Deimos was formed at the same critical distance; at this distance the tidal forces of Mars would not noticeably influence the motion of the moons. One could then assume further that for various reasons the moons moved out of their almost stable orbits—that Phobos was swept toward the planet and Deimos was swept in the opposite direction. For small displacements, the tidal forces would be very small, and a large amount of time could pass before Phobos' r became smaller than, for example, $2.1\, r_0$.

However, this is a highly improbable explanation for the origin of the moons. Why would the moons necessarily be formed at the precise distance where the condition $\omega = \Omega$ is fulfilled? All other satellites revolving about planets in the solar system are found at relatively great distances from the planets. Furthermore, it is difficult to understand why Deimos, on which tidal forces have practically no influence (because of its small mass), should move away from the planet, out of the orbit at which $\omega = \Omega$ (where, according to our supposition, it was formed).

One must bear in mind that over several billions of years, the period of rotation of Mars could have changed substantially. Such a fact would invalidate the hypothesis that satellites were formed at a distance simply determined by the contemporary value of the period of rotation of Mars. However, these considerations strongly suggest that the observed secular acceleration of Phobos cannot be explained by tidal friction in the solid body of Mars.

3. *Magnetic braking.* In principle, electromagnetic effects might lead to the observed secular acceleration of Phobos. Let us assume a satellite which is a good conductor of electricity. Also let us assume that Mars has a magnetic field. Then, the movement of the moon in the magnetic field gives rise to an electric field $E' = [v \times H]/c$. This field would polarize the satellite; that is, charges of different signs would migrate to opposite sides. The electric field of these charges in the space surrounding the satellite would be of the same order as E', so that the electric potential with respect to the ions it meets would be $x = Es$, where s is the characteristic dimension of the satellite. The value of x in volts is $300\, E's = 300\, vHs/c$. Assuming that $v = 2 \times 10^5$ cm sec^{-1}, $H = 10^{-3}$ gauss, and $s = 10^6$ cm, we find that $x = 2$ volts. Since this energy is comparable to the thermal energy of the interplanetary gas, it follows that positive ions would settle on the negatively charged surface of the satellite and all electrons would be repulsed. On the reverse side, which is positively charged, the ions would be repulsed, and some of the electrons would settle. Then the current, I, would be equal to the flow of positive ions through the hemisphere. Since the velocity of the satellite is close to the velocity of the ions, v_i, we have: $I = n_i v_i e A$, where $A \propto s^2$ is the cross-section of the satellite, e is the charge on the electron, and n_i is the number density of ions. The damping force is $f = IHs/c \simeq n_i e v_i s A H/c$, and the magnitude of the acceleration is

$$\left(\frac{dv}{dt}\right) = \frac{n_i e v_i s A H}{mc} \simeq \frac{n_i v_i e H}{c\delta},$$

where m is the mass of the satellite and δ is its density. The time of damping will be

$$t \simeq \frac{v}{dv/dt} \simeq \frac{vc\delta}{n_i v_i eH} \simeq \frac{2 \times 10^{15}\,\delta}{n_i} \text{ years,}$$

where we assume that at a distance of 6,000 km from the surface of Mars, $H = 10^{-3}$ gauss (probably an overestimate). Since $n_i < 10^3$ cm^{-3} while $\delta \simeq 2.5$ gm cm^{-3}, then $t > 5 \times 10^{10}$ years. ▽ Thus, the timescale for magnetic damping of a conducting satellite would be longer than the age of the solar system. △

If the conductivity of the satellite is sufficiently small, then the current across it will be determined by the electrical conductivity, λ, and not by the flow of interplanetary charged particles. In this case,

$$I = \lambda E' A \simeq \lambda v H A/c$$
$$f = IHs/c \simeq \lambda^2 H^2 As/c^2$$
$$t' \simeq \frac{v}{dv/dt} \simeq \frac{c^2\delta}{\lambda H^2}$$

Down to $\lambda \leqq 10^9$ sec^{-1} (which is significantly greater than the conductivity of rocks), $t' > t$. For $\lambda > 10^{10}$ sec^{-1}, taking into account the polarization, the time of the electromagnetic drag will be determined by t.

In summary, we must conclude that it would be impossible to explain the observed secular acceleration of Phobos by magnetic forces.

4. *Radiation pressure.* We could try to explain the secular acceleration by the Poynting-Robertson effect. Owing to the aberration of light, the force of light pressure on a moving body will have a component directed against the motion, which leads to a continuous drag on the body. This is known as the Poynting-Robertson effect. Due to this effect, dust with dimensions greater than 0.5μ which revolve in orbits about the sun will fall into the sun ▽ in a time shorter than the age of the solar system. △ If an ordinary particle has a dimension which is less than 0.6μ (but greater than 0.2μ), the force of the light pressure will exceed the force of gravitational attraction. Such particles will be ejected beyond the limits of the solar system (see Chapter 15).

However, I am convinced that this effect, which here depends both on direct solar radiation and on the light reflected from Mars, would give a secular acceleration some six to eight times less than the observed secular acceleration.

5. *Classical celestial mechanical perturbations.* Finally, we shall consider the possibility of a purely celestial mechanical explanation for the secular acceleration of Phobos. For example, the effect of the sun on Deimos, in theory, could lead to the appearance of long-period terms in the planetocentric longitude of Phobos. The perturbations of the motion of the satellites of Mars by the sun, and also their mutual perturbations, was recently investigated by the Soviet astronomer M. P. Kosachevskii. According to his calculations, the mutual perturbations are more significant than the solar perturbations; moreover, the motion of Deimos is far more strongly affected than the motion of Phobos. This is completely understandable, since Deimos is further from Mars than Phobos. The absolute magnitudes of the perturbations of the two moons, according to Kosachevskii's calculations, are very small.

Thus, all of the mechanisms we have discussed apparently cannot explain the secular acceleration of Phobos. Of course, we repeat, there is a slight possibility that Sharpless' observations are in error. However, at the present time this seems unlikely to me.

In 1959, I proposed a new and radical hypothesis concerning the motion of Phobos. ▽ Let us reconsider the discussion on page 369. There, we saw that for the secular acceleration to be explained by the action of a resisting medium, the density of this resisting medium had to be about $3 \times 10^6\ \delta$ cm^{-3}, where δ was the density of Phobos. We also saw that the expected Martian exospheric densities in the vicinity of Phobos are less than 2×10^3 cm^{-3} △ Thus, if the mean density of Phobos were about ▽ 10^{-5} gm cm^{-3}, or, with the older numbers, △ 10^{-3} gm cm^{-3}, then its secular acceleration could be explained by the resistance of the Martian exosphere.

But how can a natural satellite have such a low density? The material of which it is made must have a certain amount of rigidity, so that cohesive forces will be stronger than the gravitational tidal forces of Mars, which will tend to disrupt the satellite. Such rigidity would ordinarily exclude densities below about 0.1 gm cm^{-3}. Thus, only one possibility remains. Could Phobos be indeed rigid, on the *outside* —but hollow on the inside? A natural satellite cannot be a hollow object. Therefore, we are led to the possibility that Phobos—and possibly Deimos as well —may be artificial satellites of Mars.

▽ They would be artificial satellites on a scale surpassing the fondest dreams of contemporary rocket engineers. If the density of Phobos ranges between 10^{-3} and 10^{-5} gm cm^{-3}, then its mass must range from tens of millions to billions of tons although the solid outer shell might be no more than a foot thick. For comparison, the most massive artificial satellites hitherto launched from this planet are in the 10-ton range, and artificial satellites much beyond the 100-ton range do not seem to be feasible projects, at least for the next few decades. △ (If it turns out that the visual albedo of the Martian satellites is high—for example, 0.60 to 0.80—then their dimensions will be 2 or 3 times less than these calculations indicate, and their mass, 5 to 10 times less.)

The idea that the moons of Mars are artificial satellites may seem fantastic, at first glance. In my opinion, however, it merits serious consideration. A technical civilization substantially in advance of our own would certainly be capable of constructing and launching massive satellites. Since Mars does not have a large natural satellite such as our moon, the construction of large, artificial satellites would be of relatively greater importance to an advanced Martian civilization in its expansion into space. The launching of massive satellites from Mars would be a somewhat easier task than from Earth, because of the lower Martian gravity. ▽ Conceivably, the capture and hollowing of a small asteroid may be technically more feasible than the construction in orbit of an artificial satellite with material brought from the surface. △

It is quite possible that in several centuries the Earth will have satellites with dimensions of some kilometers. ▽ Manned orbiting laboratories in the 100-meter

size range have already been designed. △ Let us imagine that through the next several centuries, massive artificial Earth satellites are launched and maintained. ▽ Over a much longer timescale—say, 10^7 or 10^8 or 10^9 years—the evolution of human society and of life on Earth will not remain static. Perhaps mankind will destroy itself; or develop a society unconcerned with technological triumphs; perhaps a society will evolve which leaves the Earth altogether; or natural catastrophes, tectonic or climatological, may destroy civilization on Earth. We cannot reasonably assess these possibilities, but it does seem conceivable that the lifetime of our artificial satellites may exceed the lifetime of our civilization. △ These satellites would then remain as unique and striking monuments to a vanished species which had once flourished on the planet Earth.

Perhaps we are observing an analogous situation on Mars. According to the distinguished American cosmochemist Harold C. Urey, of the University of California, some billions years ago Mars may have possessed extensive oceans suitable for the origin of life, and perhaps even an oxygen atmosphere, ▽ although this latter is much less certain. △ Perhaps Phobos was launched into orbit in the heyday of a technical civilization on Mars, some hundreds of millions of years ago.

▽ The Soviet writer F. Zigel has made a more bizarre suggestion. Why, he wonders, were Phobos and Deimos not discovered by Herschel during the favorable Martian opposition of 1862, but found instead by Hall, with a smaller telescope, during the favorable opposition of 1877? The only explanation which occurs to Zigel is that the moons of Mars were launched into orbit between 1862 and 1877; it would then follow that an advanced technical civilization exists on Mars today. But the Naval Observatory telescope of 1877 was superior in several important respects to its predecessors. And the history of astronomy is full of similar incidents. After the discovery of Pluto in 1930 by Clyde Tombaugh at Lowell Observatory, the planet was found on photographic plates taken a decade earlier, with the larger telescope at other observatories. Uranus and Neptune were observed many times before their formal discovery, but their significance passed unnoticed. We have seen in Chapter 20 the unlikelihood of an extant civilization on Mars. If the moons of Mars are artificial—and we have at best only a plausibility argument to support this contention—they are much more likely mute testaments to an ancient Martian civilization than signs of a thriving contemporary society.

▽ While the birthdate of Phobos is difficult to estimate, some idea of the date of its death can be obtained more reliably. As we have seen, it is possible to compute the date on which Phobos will plunge through the lower Martian atmosphere and strike the surface; just as it is possible to compute the decay of an artificial Earth satellite. △ Careful calculations from the magnitude of the secular acceleration indicate that Phobos will impact Mars in some 10 or 20 million years. At that time, the planet itself will have existed for several billion years. This circumstance points out another difficulty in the assumption that Phobos has a natural origin, for it means that we are now observing Phobos during the last