The Open University

LIFE IN THE UNIVERSE:
THE SCIENCE OF ASTROBIOLOGY

Part 1 Edited by Iain Gilmour and Mark A. Sephton

FE PLANETARY SCIENCE AND THE SEARCH FOR LIFE PLANETARY SCIE

Background image: An all-sky image taken during the 1998 Leonid meteor shower, when the peak rate of meteors reached several hundred per hour.

Thumbnail images: (from left to right) False colour image of the North Pole showing ocean areas with abundant plankton and land areas with significant vegetation in green; part of Europa; DNA; surface of Mars.

This publication forms part of an Open University course S283 *Planetary Science and the Search for Life*. The complete list of texts which make up this course can be found on the back cover. Details of this and other Open University courses can be obtained from the Course Information and Advice Centre, PO Box 724, The Open University, Milton Keynes MK7 6ZS, United Kingdom: tel. +44 (0)1908 653231, e-mail general-enquiries@open.ac.uk

Alternatively, you may visit the Open University website at http://www.open.ac.uk where you can learn more about the wide range of courses and packs offered at all levels by The Open University.

To purchase a selection of Open University course materials visit the webshop at www.ouw.co.uk, or contact Open University Worldwide, Michael Young Building, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure.
tel. +44 (0)1908 858785; fax +44 (0)1908 858787; e-mail ouwenq@open.ac.uk

The Open University
Walton Hall, Milton Keynes
MK7 6AA

First published 2003

Edited, designed and typeset by The Open University.

Printed and bound in the United Kingdom by Bath Press, Bath.

ISBN 0 7492 5675 3

1.1

PLANETARY SCIENCE AND THE SEARCH FOR LIFE COURSE TEAM

Course Team Chair	Iain Gilmour
Academic Editors	Iain Gilmour, Neil McBride, Mark A. Sephton
Authors	Philip A. Bland Andrew Conway Iain Gilmour Barrie W. Jones Neil McBride Tony McDonnell Elaine A. Moore David A. Rothery Mark A. Sephton Mike Widdowson Ian Wright John Zarnecki
Course Manager	Jennie Neve
Course Team Assistant	Valerie Cliff
Editors	Peter Twomey Pamela Wardell
Software Designers	Fiona Thomson Will Rawes
Multimedia Producer	Kate Bradshaw
Web Developer	Hong Yu, Daniel Thorp
Graphic Designers	Debbie Crouch, Jenny Nockles
Graphic Artist	Sara Hack
Indexer	Jane Henley
Picture Researcher	Lydia Eaton
Book Assessors	Prof. F.W. Taylor, Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford. Dr Alan Penny, Rutherford Appleton Laboratory.
Course Assessor	Prof. David W. Hughes, Department of Physics and Astronomy, The University, Sheffield.

The book made use of material originally produced for the S281 *Astronomy and Planetary Science* Course Team.

CONTENTS

CHAPTER 1
ORIGIN OF LIFE

Shortly after the formation of the Earth some 4.6 Ga ago, our planet was a lifeless and inhospitable place. Yet if we examine rocks that were created about a billion years later, we can find evidence that by 3.5 Ga ago life had established a firm foothold on Earth. It is what happened in the intervening period that is the focus of this chapter. In other words, we will try to understand how life began.

During the course of this chapter you will examine how scientists have striven to define life and how unexpectedly difficult it is to perform this seemingly simple task. You will also examine the chemistry and function of entities that make up a living system. Then you will study the sites in the Universe and on the Earth where life's raw materials could have been formed before life had even begun. Finally, you will cover the mechanisms by which non-biological raw materials may have been combined into the first living organism.

1.1 What is life?

We will begin our attempts to define life in good company, as some of the world's most famous scientists have contributed to our current level of knowledge. Throughout history, questions have been raised about how and when life arose. Initially, many considered that life arose spontaneously and repeatedly on the Earth. These convictions were supported by what was thought to be the spontaneous generation of flies and maggots from rotting meat, lice from sweat, eels and fish from sea mud, and frogs and mice from moist soil. Occasionally, the idea of spontaneous generation was queried. For example in 1668, a Tuscan doctor called Francesco Redi (1627–1697) demonstrated that maggots were the larvae of flies and if the meat was kept in a sealed container, so that adult flies were excluded, no maggots appeared. However, when Dutch microscope maker Anthony van Leeuwenhoek (1632–1723) detected micro-organisms in 1676, spontaneous generation was the seemingly obvious explanation for such ubiquitous creatures. The matter was finally laid to rest in 1862 when, in an attempt to win a prize offered by the French Academy of Science, Louis Pasteur (1822–1895) (Figure 1.1) performed a convincing series of experiments. Pasteur showed that if a broth or solution was properly sterilized and excluded from contact with micro-organisms, it would remain sterile indefinitely.

Figure 1.1 Louis Pasteur (1822–1895), who disproved the idea that life could be generated spontaneously.

Pasteur had answered an important question by disproving spontaneous generation as the origin of life, but inevitably he had raised a new and more difficult question. If all life comes from existing life, where did the first life come from? Ironically, in demolishing the long-held idea that life arose spontaneously from inanimate matter, it was an inescapable and logical conclusion that the very first life may have done exactly that – arisen from non-living materials present in the Universe.

1.1.1 A definition of life

If we are to establish *when* and *how* life originated, we must first define exactly what life is.

Most biologists would identify two key features that indicate life:

- the capacity for self-replication, and
- the capacity to undergo Darwinian evolution.

Let us explore these criteria in slightly more detail. For an organism to self-replicate it must be able to produce copies of itself. For Darwinian evolution to occur, imperfections or mutations must occasionally arise during the copying process and these new mutations must be subjected to natural selection (Box 1.1). Nature favours particular characteristics under particular environmental conditions and those individuals best suited to the existing conditions are most likely to survive. For this process to bring about an evolutionary change any advantageous features brought about by mutation must be passed on to future generations.

BOX 1.1 NATURAL SELECTION AND DARWINIAN EVOLUTION

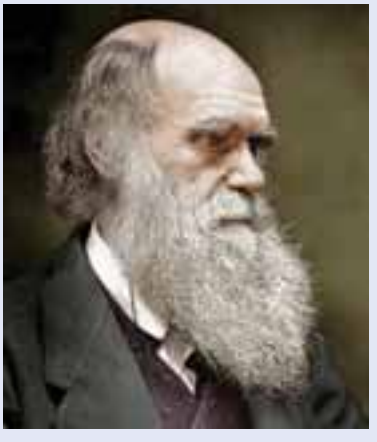


Figure 1.2 Charles Darwin (1809–1882), who established the theory of natural selection.

After graduating from Cambridge with a degree in theology at the age of 22, Charles Darwin (1809–1882) set sail as a naturalist on the British Navy's HMS *Beagle* mapping expedition (1831–1836). His voyage brought him to the Galapagos Islands in the Eastern Pacific Ocean. Within the Galapagos archipelago he found a wide variety of plants and animals. It was there that he began to formulate ideas about the process of evolution. Darwin recognized that any population consists of individuals that are all slightly different from one another. Those individuals having a variation that gives them an advantage in staying alive long enough to successfully reproduce are the ones that pass on their traits more frequently to the next generation. Subsequently, their traits become more common and the population evolves. Darwin called this 'descent with modification'.

The Galapagos finches provide an excellent example of this process. For instance, among the birds that ended up in arid environments, the ones with beaks that were better suited for eating cactus seeds got more food than those birds with beaks that were less suitable. As a result of the additional food, they were in better condition to mate. In a very real sense, nature selected the best-adapted varieties to survive and to reproduce. Darwin called this process 'natural selection'.

Darwin did not believe that the environment was producing the variation within the finch populations. He correctly thought that the variation already existed and that nature just selected the best-adapted varieties – in our example the birds with the most suitable beak shape for eating cactus seeds as against less-favourable shapes. Darwin described this process as the 'survival of the fittest'. It was not until 1859 when Darwin was 50 years old that he finally published his theory of evolution in a book entitled *The Origin of Species by Means of Natural Selection*. Today, the concept of natural selection and its influence on successive generations is called **Darwinian evolution**.

From our short list of two characteristics a working definition of *life* can be created. Gerald Joyce of the National Aeronautics and Space Administration (NASA) proposed the following definition: 'a self-sustaining chemical system capable of undergoing Darwinian evolution'. However, any definition of life is likely to fail in certain circumstances. For example, the mule is the offspring of a donkey and a horse. A mule cannot breed and therefore is incapable of taking part in the processes of self-replication and Darwinian evolution, yet few would deny that it is alive. But, for the majority of cases, our definition of life will be a satisfactory one.

For life to be self-sustaining and capable of Darwinian evolution both energy and materials must be extracted from the surrounding environment to allow growth and replication. Furthermore, some sort of living apparatus must be present to govern and facilitate the chemistry of life. In the following sections we will examine just what kinds of chemical system characterize life and what types of energy might have been available to primitive life on the early Earth.

1.1.2 Why carbon?

There is only one element that can form **molecules** of sufficient size to perform some of the functions necessary for life as we know it. This element is carbon.

Carbon can form chemical bonds with many other atoms, allowing a great deal of chemical versatility. Commonly, organic compounds also contain the elements hydrogen, oxygen, nitrogen, sulfur and phosphorus. A range of metals such as iron, magnesium and zinc also bond with carbon.

Carbon can form compounds that readily dissolve in liquid water and, as you will see shortly, water is essential for life on Earth. Elements fundamental to the development of living organisms must be able to interact readily with one another, and that occurs most readily in the presence of water.

You will often encounter the term 'organic' in relation to how life may have originated. 'Organic' usually signifies the influence of biology, but to the chemist the term simply covers the chemistry of compounds based on carbon.

All currently known life utilizes carbon-based organic compounds.

The relative abundance of the more common elements in the Universe indicates that the Universe is well stocked with the elements needed to construct these organic compounds (Table 1.1). The four most common elements that are utilized by life on Earth (hydrogen, oxygen, carbon and nitrogen) are the most abundant non-noble gas elements in the Universe. Sulfur and phosphorus (not listed, but the 15th most abundant element in the Universe) are also important for life on Earth.

QUESTION 1.1

Examine Table 1.1 and, extrapolating from our discussion of carbon above, explain why the amounts of noble gas elements in living organisms are so small.

Table 1.1 The ten most abundant elements in the Universe, Earth and life (expressed as atoms of the element per 100 000 total atoms).

Order	Universe		Whole Earth		Earth's crust		Earth's ocean		Humans	
1	H	92714	O	48880	O	60425	H	66200	H	60563
2	He	7185	Fe	18870	Si	20475	O	33100	O	25670
3	O	50	Si	14000	Al	6251	Cl	340	C	10680
4	Ne	20	Mg	12500	H	2882	Na	290	N	2440
5	N	15	S	11400	Na	21554	Mg	34	Ca	230
6	C	8	Ni	1400	Ca	1878	S	17	P	130
7	Si	2.3	Al	1300	Fe	1858	Ca	6	S	130
8	Mg	2.1	Na	640	Mg	1784	K	6	Na	75
9	Fe	1.4	Ca	460	K	1374	C	1.4	K	37
10	S	0.9	P	140	Ti	191	Si	–	Cl	33

Noble gases are highly unreactive and, until recently, were known as the inert gases. They include helium (He), neon (Ne) and argon, (Ar). These gases do not usually bind chemically with any other elements to form compounds.

1.1.3 Why water?

Liquid water also appears to be an essential requirement for life. Living systems need a medium in which molecules can dissolve and chemical reactions can take place. Water has been called the universal solvent because it performs this function so well. Few other solvents can match the abilities of water to facilitate life. Water exists as a liquid in a temperature range that is not too cold to sustain biochemical reactions and not too hot to stop many organic bonds from forming. An occasionally proposed alternative, ammonia, would be liquid on other worlds much colder than ours, but at such low temperatures chemical reactions that could lead to life would operate sluggishly and living systems may struggle to become established.

1.2 The building blocks of life

We have discovered that life on Earth relies mainly on four elements, hydrogen, oxygen, carbon and nitrogen, with smaller amounts of two other elements, sulfur and phosphorus. Yet these six elements are found in a wide variety of organic combinations. Each combination has its own role in maintaining and perpetuating living systems. To fully appreciate how living systems operate we must begin to think of our elements in terms of the molecules in which they are contained.

1.2.1 Water

Before discussing the major classes of organic molecules found in living things we will pause and consider the role of water. The water molecule is the major component of living tissues, generally accounting for 70% of their mass.

- What clues are there in Table 1.1 to suggest there are relatively large quantities of water present in living systems?
- Hydrogen and oxygen, the elements that combine to form water, are the two most abundant elements in the human body.

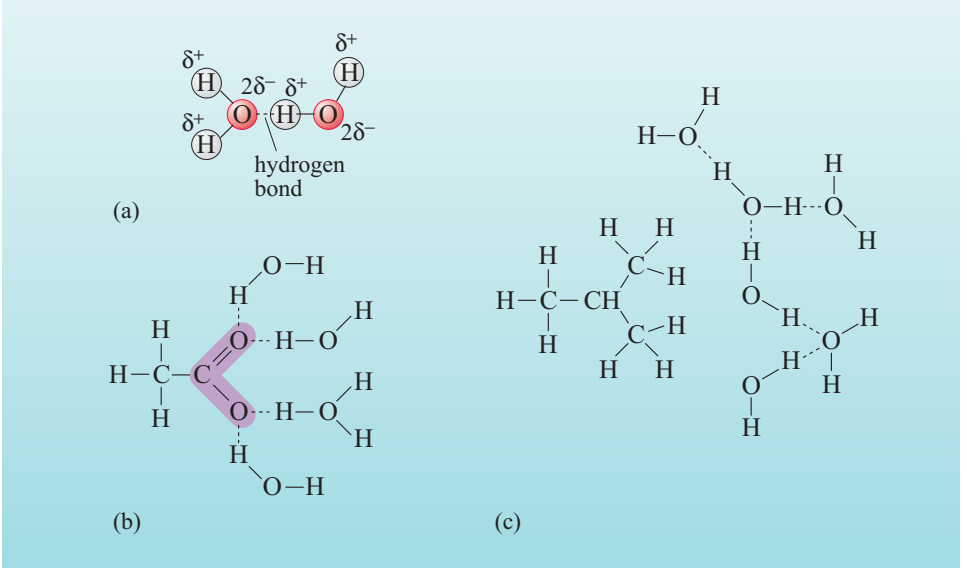


Figure 1.3 Water molecules and their interactions. (a) Water molecules carry a partial positive charge (δ^+) on their hydrogen atoms and a partial negative charge (δ^-) on their oxygen atoms that can interact to form a hydrogen bond. (b) Water molecules interact with polar organic molecules. (c) Apolar organic molecules do not interact with water molecules.

Because living systems contain so much water it is inevitable that the majority of other components will exist in an aqueous environment. Water is a **polar** solvent, in other words each side of the molecule carries a different electrical charge. In detail, the hydrogen atoms in water are positively charged while the oxygen atom is negatively charged. Because opposites attract, two water molecules can interact to form a hydrogen bond between the hydrogen of one water molecule and the oxygen of another (Figure 1.3a).

Water molecules not only interact with themselves but can also exert a significant influence on organic molecules. When considering organic solutions it is useful to remember the saying 'like dissolves like'. What this means is that polar solvents such as water will dissolve polar organic molecules (Figure 1.3b). Conversely, **apolar** (non-polar) organic molecules do not readily dissolve in water (Figure 1.3c). These characteristics allow us to define two classes of organic molecule:

- Those that are polar, have a high affinity for water, and are therefore soluble are termed hydrophilic (water lovers) molecules.
- Those that are apolar, have a low affinity for water, and are therefore relatively insoluble are called hydrophobic (water haters) molecules.

We shall see later that living systems exploit the hydrophobic and hydrophilic nature of different molecules to perform specific functions.

Table 1.2 lists the major constituents found in a bacterium. Notice the dominant constituent ism water, but other chemicals are present.

■ Are the majority of organic molecules in a living system present as small or large structures?

❑ Most of the organic molecules are very large.

Table 1.2 Types and abundances of the molecules that make up a bacterium.

	Percent of total weight	Number of types of molecule
water	70	1
inorganic ions, e.g. Na^+, K^+ and Ca^{2+}	1	20
small organic molecules (< 1000 atomic units), e.g. fatty acids, sugars, amino acids, nucleotides	7	750
large organic molecules (> 100 000 atomic units), e.g. collections of lipids, carbohydrates, proteins, nucleic acids	22	5000

An atomic unit, or more correctly atomic mass unit, is defined as one-twelfth the mass of an atom of carbon and approximates to the mass of a single proton or neutron.

So, except for water, most of the molecules in a living system are large organic molecules or 'macromolecules'. These macromolecules can be subdivided into four different types: **lipids**, **carbohydrates**, **proteins** and **nucleic acids**.

These macromolecules are usually the products of combining many individual organic molecules called **monomers** (from the Greek for single-parts) to create **polymers** (from the Greek for many-parts). Each of these types of macromolecule has a specific function in living systems. We will now examine these different types of macromolecule and the roles that they perform.

1.2.2 Lipids (fats and oils)

Lipids are a diverse group of molecules that have one hydrophobic and one hydrophilic end (Figure 1.4). Overall they are poorly soluble in water, a feature which ensures that they are rarely found as individual molecules. Lipids arrange themselves into weakly bonded aggregates that can be considered macromolecules. Lipids are a convenient and compact way to store chemical energy and the weak bonding within their macromolecular structure results in a high degree of flexibility that is useful in membranes.

1.2.3 Carbohydrates

Carbohydrates are molecules that have many hydroxyl groups (–OH) attached, as shown in Figure 1.5. These hydroxyl groups are polar so carbohydrates are soluble in water. Sugars are common carbohydrates that form ring-like structures when dissolved in water (Figure 1.5). Sugars with five carbon atoms are called pentoses while those with six carbon atoms are called hexoses. Large carbohydrate structures are called polysaccharides and consist of sugar monomers connected together. This process whereby monomers are linked together to give a polymer is called polymerization. Polymerization occurs by reactions that involve the loss of water and result in a linear or branched network as in polysaccharides. Polysaccharides are useful energy stores and can also provide structural support for organisms.

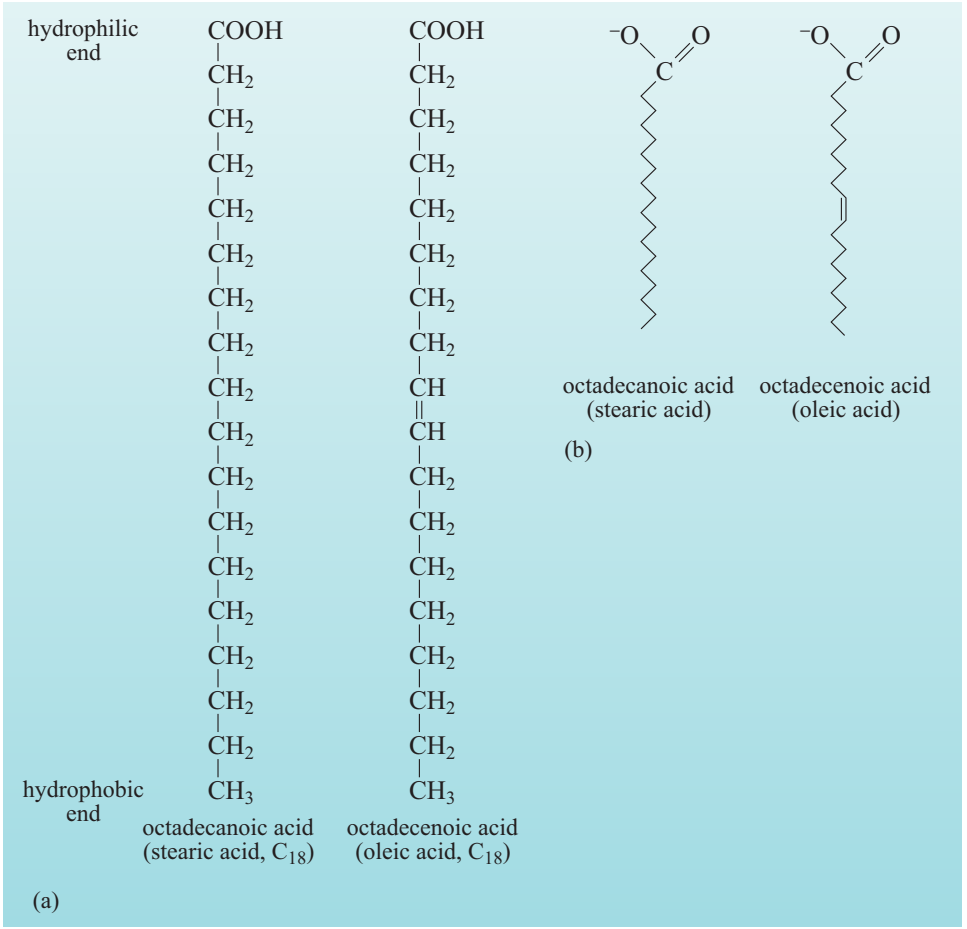


Figure 1.4 The structure of two common lipids containing eighteen carbon atoms, octadecanoic acid and octadecenoic acid: (a) detailed structures; (b) abbreviated structures.

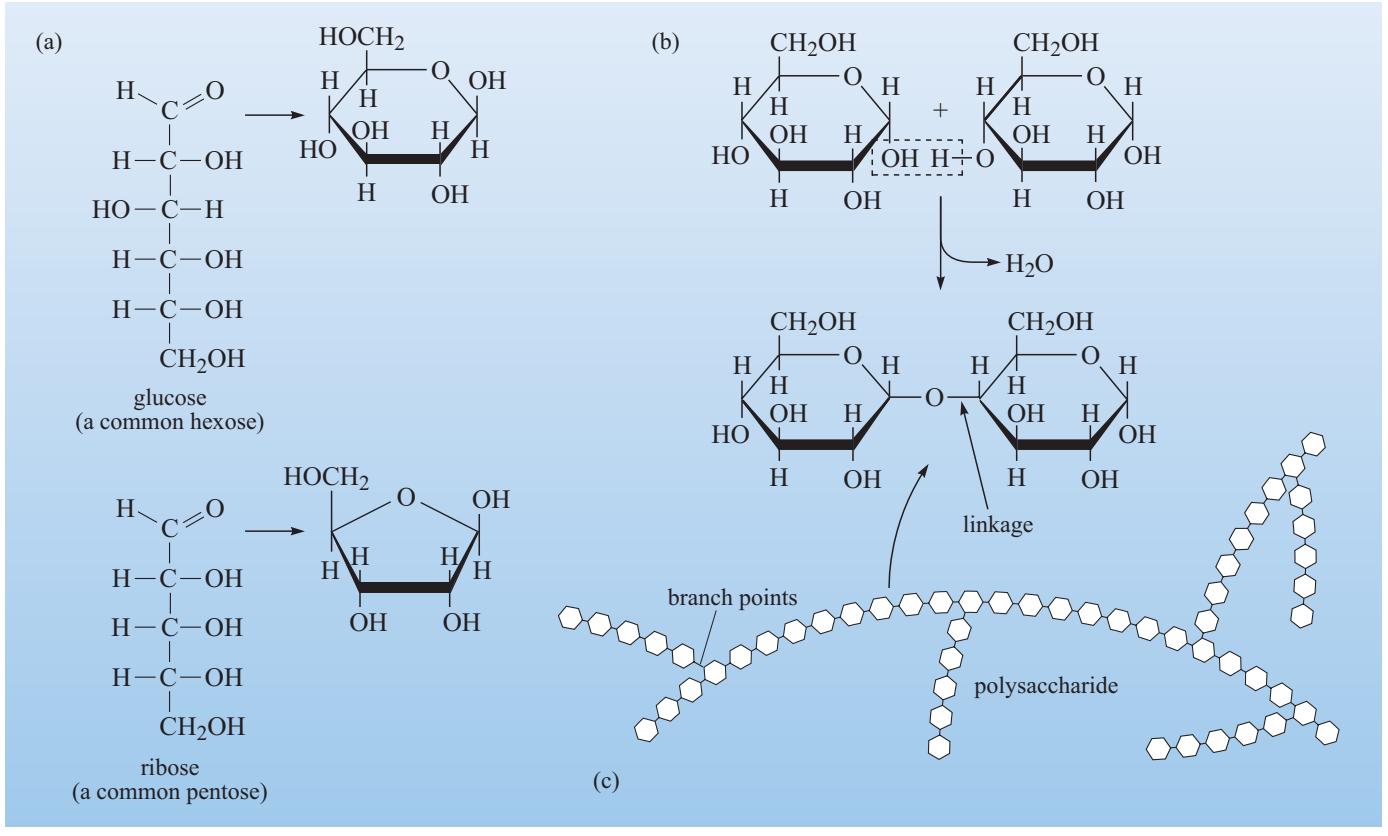


Figure 1.5 (a) The structure of two common five- and six-carbon sugar molecules. (b) Sugar monomers polymerize by simple reactions that involve the loss of water to form (c) polysaccharides.

1.2.4 Proteins

Proteins (from the Greek *proteios* or primary) are the most complex macromolecules found in living systems. They consist of long 'trains' of amino acids linked together. As with polysaccharides, they are linked together by simple reactions that involve the loss of water (Figure 1.6). There are 20 different amino acids found in the proteins of living systems and it is the particular sequence of amino acids that gives a protein its function. Proteins are perhaps the most important of life's chemicals and have an enormous number of different roles. For example, they provide structure (e.g. in human fingernails and hair) and act as **catalysts** (e.g. aiding digestion in our stomachs). Proteins with catalytic properties are called **enzymes**.

A catalyst is a substance that increases the rate of a reaction but is not itself used up in the reaction.

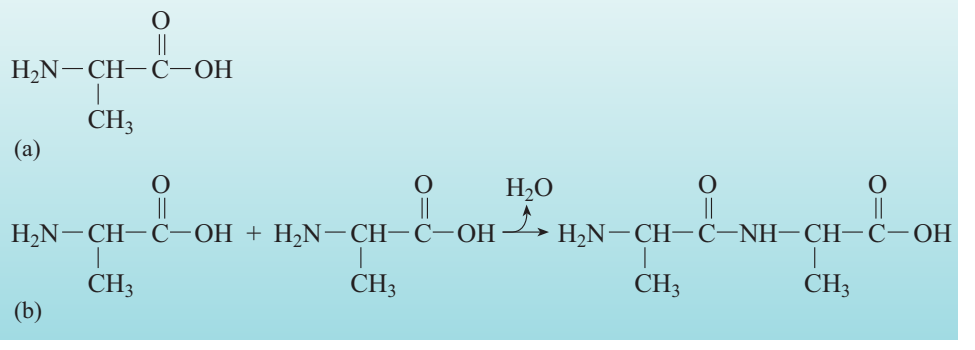


Figure 1.6 (a) An amino acid. (b) Amino acid monomers polymerize by simple reactions that involve the loss of water. Many reactions of this type will eventually produce proteins.

1.2.5 Nucleic acids

Nucleic acids are the largest macromolecules found. They exist as a collection of individual **nucleotides** (Figure 1.7a) linked together in long linear polymers (Figure 1.7b). As with sugars and amino acids, nucleotides can be linked together by simple reactions that involve the loss of water. Nucleotides contain:

- a five-carbon sugar molecule
- one or more phosphate groups
- a nitrogen-containing compound called a nitrogenous base.

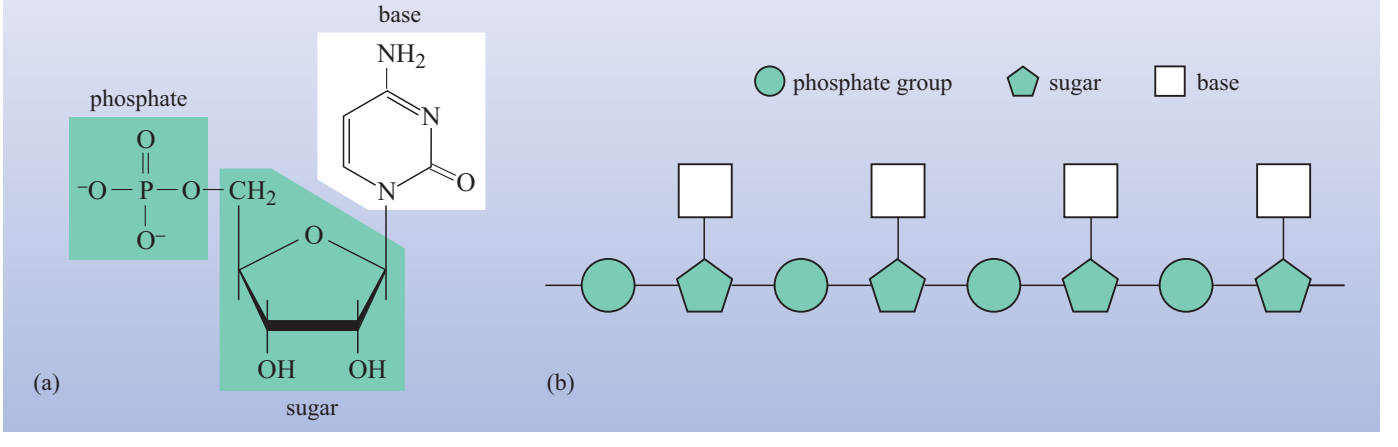


Figure 1.7 (a) The structure of a nucleotide consisting of a phosphate group, sugar molecule and nitrogenous base (cytosine in this instance). (b) Nucleotides polymerize by simple reactions that involve the loss of water to form nucleic acids.

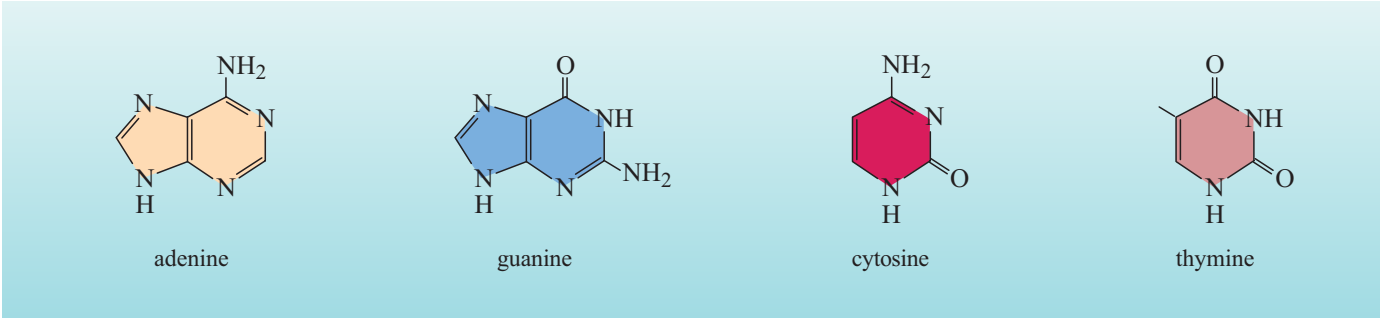


Figure 1.8 The bases found in DNA.

The most famous nucleic acid is deoxyribonucleic acid or **DNA**. Another important nucleic acid is ribonucleic acid or **RNA** (we will examine the role of RNA later in this section). Prior to 1953 it was known that DNA contained four different nucleotides, each possessing identical sugar and phosphate groups but different bases. These bases are adenine, guanine, cytosine and thymine, sometimes referred to in shorthand by the letters A, G, C and T (Figure 1.8). However, exactly how these components were arranged was unknown.

In 1953, James Watson and Francis Crick recognized that DNA consists of two long molecular strands that coil about each other to form a **double helix** (Figure 1.9). Bonds that resemble the steps of a spiral staircase connect the two helical strands. The steps consist of two nucleotides, with each nucleotide forming half of the step. The bases in the centre of the helix are joined by weak hydrogen bonds. The bases always match – adenine in one nucleotide is always paired with thymine in the other and, likewise, guanine is always linked to cytosine. Hence, the sequence of bases on one strand strictly determines the base sequence on the other.

DNA strand	DNA strand
A	T
T	A
G	C
C	G

The bases are attached to their helical strands by sugar groups, which in turn are connected together along the exterior of the helix by phosphate groups.

Figure 1.9 The DNA double helix. Note that the 'ribbons' are not real, but are there to illustrate the nature of the double helix.

■ Examine the DNA shorthand above and compare it with the idealized structure of nucleic acids in Figure 1.7. Which parts of the nucleic acid structure are omitted in the shorthand?

❑ The sugar and phosphate backbone is not represented because the types of sugar and phosphate groups do not change within a particular nucleic acid.

■ Can you suggest how this complementary system allows DNA to pass on genetic information?

❑ Since the sequence of bases on one strand of the helix determines the sequence on the other, 'unzipping' the double helix provides two templates that can be used to produce two new DNA molecules from the single parent.

Special protein enzymes separate the strands of the double helix. The single strands hook up with spare nucleotides present in the liquid surrounding the molecule. Each base in the unzipped strand latches on to its complementary base. The sugar and phosphate groups of the newly acquired nucleotides then join together into helical strands, so two identical double-helix molecules are formed, exactly like the original (Figure 1.10).

The discovery of DNA structure provided the basis for understanding one of the key characteristics of life: the mechanism that enables biological molecules to replicate themselves.

Different DNA sequences also account for the variations between individuals of the same species, and for the differences between species (Box 1.2).

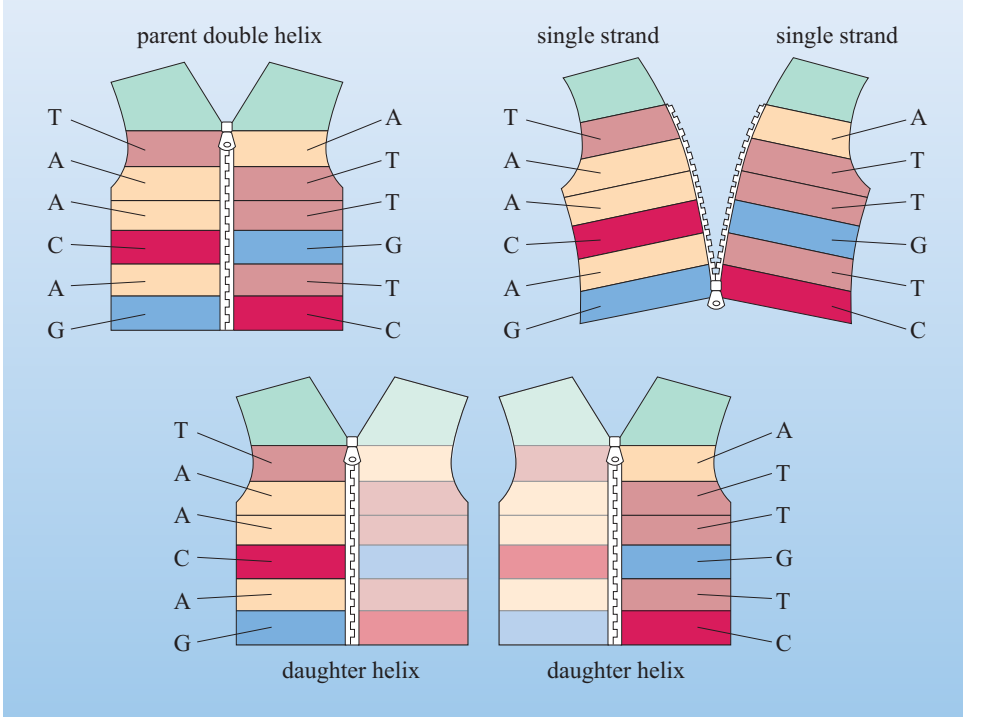


Figure 1.10 DNA replication showing how one parent double helix 'unzips' and produces two identical daughter double helices.

BOX 1.2 DNA HYBRIDIZATION – THE 'WHO'S WHO' OF GENETICS

We have talked about unzipping the DNA double helix to produce two identical DNA strands and it is reasonable to assume that single strands of DNA from the same species can be zipped back together. But what of single DNA strands from closely or distantly related species, can they be spliced together? The answer is yes – but only partly. Closely related species will have similar, but not identical, sequences of nucleotides and will splice relatively well, whilst distantly related species will have dissimilar sequences of nucleotides and will splice relatively poorly.

This concept has been exploited to produce a technique known as DNA hybridization. The DNA helix from one species is unzipped by heating and then combined with unzipped DNA from another species. When the mixture cools some of the different strands splice together. If the species are closely related the strands will almost match and the new double helix will be strongly bound together. If the species are not closely related the opposite is true – the new double helix will not be strongly bonded together. The strength of the bonding, and therefore the strength of the species relationship, is revealed when the new mixture is heated again – the weakly bound (less-related) helices unzip at lower temperatures.

DNA hybridization has revealed that some species that appear similar and were thought to be related, actually have quite different ancestries. For example, the belief that owls were akin to falcons and hawks (Figure 1.11) was mistaken as these birds are more closely related to nightjars. Similarly, starlings are not close relatives of crows but are related more to mockingbirds.

Figure 1.11 Appearances can be deceptive. The DNA of owls (centre) reveals that they are more closely related to nightjars (left) than falcons and hawks (right).

In addition to self-replication and passing on information from generation to generation, DNA also governs protein synthesis. DNA contains a set of 'instructions' called the **genetic code** which is expressed by the sequence of bases in the molecule. For example ATGC would be one part of a genetic code, while ATGG would be another. The genetic code directs the production of thousands of proteins needed for the structure and function of living systems. In the process of protein synthesis, the DNA message is first transcribed (copied) into an RNA message. RNA is very similar to DNA, but has slight differences. The sugar component in RNA is ribose rather than deoxyribose, and the base uracil is present

instead of thymine (Figure 1.12). Hence, the four RNA bases are adenine, guanine, cytosine and uracil or in shorthand A, G, C and U. When bonding with DNA, uracil replaces thymine and forms a base pair with the adenine of DNA.

DNA strand	RNA strand
A	U
T	A
G	C
C	G

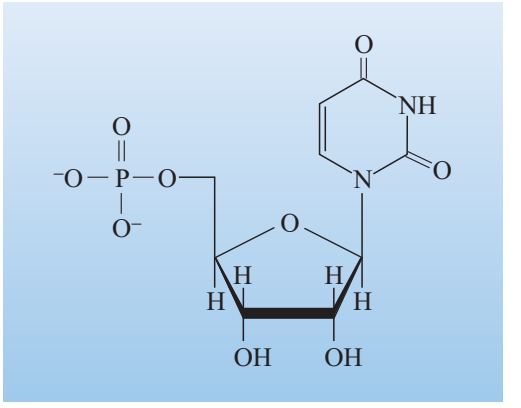


Figure 1.12 An RNA nucleotide containing the sugar and base that makes RNA distinct from DNA.

During transcription, the DNA makes messenger RNA (mRNA). To do this, the DNA helix first unzips as if to replicate itself. Next, instead of each of the nucleotides seeking out a matching DNA nucleotide to build a new DNA double helix, they seek out RNA nucleotides to produce a strand of mRNA. The mRNA strand is then released and the DNA helix zips itself together again. The released mRNA carries its own version of the DNA sequence into a region that contains free amino acids where molecular factories called ribosomes use the mRNA to combine amino acids into long protein chains.

1.2.6 The cell

Many different molecules must be in close association for living systems to operate. This is because in chemistry the rate of reaction generally increases with the concentration of the reactants. Yet what is there to stop molecules simply drifting off in solution and bringing a halt to the chemistry of life? The answer is the cell. In its simplest form, a cell is a small bag of molecules that is separated from the outside world (Figure 1.13). At the centre of the cell, strands of DNA are devoted to the storage and use of genetic information. The DNA is surrounded by the cytosol, which is a salt water solution containing enzymes and the ribosomes. The cell contents are surrounded by a soft membrane. This is called the cell membrane and consists of lipids and proteins. The cell membrane restricts the movement of molecules into and out of the cell and thereby protects the cell's contents. Finally, a tough cell wall consisting of carbohydrate molecules and short chains of amino acids provides the cell's rigidity. So cells provide an environment in which biochemical processes can occur and genetic information can be stored. Cells are the basic structural unit of all present-day organisms on the Earth, but vary in number, shape, size and function. For instance, bacteria are single-celled organisms whereas humans contain around 10^{22} cells.

Simple cells such as the one in Figure 1.13 can reproduce by splitting in two. This process begins when DNA is replicated and the two new DNA molecules attach themselves to different parts of the cell membrane. Next the cell begins to divide, separating the two DNA-containing regions. Finally, when cell division is complete, two identical daughter cells have been produced from the parent.

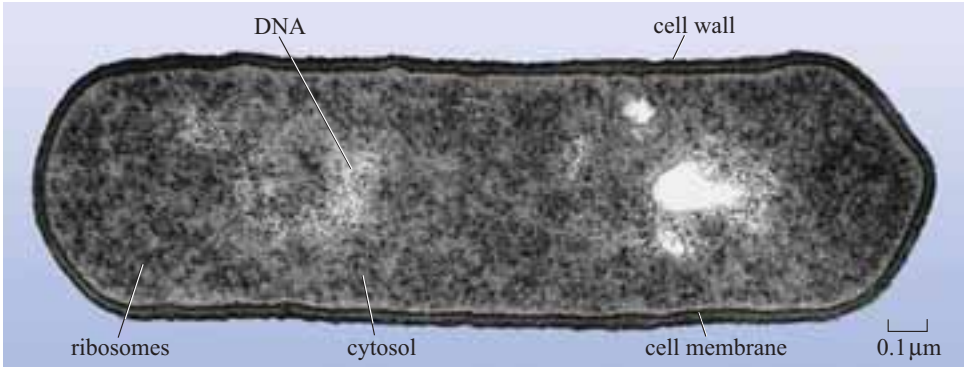


Figure 1.13 A simple cell.

1.3 How to study the origins and remains of life

Now that we have explored what we mean by life, what is necessary for life to exist and what living organisms are made of, we can turn to the detection of life. At this point, it is also appropriate to consider how we will perform our investigations into life's origins. As you will see, there is more than one approach to the problem.

1.3.1 Identifying past and present life

'Biological marker', or 'biomarker' for short, is a term initially used in petroleum exploration. Petroleum geochemists attempt to discover when and where the correct environments existed in the geological past to produce and accumulate fossil fuels. One of their most useful tools is the recognition of molecular fossils or biomarkers that are specific to particular organisms in organic-rich rocks. The value of a biomarker increases if the organism was restricted to a certain environment, thereby increasing its diagnostic value.

Recently, astrobiologists have adopted the term biomarker and its definition has been extended. Today, the word 'biomarker' is no longer used exclusively for organic material but is used for any evidence that indicates present or past life detected either in situ or remotely. In 1999, astrobiologists David Des Marais and Malcolm Walter listed the following categories of biomarker:

1 Cellular remains.

2 Textural fabrics in sediments that record the structure and/or function of biological communities.

3 Biologically produced (biogenic) organic matter.

4 Minerals whose deposition has been affected by biological processes.

5 Stable isotopic patterns that reflect biological activity.

6 Atmospheric constituents whose relative concentrations require a biological source.

A brief glance at these criteria will reveal the extreme subjectivity included in the definition. Establishing whether or not textural fabrics or organic matter in a sample is biogenic might be a very difficult process. For example, aromatic hydrocarbons are a class of organic molecule that can be generated by heat and pressure on the

Aromatic hydrocarbons are molecules built up of units containing six carbon atoms joined in a ring by alternating single and double bonds. If several carbon rings are present the term polyaromatic hydrocarbon, or PAH for short, is used.

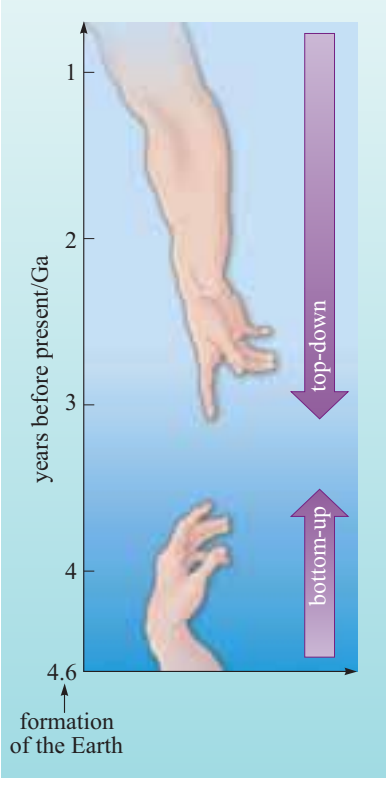


Figure 1.14 The two approaches to the study of life's origins.

biological remains of living organisms. They are a major component of coal. Coal comprises the fossil remains of land plants. The problem, however, is that where aromatic hydrocarbons are concerned, many roads lead to Rome. These molecules can just as easily be produced in internal combustion engines, garden barbecues or giant stars. It is easy to imagine the controversy that could arise from detecting such organic compounds and attributing their presence to once-living organisms.

1.3.2 Two approaches to the origin of life

Studies into life's origins can be categorized into two general types of enquiry:

- a 'bottom-up' strategy, and
- a 'top-down' strategy.

The bottom-up approach focuses on a collection of inanimate elements, molecules and minerals with known properties and attempts to figure out how they may have been combined in the past to create a living organism. By contrast, the top-down method looks at present-day biology and uses the information to extrapolate back towards the simplest living entities. In the following sections we will use both the bottom-up and top-down tactics to arrive as close as possible to an answer to the question of how life on Earth arose. Of course, there is no guarantee that we can take both approaches far enough so that the two arrive at a mutual destination (Figure 1.14).

1.4 Organic matter in the Universe

Let us begin our look at the origin of life using the bottom-up approach. Organic matter is a fundamental constituent of living systems and represents one of the inanimate substances that life must have been generated from on the early Earth. Furthermore, it is inevitable that the distribution of organic matter in the Universe will have a direct bearing on where life could originate. So, to put our own planet in context, we will now examine the major environments in which it is thought that organic matter is created.

The production of organic matter was occurring in our region of the Galaxy long before our Solar System formed. The circumstellar (surrounding a star) envelopes around carbon-rich red giant stars churn out large amounts of extraterrestrial organic molecules (Figure 1.15). It has been postulated that the chemical reactions that form organic molecules in these regions are similar to those observed in a simple candle flame on Earth and the dominant products are aromatic hydrocarbons. Figure 1.16 presents infrared spectra from a circumstellar envelope around IRAS21282+5050. The close match between the spectra of the starlight and two laboratory aromatic hydrocarbon standards suggests that these organic molecules are common constituents of that part of space.

The name IRAS21282+5050 denotes that the object was identified in the data returned by the InfraRed Astronomical Satellite (IRAS) which surveyed the sky in 1983. The numbers refer to the object's position on the sky.

Figure 1.15 A circumstellar envelope in which large amounts of extraterrestrial organic matter are found.

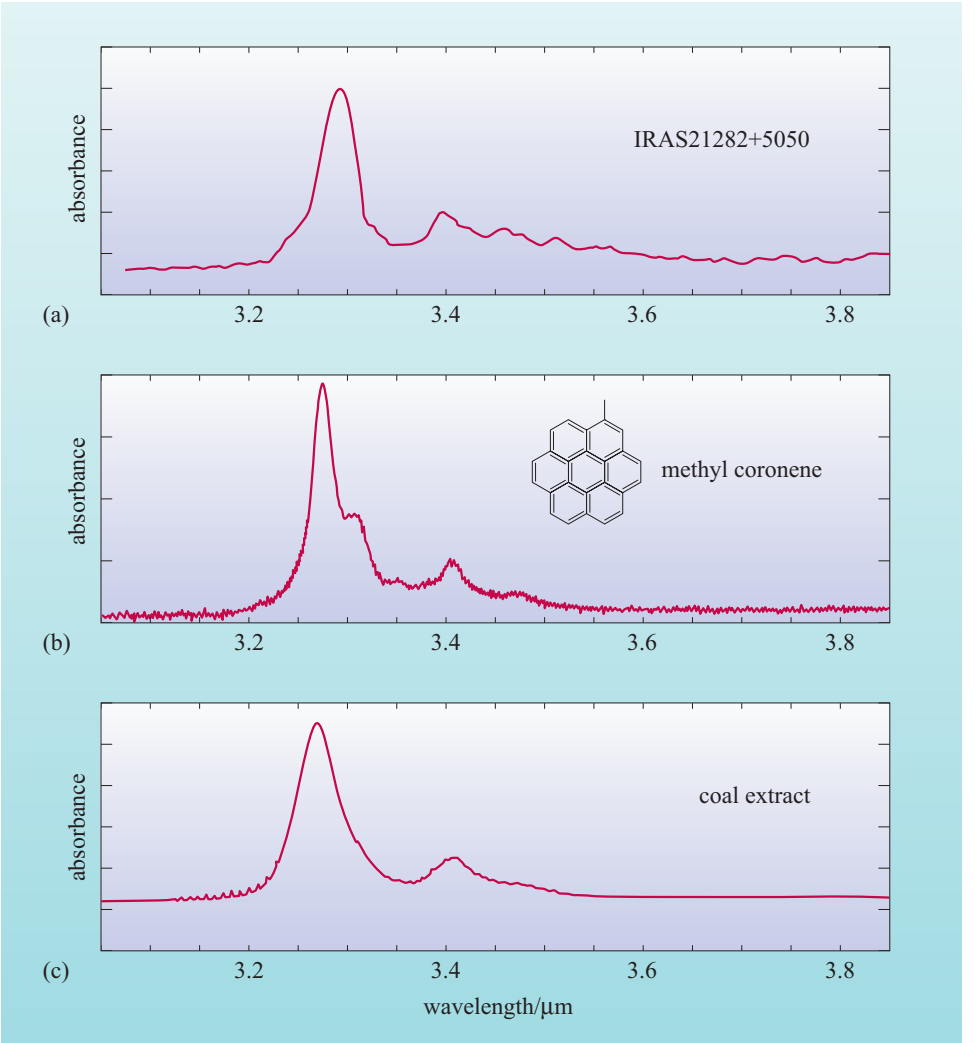


Figure 1.16 Partial infrared spectra of a circumstellar envelope from IRAS21282+5050 compared to a single aromatic hydrocarbon (methyl coronene) and an aromatic hydrocarbon mixture (coal extract).

The star's wind expels these molecules into interstellar space. From there they can be caught up in other environments where organic matter may be created. Some of these environments are molecular clouds (Figure 1.17) that represent the coldest (10–20 K) and densest parts of the **interstellar medium** and play a key role in the evolution of the Galaxy. Every star and planetary system was formed inside a molecular cloud; the other types of interstellar clouds are too warm and diffuse to allow the generation of stars. Star formation occurs when deeply embedded clumps of interstellar gas and dust collapse under their own gravitational attraction.

Numerous different molecules have been identified in interstellar clouds and circumstellar envelopes (Table 1.3) and the list continues to expand. It may seem surprising that the largest quantity of organic molecules in our Galaxy is found not on the Earth but in giant molecular clouds.

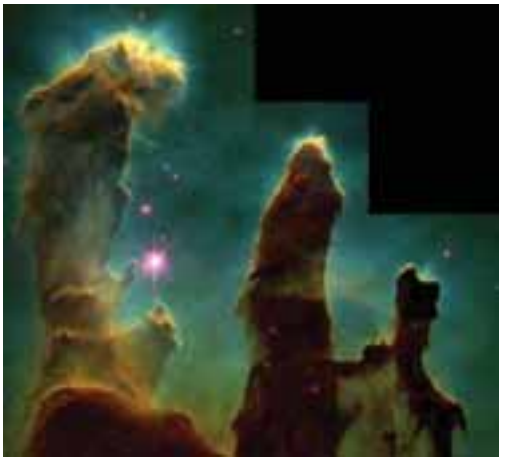


Figure 1.17 Hubble space telescope images of molecular clouds in the Eagle Nebula. The gaseous clouds are several light years long.

Table 1.3 The molecules detected in interstellar space and in circumstellar envelopes. Note: D is deuterium, a form of hydrogen.

hydrogen species					
H_2	HD	H_3^+	H_2D^+		
hydrogen and carbon compounds					
CH	CH^+	C_2	CH_2	C_2H	C_3
CH_3	C_2H_2	C_3H (lin)	C_3H (circ)	CH_4	C_3H_2 (circ)
H_2CCC (lin)	C_4H	C_5	C_2H_4	C_5H	H_2C_4 (lin)
CH_3C_2H	C_6H	H_2C_6	C_7H	CH_3C_4H	C_8H
hydrogen, carbon and oxygen compounds					
OH	CO	CO^+	H_2O	HCO	HCO^+
HOC^+	C_2O	CO_2	H_3O^+	$HOCO^+$	H_2CO
C_3O	CH_2CO	HCOOH	H_2COH^+	CH_3OH	HC_2CHO
C_5O	CH_3CHO	C_2H_4O (circ)	CH_3OCHO	CH_2OHCHO	CH_3COOH
CH_3OCH_3	CH_3CH_2OH	$(CH_3)_2CO$			
hydrogen, carbon and nitrogen compounds					
NH	CN	NH_2	HCN	HNC	N_2H^+
NH_3	$HCNH^+$	H_2CN	HCCN	C_3N	CH_2CN
CH_2NH	HC_2CN	HC_2NC	NH_2CN	C_3NH	CH_3CN
CH_3NC	HC_3NH^+	C_5N	CH_3NH_2	CH_2CHCN	HC_5N
CH_3C_3N	CH_3CH_2CN	HC_7N	CH_3C_5N	HC_9N	$HC_{11}N$
hydrogen, carbon (possibly), nitrogen and oxygen compounds					
NO	HNO	N_2O	HNCO	NH_2CHO	
other species					
SH	CS	SO	SO^+	NS	SiH
SiC	SiN	SiO	SiS	HCl	NaCl
AlCl	KCl	HF	AlF	CP	PN
H_2S	C_2S	SO_2	OCS	HCS^+	SiC_2(circ)
NaCN	MgCN	MgNC	H_2CS	HNCS	C_3S
$HSiC_2$	SiC_3	SiH_4	SiC_4	CH_3SH	C_5S

Note: (circ) denotes circular molecules and (lin) denotes linear molecules.

■ Using Table 1.3, which is the smallest molecule and which is the largest molecule detected in interstellar and circumstellar environments?

❑ H_2 is the smallest molecule and $HC_{11}N$ is the largest molecule.

These various types of molecules are formed through a complicated network of chemical reactions inside the interstellar or circumstellar clouds where they are found. In dense molecular clouds the temperatures are so low that any gas hitting a dust grain of solid matter will immediately freeze out to form an icy mantle. Once the organic compounds are attached to a grain, chemical reactions are catalysed by the grain surface and any reaction products are processed further by ultraviolet and cosmic rays. The products of grain mantle chemistry may have an opportunity to take part further in organic chemistry when they are incorporated in the warm (200–400 K) and dense (>100 H_2 molecules/cm^{-3}) areas of gas located around recently formed stars. These areas have been shown to be the most chemically diverse regions in the interstellar medium and are called hot cores. Many of the species in Table 1.3 have only been seen in hot cores. Various interesting chemical compounds are produced when the icy grain mantles are heated in hot cores. When a star is formed nearby, its radiation evaporates the ice and the molecules return to the gas phase.

Once star formation is underway, a spinning disc of dust and gas called a **solar nebula** is produced. The solar nebula inherits a variety of organic molecules from its molecular cloud (Figure 1.18) although some organic matter may be synthesized anew. Processes similar to the gas phase reactions proposed for circumstellar shells and the grain catalysed reactions in molecular clouds may have operated in the nebula. Eventually, the solar nebula is replaced by a solar system containing stars and planets.

Figure 1.18 An artist's impression of the solar nebula – a rotating disk of dust and gas that gave rise to the Sun and planets.

1.5 Synthesis of organic molecules on the early Earth

Following the formation of a solar nebula, the surfaces of newly formed planets and their atmospheres provide the next opportunity for the generation of organic molecules. It is the processes taking place on planets that we turn to next, by considering in detail the reactions that may have occurred following the birth of the planet we understand best, our own Earth.

1.5.1 Energy sources

It appears that much of the chemistry of the Universe is organic chemistry and there are a number of extraterrestrial environments that are capable of generating organic compounds that may be biologically useful. But these organic compounds are relatively simple and are a far cry from the highly organized organic systems found in living organisms. The laws of physics dictate that systems will inevitably trend towards disorder if left to their own devices. Energy is required to generate and sustain order. So to understand how life began on the early Earth we must appreciate what energy sources were available.

Energy may have had two roles in the origin of life: it could have fuelled reactions that synthesized organic matter on the early Earth and it was certainly utilized at some point to sustain primitive life.

Table 1.4 lists the main sources of energy available on the present-day Earth.

Table 1.4 Present-day sources of energy averaged over the Earth.

Source	Energy /J m^{-2} yr^{-1}
total radiation from the Sun	1090000.0
ultraviolet light	1680.0
electric discharges (lightning)	1.68
cosmic rays	0.0006
radioactivity (to 1 km depth)	0.33
volcanoes	0.05
shock waves (atmospheric entry)	0.46

■ What is the largest source of energy on the Earth, and by what factor does it exceed the next largest?

❑ Sunlight is the largest energy source, by a factor of 1 090 000/1680 = 650.

The strength of the Sun on the early Earth is thought to have been up to 20–30% weaker than it is today, but it would still have been the major energy source available for the synthesis of organic matter. As we shall see later, electrical discharges have been used in many experiments that have attempted to recreate the synthesis of organic molecules on the early Earth. But given the small amounts of energy provided in this way, it seems unlikely that they were important sources of energy for organic synthesis. The decay of radioactive forms of uranium and potassium would have provided heat from the Earth's interior. Primordial heat would have been generated as the Earth's accretion released gravitational energy and volcanic activity would have led to the eruption of lavas at temperatures well over 1000 °C. The shock waves generated as meteors and **meteorites** passed through the atmosphere would have also contributed to the energy available to synthesize molecules. Yet all of these energy sources would have been small relative to that supplied by the Sun.

1.5.2 Miller's origin of life experiment

In the early 1950s, Stanley Miller was working for his PhD at the University of Chicago under the guidance of Harold C. Urey (1893–1981). In an attempt to recreate the types of chemical reactions that may have occurred on the early Earth, Miller used a flask of water to represent the primordial ocean. He then heated the flask. Water vapour circulated through the apparatus, setting his primitive hydrosphere in motion (Figure 1.19). Another flask, placed higher than the one containing water, represented the atmosphere and contained methane (CH_4), ammonia (NH_3) and hydrogen (H_2) all of which became mixed with the invading water vapour. Next, in a move reminiscent of a low-budget Frankenstein movie, he subjected the gases to a continuous electrical discharge that represented lightning. The electrical energy caused the gases to interact and the reaction products accumulated lower down in a water-filled trap. The Miller–Urey experiment was allowed to run for one week. When the reaction products were analysed it became clear that a number of organic compounds needed for life, notably amino acids, had been produced relatively simply and abiotically in a reducing atmosphere. The ease at which these compounds could be produced suggested that they should be abundant and widespread in the Universe.

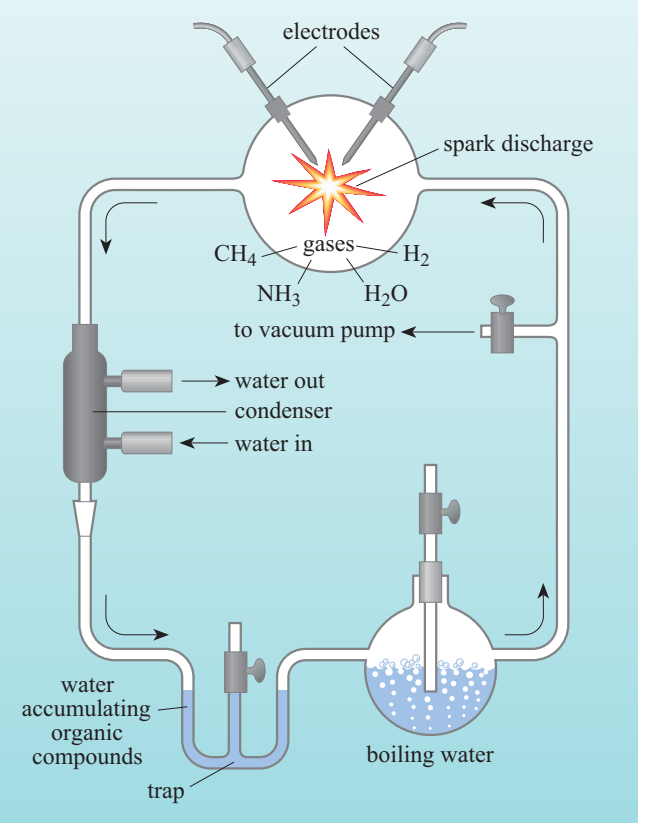


Figure 1.19 Miller and Urey's apparatus used in the abiotic synthesis of amino acids. The lower flask containing boiling water represents the primordial ocean. Water vapour enters the upper flask, representing the primordial atmosphere, and mixes with methane, hydrogen and ammonia. Electrical discharges cause the gases to combine into amino acids, which then accumulate in a water-filled trap.

1.5.3 The fall of the Murchison meteorite

On a Sunday morning in September 1969 the tranquillity of a small town called Murchison, near Melbourne, Victoria in Australia was shattered by a sonic boom that heralded the arrival of a rare type of carbon-rich meteorite. The Murchison meteorite was a **carbonaceous chondrite**, a piece of **asteroid** that had sat out in space, somewhere between Mars and Jupiter, and had remained unaltered since shortly after the birth of the Solar System (Figure 1.20). The first suggestions that the Murchison meteorite contained organic molecules came from the initial eyewitness reports that commented on solvent smells emanating from the stone. The early investigations of the organic inventory of the Murchison meteorite were performed in laboratories that had been preparing for the return of samples from the Apollo lunar missions. Several classes of organic compounds were quickly recognized, including amino acids. These discoveries indicated that the early Solar System must have been a place in which much organic chemistry was taking place. Table 1.5 compares the types and amounts of amino acids synthesized in the Miller–Urey experiment to those found in the Murchison meteorite.

■ How good a match is the amino acid content of the Miller–Urey products and the Murchison meteorite?

❑ The match is remarkably good. Both the types and abundances of amino acids appear similar.

The discovery of similar organic compounds in the Murchison meteorite and Miller–Urey experiment supported the idea that the production of the basic organic building blocks of life is a widespread and probably common feature of the Universe. It appeared that simple organic molecules would have been available in reasonable concentrations in the early Solar System and on the early Earth. It was not difficult to imagine that, eventually, simple molecules would be polymerized into macromolecules that would achieve the properties of life.

The Miller–Urey experiments have been reproduced many times using different energy sources and slightly different starting mixtures. Each time, biologically useful small molecules have been created from mixtures of reduced gases. However, as you'll see in Section 2.4.4, recent models of atmospheric evolution indicate that the early Earth would not have had a methane- and ammonia-rich reducing atmosphere because these gases are easily destroyed by sunlight. It now seems that the more stable molecules carbon Dioxide, nitrogen and water dominated the Earth's early atmosphere. Under these less reducing conditions Miller–Urey synthesis is much more difficult. So it appears that the environment of the early Earth might have been fit for sustaining life but less suitable for the in situ production of life's organic raw materials.

Figure 1.20 The Murchison carbonaceous chondrite, a type of meteorite that preserves organic matter from the early Solar System.

Table 1.5 Abundances of amino acids synthesized in the Miller–Urey experiment and those found in the Murchison meteorite. The number of dots represents relative abundance. Those amino acids used by life (i.e. in proteins) are indicated.

Amino acid	Abundance of amino acids		Found in proteins on Earth
	synthesized in the Miller–Urey experiment	Found in the Murchison meteorite	
glycine	●●●●	●●●●	yes
alanine	●●●●	●●●●	yes
α-amino-*N*-butyric acid	●●●	●●●●	no
α-aminoisobutyric acid	●●●●	●●	no
valine	●●●	●●	yes
norvaline	●●●	●●●	no
isovaline	●●	●●	no
proline	●●●	●	yes
pipecolic acid	●	●	no
aspartic acid	●●●	●●●	yes
glutamic acid	●●●	●●●	yes
β-alanine	●●	●●	no
β-amino-*N*-butyric acid	●●	●●	no
β-aminoisobutyric acid	●	●	no
γ-aminobutyric acid	●	●●	no
sarcosine	●●	●●●	no
N-ethylglycine	●●	●●	no
N-methylalanine	●●	●●	no

1.6 Delivery of extraterrestrial organic matter to the early Earth

1.6.1 We are stardust

With a non-reducing atmosphere on the early Earth, Miller–Urey reactions could not have produced the large amounts of organic matter that provided the raw materials for life. In this scenario, the early Earth would have been bereft of its own organic matter. In fact, in the early Solar System the paucity of organic matter available to create primitive life was not a problem exclusive to the early Earth. Figure 1.21 shows a plot of carbon abundance in the Solar System. In general, carbon can be considered as an indicator of the amounts of abiotic organic matter. Biotic organic matter (life) is also indicated. Inwards of the **asteroid belt** the amount of organic matter declines abruptly. So it appears that during the time that life originated, the whole inner Solar System contained little organic matter.

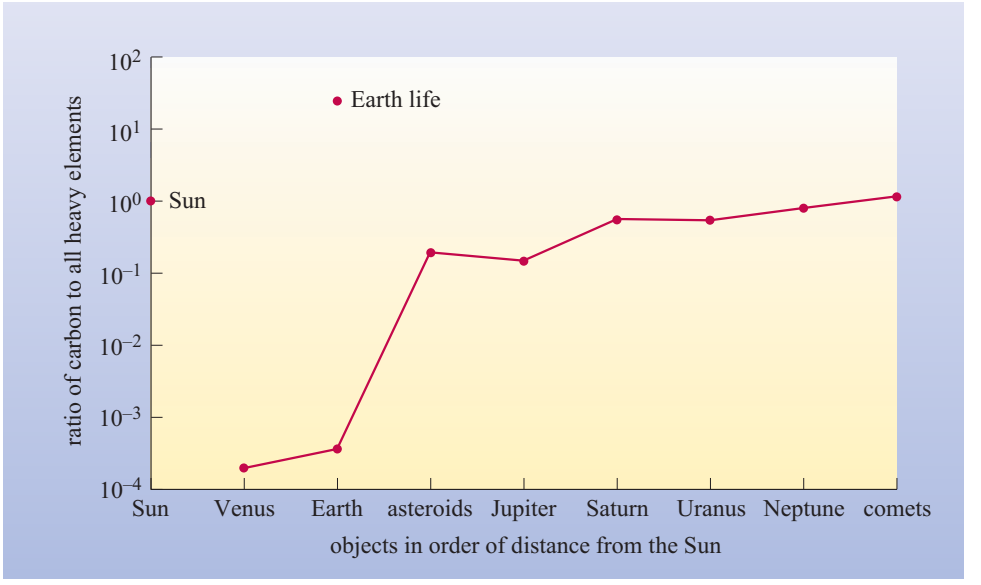


Figure 1.21 Ratio of carbon to heavy elements (all elements more massive than H and He) for various Solar System objects. The horizontal axis is not to scale.

We have mentioned that, in addition to organic matter, liquid water is a prerequisite for life. Liquid water has probably been stable on the Earth's surface since early in Earth's history and may have once been stable on the surface of Mars. Further out in the Solar System, liquid water may be maintained in subsurface environments of icy moons, but the zone in which liquid water has been stable on planetary surfaces over the last 4.6 Ga has not extended beyond about 1.7 AU (see Section 2.3.2).

This is a paradoxical situation as the two ingredients needed for the recipe of life (water and organic matter) appear, in a very general sense, to occupy different areas of our Solar System. In 1961 Juan Oró proposed a solution to the paradox of life's two key components occupying different parts of the Solar System.

■ Look again at Table 1.1. Is the elemental composition of life more similar to the Earth and its crust, or to the Cosmos as a whole?

❑ Life has elemental abundances that are more similar to the Cosmos than the Earth.

This relationship did not go unnoticed by Oró and other astrobiologists and led to the proposal that life could have been kick-started by organic matter delivered to the early Earth by extraterrestrial objects. In effect, meteorite and **comet** impacts would simply bring parts of the organic-rich region of the Solar System to those areas in which liquid water could survive. Table 1.6 compares the types of organic molecules found in living systems and those found in the abiotic mixtures in meteorites.

■ What is the main difference between the organic constituents of life and those of the Murchison meteorite?

❑ Meteorites contain simple organic molecules (monomers) whereas life also contains more complex polymerized versions of these molecules (polymers).

Table 1.6 The biological role and types of organic molecules (both monomers and polymers) found in life and in meteorites.

	Role	Life	Murchison meteorite
water	solvent	yes	yes
lipids (hydrocarbons and acids)	membranes, energy storage	yes	yes
sugars (monosaccharides)	support, energy storage	yes	yes
polysaccharides (polymerized sugars)		yes	no
amino acids	many (support, enzymes, etc.)	yes	yes
proteins (polymerized amino acids)		yes	no
phosphate	genetic information	yes	yes
nitrogenous bases		yes	yes
nucleic acids (polymerized sugars, phosphates and nitrogenous bases)		yes	no

QUESTION 1.2

Table 1.7 gives present-day mass ranges, estimated accretion rates and carbon contents of various extraterrestrial objects that fall to Earth today. The upper part of the table gives data for carbon in any form, i.e. both inorganic and organic forms. The lower part of the table gives data for carbon present only as organic matter. Using the data provided calculate the answers to the following questions.

(a) Complete the table by calculating the accretion rates of total carbon and organic carbon for each type of object listed.

(b) For total meteoritic matter, what object represents the greatest source of carbon per year? Is this carbon likely to arrive steadily over time?

(c) For organic matter, what object represents the greatest source of carbon per year? Is this carbon likely to arrive steadily over time?

(d) How does the total meteor carbon accretion rate compare with the organic carbon accretion rate for meteors? What does this tell you about the overall state of carbon in meteors?

(e) How much *total carbon* would be supplied by meteoritic matter in (i) ten years, (ii) a hundred years, and (iii) a hundred thousand years?

(f) How much *organic matter* would be supplied by meteoritic matter in (i) ten years, (ii) a hundred years, and (iii) a hundred thousand years?

Table 1.7 Accretion rates on Earth today.

Sources	Mass range /kg	Mass accretion rate (estimated) /10^6 kg yr^{-1}	Carbon %	Carbon accretion rate /10^6 kg yr^{-1}
meteoritic matter				
meteors (from comets)	10^{-17} to 10^{-1}	16.0	10.0	
meteorites	10^{-2} to 10^{5}	0.058	1.3	
crater-forming bodies	10^{5} to 10^{15}	62.0	4.2	
unmelted material contributing organic matter				
meteors (from comets)	10^{-15} to 10^{-9}	3.2	10.0	
meteorites, non-carbonaceous	10^{-2} to 10^{5}	2.9×10^{-3}	0.1	
meteorites, carbonaceous	10^{-2} to 10^{5}	1.9×10^{-4}	2.5	

QUESTION 1.3

The mass of carbon in the biosphere is 6×10^{14} kg. How long would it take for meteoritic materials to supply (a) a similar amount of carbon, and (b) a similar amount of organic carbon? Would the supply of extraterrestrial carbon and organic carbon to the early Earth have been greater or lesser than that of today?

The answer to Question 1.3 indicates that significant amounts of carbon arrive in the form of meteorites and meteors but only a fraction of this is organic matter. The largest outside contributors to the Earth's organic inventory are the meteors which rain down steadily on the planet but at present-day rates it would still take around a million years to create the equivalent carbon content of Earth's current biosphere. In addition to any organic matter contained within meteorites and meteors, the shock waves generated from these objects travelling through the Earth's atmosphere may have forced gases to combine, producing organic matter. Furthermore, when large meteorites arrive at the Earth's surface they, and the rocks they hit, may be vaporized and organic matter generated as the gases recombine. On the early Earth between 4 Ga and 3.8 Ga ago all of the processes outlined above would have been more relevant, simply because the rate at which meteors and meteorites arrived at the planet's surface would have been much higher. Evidence for this comes from the Moon where craters indicate that the asteroid and comet impacts that scarred the lunar surface reveal a final flurry of devastation. It is reasonable to assume that the Earth experienced a similar punishment. This period is called the **late heavy bombardment** and the supply of extraterrestrial objects and their associated organic matter to the Earth's surface would have been much greater and the raw materials of life more abundant.

1.6.2 Chirality

The connection between extraterrestrial organic matter and terrestrial life is not proven, but the detailed chemical properties of amino acids found in meteorites and those found in life reveal a startling similarity. Molecules with identical chemical formulae can differ in how their atoms are arranged and these different structural arrangements are called **isomers**. Some isomers separate into left-handed and right-handed forms. This chemical property is **chirality** (pronounced 'ky–rality'). Life exhibits a preference when it comes to the use of chiral molecules.

What is chirality?

If you hold a plain ball up to a mirror, the mirror image of the ball will be identical to the ball itself. It is easy to imagine taking the reflected image of the ball and superimposing it on the real item (Figure 1.22a). However, mirror images are not always identical. For example, place your hands side-by-side in front of you with your palms facing up. You will see that your hands are mirror images of each other but they are not the same (Figure 1.22b). It is for this reason that we need right-hand and left-hand gloves. Try laying one of your hands on top of the other, again palms up. You will see that the thumbs and fingers do not lie in the same position on both hands. In other words they are non-superimposable. Any object whose mirror image is non-superimposable, such as a hand, will be chiral. In fact, the word *chiral* is Greek for 'hand-like'. Any object whose mirror image is superimposable, such as a ball, is termed 'achiral'.

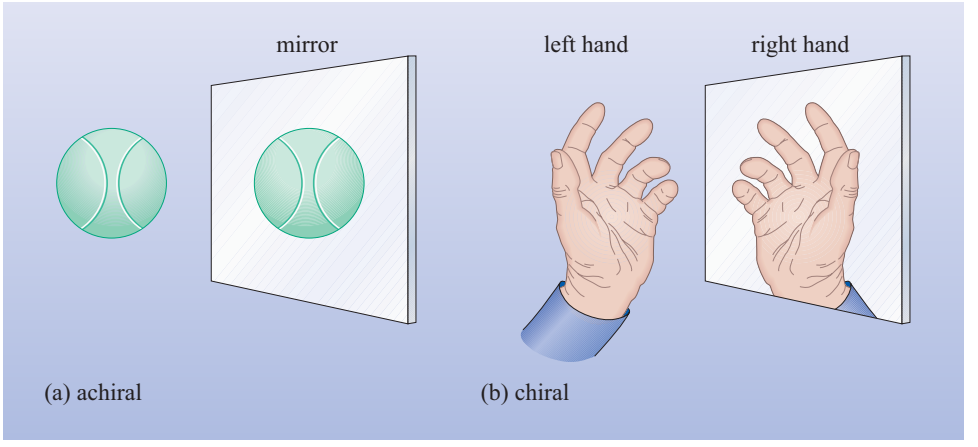


Figure 1.22 (a) The mirror image of a ball is superimposable on the original object and is therefore achiral. (b) The mirror image of a hand is not superimposable on the original and is therefore chiral.

■ What other human appendages are non-superimposable or chiral?

❑ Feet are chiral objects. If you need convincing of this point, try to put your left shoe on your right foot. Ears are also chiral.

Molecules are equally capable of exhibiting achirality and chirality. When all the carbon atoms in a molecule have less than four different structures attached to them then the molecule will be superimposable on its mirror image and will be achiral. For example, the molecule in Figure 1.23a has a mirror image that can be superimposed on the original molecule. However, whenever a molecule has a carbon atom that has four different structures bonded to it, it will not be superimposable on its mirror image and will be chiral. For example, the mirror image of the molecule in Figure 1.23b cannot be superimposed on the original molecule.

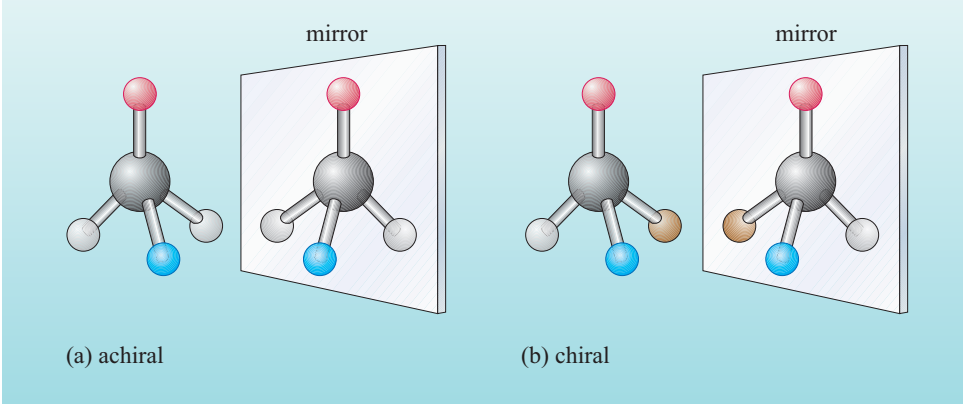


Figure 1.23 (a) An achiral molecule. Its mirror image can be superimposed on the original. (b) A chiral molecule. The mirror image cannot be superimposed on the original.

When discussing chiral molecules it is usual to talk in terms of left-handed and right-handed forms. The right-handed forms are often abbreviated to D (dextro) and the left-handed forms to L (levo).

We can recognize achiral and chiral molecules in amino acids. The simplest amino acid glycine (Figure 1.24a) has a superimposable mirror image and is achiral. More complex amino acids such as alanine (Figure 1.24b) have a non-superimposable mirror image and are chiral. In other words, alanine can be present in left-handed and right-handed forms (Figure 1.25).

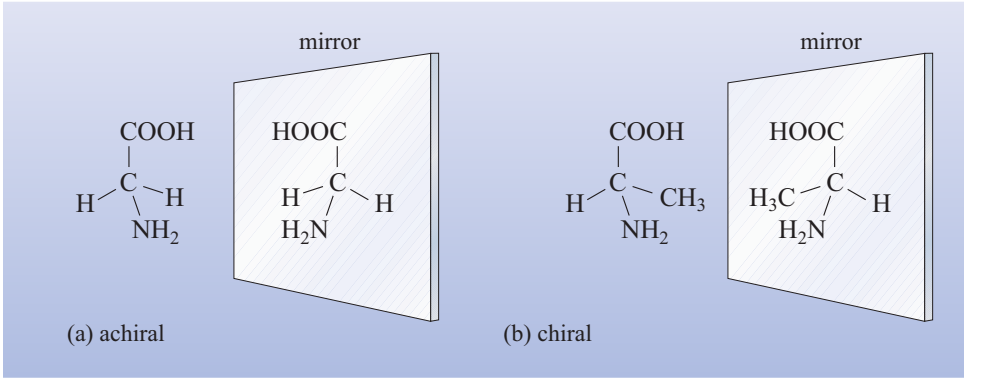


Figure 1.24 (a) The achiral amino acid glycine. Its mirror image can be superimposed on the original. (b) The chiral amino acid alanine. The mirror image cannot be superimposed on the original.

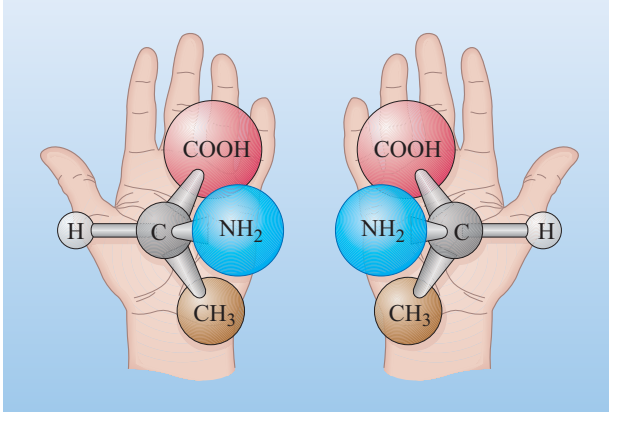


Figure 1.25 The chiral amino acid alanine can be present as left- and right-handed forms.

Chirality and life

In the absence of life, the chemical reactions that make amino acids generally create equal numbers of left- and right-handed forms and such assemblages are called **racemic** mixtures. Yet what is remarkable is that all life on Earth uses only the left-handed forms of amino acids for producing proteins. Proteins are found in every living thing and there can be no life without them. Proteins are constructed from a number of amino acids, and of the 20 amino acids used to make proteins, 19 can exist as left-handed and right-handed forms. The exception is the simple molecule glycine (Figure 1.24a) which has a central carbon atom with fewer than four different structures attached and therefore is achiral.

Mixing left- and right-handed forms of amino acids in proteins would produce structures that hindered the proteins from performing their biological functions. So at some point in the Earth's history, life must have begun with the left- rather than the right-handed amino acids. Once underway, biology was locked into this preference.

The origin of chirality

For many years it was thought that life's use of left-handed molecules was simply a result of an original random chance on the early Earth. However, recent research on organic-rich meteorites has suggested that this preference was inherited from life's starting materials that may have come from space. It has been discovered that an excess of the left-handed form of amino acids is present in meteorites, such as the Murchison meteorite, indicating that these amino acids existed in Solar System material before there was life on Earth.

It is thought the culprit for generating a left-handed excess in extraterrestrial organic molecules is a particular type of starlight called **ultraviolet circularly polarized light**, or UVCPL for short. With UVCPL, the electric field direction rotates along the beam. As rotation can occur in a left- or right-hand direction, UVCPL is a chiral phenomenon. Chiral substances have different absorption intensities for left and right UVCPL. Since photolysis (destruction by light) occurs only when light photons are absorbed, UVCPL light destroys one form of the molecules more readily than the other form.

Polarized light has light waves that have electromagnetic vibrations in only one direction. Ordinary light vibrates in all directions perpendicular to the direction of propagation.

Recent work has detected large amounts of UVCPL from the star-forming regions of Orion (Figure 1.26) which will be imposing a chiral preference on any abiotic molecules present there. As this part of the Orion nebula is a site for star formation, chiral amino acids may be available in the soon to be created star and planetary systems. It is tempting to imagine our own Solar System forming in such a region of circularly polarized light, ultimately leading to the excess of left-handed amino acids that we see in meteorites.

It appears therefore that life's left-handed preference may have originated by the action of UVCPL on chiral molecules. When the Solar System formed, some of this organic material arrived on the early Earth via impacts of comets, meteorites and dust particles. These molecules were then part of the **prebiotic** material available for the origin of life, and may have tipped the scales for life to develop with left-handed amino acids. Yet some scientists have suggested that extraterrestrial environments may not only have delivered organic matter to the early Earth but also viable micro-organisms themselves, a theory termed **panspermia** (Box 1.3).

Figure 1.26 Reflection nebulae in Orion, a good source of UVCPL. At the top of the picture is a loose grouping of bright stars. The fronds beneath these stars are the reflection nebulae.

BOX 1.3 PANSPERMIA

In 1908 a Swedish chemist, Svante Arrhenius (1859–1927), published a book which contained the proposal that life in the form of spores could survive in space and be spread from one solar system to another. Spores drifting in the upper atmosphere of a life-rich planet would be forced into interstellar space by the pressure of a nearby star's radiation. Eventually, some of the spores would fall upon another planet where they would flourish and the process would begin again. However, the theory soon attracted criticism because of the large doses of fatal radiation that would have been encountered during a lengthy trip through space.

William Thomson (Lord Kelvin) (1824–1907) proposed a variation on panspermia in which spores are carried through space by meteorites. Although it is unknown whether impact events could launch rocks between solar systems, transport to and from worlds within solar systems occurs frequently. Evidence for interplanetary transport is present in the form of meteorites on Earth that have come from the surfaces of Mars and the Moon. The most likely candidates for a source of organisms in the early Solar System are Mars and Venus. Both have hostile surface environments at the present day but over 3 Ga ago the situation was likely to have been very different and the exteriors of both worlds more hospitable to life.

In 1996 NASA scientists made the highly controversial announcement that a Martian meteorite contained fossil evidence of alien micro-organisms. The meteorite, Alan Hills (ALH) 84001, has a complex and fascinating history:

4.5 Ga	crystallized from magma on Mars
4.0 Ga	battered but not ejected by an asteroid impact
3.6–1.8 Ga	altered by water to produce carbonate minerals
16 Ma	blasted into space by an asteroid impact
1984	discovered in Antarctica
1996	NASA announced the discovery of Martian life

Although the claims of fossil life in (ALH) 84001 have now largely been discounted, the concept that micro-organisms could be transported between planets has begun to attract serious scientific attention.

1.6.3 Keeping your concentration – organic matter accumulation

The amount of organic material estimated to have fallen to Earth per hundred million years around the time of the origin of life is 10^{16}–10^{18} kg. If this material were spread across the surface of the Earth, it would form a layer ranging from 1.6 cm to 1.6 m in thickness. Although this amount of material would represent a significant source of organic carbon in the prebiotic environment if it all survived and accumulated, most of the cometary and meteoritic infall surviving atmospheric entry would presumably fall into oceans and be buried in sediments.

EXAMPLE

Calculate how much organic carbon would be delivered to the Earth each 100 Ma using the present-day rates in Table 1.7. Explain any differences between your calculated value and the value given for the early Earth.

SOLUTION

Present day rate of organic matter accumulation

$= (0.32 \times 10^{6}) + (2.9 \times 10^{-6}) + (4.7 \times 10^{-6})$

$= 0.320 \times 10^{6}$ kg yr^{-1}.

Over 100 Ma the amount of organic matter delivered to Earth = 3.20×10^{13} kg.

These rates are less because the early Earth was in the middle of the late heavy bombardment.

As stated earlier, the rate of a chemical reaction increases with the concentration of reactants. So some form of concentration mechanism would have been necessary for the organic matter arriving in the form of meteorites, meteors and comets to take part in chemical reactions that may have led to the production of a living system. After all for two molecules to react, they have to come into contact. Several possible concentration mechanisms can be considered.

1 Marginal marine environments such as lagoons or tidal pools provide a means of concentrating dilute solutions. The solutions are temporarily cut off from the main ocean and evaporation causes the residual liquid to contain a higher proportion of organic molecules.

2 Freezing an aqueous solution also causes an increase in the concentration of any dissolved organic compounds because the water freezes first.

3 The surfaces of clays and other minerals provide sites for trapping organic matter. Clays in particular are useful minerals as they can accommodate organic molecules on and within their structure.

Concentration processes are important for providing enough localized raw materials for the creation of primitive living systems – in much the same way as it is impossible to build a house if the bricks and mortar are repeatedly scattered over a wide area. Once the raw materials required for the origin of life were in place, the serious work of construction could begin.

1.7 Achieving complexity

The mechanisms by which organic molecules could have been assembled into complex living organisms are very poorly understood. Yet it is interesting to examine how these fundamental steps may have occurred. In this section you will encounter several processes that could have aided the development of increasingly complex organic systems that, eventually, became the direct forerunner of life.

1.7.1 The ties that bind – creating polymers and macromolecules

The data in Table 1.2 indicate that macromolecules are important to life. Initially, this appears to be a cause for concern for our ability to understand the origin of life as great complexity appears necessary for living systems to function. It is appropriate therefore to consider just how difficult it is to generate macromolecules from simple organic molecules.

■ Carefully examine Figures 1.5 and 1.6 and state how the production of polymers from simple monomers occurs.

❑ The polymerization of monomers occurs by a reaction involving the loss of a water molecule.

For example, the –OH groups from two sugar monomers can combine to form a bond following the release of a water molecule (Figure 1.5). Similarly, the combination of the $-NH_2$ group and –COOH group of two amino acids also forms a bond following the release of water (Figure 1.6). Similar reactions also occur during the polymerization of nucleic acids.

■ Following the formation of a bond between sugar monomers in Figure 1.5, are all of the reactive –OH groups used up?

❑ No, the ends of the new larger molecule still contain reactive OH groups.

In a similar fashion, following the formation of a bond between two amino acids, the ends of the new larger molecule still contain reactive $-NH_2$ and –COOH groups. These features ensure that polymerization reactions can continue indefinitely to form larger and larger organic structures.

So large molecules can be generated by simple polymerization reactions that involve the loss of water. The simplicity and efficiency of this process indicates that it is highly likely that most primitive polymerizations occurred in this way. The high level of order and complexity of some of the macromolecules used in living systems may require more sophisticated methods, and life makes good use of the catalytic properties of protein enzymes. Enzymes make the chemistry of life more efficient, but with our simple water-loss reactions we have at least made a start towards organic complexity.

1.7.2 Formation of boundary layers

As we discussed in Section 1.2.6, all life today has cells, tiny packets of chemicals surrounded by membranous boundary layers. Hence, a question we musk ask is how did cellular life arise? There were no large molecules like nucleic acids and proteins available on the prebiotic Earth to control the assembly processes characteristic of life. So the first forms of life must have arisen through a self-assembly process.

Let us explore how such processes operate. We have talked about hydrophobic and hydrophilic compounds but certain kinds of organic compounds have both properties at either end of the molecule: a polar hydrophilic head and a hydrophobic tail. These compounds are said to be **amphiphiles**, which means they 'love both'. The polar heads carry a small electrical charge, which makes them soluble in polar solvents such as water. The uncharged tails are much less soluble in water.

If amphiphilic molecules are added to water they tend to sit at the surface with the hydrophilic heads in the water and the hydrophobic tails in the air. In this way they can create a single layer of molecules – a monolayer, as shown in Figure 1.27. The **monolayer** is just one molecule thick, so can be thought of as a two-dimensional surface, or membrane.

- What form do you think amphiphilic molecules will take when introduced within the water by shaking the mixture?
- They tend to gather together and form small spherical structures where the hydrophilic heads face the water while the hydrophobic heads are tucked inside, shielded from the water (Figure 1.28). These spherical structures are termed '**micells**'.

Larger collections of amphiphilic molecules can form a double-layer structure. This arrangement is called a **bilayer** (Figure 1.29a). Imagine two sheets of molecules sandwiched together so that all the hydrophilic heads are on the outside (in contact with the water) and the hydrophobic tails are inside (away from the water). Just

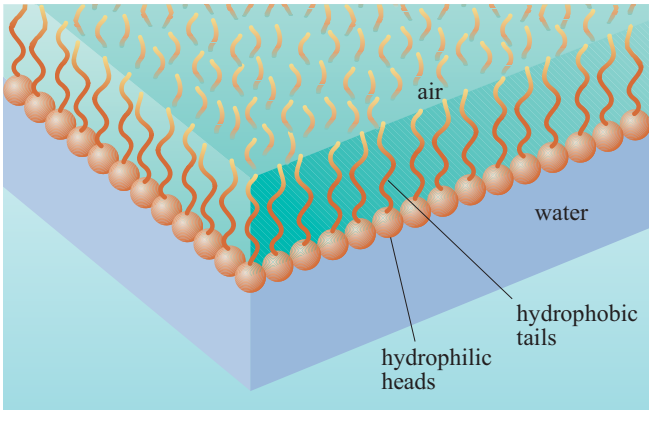


Figure 1.27 A lipid monolayer – a single layer of lipid molecules with hydrophilic ends in the water and hydrophobic ends in the air.

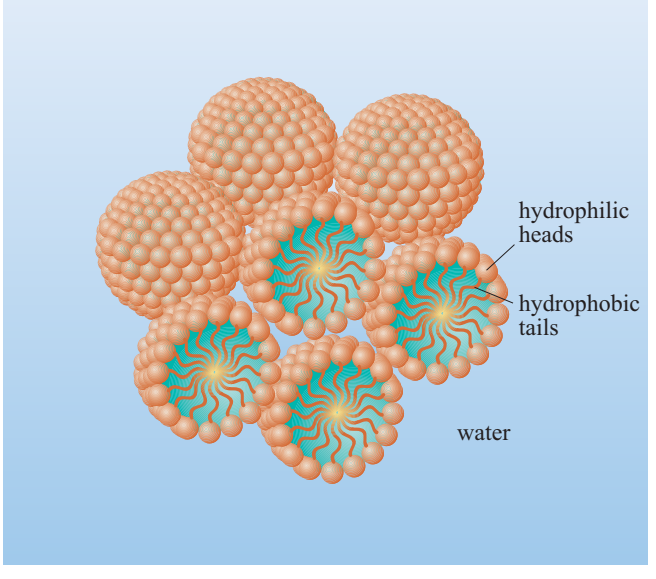


Figure 1.28 A lipid micell – spherical collections of lipids with hydrophilic heads on the exterior and hydrophobic tails in the interior.

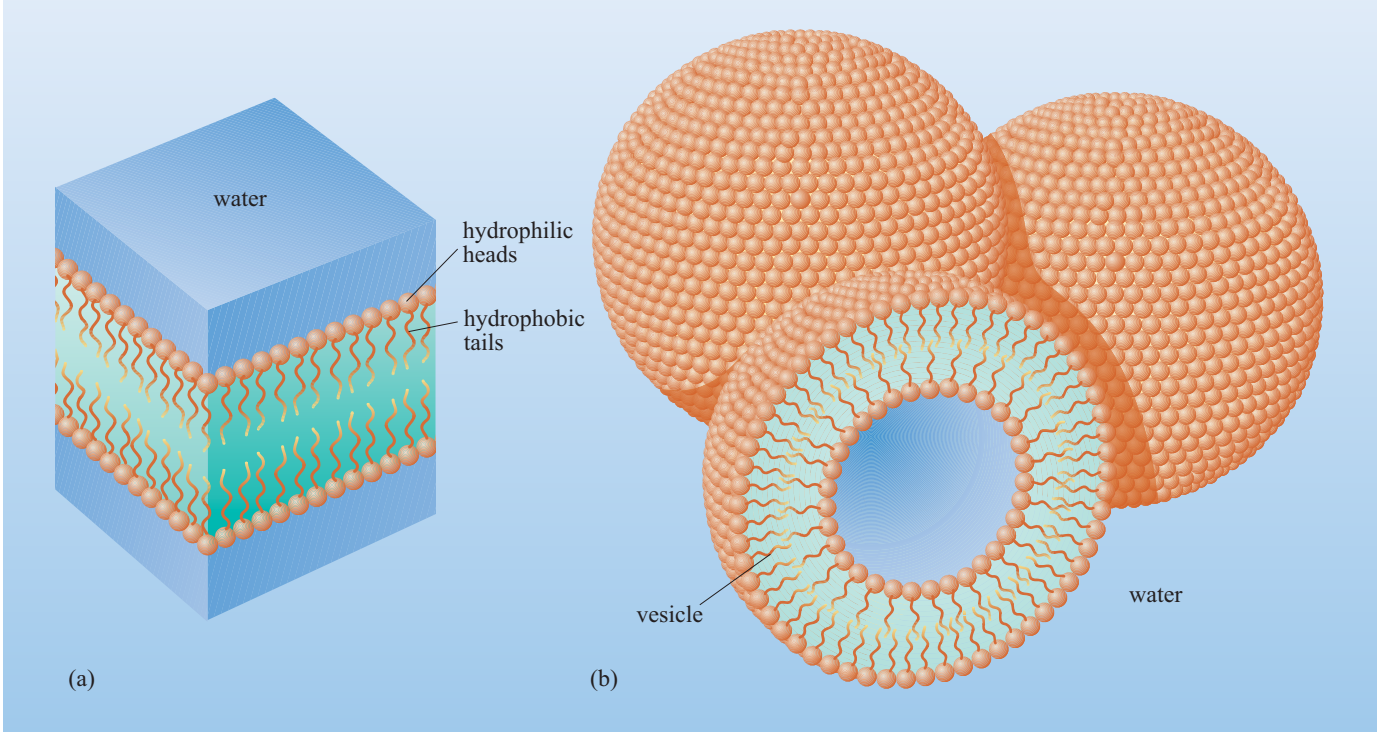
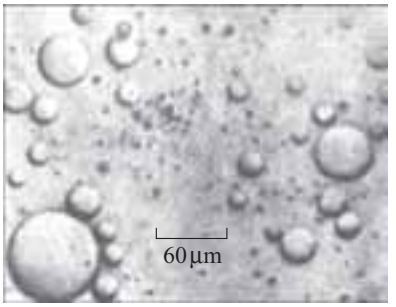


Figure 1.29 (a) A lipid bilayer and (b) bilayer vesicle.

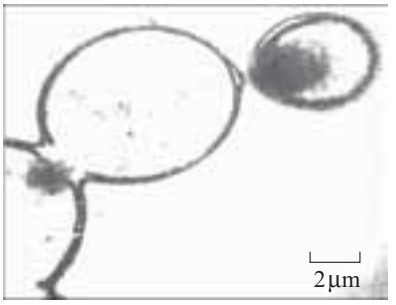
like monolayers, bilayers are a form of membrane and while the spherical form of a single layer of molecules is called a micell the double layer equivalent is termed a '**bilayer vesicle**' (Figure 1.29b). Bilayers are particularly interesting because many organisms use this type of membrane to preserve the integrity of their cells. Monolayers, micells, bilayers and bilayer vesicles are structures that form spontaneously and perhaps provided the original membrane-bounded environment required for cellular life to begin.

The importance of membranes to primitive life has led to a number of experimental investigations aimed at determining how they could have formed under prebiotic conditions. In 1924, Alexander Oparin (1894–1980), a Russian biochemist, showed that proteins, when added to water, group together to form droplets (Figure 1.30a). These droplets are called **coacervates**, a name derived from the Latin for clustered or heaped. Coacervates form in solutions of many different polymers, including proteins, nucleic acids and polysaccharides. Coacervation is a property of physical chemistry related to the polarity of molecules and their ability to form hydrogen bonds in water. Many substances when added to the coacervate preparation can become preferentially incorporated into the droplets, providing a means by which prebiotic chemical factories could have been constructed.

In 1958, Sidney Fox heated dry mixtures of amino acids causing their polymerization by reactions involving the loss of water. The amino acid polymers resembled proteins and so Fox called the new molecules 'proteinoids'. When these proteinoids were dissolved in hot water and then the solution cooled, the proteinoids formed small spheres about 2 μm in diameter which Fox termed '**microspheres**'. The microspheres displayed a double wall resembling a biological membrane and could shrink or swell, depending on the salt concentration of the water. If left for several weeks the microspheres absorbed more proteinoid material from the solution and produced buds which occasionally separated to form second generation microspheres (Figure 1.30b).



(a)



(b)

Figure 1.30 (a) Coacervates and (b) proteinoid microspheres.

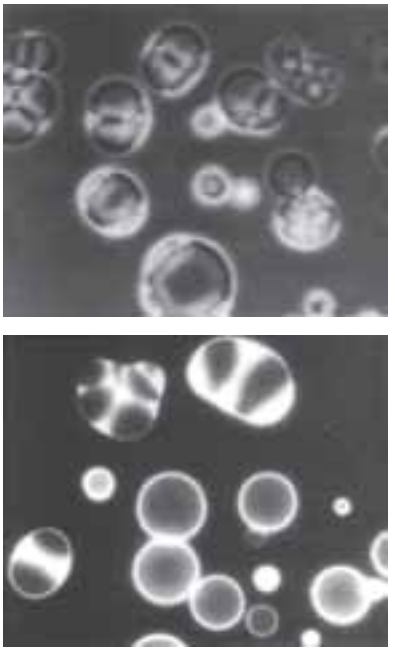


Figure 1.31 Bilayers generated from the Murchison meteorite

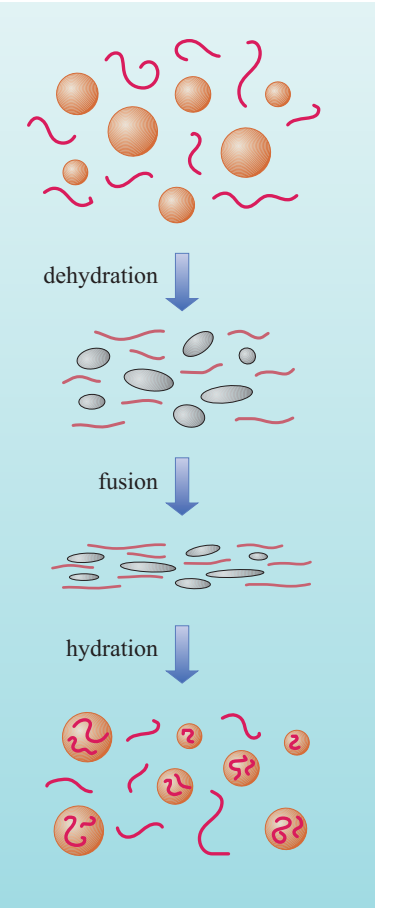


Figure 1.32 The dehydration and incorporation of molecules, and rehydration of membranes.

In 1985 David Deamer considered that amphiphilic molecules might have been delivered to the early Earth through extraterrestrial infall. He extracted organic matter from the Murchison meteorite and added it to water to explore the ability of meteoritic organic matter to form boundary layers. The Murchison molecules formed membrane-bound bubbles (Figure 1.31) providing strong evidence that, on the early Earth, mixtures of abiotic organic compounds could have helped to form membranes for primitive cellular life.

You can imagine how these membranes could have been mixed with a collection of molecules and then dehydrated on the early Earth. When hydrated again small chemical factories may have developed into primitive cells (Figure 1.32).

1.7.3 The role of minerals

The development of the first living system must have involved a sequence of chemical transformations which achieved a greater level of structure and complexity than the available starting materials. Many believe that minerals served a number of critical functions in this process. We can identify four key roles that minerals could have played in the origin of life: protection, support, selection and catalysis. Let's look at each of these in turn.

Minerals could have acted as hosts for assembling chemical systems thereby *protecting* them from dispersal and destruction. For example volcanic rock contains many small air pockets created by expanding gases while the rock was still molten (Figure 1.33) and some common minerals develop microscopic pits following weathering. Tiny compartments such as these could have housed small chemical mixtures which may have taken the first steps towards organized life.

Figure 1.33 Volcanic rock containing small air pockets that were created by expanding gases while the rock was molten.

Figure 1.34 Iron sulfide, which may have acted as a template, catalyst and energy source for the production of biological molecules.

The surfaces of minerals could have acted as a *support* structure for molecules to accumulate and interact. An effective way to assemble molecules from a dilute solution is to concentrate them on a flat surface. Experiments have been performed where solutions containing amino acids are evaporated in a vessel containing clays. The amino acids concentrate on the clay surfaces and then polymerize into short protein-like chains.

Minerals may have aided the *selectivity* of certain biologically useful molecules. Many have crystal faces that are mirror images of each other. Minerals such as calcite bond strongly to amino acids and when calcite crystals are submerged in racemic solutions of amino acids the left- and right-handed forms of the amino acids bond to different crystal faces. It is plausible that under the right conditions this selection and concentration process allowed protein-like molecules to form which were exclusively right or left handed. At some point natural selection chose the left-handed molecules and all subsequent life inherited this trait.

Minerals can act as *catalysts*. One of the elements that is required to generate biologically useful materials is nitrogen. However, although nitrogen is abundant in the Earth's atmosphere it is present as unreactive nitrogen gas. Primitive organisms must have found a way of converting nitrogen gas to a form assimilable by life. In industrial processes nitrogen and hydrogen are passed over metallic surfaces to generate ammonia. If similar reactions took place on the early Earth, the ammonia would have been a valuable source of nitrogen for biological reactions. This may have occurred in **hydrothermal vents** where nitrogen and hydrogen are passed over iron oxide surfaces.

Perhaps all of these functions operated simultaneously on the same mineral. For example, in 1988 Gunter Wächtershäuser, a German patent lawyer, suggested that iron and nickel sulfides could have served as a template, catalyst and energy source for the production of biological molecules (Figure 1.34). He took his proposal further by suggesting that the first living things may have been coatings stuck to the surfaces of these crystals. If the mineral acted as a catalyst it is possible that primitive metabolism proceeded in the absence of enzymes.

1.8 From chemical to biological systems

1.8.1 The RNA world

Recall from Section 1.2 that life as we know it uses DNA as a store of genetic information and RNA as the messenger that carries the information out into the cell. Importantly, nucleic acids have the genetic information necessary to reproduce themselves but need proteins to catalyse the reaction. Conversely, proteins can catalyse reactions but cannot reproduce without the information supplied by nucleic acids. This poses us with a dilemma in our bottom-up study of the origin of life – which one of these three key molecules (DNA, RNA and protein enzymes) could have existed without the other two? This has been called the 'chicken and egg' paradox.

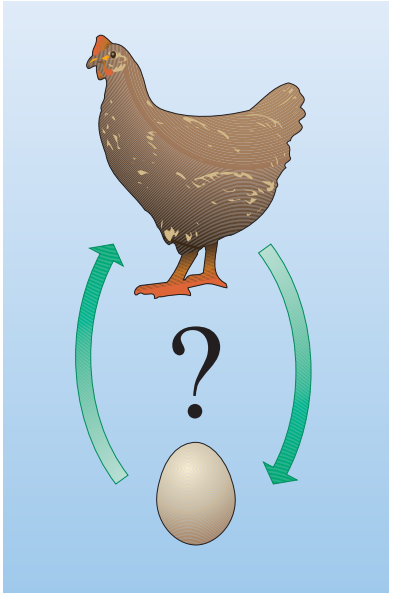


Figure 1.35 The 'chicken and egg' paradox. Which came first, proteins which catalyse the production of nucleic acids, or nucleic acids that contain the genetic information needed to produce proteins?

In the mid-1980s it was discovered that RNA, unlike DNA, can perform some of the enzymatic functions needed for replication. The evidence came in a discovery made independently by the US biochemists Sidney Altman and Thomas Cech, which led to them being jointly awarded the Nobel Prize for Chemistry in 1989. RNA molecules that have catalytic properties similar to enzymes are called 'ribozymes'. In principle, because RNA molecules could store genetic information and act as catalysts, they would make proteins unnecessary for simple life. So a simpler 'RNA world' may have preceded the DNA-plus-protein world of today. RNA might have been able to replicate and evolve without specialized proteins – there are several observations that support this proposition:

- The nucleotides in RNA are more readily synthesized than the nucleotides in DNA.
- It is easy to imagine that DNA evolved from RNA and then, on account of its greater stability, DNA took over the RNA role.
- RNA is likely to have evolved before proteins because no plausible scenario can be envisaged where proteins can replicate in the absence of RNA.

Eventually, the evolving RNA organisms began transcribing DNA, which was a much more efficient replicator. The ability for RNA to create DNA is vividly illustrated by retroviruses (Box 1.4). Natural selection then saw to it that the more proficient DNA-plus-protein world outcompeted its parent RNA.

BOX 1.4 RNA AND RETROVIRUSES

A virus is a parasite without a cellular home. Viruses are fragments of nucleic acid within protein coats and are small (10–200 nm). When without a host, virus particles do not carry out the functions of living cells, such as **respiration** and growth. Yet once within a host cell they steal the cell's chemical energy and hijack its ability to synthesize protein and nucleic acids in order to replicate themselves.

Some viruses do not kill the host cells but persist within them in one form or another. Cancer-causing viruses and the human immunodeficiency virus (HIV) are of this type and evolve at about a million times the rate of nuclear DNA. These are 'retroviruses'. They reverse the normal cellular process of transcribing DNA into RNA: they multiply by transcribing RNA into DNA, which then takes over the cellular machinery to make more viral RNA.

The ability of RNA to transcribe DNA is a valuable piece of supporting evidence for the proposal of a past RNA world. It is tempting to consider retroviruses as the legacy of our darkest ancestors that can still wreak havoc in our modern complex biosphere.

1.8.2 Primitive biochemistries

In Section 1.5.1 we discussed the possible sources of energy available for life. These energy sources are captured by life and then utilized via metabolic processes. The most important source of energy on the Earth today is sunlight and this energy is captured by a process called **photosynthesis**. Photosynthesis is the production of carbohydrates from water and carbon dioxide. A photosynthetic reaction with a small energy barrier is one based on sulfur.

$$n\mathrm{CO_2} + n\mathrm{H_2S} \longrightarrow (\mathrm{CH_2O})_n + n\mathrm{H_2O} + n\mathrm{S} \tag{1.1}$$

Yet the photosynthesis reaction used most commonly today by plants is based on water.

$$n\mathrm{CO_2} + n\mathrm{H_2O} + \text{energy} \longrightarrow (\mathrm{CH_2O})_n + n\mathrm{O_2} \tag{1.2}$$

When organisms are in environments where sunlight is unavailable, alternative mechanisms must be employed to generate organic compounds. For example, in 1977 scientists in the submersible *ALVIN* were studying a mid-ocean ridge near the Galapagos Islands in the Pacific Ocean. They discovered underwater volcanoes with deep-sea hydrothermal vents (Figure 1.36) that were populated with a range of organisms. Seawater circulating through new, hot ocean crust at mid-ocean ridges is heated, and this hot seawater dissolves and exchanges chemicals with the rock. In some places along the ridges, this mineral-rich hot water vents back into the sea at temperatures of up to about 400 °C and only the great pressures generated by the column of seawater above stops the water from boiling. These sources of mineral-rich hot seawater support communities of organisms in which life depends not on light energy, but on chemical energy. The synthesis of organic matter in this way is called **chemosynthesis**. Today we know that hydrothermal systems are not restricted to mid-ocean ridges. They also occur deep in the Earth's crust where water and heat are present. Recently, it has been discovered that these hydrothermal regions also host simple chemosynthetic life forms. These ecosystems are now termed the 'deep hot biosphere' and it appears that the amount of organic matter deep in the Earth may actually rival that at the surface. You will encounter some of these unusual subsurface dwellers again in Section 2.5. Organisms that utilize photosynthesis and chemosynthesis to generate organic compounds from energy and simple inorganic substances are termed **autotrophs** (from the Greek *auto* meaning 'self' and *troph* meaning 'feed').

The carbohydrates generated by the energy-capturing processes of photosynthesis and chemosynthesis are used to generate energy-rich phosphate bonds. This stored chemical energy can then be tapped by the organism when needed. **Fermentation** and respiration are the two most common forms of metabolism on the Earth today. In fermentation, the carbohydrate glucose ($\mathrm{C_6H_{12}O_6}$) is transformed into both $\mathrm{CO_2}$ and ethanol ($\mathrm{CH_3CH_2OH}$), or lactic acid ($\mathrm{C_3H_6O_3}$) for a net gain of two energy-rich phosphate bonds. In respiration, the free oxygen in the Earth's atmosphere is used

Figure 1.36 A deep-sea hydrothermal (hot water) vent.

to extract more energy from the glucose molecules than is obtained by fermentation. The glucose is transformed into CO_2 and water with a net gain of 36 energy-rich phosphate bonds – a significant improvement.

$$(CH_2O)_n + nO_2 \longrightarrow nCO_2 + nH_2O + \text{energy} \qquad (1.3)$$

The metabolic mechanisms outlined above allow organisms to capture and store energy for use in the complex chemical reactions that keep biochemical systems operating. But what of organisms that seek to obtain their sustenance by consuming autotrophs and taking advantage of the hard work they have performed? These organisms are termed **heterotrophs** (from the Greek *hetero* meaning 'different').

1.9 The top-down approach – molecular phylogeny

Up to now, we have focused on the bottom-up approach to understanding the origin of life, that is attempting to build life from scratch. Now we turn to the top-down approach. We will try to extrapolate as far back as we can go to the origin of life by using information contained in the Earth's extant biology. Life on Earth has a history that extends back over almost 4 Ga and it has long been believed that this evolutionary history started with a single and simple common ancestor. Expressed another way, all life on Earth is related. Darwin subscribed to this idea and, in 1857, expressed the view that a time would come 'when we shall have very fairly true genealogical trees of each great kingdom of nature'.

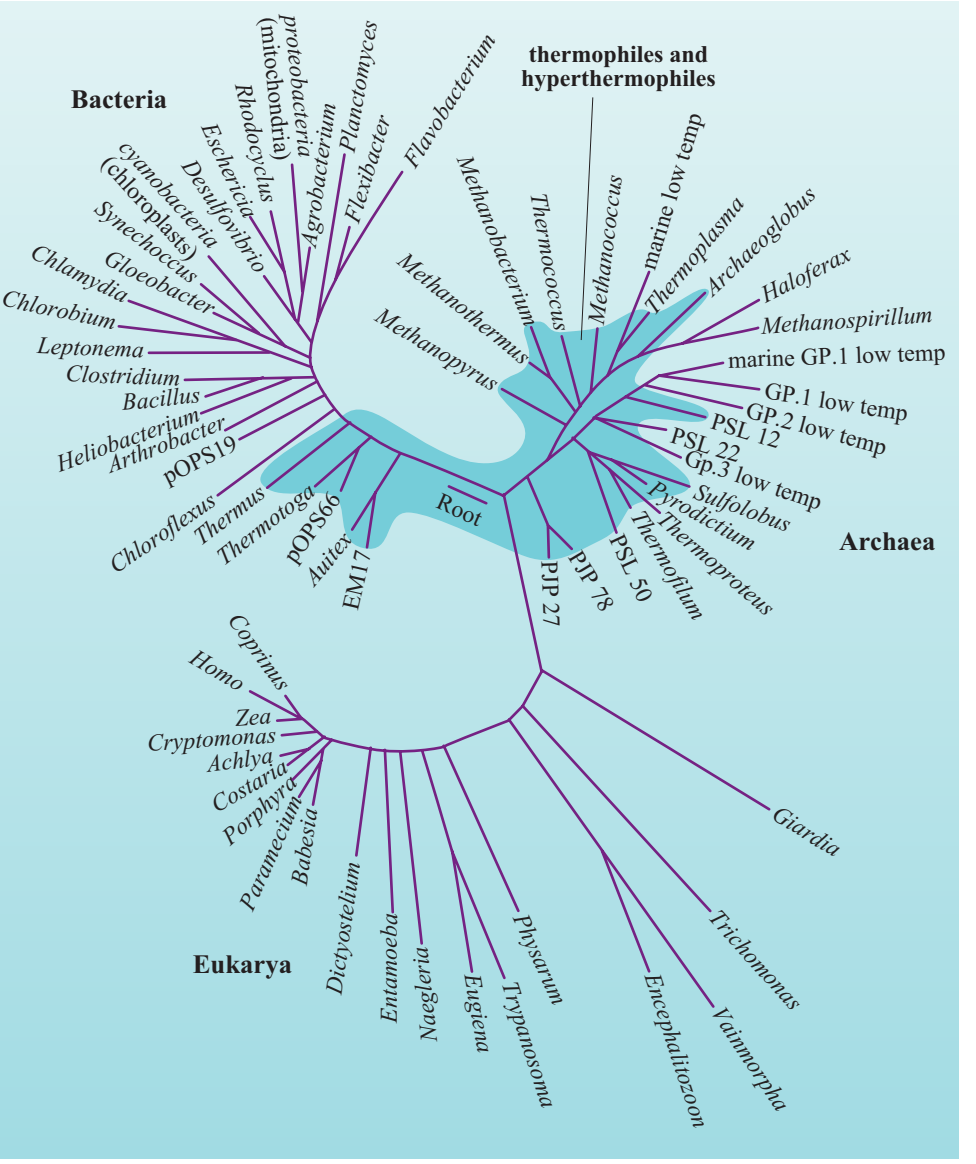


Figure 1.37 The universal tree of life based on ribosomal RNA data. Lengths of lines separating pairs of organisms correspond to the genetic differences between them. Life is divided into three domains (Bacteria, Archaea and Eukarya) as a result of constructing this tree. Deep, short branches indicated in the centre of the tree are populated by thermophilic and hyperthermophilic organisms.

Today Darwin's dream has been realized with the advent of genetically based **phylogenetic trees**. Life, it seems, does not reject what evolution has created, but simply builds on what has gone before. The biological record of this continuous addition and modification is present in genetic material, namely the sequence of nucleotides in RNA and DNA. So the basic method involved in constructing a phylogenetic tree is to examine similar molecules in different creatures and, if parts are found to be alike, then those parts must have been inherited by organisms from a common ancestor.

One of the more useful trees is built on the genetic information obtained from small sub-units of ribosomal RNA and comparisons between different organisms reveal a hierarchy of evolutionary innovation (Figure 1.37). The longer a branch, the greater the difference in ribosomal RNA sequences between the organisms at the start and the end. Three clear domains are evident: Bacteria, Archaea and Eukarya. Within the three domains, the branches with names associated with them refer to species, or groups of species, more closely related to each other than to other groups on the tree. Following these branches back to the points where they join other branches leads to the ancestral species of the named groups. Branches that lead back to the same main branch represent species that share a common ancestor. When studying the origin of life it is the *root* of this phylogenetic tree that interests us most.

Note that the organisms closest to the centre of the tree, those that populate the deepest and shortest branches, are the **thermophiles** and **hyperthermophiles**. These are heat-loving microscopic organisms found near hot springs and deep-sea hydrothermal vents. You will study these in more detail in Section 2.5.

- What does the occurrence of thermophiles and hyperthermophiles at the centre of the phylogenetic tree imply about the course of evolution of life on Earth?
- One interpretation of the ribosomal RNA tree is that the course of evolution has generally moved from high to low temperatures.

Another important feature of the ribosomal RNA tree is that the majority of the deepest branching organisms do not use light as an energy source. This suggests that photosynthesis may be a later development than processes utilizing geochemical energy sources. So can we use the tree to draw conclusions about the nature of the earliest life on Earth?

The phylogenetic tree seems to be telling us that our **last common ancestor** may have been similar to heat-loving chemosynthetic organisms that populate hydrothermal vents today.

Yet the term 'last common ancestor' highlights the fact that this was not necessarily the first organism on Earth.

It is possible that life began in a low-temperature environment then through adaptation colonized deep-sea hydrothermal systems. Once the hydrothermal environments were colonized, an ocean-sterilizing impact could have wiped out all but the heat-loving vent life which was then free to evolve and repopulate the now vacant cooler environments.

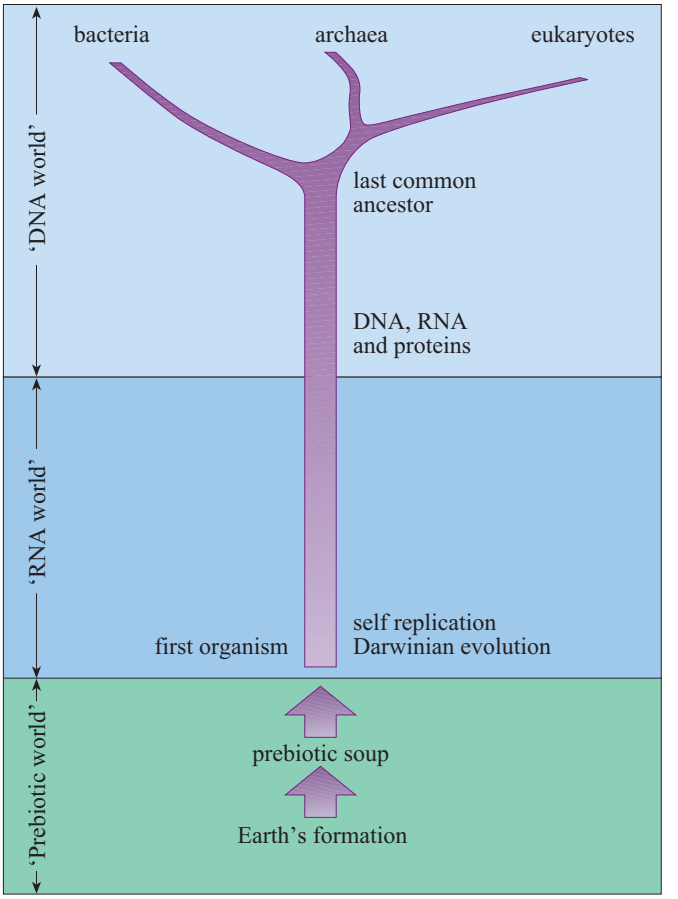


Figure 1.38 A synthesis of information from our bottom-up and top-down approaches to the origin of life.

1.10 A synthesis on the origins of life

In this chapter we have examined what life is and what it requires to exist. We have also attempted to construct plausible scenarios for the origin of life on Earth. But just how close have we come to understanding how life on Earth actually started? Figure 1.38 is a synthesis of the various lines of evidence we have put carefully into place. You will see that information from both the bottom-up and top-down approaches have been incorporated into the scheme. Our summary begins at the bottom, with the Earth's formation and the generation of a prebiotic soup, perhaps aided by the influx of organic matter from extraterrestrial objects. At some point in time, the prebiotic world produces the first living organism, possibly using minerals as catalysts or templates, and we are now in a biotic world in which self-replication and Darwinian evolution can take place. RNA probably played a significant role in primitive organisms, acting as both the store of genetic information and the catalyst for replication. Once DNA supplanted RNA as the main harbinger of genetic information the RNA world came to an end. Now, in the DNA world, the labour of life is shared between three molecules – DNA, RNA and proteins. No record of the first living organism or the first DNA-using organism remains so we must now take a giant leap to meet the top-down record provided by molecular phylogeny. The last common ancestor of life on Earth was probably a heat-loving organism similar to those found today at deep-sea hydrothermal vents.

1.11 Summary of Chapter 1

- Life can be described as a system that has the capacity to undergo self-replication and evolution. However, any definition of life is likely to be inadequate in specific circumstances.
- All known life is based on water and carbon. The constituent elements of life, the so-called biogenic elements, are abundant in the Universe.
- Life utilizes the ability of carbon to bond with many other elements and itself to create a variety of organic compounds that have specific biological functions. The main molecules of life are large macromolecules and include lipids, carbohydrates, proteins and nucleic acids.
- There are a number of extraterrestrial environments in which organic matter is produced without the aid of biological processes. These environments include the shells of carbon stars, molecular clouds and the solar nebulae.
- The organic products of extraterrestrial environments would have rained down on the Earth close to the time of the origin of life, adding to any organic matter synthesized on the Earth itself. Common characteristics, such as a left-handed preference in amino acid structures, appear to suggest a link between terrestrial biotic and extraterrestrial abiotic organic matter.
- Molecular complexity may have initially been achieved by simple chemical reactions although later, for greater efficiency, protein enzymes would have catalysed the reactions. Chemical reactions would also have been promoted by certain concentration mechanisms and the encapsulation of molecules within a membranous boundary layer.
- Early organisms are unlikely to have used the DNA, RNA and protein-based biochemistry common on the Earth today. It is more likely that RNA performed the functions of genetic information store and catalyst.
- The capture of energy by life involves the production of carbohydrates by autotrophic mechanisms. These carbohydrates are transformed into energy-rich phosphate bonds by metabolic processes such as fermentation and respiration.
- Molecular phylogeny indicates that the last common ancestor to all life on Earth was a heat-loving organism similar to those organisms that populate today's deep-sea hydrothermal vents.

CHAPTER 2
A HABITABLE WORLD

2.1 Introduction

In the previous chapter, you examined current theories as to how life on Earth might have originated from simple biogenic precursors. This raises the obvious question: why Earth? What, if anything, was special about conditions on the early Earth that enabled life to originate and evolve on this planet? Is an Earth-like planet essential for life to evolve elsewhere in the Universe? In this chapter you will examine what it is that makes the Earth a habitable planet, what conditions were like on the early Earth, and whether the life that has evolved on Earth can provide information as to the likelihood of life arising elsewhere in our own Solar System or beyond.

The Earth of today (Figure 2.1) is a very different place from the Earth that existed some 4.5 Ga ago shortly after the Solar System formed. Examining our planet from space, there are several pointers to the presence of conditions favourable to life and even the existence of life itself on its surface. One clue is the presence of liquid water, although this alone is not sufficient. However, a lot can be learned from a planet's colour, or more accurately the regions of the electromagnetic spectrum that the Earth reflects back into space. The Earth fluoresces in ultraviolet light, indicating that it has an atmosphere that contains almost 20% oxygen. This is demonstrated in Figure 2.2, a false colour image that illustrates how oxygen atoms in the Earth's atmosphere fluoresce as they absorb the Sun's ultraviolet radiation. Equally significant is the fact that green light (associated with complex organic molecules formed largely of carbon, in combination with lesser amounts of hydrogen, nitrogen, sulfur and other elements) is reflected from large areas of both land and sea (Figure 2.3).

The atmosphere is not a stable mixture of chemicals. Unless the carbon-rich compounds are continuously regenerated, nearly all of the organic material reflecting green light would have decomposed in the oxygen-rich atmosphere within a few hundred years.

Figure 2.1 From 4 million miles away on 16 December 1992, NASA's Galileo spacecraft, on its way to Jupiter, took this picture of the Earth–Moon system. The bright, sunlit half of the Earth contrasts strongly with the darker subdued colours of the Moon.

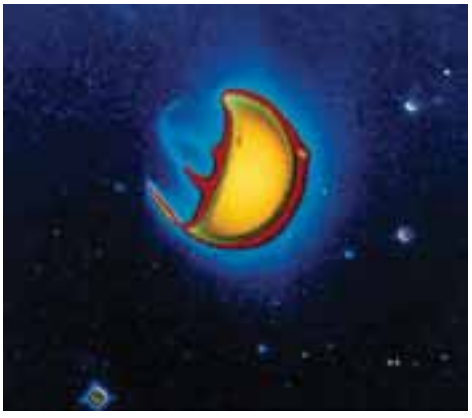


Figure 2.2 This false colour image shows how the Earth glows in ultraviolet (UV) light. Reacting to, and absorbing the Sun's ultraviolet radiation, oxygen atoms fluoresce, appearing in Figure 2.2 as a cloak of gold closest to the Earth's surface, where the oxygen gas lies heaviest. As it thins with altitude, oxygen is coloured green, red, and finally blue. This image was taken with the Far UV Camera/ Spectrograph deployed and left on the Moon by the crew of Apollo 16.

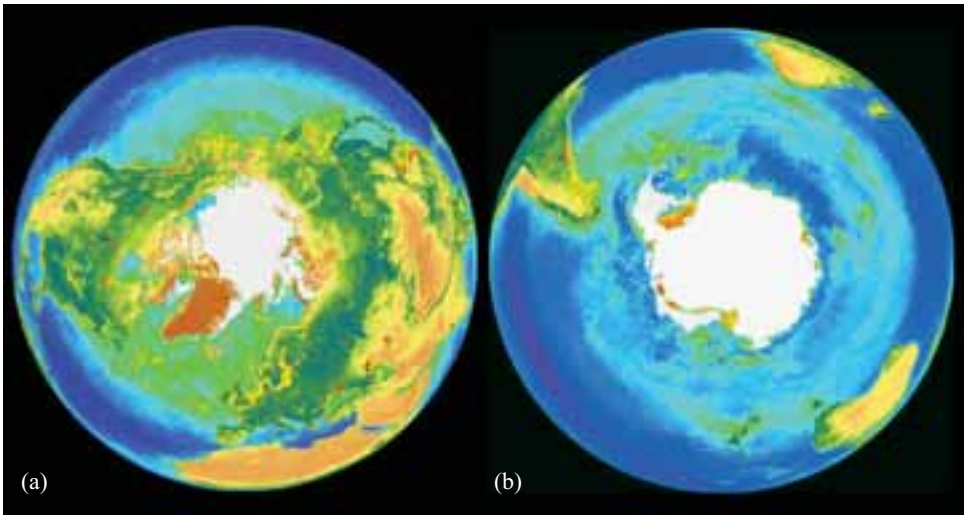


Figure 2.3 Oceans and life. This false colour image compiled from the SeaWiFS (Sea-viewing Wide Field-of-view Sensor) instrument on board the orbiting SeaStar satellite records reflectance spectra from the Earth's oceans. In these North (a) and South (b) Pole projections the colour depends on how sunlight is reflected by free-floating phytoplankton – photosynthesizing organisms that contain chlorophyll. Chlorophyll absorbs blue and red light and reflects green so that ocean areas with abundant plankton are shown in green as are land areas with significant vegetation.

This would remove some of the oxygen from the atmosphere while the rest would have been used up in chemical weathering reactions (oxidation) involving the rocks of the Earth's surface. It seems, therefore, that some process operating on the Earth's surface continually regenerates complex carbon-rich compounds and maintains the oxygen content of the atmosphere.

Of course, we know that there is life on Earth: the carbon-rich compounds make up the plants and animals and these in turn use oxygen for respiration. So can we use the Earth as a model for our search for life elsewhere in the Solar System and beyond?

2.2 Defining a habitable planet

Life on earth has managed to survive a number of major climate changes such as large-scale glaciations. However, on occasions, dramatic changes in the Earth's environment have caused mass extinction events, wiping out large numbers of species.

Of necessity we will adopt the Earth as the reference of a habitable terrestrial planet since it is the only example we have. At least within our own Solar System, it appears that the Earth's habitability may be near optimal, especially for complex life. We also need to be clear what we mean by habitable. The conditions needed to sustain Earth-like animal life that uses oxygen are quite different from the much broader range of conditions that can support microbial life. For the former a habitable terrestrial planet would require an ocean and some dry land, moderately high O_2 (and low CO_2) abundance, and a reasonably stable climate. The moderately high O_2 and low CO_2 is a requirement for large mobile life on physiological grounds and also for the production of an ozone layer to provide shielding from the effects of harmful ultraviolet radiation. The Earth's oceans effectively regulate the planet's temperature on a global scale via the operation of a water cycle, which interacts with processes such as plate tectonics, and chemical weathering on land. Earth's long-term climate stability results from many astrophysical and geophysical constraints, including stellar evolution, comet and asteroid impact rate, the presence of a large natural satellite, and a long-term planetary heat source to drive plate tectonics. However, larger plants and animals have only been around on Earth for the last 500 Ma. For over 3 Ga the Earth has supported microbial life which can survive and evolve under much more extreme conditions, a topic you will examine in greater detail in Section 2.5. Even these simple forms of life share some common requirements: the presence of liquid water and long-term environmental stability (i.e. environmental conditions that have never been so extreme as to extinguish all life).

2.3 Habitable zones

2.3.1 Water and light

We have identified two properties that sharply differentiate the Earth from other planets in our Solar System and that enable it to support the abundant life on its surface. These are the liquid water that covers much of its surface and the planetary environment that maintains it. However, liquid water is rare in the Solar System; as you will see in Chapter 3 there is evidence, shown in Figure 2.4, that it once existed on Mars and you will explore the possibility that it may exist below the surface of Jupiter's satellite Europa in Section 4.1. For now, however, we will concern ourselves with the Earth and what makes it such a habitable planet.

Pure water exists as a liquid between 273 K and 373 K unless the pressure is too low, in which case the water sublimes to water vapour. The presence of liquid water on a planetary surface could therefore be used as a simple requirement for us to consider the planet as being habitable. As you will see, this single factor is unlikely to be either necessary or sufficient. However, it provides a useful guide for the conditions needed to support Earth-type life on terrestrial planets (or sufficiently large moons) in orbit around a star.

A **circumstellar habitable zone** is defined as encompassing the range of distances from a star for which liquid water can exist on a planetary surface.

The primary consideration in determining a planet's habitability is therefore temperature.

- What determines the average temperature of the atmosphere and surface of a terrestrial planet?
- The balance between incoming solar radiation and thermal emission from the planet.

The amount of sunlight received by a planet is determined by its distance from the star that it orbits and the amount of energy emitted by that star: its **luminosity**. The luminosity of a star represents the total output of radiant energy per second. In comparison with the Sun, therefore, a star 100 times as luminous would emit 100 times as much energy per second as the Sun does.

- Relative to incoming energy, how much energy must a planet radiate back into space in order to remain in equilibrium with its surroundings?
- For a planet not to get any hotter or cooler it must radiate the same amount of energy that it absorbs.

If we therefore assume that a planet undergoes no net heating or cooling in the short term, it is possible to estimate the temperature necessary for a planet to re-radiate all of the energy absorbed by the atmosphere and the surface (Box 2.1).

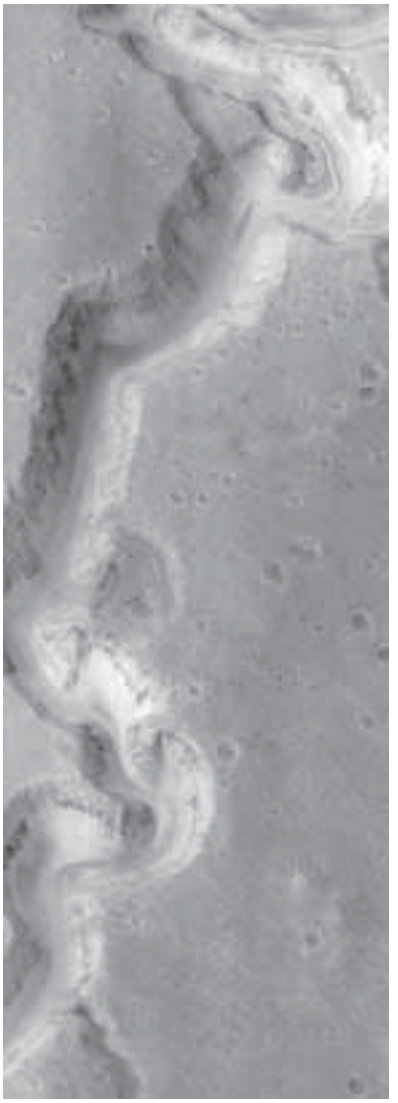


Figure 2.4 This Mars Global Surveyor image shows a portion of the meandering canyons of the Nanedi Valles system on Mars. The valley is about 2.5 km wide. The floor of the valley in the upper-right corner of the image exhibits a small channel, 200 m in width, which is covered by dunes and debris elsewhere on the valley floor. The presence of this channel is interpreted as indicating that the valley might have been carved by water that flowed through this system for an extended period of time.

BOX 2.1 DETERMINING A PLANET'S EFFECTIVE TEMPERATURE

The temperature of the surface and atmosphere of a planet is determined by the balance between the energy that is absorbed and the energy that is emitted.

If a planet has no significant source of internal heat then the source of most of the energy reaching the atmosphere and the surface is the Sun. The solar flux density at the top of the Earth's atmosphere is about 1.38×10^3 W m^{-2}. Some of this energy is reflected back to space, the atmosphere absorbs some, and the rest reaches the surface, where it is either reflected or absorbed. The absorbed radiation heats the surface, which then re-radiates this energy, mainly in the infrared region.

On the assumption that a planet undergoes no net heating or cooling in the short term, it is possible to estimate the temperature necessary for a planet to re-radiate all of the energy absorbed by the atmosphere and the surface. This temperature, called the effective temperature, T_e, is defined as follows:

$$T_e^4 = \frac{L}{4\pi R^2 \times 5.67 \times 10^{-8}} \qquad (2.1)$$

where L is the *total* power radiated by the planet in watts, R is the radius of the planet in metres (its surface area is $4\pi R^2$, and radiation is emitted from the whole surface) And 5.67×10^{-8} is a constant which has the units Wm^{-2} K^{-4}.

This equation was originally derived for thermal sources or black bodies, but it now serves to define effective temperature, regardless of the form of the spectrum of the emitted radiation.

- Look at Equation 2.1. How will the effective temperature of a planet vary as a function of the luminosity of its star?
- The effective temperature in K is raised to the power four, it will therefore vary as the fourth root of the luminosity of the planet's star.

This is illustrated in Figure 2.5 which shows how the effective temperature (T_e) of an Earth-sized black body would vary if it were in orbit around stars of differing luminosity.

The balance between the radiation absorbed and emitted by a planet determines the temperature of its surface and atmosphere. Thus, in order to estimate T_e, the power lost by radiation must be equated with that absorbed from solar radiation. Since a star's radiation arrives from one direction, a planet is heated over only half of its surface at any time. The planet therefore casts a disc-shaped shadow of area πR^2, where R is the radius of the planet. The power absorbed depends on this area, and also on the solar flux density at the distance of the planet from the Sun.

- Is all of the solar radiation that reaches a planet absorbed?
- No, a fraction of solar radiation is also reflected.

The total fraction of solar radiation that is reflected by a planet is called the **albedo**, a. The total fraction absorbed is simply $(1 - a)$. By equating the power, L, radiated by the Earth with the solar power absorbed, which can be readily estimated independently, an effective temperature of 255 K can be estimated from Equation 2.1.

Another way of looking at the effect of stellar luminosity on a planet's effective temperature is to consider the distance of a planet from its star. For example, if we replaced our Sun with a star of greater luminosity, how far from that star would

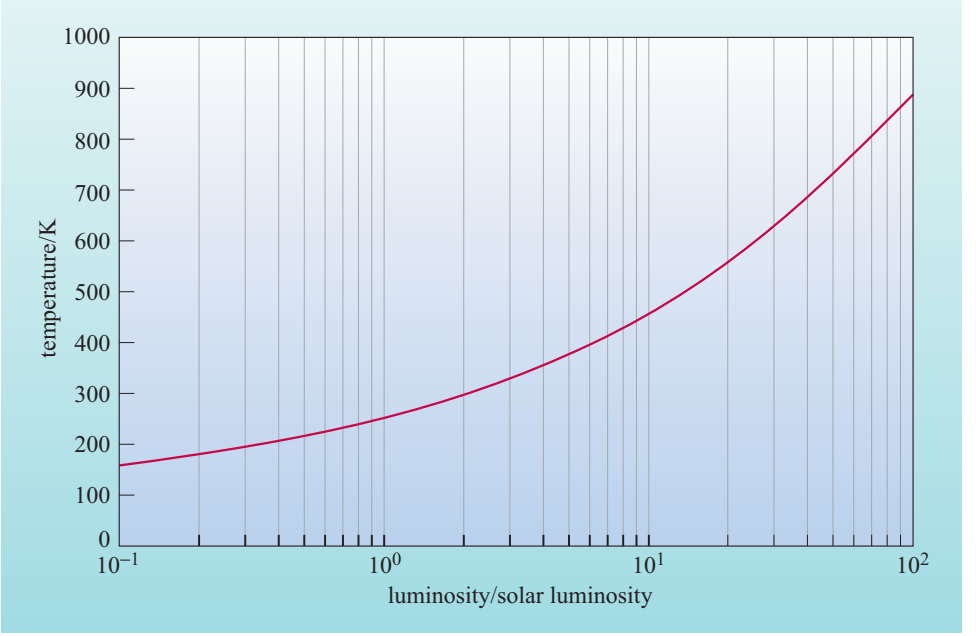


Figure 2.5 The variation in effective temperature (T_e) of an Earth-sized black body at 1 AU from stars of different luminosities. Luminosity is expressed relative to that of the Sun.

Earth have to be to maintain its current effective temperature? Since energy is conserved, electromagnetic radiation is not diminished as it travels through space. Consider a light source suspended in space that emits a flash of light that consists of a specific amount of energy. That energy will travel outwards in all directions from the light source, like a rapidly expanding sphere. At any particular moment the total energy in the expanding sphere is exactly the same as the energy initially emitted by the light source.

■ What will happen to the surface area of the sphere as its radius increases?

❑ As the sphere expands its surface area increases.

Thus the initial energy is spread over a larger and larger area of space so that the amount of energy in a square metre of the surface of the sphere decreases. Since the surface area of the sphere is related to its distance from the light source then it follows that the further a planet is from its star, the less energy it will receive and the lower its effective temperatures will be. Thus the amount of energy (E_{in}) received by a planet, referred to as its solar flux density, is defined by:

$$E_{in} = \frac{\text{luminosity}}{4\pi R^2} \qquad (2.2)$$

where R is the distance of the planet from the star. The distance scales with the square root of the star's luminosity, assuming that the effective temperature is to remain the same, and is illustrated in Figure 2.6. So if we were to replace the Sun by a star 10 times as luminous, the Earth would have to increase the radius of its orbit by the square root of 10, that is 3.16, if we wanted to maintain its present effective temperature at 255 K. This would correspond to an orbit in the middle of the asteroid belt between the orbits of Mars and Jupiter.

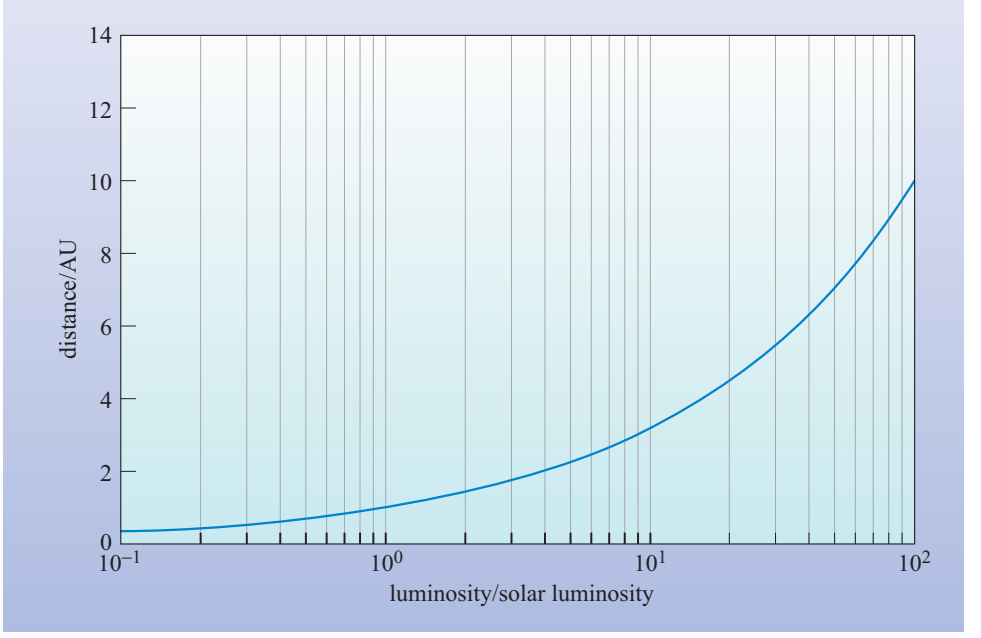


Figure 2.6 The distance from a star that is required to keep an Earth-sized body with an effective temperature of 255 K with increasing stellar luminosity.

- The mean surface temperature of the Earth today is 288 K, some 33 K higher than its effective temperature. Why is this?
- The effective temperature calculation in Equation 2.1 does not take account of the trapping of heat energy by the Earth's atmosphere, the so-called greenhouse effect, which will raise the surface temperature as will internal heat from the Earth's interior.

We will look at the consequences of albedo and the greenhouse effect on the extent of circumstellar habitable zones in Section 2.3.2.

If we were to replace our Sun with a star ten times as luminous as our Sun, the effective temperature on the Earth would increase as the fourth root of 10, that is 1.78. So the effective temperature on our planet would become (255 × 1.78) = 453 K, which is about 180 °C.

QUESTION 2.1

What would the effective temperature of the Earth be, if the Sun were to be replaced by a star 10 000 times as luminous as the Sun?

QUESTION 2.2

Where would the Earth have to orbit to have today's temperatures, if the Sun were to be replaced by a star 10 000 times as luminous?

2.3.2 The Sun's habitable zone

In fact, the Sun's luminosity has not remained constant throughout the history of the Solar System. In common with all **main sequence stars**, it has slowly increased from a level around 4 Ga ago estimated to be around 70% of its present value.

■ What effect will a slowly increasing luminosity have on a star's circumstellar habitable zone?

❑ The habitable zone will migrate away from the star.

One consequence of the increase in the Sun's luminosity is that there has been a region around the Sun that has remained habitable throughout the history of the Solar System (Figure 2.7).

This region, in which a planet may reside and maintain liquid water throughout most of a star's life, is called the **continuous habitable zone.**

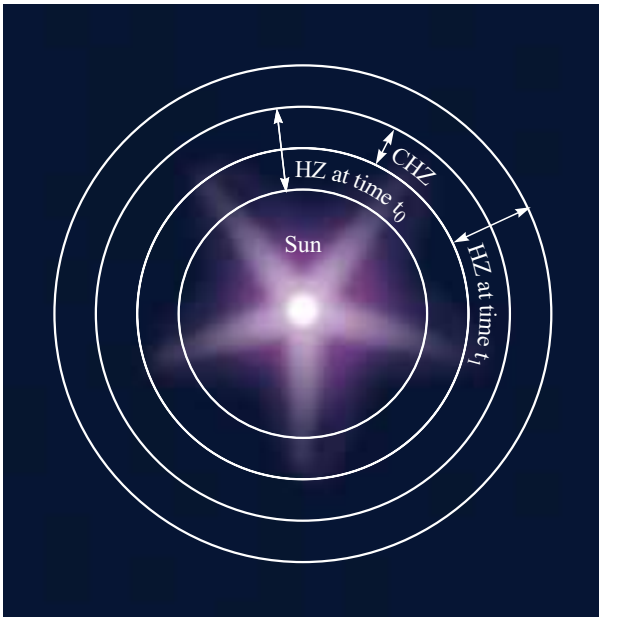


Figure 2.7 The habitable zone around a star will move outwards as the star's luminosity increases. The region that remains continuously habitable is the continuous habitable zone, CHZ.

We could use the relationships given in Equations 2.1 and 2.2 to determine the inner and outer radii of the Sun's habitable zone by equating the total power radiated by a planet (L in Equation 2.1, which is proportional to its effective temperature raised to the fourth power, T_e^4) to the solar flux density (proportional to the star's luminosity/R^2, from Equation 2.2). By using lower and upper temperatures of 273 K and 373 K for the freezing point and boiling point of water it would be possible to determine the inner and outer radii of the habitable zone. In essence, this is what was done when the concept of a habitable zone was first proposed in the late 1950s.

However, we've already seen that this is a simplistic approach since we obtain an effective temperature for the Earth of 255 K, well below the freezing point of water. It is important to note that the Earth is *not* an ideal black body. The majority of energy from the Sun is incident on the Earth at visible wavelengths, to which the Earth's atmosphere is transparent. However, the Earth radiates this heat away at infrared wavelengths. Since our atmosphere is not completely transparent at these wavelengths, because the so-called greenhouse gases (carbon dioxide, methane, water vapour and nitrous oxide) and the chlorofluorocarbons absorb infrared light, the planet is therefore forced to warm up in order to remain at equilibrium.

■ What will be the consequences of variations in planetary albedo and the greenhouse effect on the extent of the habitable zone around a star?

❑ By reflecting a portion of stellar luminosity back into space, higher planetary albedos would move the inner edge of the habitable zone towards the star. On the other hand, the greenhouse effect, by raising a planet's temperature, would extend the outer edge of the habitable zone away from the star.

In the early 1990s, James Kasting and his co-workers proposed a means of determining the Sun's habitable zone that took into account the effects of albedo and greenhouse gases. They used a climate model to estimate the width of the habitable zone around our Sun and around other main sequence stars. They employed the basic premise that they were dealing with Earth-like planets with $CO_2/H_2O/N_2$ atmospheres and that habitability required the presence of liquid water

on the planet's surface. The inner edge of the habitable zone was determined in this model by loss of water via its breakdown to oxygen and hydrogen by photolysis and the loss of hydrogen to space. With the water gone, then life as we understand it would not be possible. Kasting's climate model gives an estimate for the inner edge of both the habitable zone and continuous habitable zone in our own Solar System of 0.95 AU.

The distance at which CO_2 and other greenhouse gases can no longer compensate for the lower solar flux determines the outer edge of the habitable zone. Kasting's model indicated that the width of the habitable zone is greatly extended by the existence of a natural *carbon dioxide thermostat* that tends to regulate the temperatures of Earth-like planets, keeping them from getting too hot or too cold for liquid water to exist. He suggested that atmospheric CO_2 levels would tend to rise as a planet's surface becomes colder. The reason is that removal of CO_2 by silicate weathering, followed by carbonate deposition (Box 2.2), should slow down as the climate cools, and would cease almost entirely if the planet were to globally freeze. On planets such as Earth that have abundant carbon (in carbonate rocks) and some mechanism for recycling this carbon, for example plate tectonics, volcanism should provide a more-or-less continuous input of CO_2 into the atmosphere.

BOX 2.2 SOURCES AND SINKS OF CARBON DIOXIDE

Major CO_2 sources

On planets like Earth that have a means of recycling carbon, decarbonation (thermally decomposing carbon-containing rock and releasing CO_2) and volcanic outgassing are major sources of CO_2. Oceanic sea floor is continuously created at mid-ocean ridges and destroyed in subduction zones, where oceanic plates are pushed into the Earth's mantle. Here, calcium carbonate ($CaCO_3$) and silicate minerals are heated to high temperatures and pressures and chemically react with each other to produce CO_2:

$$\text{silicate minerals} + CaCO_3 \rightarrow \text{'new' silicate minerals} + CO_2 \qquad (2.3)$$

The CO_2 and other volatiles produced by decarbonation escape and are ultimately emitted to the atmosphere by volcanoes, hot springs etc.

■ Can you think of another major reservoir of carbon on Earth?

❑ Organic carbon.

On Earth, deeply buried organic sediments can, over millions of years, become uplifted to the surface during episodes of mountain building. Once exposed, the organic carbon in these sediments can be oxidized, returning CO_2 to the atmosphere. This provides a major additional source of CO_2. The organic carbon on Earth is overwhelmingly the result of biological processes; its recycling is an example of how the presence of life itself has a role to play in modifying planetary environments.

Major CO_2 sinks

Carbon dioxide in the atmosphere dissolves in rainwater to produce a weak acid called carbonic acid (H_2CO_3). In contact with rock at the surface, the carbonic acid can remove ions such as calcium and sodium from parent minerals – this process is called chemical weathering and it results in the production of bicarbonate ions (HCO_3^-). If the rock is a silicate, the reaction can be simplified as:

$$CaSiO_3 + 2H_2CO_3 + H_2O \rightarrow Ca^{2+} + 2HCO_3^- + H_4SiO_4 \quad (2.4)$$

(carbonic acid) (silicic acid)

In water, the calcium ions released by weathering can recombine with bicarbonate ions to form calcium carbonate as a solid deposit:

$$Ca^{2+} + 2HCO_3^- \rightarrow CaCO_3 + CO_2 + H_2O \quad (2.5)$$

Note that for every two atoms of carbon removed from the atmosphere as CO_2 dissolved in rainwater, one atom of carbon is precipitated as calcium carbonate and the other is returned to the atmosphere as CO_2 gas.

■ Can you think of another major sink of CO_2 on Earth?

❑ Photosynthesis (Equation 1.2) removes CO_2 from the atmosphere (or CO_2 dissolved in water) and converts it to organic matter, which can then be buried to form fossil organic matter.

Kasting's climate model gave two estimates for the outer edge of the Sun's habitable zone. The first, based on the point at which CO_2 would start to condense from the atmosphere, gave a limit to the outer edge of the Sun's habitable zone of 1.37 AU. The second estimate considered the point at which a maximum greenhouse effect would operate, i.e. the point where there would be enough CO_2 and H_2O in a planet's atmosphere to raise temperatures to 273 K, and gave a limit to the outer edge of the habitable zone of 1.67 AU.

■ Where do the orbits of Venus and Mars fall in relation to the present habitable zone?

❑ Venus, with a mean distance from the Sun of 0.72 AU, is well inward of the inner boundary of the habitable zone for our Solar System. Yet the runaway greenhouse effect on that planet has resulted in surface temperatures some 500 K higher than its effective temperature. Mars, with a mean distance from the Sun of 1.52 AU, falls within the 'maximum greenhouse' limit of the habitable zone, but remains outside the first CO_2 condensation limit.

It is evident from the preceding discussion that trying to establish firm limits to the extent of a habitable zone, even in our own Solar System, requires a considerable understanding of processes on the planets concerned if the models are to be accurate. Indeed, Kasting's model gave an estimate for the width of the 4.6-Ga continuously habitable zone as 0.95 AU to 1.15 AU.

- Does Kasting's model for the extent of the continuous habitable zone fit with our knowledge of early Mars?
- We've already seen in Figure 2.4 that there is evidence of channels in the old cratered terrain that indicate that flowing liquid was widespread on Mars during the first 1–2 Ga of the Solar System.

Early Mars remains a real puzzle. Although it lies beyond the continuous habitable zone in Kasting's model, its surface was once carved by streams of some flowing liquid. Whether this implies that the early Martian climate was warm, or whether it was kept warm by geothermal heat, is still debated. If the climate was indeed warm, then the models are overlooking a key element of the climate system. Two additional warming mechanisms have been suggested:

1 The presence of additional greenhouse gases, especially CH_4. It is estimated that 0.1–1% methane may have been sufficient to supply the additional greenhouse warming.

2 The presence of CO_2 ice clouds analogous to cirrus clouds on Earth, which can create a substantial greenhouse effect. Such clouds primarily scatter outgoing infrared radiation and their net effect is to warm since they scatter more efficiently at infrared wavelengths than at solar wavelengths. We'll return to the environment of early Mars in the next chapter.

QUESTION 2.3

Why is Mars presently too cold to sustain life? (Hint: Consider the role that carbon sources and sinks play in regulating the Earth's climate.)

Plate tectonics or volcanism is required to recycle carbon.

Your answer to Question 2.3 should suggest a possible additional requirement for a planet to remain habitable, namely that it is large enough to maintain active plate tectonics or at least some form of volcanism throughout its lifetime. Where this cut-off lies is uncertain, but it is somewhere between 1 and 0.1 times the mass of the Earth as Mars is about one-tenth of the Earth's mass. We'll look at the role of plate tectonics on maintaining Earth's habitability in more detail in Section 2.4.2.

Are there circumstances in which liquid water may exist beyond the strict definition of a star's circumstellar habitable zone? So far, we've been concerned almost entirely with temperatures related to solar radiation, and the subsequent possibility of liquid water, on a planet's surface. However, tidal heating of the satellites around giant planets, such as Jupiter's satellite Europa, has raised the possibility of liquid water existing below the surface of this ice-covered satellite. The interest in Europa comes from information and images acquired by the NASA Galileo spacecraft, which revealed its surface to be one of the brightest in the Solar System as a result of the satellite having a water-ice crust, 150 km in thickness. As you'll examine in detail in Chapter 4, evidence for cryovolcanism on Europa's surface has led to the suggestion of a liquid or semi-liquid water layer beneath a crust of ice.

2.3.3 Habitable zones elsewhere in the Universe

We'll examine the possibility of habitable planets around other stars in more detail in Chapter 8. Here, we'll consider some of the characteristics of a star that will determine the extent of its habitable zone. We've already seen how changes in our Sun's luminosity throughout its history lead to the concept of a continuous habitable zone. However, the mass of a star will also determine both the size and the duration of a circumstellar habitable zone. The types of stars that can support Earth-type life on planets may be limited to those of lower masses since only these stars have long enough lives as stable luminous stars for planets to form and complex life to evolve. Although all main sequence stars generate luminous energy by converting hydrogen into helium through thermonuclear fusion, stars more massive than 1.5 times that of the Sun age too quickly to support the development of complex Earth-type life. On the other hand, stars with less than half of the Sun's mass (e.g. smaller stars like our Sun's nearest neighbour Proxima Centauri) are likely to tidally lock planets that are orbiting close enough to have liquid water on their surface and may eventually cause the destruction of a life-sustaining atmosphere through condensation on the cold, perpetually dark side of the planet. The extent of the habitable zone around stars of different masses is summarized in Figure 2.8.

Tidal locking is the synchronous rotation of the star and planet.

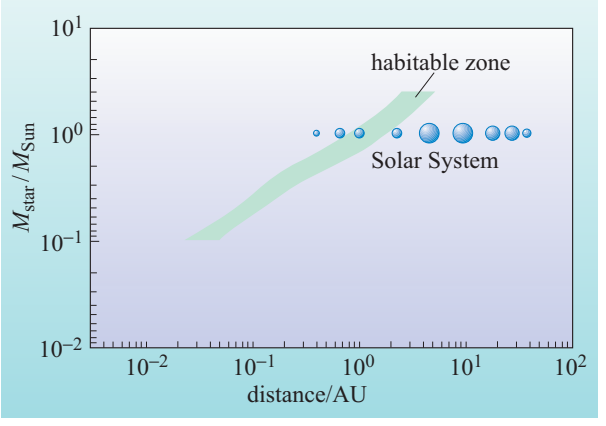


Figure 2.8 The extent of the continuously habitable zone around main sequence stars is bounded by the range of distances from a star for which liquid water would exist and by the mass of the star. Stars more massive than about 1.5 times the mass of the Sun evolve too quickly so planets would not have enough time to form complex life. Stars less than about 0.1 times the mass of the Sun will tidally lock planets close enough to have liquid water as well as subjecting them to stellar flares.

About half the stars in our Galaxy are in **binary systems**. In such systems, two kinds of stable planetary orbits exist: those in which the planet orbits both stars, called close binaries, and those in which the planet orbits one star well separated from its companion, called wide binaries. In each case only certain stable orbits exist: a planet must not be located too far away from either one star or too close to two 'home' stars or its orbit will be unstable. If that distance exceeds about one-fifth of the closest approach of the other star, then the gravitational pull of that second star can disrupt the orbit of the planet. In these cases the habitable zone would have to be calculated on an individual basis.

Scientists have recently proposed a theory that argues that certain regions of a Galaxy are more amenable to the development of complex life that others. In effect, they suggest that there are **galactic habitable zones**. Our own Milky Way Galaxy is unusual in that it is one of the more massive galaxies in the nearby Universe, and our own Sun has a relatively high concentration of elements heavier than helium, He. Based on studies of extrasolar planets, astronomers have noted that stars with higher concentrations of elements heavier than He are more likely to have planets

orbiting around them. Our Sun is also located in the outer region of the Galaxy, which protects our Solar System from the gravitational forces and increased levels of radiation associated with the large number of stars clustered near the galactic centre. The Sun's orbit about the centre of our Galaxy also means it tends to avoid the Galaxy's spiral arms, where the increased density of gases and interstellar matter leads to the formation of new stars, since it is rotating around the centre of the Galaxy at roughly the same speed as the spiral arms.

2.4 The environment on the early Earth

2.4.1 A habitable planet?

The rock record of the Earth's crust extends back only about 89% of the history of the Earth, to around 4 Ga. The oldest rocks available for study on Earth come from exposures in western Greenland, near Isua. These include ancient sediments and volcanic lavas that have since been subjected to complex folding and intrusion by younger igneous rocks as a result of subsequent geological activity (Figure 2.9). When dated using radiometric techniques, the rocks turned out to have an age of 3.8 Ga. The record can be improved somewhat if we include the information that can be gleaned from rare detrital zircon grains isolated from rocks in Australia that have ages of up to 4.27 Ga, which extend back to around 94% of Earth's history. However, the Isua rocks, to date, provide us with the best information we have as to what conditions may have been like on the early Earth.

Radiometric dating uses the degree of radioactive decay of certain isotopes to arrive at an age of mineral or rock formation.

The rocks at Isua do not differ significantly from many of the rocks formed in the more recent geological past. They contain sedimentary rocks such as limestones and sandstones together with volcanic lavas.

Figure 2.9 Amongst the oldest rocks on Earth is this 3.8 Ga-old metamorphic gneiss from western Greenland. This specimen is about 30 cm across.

■ What inference can be drawn from the presence of limestones about conditions on the early Earth?

❑ Limestones are deposited in water, either as a chemical precipitate of calcium carbonate or, in the more recent geological past, from the shells of carbonate-producing organisms. Therefore liquid water existed when the sediments were deposited.

The limestones from Isua do not contain any fossils since they were produced inorganically. The sandstones found at Isua also contain evidence of having been deposited under water and the lavas have what is called a *pillow structure*, which indicates that they cooled under water. The rocks also contain traces of carbon with a particular isotopic signature, which have been interpreted as evidence of early life-forms (see Section 2.4.4).

The Isua rocks allow us to make several inferences about conditions on the early Earth. The observation that sediments (including limestones and sandstones) and lavas were deposited and erupted under water indicates that there must have been bodies of liquid water at the Earth's surface, possibly even ocean basins. To form such sediments, land areas (albeit small) would have to have been exposed to weathering and erosion, the products of those processes being transported to our (inferred) ocean basins. Sandstones are primarily composed of quartz crystals, which implies that the land areas being weathered would have been broadly similar to the upper parts of present-day continental areas. For example, granite is a common rock that occurs in continental areas and contains a good deal of quartz. When weathered and eroded it produces quartz-rich sands. It seems, therefore, that the geological processes and cycles that we recognize today were operating on the Earth at least 4 Ga ago, albeit with some significant differences in detail.

These inferences are supported by the discovery in Western Australia in 1995 of the oldest preserved terrestrial landscape so far known. It is an area of continental crust that was being weathered and eroded some 3.5 Ga ago, before subsiding and being overlain by lavas together with some shallow water sediments. The sedimentary rocks include carbonate and sulfate minerals produced by the evaporation of seawater. Such ancient erosion surfaces provide evidence that areas of continental crust were above sea-level at a very early stage of the Earth's geological history.

The ancient sediments at Isua do not provide us with direct evidence of the composition of the Earth's early atmosphere, but they do enable us to be fairly certain that temperatures at the Earth's surface must have been within a few tens of degrees of those prevailing today.

■ What does the occurrence of sedimentary rocks at Isua tell us about the surface temperature of the Earth 3.8 Ga ago?

❑ For weathering, erosion and deposition of sediments to occur, there must have been both rain and liquid water present, implying surface temperatures of between 273 K and 373 K.

However, as you saw in Section 2.3.2, stellar evolution models suggest that solar luminosity was a factor of 25–30% lower during the early history of the Solar System than today, so why wasn't the surface of the early Earth frozen?

The problem of keeping Earth's surface temperature above 273 K in the early Solar System is often referred to as the faint young Sun problem.

We've already made one important inference from the Isua rocks that helps explain this: the evidence that geological cycles were in operation 4 Ga ago. This implies that the Earth's internal structure was very similar to what we believe it to be today: the Earth had acquired its layered structure, as summarized in Figure 2.10, by 4 Ga ago.

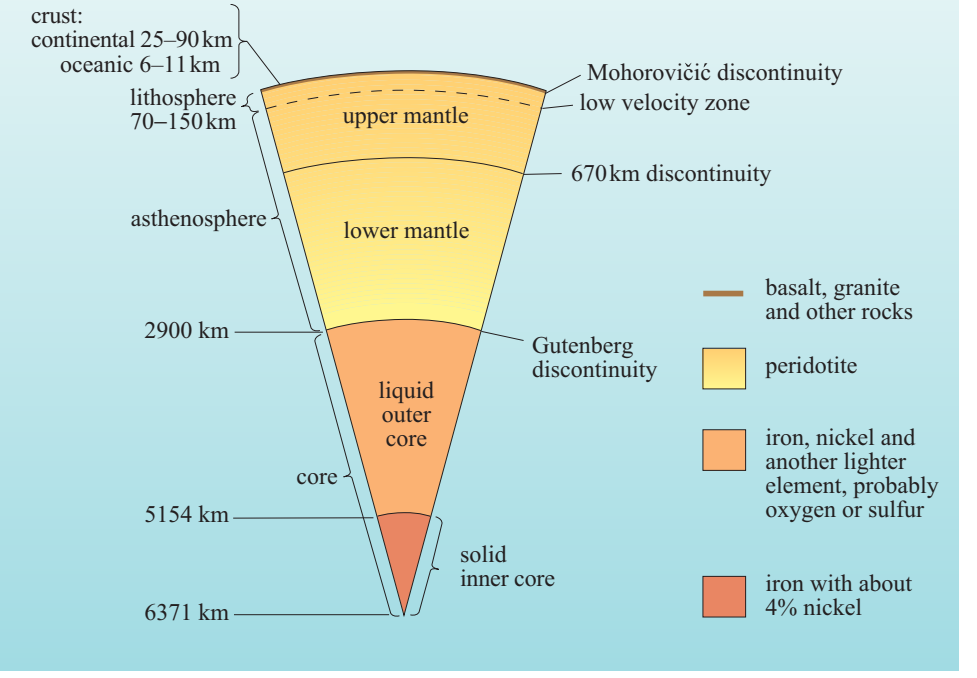


Figure 2.10 A schematic slice through the Earth, showing the major compositional features of the principal layers.

2.4.2 Plate tectonics on the early Earth?

Plate tectonic activity appears to be essential for maintaining the habitability of the Earth, since it continually recycles oceanic crust back into the mantle at subduction zones (Figure 2.11), and continually regenerates it at ocean ridges by the solidification of newly generated magma, some of which is erupted as lava on the sea-bed, to form the upper layers of the ocean crust. Without such a process, CO_2 and other atmospheric constituents would not be recycled back to the atmosphere. On planets like Earth that have abundant carbon (in carbonate rocks) and some mechanism, like plate tectonics, for recycling this carbon, volcanism should provide a more-or-less continuous input of CO_2 into the atmosphere.

■ Why is CO_2 such an important gas?

❑ As you saw in Section 2.3, CO_2 is a greenhouse gas that raises the temperature of the Earth's surface and atmosphere by retaining some of the energy received from the Sun.

The three processes that transfer internal heat to the Earth's surface and drive plate tectonic activity on the Earth are conduction, convection and advection. We know that modern internal Earth processes are mainly driven by heat that originates in approximately equal measures from two sources:

1 Heat energy from the radioactive decay of unstable isotopes, notably those of potassium, uranium and thorium.

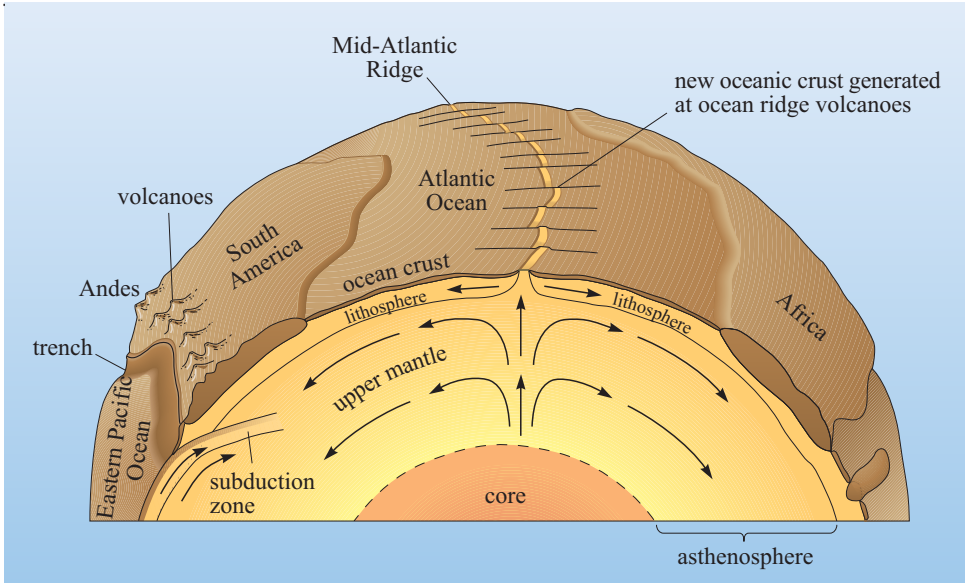


Figure 2.11 Schematic cross-section of the present-day plate tectonic cycle.

2 The remaining primordial heat.

This internal heat has somehow to be lost to space by thermal radiation from the surface. This can only happen if the heat is transferred upwards to the surface by conduction, convection or advection. However, silicate rocks are poor heat conductors so that convection and advection are the main processes by which the Earth transfers heat from its interior to its surface.

The rate at which a planet's internal heat is lost is therefore critical to the continuation of its tectonic activity.

■ What will determine the *rate* at which a planet loses its internal heat?

❑ Key factors are its size (larger bodies lose heat more slowly and will therefore remain active longer) and its composition. These factors determine the amount of heat available and the body's ability to convect.

While volcanism played a major role in the early history of Mars, the Moon, and probably Mercury, their small sizes relative to Earth resulted in the loss of internal heat at a much faster rate.

Tectonic activity on a planet will also be determined by its composition since rocks of different composition have different physical properties, which when heated will influence the ability of a body to convect. Under high pressures and temperatures, most solids can behave as highly viscous fluids, given enough time. When rocks are heated, they expand and their density decreases, making them more buoyant relative to cooler and denser ones. Thus, while hot rocks slowly rise in some regions, cooler rocks slowly descend in others, in a system of convecting cells similar to those shown in Figure 2.12 (but the convecting cells are considerably less uniform).

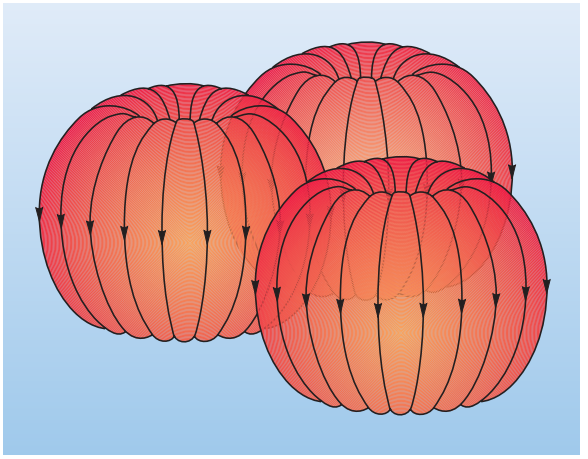


Figure 2.12 Highly schematic representation of simple convection cells, represented by arrowed flow lines. Hotter material is transferred upwards from below and cooler material is transferred downwards from above.

A planet's chemical composition will determine the amounts of heat-producing radioactive elements present and affect the likelihood of internal convection.

■ Would we expect the early Earth to have the same rate of convection as it has at the present day?

❑ No, since one of the heat sources comes from radioactive decay then the total amount of unstable isotopes will decrease with time. There would have been more radioactive isotopes decaying to produce heat in the early Earth than there are now, so there would have been more heat to lose and rates of convection would have been greater.

It has been estimated that there was about five times more heat being produced in the Earth's interior 4 Ga ago that there is today. However, not all of this heat came from radioactive heat-producing elements. In addition to primordial heat from accretion, a substantial amount of gravitational energy was released as heat when huge amounts of iron and nickel sank to the middle of the Earth, forming the core, probably about 4.5 Ga ago.

A greater degree of internal heating is one possible solution to keeping the Earth's surface temperature above 273 K in response to the reduced level of solar luminosity in the early Solar System. Plate tectonics has certainly played a role in keeping the Earth habitable, as you saw in Section 2.3.2, through the recycling of carbon. However, some scientists believe that there have been times when the mechanisms that maintain the constancy of the habitability of the Earth may have faltered with periods of dramatic climate change in which ice entombed the whole planet (Box 2.3).

BOX 2.3 SNOWBALL EARTH

We have emphasized the importance of the Earth's surface temperatures and how the cycling of carbon has helped maintain liquid water on the Earth's surface. However, what happens if that cycle is disrupted? Scientists have been aware for some time that the geological record suggests that the Earth experienced periods of dramatic climate change, as many as four times between 750 Ma and 580 Ma years ago, that resulted in periods of global glaciations: a hypothesis that has been given the name **Snowball Earth**.

Icehouse and greenhouse

The evidence for climate change comes from thick layers of sedimentary rocks deposited 750–580 Ma ago. But these sedimentary rocks are apparently full of contradictions. Let's take, for example, glacial deposits that were laid down at sea-level near the Earth's Equator at the time. Today, a glacier would have to be at an altitude of more than 5 000 m to survive in the tropics. Interspersed with the glacial deposits are layers of iron-rich rocks that should only have formed if the Earth's atmosphere and oceans contained very little oxygen – however, the atmosphere at the time would have had a composition not too different from that of today. Equally puzzling was the observation that immediately overlying the glacial deposits were layers of carbonates typically found in tropical environments today. If glaciers had extended all the way to the Earth's Equator, in effect covering the planet in ice, how did it manage to warm up again so rapidly?

These contradictions started to make sense when scientists began to contemplate that the Earth may

have experienced periods of severe climate change. The Snowball Earth hypothesis considers an Earth that was globally frozen for periods of 10 Ma or more. Heat escaping from the Earth's interior prevents the oceans from freezing to the bottom. However, surface temperatures drop to around 223 K and ice forms to a thickness of a kilometre or more. Under such conditions, all but a small fraction of the Earth's primitive organisms become extinct.

- What effect would an ice-covered Earth have on the major sinks for CO_2 from the Earth's atmosphere?
- Recall from Box 2.2, that CO_2 in the atmosphere is removed through silicate weathering followed by carbonate depositions. An ice-covered Earth would halt this process.

However, volcanic activity would not stop on an ice-covered Earth so that volcanic outgassing of CO_2 would continue. The Snowball Earth hypothesis suggests that it would accumulate to 350 times present-day levels of CO_2 creating severe greenhouse conditions that would warm the planet and melt the ice in perhaps as little as a few hundred years. Organisms that survived the icehouse must now endure a hothouse.

The Snowball Earth hypothesis also explains the occurrence of what are normally very rare iron-rich layers between the glacial deposits. These layers are analogous to the iron formations found much earlier in the Earth's history (see Section 2.4.4) when the oceans and atmosphere contained very little oxygen, and iron could readily dissolve. However, given several Ma of ice cover the oceans would be deprived of oxygen, so that dissolved iron expelled from seafloor hot springs could accumulate in the water. Once a CO_2-induced greenhouse effect began melting the ice, oxygen would again mix with the seawater and force the iron to precipitate out.

An explosion of life

Did the recovery of the Earth's climate following these huge glaciations 750–580 Ma ago pave the way for the explosion of complex multicellular animal life that happened shortly thereafter? Eukaryotes, cells with a membrane-bound nucleus and from which all plants and animals descended, emerged almost 1.8 Ga ago. However, the most complex organisms that had evolved when the first of these large glaciations occurred were filamentous algae and simple unicellular protozoa. It has always puzzled scientists why it took so long for these primitive organisms to diversify into the more complex organisms that suddenly appear in the fossil record at around 670 Ma (Figure 2.21).

A series of such global freezing events followed by equally unpleasant greenhouse conditions would certainly have had a dramatic effect on the evolution of life on Earth, effectively filtering out earlier forms of eukaryotes. All of the eukaryotes around today would thus derive from the survivors of a Snowball Earth. Some measure of the impact these conditions would have had on the evolution of eukaryotes may be evident from the phylogenetic tree (Figure 1.37). This depicts the phylogeny of the eukaryotes as a delayed radiation at the end of a long, unbranched stem. The lack of any earlier branching may indicate that any pre-existing eukaryotic ancestors were, in effect, pruned by the Snowball Earth episodes. Those that survived such global glaciations may have done so by taking refuge near the surface of the ice where photosynthesis could be maintained or on the sea floor near energy-rich hydrothermal vents.

2.4.3 The Earth's early hydrosphere

So far we have emphasized the importance of liquid water both to the origin of life on Earth and to the habitability of an Earth-like planet. But where did the Earth get its water from? The distribution of water in the inner Solar System is poorly understood, but may be roughly in scale with the size of the body. For those bodies from which we have samples, Earth contains the most water, followed by Mars, while other bodies such as the Moon and some asteroids, from which we have meteorite samples, are relatively dry. We have no known samples of Venus or Mercury, but the direct detection of water in the Venusian atmosphere suggests that Venus did contain water when it first formed.

Until recently, there have been competing views as to the origin of the Earth's water. One view held that the Earth accreted as a dry body and its water was subsequently added through cometary impact. The competing viewpoint held that the Earth inherited its water from water-bearing minerals in the un-degassed interiors of planetary embryos. However, evidence from hydrogen isotopes (Box 2.4) suggests that comets are unlikely to be the source of the Earth's water.

The ratio of the two stable isotopes of hydrogen in comets is not the same as the ratio for terrestrial ocean water, implying that the Earth did not obtain the bulk of its water by cometary impact after the end of accretion.

BOX 2.4 OTHER SOURCES OF THE EARTH'S WATER?

Water from comets?

Given that comets contain significant quantities of water, the idea that they provided the Earth's water may seem feasible. However, there is evidence from the isotopes of hydrogen that can be used as an argument against this viewpoint. Table 2.1 lists the ratio of the two stable isotopes of hydrogen (^{1}H, hydrogen; ^{2}H, deuterium – abbreviated to D) in comets 1P/Halley and Hyakutake and of the Earth's oceans (the overwhelming majority of water in Earth exists in its oceans). The ratio of the isotopes ^{2}H to ^{1}H is usually referred to as the deuterium/hydrogen ratio (abbreviated to D/H).

Table 2.1 D/H ratio (^{2}H/^{1}H) of water in comets 1P/Halley, Hyakutake and the Bulk Earth.

Body	D/H isotopic ratio
Bulk Earth	1.5×10^{-4}
Comet 1P/Halley	3.16×10^{-4}
Comet Hyakutake	2.82×10^{-4}

These data indicate that comets contain roughly twice as much deuterium relative to ^{1}H as the Earth's oceans. Thus, if the two comets listed in Table 2.1 are broadly representative of all comets, which is plausible but as yet not testable, then most terrestrial water must have come from sources other than comets.

Water from the solar nebula?

It also seems unlikely that the early Earth scavenged volatiles such as H_2O, CO_2 and N_2 directly from the solar nebula. This is because the relative concentrations of other volatiles, notably the rare gases Ne, Ar, Kr, and Xe, were much higher in the solar nebula than in the present atmospheres of the terrestrial planets. The evidence for this is illustrated in Figure 2.13, which shows the noble gas abundances in the atmospheres of the terrestrial planets relative to the solar composition (which represents that of the primordial solar nebula). It would be difficult for the terrestrial planets to preferentially lose rare gases but retain other volatiles.

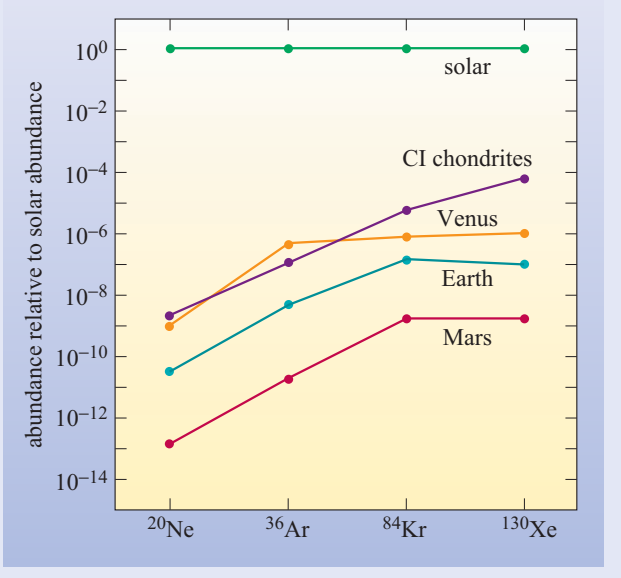


Figure 2.13 Noble gas abundances in the atmospheres of Mars, Earth and Venus relative to solar compositions. The data are plotted as the concentration of the rare gas relative to silicon divided by the corresponding solar ratio.

At present, the most plausible model for the origin of volatile materials on the early Earth is from water-bearing grains that became incorporated in planetesimals and eventually planetary embryos. This model is not without its problems – one uncertainty is whether water-bearing planetesimals could have formed at 1 AU or whether they could only have formed at distant parts of the solar nebula, for example at the asteroid belt.

■ Can you suggest how water could have been incorporated in material that condensed from the solar nebula at around 1 AU?

❑ It could have been incorporated in hydrated minerals, which condense at higher temperatures than the temperature at which water condenses to ice.

Evidence for the role of hydrated minerals comes from their presence in meteorites, for example in the carbonaceous meteorites such as Murchison that you met in Section 1.5.3. Current ideas for the origin of the Earth's hydrosphere are summarized in Figure 2.14. Large amounts of volatiles became incorporated into

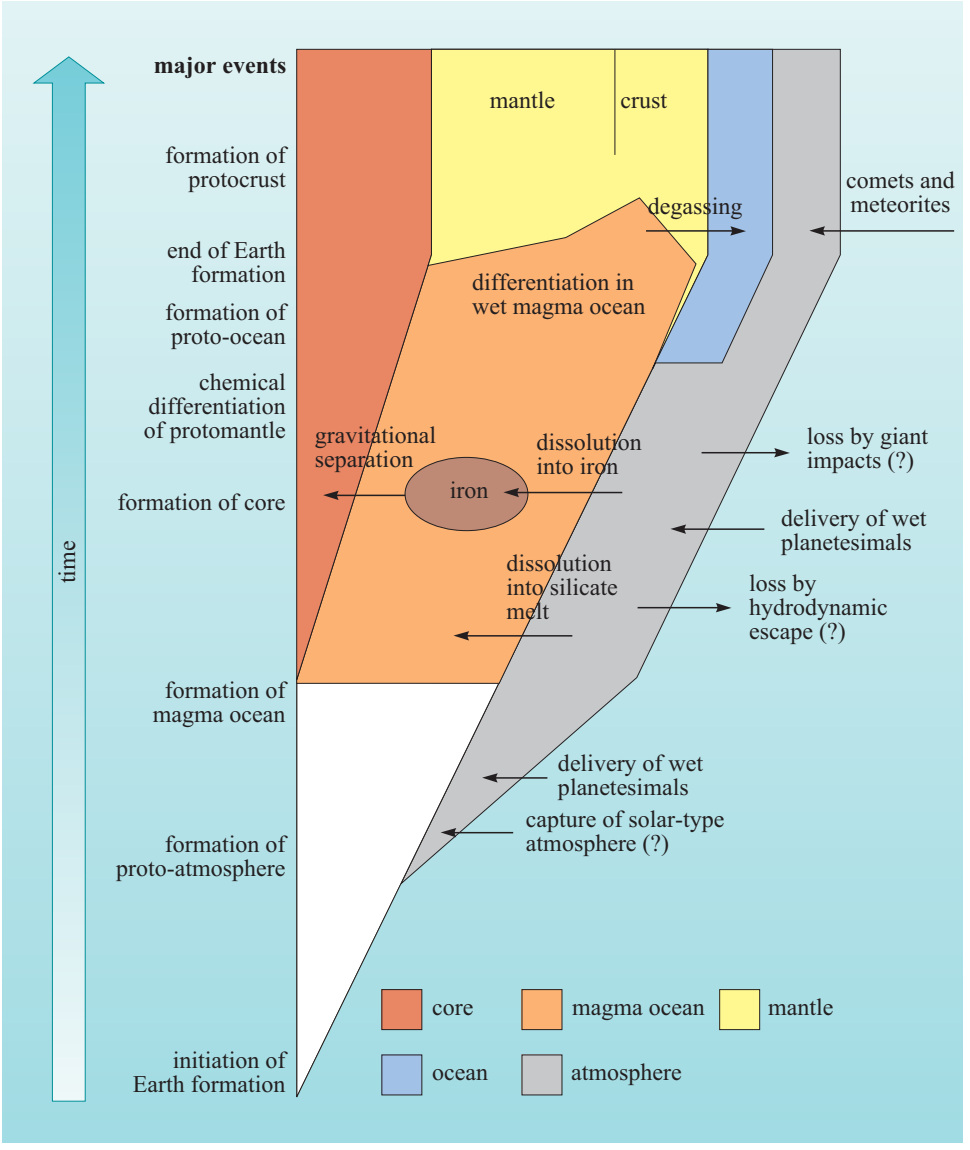


Figure 2.14 Schematic diagram of the behaviour of water during the early evolution of Earth.

the Earth as it formed. However, it seems likely that the heat generated by accretion, as well as during the formation of the Earth's core and the Moon, may have driven off much of these original volatile materials, while some volatile materials may have been incorporated into the Earth's core. The volatile materials that remained within the Earth have been outgassing from its interior ever since. Outgassing rates must have been much greater on the early Earth than they are now because rates of internal convection were greater. Most of the Earth's atmosphere and hydrosphere had probably originated from within the Earth by around 4 Ga ago, after formation of the core and the 'birth' of the Moon.

2.4.4 The Earth's early atmosphere

There is considerable debate over the oxidation state of the Earth' early atmosphere. What evidence we do have comes from the geological record. For example, the occurrence of oxidized sulfur compounds as sulfate (SO_4) deposits in rocks 3.5 Ga old from Australia implies that non-reducing conditions existed in some places, although many of the Earth's oldest rocks contain large quantities of the more reduced sulfur compound, pyrite (FeS). Other lines of evidence, such as the composition of gases emitted from modern volcanoes, our knowledge of the present composition of the atmosphere, and observations of conditions on other planets, have led to the conclusion that, after formation of the core, the Earth's early atmosphere became dominated by nitrogen (N_2), carbon dioxide (CO_2), possibly sulfur oxides (especially SO_2), and water vapour (as there were oceans present).

For about the first billion years of Earth's history, oxygen (O_2) was only present in trace amounts as a result of the breakdown of water vapour by UV radiation.

However, this inorganic mechanism of releasing oxygen compounds into the atmosphere produces only tiny amounts of free oxygen in the atmosphere. Since the Earth's present atmosphere is oxygen-rich and because all higher forms of life require free oxygen, there is an obvious need for some other, much more powerful source of oxygen. The most plausible source is oxygen-producing photosynthesis (referred to as oxygenic photosynthesis), a process that first evolved in cyanobacteria (or their immediate precursors) during the Archaean Era (from 3.8 Ga to 2.5 Ga ago).

One consequence of very low O_2 levels in the Earth's early atmosphere would have been the lack of an ozone layer to absorb incoming ultraviolet radiation from the Sun and prevent it from reaching the Earth's surface (as it does today).

- What effect might the lack of an ozone layer have on the development of early life?

- Higher levels of UV radiation at the Earth's surface suggest that life would stand a better chance of survival in more protected environments, for example deeper water or within sediment.

However, clouds also reflect sunlight, including ultraviolet radiation, and the lack of an ozone layer may have resulted in a different thermal profile in the early atmosphere of the Earth. On the present-day Earth, the temperature of the atmosphere rises above the tropopause (at around 20 km) because the ozone layer

absorbs ultraviolet light and gains energy. This energy is emitted as heat, effectively providing an invisible 'lid' for convection cells in the atmosphere. Without an ozone layer, atmospheric temperature could have continued to decrease to much greater altitudes on the early Earth. This would have allowed atmospheric convection cells to extend to much greater altitudes and, consequently, cloud formation may have extended to considerably greater heights than it does today. It is therefore possible that some shielding from ultraviolet light was provided by clouds, although we have no way of knowing how extensive any cloud cover would have been.

In Section 2.4.1 you saw that there were certainly bodies of water (possibly oceans) at least by 3.8 Ga ago, and also some land areas to provide sediments. Land formed from granitic material (like modern continental crust) must have been more limited in extent or we would have expected to find more examples of ancient granitic rocks on Earth. There may also have been small islands produced by basaltic volcanoes, which would not be preserved but would have been subducted and recycled back into the mantle.

Volcanic activity would have resulted in one particular environment wherever there were bodies of surface water. That is hydrothermal circulation, resulting from the circulation of seawater through cracks and fissures in hot basaltic rocks where lavas are being erupted to form new ocean crust or build volcanoes on the ocean floor. As you saw in Section 1.7.3, hydrothermal environments may have played a significant role in the origin of life on Earth.

The geological record also provides possible evidence for the changing nature of the Earth's early atmosphere as a result of the appearance of life. We'll examine evidence from carbon isotopes as to when oxygenic photosynthesis originated in the next section. However, one line of evidence for the appearance of free oxygen during the Archaean Era comes from **banded iron formations (BIFs)**. BIFs are amongst the oldest rocks on Earth. They occur throughout the world and are vast in extent. The example shown in Figure 2.15 from the Hamersley Ranges in Western Australia occupies a basin more than 300 km in diameter.

Figure 2.15 Mount Tom Price iron ore mine, Hamersley Ranges, Western Australia.

Banded iron formations are exactly what their name implies: they are rock formations that are characterized by finely banded dark brown, iron-rich layers alternating with lighter-coloured iron-poor layers, the layers range in thickness from less than a millimetre to about a centimetre. The iron-rich bands contain the highly insoluble iron oxides: haematite (Fe_2O_3), limonite ($Fe_2O_3.3H_2O$) and magnetite (Fe_3O_4). Chert, a rock composed of precipitated silica, occupies the iron-poor bands. Individual bands, often only a few millimetres thick, can extend for tens of kilometres. How BIFs formed is not entirely clear; we have no modern analogues to guide us. However, the involvement of iron oxides suggests that the process that led to the formation of BIFs must have affected the oxidation state of iron and hence BIFs could contain information about the oxidation state of the Earth's ocean and atmosphere at the time they formed.

QUESTION 2.4

The total mass of BIF deposits older than 2.5 Ga is estimated at 3.3×10^{16} kg. If a typical BIF consists of 30% haematite (Fe_2O_3), how much oxygen would have been incorporated in BIFs prior to 2.5 Ga ago?

Your answer to Question 2.4 shows that whatever the chemistry occurring in the formation of BIFs, large amounts of oxygen were incorporated into BIFs very early in the Earth's history. One interpretation is that this suggests oxygen was available in the shallow seas where most BIFs were formed.

Most theories for the origin of BIFs involve a significant role for hydrothermal activity on the early Earth. The seawater flowing through these hydrothermal systems would have dissolved iron-containing minerals so that iron in a reduced form was subsequently injected into the deep ocean through hydrothermal vents. It is generally accepted that the deep ocean on the early Earth was extremely oxygen deficient or anoxic so that iron escaped oxidation and precipitation at the vents themselves but was deposited in much shallower, more oxygenated water.

How did the iron get from the deep ocean to shallow water, crossing large expanses of oceans in so doing? One idea is that the iron was actually consumed by bacteria which flourished near the vents and that these bacteria then drifted away in vast colonies into shallow water where they died, depositing a thin film of organic-rich material. After a while the organic material would have been recycled, leaving the iron behind in its highly insoluble oxide form.

- Hydrothermal vents provide a satisfactory source for the large amounts of iron involved in the formation of BIFs, but where might the large amounts of oxygen come from?
- Since oxygenic photosynthesis would have been a powerful generator of free oxygen, a strongly oxidizing local environment would result wherever it occurred.

2.4.5 Palaeontological and geochemical evidence for early life on Earth

The only direct evidence we have for life on the early Earth comes from palaeontological and carbon isotopic data obtained from the preserved geological rock record. These data suggest that life *may* have already been established by 3.5 Ga ago, possibly 3.8 Ga ago, which in turn suggests that life may actually have originated by 4 Ga ago. However, as you will see, these data are by no means unchallenged and are the subject of intense scientific debate.

If life was present on Earth 3.8 Ga ago, it is generally assumed that it would have arisen after the formation of the Earth's crust and oceans and after the end of the late heavy bombardment, i.e. some 4 Ga ago.

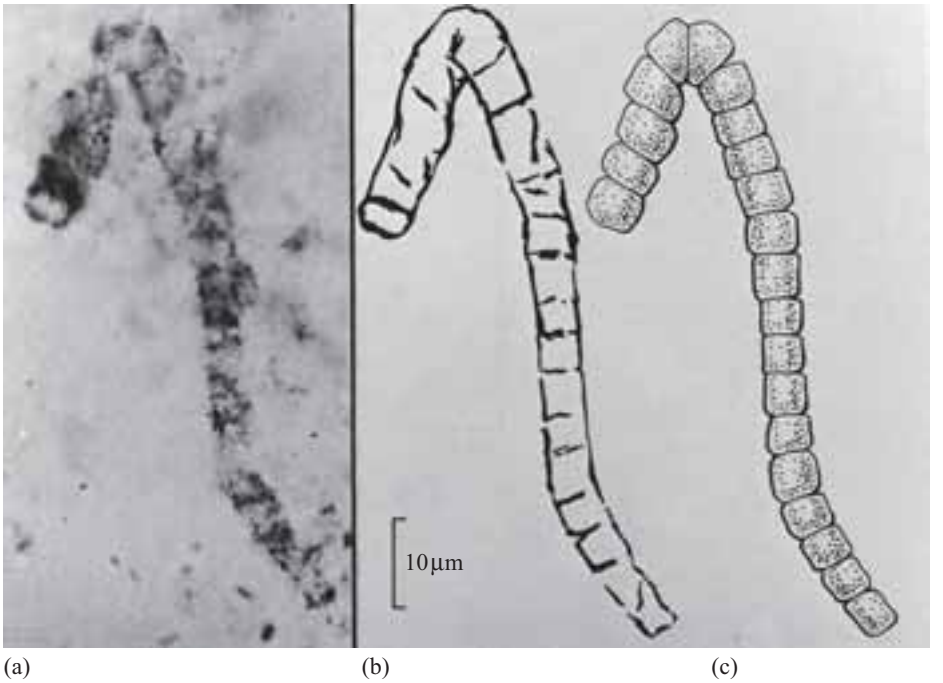
Stromatolites

In shallow coastal waters, characteristic mound-shaped structures called **stromatolites** (Figure 2.16) are built up by the accumulation of sediments that consist of thin gelatinous mats alternating with thin layers of calcium carbonate. The organisms that form these mats, and precipitate the calcium carbonate, include the simple photosynthesizing nitrogen-fixing cyanobacteria or blue–green algae. In cross-section, stromatolites have a layered structure like a stack of pancakes. Identical fossil structures occur in a variety of rocks that were produced in the first 2 Ga of the Earth's history. The oldest putative stromatolites so far reported come from the 3.46 Ga-old Apex cherts of the Warrawoona Group in Western Australia (Figure 2.17). The organisms that formed these stromatolites may also have been cyanobacteria, but this is not absolutely certain because some modern photosynthetic but not oxygen-producing bacteria form rather similar structures. This is an important point with far-reaching implications for the environment of the early Earth. Cyanobacteria (but not some other forms of photosynthetic bacteria) carry out oxygenic photosynthesis and release oxygen into the environment. The evidence that these 3.46 Ga-old rocks might contain the Earth's oldest fossils comes from structures resembling remarkably well-preserved bacterial and cyanobacterial microfossils that were first reported by the University of Los Angeles geologist Bill Schopf in the early 1980s. Schopf believes that the structures he observes are microfossils that contain evidence of *cells* and in some cases there are filaments and spherical structures that look remarkably like modern cyanobacteria (Figure 2.18). However, this interpretation has recently been challenged (Box 2.5).

Structures similar to stromatolites can also be produced by non-biological processes so fossil stromatolite occurrences can be controversial.

Figure 2.16 Modern stromatolites in Shark Bay, Eastern Australia.

Figure 2.17 Cross-section showing the layered structure of a stromatolite, 3.46 Ga old, from the Warrawoona Group, Western Australia.



(a) (b) (c)

Figure 2.18 Segmented, filamentous, carbonaceous structures found in 3.46 Ga old fossil stromatolites resembling modern cyanobacteria: (a) photomicrograph, (b) line drawing, (c) reconstruction.

BOX 2.5 FOSSILS OR ARTEFACTS?

The interpretation that structures preserved in the 3.46 Ga old Apex cherts of the Warrawoona Group in Western Australia represent preserved microfossils of cyanobacteria was challenged in 2002 by a team of scientists led by Martin Brasier from the University of Oxford. Brasier and his colleagues believe the structures to be artefacts formed from carbon deposited by hydrothermal fluids passing through the rocks. They argue that the carbonaceous, filamentous structures are not consistent with those that might be expected from cyanobacteria, being much larger and containing branches not found in cyanobacteria. They also question whether the carbon found in the Apex cherts is of biological origin. If it is of biological origin, the estimated hydrothermal temperatures of between 250 °C and 350 °C would require extremely heat-tolerant organisms and they suggest and that the carbon in these rocks may be produced by the non-biological catalytic synthesis of organic material.

Although biochemical remains are rare in rocks preserved from the early Earth, a portion of the organic matter they once contained is often preserved as a macromolecular material called *kerogen*, a word derived from the Greek *keros* meaning wax and used to describe the organic matter present in sediments. Most ancient sediments have been heated up by deep burial since they were deposited more than 3 Ga ago, and the kerogen in them is often highly altered. This makes much of the structural chemical information it once contained indecipherable. However, it is possible to use the abundance of the carbon isotopes in ancient kerogen as a tool for unravelling the biochemistry of the ancient Earth. Box 2.6 summarizes how biological processes will determine the relative ratios of ^{13}C to ^{12}C, expressed as $\delta^{13}C$ values, in living organisms. Since the carbon from those organisms is preserved when they die and can then become incorporated as kerogen in rocks, we can use the $\delta^{13}C$ values of kerogen to draw inferences about ancient biological processes on Earth and, should measurements become available, on other planets. The important point to take from Box 2.6 is that if a sample of organic matter is enriched in ^{12}C and depleted in ^{13}C relative to the inorganic carbonate standard, *the value of* $\delta^{13}C$ *will be negative*.

BOX 2.6 CARBON ISOTOPES AS INDICATORS OF BIOLOGICAL PROCESSES

Carbon isotopes

Carbon consists of a number of isotopes of which two, ^{12}C and ^{13}C, are stable. ^{12}C contains 6 neutrons and 6 protons; ^{13}C contains 7 neutrons and 6 protons. You may have come across the technique of carbon dating that uses one of the radioactive isotopes of carbon, ^{14}C, to determine the age of archaeological remains (^{14}C is an unstable isotope). Since stable isotopes do not undergo radioactive decay, they cannot be used to date rocks or other materials. However, physical processes do affect the ratio of the stable isotopes of an element. For example, in chemical reactions it takes less energy to break the bond $^{12}C—^{12}C$ than $^{12}C—^{13}C$ and, similarly, it takes less energy to make a bond between two ^{12}C atoms than between a ^{12}C atom and a ^{13}C atom.

- If a simple chemical reaction involves the making of carbon bonds, will the products of the reaction contain more or less ^{12}C than the starting materials?
- Since it takes less energy to make a $^{12}C—^{12}C$ bond than a $^{12}C—^{13}C$ bond, the reaction will preferentially incorporate more ^{12}C into the products. The products will contain more ^{12}C and we would refer to them as being ^{12}C-enriched.

In any chemical reaction, molecules bearing the lighter isotope, ^{12}C, will, in general, react slightly more readily than those with the heavy isotope, ^{13}C.

The reason why bonds between elements containing the lighter isotope react more readily than those with the heavier one is due to differences in their physical properties (e.g. density, vapour pressure, boiling point and melting point) due to the greater vibrational energy of the lighter isotope, although the chemical properties of the isotopes of an element are the same.

Since biology involves a wide range of physical and chemical processes, it should come as no surprise that biological processes affect the isotopes of carbon. We refer to the separation of isotopes of an element during naturally occurring processes as a result of the mass differences between their nuclei as **isotope fractionation**. You should note, however, that most natural processes are not capable of completely separating the isotopes of an element, rather they tend to concentrate one isotope in preference to another.

On Earth, the natural abundance of ^{13}C is roughly one-ninetieth that of ^{12}C, that is ^{12}C is considerably more abundant. It is the ratio between these two isotopes that we are interested in and, by convention, we refer to the ratio of the minor isotope to the major isotope, i.e. $^{13}C/^{12}C$. Thus a typical carbonate rock on Earth might have a $^{13}C/^{12}C$ ratio of 0.01123722, while carbon from a living organism that has had its ratio of ^{13}C to ^{12}C affected by biological processes might have a $^{13}C/^{12}C$ ratio of 0.0109563. Differences in the natural abundance of carbon stable isotopes only begin to express themselves in three figures after the decimal point, i.e. a few parts per thousand variation. It is obviously not very practical to have to refer to the ratio between ^{13}C and ^{12}C using such small numbers. Furthermore, determining the absolute ratio of two isotopes is analytically very difficult so scientists adopt a standard and measure the ratio of $^{13}C/^{12}C$ relative to that standard. Enrichment or depletion in ^{13}C is then expressed in terms of a $\delta^{13}C$ value (δ is the lower-case Greek letter 'delta'). The ratio of $^{13}C/^{12}C$ for the sample being investigated is compared to that of a carbonate standard:

$$\frac{^{13}C/^{12}C \text{ ratio of sample}}{^{13}C/^{12}C \text{ ratio of standard}}$$

One is subtracted from this value and the whole is multiplied by 1000 to give a $\delta^{13}C$ value in terms of parts per thousand (‰):

$$\delta^{13}C = \left[\frac{^{13}C/^{12}C \text{ sample}}{^{13}C/^{12}C \text{ standard}} - 1\right] \times 1000 \qquad (2.6)$$

For carbon, the standard used is a Cretaceous fossil belemnite known as the Pee Dee Belemnite (abbreviated to PDB) after the rock formation from where it was recovered. It has a $^{13}C/^{12}C$ ratio of 0.01123722.

- Using Equation 2.6, calculate the $\delta^{13}C$ value of a sample of organic matter with a $^{13}C/^{12}C$ ratio of 0.0109563.

❑ Substituting the values in Equation 2.6 we get:

$$\left[\frac{0.0109563}{0.01123722} - 1\right] \times 1000 = -25.0\,‰$$

One of the major carbon stable isotope fractionations that is the result of a biological process is that due to autotrophic photosynthesis. This is a complex process, but we can think of it as occurring in two stages: in the first stage carbon dioxide from the atmosphere is 'imported' into the cell; in the second stage, an enzyme known as ribulose bisphosphate carboxylase (abbreviated to **Rubisco**) is involved in forming carbohydrates $(CH_2O)_n$ from the imported CO_2. Both stages favour the incorporation of ^{12}C over ^{13}C so that the carbohydrates will have more negative $\delta^{13}C$ values than the CO_2.

- Using Equation 2.6, will the value of $\delta^{13}C$ be negative, positive or zero when the ratio $^{13}C/^{12}C$ is (a) equal for standard and sample? (b) Greater in the sample? (c) Lower in the sample?

❑ Because one is subtracted from the ratio of ratios, for (a) the value will be zero; for (b) it will be positive and for (c) it will be negative.

The fractionation of the carbon stable isotopes produced by photosynthesis can be used as a biomarker for past biological processes since the organic matter produced by living organisms can become incorporated in rocks as organic carbon

(abbreviated to C_{org}). However, it is important to realize that we also need an isotopic value for the source carbon used in photosynthesis, i.e. atmospheric CO_2 or CO_2 dissolved in ocean water. Direct measurements of the $\delta^{13}C$ values of past atmospheric CO_2 or oceanic dissolved CO_2 are rarely possible. Instead, scientists use the $\delta^{13}C$ value of carbon from carbonate rocks (abbreviated to C_{carb}) which represents oxidized carbon and so reflects the $\delta^{13}C$ values of the atmosphere at the time of carbonate rock formation. C_{carb} $\delta^{13}C$ values have remained around 0‰ throughout Earth history (Figure 2.20), from which scientists infer that the $\delta^{13}C$ value of atmospheric CO_2 has remained roughly constant over the last 3.8 Ga.

Biological processes result in large isotopic fractionations, i.e. they will preferentially use one isotope of carbon over another. Indeed, biological processes are the most important cause of variations in the isotopic composition of carbon on Earth. For the most part, the largest fractionation in carbon isotopes occurs during the initial production of organic matter by autotrophs. In general autotrophic organisms preferentially use the isotope ^{12}C over ^{13}C when they fix carbon, for example from atmospheric CO_2, to produce organic matter.

- If we were to analyse the carbon isotopic composition of a blade of grass, would it contain more ^{12}C or less ^{12}C that the carbon dioxide it fixes from the atmosphere?

- The grass, in common with all plants, will preferentially use ^{12}C over ^{13}C, so it will contain more ^{12}C than atmospheric CO_2. The grass will therefore have a more negative $\delta^{13}C$ value than atmospheric CO_2.

Figure 2.19 shows the range of $\delta^{13}C$ values for living (extant) autotrophs, present-day marine bicarbonate (in solution) and atmospheric CO_2. There is quite a range of values among autotrophs, reflecting their different styles of carbon fixing-reactions, but the $\delta^{13}C$ values are mostly between about –10‰ and –40‰.

Atmospheric CO_2 and marine bicarbonate are the principal carbon sources used in photosynthesis.

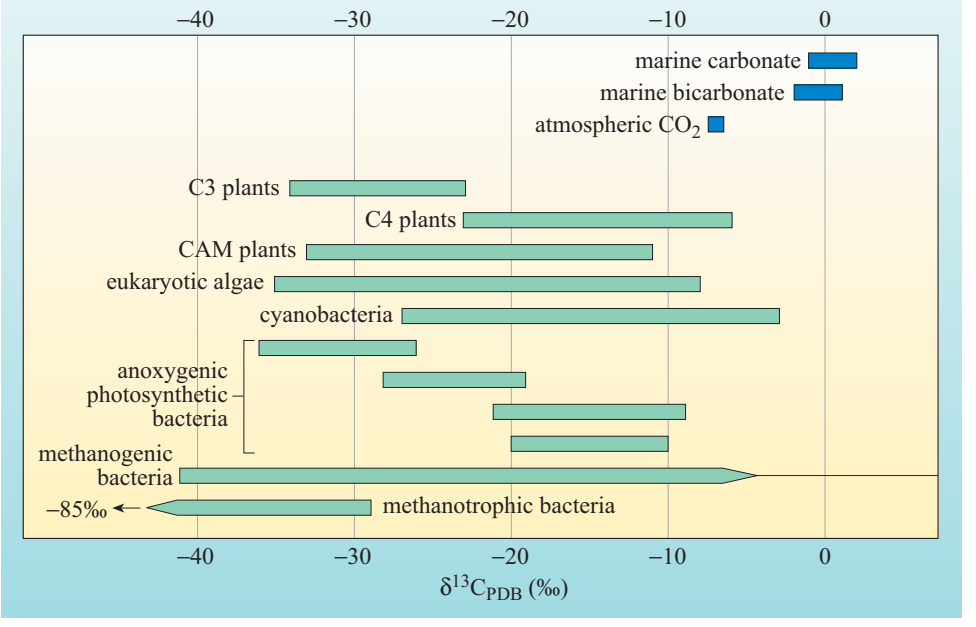


Figure 2.19 Carbon isotope ranges of major groups of higher plants and micro-organisms compared with the respective ranges of the principal inorganic carbon species in the environment.

Figure 2.20 summarizes more than 10 000 measurements of the carbon isotopic composition of sediments of all ages. There are two groups of values shown on this figure, C_{carb}, which denotes the values obtained from carbonate rocks and C_{org}, which denotes the values measured for kerogen isolated from sediments. The difference between the two has remained roughly constant throughout the Earth's history and reflects the difference between atmospheric CO_2 and marine bicarbonate, the main sources of carbon for photosynthesis, and the organic carbon of autotrophs.

The Isua rocks at 3.8 Ga are the oldest rocks for which carbon isotope data have been obtained. Some scientists believe that the data from Isua indicate that there were biological processes operating on the Earth at that time. Figure 2.20 shows that the Isua rocks have both more negative carbonate (C_{carb}) $\delta^{13}C$ values and less negative organic (C_{org}) $\delta^{13}C$ values than those in later sediments. This could be because the rocks have been metamorphosed, i.e. subjected to high temperature and pressure, which is known to affect the carbon isotope ratios. The original $\delta^{13}C$ values, before metamorphism, may have been the same as those in younger rocks so it is possible that autotrophs were thriving some 3.8 Ga ago.

More recent analyses, though controversial due to the possibility of contamination, measured the carbon isotopic composition of small carbon inclusions in single grains of the mineral apatite from 3.85 Ga-old Isua formation rocks and gave more negative $\delta^{13}C$ values (their ranges are shown as A1 and A2 on Figure 2.20). The results gave $\delta^{13}C$ values ranging from –21‰ to –41‰, which fall well within the biological ranges shown in Figure 2.19 and are consistent with the archaebacteria, for example methanotrophs. Such organisms figured prominently in some of the Earth's oldest microbial ecosystems. For example, Figure 2.20 also shows a pronounced negative excursion in the $\delta^{13}C$ record of C_{org} at around 2.7 Ga ago with $\delta^{13}C$ values reaching as low as –60‰ in rocks from the Fortescue Formation in Australia. These excursions have been interpreted as evidence of methane-using bacteria, which often have very negative $\delta^{13}C$ values.

A methanotroph is a bacterium that can use methane as a nutrient.

Many scientists believe that the sedimentary carbon isotope record shown in Figure 2.20 can best be interpreted as representing evidence that biological fixation of carbon by autotrophic organisms had been established by 3.8 Ga ago, having

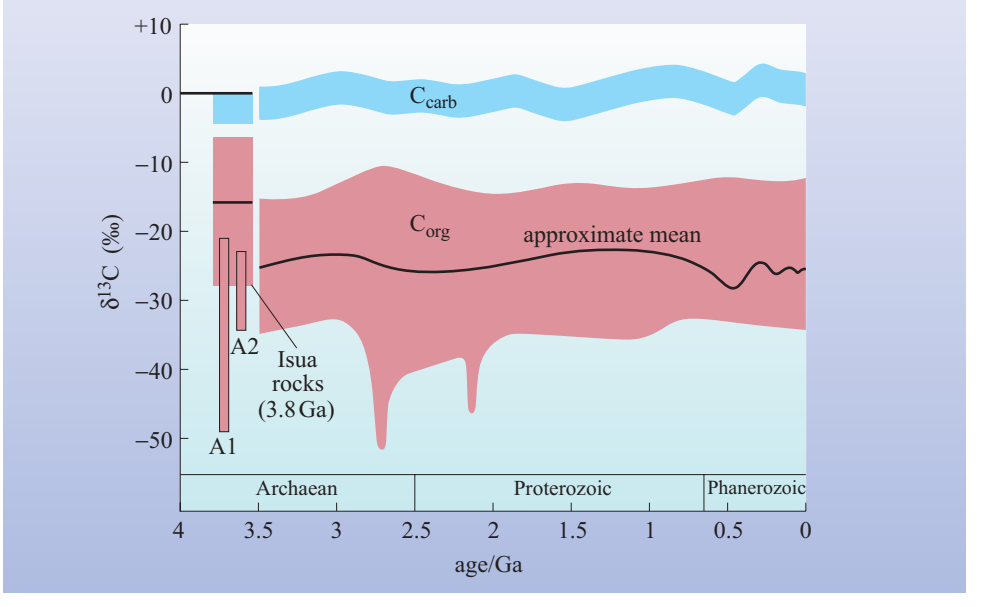


Figure 2.20 Carbon isotope compositions as $\delta^{13}C$ values for sedimentary carbonate (C_{carb}) and kerogen (C_{org}) over 3.8 Ga of the Earth's history. Superimposed on the envelope for whole rock analyses are the ranges obtained for carbon inclusions in apatite grains (A1 and A2) from two iron-rich 3.85 Ga-old Isua formation rocks.

been fully operational by the time of the formation of the Earth's oldest sediments. However, this idea was challenged in 2002 when geochemical data suggested that some of the rocks from Greenland were not BIFs as had been previously thought, and therefore likely to preserve biologically derived carbon, but were igneous in origin and had been given the superficial appearance of a BIF rock as a result of the passage of fluids through the rock. It was argued that non-biological processes involving fluids and inorganic iron carbonate produced the carbon in these rocks. However, to fully invalidate the biological interpretation of the Earth's early $\delta^{13}C$ record will require evidence for an inorganic process operating on a global scale that can mimic, both in direction and magnitude, the isotopic fractionation associated with autotrophic carbon fixation. Evidence of the Earth's earliest biosphere remains a topic of contentious debate amongst scientists.

QUESTION 2.5

Based on what you have studied in Chapters 1 and 2 so far, outline a scenario for the emergence of life based on the following lines of evidence:

- Geological evidence that familiar geological and geochemical processes were operating on the early Earth.
- Evidence in favour of the Earth's internal heat acting as a source of energy for the first autotrophic metabolic reactions to appear as opposed to external inputs of energy into a reduced atmosphere that led to the appearance of heterotrophic organisms.
- Geological and phylogenetic evidence that suggests that hydrothermal systems were a key environment on the early Earth.

QUESTION 2.6

From the discussion in Sections 2.4.1 to 2.4.5, draw up a list of conditions on the early Earth that would have affected the potential for the emergence of life in hydrothermal systems. How do these conditions compare with those of the present day?

2.4.6 Evolving complexity

A common question that arises when considering the possibility of life elsewhere in the Universe is that of the existence of life-forms like the ones that have evolved on Earth. The evolutionary trends of life on Earth form the main basis we have for hypotheses on the nature of life elsewhere. They are marked by a series of major changes in the size, form and complexity of organisms and major expansions in diversity that have produced the enormous variety of species that populate the fossil record (Figure 2.21). This record, combined with the phylogenetic tree (Figure 1.37), form the foundation for inferences about the sequence and direction of evolution.

■ With reference to Figure 2.21, how has the size of organisms changed throughout the Earth's history?

❑ For the first 2.5–3 Ga of life on Earth, most species did not exceed a few millimetres in size and most were generally substantially smaller. In the last 600 Ma, however, the evolution of larger and more complex organisms has occurred.

Bilaterians are animals that are symmetric about a central axis.

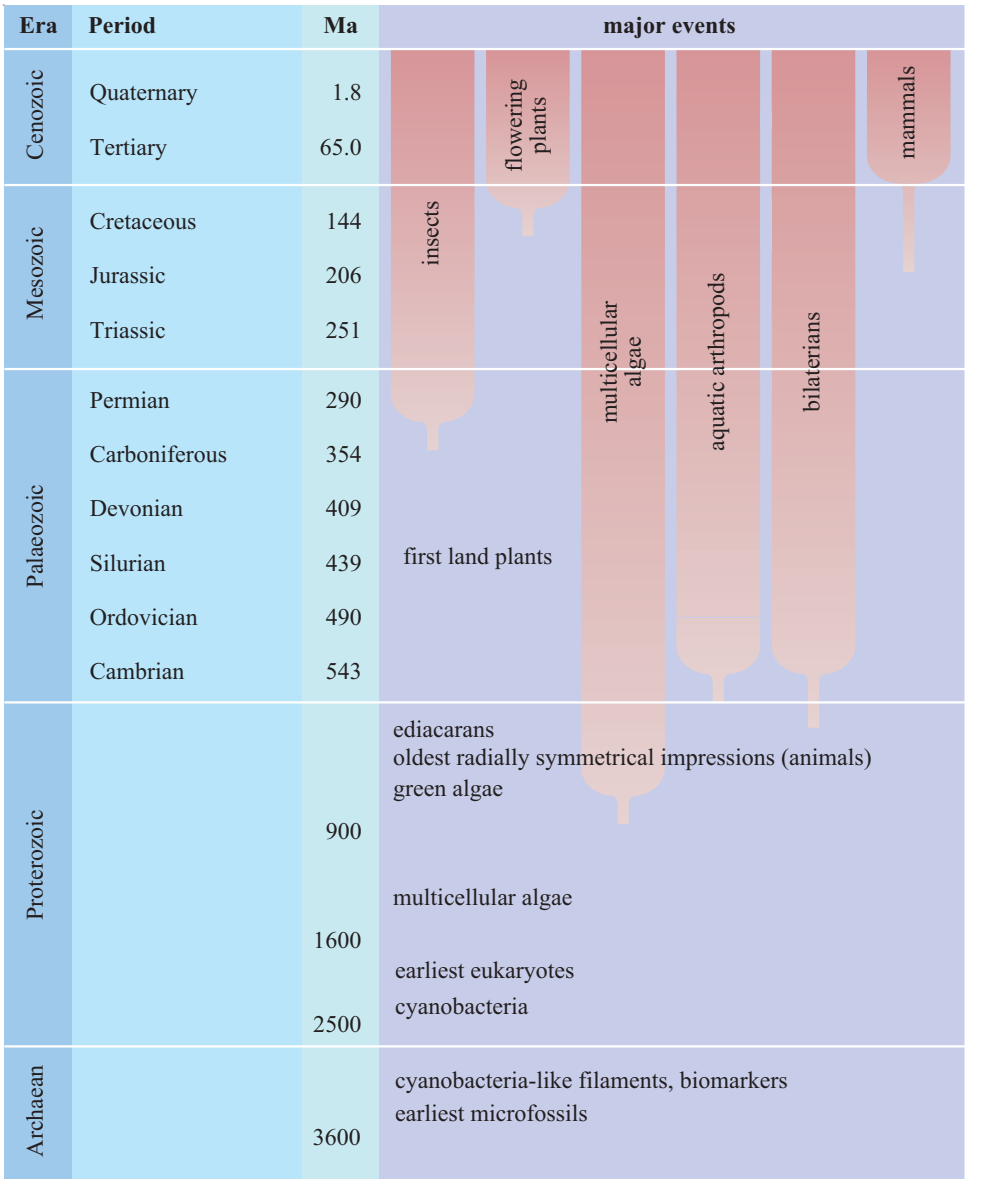


Figure 2.21 History of major evolutionary events in the geological record of Earth.

The size of organisms increased significantly with the evolution of multicellular forms. In algae and bacteria, one of the simplest ways to form a multicellular organism was for the products of cell division (see Section 1.2.6) to remain together and produce long filaments. Evidence from the fossil record suggests that many early multicellular eukaryotes were indeed millimetre-sized, linear or branched filamentous forms. By around 600 Ma ago, the fossil record suggests the presence of millimetre-sized radially symmetric life-forms. From 600 Ma onwards, major changes in the size of organisms started to appear. The Ediacaran fauna are a distinctive group of fossils that arose about 670 Ma ago. They are named after Australia's Ediacara hills, where they were first discovered. They comprised tubular, frond-like, radially symmetric forms and generally reached several centimetres in size, with some as large as 1 m (Figure 2.22).

(a)

(b)

(c)

(d)

Figure 2.22 Examples of Ediacaran fauna fossils. (a) *Dickinsonia*, an elongate pancake-shaped worm; (b) *Cyclomedusa*, a jellyfish; (c) *Tribrachidium*, a bun-shaped organism with three spiral tracts on its upper surface; (d) *Inkrylovia*, an elongate bag-like form with transverse partitions.

The size of organisms expanded rapidly after around 500 Ma ago, with algae and sponges reaching up to 50 cm in size. Subsequently, the sizes of animals increased by another two orders of magnitude to produce such giants as the dinosaurs and the larger mammals that have evolved in the last 65 Ma. Along with an increase in size, life on Earth has also increased in diversity since its origin. However, this has not been a steady and continuous increase. Major extinctions have caused marked reductions in the diversity of life at several periods in the Earth's history. The most recent major or mass extinction occurred 65 Ma ago and is thought to have been caused by the impact of a large comet or asteroid, a reminder of events that may have frustrated the evolution of life on the early Earth.

QUESTION 2.7

As a thinking exercise, suggest how the evolutionary trends observed for life on Earth might help us answer the question about the possible nature of life elsewhere. What assumptions do you think we need to make before undertaking such extrapolations?

2.5 Life on the edge

2.5.1 Introduction

From the preceding discussion it seems that life on the early Earth may well have arisen in an environment that we would consider extreme when compared to those that exist on much of our planet today. Evidence from both the geological record and phylogeny suggests that the first organisms on Earth may have been the heat-tolerant thermophiles or hyperthermophiles. **Extremophile** is a term used to describe the many micro-organisms that are capable of different degrees of adaptability to the extreme range of living conditions available on Earth. Such environments and organisms are likely candidates for having given rise to life on Earth. These general considerations give some support to the idea that similar ecosystems may have emerged elsewhere as well.

Must extremophiles actually 'love' (as the suffix *-phile* implies) their environment or merely tolerate it? It is certainly easier to determine in the laboratory if an organism will simply tolerate extreme conditions and it is common to find organisms in extreme environments that tolerate rather than love them. However, the number of known true extreme-loving organisms from a variety of environments is increasing, confirming that life can exist, and indeed thrive, under those conditions.

Figure 2.23 Recovery of the camera from the Surveyor 3 spacecraft by the Apollo 16 astronaut, Pete Conrad. When returned to Earth, a strain of the bacterium *Streptococcus* that was isolated from foam inside the camera was found to have survived exposure on the lunar surface.

The ability of life to survive in extreme environments has been dramatically demonstrated by the recovery of bacteria exposed to the hostile environment of the lunar surface by the astronauts of Apollo 16 (Figure 2.23). Apollo 16 astronauts recovered a camera from the Surveyor 3 spacecraft that had landed on the Moon two and a half years earlier. When returned to Earth, foam from the inside of the camera was examined to see if any bacteria on it had survived their journey to the lunar surface. The 50–100 organisms recovered survived launch, space vacuum, 3 years of radiation exposure, deep-freeze at an average temperature of 20 K, and no nutrient, water or energy source. However, the organisms effectively did nothing while they sat on the lunar surface, they were in effect freeze-dried. These were not extremophiles, they merely survived. An important observation about extremophiles is that these organisms do not merely tolerate their lot; they do best in their punishing habitats and, in many cases, require one or more extremes in order to reproduce at all.

Studies of extremophiles are responsible for a marked change in evolutionary theory that has given rise to the phylogenetic tree you met in Section 1.9. It had been thought that living organisms could

be grouped into two basic domains: bacteria, whose simple cells lack a nucleus, and eukarya, whose cells are more complex. We now know that a third group, the archaea, exists. Anatomically, the archaea lack a nucleus and closely resemble bacteria – some of their genes have similar counterparts in bacteria, a sign that the two groups function similarly in some ways. But the archaea also possess genes otherwise found only in eukarya, and a large fraction of their genes appear to be unique. These unshared genes establish the archaea's separate identity.

So what are the physical limits to life on Earth and what sort of organisms thrive under extreme conditions? Table 2.2 summarizes our present state of knowledge of the physical limits to life. We'll examine the organisms that live in these environments in the following sections.

Table 2.2 The physical limits for life on Earth, with examples of some of the organisms associated with particular environments.

Environment	Limiting conditions	Type	Example
temperature	<15 °C	psychrophiles	
	15–50 °C	mesophiles	*Homo sapiens*
	50–80 °C	thermophiles	*Thermoplasma* can reproduce at >45 °C
	80–115 °C	hyperthermophiles	*Pyrolobus fumarii* (113 °C)
radiation			*Deinococcus radiodurans*
salinity	15–37.5% NaCl	halophiles	
ph	0.7–4	acidophiles	
	8–12.5	alkalophiles	
dessication	anhydrobiotic	xerophiles	nematodes, microbes, fungi, lichens
pressure	pressure-loving – up to 130 MPa	piezophiles	
	weight-loving	barophiles	
vacuum	tolerates vacuum		microbes, insects, seeds
oxygen	cannot tolerate O_2	anaerobes	
	tolerates some O_2	microaerophiles	
	requires O_2	aerobes	*Homo sapiens*
chemical extremes	gases		*C. caldarium* (pure CO_2)
	can tolerate high concentrations of metals		

2.5.2 Temperature

Temperature presents a range of challenges to living organisms. The structural breakdown of cells caused by the formation of ice crystals in sensitive plants can be readily witnessed in those parts of the world that experience cold winters or even just the occasional frosty night. At the other extreme, high temperatures result in the structural breakdown of biological molecules such as proteins and nucleic acids, a process known as denaturation. High temperatures increase the rate at which material diffuses through cell membranes, the membrane fluidity. Temperatures of 100 °C can disrupt the structural integrity of cell membranes to the extent that they leak important cellular constituents.

Life on Earth has adapted to a surprising range of temperatures (Figure 2.24). Although the majority of organisms grow best at moderate temperatures of between 20 °C and 45 °C (the **mesophiles** in Figure 2.24), the temperature preferences of other organisms range from hyperthermophiles (able to reproduce at temperatures >80 °C) to psychrophiles where maximum growth occurs at temperatures <15 °C.

Thermophilic organisms are among the most studied extremophiles. The archaea *Thermoplasma* (Figure 1.37), for example, found in volcanic hot springs, can reproduce at temperatures in excess of 45 °C. Hyperthermophiles, such as the archaea *Sulfolobus* (Figure 1.37) have been recovered from environments where they are exposed to temperatures in excess of 100 °C. By comparison, most regular bacteria thrive at temperatures between 25 °C and 40 °C. No multicellular animals or plants are known that can tolerate temperatures above 50 °C and no microbial eukarya are able to tolerate long-term exposure to temperatures above 60 °C.

Thermophiles that are content at temperatures up to 60 °C have been known for a long time, but true extremophiles, those able to flourish in greater heat, were first discovered in the 1960s during a study of microbial life in hot springs and other waters of Yellowstone National Park in the USA. The first extremophile reported to be capable of growth at temperatures greater than 70 °C was the bacterium *Thermus* (Figure 1.37).

To date, more than 50 species of hyperthermophiles have been isolated, the most resistant of which, *Pyrolobus fumarii*, grows in the walls of black smokers on the ocean floor. It reproduces best in an environment of about 105 °C; it won't grow at all at temperatures below 90 °C. Another hyperthermophile that lives in deep-sea hydrothermal systems is the methane-producing archaean *Methanopyrus* (Figure 1.37).

QUESTION 2.8

Look at Figure 1.37. Where on the phylogenetic tree can species such as *Methanopyrus, Thermoplasma,* and *Sulfolobus* be found? How does their position 'fit' with the concept of the last common ancestor you met in Section 1.9?

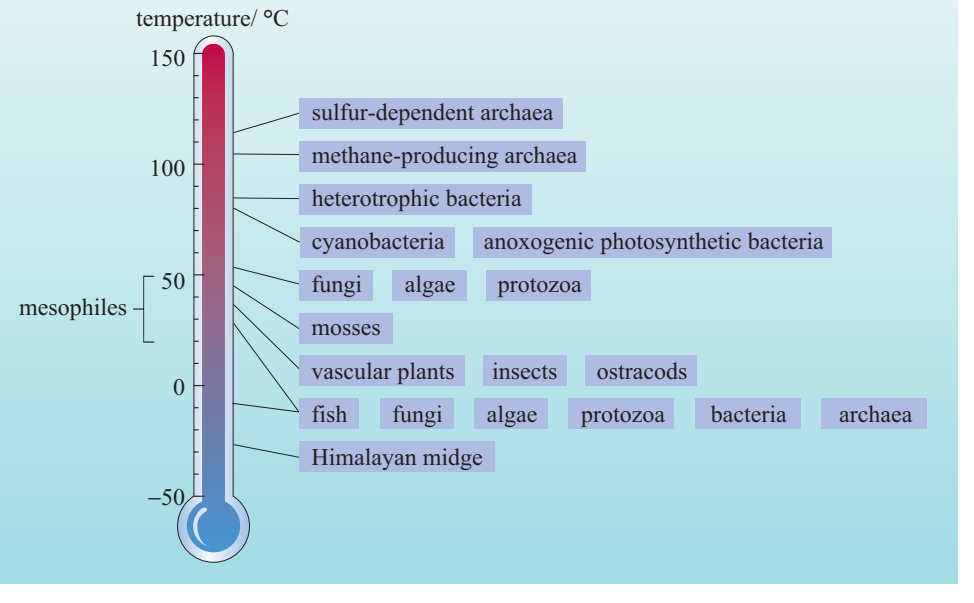


Figure 2.24 The temperature limits for major groups of organisms.

What is the upper temperature limit for life? Do 'super-hyperthermophiles' capable of growth at 200 °C or 300 °C exist? At present we do not know, although it seems likely that the limit will be about 140–150 °C; this is the maximum temperature at which activity has been observed for hyperthermophile enzymes. Above this temperature, proteins and nucleic acids denature, so that a loss in the integrity of DNA and other essential molecules would probably prevent reproduction.

Denaturation involves the unfolding of the double helix of DNA.

So how have organisms adapted to these high temperatures? Since high temperatures increase membrane fluidity, one adaptation is to change the composition of the membrane to reduce that fluidity. For example bacteria will alter the ratio of different lipids in their membranes in response to the temperature at which they are grown. Thermophiles and hyperthermophiles have also evolved proteins that are better able to cope with higher temperatures. The DNA of hyperthermophiles, which would otherwise denature at temperatures above 70 °C, is more stable *in vivo* ('in life') than that of mesophiles. This reflects the fact that the G–C pair of nucleic acids is more thermally stable than the A–T or A–U pairs because of the additional hydrogen bond. Elevated G + C to A + T or A + U ratios are found in the ribosomal and transfer RNAs of thermophiles.

On Earth, cold environments are actually more common than hot ones. The Earth's oceans maintain an average temperature of 1–3 °C. However, large areas of the Earth's surface are permanently frozen or are unfrozen for only a few weeks in summer. Some of these frozen environments support life in the form of psychrophiles. Representatives of all major groups of organism are known from environments with temperatures just below 0 °C. Freezing in liquid nitrogen at a temperature of –196 °C can preserve many microbes successfully. However, the lowest recorded temperature for active microbial communities is substantially higher, at –18 °C. A typical example of a psychrophile is the bacterium *Polaromonas vacuolata*: its optimal temperature for growth is 4 °C and it finds temperatures above 12 °C too warm for reproduction. Liquid water is both a solvent for life and an important reactant or product in most metabolic processes. When water freezes the resulting ice crystals can rip cell membranes apart, and solution chemistry stops in the absence of liquid water. Freezing of the water inside cells is almost invariably lethal. The only exception to this rule reported so far, outside of cryopreservation, is the nematode *Panagrolaimus davidi*, which can withstand freezing of all body water.

As with thermophilic organisms, psychrophiles have evolved adaptations to the problems of adverse temperatures. While high temperatures increase membrane fluidity, low temperatures result in a decrease in membrane fluidity. In response, psychrophilic organisms adjust the ratios of lipids in their membranes to improve the fluidity of the cell membrane. Two principal adaptations have evolved to deal with temperatures below the freezing point of water: the protection of cells from ice formation by preventing freezing or, if ice does form, protection of the cells during thawing. One way organisms prevent ice forming is to accumulate soluble compounds that can depress the normal freezing point of water. Increased concentrations of salts and sugars can achieve this, but organisms also produce relatively inert molecules, notably glycerol, specifically for this purpose. For example, high concentrations of glycerol can enable the survival of some invertebrates to temperatures as low as –60 °C. The teleost fish, which inhabit polar regions, manufacture specific proteins which can effectively act as antifreeze agents by binding to the edges of ice crystal lattices and preventing the addition of further water molecules. This phenomenon, known as **thermal hysteresis**, depresses the freezing point of water well below its melting point, hence these proteins are known as thermal hysteresis proteins.

2.5.3 Radiation

Radiation is energy in the form of waves or particles, such as electromagnetic radiation (e.g. gamma rays, X-rays, ultraviolet radiation, visible light, or infrared radiation) and particles (neutrons, protons, electrons, or alpha particles). While very high levels of radiation do not generally occur naturally on Earth, the effects of radiation on living organisms have been well studied as a result of research on the use of radiation in medicine and on the consequences of human activity ranging from warfare to space travel. UV and ionizing radiation can cause serious damage to DNA by modifying the nucleotide bases or causing single or double-strand breaks in DNA. One organism that is known to withstand exceptional levels of ionizing radiation and that probably qualifies as a radiation extremophile is the bacterium *Deinococcus radiodurans*, first discovered in 1956. It can withstand exceptionally high doses of UV and gamma radiation. The ability to survive such extreme environments is attributed to *D. radiodurans*' ability to repair damaged DNA since it has redundant strands of DNA, i.e. its genetic code repeats itself many times. This enables damage in one area, e.g. the double-strand breaks in DNA caused by ionizing radiation, to be recognized and quickly repaired. This extraordinary resistance is thought to be a consequence of evolutionary adaptations to cope with extreme desiccation so that *D. radiodurans* may in fact be a xerophile (see Table 2.2).

2.5.4 pH

pH, which ranges on a logarithmic scale from 0 to 14, measures the concentration of hydrogen ions (H^+) in solution. Biological processes tend to occur towards the middle range of the pH scale so that typical environmental pH values also fall within this range, e.g. the pH of seawater is ~8.2. Some extremophiles are known that prefer highly acidic or alkaline conditions, the **acidophiles** and **alkaliphiles**. Acidophiles thrive in the rare habitats having a pH of between 0.7 and 4, and alkaliphiles favour habitats with a pH between 8 and 12.5.

Highly acidic environments can occur naturally from geochemical activities. For example, the production of sulfur-rich gases in deep-sea hydrothermal vents and at some hot springs. However, acidophiles are not able to tolerate a significant increase in acidity inside their cells, where it would destroy important molecules such as DNA. Thus they survive by keeping the acid out. But the defensive molecules that provide this protection, as well as others that come into contact with the environment, must be able to operate in extreme acidity. Indeed, enzymes have been isolated from acidophiles that are able to work at a pH of less than 1.

Alkaliphiles live in soils laden with carbonate and in so-called soda lakes, such as those found in Egypt, the Rift Valley of Africa and the western USA. Above a pH of 8 or so, certain molecules, notably those made of RNA, break down. Consequently, alkaliphiles, like acidophiles, maintain neutrality inside their cells.

2.5.5 Salinity

Organisms can live within a range of salinities, from essentially distilled water to saturated salt solutions. **Halophiles** are organisms that require high concentrations of salt in order to live. Their optimal NaCl concentrations for growth range from twice to nearly five times the salt concentration of seawater. They are found in habitats like the Great Salt Lake (Figure 2.25), Dead Sea and salterns (evaporation basins for obtaining salt). Some high-salinity environments are also extremely

alkaline because weathering of sodium carbonate and certain other salts can release ions that produce alkalinity. Not surprisingly, microbes in those environments are adapted to both high alkalinity and high salinity.

Halophiles have a particular adaptation that allows them to tolerate a high salt environment. Under normal conditions, water tends to flow across a semi-permeable membrane such as a cell wall from areas of low salt concentration to areas of higher concentration, a process known as **osmosis**. Thus a cell suspended in a very salty solution will lose water and become dehydrated unless it contains a higher concentration of salt (or some other solute) than its environment. Halophiles contend with this problem either by producing large amounts of an internal solute or by retaining a solute extracted from outside the cell. For example, the archaean *Halobacterium salinarum* concentrates potassium chloride in its interior. As with the hyperthermophiles, these adaptations will not work under more normal salinities.

Figure 2.25 The Great Salt Lake, Utah, seen from the Shuttle Atlantis. The area of the image is around 200 km by 200 km.

2.5.6 Dessication

You saw in Sections 1.1.3 and 2.1 that its high melting and boiling points and the wide range of temperatures over which it remains liquid makes water an essential solvent for life. Water limitation therefore represents a particularly extreme environment for life. Some organisms can tolerate extreme desiccation by entering a state of apparent suspended animation known as **anhydrobiosis**, characterized by little intracellular water and no metabolic activity. It is well documented in organisms such as bacteria, yeast, fungi and plants and animals associated with environments where the water-film essential for active life is often transient and sporadic. When the film dries out these organisms appear to be dead for periods of days, weeks, or even years until moisture returns, when they 'come back to life' and resume their normal activities.

2.5.7 Pressure

Terrestrial plants and animals at the Earth's surface have evolved at normal atmospheric pressure (101 kPa = 1 atmosphere). However, hydrostatic pressure increases with depth in the oceans so marine organisms may have to deal with much higher pressures. Atmospheric pressure also decreases with altitude, so that by 10 km above sea-level, it is around one quarter of the atmospheric pressure at sea-level.

■ What effect will decreasing atmospheric pressure have on the boiling point of water?

❑ The boiling point of water decreases with decreasing pressure.

Conversely, the boiling point of water increases with increasing pressure so that water in the Earth's deepest ocean basins will remain liquid at temperatures as high as 400 °C.

Pressure presents problems to life because it forces volume changes, for example when pressure increases, the molecules in cell membranes pack more tightly, resulting in decreased membrane fluidity. Organisms that can tolerate high

pressures have often adapted the compositions of their cell membranes to increase fluidity. Similarly, any biochemical reaction that results in an increase in volume, as many do, will be inhibited by an increase in pressure. Pressure-loving piezophiles (see Table 2.2) have been recovered from the Earth's deepest sea floor, the Mariana Trench, where they thrive at pressures of 70–80 MPa, but will not grow at pressures below 50 MPa.

Gravity also has an effect on the forces experienced by an organism. However, until recently, all organisms on Earth have lived at 1 *g*. The advent of space exploration means that humans have had to deal with a range of different gravity regimes, from the variable *g* experienced during launch to microgravity environments on board the International Space Station (ISS). Although most research concerned with microgravity has concerned human health, studies on board the ISS have demonstrated that gravity plays an important role in a variety of biological processes. Some effects of microgravity were expected in organisms that were adept at perceiving gravity, such as the root tips in plants. What was unknown was whether gravity played a role at the sub-cellular level where the force of gravity is almost negligible when compared with the forces governing molecular interactions. Scientists now believe that there are conditions in which the weightless environment influences the cellular machinery fundamentally resulting in specific changes to cell membranes and the reproduction of micro-organisms.

2.5.8 Oxygen

For much of its early history the Earth was an anaerobic environment. Today oxygen plays a crucial role in life on Earth and organisms inhabit environments ranging from strictly anaerobic to aerobic. Oxygen plays a key part in the mechanisms that sustain plant and animal life, photosynthesis and respiration and it is the subtle balance between the consumption of oxygen in respiration and its production in photosynthesis that is critical for the stability of the oxygen level in the Earth's atmosphere. Aerobic metabolism is far more efficient than anaerobic metabolism, but it comes at a price. Molecular oxygen can cause considerable oxidative damage to living organisms and has been implicated in a variety of human health problems, from cancer to ageing. UV radiation can produce reactive oxygen species such as hydrogen peroxide (H_2O_2) and the same species can be produced in aerobic metabolism. As a result, some organisms have evolved mechanisms to avoid or repair the effects of oxidative damage by producing antioxidants.

2.6 Extreme environments

The shear diversity of life on Earth makes it impossible to do a complete survey of even the Earth's more extreme environments in a few pages.

> The continued discovery of new extreme environments and the organisms that inhabit them has made more plausible the search for life on other bodies in the Solar System such as Mars and Europa.

Given that many of these environments that appear to be extreme on Earth may be analogous to the normal environments for other planetary bodies, we'll examine a few of them that may have a role to play in either the origin of life or providing suitable habitats in otherwise hostile environments.

Hotsprings

Hotsprings and geysers such as those in volcanic areas of New Zealand (Figure 2.26) are characterized by hot water, steam and sometimes low pH and toxic metals such as mercury. They are, nonetheless, environments that sustain a remarkably diverse range of life. The range of colours visible in Figure 2.26 reflects different algal populations growing around the Waiotapu hot springs in New Zealand.

Figure 2.26 Hot spring in Waiotapu Park, Rotorua in North Island, New Zealand. The various colours around the edge are due to microbial mats formed by organisms that thrive in different temperature and pH environments.

The deep sea

The deep-sea environment has high pressures and both heat and cold. In the vicinity of hydrothermal vents water temperatures may be as high as 400 °C.

■ Why does water not boil at hydrothermal vents?

❑ Hydrostatic pressure keeps the water liquid as it raises its boiling point.

Hydrothermal vents have pH ranges from around 3 to 8 and as you saw in Section 1.7.3 were possibly critical to the evolution of early life on Earth, a conclusion supported by phylogenetic evidence that suggests that thermophiles were the last common ancestor (Section 1.9).

The fact that life can exist, and indeed thrive, at depth in the Earth's oceans without the need for photosynthesis has significant implications for the plausibility of environments for life elsewhere in the Solar System. As you'll see in Chapter 4, Jupiter's moon Europa may harbour a subsurface ocean of liquid water that lies below a layer of ice too thick to allow photosynthesis.

Hypersaline environments

Hypersaline environments include salt-flats, evaporation ponds, natural lakes (e.g. the Great Salt Lake in Utah, USA) and deep-sea hypersaline basins. Halophilic organisms are often the dominant organisms in these environments, tolerating salinities of more than 30%.

Evaporites

Evaporite deposits consisting of the minerals halite (NaCl), gypsum ($CaSO_4.2H_2O$) or anhydrite ($CaSO_4$), are well known from the geological record and are still widespread. These deposits frequently contain algal or bacterial communities and micro-organisms have been found trapped in inclusions inside crystals where they have been observed to remain viable for periods up to a year. This has led scientists to speculate that bacteria could survive for millions of years in the inclusions of salt deposits, a controversial but intriguing suggestion.

Deserts

Deserts are extremely dry and can be either hot or cold. Water is always the limiting factor in such ecosystems. The Atacama desert in Chile is one of the hottest and driest areas while the coldest and driest places on Earth are the so-called dry valleys of Antarctica. The primary inhabitants of both kinds of desert ecosystems are bacteria, algae and fungi that live on or a few millimetres below the surfaces of rocks. Organisms that have adapted to living on, or beneath, the surfaces of rocks are referred to as **endoliths**, meaning literally 'within rock'. Those that exist a few millimetres below the surface of a rock are also known as **cryptoendoliths**.

- Can you think of an advantage to living inside a rock?

❑ For simple organisms, adopting an endolithic lifestyle can provide protection against extremes of desert temperature. Protection from UV radiation is also an advantage.

The atmosphere

Airborne micro-organisms are known to exist. They have to withstand desiccation and exposure to UV radiation to survive. However, it is less clear whether these organisms constitute a viable aerial ecosystem or are merely the dormant spores of surface biota.

Ice, permafrost and snow

Microbial life frequently uses frozen water as a habitat and pink algal blooms on snow and ice are common features. However, ice environments essentially contain survivors – it seems unlikely that the inhabitants of these environments actually prefer it, but have found themselves trapped in the ice and are more resistant than other organisms that perished.

Subsurface environments

As you will see in Chapter 3 when we examine the possibility of life on Mars, it would be hard for life to survive the harsh conditions present today on the Martian surface. One or two of the organisms we've examined in this section could hypothetically withstand one or more of the Martian extremes, though they would need some protection. However, Mars, for the most part, is frigid, it receives 43% as much solar radiation as the Earth but the thin CO_2-rich atmosphere absorbs little harmful UV radiation and atmospheric pressure is too low for water to be stable at the surface.

■ Of the various extremophiles covered in this section, which do you think might prove most resistant to the arid and exposed conditions on the Martian surface?

❑ One of the toughest we've mentioned is *Deinococcus radiodurans*, which has evolved to cope with high radiation conditions and desiccation.

As you'll learn in Chapter 3, the search for extant life on Mars is therefore focused on the possibility of a subsurface biota. The plausibility of subsurface life on other planets has been enhanced by the discovery on Earth of subsurface microbial ecosystems that are not dependent on photosynthesis. Instead, these microbes appear to thrive on chemical energy in basalt, a rock common to Earth, Mars and other terrestrial planets, but which contains little of the organic nutrients that normally feed micro-organisms. These microbes were found in groundwater samples taken more than 1 000 metres below the surface of the Columbia Basin basalt flows in the western USA. Called a subsurface lithoautotrophic microbial ecosystem, given the acronym **SLiME**, this community of microbes exists on a diet of mostly hydrogen. Hydrogen-utilizing bacteria have been identified before, but SLiME appears to get its hydrogen in an unusual way. Other microbes depend on organic carbon or hydrogen from decaying plant matter that was originally generated from photosynthesis, but SLiME apparently consume hydrogen given off by a reaction between basalt and groundwater passing through the rocks. The basalt–water–microbial relationship has been confirmed in the laboratory when microbes were grown in a basalt–water mixture.

The basalt connection suggests that it might be possible for micro-organisms to exist in the Martian subsurface. Until recently, it was believed that all water under the surface of Mars was frozen, but hydrogeological data from the Mars Global Surveyor suggests that liquid water may flow today under the Martian surface (see Section 3.4). If this is confirmed, then the subsurface conditions on Mars are similar to the basalt and water subsurface environments of the Columbia Basin area. This does not mean there is life on Mars, but if SLiME can exist here, then, in theory, it could exist on Mars too.

2.7 Summary of Chapter 2

- Circumstellar habitable zones encompass the range of distances from a star for which liquid water can exist on a planetary surface. A continuous habitable zone defines the region around a star in which a planet may reside and maintain liquid water on its surface throughout most of its life.
- Plate tectonics have played a crucial role in maintaining the Earth's habitability throughout its history through the recycling of carbon by decarbonation and the outgassing of CO_2.
- The Earth probably acquired its water (and hence its oceans) from hydrated minerals that eventually became incorporated in the interiors of planetary embryos.
- The Earth's early atmosphere was dominated by N_2, CO_2, possibly sulfur oxides such as SO_2, and water vapour, but only trace amounts of O_2. However, that atmosphere began to change with the emergence of life. The geological record contains evidence in the form of banded iron formations for the emergence of oxygenic photosynthesis during the first billion years of the Earth's history.

- Geological evidence, in the form of stromatolites, provides some evidence for the emergence of life possibly as early as 3.5 Ga ago. Carbon isotopes provide evidence for the operation of biological carbon fixation in rocks as old as 3.8 Ga. However, both lines of evidence are controversial.
- Life on Earth has become progressively more complex, and larger. However, large organisms evolved comparatively recently over the last 600 Ma.
- Extremophiles are organisms that have adapted to living under some of the more extreme environmental conditions on Earth such as high and low temperatures, extremes of pH and salinity, and high radiation environments. Adaptations, particularly to their cell membranes, enable these organisms to thrive in otherwise hostile environments.
- Many of the Earth's extreme environments, and the organisms that inhabit them, provide useful analogues for understanding how life may be able to exist on other planetary bodies in our Solar System and beyond.

CHAPTER 3
MARS

3.1 Introduction: Mars and life

Of all the bodies in the Solar System other than Earth, it has almost always been Mars that has dominated discussions of the possible existence of life, whether it is extinct or still present (extant). Even in ancient times, both mythology and informed thinking suggested that Mars might be inhabited. This belief prevailed through the Middle Ages even up to the modern scientific epoch. By the advent of the space age, the idea of any advanced life forms had all but disappeared. But even in 1960, when the first space mission to Mars was attempted, the seasonal changes in colour which had been observed from the Earth were still taken by some to indicate that Mars was vegetated.

Arguably the most notable (and possibly notorious) of observers of Mars was the 19th century Italian astronomer Giovanni Schiaparelli (Figure 3.1) who, apart from his scientific contributions, made an everlasting contribution to our cultural awareness of Mars. At a time when astronomical observations relied on attention to detail and outstanding eyesight, he would draw features which photographic plates could not record. He announced, in 1878, that he had observed extensive straight lines or streaks on the surface of Mars (Figure 3.2). Schiaparelli treated his findings with great caution and at first doubted his own observations. But successive observations convinced him of their veracity. In his description of these observations, he used the Italian word 'canali' which means 'channels' or 'grooves'. This generated enormous interest and many scientists and writers in the English-speaking world translated this word as 'canals', implying artificial waterways presumably created by intelligent beings. (Who could have imagined then that some 90 years later, spacecraft would photograph canyons and valleys

Figure 3.1 Giovanni Virginio Schiaparelli (1835–1910), Italian astronomer and director of the Milan Observatory, who discovered the relationship between comets and meteor streams in 1866. He was most famous for his meticulous observations of Mars (1877–90), including features that became known as 'Martian canals' (Figure 3.2). He continued to observe Mars faithfully until his eyesight failed.

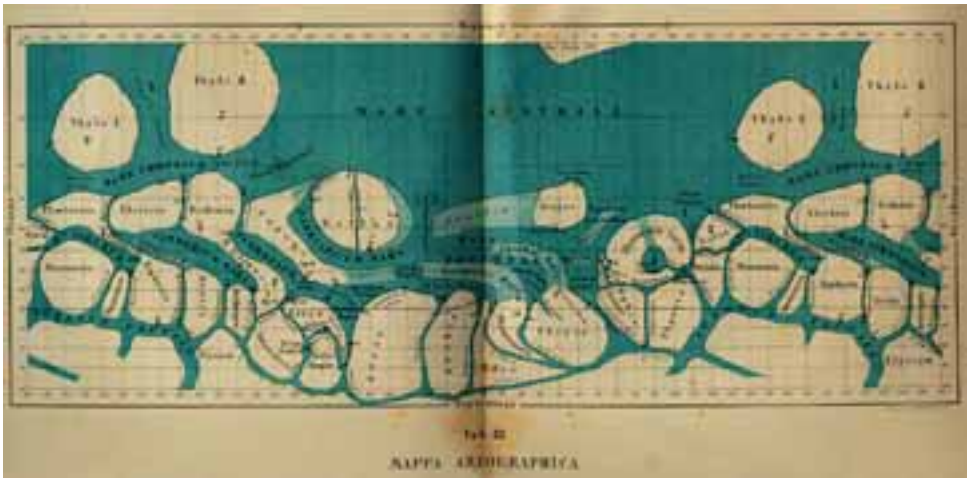


Figure 3.2 One of Schiaparelli's many sketches of Mars, this one completed in 1881, based on his telescopic observations. These led to a 'craze' amongst astronomers, both amateur and professional, to search for evidence for intelligent life and extravagant claims by some for positive evidence. Schiaparelli remained rather moderate in his assertions about 'canals', suggesting that they might be natural rather than artificial structures.

on Mars hundreds of kilometres long and produced at one time by flowing water?) For nearly a century the idea of life on Mars was embedded in popular imagination and was a fertile source for both serious literature (for example H. G. Wells' *War of the Worlds*, first published in 1898), and pulp fiction in its many manifestations (Figure 3.3). One of those most inspired by Schiaparelli's observations was Percival Lowell (1855–1916), an American polymath who devoted his energy (and fortune!) to the study of Mars (Figure 3.4). He advanced a theory in his lectures and writings that the 'canals' were a result of attempts by struggling Martian inhabitants to irrigate the planet from the melting polar icecaps!

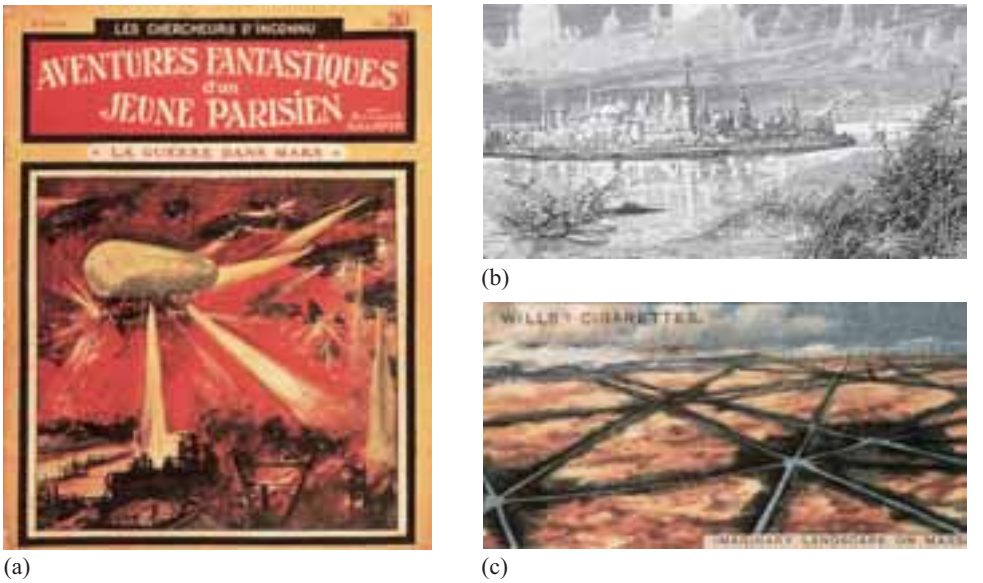


Figure 3.3 Illustrations from several examples of the treatment of life on Mars in literature. (a) The cover picture by an unknown artist of *La Guerre dans Mars,* (b) an illustration of a floating Martian city by Paul Handy in *Letters from the Planets* (1890) and (c) a cigarette card from Will's cigarettes.

Figure 3.4 Percival Lowell (1855–1916), a Boston-born American, who spent his life devoted to business, travel, literature and astronomy. He became widely known for his theory that the Martian 'canals' were a result of attempts by the struggling inhabitants to irrigate the planet from the melting polar icecaps. He founded the great observatory that bears his name at Flagstaff, Arizona, initially with the exclusive intention of confirming the presence of advanced life forms on the planet. Although his theories met with widespread opposition, he received numerous honours during his life. Nearly 14 years after Lowell's death, Clyde Tombaugh discovered the planet Pluto from the Lowell observatory, a discovery for which Lowell had paved the way with his calculations concerning the gravitational perturbations to the movement of the planet Neptune.

In recent years, as a result of developments in our knowledge of life on Earth, the known abundance of the elements, the fundamentals of organic chemistry and our knowledge of the Martian environment, the belief that some form of life has at some time existed on Mars has been strengthened in the view of many (though not all) planetary scientists. However, despite enormous steps forward in both knowledge and understanding, this is still a matter of some controversy, and unequivocal evidence is still awaited. In fact, in recent times, the consensus about life on Mars has ebbed and flowed with its perceived likelihood rising and falling as the latest results and theories, both from Mars-based measurements or developments here on Earth, are digested.

But without any detailed knowledge of Mars, are there any reasons for believing that Mars might be a habitat for life?

■ Recall from Section 2.3.2 the position of the outer edge of the Sun's habitable zone. How does this relate to the position of Mars?

❑ Models for the outer edge of the present habitable zone place it between 1.37 AU (where CO_2 starts to condense) and 1.67 AU (the point at which a maximum greenhouse effect would operate). Mars is located at 1.52 AU, placing it inside the outermost estimate of the Sun's habitable zone.

There is one prerequisite for life which has dominated the issue of the existence or not of life on Mars.

■ What is this prerequisite?

❑ The need for liquid water (see Section 1.2.1).

The search for water has thus become inextricably linked to the search for life itself. Whereas it is clear that water did once exist on Mars (Figure 2.4), the question of when it disappeared, or even if it has completely disappeared, has become a dominant issue. As one of the scientists involved in this search has said:

> 'Following the water makes sense if you're prospecting for biology. If we could find evidence of preserved liquid water on Mars, that would be the Holy Grail.'

3.2 Background

Up to the end of the year 2002, over 30 space missions had attempted to explore Mars (Table 3.1). Starting with an unsuccessful Soviet mission (official designation Mars 1960A), which failed on launch in 1960, the space exploration of Mars was dogged by many failures, especially in the early days of the space age.

However, a series of extremely successful missions, coupled with Earth-based telescopic observations, has given us a fairly thorough picture of the basic facts about Mars. The first successful space mission was the Mariner 4 fly-by in 1964 and there followed various missions (both from the USA and the USSR) over the next 11 years, leading to the successful Viking 1 & 2 orbiters and landers. However, the results from the early missions were not encouraging for the proponents of

Table 3.1 Spacecraft Missions to Mars ordered by date of launch.

Mission	Launch Date	Remarks
Mars 1960A (USSR)	10 Oct 1960	attempted fly-by; launch failure
Mars 1960B (USSR)	14 Oct 1960	attempted fly-by; launch failure
Mars 1962A (USSR)	24 Oct 1962	attempted fly-by; failed to leave Earth orbit
Mars 1 (USSR)	1 Nov 1962	fly-by; lost contact in transit
Mars 1962B (USSR)	4 Nov 1962	attempted lander; failed to leave Earth orbit
Mariner 3 (USA)	5 Nov 1964	attempted fly-by
Mariner 4 (USA)	28 Nov 1964	fly-by, imaging
Zond 2 (USSR)	30 Nov 1964	fly-by, lost contact in transit
Mariner 6 (USA)	24 Feb 1969	fly-by, imaging & atmospheric measurements
Mariner 7 (USA)	27 Mar 1969	fly-by, imaging & atmospheric measurements
Mars 1969A (USSR)	27 Mar 1969	attempted lander; launch failure
Mars 1969B (USSR)	2 Apr 1969	attempted lander; launch failure
Mariner 8 (USA)	8 May 1971	attempted lander; launch failure
Cosmos 419 (USSR)	10 May 1971	attempted orbiter/lander
Mars 2 (USSR)	19 May 1971	orbiter; lander crashed on surface
Mars 3 (USSR)	28 May 1971	orbiter; lander lost contact
Mariner 9 (USA)	30 May 1971	orbiter; imaging of Mars, Phobos and Deimos
Mars 4 (USSR)	21 July 1973	fly-by imaging; attempted orbiter
Mars 5 (USSR)	25 July 1973	orbiter; imaging
Mars 6 (USSR)	5 Aug 1973	orbiter; lander lost contact, some data
Mars 7 (USSR)	9 Aug 1973	orbiter; attempted lander
Viking 1 (USA)	20 Aug 1975	orbiter and lander in Chryse Planitia
Viking 2 (USA)	9 Sept 1975	orbiter and lander in Utopia Planitia
Phobos 1 (USSR)	7 July 1988	attempted Mars orbiter and Phobos landers
Phobos 2 (USSR)	12 July 1988	Mars orbiter, some imaging before failure; Phobos landers failed
Mars Observer (USA)	25 Sept 1992	orbiter; contact lost during Mars orbit entry
Mars Global Surveyor (USA)	7 Nov 1996	orbiter, arrived 12 Sept 1997
Mars 96 (Russia)	16 Nov 1996	attempted orbiter/landers; launch failure
Mars Pathfinder (USA)	4 Dec 1996	lander/rover, landed 4 July 1997 in Ares Vallis
Planet-B, Nozomi (Japan)	4 July 1998	orbiter, atmospheric probe; arrival delayed to Dec 2003
Mars Climate Observer (USA)	11 Dec 1998	orbiter, lost on arrival at Mars 23 Sept 1999
Mars Polar Lander/Deep Space 2 (USA)	3 Jan 1999	lander/descent probes, lost on arrival 3 Dec 1999
Mars Odyssey (USA)	7 Apr 2001	orbiter, currently conducting prime mission of science mapping
Mars Express/Beagle 2 (European Space Agency)	June 2003	orbiter/lander, arrival Dec 2003, prime aim is search for water & life from orbit and the surface
Mars Exploration Rovers (USA)	May–July 2003	2 identical rover missions to separate landing sites. Arrival Jan 2004, prime aim is search for past water

Mars as a habitat for life. It was revealed as a freeze-dried wasteland with a piteously thin atmosphere, bathed in lethal ultraviolet radiation and exposed to cosmic radiation because of an insufficient magnetic field to deflect it.

The basic physical parameters of Mars are: a radius approximately one-half that of Earth (3396 km), resulting in a surface area similar to the land area on Earth, and a mass of around one-tenth (6.419×10^{23} kg) of Earth's.

QUESTION 3.1

From these basic figures, estimate the surface gravity on Mars as a fraction of Earth's surface gravity. (*Hint*: Surface gravity is given by GM/R^2 where M and R are respectively a planet's mass and radius and G is the gravitational constant.)

The surface of Mars is rather cold. At low latitudes, a typical daily temperature range is from −100 °C to +17 °C with a mean of about −60 °C. In winter, the temperatures at the pole can fall to −125 °C.

- Can you think of a simple reason for the low temperature on the Martian surface?

- The most obvious is Mars's distance from the Sun, namely 1.5 AU, from Box 2.1, so the solar flux is $1/(1.5)^2$ times the solar flux at Earth so sunlight on Mars is 2.25 times weaker than on Earth.

In addition, the atmosphere of Mars is extremely thin with an average surface pressure of 6.3 mbar (see Box 3.1), which means that there is a much smaller greenhouse effect operating on Mars than on Earth.

But note that the actual surface pressure varies by up to 2.4 mbar due to seasonal temperature changes and there is a further variation due to differences in height of the Martian surface.

BOX 3.1 UNITS OF PRESSURE

The SI unit of pressure is the pascal, abbreviated to Pa:

$$1\ \text{Pa} = 1\ \text{N m}^{-2} = 1\ \text{kg m}^{-1}\ \text{s}^{-2}$$

You may be more familiar with pressure measured in bar or millibars (1 mbar = 10^{-3} bar), units which are commonly used in meteorology. 1 bar is the mean atmospheric pressure at sea-level on Earth and the conversion to pascals is given by:

$$1\ \text{bar} = 10^5\ \text{Pa}$$

■ How does the surface pressure of Mars compare with that on Earth?

❑ The surface pressure on Earth is about 1000 mbar, so the Martian surface pressure is less than 1% of that on Earth.

■ Express the average surface pressure on Mars in SI units.

❑ Average surface pressure = $(6.3 \times 10^{-3} \times 10^{5})$ Pa = 6.3×10^{2} Pa.

Table 3.2 Composition[a] of the **troposphere** of Mars with sources and sinks of the components where known.

Gas	Volume ratio[b]	Major source[c]	Major sink
CO_2	9.53×10^{-1}	Evaporation, outgassing	Condensation
N_2	2.7×10^{-2}	Outgassing	Escape (as N)
^{40}Ar	1.6×10^{-2}	Outgassing	
O_2	1.3×10^{-3}	CO_2 photodissociation (3.2–3.4)	Photoreduction
CO	7×10^{-4}	CO_2 photodissociation (3.2)	Photooxidation
H_2O	3×10^{-4}	Evaporation, desorption	Condensation, adsorption
^{36}Ar	5×10^{-6}	Outgassing	
Ne	2.5×10^{-6}	Outgassing	
Kr	3×10^{-7}	Outgassing	
Xe	8×10^{-8}	Outgassing	
O_3	$(0.1 \text{ to } 20) \times 10^{-8}$	Photochemistry (3.6)	Photochemistry
NO	7×10^{-5} (at 120 km)	Photochemistry	Photochemistry

[a] Values at the surface unless indicated otherwise.

[b] The **volume ratio** is the fraction by *number* of the atoms or molecules of a species present. Chemists often refer to this as the mole fraction. It is also called the *volume mixing ratio* by some atmospheric scientists. When multiplied by the atmospheric pressure, it gives a quantity called the **partial pressure**, which may be envisaged as the fractional contribution of a component to the total pressure.

[c] Numbers in brackets refer to equation numbers in the text.

The atmosphere of Mars is composed mainly of carbon dioxide, with only a few per cent of N_2 and very minor amounts of other gases, including H_2O (Table 3.2). Being composed largely of CO_2, the atmosphere of Mars resembles that of Venus, especially at high altitude, although the **column mass** (Box 3.2) is very different (see Table A1).

Permafrost is a term used to describe permanently frozen soil, subsoil or other deposits.

In addition to these atmospheric components, reservoirs of H_2O and CO_2 are contained in the polar ice caps (see Figure 3.5) and permafrost. So, not surprisingly, the Viking missions (see Section 3.3) observed relatively large amounts of H_2O in the atmosphere close to the north polar cap, especially during summer when the

BOX 3.2 COLUMN MASS

This parameter gives the mass of gas in a column of unit cross-sectional area (i.e. 1 m^2) extending from the surface of the planet vertically upwards to the very top of the atmosphere. If we know the atmospheric pressure, P, and the gravitational acceleration, g, at the surface, we can calculate the column mass, M_c, as follows:

Pressure = force/area; force = mass × acceleration.

Therefore, pressure = (mass × acceleration)/area.

We can identify (mass/area) with the column mass, M_c.

Therefore, we finally obtain: pressure = column mass × acceleration, or

$$M_c = P/g \qquad (3.1)$$

■ What units are used for column mass?

❑ Column mass is mass divided by area so the units are kg m^{-2}.

Figure 3.5 This image, one of the best of Mars taken from the Earth (or Earth-orbit), was acquired by the Hubble Space Telescope in 1997. It shows many features. For example, the north polar CO_2 ice cap is clearly visible – at the time the image was taken, at the end of Martian spring, it was rapidly receding to reveal the much smaller permanent water-ice cap. It also reveals the circular, dark sea of sand dunes (Olympia Planitia) that surrounds the north pole. Another major feature is the large dark area (Syrtis Major Planitia) just below the centre. Near the southern extremity, clouds of water-ice obscure the giant impact basin, Hellas.

cap is evaporating. Rather less enhancement of H_2O was observed at the south polar cap in its summer. The northern polar cap is believed to consist of CO_2 overlying a residual cap of H_2O ice about 600 km across which is exposed in summer. In contrast, the south polar cap appears to be composed predominantly of CO_2.

But why should the frost or ice, when its temperature is raised, turn straight into the gas or vapour state rather than liquid. To understand this, you need to make use of a **phase diagram** (Box 3.3).

BOX 3.3 PHASE DIAGRAMS

The relationship between the solid, liquid and vapour states of a given substance can be depicted diagrammatically by what is known as the phase diagram (or phase equilibrium diagram) of that substance. Figure 3.6 is a typical example, in which pressure is plotted against temperature for water. The point O, the triple point, is unique, representing the only conditions under which all three phases are in equilibrium with each other. The lines OA, OB and OC show how the equilibrium condition between any two particular phases varies with pressure and temperature.

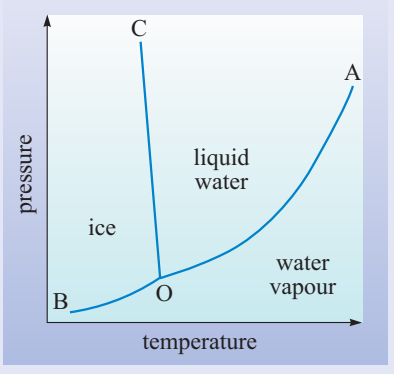


Figure 3.6 The phase diagram for water.

Consider the ambient conditions at the Martian northern polar cap in the context of Figure 3.6. During winter, the conditions correspond to a point low down in the region marked 'ice'. As the temperature rises during spring, the point on the diagram moves to the right through the line OB to the region marked 'water vapour'. So water passes straight from solid (ice) to the gas (water vapour) phase. Water cannot exist in stable form as a liquid at this pressure. Note however that water can exist as a liquid in an unstable form, for example if released catastrophically (at a rate faster than it can evaporate) or if protected by a self-generating skin of ice.

QUESTION 3.2

Given that the triple point of H_2O corresponds to a temperature of 0.01 °C and a pressure of 6.13 mbar, explain with the aid of the phase diagram of H_2O, why liquid water in stable form cannot exist in equilibrium under the average conditions on the surface of Mars. Can you suggest any conditions on Mars under which water might exist in stable form as a liquid?

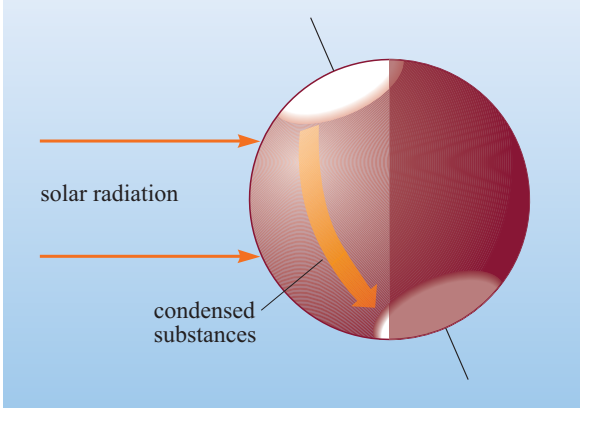


Figure 3.7 During summer in the Martian northern hemisphere, condensed substances (mostly CO_2 with some H_2O) evaporate and migrate to the southern pole, where they condense as ices.

There is a continual seasonal exchange of CO_2 and H_2O between the polar caps via the atmosphere (Figure 3.7). Similarly, the soil exchanges CO_2 and H_2O with the atmosphere seasonally. The amount of these gases absorbed in the surface may be many times the amount in the atmosphere and, via the atmosphere, they exchange with the polar caps. Longer-term variations in the axial inclination and orbit of Mars suggest that the atmospheric pressure may have varied from about 10^{-3} bar to 2×10^{-2} bar, as a result of contributions from the absorbed gas.

This could have had a potentially important implication for the issue of the atmospheric composition of Mars. Currently, the axis of rotation is inclined at 25° to the **ecliptic plane**, similar to the situation on Earth, meaning that the two planets have similar seasons. However, it is believed that the axial inclination of Mars has varied in a chaotic fashion over at least the last 10 Ma. (The axial inclination or obliquity of Mars is influenced by the changes in the gravitational pull of Jupiter and, unlike the case for Earth, it is not stabilized by the presence of a large and nearby Moon.) This means that, according to calculations, this angle has ranged from 15° to 35° (see Figure 3.8).

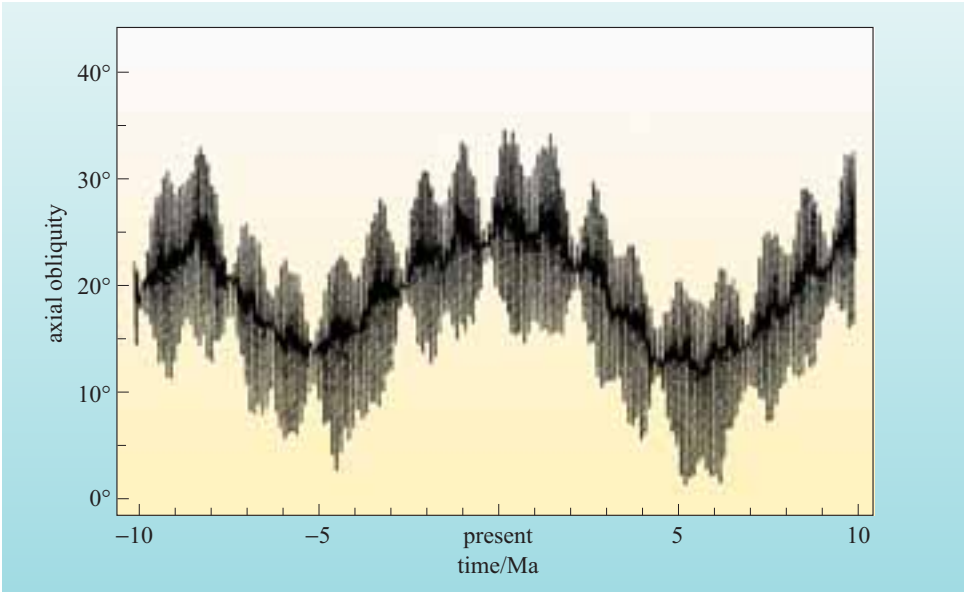


Figure 3.8 Theoretical modelling shows that the Martian axial inclination (or obliquity) should have oscillated with a period of about 100 000 years, with larger amplitude oscillations occurring approximately every million years. In addition, there is a longer-term variation with a period of some 10 Ma. These oscillations modify the energy balance at the poles, causing CO_2 and H_2O to move in and out of the polar regions.

■ Why should this variation of axial inclination be significant?

❑ Its value will affect the amount of sunlight falling on the polar caps, causing changes in temperature and thus modifying the exchange of CO_2 and H_2O between the solid and gaseous phase (i.e. between the polar caps and the atmosphere).

The oscillations are predicted to have a period of about 100 000 years, with the oscillation having a maximum amplitude every million years or so. On top of this is a slow variation with a period of about 10 Ma.

This factor could be significant for the issue of the survival of life on Mars. When the obliquity is at a minimum, the Sun does not rise so high above the horizon at the poles during the summer, resulting in the poles having permanent CO_2 ice caps, since, as on Earth, these regions would receive little sunlight. Conversely, at the other extreme, the CO_2 and H_2O from the polar caps would vapourize and be released into the atmosphere, perhaps raising the pressure sufficiently for liquid water to be stable for short periods. During such periods, any subsurface microbes that might exist could migrate to the surface. Thus even in recent times, there could have been habitable environments on the surface, such as lakes and springs. When water sublimates from the summertime polar cap, it is redistributed globally by the winds, and an annual exchange takes place between the southern and northern caps. At present, several tenths of a millimetre of ice can sublimate from the northern polar cap during summer. However, during times of high obliquity, maybe as much as several tens of centimetres might be removed each year.

Interestingly, evidence for such quasiperiodic climatic changes appears to be shown by the layered deposits of dust and ice at the Martian poles (Figure 3.9). The individual layers are too thick to have formed in a single year. They have probably resulted from the net exchange of water between the two polar caps over 10 000 years or more.

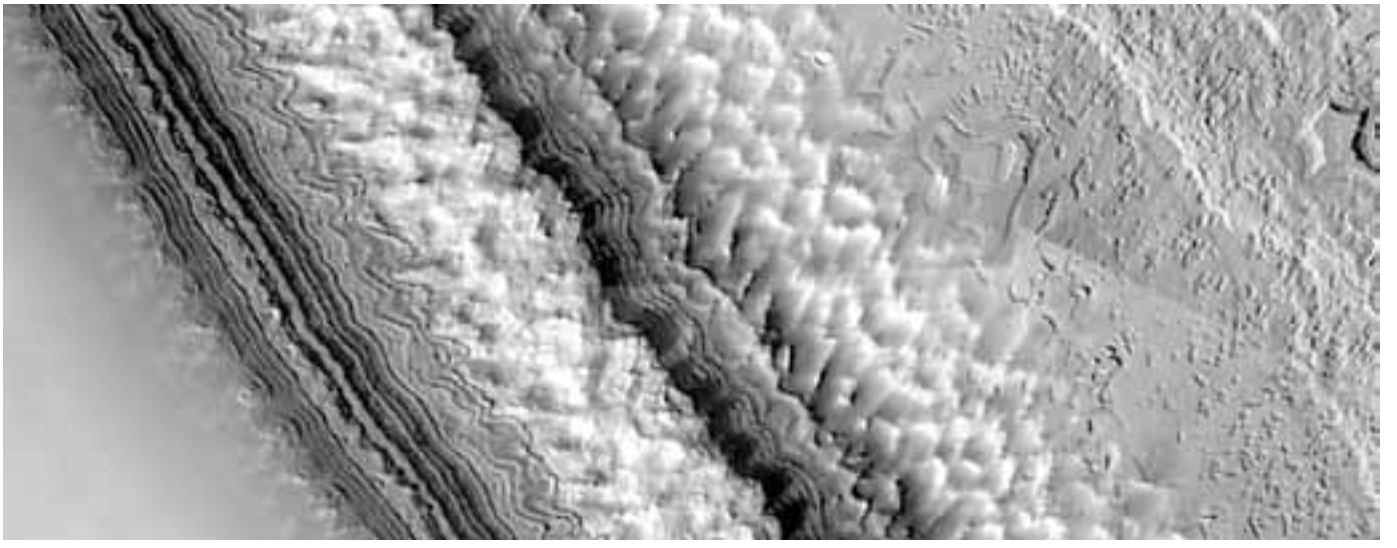


Figure 3.9 Alternating layers of ice (water and carbon dioxide?) and dust (or ice and dust mixed together in differing proportions) are visible in this exotic image of part of the southern polar cap. These layers probably formed as a result of the varying axial inclination (obliquity), and they reflect climatic changes over perhaps the last billion years. This terrain has, to the best of our knowledge, no parallel elsewhere in the Solar System and its structure and appearance are not yet fully explained. The image was produced from a mosaic of images obtained by the Mars Orbiter Camera on the Mars Global Surveyor in October 1999. The region is at latitude 87° S and the image covers an area of about 10 km × 4 km.

Being of such low column mass, the Martian atmosphere is completely exposed to solar UV radiation right to the surface. We therefore expect that CO_2 will dissociate:

$$CO_2 + \text{photon} \longrightarrow CO + O \qquad (3.2)$$

Equation 3.2 is an example of photodissociation or photolysis. In such a process, a molecule dissociates as a result of the absorption of a photon.

However, there is a problem of timescale. In the absence of other reactions, the entire atmosphere of CO_2 would be destroyed in about 3000 years.

The key to the chemistry that maintains CO_2 in the Martian atmosphere is the intervention of H_2O. Thus, recombination of CO and O takes place mainly in the lower atmosphere through the intervention of H, OH and HO_2 (hydroperoxyl). These species occur when H_2O is **photodissociated**. As in the Earth's troposphere, OH acts to **oxidize** CO, that is it reacts to increase the proportion of oxygen in a compound (see Box 3.4). Starting with atomic hydrogen, a sequence of reactions leading to CO_2 can be deduced from laboratory studies according to Equations 3.3 to 3.5:

$$H + O_2 + M \longrightarrow HO_2 + M \qquad (3.3)$$

$$O + HO_2 \longrightarrow O_2 + OH \qquad (3.4)$$

$$CO + OH \longrightarrow CO_2 + H \qquad (3.5)$$

■ What happens to the substance M that enters Equation 3.3 on the left-hand side of the equation?

❑ Nothing! M is an example of a catalyst so it reappears unscathed on the right-hand side.

BOX 3.4 OXIDATION AND REDUCTION

Oxygen occurs in combination with other elements in substances, such as H_2O and CO_2 that are present in the Martian atmosphere. In carbon dioxide, the ratio of oxygen to carbon (O : C) reaches its highest in compounds of these two elements. So in CO_2, the carbon is described as oxidized.

The Earth's atmosphere also contains compounds of carbon in which there is no oxygen. In one example, methane, CH_4, the ratio of hydrogen to carbon (H : C) has its highest value for compounds of these two elements: the carbon is now a **reduced** substance. When combined with the maximum amount of hydrogen, an element is said to be reduced.

Oxygen, which is responsible for the conversion of carbon-containing substances into CO_2, is an **oxidizing** substance. Hydrogen, which can convert carbon-containing substances to CH_4, is called a reducing substance. The conversion processes are referred to, respectively, as **oxidation** and **reduction**. Intermediate levels of oxidation (or reduction) are represented by substances such as carbon monoxide, CO, which also occurs in the Martian atmosphere. Oxidized and reduced are relative terms: for example, CO is oxidized relative to CH_4, but reduced relative to CO_2.

The net effect of these three reactions is the conversion of CO and O into CO_2. It is just one of a number of schemes that can be devised using known chemical reactions of OH and HO_2 that are thought to be effective in regenerating CO_2 in the Martian atmosphere. Note that this sequence involves atomic oxygen, present from the photodissociation of CO_2. Its reaction with molecular oxygen, O_2, accounts for the presence of small amounts of ozone, O_3, via Equation 3.6, one of a set of four reactions known as the Chapman scheme (named after English geophysicist Sydney Chapman, 1888–1970):

$$O + O_2 + M \longrightarrow O_3 + M \qquad (3.6)$$

Both O_3 and HO_2 are powerful oxidizing molecules. These and other oxidizing molecules have effects beyond interactions with atmospheric components, which will be considered later in the context of the Viking space mission.

Another feature of the atmosphere is dust storms of size 100 km to 1000 km which are found to be relatively common. Occasionally these can grow to planet-wide proportions, enveloping the entire planet in a shroud of dust. Dust particles raised by Martian winds are typically 1 μm across and can remain aloft for weeks before settling to the surface. The redistribution of dust by these storms plays an important role in the centimetre- and metre-scale geology of the surface.

It now seems likely therefore, that the most extreme environments on Earth in which organisms can replicate (as illustrated by the extremophiles that you met in Section 2.5) are notably less extreme than the environments that occur on the surface of Mars. A logical conclusion may be that it is very unlikely that any terrestrial organism could grow on the surface of Mars.

■ In view of what you have learned so far about the Martian environment, where might be the only sensible place to look for signs of **microbial** life?

❑ There may be suitable habitats under the ground, just as there are on Earth.

3.3 Viking: the first search for life

By the mid-1970s, with the initial spacecraft surveys of Mars having been completed, NASA decided that the time was right for a major mission to Mars, specifically to address the issue of life. The answer was the Viking project, which consisted of two identical spacecraft, Vikings 1 & 2. Each consisted of an orbiter and lander, with the latter including a suite of experiments designed specifically for studying past or present life on Mars. The prime purpose of the orbiters was to map the surface and to identify suitable landing sites for the landers. Both craft were launched in 1975, with Viking 1 reaching Mars in July 1976 and Viking 2 several weeks later.

Based on the assumptions that Martian life, if it exists, would be carbon-based, its chemical composition would be similar to that of terrestrial life, and it would most likely metabolize simple organic compounds, a life detection instrument package of mass 15.5 kg was carried on the landers. It consisted of three experiments to detect metabolic activity of potential microbial soil communities. They were:

- the Pyrolytic Release experiment (PR) which tested for carbon fixation,
- the Gas Exchange experiment (GEX) which tested for metabolic production of gaseous by-products in the presence of water and nutrients as produced during respiration,
- the Labelled Release experiment (LR) which tested for metabolic activity.

Other instruments such as the gas chromatograph-mass spectrometer (GCMS) and the X-ray fluorescence experiment supported these biology experiments. The former was capable of detecting organic residues in the Martian soil down to parts per billion for some compounds, and the latter could analyse the elemental composition of the Martian surface soil for elements heavier than magnesium (Mg).

Of the three Viking biology experiments, only the PR experiment simulated actual Martian surface conditions and did not use water. In this experiment, a 0.25 cm^3 soil sample was incubated in a simulated atmosphere of CO_2 and CO (carried from Earth) labelled with radioactive ^{14}C. A xenon arc lamp provided simulated sunlight. After 5 days, the atmosphere was removed and the soil sample heated to 625 °C to break down any organic material, and the resulting gases were passed through a ^{14}C detector to see if any organisms had ingested the radioactive CO_2.

The **Gas Exchange experiment** sought to detect alterations in the composition of the gases in the test chamber as a result of biological activity. The procedure involved partially submerging a 1 cm^3 sample of soil in a complex mixture of compounds the investigators called ‘chicken soup’. The soil was then incubated for at least 12 days in a simulated Martian atmosphere of CO_2, with helium and krypton added. Gases that might be emitted from organisms consuming the nutrient were then detected by a gas chromatograph – this instrument could detect CO_2, O_2, CH_4, H_2, and N_2.

The LR experiment moistened a 0.5 cm^3 sample of soil with 1 cm^3 of a nutrient consisting of distilled water and organic compounds. The organic compounds had been labelled with radioactive ^{14}C. After moistening, the sample was allowed to incubate for at least 10 days, and any micro-organisms would hopefully consume the nutrient and give off gases containing the ^{14}C, which would then be detected. An example of such a process (respiration) is shown in Equation 3.7. (Terrestrial organisms would give off CO_2, CO, or methane CH_4.)

$$(^{14}CH_2O)_n + nO_2 \longrightarrow n^{14}CO_2 + nH_2O \qquad (3.7)$$

Ironically, it was the GCMS rather than any of the biology experiments that arguably produced the most important result for the detection of life. It discovered no sign of any organic compound on the surface of Mars. This result came as a complete surprise as organic compounds are known to be present in space (for example, in meteorites). Proof that the GCMS was definitely working came from the fact that it was able to detect traces of the cleaning solvents that had been used to sterilize it prior to launch (see Box 3.7).

The total absence of organic material on the surface made the results of the biology experiments equivocal, since metabolism involving organic compounds were what those experiments were designed to detect. However, the results from the biology experiments were sufficiently confusing to be worth examining.

To reduce the chance of erroneous positive results, the biology experiments not only had to detect life in a soil sample, they had to *fail* to detect it in another soil sample that had been heat-sterilized (the so-called *control* sample). Had terrestrial life been tested with the Viking biology experiments, the results in Table 3.3 would have been expected.

Table 3.3 Results of testing terrestrial life with the Viking biology experiments.

	Response for sample	Response for heat-sterilized control
GEX	oxygen or CO_2 emitted	none
LR	labelled gas emitted	none
PR	carbon detected	none

If life were completely absent from Mars, as the GCMS results suggested, the expected results of the Viking biology experiments would have been as in Table 3.4.

Table 3.4 Results of testing with the Viking biology experiments if life were completely absent from Mars.

	Response for sample	Response for heat-sterilized control
GEX	none	none
LR	none	none
PR	none	none

The actual results from Mars, in a highly simplified form, are given in Table 3.5.

Table 3.5 Actual results of tests on Mars with the Viking biology experiments.

	Response for sample	Response for heat-sterilized control
GEX	oxygen emitted	oxygen emitted
LR	labelled gas emitted	none
PR	carbon detected	carbon detected

The fact that both the GEX and PR experiments produced positive results even with the control sample indicates that non-biological processes were operating. Ensuing laboratory experiments on Earth involving the exposure of materials thought to be similar to Martian soil (oxides or superoxides) to UV radiation in the presence of a Martian-type atmosphere generated peroxides in or absorbed on the soil and, moreover, reproduced the results of the Viking lander experiments. Oxidized iron could act as a catalyst to produce the results seen by the PR experiment. It is likely that the surface of Mars is highly oxidized, the red colour (Figure 3.10) supporting the contention that iron is in this form. The oxidizing atmosphere of the Earth also ensures that iron is usually in similar form at the Earth's surface.

Only the LR experiment appears to have met the criteria for life detection, but it does this rather ambiguously. When the nutrient was first injected, there was a rapid increase in the amount of labelled gas emitted. Subsequent injections of nutrient caused the amount of gas to decrease initially (which is surprising if biological processes were at work) but then to increase slowly. No response was seen in the control sample sterilized at the highest temperature (160 °C). While there is still some controversy, the consensus is that the LR results can also be explained non-biologically.

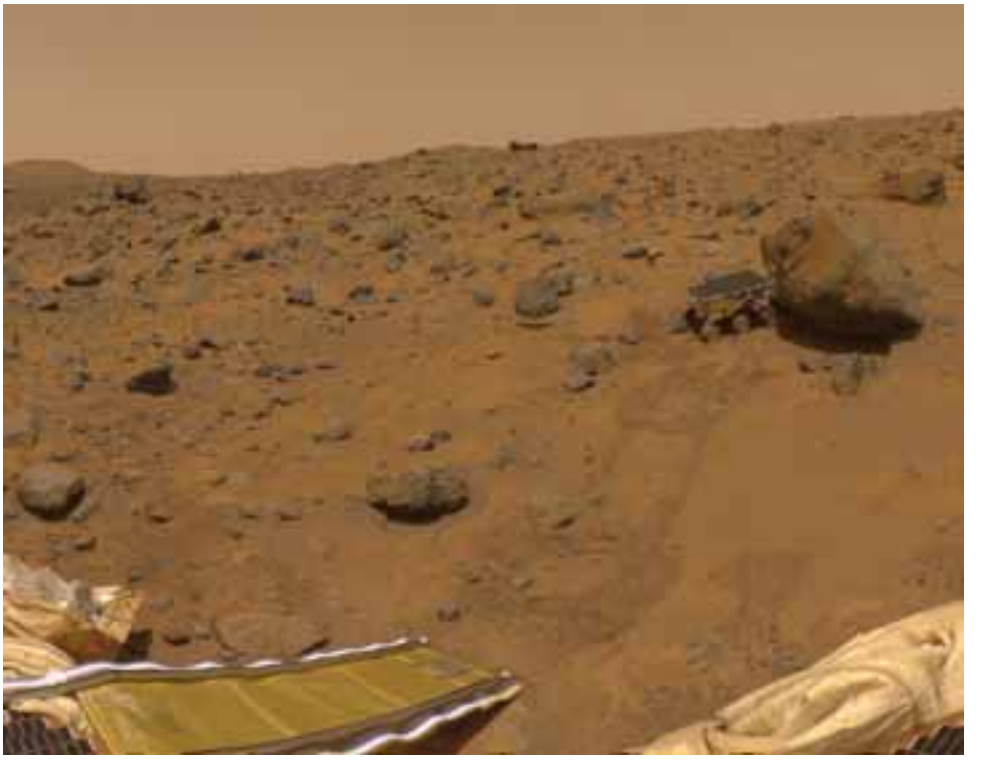


Figure 3.10 A spectacular panorama taken by Mars Pathfinder a few days after landing in July 1997. The camera has a resolution of 2 mm at a distance of 2 m. The feature on the horizon at left is one of the so-called 'Twin Peaks' which are at a distance of some 1 to 2 km. (The other 'twin' is just out of shot.) The Sojourner Rover is seen analysing the rock nicknamed 'Yogi' some 3 to 4 m from the lander. The characteristic colour is suggestive of iron being in oxidized form.

To summarize, all three Viking biology experiments gave results indicative of active chemical processes when samples of Martian soil were subjected to incubation under the conditions imposed on them. However, the experiments failed to detect any organic matter in the Mars soil, either at the surface or from samples collected a few centimetres below the surface. The indications were that strong oxidative processes were at work at the surface. Subsequent theoretical work has shown that **photochemical** processes, as well as the effects of oxidants such as hydrogen peroxide are likely to be responsible for the destruction of all such material in the surface region.

The enigma of an active organic chemistry in the absence of life has not been fully explained and, as a result of the Viking biology experiments, the scientific community was split into those who denied (or at least strongly questioned) the existence of life on Mars, and those who did not rule out the existence of certain biotic 'oases' on Mars. However, it is clear that views of the issue of life on Mars were dominated for nearly 20 years by the results from the Viking biology experiments.

QUESTION 3.3

(a) The view of most scientists after the Viking biology experiments were performed on the surface of Mars was that they had failed to detect any positive indications of life. What arguments could be used against the notion that these experiments had ruled out all possibilities of life?

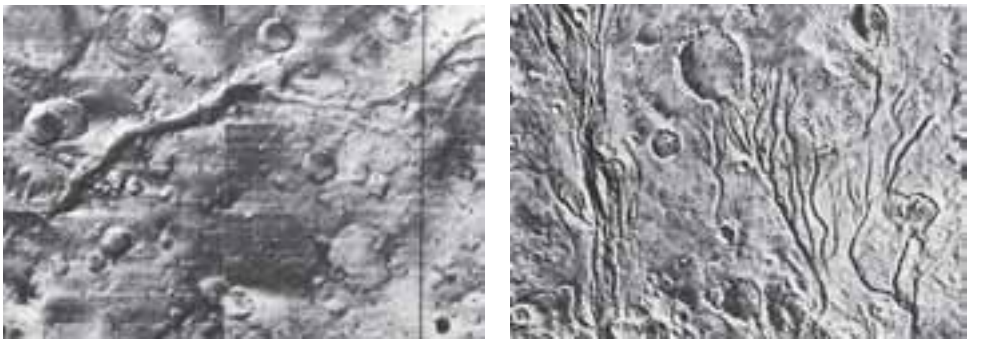
(b) These experiments employed a control sample against which results from Martian soil samples were compared. If this control sample had not been used, how might the results from the Martian samples have been interpreted?

3.4 Water, water everywhere?

Images of the surface of Mars, captured by the early space missions (Figure 3.11), showed features that resemble canyons and valleys, whereas others look distinctly like channels on Earth, indications that water once flowed as a liquid on Mars. The cratering timescale of the regions that contain these features has been used to estimate when these valleys were formed. Although different interpretations have been made of this evidence, it does appear that the valleys and gullies were not formed during a single, early event. Channels were formed during the period of late heavy bombardment, which occurred between 4 Ga and 3.8 Ga ago (see Section 1.6.1) but there is also evidence of water flow later at about 3 Ga ago, and also of very recent flow.

3.4.1 Evidence for past and recent water

Although the first direct evidence for the existence of surface water came as early as 1972 from images from the Mariner 9 spacecraft (see Figure 3.11a), the real progress in our knowledge of the existence of water on Mars has come with the results from two recent spacecraft, namely the Mars Global Surveyor and Mars Odyssey. They have produced an enormous body of evidence which reinforces the previously held idea that Mars once possessed significant bodies of water – but,



(a) (b)

(c)

Figure 3.11 Images of the Martian surface from early space missions, showing a variety of features indicative of flowing water. (a) Part of a 700 km long channel in the heavily cratered southern highland region (20° S, 184° W) discovered by Mariner 9 in 1972. Channels like these provide firm evidence of an episode of erosion by flowing water, very early in Mars's history, before heavy bombardment had ceased entirely. (b) River valleys and impact craters. (c) A Viking Orbiter image showing the giant 'outflow channel' Ares Vallis (20° N, 33° W), a result of catastrophic flooding. The largest impact crater visible is 62 km in diameter. Clearly visible is the presence of streamlined islands with pointed prows upstream and long tapering tails downstream.

further than this, the evidence points to this water having existed quite recently or, more speculatively, that it is even still existing. First we should look at evidence which has come from the Mars Global Surveyor (MGS) spacecraft (see Box 3.5).

BOX 3.5 THE MARS GLOBAL SURVEYOR AND MARS ODYSSEY MISSIONS

These are two of NASA's recent armada of spacecraft sent to Mars. Both are orbiting spacecraft only whose vital statistics are given in Table 3.6 below.

Table 3.6 Brief mission specifications for Mars Global Surveyor and Mars Odyssey.

	Mars Global Surveyor	Mars Odyssey
Launch	7 November 1996	7 April 2001
Arrival	12 September 1997	24 October 2001
Mass (kg)	767	758
Lifetime	The primary science mission ended on 31 January 2001, but was prolonged into an 'extended mission phase'.	The primary science mission ends in August 2004.
Primary scientific instruments	Orbiter Camera Orbiter Laser Altimeter Thermal Emission Spectrometer	Gamma-Ray Spectrometer Thermal Emission Imaging System Radiation Environment Experiment

Mars Global Surveyor became the first successful orbiter around Mars in 20 years when it entered orbit on 12 September 1997. One and a half years was spent, as planned, in trimming its orbit from an eccentric ellipse to a circle, so that its primary mapping mission started in March 1999. It studied Mars from a low-altitude, nearly polar orbit over one complete Martian year (a year on Mars is 687 Earth days or about 2 Earth years). At the time of writing, Mars Global Surveyor has returned more data than all previous Mars missions combined!

Mars Odyssey is targeted primarily to study the composition of the surface of Mars and to detect water and shallow buried ice. It also collects data on the radiation environment to help assess potential risks to any future human exploration and can act as a communications relay for future Mars landers. Its high-gain antenna unfurled on 6 February 2002, and its instruments began mapping Mars at the end of that month. Odyssey's Thermal Emission Imaging Sensor camera is imaging Mars simultaneously at numerous infrared wavelengths (from 8 to 20 μm) with unprecedented resolution (Box 3.6), even down to the size of a football pitch, seeking thermal and mineral 'fingerprints' hinting at 'seeps' (these are dark streaks seen in MGS Mars Orbital Camera images and possibly created by gradual seepage of liquid water), volcanic vents, or underground reservoirs.

Mars Global Surveyor

One of the great improvements offered by the instruments on the Mars Global Surveyor was the **resolution** (see Box 3.6) of its camera, the Mars Orbiter Camera (MOC). In fact, the smallest feature detectable by this instrument on Mars's surface was of length 1.4 m. Over 60 000 images have been produced and of these, over 200 of them show some very interesting features, which throw more light on the question 'Where did the water go?'

BOX 3.6 RESOLUTION ON SPACECRAFT IMAGES

Resolution is an optical term, referring to the most closely-spaced objects that can be separated. Low resolution (or coarse resolution) means that closely spaced objects cannot be distinguished, whereas high-resolution images reveal fine detail. In the case of astronomical images of the sky, resolution is conventionally expressed as fractions of a degree, but resolution on planetary surfaces is more usefully expressed in terms of true distance on the ground.

In a digital image, the detail that can be seen usually depends on the size of the picture elements or *pixels* of which the image is composed. The terms resolution and pixel size are often treated as having the same meaning, although they are not strictly identical from an optical perspective. The highest resolution images of Mars obtained by the MGS MOC have pixels that are 1.4 m across, but some of the surface has been imaged with the lowest (i.e. worst) MOC resolution, namely 230 m.

QUESTION 3.4

A planetary image is produced by an orbiting spacecraft camera employing 512 pixels from top to bottom of the image and 1024 pixels from side to side. The area imaged corresponds to a scale of 4.5 km from top to bottom and 12.7 km from side to side respectively on the surface of the planet.

(a) What is the resolution (expressed as pixel size) of the imaging system in the configuration described?

(b) In this configuration, would it be possible to distinguish (or resolve) (i) 500 m-scale impact craters and (ii) 1 m-scale boulders?

(c) What happens (qualitatively) to the resolution if the same imaging system is used from a higher orbit?

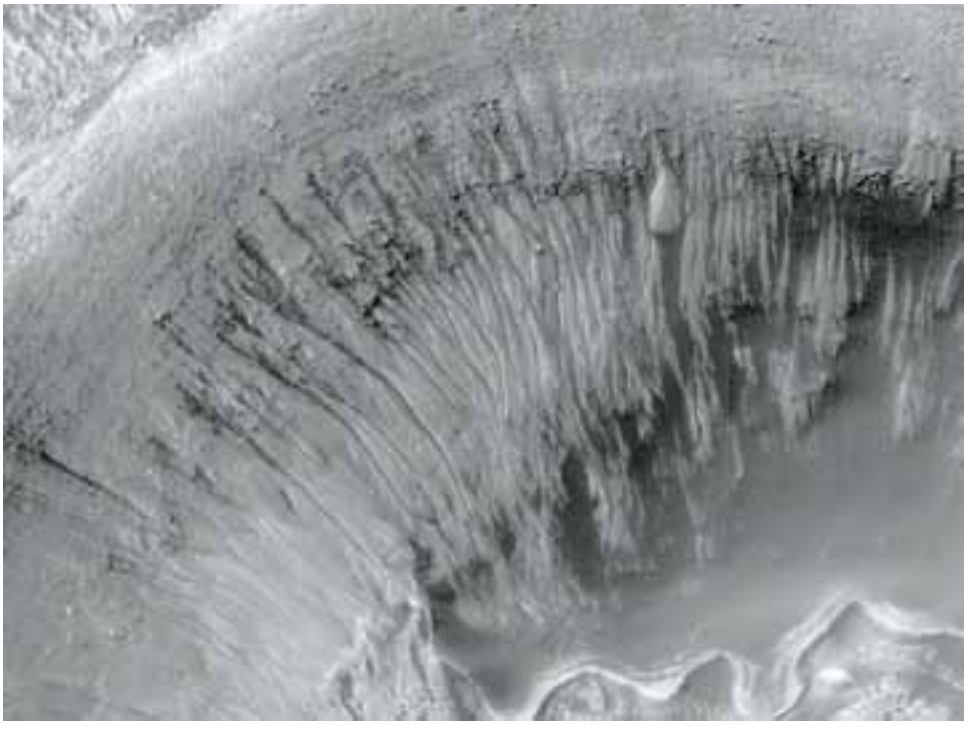
They pinpointed hundreds of delicately structured gully systems (Figure 3.12). Individual gullies are just 10 metres wide (earlier missions couldn't detect such small features because of their inferior resolution) and a whole system might cover an area of only a dozen football pitches. Most are in the southern hemisphere and nearly all occur between latitudes 30° and 70°. Their sculpted terrain, cut-bank patterns, and fan-shaped accumulations of debris look remarkably similar to flash-flood gully washes in deserts on Earth (Figure 3.13). However, the headwalls of the gullies (i.e. starts of gullies) rarely have tributaries and are unlike systems fed by precipitation, so the cause appears not to be the same as flash floods on Earth.

Many (though not all) of the gully systems appear on the shaded sides of hills facing the polar ice caps. Their geometry suggests that 'swimming-pool volumes of water could be entombed underground until suddenly it's warm enough for an ice plug to burst, letting all the water rush down the slopes,' to quote the lead scientist of the MOC team. Such a scenario is shown in Figure 3.14. In this model, underground liquid water is trapped behind an ice barrier or 'plug' which has formed on the shadowed slopes of craters and ravines. Salts dissolved in the water behind the plug could help it to remain liquid as salts have the potential to lower the freezing point of water significantly. (Remember that the phase diagram for water shown in Figure 3.6 is for *pure* water – that for salty water will be different.) Ultimately, when the damn breaks, a flood is sent down the gully, resulting in the observed patterns.

Many of the gully systems look extraordinarily recent (less than 1 Ma old) – sharply carved and crossing older, wind-scoured features. Their appearance is so fresh, in fact, that some planetary geologists think that Mars may have undergone massive, short-term climate changes, where water could come and go in hundreds of years. Indeed, scientists wonder whether liquid water might exist on Mars now, buried in some areas perhaps 500 metres underground.

MOC's findings are corroborated by data from another instrument on the spacecraft, the Mars Orbiter Laser Altimeter (MOLA). For 27 months – longer than a Martian year – MOLA gauged the daily height of the polar icecaps, meticulously recording how much frozen material accumulated in winter and eroded (sublimated or evaporated) in summer in each hemisphere. MOLA showed that each ice cap has a volume as great as the Greenland ice cap on Earth (about 2.5×10^6 km^3).

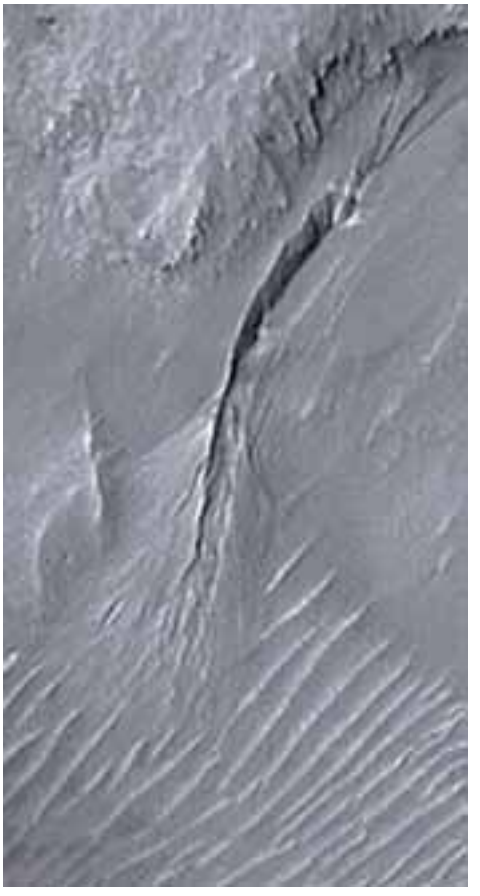
Although the upper crust of the icecaps is clearly carbon dioxide, scientists are now convinced that much of both caps' supporting mass must be frozen water because structurally, dry ice (i.e. frozen CO_2) can't support the mass of a 3 km high polar cap. MOLA and MOC measured how the polar caps shrink in each hemisphere's summer. They shrink so much, in fact, that if the observed trends were continued for just a few centuries, nearly one-third of each polar cap could evaporate into Mars's atmosphere. That would pump the atmospheric pressure up from 6 mbar to 30 or 40 mbar (remember the Earth's atmospheric pressure is about 1000 mb) which is high enough pressure for liquid water to be stable on the planet's surface under certain temperature conditions. Thus, perhaps as recently as just a century or two ago, Mars might have been clement enough for ponds of water to have dotted its surface like desert oases, and current trends suggest it might become so again. All these observations reopen the venerable question: was there – or is there – life on Mars?



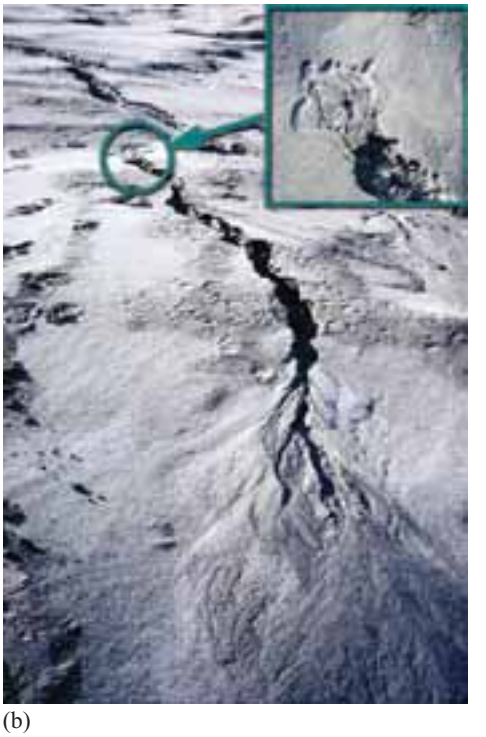
(a)

(b)

Figure 3.12 Two examples of Martian gullies observed by the Mars Global Surveyor. (a) Gullies in the northern wall of the Newton crater in the northern hemisphere. The width of this image corresponds to a real distance of 6.5 km on Mars. (b) Gullies at 70° S in polar pit walls.



(a)



(b)

Figure 3.13 (a) The accumulated debris (or 'apron') from this gully on Mars covers sand dunes that may have formed less than a century ago. The width of this image corresponds to a real distance of 1.6 km on Mars. (b) For comparison, an apron on Earth is shown. In this example, rain water flowing under and seeping along the base of a recently-deposited volcanic ash layer (at Mount St Helens) has created the gully.

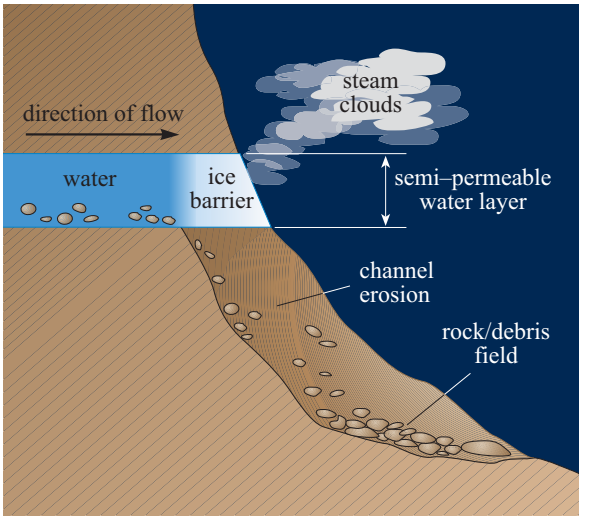


Figure 3.14 A possible model for the formation of the characteristic channels and aprons of Martian gullies.

Mars Odyssey

This mission started its main scientific tasks at the end of February 2002 and in a very short time began to produce outstanding results. We shall focus here on some very important measurements that are of fundamental importance. These are based on observations made with the Gamma-Ray Spectrometer. First we shall look at this instrument in more detail. The Mars Odyssey Gamma-Ray Spectrometer (GRS) is a suite of three different sensors that share a common electronics box and complimentary scientific objectives. The instruments are the GRS proper, the Neutron Spectrometer (NS) and the High-Energy Neutron Detector (HEND). This instrument is a follow-on instrument to the GRS that was lost when the Mars Observer Mission (Table 3.1) failed in 1993.

How GRS works

When exposed to cosmic rays (charged particles in space that come from the stars including our Sun), chemical elements in soils and rocks emit uniquely identifiable signatures in the form of gamma-rays and neutrons. The gamma-ray spectrometer analyses these signatures coming from the elements present in the Martian soil. By making these measurements, it is possible to determine which elements are present, how abundant they are and how they are distributed around the planet's surface.

The incoming cosmic rays collide with the nuclei of some of the atoms in the soil and, in some cases, they release neutrons as a result. These neutrons have high energies (and are referred to as 'fast' neutrons), and they scatter and collide with other atoms, some of which are excited to a higher energy state than usual. This extra energy can then be released in the form of gamma-rays so that the atom can return to its normal unexcited energy state. The energy E of the gamma-ray (γ) is characteristic of the atom from which it was released – in other words, it is characteristic of its parent element. Some elements such as potassium, thorium and uranium are naturally radioactive so that they don't require an external source, such as cosmic rays, to excite them.

Now the neutrons generated in the initial interactions can undergo further reactions themselves. When they collide with the nuclei of other atoms, they might lose energy, slow down, and eventually become thermalized, which means that they are moving at speeds comparable to the speed at which atoms on the surface are moving. This process is known as **moderation**; hydrogen atoms are especially important in moderating neutrons because the two have nearly identical masses. The various processes discussed here are illustrated in Figure 3.15.

'Thermal', 'epithermal' and 'fast' are terms used to describe neutrons of progressively higher energy.

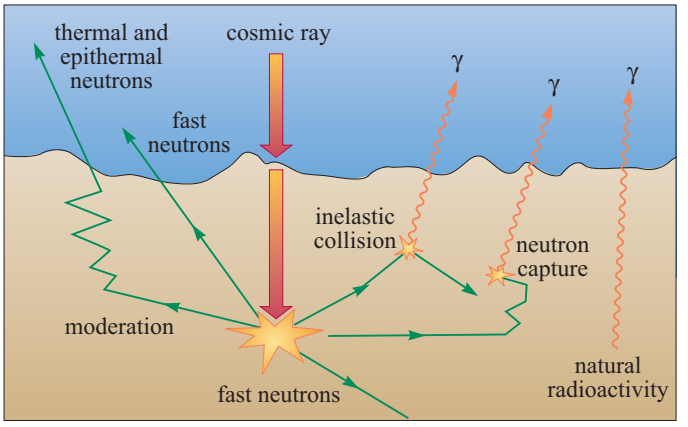


Figure 3.15 Nuclear radiation from a planetary surface produced by the interaction of incident cosmic rays with the surface. The products of the interactions described in the text are neutrons of varying energy (thermal, epithermal and fast in increasing energy) and gamma-rays (γ).

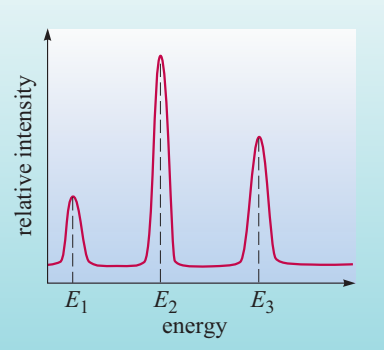


Figure 3.16 A hypothetical γ-ray spectrum in which the number of γ-rays detected is plotted against energy (energy units in MeV rather than units of wavelength or frequency are used by convention when considering γ-rays).

The γ-rays show up as sharp emission lines, as they are termed, in the spectrum recorded by the instrument. The energy of these emissions in the spectrum indicates which elements are present, while the intensity of the particular spectral line, as it is called, reveals the element's concentration. This is illustrated in Figure 3.16. Here, a γ-ray spectrometer has analysed each incoming γ-ray to determine its energy. After a sufficient number have been analysed in this way, they can be shown on a diagram where the number of γ-rays detected (or number of 'counts' as they are often referred to) is plotted against energy to produce a spectrum. In this example, three distinct peaks, or 'lines', are clearly seen in the spectrum at three different γ-ray energies, E_1, E_2 and E_3. Usually these will correspond to three different elements. The height or amplitude of each line, which is related to the number of γ-rays emitted and detected, is proportional to the abundance of that particular element.

■ How do you think that the distribution of elements across a planetary surface can be measured?

❑ By placing the detector on an orbiting spacecraft and arranging for it to look downwards to the planetary surface, the distribution of elements on the surface can be mapped.

As soon as the GRS started to take observations, it began to produce startling results implying the existence of significant quantities of hydrogen. In fact, the NS and HEND instruments also gave indications of concentrations of hydrogen below the surface.

■ What do you think is the likely condition of hydrogen?

❑ Considering the oxidizing nature of the Martian surface, the most probable condition is in H_2O, water.

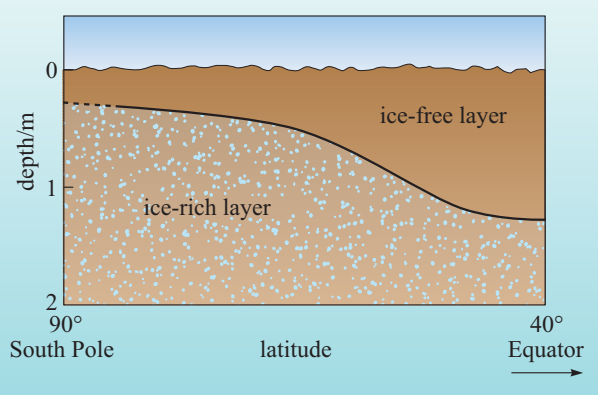


Figure 3.17 A cross-section of the near-surface regions of the Martian southern hemisphere as implied by the GRS data. The ice-rich layer is approximately 0.6 m below the surface at a latitude of 60° S and extends to within 0.3 m of the surface at 75° S.

The data from any one of the three instruments are quite difficult to interpret uniquely on their own, but when all three are put together, it becomes possible to come up with a model which explains all the results. The best explanation of the data is that there exists a topmost layer that is hydrogen-poor which overlies a layer, which is hydrogen-rich. The data suggest that the thickness of the upper layer decreases with decreasing distance to the pole. The hydrogen-rich regions appear to correlate with regions of predicted ice stability. This latter fact in particular strongly supported the contention that the host of the hydrogen in the subsurface layer is ice. A schematic of the possible distribution is shown in Figure 3.17.

Results from GRS data also seem to show that at low to middle latitudes, large areas of Mars contain slightly enhanced quantities of hydrogen. The interpretation of these data is still underway but it seems that in this case, the relatively small amount of hydrogen does not correspond to water-ice but is more likely to be in the form of H_2O or OH chemically bound to minerals in the soil.

Perhaps with these latest results from the Mars Odyssey GRS, we are at least partially answering the question 'Where did all the water go?' But there is still a great deal to do. These results, for example, have only been able to comment on the situation in the topmost metre or so. We need to know what is happening at greater depths.

3.4.2 Evolution of the Martian atmosphere: the role of water

Now that we have established the very strong evidence for past, recent (and maybe even present) water on Mars, the remaining part of this section will be devoted to considerations of the evolution of the Martian atmosphere and the role of water.

One result of the terrestrial planets having formed from similar materials and having experienced similar interior processes is that we would expect the atmospheres of Venus, Earth and Mars to consist of similar gases. There are indeed similarities but the proportions and amounts are very different. In fact, differences in composition and abundance appear to outweigh the similarities. Let us consider as an example CO_2 in the atmospheres of Earth and Venus. The proportions by volume are 0.03% and 96% respectively, clearly widely divergent. However, the differences between the two are much less significant if consideration is also given to the material contained in the interior. On Earth for example, CO_2 is removed very efficiently from the atmosphere by the oceans, from which it precipitates onto the seafloor and ends up as limestone. If we include CO_2 contained in terrestrial deposits of limestone (the white cliffs of Dover are an example), then the total amounts on Venus and Earth are similar. The same is true for nitrogen. Table 3.7 shows the relative quantities of **volatiles** for Venus, Earth and Mars, expressed as a fraction of the planet's total mass.

Table 3.7 Abundances of volatiles on Venus, Earth and Mars expressed as a fraction of each planet's total mass and including CO_2 trapped in limestone and other carbonates.

	CO_2	H_2O	N_2
Venus	9.6×10^{-5}	$>2 \times 10^{-5}$	2×10^{-6}
Earth	16×10^{-5}	2.8×10^{-4}	2.4×10^{-6}
Mars	$>3.5 \times 10^{-8}$	$>5 \times 10^{-6}$	4×10^{-8}

■ In general terms, how do the relative volatile abundances of Venus and Earth compare? And how do these compare with those of Mars?

❑ The relative volatile abundances for Earth and Venus are (approximately) similar. Those of Mars are considerably lower.

■ Which factors could be responsible for the lower volatile budget of Mars?

❑ If, as is supposed, the atmosphere and surface volatiles derived from outgassing, it could be that Mars originally possessed a smaller amount of volatile material, or that it has not outgassed to the same extent as the other planets.

The temperature gradient in the solar nebula would militate against a low original content, since material condensing at the distance of Mars, being further from the protoSun than either Venus or Earth, and thus at a lower temperature, might be expected to be richer in volatiles. Therefore less complete outgassing seems more likely. If this is the case, the Martian atmosphere should never have been as extensive as that of Venus or Earth.

There is, however, evidence of former periods when a more substantial atmosphere existed than at present. As we have seen, the atmospheric pressure is currently so low that liquid water does not exist stably, the transition from solid to gas occurring directly. So a higher atmospheric pressure is probably needed to sustain the water which has created some of the observed water-formed features.

The most plausible explanation is that the surface temperature of Mars has been higher at some epoch in the past than it is at present. A higher temperature would result in the evaporation of some of the condensed volatiles, for example H_2O and especially CO_2. Mars would then have possessed a more substantial atmosphere. Subsequent loss of some of this atmosphere would have led to a cooling of the planet, because the surface temperature is strongly determined by the bulk of the atmosphere. Loss of water or other atmospheric components would have occurred mostly by thermal escape (but also by various other processes). This is a sufficiently important process that it is worth digressing briefly to consider it.

Ultimately, it is the strength of the gravitational field at its surface that dictates whether or not an object in the Solar System can retain an atmosphere: the stronger the field is, the stronger the gravitational forces acting on the molecules in the atmosphere. This leads to the notion of **escape velocity**, the smallest upward speed that any object (spacecraft or molecule) must have to escape from a body. The escape speed v_{esc} for a body of mass M and radius R is given by:

$$v_{esc} = \sqrt{\frac{2GM}{R}} \tag{3.8}$$

where G is the gravitational constant. Whether atmospheric molecules have sufficient speed depends on the temperature. As the temperature of a gas increases, its molecules move around more quickly, and the average speed of its molecules increases. Some fraction of the molecules will always be travelling fast enough to overcome gravitational forces, allowing them to escape to space. At low temperatures, this proportion is negligible, but at higher temperatures it becomes progressively more signicant, until most molecules exceed the escape speed for the planetary body. Note that the relevant temperature is that at a level in the upper atmosphere above which the atmosphere is so thin that a molecule moving outwards has little chance of colliding with another, and so *will* escape if it has sufficient speed.

'Planetary body' is a handy term that can be used to encompass planets, satellites and asteroids.

Different gases have different molecular masses, so their average speeds are different at a given temperature.

In order for a planetary body to retain a particular gas in its atmosphere for a period of time of the same order as the age of the Solar System, the average speed of the molecules in the gas should be less than about one-sixth of the escape speed. If the average speed exceeds one-sixth of the escape speed, a signicant proportion of molecules will be moving faster, and will be lost.

This condition is achieved on only a few planets and satellites. Mercury is so close to the Sun and so hot that average molecular speeds for all common gases are too great. Titan, which is a similar size to Mercury, is much less dense ($\approx 1.9 \times 10^3\,\mathrm{kg\,m^{-3}}$). This in turn means that its mass is less than one-half of Mercury's, and therefore its surface gravity and escape speed are lower. Titan, however, is also so far from the Sun that its temperature is only about 100 K. Thus, Titan can retain a dense atmosphere.

QUESTION 3.5

(a) The average (or, more correctly, root mean square) speed of a gas molecule is given by

$$v = \sqrt{\frac{3kT}{m}}$$

where m is the molecular mass. Based on this formula, what can you say about the likelihood of different gases being lost from a planetary atmosphere by thermal escape?

(b) In the Martian atmosphere, calculate the relative average speeds of the two most common constituents of the atmosphere?

Returning now to Mars, can the remaining atmosphere tell us anything about previous episodes of loss? Well, the answer is yes. It should carry the signature of this loss in the enrichment of heavier isotopes – lighter isotopes having been preferentially lost. Of particular interest in this context is the isotopic analysis of volatiles from Mars as measured in certain meteorites for which there is compelling evidence of Martian origin. In the following section, we shall examine this isotopic record of past climatic conditions on Mars. But first, we need to consider the evidence that these meteorites originate from Mars.

3.4.3 Meteorites from Mars

EET A79001 is a meteorite collected from Antarctica. Lets first consider the lettering and numbering system used in its designation. 'EET' refers to the collection site (which, in this instance, was Elephant Moraine in the Antarctic), 'A' designates the collection trip and '79' is the year of collection (1979). The identifying number, 001, signifies that it was the first meteorite to be classied upon return of the samples to the curatorial facility. When the collectors, who are often meteorite researchers, spot rare or otherwise unusual samples in Antarctica, a note is made to give them priority treatment during the preliminary classication procedure. EET A79001 was one such sample, and on this particular occasion the decision to promote its investigation could not have been more justied, since the meteorite is now widely believed to come from Mars.

EET A79001 is a so-called shergottite. A photograph of EET A79001 is shown in Figure 3.18. Shergottites are linked with two other categories of meteorites known as nakhlites (pronounced '*nahk*-lights') and chassignites (pronounced '*sha*-sig-nights'). These samples, around 30 in number, are collectively referred to as the SNC meteorites (where

Figure 3.18 The meteorite EET A79001.

S, N and C denote shergottites, nakhlites and chassignites). The names derive from the discovery locations, namely Shergotty (India), Nakhla (Egypt) and Chassigny (France). All of the SNC meteorites are igneous rocks formed by crystallization from magma. In appearance, the shergottites are medium-grained rocks of basaltic composition. However, one important fact distinguishes most SNC meteorites – they have relatively young formation ages, in the case of EET A79001 about 0.2 Ga ago.

- The crystallization ages of most of the SNC meteorites range from 0.2 Ga to 1.3 Ga. What are the implications of this?

❑ These meteorites were formed late in the history of the Solar System. Wherever they were formed, at least some part of their parent body had to have been melted 0.2 Ga ago (or remained molten until this time).

The ages of most meteorites cluster around 4.5 Ga and represent samples that formed in parent bodies of asteroidal size (i.e. a few hundred kilometres in diameter). Asteroids were heated and cooled relatively early in the history of the Solar System. Some meteorites have younger ages than 4.5 Ga, but these are the result of later impact melting which acts to re-set the radiometric dating systems. It is apparent from the textures of SNC meteorites that they are not impact-produced melts. The only reasonable environment that retains sufficient heat to produce melting 0.2 Ga ago is a parent body of planetary dimensions, so there can only be a few candidates for the source of SNC meteorites – i.e. Mercury, Venus, Earth, the Moon, Mars, or Io.

- Can you think of ways in which the SNC meteorites could be removed from the surface of the planet?

❑ Ejection caused by a volcanic eruption or an impact.

Volcanic processes can be rejected on various grounds. The most plausible mechanism involves removal following an impact onto a planetary-sized surface by a meteoroid or comet. A schematic representation of calculations pertaining to large impact craters is shown in Figure 3.19. In this model, the incoming projectile, in addition to pulverizing the target rocks and producing a crater, also causes a very thin layer of ejecta to be propelled away from the impact site. Some of these pieces of ejecta, which are unmelted and unshocked, can escape from the planet's surface. Other escaping ejecta components could be melted or shocked.

- If fragments ejected from another body can reach the Earth, what is the most likely source and why?

❑ The Moon – simply because it is so much closer than any other solid body.

Meteorites of lunar origin have now been unambiguously found on the Earth, numbering around 40 in total.

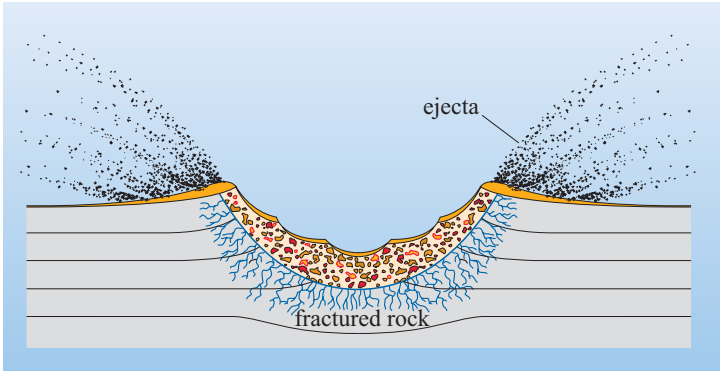


Figure 3.19 Schematic diagram of a crater-forming event on a planetary-sized body. No physical parameters are given since this is a generalized case (although imagine that it pertains to projectiles of kilometre size travelling at around 10 km s^{-1}). Note that the impact produces melting and intense fragmentation in the target rocks, with fragment size increasing away from the impact site. However, in addition, unshocked and unmelted materials can be ejected from a very thin surface layer. Close to the impact site, these may be travelling sufficiently fast to overcome the escape velocity of the planet and are, as a consequence, ejected from the planet.

■ Why can we be so certain that these fragments originate from the Moon?

❑ The Moon is the only body for which there are returned samples with which to compare. It transpired that chemical compositions, isotope ratios, minerals, and textures of the lunar meteorites are all similar to those of samples collected on the Moon during the Apollo missions. Taken together, these various characteristics are different from those of any other type of meteorite or terrestrial rock.

Having established unambiguously that certain meteorites found on Earth have originated from another body in the Solar System, we can now turn again to EET A79001. Its characteristic features bore little resemblance to any returned lunar samples so the Moon was unlikely to be the origin. In addition, lunar volcanism ceased about 3.2 Ga ago. Io's distance from Earth and its proximity to Jupiter make it dynamically very unlikely that ejecta from its surface reaches the Earth. Mercury, despite being relatively close to the Earth, is so close to the Sun that it is unlikely that material ejected from its surface could be transported to the orbit of Earth. Venus has a dense atmosphere (Table A1) and a large gravitational field (about the same as Earth) – removal of material from this planet would require speeds so high that frictional heating in the atmosphere would cause severe melting (or even complete vaporization in the case of small pieces of rock).

■ Based on the above arguments, what do you think is the likely parent body of the SNC meteorites?

❑ Mars would seem to be a good possibility, if only by default.

In support of this it should be noted that Mars has a tenuous atmosphere (surface pressure currently 6 mb) and the gravitational field is comparatively low (less than one-half that of the Earth; see Question 3.1) – thus, solid materials could be ejected without vaporization – and it is a lot closer than Io.

However, there is some very specific evidence that quite unambiguously points to Mars as the parent body of EET A79001. We have already stated that SNC meteorites are not impact melts but are the result of igneous activity, analogous in many ways to rocks formed at or near the Earth's surface. However, SNC meteorites were subjected to an impact event, which was violent enough to accelerate them to a speed greater than the escape velocity of Mars (5 km s^{-1}).

Intuitively therefore, it may be expected that the meteorites would show some evidence of this process. Indeed, shergottites and chassignites record the effects of shock, though the nakhlites are unshocked (which poses a constraint on theoretical modelling of the impact event).

During the impact event that is thought to have removed EET A79001, localized melting occurred within the sample. These melts cooled extremely rapidly to form a glass containing trapped atmospheric gases. Analyses of these trapped gases showed them to be chemically and isotopically distinctive. The abundances of the gases contained within the shock-produced glass from EET A79001 are plotted in Figure 3.20 against the abundances of gases in the Martian atmosphere as determined by the Viking and Pathfinder missions (these are the same as the abundances shown in Table 3.2 but expressed as the number of particles per m^3 rather than as the volume ratio). The line represents points whose values are the same on both axes.

■ What can you conclude from Figure 3.20 about the abundances of gases in EET A79001 and the Martian atmosphere?

❑ The data points show near perfect correlation between the two sets of abundances, suggesting that EET A79001 contains trapped Martian atmosphere.

The evidence is compelling and the conclusion of this evidence is therefore that EET A79001 and, in fact, all the SNC meteorites originate from Mars.

QUESTION 3.6

(a) Place in ascending order of escape velocity Mars, Venus and the Moon.

(b) What factors other than escape velocity dictate the likelihood of material from these bodies being ejected by surface impact and reaching the Earth?

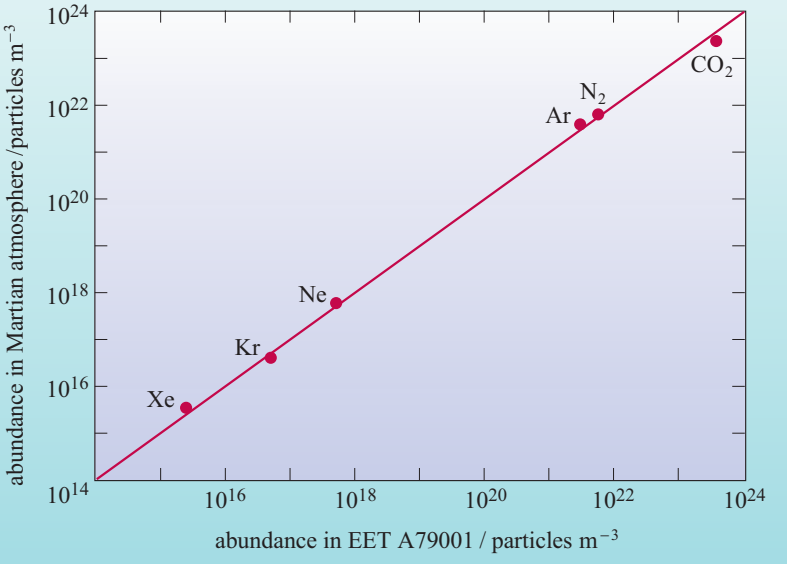


Figure 3.20 Plot of the abundances of gases in the Martian atmosphere (measured by spacecraft landers) versus the abundances in the glass from EET A79001. Both axes are logarithmic. Note the extremely good correlation, which forms part of the evidence supporting a Martian origin for SNC meteorites.

The Martian meteorite EET A79001 can give us very strong clues as to the history and evolution of water on Mars. The shock-produced glass in EET A79001 contains a variety of gases. Through analyses of the noble gases, a theoretical model has been devised to describe the early evolution of the Martian atmosphere. Before considering this further it should be noted that the isotope composition of xenon in the EET A79001 glass shows that the Martian atmosphere includes gases attributable to carbonaceous chondrites, a primitive class of meteorite (this can also be inferred from an assessment of Viking data, or by looking in detail at the chemical compositions of the SNC meteorites). It is considered that Mars received its presently observed complement of volatiles from an inux of carbonaceous chondrite-like material near the end of its formation. Outgassing from this veneer of chondritic material subsequently formed the early Martian atmosphere.

You have already seen that observations of the Martian surface show considerable evidence for the action of fluid flow (see Figure 3.11). At least two episodes of water flow and subsequent loss from the atmosphere to space have been identied. At the present time, photodissociation of water vapour results in the loss of hydrogen and oxygen. This mechanism has been operating on Mars since its formation. In earlier times, because the ultraviolet flux from the Sun was comparatively high, resulting in enhanced levels of photodissociation, the process was more efficient. Even so, the rate is too slow to account for all the water assumed to be lost. Quite clearly another process was responsible.

Over the first 100 Ma of Martian evolution, water was readily converted to hydrogen by reaction with, for instance, iron–nickel metal. The vast amounts of hydrogen produced in this way were very quickly lost from the planet by thermal escape as a rapidly moving flow of gas. During this time, gases heavier than hydrogen were swept away by the rapid flow, and so were also lost from the planet – a process known as **hydrodynamic escape**. The details of this are complicated but its imprint can now be observed in the abundances of noble gases and the isotope composition of xenon in the glass of EET A79001.

■ With the knowledge that hydrogen has been lost from the Martian atmosphere, how would you anticipate its isotope composition has evolved with time, given that hydrogen has two isotopes, namely the lighter isotope of hydrogen (^{1}H, hydrogen) and the heavier isotope (^{2}H, deuterium)?

❑ It would be expected that ^{1}H, would be preferentially lost from the atmosphere compared with the heavier isotope ^{2}H, deuterium, resulting in an increase in the D/H ratio with time.

Measurements of water vapour in the Martian atmosphere show a D/H ratio which is about 5 times that of water in the Earth's oceans. This measurement spectacularly supports the contention of the loss of hydrogen to space, and implies that water vapour may have been carried with it by hydrodynamic escape. So our model of a Martian surface and atmosphere which at some periods was much warmer and which showed extensive surface water deposits is strongly supported.

3.4.4 Conclusions

In summary, the evidence of atmospheric evolution on Mars points to a planet in which outgassing is less complete than on Venus and Earth, and in which the atmosphere has been lost partly to space and then to the surface as the temperature consequently fell.

You have seen that the conclusions of the initial exploration of Mars by spacecraft were not particularly encouraging for those who wished to see Mars as a habitat for past or present life forms. And these conclusions were also supported by the results of the Viking biology experiments, despite their ambiguity.

But subsequent observations by the Mars Global Surveyor and the Mars Odyssey have provided evidence of extensive bodies of frozen water below the surface as well as evidence of flowing water in the recent past – and, more controversially, the interpretation that mechanisms which produce this flowing water may be active even today. Additionally, there is separate, strong evidence from Martian meteorites for past periods of extensive water followed by escape to space. This evidence is in the form of the isotopic signature of hydrogen. So the pessimistic conclusions of most scientists following the results from Viking have not been completely vindicated. Now there is a significant body of opinion which believes that conditions in the past were conducive for simple life forms to have existed on Mars, and even more controversially, that they may have survived in certain protected 'oases'. These arguments form the foundations for the space experiments planned and in preparation over the next decade to search for evidence of past or present life.

Before life could originate on Mars (or anywhere) certain prerequisites were needed.

- What were the prerequisites?
- Water, organic materials, an energy source and a site in which to concentrate them.

Let's consider the hypothesized earliest period of extensive water flow on the surface of Mars which extended to around 3.8 Ga ago (around the same time as the end of the period of heavy asteroid bombardment). From our knowledge of the evolution of the Martian atmosphere, it seems likely that at this time the above prerequisites would have been available on Mars as on Earth. We've seen that on Earth, life would have originated between the end of the period of heavy bombardment, 4.0 Ga to 3.8 Ga ago and before 3.85 Ga to 3.5 Ga ago. This gives a 'window' of between 100 Ma and 500 Ma when life would have evolved. On Mars, this opportunity would have been somewhat shorter as the surface water was beginning to vanish rather rapidly by 3.8 Ga ago. In fact, the major uncertainty in this scenario appears to be whether liquid water was available for long enough and abundantly enough for life to arise.

3.5 The ALH 84001 story: evidence of life in a Martian meteorite?

You now know almost unequivocally that meteorites of Martian origin exist on Earth. We shall now proceed to consider the most famous of all Martian meteorites and its significance for the question of life on Mars. Until 7 August 1996, the name ALH 84001 was probably known to only a relatively small number of planetary scientists worldwide.

■ What is the significance of the designation ALH 84001?

❑ '84' signifies the year of recovery of the meteorite, namely 1984, and '001' signifies that this was the first meteorite to be classified upon return of the samples to the curatorial facility. ALH indicates the locality of recovery, which in this case was Allan Hills in Antarctica.

But on that day, everything changed. A press conference was held by a group of scientists led by David McKay, Everett Gibson and Kathy Thomas-Keptra of NASA's Johnson Space Centre to announce that certain characteristics in this meteorite, known to have come from Mars, were most likely interpreted as being the relics of ancient Martian microbial life. The effect was dramatic. Newspapers, radio and TV around the world rushed to declare 'Life found on Mars!' (or its equivalent in many languages) and with their own particular stress. It seemed that humanity's desire to find that we are not alone (or at least might not have been alone at some time in the past) had at last been answered. But had it? Well not everyone in the scientific community thinks so. In fact, it is true to say that the majority of informed opinion does not agree with the conclusions of the authors of the scientific paper that presented the full analysis of their case.

Why did a 1.9 kg potato-sized lump of rock cause such a sensation? The *recent* history of ALH 84001 started when it was discovered in 1984 in the Allan Hills region of Antarctica.

■ Why are so many of the world's meteorite collection found in Antarctica?

❑ It's a combination of the environmental conditions (e.g. uniform surface, etc) and the lack of human activity that makes Antarctica a good location for meteorite detection and collection. Of the world's collection of more than 22 000 meteorites, by far the majority have come from this source.

The Martian origin of ALH 84001 was not recognized until 1993, when it was realized that the SNC meteorites (now numbering around 30), were almost certainly from Mars. But what of the history of ALH 84001 before it was found in Antarctica in 1984? Here, we shall quote from a 1997 article by several of the team who undertook the original work:

> 'The meteorite timeline begins with the crystallization of the rock on the surface of Mars, during the first 1 percent of the planet's history. Less than a billion years later the rock was shocked and fractured by meteoritic collisions. Some time after these impacts, a water-rich fluid flowed through the fractures, and tiny globules of carbonate minerals formed in them. At the same time, molecular by-products, such as hydrocarbons, of the decay of living organisms were deposited in or near the globules by that fluid. Impacts on the surface of Mars continued to shock the rock, fracturing the globules, before a powerful collision ejected the rocks into space. After falling to Earth, the meteorite lay in the Antarctic for millennia before it was found and its momentous history revealed.'
>
> Gibson, E. K. et al. (December 1997) 'The Case for Relic Life on Mars', *Scientific American*, pp. 58–65.

■ How old is the ALH 84001 meteorite?

❑ According to the quotation above, ALH 84001 formed 'during the first 1 percent of the planet's history'. Since Mars, like all the planets, formed around 4.5 Ga ago, ALH 84001 formed around 4.5×10^7 years later. This means that it has an age of nearly 4.5 Ga. This is in contrast to most SNC meteorites which have an age in the range 0.2 Ga to 1.3 Ga.

But how did they reveal this 'momentous history', namely that ALH 84001 contained evidence for the existence of living organisms? Their conclusion was based on five separate strands of evidence, each of which on its own was not compelling, but, taken together, according to the authors, were highly convincing. Almost all of their evidence comes from carbonate globules that are found on the surface of a fracture through which fluid flowed and deposited these globules. (See Figure 3.21.)

The five lines of evidence, as presented by David McKay and his colleagues can be summarized as follows:

(a) The carbonate is in the form of 'globules' comparable to crystal aggregates known to be produced by bacteria on Earth.

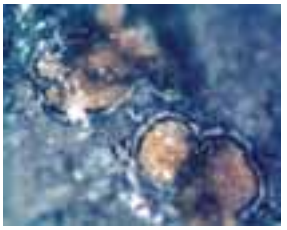
(b) Perhaps the most visually compelling evidence are objects that seem to be the fossilized remains of microbes themselves. Nanometre-scale carbonate structures, shown in Figure 3.22a in the globules resemble fossil spheroidal, rod-shaped and filamentary bacteria. The segmented object shown in Figure 3.22a is 380 nm long. This can be compared with terrestrial samples such as that shown in Figure 3.22b. The approximately vertical feature just to the right of centre is believed to be a minute fossil and is also 380 nm long. It was found 400 m below the Earth's surface in a formation known as Columbia River Basalt. Some curved structures in the meteorite have lengths in the range 500 to 700 nm. Others, for example ovoids (meaning egg-shaped), are as small as 30 nm in length. These are about 10 times smaller than terrestrial objects that are usually interpreted as bacteria. However, McKay et al. argue that typical cells often have small appendages attached, of sizes similar to these small features found within ALH 84001. The suggestion is that some of these features are fragments or parts of larger units.

(c) Inside the carbonate globules, they found fine-grained particles of magnetite (Fe_3O_4) and iron sulfide within the size range 10 to 100 nm. Using sophisticated analysis techniques involving microscopy and spectroscopy, they found that the size, purity, shapes and crystal structure of all the magnetites were typical of magnetites produced by bacteria on Earth. Such particles on Earth are known as magnetofossils. The magnetites within ALH 84001 are typically 40 to 60 nm in size. Intriguingly, some of them in ALH 84001 are arranged in chains, similar to pearls in a necklace. Terrestrial bacteria often produce magnetite in just this pattern, because as they biologically process iron and oxygen from water, they produce crystals that align themselves with the Earth's magnetic field.

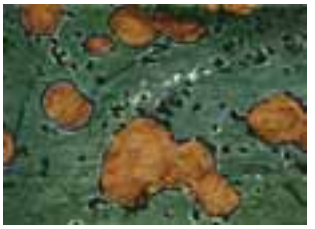
(d) Another strand of evidence is that the carbonate, iron sulfide and iron oxide minerals occur together but would not be stable under any one set of physical conditions suggesting formation in the 'non-equilibrium' conditions characteristic of life.

(a)

(b)



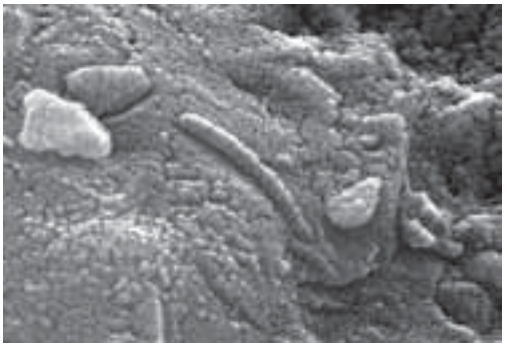
(c)



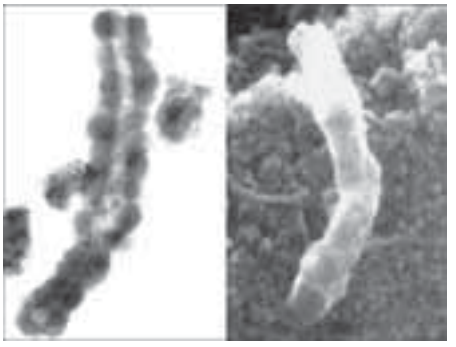
(d)

(e)

Figure 3.21 (a) ALH 84001 shown in its entirety at the top. A cut through the rock, showing a cross section, is shown at middle. Just right of centre is a vertical crack through which fluid flowed and deposited globules of carbonate minerals. A 2 mm long fragment is shown at bottom left, containing several of the globules, which are about 200 μm in size. One of these is shown magnified at bottom right.



(a)



(b)

Figure 3.22 (a) The segmented object (380 nm long) was found in the Martian meteorite ALH 84001 and has been interpreted as a fossil microbe. It is claimed to resemble fossilized bacteria found on Earth. For example, the near-vertical objects in (b), which are of the same length as the object in (a), were found at a depth of 400 m below the terrestrial surface.

(e) There is organic matter including complex hydrocarbons that could have been produced by living organisms. Now organic molecules have been found in other meteorites which are known to have come from asteroidal sources in interplanetary space, an environment hardly likely to support life, so why should this fact be significant. McKay's team argue that the type and relative abundance of the specific organic molecules are suggestive of life processes. When living organisms die and decay in the terrestrial context, they create hydrocarbons associated with coal, peat and petroleum. Many of these belong to a class of organic molecules known as polycyclic aromatic hydrocarbons (PAHs). There are literally thousands of PAHs but their presence does not necessarily demonstrate that biological processes have occurred. In ALH 84001, the PAHs are always found in carbonate-rich regions, including the globules. They contain a relatively small number of different PAH types, all of which have been identified in the decay products of microbes. Significantly, these PAHs were located inside the meteorite where terrestrial contamination is unlikely to have occurred.

All these features occur together in the carbonate veins. The team argue that these features are indigenous to the meteorite and 3.9 Ga old, and therefore that there was life on Mars at that time. There is however much scepticism in the scientific community as to the veracity of this conclusion, and much research on this subject has been instigated since 1996. Amongst the objections to these conclusions are:

Figure 3.23 A barium carbonate crystal aggregate grown in a silica gel, in the absence of any organisms. Compare with the hypothesized microfossil in Figure 3.22a.

(a) It is always difficult to find compelling evidence of a microbial origin for such simple structures which might equally plausibly be chemical and mineralogical artefacts (see Figure 3.23).

(b) Convincing fossils of this age are extremely rare on Earth, and very hard to find even in systematic searches of well-exposed regions of well-preserved rocks. So it is very surprising that with a sample of little more than 20 rocks from Mars, one were to contain fossils.

(c) The original research compared the carbonate microstructures (evidence line (b)) to fossil bacteria but made no attempt to make comparisons with non-biogenic mineralic structures that could be mistaken for fossil bacteria.

(d) A recent study suggests that the carbonates and other materials in the veins crystallized at 200 to 500 °C from minerals melted at the moment of impact that ejected ALH 84001 into space. Experimentally produced melts of this type generate carbonate globules like those in the meteorite. If the fractures and the carbonates that fill them formed during a high-velocity impact on Mars, then the structures and compounds within them are extremely unlikely to be fossils.

(e) There are also tiny crystals of the iron sulfide mineral greigite in the carbonate globules. These are interpreted by McKay and his team as products of bacteria. On Earth, there are bacteria that gain their energy for life by converting dissolved sulfates to sulfides. The sulfide then reacts with any iron present to produce iron sulfide minerals. This is a very common process in oceanic sediments. However, there are also natural chemical processes that produce iron sulfides, so they need not be biogenic.

In fact, at the time of writing, it seems that of the five lines of evidence presented above, four can be explained without the necessity for a biogenic origin. However, one, namely (c), remains controversial. This concerns the origin of the magnetite crystals, present in abundance within the carbonate globules. In fact, within the space of a few months, scientific results on these features were published by two different groups of scientists with diametrically opposing conclusions. Both groups have studied primarily the shapes and other features of the magnetite crystals. One group claims that the Martian magnetites are physically and chemically identical to magnetites produced by a certain terrestrial bacteria strain. Conversely, the second group, using new microscopic measurements, state that 'the crystallographic and morphological evidence is inadequate to support the inference of former life on Mars'.

The original team has put forward arguments to counter the objections described above. However, the present opinion of the scientific community is that while the evidence for life is intriguing and demands further study, it is not compelling. It is probable that the argument will only be finally resolved by results from future missions to Mars. One of the most significant results from ALH 84001 was the impetus it gave to studies of life on Mars.

3.6 Planetary protection

We shall now consider an issue which is critical to all *in situ* searches for life on other planetary bodies.

- Suppose you send a spacecraft to another Solar System body with an instrument designed to detect the signs of life, and that it is successful in its aim. How can you be sure that what you have detected is from the body that you have visited and not carried inadvertently by your spacecraft from Earth?

❑ Well you can't be sure unless you have been meticulous in ensuring that the chance of the spacecraft carrying micro-organisms from Earth and subsequently 'contaminating' the landing site has been eliminated.

After all, you have already come across one case where the transfer of such micro-organisms from Earth could have taken place. Figure 2.23 shows the camera from the Surveyor-3 spacecraft. On return to Earth after 2.5 years on the lunar surface, living terrestrial organisms were found in the foam inside the camera – the potential for contaminating other bodies had been clearly demonstrated.

The pursuit of preventing such contamination and all the issues relating to it are generally referred to as 'planetary protection'. There even exists a United Nations Treaty to which most space-active nations are signatories that states 'Parties to the treaty shall pursue studies of outer space……so as to avoid their harmful contamination and also adverse changes in the environment of the Earth……'.

United Nations (1967) Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies. U.N. Document No. 6347.

Remember, however, that contamination from the Earth can probably never be 100% eliminated – we can only minimize the chance of it occurring. This means that the results of all in situ experiments designed to detect extraterrestrial life must be treated with great care.

The major focus of planetary protection procedures and protocols has been on preserving other planetary environments from contamination by organisms from Earth that might grow there and thus obscure *forever* any efforts to understand the origin and evolution of life at locations other than Earth. But planetary protection is concerned not only with forward contamination, as it is called, but also back contamination, namely the return to Earth, and release into the biosphere, of potentially harmful organisms or substances. However, it is the former on which we shall concentrate here.

Clearly, different space missions to different environments and to perform different tasks do not all require the same degree of planetary protection.

■ Can you suggest what factors dictate the degree of planetary protection activity required?

❑ Factors to be considered include:

(i) the target body,

(ii) whether the spacecraft is intended to land on the target,

(iii) whether the spacecraft will return to Earth.

In the earlier days of space research, the analysis of the likelihood of contamination was implemented using a probabilistic approach. More specifically, the probability of contamination, P_c, was divided into two components, namely $P_c = P_t \times P_g$ where P_t is the probability of an organism's survival during transit from Earth surface to the surface of the target body and P_g is the probability of growth (and reproduction) of a contaminating organism on the target body.

Bioload (or bioburden) is a term used to describe the number of viable organisms present on an item.

The aim then would be, by controlling the bioload of the spacecraft and by determining the value of P_t and P_g by measurement and analysis, to ensure that the value of P_c for the spacecraft's entire bioload is less than a predefined value, often taken to be 10^{-3}.

■ What does a value for $P_c = 10^{-3}$ mean?

❑ It means that if 1000 such missions were sent to a particular planet, one of them would result in contamination.

However, this approach gradually fell into disrepute due to the difficulty of realistically estimating some of the probability terms. Instead, a simplified more robust approach has gradually been adopted.

Through a worldwide organization of scientists known as COSPAR (Committee on Space Research), the world's space-faring nations, and others, have agreed a unified approach to planetary protection. To simplify matters, all space missions that travel to a body in the Solar System are categorized into one of five categories, depending on the risk of contamination, both forward and backward. These categories, I to V in increasing order of risk, each have a different series of requirements concerning planetary protection associated with it. The categories (approved in 1984) are as follows:

- Category I missions include any mission to a target planet that is not of direct interest for understanding the process of chemical evolution. In effect, no protection of such planets (such as Mercury for example) is warranted, and so no planetary protection requirements are imposed.
- Category II missions are all types of missions to those target planets that are of significant interest for understanding the process of chemical evolution, but for which there is only a remote chance that contamination carried by a spacecraft could jeopardize future exploration. The concern is primarily over unintentional impact, since these missions are not designed to land.
- Category III missions are certain types of missions (fly-by and orbiter) to a target planet of interest for understanding the chemical evolution and/or the origins of life, or for which scientific opinion suggests a significant chance of contamination that could jeopardize a future biological experiment.
- Category IV missions are certain types of missions (mostly probe and lander) to a target planet of interest for understanding chemical evolution and/or the origins of life, or for which scientific opinion suggests a significant chance of contamination that could jeopardize future biological experiments.
- Category V missions include all Earth-return missions. The concern is for the protection of the terrestrial system as well as the scientific integrity of the returned sample.

One result of the application of the COSPAR regulations is that once the appropriate category for a spacecraft has been determined, it is possible to determine the maximum allowed bioload for that particular spacecraft. As an example, when these regulations are applied to a Category IV mission to Mars, a maximum of 300 000 organisms are allowed at launch. When the effects of the journey to Mars are taken into account, the intent is that the spacecraft is effectively 'biologically inert' when it arrives. To put the value of 300 000 into context, it should be appreciated that one sneeze disperses something like 1000 000 organisms.

■ There are three broad sources of microbiological contamination when a spacecraft is being assembled. Can you suggest what these are?

❑ They are:

the environment in which the spacecraft is assembled (5%),

the materials out of which the spacecraft is constructed (15%),

the people who assemble the spacecraft (80%).

The figures in brackets represent the approximate contribution of each to the total bioload for a spacecraft built in a clean-room environment.

There is a wide range of techniques available for reducing the bioload resulting from each of these categories. For example, contact between humans and the spacecraft is kept to a minimum. When absolutely necessary, techniques are used to ensure minimal transfer of organisms by the use of barrier clothing and appropriate procedures. The environment in which the spacecraft is constructed is very rigorously controlled by the use of so-called clean-room techniques. For a Category IV mission, a typical environment would be classified as 'Class 100' which means that there are less than 100 particles (including micro-organisms) of size greater than

0.5 μm within every cubic foot of air within the controlled environment (historically, SI and Imperial units have been mixed in this definition). The main active sterilizing techniques that are used for planetary protection applications, in increasing order of complexity (and effectiveness) are given in Table 3.8.

Table 3.8 Sterilizing techniques used in planetary protection.

Technique	Notes:
Cleaning of individual components	Levels and type of cleaning will depend on the particular measurements being performed during the mission. Typical cleaning techniques include detergent cleaning, solvent cleaning and hot helium purge.
Surface sterilization	Same techniques as for individual components.
Lander sterilization	Moist or dry heat which kills micro-organisms principally by oxidation. In this technique, sufficient heat has to be delivered throughout the whole of the sample. In addition, all materials within the sample must be able to withstand high temperatures (for example, 135 °C for 8 hours or 125 °C for about 50 hours).Gas plasma sterilization in which the sample is immersed in a hydrogen peroxide plasma. It is believed that this technique is less damaging to electronic components than heating techniques. Gamma-radiation. Some electronic components and also optical glasses can be damaged by this process.

An example of the effect of moist heat sterilization is shown in Figure 3.24.

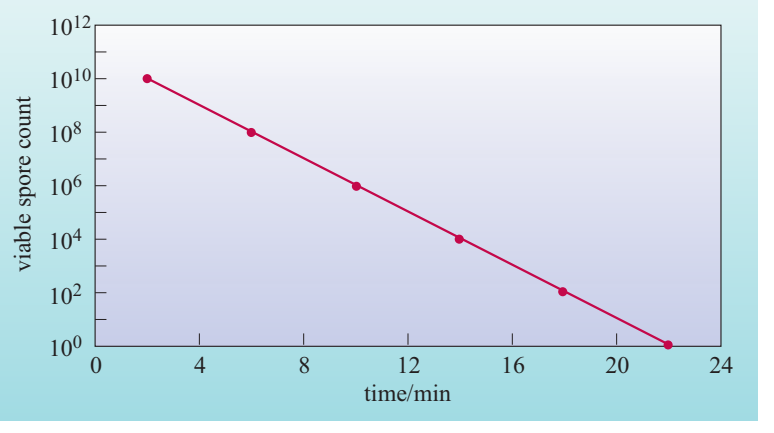


Figure 3.24 A graph showing the survival rate of spores on a sample as a function of time as a result of moist heat sterilization at 120 °C.

QUESTION 3.7

Place each of the following (hypothetical) space missions in the appropriate planetary protection category (I to V), giving your reasons in each case:

(i) cometary nucleus lander,

(ii) Mercury orbiter mission,

(iii) Jupiter orbiter with Mars fly-by en route,

(iv) Mars orbiter and lander,

(v) mission to return cometary dust to Earth.

3.7 Habitats for life

We shall end this chapter with a brief general consideration of likely Martian habitats for life.

■ If we wish to search for signs of extinct or extant life on Mars, and knowing what we do about the prerequisites for life, where should we look?

❑ Anywhere there has been or is water.

However, there are factors, which militate against the development of life in the presence of water. The short wave ultraviolet radiation (see Box 3.7) and the oxidation due to surface peroxide means that the dry dusty soil on the surface is unlikely to be a suitable habitat for life. But terrestrial experience (see Chapter 2) suggests that microscopic life is capable of colonizing even the most extreme of occurrences of liquid water. For example, extremes of pH, salinity, temperature and pressure are not necessarily barriers. However, you have already seen that Mars has certainly experienced periods when the atmosphere was thicker and the surface warmer. Maybe therefore during these epochs, the surface was sufficiently protected for life to have developed there. So any environment indicative of water deposits might have been a suitable environment. These include lakes, thermal springs and glaciers – evidence of all these features are found on the present day surface of Mars.

BOX 3.7 MARS AND ULTRAVIOLET RADIATION

The thin atmosphere and the low concentration of ozone and oxygen means that the Martian surface is exposed to high levels of harmful solar UV radiation. Figure 3.25 shows the (measured) UV flux at the surface of the Earth and the (predicted – as it has never yet been measured directly) flux at the surface of Mars.

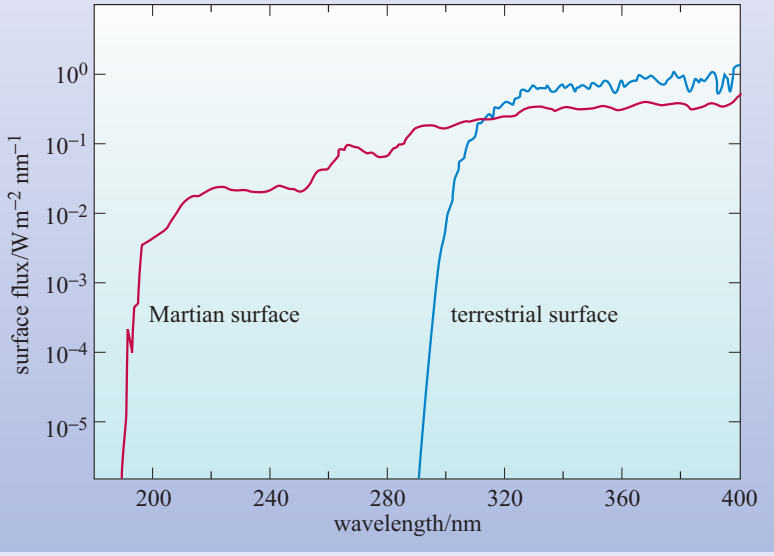


Figure 3.25 The fluxes of UV radiation at the surface of the Earth and Mars.

■ What is the most obvious difference between the two spectra shown on Figure 3.25.

❑ The terrestrial UV spectrum 'cuts-off' at about 290 nm but on Mars it extends right down to below 200 nm.

This factor has a dramatic effect. The ultraviolet spectrum is traditionally divided into three regions, namely UV-A (315–400 nm), UV-B (280–315 nm) and UV-C (200–280 nm). Interaction of UV radiation with organic material occurs across the entire range, but varies in severity as a function of wavelength. Long wavelength UV-A is the least damaging, responsible for common effects such as tanning of the skin. UV-B (280–315 nm) is partially obscured on Earth, and is responsible for increased biological damage, causing effects such as sunburn. UV-C is the extreme case, accounting for alteration and mutation of biological organisms at the genetic level resulting in severe mutation and in some cases complete destruction. As can be seen from Figure 3.25 life on Earth is shielded from this harmful short wave UV radiation, while on Mars the atmosphere offers little protection. Environments which are protected from this radiation, such as beneath rocks or even below the surface, should therefore be favoured as habitats for life.

These environments should therefore all be amongst the targets for future searches for extinct life on Mars. Other likely targets would include evaporite deposits that are produced when a lake shrinks or disappears by evaporation – these deposits can capture other constituents including biogenic constituents. Such evaporites may have a long residence time on Mars due to the hypothesized early termination of any hydrologic cycle (involving precipitation), the absence of tectonic activity and the current dry surface conditions.

Hydrologic cycle is the term used to describe the process by which water is cyclically transferred from a planet's surface (mostly by evaporation) into the atmosphere, and back to the surface by precipitation.

For present day (extant) life, the most likely environments could be quite different.

■ In view of what we know of present day conditions on Mars, where are the most probable environments for the survival of extant life forms?

❑ In subsurface environments, protected from UV and the oxidizing nature of the surface.

Critically, it appears that certain extremophiles do not require sunlight to thrive – they derive their energy from chemical reactions. In fact on Earth, some populations have been found that have existed completely isolated from the surface for millions of years. This seems to open the door to subsurface ecologies on other worlds and may have profound implications for environments such as Europa (see Chapter 4). So on Mars, all subsurface environments that may harbour water should be regarded as candidates for the development of extant life, such as those identified by the Mars Global Surveyor GRS instrument (as described in Section 3.4.1).

The above discussion makes the implicit assumption that if life has existed on Mars, it was initiated spontaneously there. But we know from the very existence of the SNC meteorites that this might not necessarily be the case.

■ Why are the SNC meteorites relevant to this issue?

❑ They show that, in principle, life from another origin (for example, the Earth) could have been transferred to Mars on ejecta from Earth by an impact. Although highly speculative, this should not be completely discounted. This could have implications for likely habitats for the development of life on Mars. Such habitats might have been therefore not the optimum ones but ones that were colonized by chance.

In addition to the above general considerations for landing site selection for any future space missions designed to search for Martian life, other more specific considerations include:

1 Concentration – It is quite important that the material indicating the presence of biota is concentrated in appreciable amounts since the first searches on the surface of Mars will probably have only limited capability for mobility and surface coverage.

2 Preservation – The preservation of the evidence is of paramount importance. Microbial material should be fossilized rapidly – this is likely to occur in an environment such as hydrothermal springs. Organic molecules and other chemical evidence may require rapid burial to avoid alteration at the surface.

3 Thin dust cover – MGS images have confirmed that the Martian surface is extensively affected by dust deposition (see also the Mars Pathfinder image in Figure 3.10 where fine-grained sediment drifts cover much of the scene, and in some places the environment has been clearly scoured and sculpted by wind). The wind-blown detritus can cover large parts of the planetary surface and the selected landing site should thus show indications of as thin as possible dust coverage.

4 Area of the target – The target for exploration should ideally be really extensive in order to be within the landing uncertainties for any lander, as it is not possible yet to land on Mars with perfect targeting precision.

These factors are not merely of academic interest. The next 20 years will see a plethora of space missions designed to try to answer once and for all the question of whether life ever has existed (or does exist) on Mars. The first of these is Mars Express carrying the Beagle 2. Later, we can anticipate a Mars sample/return mission in which material from Mars will be returned to Earth for the most sophisticated of analysis techniques which are available in terrestrial laboratories and ultimately a human exploration mission. By then, we can confidently hope that the question of whether life on Mars has ever existed will be finally settled.

3.8 Summary of Chapter 3

- Historical beliefs that Mars may be vegetated and inhabited were contradicted by observations from early space probes. These showed Mars to be a cold, arid habitat with a thin atmosphere unable to provide much protection against biologically harmful ultraviolet radiation and therefore superficially unlikely to be a suitable locale for life. Furthermore, it appears that water cannot exist in equilibrium under the average conditions on the surface of Mars.

- The apparently negative results from the Viking mission biology experiments appeared to kill off any speculation concerning the existence of life, despite some ambiguities. The experiments failed to detect any organic matter in the Mars soil, either at the surface or from samples collected a few centimetres below the surface. The results were interpreted as meaning that strong oxidative processes were at work at the surface.
- Evidence from subsequent space missions, such as Mars Global Surveyor and Mars Odyssey, has confirmed previous suspicions of the past existence of significant bodies of surface water and their possible existence even in recent times. They also provided strong evidence for the present existence of large deposits of subsurface water-ice.
- Evidence for earlier periods of a thicker atmosphere with a more significant presence of water also comes from analysis of the Martian meteorite EET A79001.
- Coupled with the discovery of various terrestrial extremophiles, this evidence has resulted in a re-assessment of the possibility of extant or extinct life on Mars.
- Claims for direct evidence of Martian fossils in the meteorite ALH 84001have remained highly controversial within the scientific community with opinion as to their veracity being divided. Most scientists believe that the observed features can be explained by non-biological processes.
- Future space missions to Mars will continue to search for evidence of extinct or extant life forms. The next such mission is Mars Express which carries the Beagle 2 lander.

CHAPTER 4
ICY BODIES: EUROPA AND ELSEWHERE

4.1 Introduction

Until the 1980s, the icy satellites of the outer planets were scarcely thought of as places where life could ever have existed. Few could have imagined that one of them, Europa, would within twenty years have become the rival of Mars as a priority for astrobiological study. This chapter recounts the history of our changing perceptions of the icy satellites, examines the available evidence for their internal structures, and considers the niches offered for life to begin and to be sustained. In this context, the 'habitable zone' embraces settings devoid of both sunlight and an atmosphere. These are areas where life could survive on the energy from chemical reactions made possible by the discharge of hot chemically enriched fluids through vents on the floor of an ocean capped by a thick layer of ice.

'Ice' does not necessarily mean just frozen water. In the outer Solar System, although H_2O is usually the dominant component, ice can incorporate other frozen volatiles such as NH_3, CO_2, CO, CH_4 and N_2.

4.1.1 Satellite discoveries

All the giant planets have satellites. Jupiter's four largest satellites were discovered in 1610 by Galileo Galilei (Figure 4.1), using one of the first telescopes to be pointed at the night sky. These are now known as the **Galilean satellites**. They are much bigger than Jupiter's other satellites, the first of which was not discovered until 1892. Saturn's largest satellite, Titan, was discovered in 1655, and four more had been found by 1700.

Sir William Herschel (Figure 4.2) discovered the first two of Uranus's satellites in 1787, less than six years after he had discovered the planet itself. Neptune's largest satellite, Triton, was discovered by William Lassell (Figure 4.3) in 1846 – within three weeks of the planet being identified. Smaller and fainter satellites continued to be found. By 1950 the known tally of outer planet satellites was Jupiter, eleven; Saturn, nine; Uranus, five; and Neptune, two.

Discoveries of lesser satellites only a few kilometres across continue to be made. In the competition to be the planet with the largest number of known satellites, the lead has changed several times between Jupiter, Saturn and Uranus. However, all the satellites of the giant planets that are large enough for their own gravity to pull

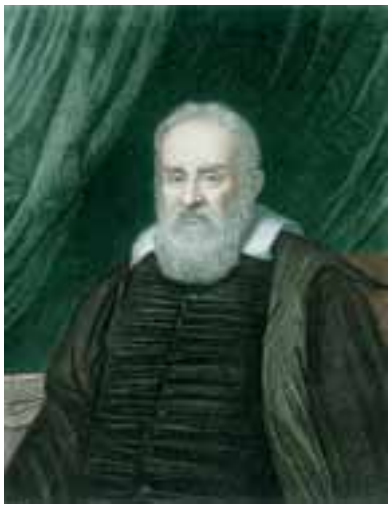


Figure 4.1 Galileo Galilei, 1564–1642. Pisa-born pioneer of the experimental scientific method, whose analysis of motion paved the way for Isaac Newton's work. Galileo used one of the first telescopes to discover the four largest of Jupiter's satellites and the phases of Venus. His consequent support for the theory that the Earth moves around the Sun led to his imprisonment for heresy in 1633.

Figure 4.2 Sir William Herschel, 1738–1822. Born in Hanover, Herschel moved to England as a young man to work as a musician. He became an astronomer and was elected a Fellow of the Royal Society in 1781, on the strength of his lunar observations and his discovery of Uranus. Using his own 48-inch (122 cm) reflecting telescope, he discovered Titania and Oberon (satellites of Uranus) in 1787 and then Enceladus and Mimas (satellites of Saturn) in 1789.

Figure 4.3 William Lassell, 1799–1880. A Liverpool businessman who made his fortune in the brewing trade. He designed and built his own telescopes, including a 24-inch (61 cm) reflector, with which he discovered Triton in 1846 and two satellites of Uranus (Ariel and Umbriel) in 1851.

them into a near-spherical shape have certainly been found. For an icy body, this means the satellite must have a radius of more than about 200 km. These larger bodies are the satellites of greatest potential for astrobiology, and their basic properties are listed in Table 4.1. Two of these satellites are larger than the planet Mercury, but not so massive, because their densities are less. Four are bigger and more massive than the Moon, and a total of six are bigger and more massive than Pluto. Pluto itself (discovered in 1930) and its satellite Charon (discovered in 1978) share many of the characteristics of the large icy satellites, and so they are also listed in the table.

4.1.2 Satellite systems and their origins

The satellite systems of the giant planets have several features in common. Most satellites are in synchronous rotation, always keeping the same face towards their planet. Irregularly shaped moonlets associated with the ring system orbit closest to the planet. They travel in near-circular prograde orbits in the planet's equatorial plane. These moonlets (like the rings) are believed to be fragments of larger satellites that were destroyed by collisions or tidal forces (Figures 4.4 and 4.5). Most are bright and presumed to be icy in composition.

'Prograde' in this sense means orbiting in the same direction as the planet's spin.

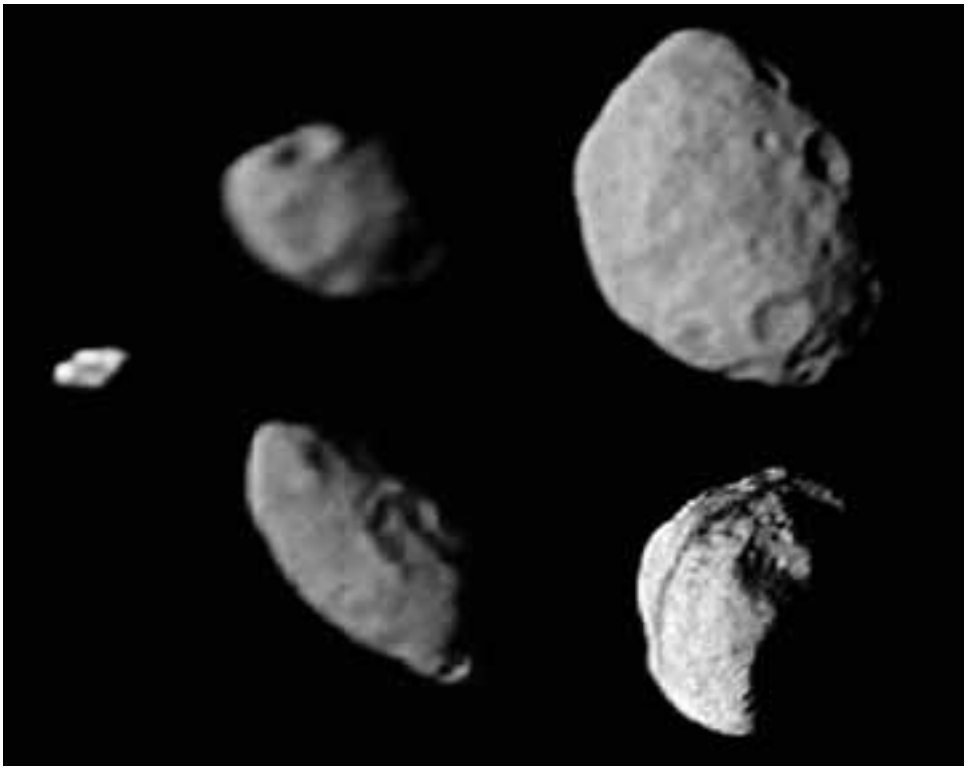


Figure 4.4 Five of Saturn's innermost satellites as imaged by the Voyager spaceprobes, shown at their correct relative sizes. From left to right: Atlas, Pandora (above) and Prometheus (below), Janus (above) and Epimetheus (below). Janus is 99 km in length. The dark line across Epimetheus is the shadow of one of the narrowest of Saturn's rings.

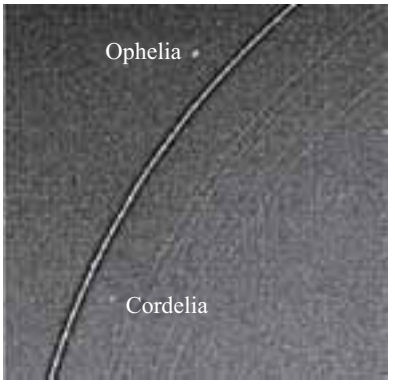


Figure 4.5 A Voyager 2 image showing Uranus's outermost and most prominent ring, which is kept in shape by the small satellites Ophelia and Cordelia (each less than about 30 km across), one of which orbits 2000 km inside the ring and the other 2000 km beyond it. The ring consists of millions of dark, dusty boulders, mostly 10 cm to 10 m in size. Four fainter rings are just visible within the orbit of Cordelia.

Table 4.1 Basic data for the satellites of the outer planets. In the *orbital period* column, R indicates retrograde orbits. Values were up to date in March 2003, but are subject to revision. Where two or more values are given in the *radius* column, these indicate a non-spherical satellite and are the dimensions (semi-major axes) of the best-fit ellipsoids to the satellite's actual shape. The numbers of small satellites are correct as of early 2003, but are subject to change as new discoveries are made.

Planet	Satellite	Mean distance from planet/10^3 km	Orbital period/ Earth days	Radius/km	Mass/10^{20} kg	Density/ 10^3 kgm^{-3}
Jupiter	4 inner	<221.9	<0.675	<125	–	–
	Io	421.6	1.77	1821	893	3.53
	Europa	670.9	3.55	1565	480	2.99
	Ganymede	1070	7.15	2634	1482	1.94
	Callisto	1883	16.7	2403	1076	1.83
	40 outer	>7435	>130	<85	–	–
Saturn*	6 inner	<151.4	<0.695	<99	–	–
	Mimas	185.5	0.942	199	0.375	1.14
	Enceladus	238.0	1.37	249	0.649	1.00
	Tethys	294.7	1.89	530	6.28	1.00
	Dione	377.4	2.74	560	10.5	1.44
	Rhea	527.0	4.52	764	23.1	1.24
	Titan	1221.9	16.0	2575	1346	1.88
	Hyperion	1481.1	21.3	165 × 113	0.11	1.1
	Iapetus	3561.3	79.3	718	16	1.0
	Phoebe	12952	551R	115 × 105	0.007	2.3
	12 outer	>11300	>449	<16	–	–
Uranus	11 inner	<86.0	<0.762	<77	–	–
	Miranda	129.8	1.42	236	0.659	1.20
	Ariel	191.2	2.52	579	13.5	1.67
	Umbriel	266.0	4.14	585	11.7	1.40
	Titania	435.8	8.71	789	35.3	1.71
	Oberon	582.6	13.5	761	30.1	1.63
	6 outer	>7169	>579R	<80	–	–
Neptune	5 inner	<73.5	<0.55	<104	–	–
	Proteus	117.6	1.12	218 × 208 × 201	0.49	1.3
	Triton	354.7	5.88R	1353	215	2.05
	Nereid	5513	360	170	0.3	1.5
	3 outer	>20200	>2516	<40	–	–
Pluto		–	–	1150	131	2.0
	Charon	19.4	6.39	586	16.1	1.9

*Saturn has three other tiny satellites: Telesto and Calypso that share the orbit of Tethys, and Helene sharing the orbit of Dione.

Orbiting further from each planet come all the satellites large enough to be spherical (or nearly so) in shape, typically in near-circular prograde orbits close to the planet's equatorial plane. These satellites probably grew within a disc of gas and dust that surrounded the planet in the later stages of its growth, mimicking in miniature the birth of the terrestrial planets from the solar nebula. Neptune's large satellite, Triton, is an exception (Figure 4.6). This has a retrograde orbit, and may be a Pluto-like Kuiper Belt object that was captured into orbit around Neptune some billions of years ago.

Beyond its large satellites each giant planet has a second collection of small irregular-shaped satellites, travelling in elongated, inclined and in many cases retrograde orbits. Most are dark bodies, rich in silicates and/or carbon compounds. These satellites are likely to be captured comets or asteroids (Figure 4.7).

We have used the term 'Kuiper Belt' but you may also see it called the 'Edgeworth–Kuiper Belt'. Kenneth Edgeworth, a British astronomer, published similar ideas a few years prior to Kuiper, but this only really came to light after the term 'Kuiper Belt' had become widely established.

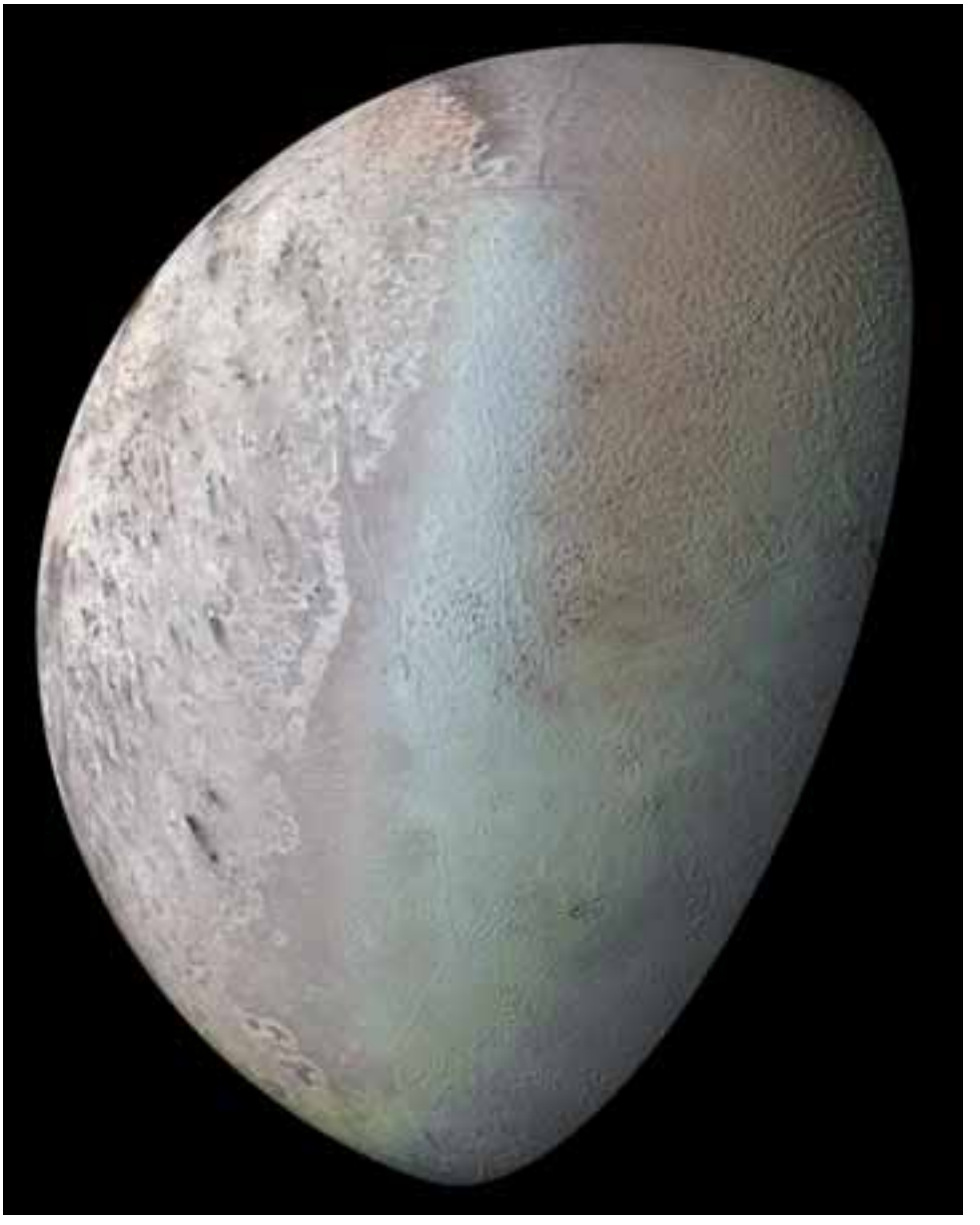


Figure 4.6 A Voyager 2 image of the sunlit part of the Neptune-facing hemisphere of Triton. The south polar cap of bright nitrogen-ice is marked by streaks of sooty (carbon-rich) material erupted from geysers. The rugged surface beyond the polar cap is methane-rich ice, contaminated by nitrogen, carbon dioxide, carbon monoxide and water.

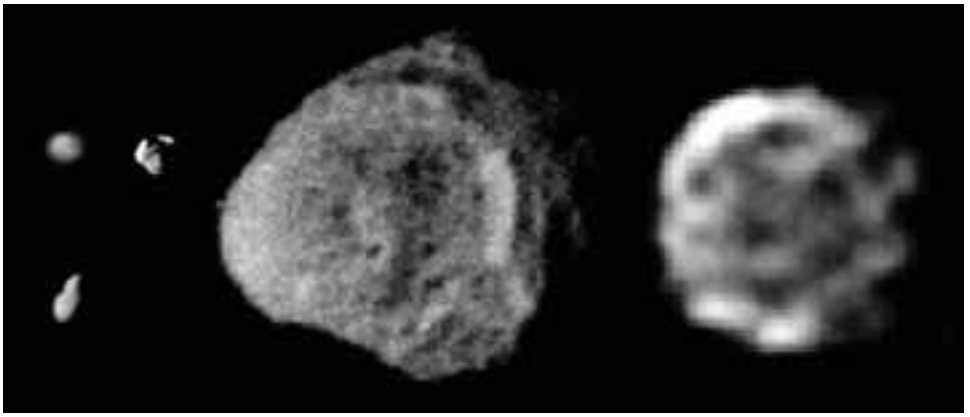


Figure 4.7 Some of the outer small satellites of Saturn shown at their correct relative sizes. From left to right: Telesto (above) and Calypso (below), Helene, Hyperion and Phoebe. Hyperion is 185 km in length.

4.1.3 Unravelling the natures of the large satellites

Before the dawn of the space age, relatively little could be discovered about even the large satellites. Their orbits were well known, and from the subtle orbital perturbations caused by neighbouring satellites it was possible to deduce their masses. Measurements of their sizes enabled densities to be calculated to within about 20% of the currently accepted values for the Galilean satellites, and with rather less certainty for the large satellites of the other giant planets. However, it was clear that, except for Io and Europa, these bodies are not dense enough to be composed largely of rock like the terrestrial planets.

During the 1950s, spectroscopic studies by Gerard Kuiper (Figure 4.8), the discoverer of Titan's atmosphere (Section 5.2), showed that the surface of Europa is mostly clean bright water-ice, whereas that of Ganymede (which has a lower albedo) is water-ice darkened by a dusty contaminant. Spectroscopic studies have now revealed that ice dominates the surfaces of all the large satellites except Io, which is effectively a terrestrial planet in orbit about Jupiter. In the Jupiter system, the **ice** is dominantly frozen water, but with increasing distance from the Sun it becomes mixed with more volatile ices. There is indirect evidence for ammonia in the ices of Uranus's satellites, and on Neptune's large satellite Triton spectroscopic observations have detected frozen nitrogen, carbon dioxide, carbon monoxide and methane. A similar mixture to that on Triton coats Pluto's surface.

We use the term 'water-ice' where necessary to make it clear that we mean frozen water, as opposed to any other kind of ice.

Figure 4.8 Gerard Kuiper, 1905–1973. Dutch-born American planetary scientist who discovered Titan's atmosphere in 1944 and subsequently used spectroscopy to identify carbon dioxide in the atmosphere of Mars and ice on the surfaces of Europa and Ganymede. He discovered Miranda (Uranus) in 1948 and Nereid (Neptune) in 1949. In 1951 he suggested that there should be a zone of primordial debris beyond the orbit of Neptune. Although the first body in this zone was not discovered until nearly twenty years after his death, it is generally known as the Kuiper belt.

All of this is consistent with our understanding of the nature of the materials from which the Solar System formed, under conditions of progressively lower temperatures at greater distances from the Sun.

The icy satellites came to be regarded as worlds made of ice mixed with rock because their densities are greater than any variety of ice. This was because the silicate minerals that form rock constitute the most abundant denser material known to exist in the Solar System. Whether these satellites are differentiated bodies with the rock forming a dense core surrounded by a less-dense icy mantle, or whether they are undifferentiated uniform mixtures of rock and ice was assumed to depend on their accretion histories. An undifferentiated structure would imply homogenous accretion (rock and ice simultaneously) combined with insufficient heating to trigger differentiation. A differentiated structure could result from heterogeneous accretion (rock first, then ice) or from homogenous accretion if the rate of energy release during the accretion process generated enough heat to melt or at least mobilize the ice.

QUESTION 4.1

If a body of average density ρ_{av} consists of a mixture of just two components, a dense one with density ρ_{dense} and a light one with density ρ_{light}, the way to work out what fraction of the body's volume is made of each is as follows. Let the fraction made of the dense component be x. The fraction made of the light component must then be $(1 - x)$.

There is a simple equation relating these values:

$$\rho_{av} = x\,\rho_{dense} + (1 - x)\rho_{light} \qquad (4.1)$$

(a) Use Equation 4.1 to calculate the fraction of Callisto's volume occupied by rock, given that Callisto's average density is $1.83 \times 10^3\ \text{kg m}^{-3}$. Assume the density of rock to be similar to that of chondritic meteorites, which is about $3.1 \times 10^3\ \text{kg m}^{-3}$ and the density of ice to be about $0.95 \times 10^3\ \text{kg m}^{-3}$.

(b) Suggest some factors that could make the value calculated in this way unreliable.

Irrespective of whether the rock is dispersed or concentrated, the total rock content of these bodies is too low for radiogenic heating, by the decay of radioactive elements contained within the rock, to provide sufficient heat to mobilize their interiors and refresh their surfaces. In the 1960s, the average surface temperatures of the Galilean satellites were established to be lower than –150 °C using infrared telescopes. This is so low that the ice near the surface must have comparable mechanical properties to rock near the surface of a terrestrial planet. Such ice is far too cold to behave like glacier ice on Earth, which is capable of flowing downhill under its own weight. Thus, whatever their internal structure and their mode of origin, all the icy satellites at Jupiter and beyond (where surface temperatures are even lower) were assumed to have long been geologically dead, with the implication that they must be densely covered by impact craters that have built up during the past four billion years.

Just how wrong some of these suppositions were did not become apparent until close-up images of the satellites of the outer planets were sent back by spacecraft. Only the merest hints were provided by the blurry images returned by the first probes to visit Jupiter, Pioneers 10 and 11 in 1973 and 1974. The situation became much clearer thanks to the remarkable tours of the outer Solar System accomplished by the two probes of NASA's Voyager series, beginning with Voyager 1's encounter with Jupiter in March 1979 and ending with Voyager 2's fly-by of Neptune in August 1989 (Box 4.1). These revealed a startling diversity of landscapes on the icy satellites. Some are indeed heavily cratered, and look much like what most people expected (Figure 4.10). But others have a complex variety of terrain types, showing relatively few impact craters but many signs that faulting, flooding and other resurfacing processes have acted to disrupt or bury any ancient heavily cratered terrains that may formerly have existed (Figure 4.11).

BOX 4.1 THE VOYAGER PROJECT

In 1977, NASA launched two probes named Voyager to explore the outer Solar System (Figure 4.9). Voyager 1 flew through the Jupiter system in March 1979, and used Jupiter's gravity to redirect its trajectory towards Saturn, which it passed in November 1980. Voyager 2 used the same 'gravity assist' tactics to visit all four giant planets in turn, beginning with Jupiter in July 1979 and concluding with Neptune in August 1989.

Each of the Voyager probes weighed 825 kg, of which 105 kg was scientific instruments. These included cameras, spectrometers, polarimeters (to measure polarization of reflected radiation) and magnetometers. Because it was designed to travel so far from the Sun, power was provided not by solar panels but by the heat produced by radioactive decay in a plutonium-rich thermoelectric generator.

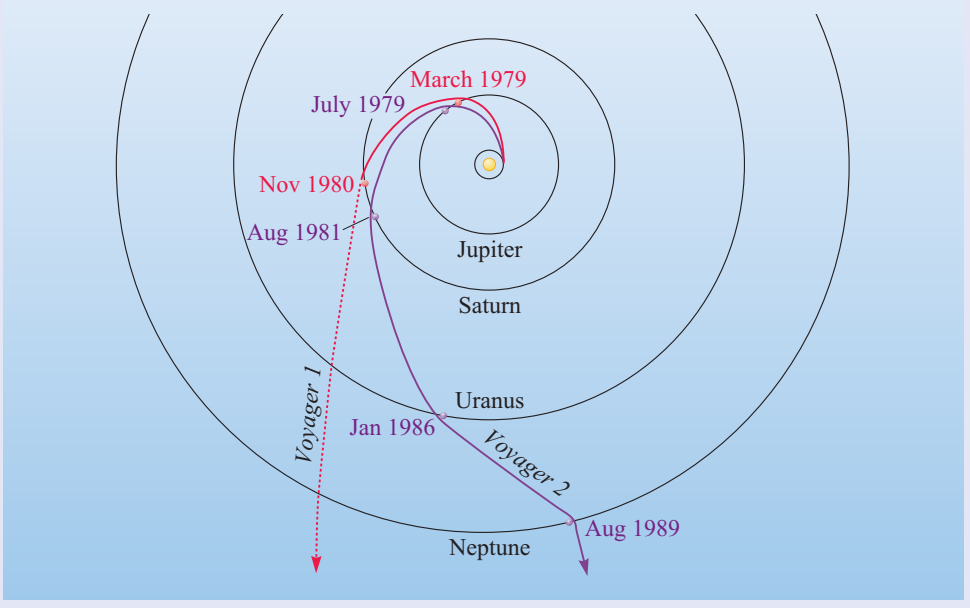
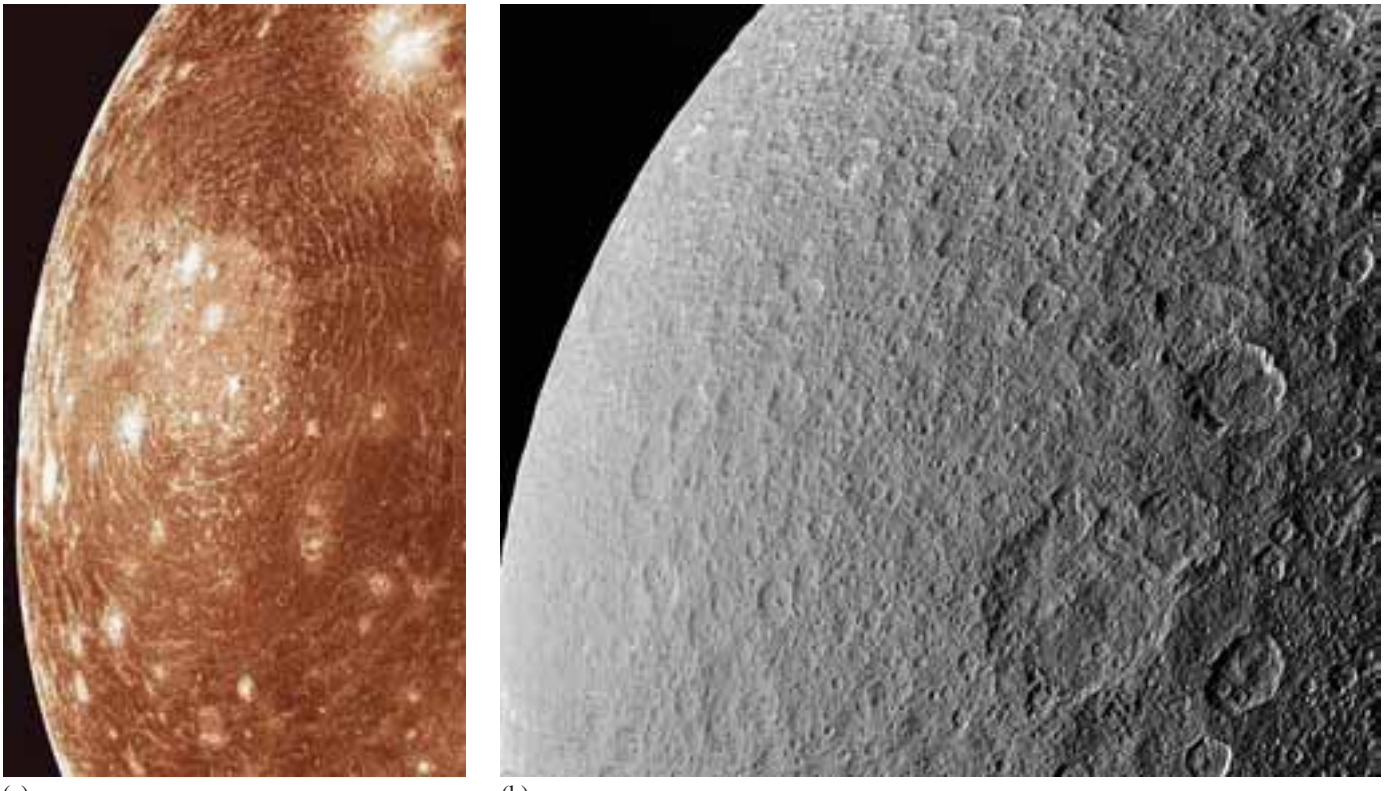


Figure 4.9 The trajectories of the two Voyager spacecraft. Voyager 1's encounter with Saturn flung it onward above the plane of the Solar System. After Neptune, Voyager 2's course took it below the plane of the Solar System.



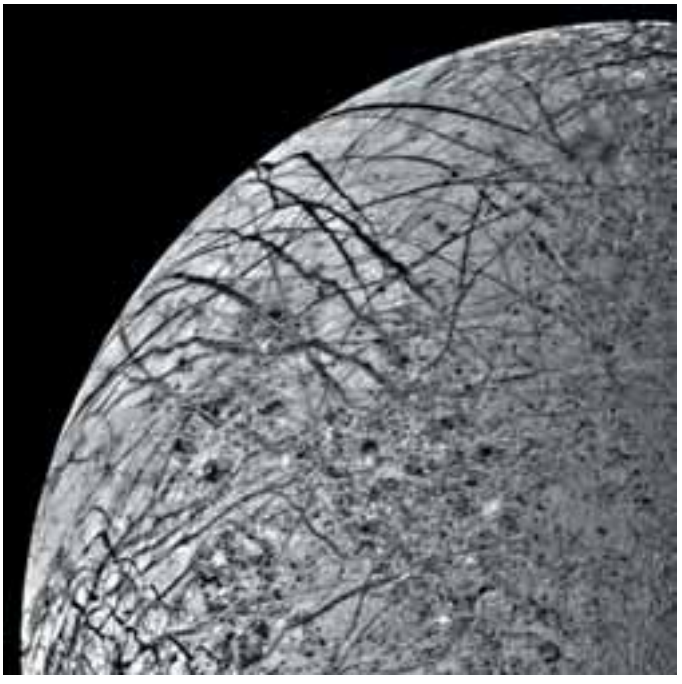
(a) (b)

Figure 4.10 Two icy satellites whose heavily cratered appearance suggests passive worlds with little or no geological activity: (a) Callisto (image 3000 km from top to bottom) (b) part of Rhea (image 600 km from top to bottom).

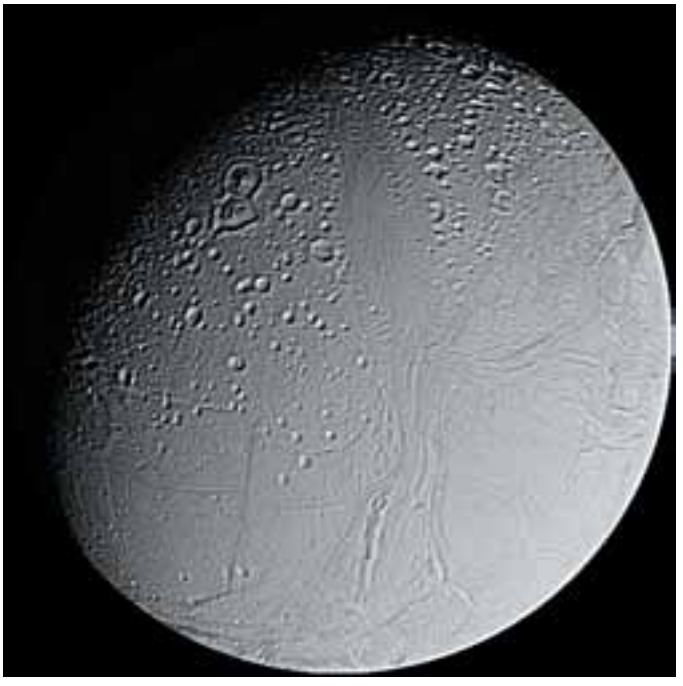
With the exception of Titan, no satellite has an atmosphere thick enough to protect its surface from bombardment. The surface of an icy satellite scatters sunlight fairly evenly in all directions, which means that not even the youngest surface consists of a continuous sheet of smooth ice. Instead, any ice that was a continuous sheet originally has become broken (presumably by meteorite and micrometeorite impact) into a mass of granular fragments, with a wide range of particle sizes, in the same way that the lunar surface consists of a **regolith** of rock debris. Presumably the icy regolith is thinner (only a few particles in thickness) on the youngest icy surfaces and thickest (several metres or more) on the oldest surfaces.

Unfortunately, the resurfacing events on satellites such as those in Figure 4.11 are impossible to date. The lunar cratering timescale, which has been calibrated radiometrically (i.e., using dating methods based on the decay of radioactive isotopes), cannot be applied in the outer Solar System. This is because we can have no expectation that the Moon (1 AU from the Sun) suffered the same rate of impact bombardment as a satellite of Jupiter (5 AU from the Sun) or a satellite of Saturn at a range of nearly 10 AU. Indeed, when the size–frequency distributions of craters on the icy satellites are examined, it is found that the pattern of distribution of craters versus size range differs from the satellites of one giant planet to the next, and that each is different to that found on the Moon. This is convincing proof that *different populations* of impactors affected each region of the Solar System, so it is likely that cratering *rates* also behaved differently in each region. We can imagine a general decrease over time, but there may have been localized flurries of cratering, such as would be caused by the impact of debris originating from a nearby satellite

'Size–frequency distribution' is a term used to describe the relative numbers of objects (in this case, craters) across a range of sizes.



(a)



(b)

(c)

(d)

Figure 4.11 Some icy satellites that shattered preconceptions. (a) Europa, which at this scale appears practically devoid of impact craters and must therefore have a very young (<100 Ma) surface; (b) Enceladus, where heavily cratered terrain is cut by tracts of a crater-poor and therefore younger surface; (c) Miranda, whose surface is a patchwork of contrasting terrains; (d) Ariel, where the ancient heavily cratered terrain is disrupted by fault-bounded valleys, some of whose floors have become flooded by icy (cryovolcanic) 'lava' flows.

that had been broken apart by a single, random, exceptionally large impact. Thus, while we can be reasonably confident that a less densely cratered surface is younger than a more densely cratered surface when comparing between satellites of the *same* giant planet, we cannot make any such comparison between a less densely cratered surface on a satellite of, say, Jupiter, and a more densely cratered surface of a satellite of, say, Saturn.

On many individual satellites, the differences in crater density between the oldest terrain (which may be a surface that is four billion years old) and the youngest resurfaced terrain suggest that the latter is considerably younger. In order for their surfaces to have been regenerated from within, these satellites must have experienced internal heating.

There are several reasons why radiogenic heating can be ruled out as a significant factor in this heat generation:

- As established previously, the total rock content of these satellites is far too low to generate enough heat, unless the rock has some implausibly weird composition,with ten to a hundred times more radioactive elements than chondritic meteorites.
- Any exotic rock composition would have to vary enormously within a single satellite system to explain the geologically complex surface of Enceladus (Figure 4.11b) in contrast to the passive densely cratered surface of its fellow satellite Rhea (Figure 4.10b). Rhea is twenty-five times more massive and ought to be producing more radiogenic heat than tiny Enceladus.
- Radiogenic heating ought to decay gradually over time, which is not reflected in the resurfacing histories of the more active satellites.

4.1.4 The discovery of tidal heating

The Voyager fly-bys of the Jupiter system convinced planetary scientists that former preconceptions about ‘dead’ globes were wrong – even before Voyager 1 had got as far as Saturn, the mission had enabled them to identify a new heating mechanism to explain the discrepancies. The ease with which this revolution in thought was brought about was thanks to some of the Voyager images of Io, Jupiter’s innermost Galilean satellite. Io is only a fraction larger and denser than the Moon, and so by rights Io should have been geologically quiet for the past 2–3 billion years. However, Io has a surface so young that (even with the more detailed images obtained subsequently) even the youngest impact craters have been erased and there are usually several volcanoes erupting simultaneously (Figure 4.12). Some of the eruptions are large enough to track using modern infrared telescopes in high-altitude observatories such as on the summit of Mauna Kea, Hawaii.

Most planetary scientists were staggered to find active volcanoes on Io, but not the authors of a paper that had been published in the journal *Science* just a few days before Voyager 1 arrived at Jupiter. In this paper, Stanton Peale and colleagues proposed that Io’s interior should be largely molten because of heat generated by the tidal stressing experienced by Io as it orbits Jupiter. Although the degree of melting within Io remains open to debate, tidal heating (Box 4.2) was rapidly accepted as the power source for Io’s volcanoes and for the episodes of resurfacing on the icy satellites.

Figure 4.12 Eruption plumes on Io. This image was recorded on 28 June 1997 by the Galileo Orbiter, but it shows similar processes to those revealed 18 years previously by Voyager 1. The surface is dominated by lava flows, stained yellow and red by oxides of sulfur. An eruption plume can be seen rising 140 km above the limb (from a volcano named Pillan Patera), and a second eruption plume at the volcano Prometheus is seen from directly above near the centre of the disc. This is enlarged in the inset at the upper left. The bluish ring is the outline of the plume, which casts a reddish shadow onto the surface to its right. The black surface feature beneath the right-hand part of the plume is a lava flow that has erupted continually since 1979.

BOX 4.2 TIDAL HEATING OF SATELLITES

When a major satellite is orbiting a giant planet, the tidal attraction of the planet distorts the shape of the satellite. This creates a tidal bulge centred on the side facing the planet and an equal bulge centred on the opposite face. The size of these bulges depends on the mass and proximity of the planet (tidal force is inversely proportional to the cube of the orbital radius), and on the strength of the material of which the satellite is made. In the extreme case of Io, the bulges are several kilometres high. Distortion of the globe associated with changes in location or size of the tidal bulges is what leads to tidal heating. The heat is generated by a kind of internal friction. This is the phenomenon that occurs if you take a bar of relatively weak metal and flex it backwards and forwards at a single point. The bent portion soon becomes hot to the touch. You can easily observe this for yourself if you are willing to sacrifice a wire coathanger to the cause of science.

If a satellite were to be rotating faster than its orbital period, the tidal bulges would have to migrate around the satellite in order to try to stay lined up with the planet. The continual distortion of the globe required for this to happen would generate an enormous amount of heat. Such a situation may have occurred very early in a satellite's life (or in the case of Triton, shortly after capture), but in most cases it would take tidal forces only a few million years to slow a satellite's rate of spin until it exactly matched its orbital period. This is why virtually all large satellites are now in synchronous rotation.

Tidal drag also tends to coax a satellite into an exactly circular orbit. This is the fate of any single satellite orbiting a sufficiently massive planet, and when it has been achieved the tidal stresses become constant and there is no more tidal heating. However, while a satellite's orbit is still elliptical then, even with

synchronous rotation, there remain two reasons why the tidal stresses continue to vary, which allows tidal heating to continue.

1 In an elliptical orbit, the distance between planet and satellite is continually changing, and so the strength of the tidal force producing the tidal bulges varies accordingly. The bulges are slightly higher when the satellite is closer to the planet and lower when it is further away.

2 In an elliptical orbit, a satellite's speed varies with its distance from the planet (in accordance with Kepler's second law). However, the rate of the satellite's axial spin remains constant. Thus although for every orbit completed the satellite rotates exactly once, during the closest part of its orbit its rotation lags slightly behind its orbital motion, and during the furthest part of its orbit its rotation is slightly ahead of its orbital motion. Consequently, as seen from the planet, the satellite does not show exactly the same face throughout its orbit, rather it swings slightly from side to side. The tidal bulges are raised by forces acting directly on a line through the centres of the two bodies, and so their locations oscillate east and west across the satellite's surface.

So, for a satellite in an elliptical orbit, both the continual variation in the heights and the oscillation in the locations of the bulges deform the satellite's interior, and so cause heating. The reason why none of the orbits of the satellites of the giant planets has yet become exactly circular is that every satellite has neighbours. Mutual perturbations each time an inner satellite overtakes an outer (and therefore slower) satellite keep the orbits slightly elliptical, despite the tidal force from the planet.

This effect is magnified when satellites are in a situation of **orbital resonance**, i.e. where the orbital periods of satellites in adjacent orbits are simple ratios. This is particularly strong among the three inner Galilean satellites: Europa completes exactly one orbit for every two orbits by Io, and Ganymede in turn has exactly twice the orbital period of Europa. The resulting exaggerated eccentricity of the orbits, described as **forced eccentricity**, is slight (0.04 in the case of Io and 0.01 for Europa), but sufficient to power Io's volcanoes and the young (probably continuing) activity on Europa. It also explains why Ganymede shows plenty of signs of past geological activity, whereas Callisto shows few or no signs. (Although three times Callisto's orbital period is almost exactly seven times Ganymede's orbital period, this 7:3 orbital resonance does not lead to sufficient forced eccentricity of Callisto's orbit to lead to tidal heating, especially as Callisto is relatively far from Jupiter.)

Of the icy satellites, Europa (Figure 4.11a) has the youngest icy surface certainly in the Jupiter system and probably in the entire outer Solar System. Density models, supported now by more specific observations, suggest that Europa has about 100 km of icy material overlying a rocky interior. The rate of tidal heating within Europa must be less than in Io, because Europa is further from Jupiter and has a less eccentric orbit. So, after the Voyager encounters, Europa became regarded as the ice-covered equivalent of a less-active version of Io. Certainly this could explain the fracturing and resurfacing evident on Europa's surface, and speculation

abounded as to whether the rate of heat transfer from the rocky part into the base of the ice would be sufficient to maintain an unfrozen ocean sandwiched between the ice and the rock. Essentially, the issue depends on which of the two alternative models in Figure 4.13 is correct. Europa and its possible ocean are the main focus of the bulk of this chapter.

The famous science fiction author Arthur C. Clarke was one of the first to realize the astrobiological implications of a tidally heated Europa, by analogy with communities around 'black smoker' hydrothermal vents on the Earth's ocean floor. In *2010: Odyssey Two* (published in 1982 as a sequel to the more famous *2001: A Space Odyssey*) he imagined an explorer's findings on the floor of the Europan ocean:

> …the first oasis filled him with delighted surprise. It extended for almost a kilometre around a tangled mass of pipes and chimneys deposited from mineral brines gushing from the interior. Out of that natural parody of a Gothic castle, black, scalding liquids pulsed in a slow rhythm, as if driven by the beating of some mighty heart. And, like blood, they were the authentic sign of life.
>
> The boiling fluids drove back the deadly cold leaking down from above, and formed an island of warmth on the seabed. Equally important, they brought from Europa's interior all the chemicals of life. There, in an environment where none had expected it, were energy and food, in abundance…
>
> In the tropical zone close to the contorted walls of the 'castle' were delicate, spidery structures that seemed to be the analogy of plants, though almost all were capable of movement. Crawling among these were bizarre slugs and worms, some feeding on the plants, others obtaining their food directly from the mineral-laden waters around them. At greater distances from the source of heat – the submarine fire around which all the creatures warmed themselves – were sturdier, more robust organisms, not unlike crabs or spiders.
>
> Armies of biologists could have spent lifetimes studying that one small oasis.
>
> (Clarke, 1982)

It took several years for speculations such as Clarke's to become acceptable among mainstream scientists. One reason for this is that ocean-floor hydrothermal vents had not yet been recognized as one of the most likely environments where life on Earth could have originated (see Section 1.10). Another reason is that the Voyager indications of an ocean below Europa's ice were not nearly so compelling as the evidence that has become available subsequently. However, by the late 1990s NASA was presenting Europa's astrobiological potential as the main reason why the US Senate ought to provide funding for a dedicated Europa mission, which at the time of writing has a planned launch date of 2011 at the earliest.

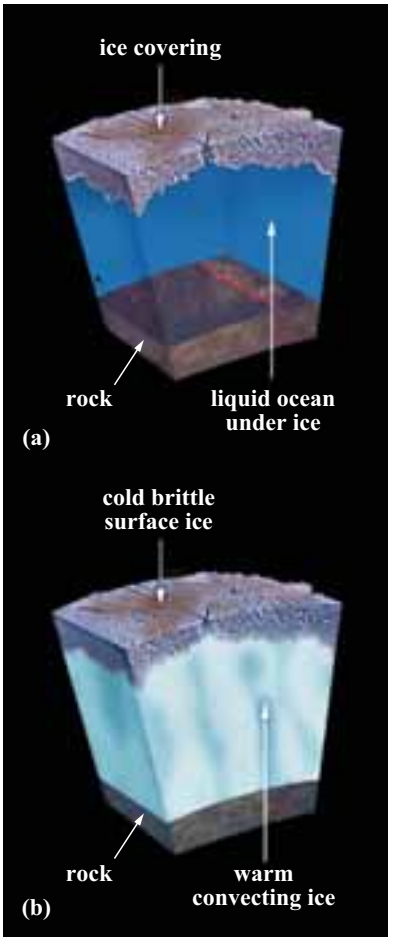


Figure 4.13 Alternative models for the nature of Europa's 'icy' layer: (a) an ocean of liquid water sandwiched between the solid ice and the rocky interior; (b) solid ice, though probably warm enough to be mobile near its base, resting directly on rock.

QUESTION 4.2

Using the orbital radii given in Table 4.1, calculate the tidal force of Jupiter on Europa as a proportion of the tidal force of Jupiter on Io.

4.1.5 The Galileo mission

It was a long time before the Voyager missions were followed up by more detailed surveys of the outer planet satellites. No Uranus or Neptune missions are planned, but a mission to Saturn called Cassini–Huygens was launched in 1997 for arrival at Saturn in 2004. You will learn about this in the next chapter. However, the Jupiter system received a similar visitor first. This was Galileo, launched in 1989, which became the first spacecraft to orbit Jupiter in December 1995. It continued to function through 2002, and was scheduled to be destroyed by plunging into Jupiter's atmosphere in September 2003. This was a planetary protection measure (Chapter 3), taken to avoid the possibility of the defunct craft eventually colliding with Europa and thereby contaminating it with any unintentional bioload.

Galileo had several close encounters with each of the Galilean satellites, providing more complete and more detailed imaging than was possible during the Voyager fly-bys, using an instrument known as the solid-state imaging (SSI) camera. It also carried a near-infrared imaging spectrometer (NIMS), which was useful for determining surface compositions (and also temperatures of Io's active lava flows), an ultraviolet spectrometer, and magnetometers that revealed the satellites' responses as they move through Jupiter's magnetosphere. Perturbations to Galileo's trajectory as it passed close to the satellites placed improved constraints on their internal density distributions, indicating dense, presumably metallic, cores at the centres of Io, Europa and Ganymede. Callisto, by contrast, was proven to be only weakly differentiated, with incomplete segregation of rock and ice (Figure 4.14). See Box 4.3 for a discussion of how terms are borrowed from the terrestrial planets to describe the compositional and mechanical layers within icy satellites.

Near-infrared means the part of the infrared spectrum that is nearest to the visible. The actual spectral range covered by NIMS was 0.7–5.2 μm.

Relying largely on Galileo observations, we will now take a detailed look at Europa, to see what we can deduce about its recent history and the possibility of a life-bearing ocean below the ice.

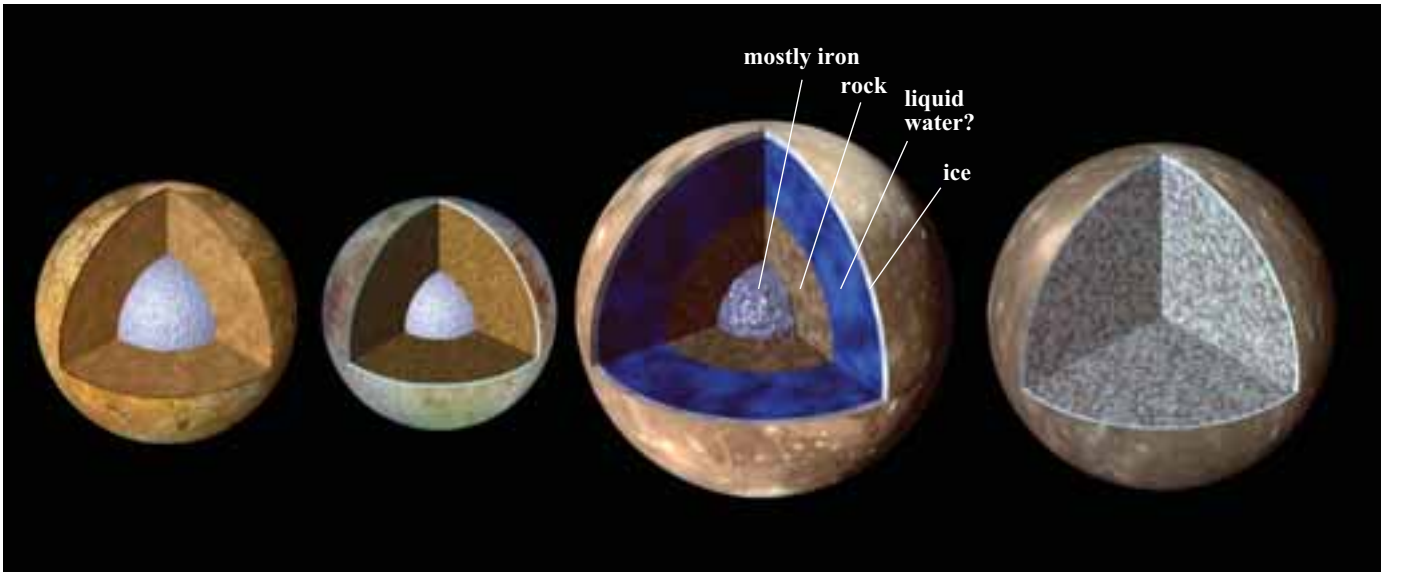


Figure 4.14 Post-Galileo models of the internal structures of the Galilean satellites (from left to right: Io, Europa, Ganymede and Callisto). (At this scale, no distinction is made between ice and liquid water in the model for Europa.)

BOX 4.3 TERMINOLOGY FOR THE LAYERED STRUCTURE OF DIFFERENTIATED ICY BODIES

In a differentiated terrestrial planet, the term *core* is used for the dense compositionally distinct inner part, which is rich in iron. This is surrounded by a rocky (silicate) *mantle*. The extreme outer part of the rocky material is referred to as the *crust* if its composition has been altered by volcanism and other recycling processes.

In a differentiated icy body, it is logical, by analogy, to regard the rocky interior as the core (and if this is itself differentiated, with an iron-rich centre, to call that the *inner core*). The icy outer part of such a body is thus the mantle, and if the outer part of the ice differs somewhat in composition from the interior, we can call this the crust. The analogy is particularly apt because Solar System ices share many important properties with silicate rock. Among these are:

1 At the prevailing near-surface temperatures, the ice is mechanically strong and rigid, like rock near the Earth's surface.

2 When caused to melt (by heat, or in some cases a decrease in pressure) ice that is water mixed with salt or water mixed with another volatile species undergoes partial melting, just like a mixture of silicate minerals in rock. The melt and the surviving crystals have different compositions, and melting begins at a lower temperature than for pure water-ice.

3 At sufficient temperatures and pressure, ice will flow without melting, and can undergo solid-state convection, just like rock in the deeper part of the Earth's mantle.

Property 2 in this list makes it likely that the outermost part of a differentiated icy body does indeed differ at least slightly in composition from its mantle, and so we can regard this differentiated ice as a true crust. Property 3 means that we can discriminate the outer rigid ice (upper mantle plus crust) from the deeper more mobile (even though solid) ice, and distinguish these by the terms *lithosphere* and *asthenosphere*, respectively, which were originally coined for the Earth.

4.2 Europa

Europa's surface is fascinating, if often perplexing, to study. One of its special characteristics is its brightness. It has an albedo of 0.7, which is exceeded among icy satellites only by Enceladus and Triton. Overall brightness is one indicator of the youth of an icy surface: the brighter the icy surface, the younger it is. Ganymede (albedo 0.45) and Callisto (albedo 0.2) are much darker. This distinction is not usually apparent when comparing images of their surfaces (for example Figures 4.10a and 4.11a, Callisto and Europa respectively), because the brightness of each image has usually been adjusted to show features on each to best advantage.

The 'albedo' of a body is simply the fraction of the incident light that is reflected. The higher the albedo, the more light is reflected, and the brighter the body appears.

The midday temperature is about 130 K (about −140 °C) at Europa's equator and about 80 K (about −190 °C) at the poles. Europa's axis of rotation is perpendicular to the plane of its orbit, which is tilted at less than half a degree relative to Jupiter's equatorial plane. Europa experiences virtually no 'seasonal' changes in illumination during its orbit about Jupiter or during Jupiter's twelve-year orbit of the Sun, because Jupiter's axial inclination is only about 3° (so Jupiter itself virtually lacks seasons too).

Galileo detected a magnetic field about Europa, which could be generated by motion within its iron core or within a salty (and therefore electrically conducting) ocean beneath the ice. The highest-resolution images of Europa sent back by Galileo have pixels representing areas about 6 m across. Such detailed images cover only a small fraction of the total surface. 9% of Europa was imaged at better than 200 m per pixel and about half the globe was imaged at better than 1 km per pixel. This was a great advance on the coverage provided by Voyager, whose best images of Europa have pixels representing areas 1.9 km across. You will examine plenty of images of Europa shortly, but first it is worth establishing what we know about the composition of the ice.

4.2.1 Ice and salt

As noted previously, Europa's near-infrared reflectance spectrum was used as long ago as the 1950s to demonstrate that its surface is mostly water-ice. More recently, spectroscopic observations by the Hubble Space Telescope and Galileo have revealed some regions where the ice appears to be salty (see below) and have also detected traces of molecular oxygen (O_2) and smaller amounts of ozone (O_3). The oxygen and ozone almost certainly result from the breakdown of water molecules in the ice brought about by exposure to charged particles (this process is known as **radiolysis**) that are channelled onto Europa by Jupiter's magnetic field, and by solar ultraviolet radiation (a process called photodissociation or **photolysis**; Section 1.6.2). Most of the oxygen and ozone is probably held within the ice (as isolated molecules trapped within ice crystals), but some may constitute an extremely tenuous atmosphere.

■ Apart from various forms of oxygen, what else would you expect to be produced when water molecules are broken down by radiation?

❑ Given that the formula for water is H_2O, hydrogen should also be produced.

Box 4.4 shows a series of reactions that could produce oxygen and hydrogen in Europa's ice.

BOX 4.4 RADIOLYTIC AND PHOTOLYTIC BREAKDOWN OF WATER MOLECULES IN ICE

The reactions that occur to generate oxygen and hydrogen within the surface ice of an icy satellite can be summarized, in simplified form, as:

$$H_2O \rightarrow H + OH$$

$$H + H \rightarrow H_2$$

$$OH + OH \rightarrow H_2O_2$$

$$2H_2O_2 \rightarrow 2H_2O + O_2$$

Europa's ozone is likely to be the product of a chain of reactions involving radiolytic and photolytic breakdown and recombination of oxygen molecules, similar to the photolytically driven reactions that generate ozone from oxygen in the Earth's stratosphere.

Hydrogen has not yet been detected on Europa, but on Ganymede, where similar 'space weathering' of exposed ice occurs, hydrogen has been found leaking away into space.

■ Suggest a simple explanation to explain why there is a lot less free hydrogen than oxygen in or above Europa's surface.

❑ Hydrogen is a much smaller and lighter atom therefore it is easier for hydrogen to escape from within the ice. Once liberated, it is so loosely bound by Europa's weak gravity that it would be lost to space much faster than oxygen or ozone.

Hydrogen peroxide (H_2O_2), which is an intermediate product of the sequence of reactions in Box 4.4, has been identified as a trace component of the ice in reflectance spectra obtained using Galileo's near-infrared imaging spectrometer. The same instrument has also revealed distortion of the absorption bands associated with water. This indicates that, in addition to forming ice crystals, some of the water molecules are bound within hydrated salt crystals. The best match to the spectra is from a mixture of magnesium and sodium salts such as magnesium sulfate hexahydrate ($MgSO_4.6H_2O$), epsomite ($MgSO_4.7H_2O$), bloedite ($MgSO_4.Na_2SO_4.4H_2O$) and natron ($Na_2CO_3.10H_2O$). The occurrence of sulfates is supported by Galileo ultraviolet spectroscopic data that indicate the presence of compounds containing a sulfur–oxygen bond.

■ Although carbonates and sulfates are fairly common salts on Earth, they are not the most abundant. What kind of salt appears to be missing on Europa, compared with the Earth?

❑ No chlorides are in the above list – note that sodium chloride (NaCl), which is the most abundant salt dissolved in the Earth's oceans, is absent.

Actually, chlorides produce no spectral features in the available part of the spectrum, so direct observational data cannot tell us whether any chlorine salts occur on Europa's surface. What the spectral mapping by Galileo did achieve, however, was to show that the distribution of salts across Europa's surface is highly non-uniform. Large expanses are relatively salt-free, but in places where the surface has been most recently and most greatly disrupted from below, the surface salt concentration reaches 99%. You will see what these areas look like shortly.

The salts occurring on Europa's surface are unlikely to be a straightforward representation of those dissolved in any ocean beneath Europa's ice – calculations have shown that the freezing process would tend to concentrate sulfates of magnesium and sodium into the ice. This is consistent with the observed preponderance of these salts at the surface. However, the concentrations of elements dissolved in Europa's ocean are largely a matter of speculation. Two of the factors that have to be considered are the composition of Europa's rocky component, and the efficiency with which each element becomes dissolved from it into the ocean. Neither of these factors is known. Although, on average, Europa's rock is likely to be similar to carbonaceous chondrites, geochemical differentiation could mean that the rock nearest to the ice–rock interface might well be very

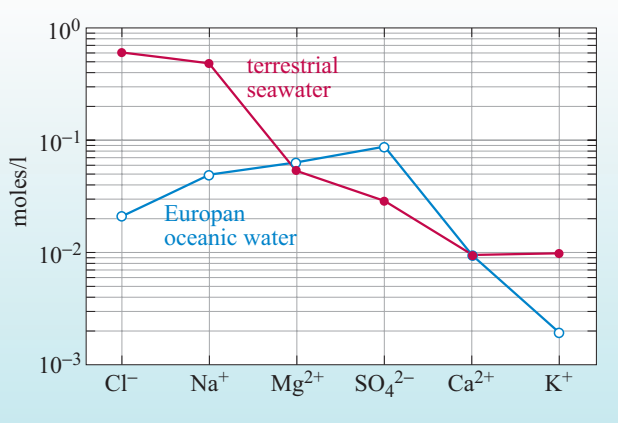


Figure 4.15 Estimated concentrations of major elements in Europan oceanic water compared with seawater on Earth.

different (as is probably the case in Io's crust, for example). The efficiency with which elements become dissolved (or sometimes reprecipitated) depends upon the temperature at which it occurs, as well as on the overall chemistry of the solution. Despite the uncertainties, attempts have been made to model the likely concentrations of dissolved elements in Europa's ocean. The results of one such model are shown in Figure 4.15.

QUESTION 4.3

According to Figure 4.15, how many more times greater is the concentration of chloride (Cl^-) in terrestrial seawater than in Europa's ocean?

4.2.2 Examining Europa's surface

It is all very well speculating about conditions in an ocean below Europa's ice, but what evidence is there that it actually exists? After all, tidal heating might not result in ice melting on a global scale, and current geophysical models of Europa's internal structure (e.g. Figure 4.14) cannot tell the difference between ice and liquid water. Fortunately, Voyager and Galileo have given us detailed images of Europa that we can use in the same way that a geologist uses aerial photographs or images from space to help decipher the processes that have shaped a particular tract of the Earth's surface.

The general view

Figure 4.16 shows a Voyager 2 image of a large region of Europa. Examine this image carefully, in order to answer Question 4.4.

BOX 4.5 LATITUDES AND LONGITUDES ON SATELLITES

As soon as the first features were discovered on the surfaces of the satellites of the outer planets, it became necessary to define co-ordinate systems to map their locations. Latitude is simple to define; it is measured in degrees north and south of the equator, which lies halfway between the satellite's poles of rotation. By convention (established by the International Astronomical Union), for a synchronously rotating satellite 0° longitude is defined to run through the centre of the planet-facing hemisphere. Longitude is normally quoted in degrees measured westwards from here, and west is always to the left when you look at a body with north towards the top.

QUESTION 4.4

(a) Study Figure 4.16 and write a short description of the kinds of features you can see on Europa's surface, noting for example relative brightness and characteristic shapes or textures. Concentrate on a simple description of appearance; we do not expect you to explain the origin or precise nature of what you can see.

(b) Try to deduce the relative ages of the features you have described.

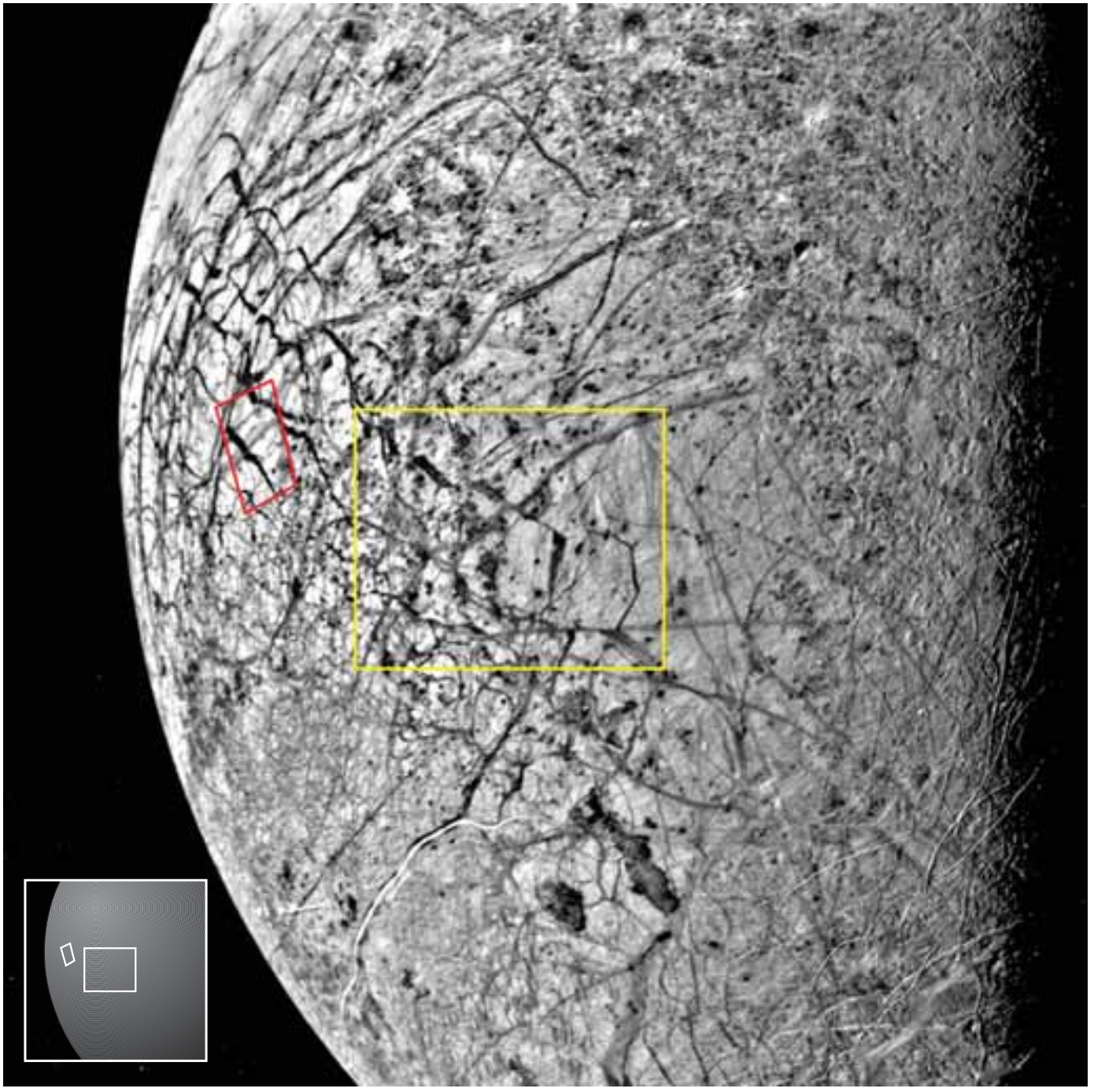
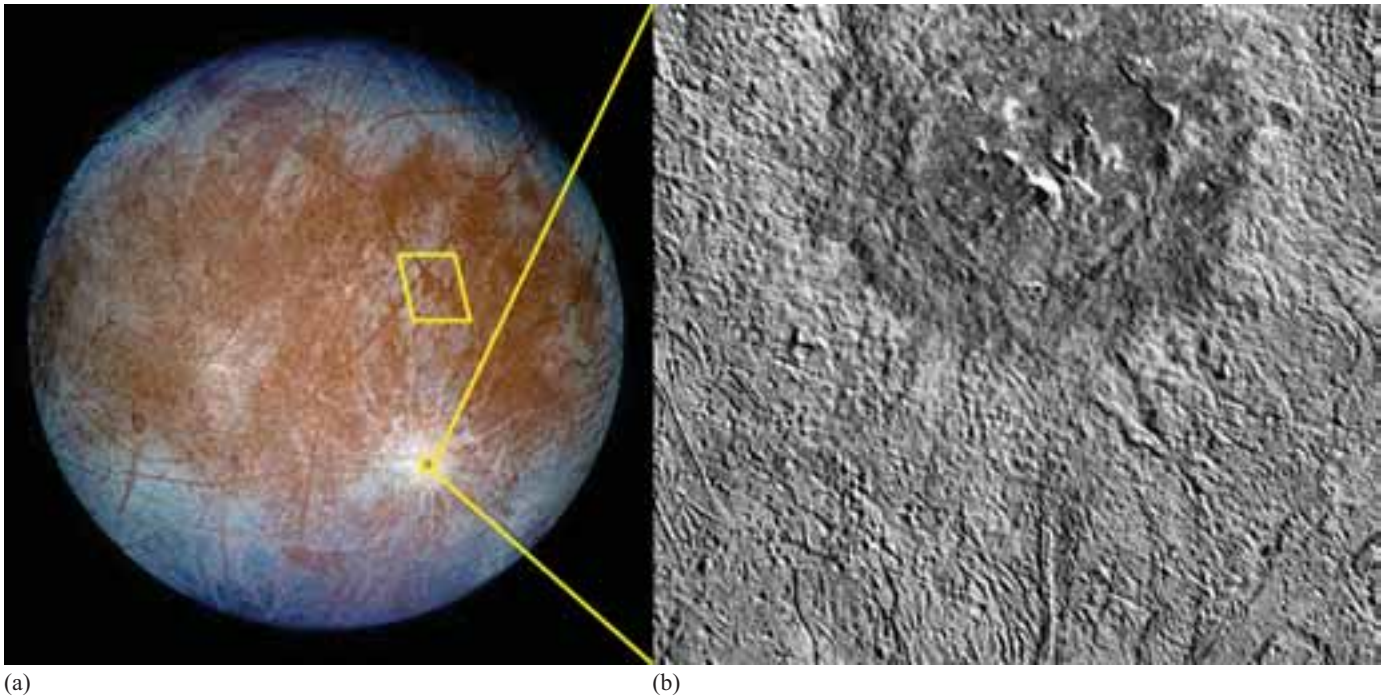


Figure 4.16 Voyager 2 image showing a region of Europa, about 3000 km across, recorded at about 2 km per pixel and centred at 10° S, 160° W (see Box 4.5 for an explanation of latitudes and longitudes on satellites). The large yellow outline shows the area covered by Figure 4.18a and the smaller red outline shows the area covered by Figure 4.19a.

In answering Question 4.4, you should have formed an impression of an original surface that (at the scale of the image) appears relatively featureless, but was subsequently cut across by processes that produced dark bands. Later, the band-disrupted terrain was itself overprinted in places to produce mottled terrain and curved ridges. The dark bands cutting across much of Europa give it the appearance of a thoroughly cracked eggshell, but please be aware that there is no evidence in Figure 4.16 (nor on any more detailed images) that these 'cracks' are open fissures in the surface. In fact, there is very little topographic relief on Europa. The curved ridges in the lower right corner of Figure 4.16 are only about 200 m high.

The crater Pwyll

You might also have noted that there are no obvious impact craters visible in Figure 4.16. In fact there are a few. One is a bright spot, 15 km in diameter, surrounded by a dark halo of ejecta that occurs 10 mm from the top edge and 65 mm from the left-hand edge of the figure. Another is a slightly larger pale feature with a discernible central peak 20 mm from the top edge and 45 mm from the right-hand edge. The youngest large crater on Europa occurs at 26° S, 271° W, which is outside the area covered by Figure 4.16. This is shown in Figure 4.17, and is named Pwyll (pronounced 'Puh-hl' or 'Poo-eel', after a character from Welsh legend, Box 4.6). Pwyll is 26 km across, and has a dark floor and a halo of equally dark ejecta extending for about 8 km beyond its rim, which was presumably excavated from below the surface. Much brighter, finely fragmented ejecta in the form of discontinuous rays can be traced for more than 1000 km, and forms the bright region surrounding the crater in the global view in Figure 4.17a. It is the high visibility of its ejecta rays that shows Pwyll to be the youngest of Europa's large craters. Statistical arguments based on the likely frequency of comet impacts onto Europa suggest that Pwyll is very unlikely to be older than about 20 million years, and is probably about 3 million years old.



(a) (b)

Figure 4.17 (a) Global view of Europa showing the location of the crater Pwyll, which is shown enlarged in (b), a Galileo SSI image recorded at 250 m per pixel. The outline superimposed on (a) indicates the area covered by Figure 4.21.

BOX 4.6 NAMES ON EUROPA AND OTHER SATELLITES

In order to avoid duplication of the names of features between bodies, and to try to achieve consistency of nomenclature on each body, the International Astronomical Union has established a naming convention for each planetary body in the Solar System. Names on Europa are drawn from Celtic gods, heroes, and myths; people and places associated with the Europa myth; and place names from ancient Egypt.

The crater on Europa called Pwyll is a character from Welsh legend, who appears in the mediaeval collection of tales known as *The Mabinogion*.

By contrast, features on Io are named after gods and heroes associated with fire, sun, thunder and volcanoes, and also people and places associated with the Io myth and Dante's *Inferno*.

Incidentally, the Galilean satellites themselves and many of Jupiter's smaller satellites are named after mythological characters (of various genders and species) who, to put it delicately, became 'romantically entangled' with the god Jupiter.

QUESTION 4.5

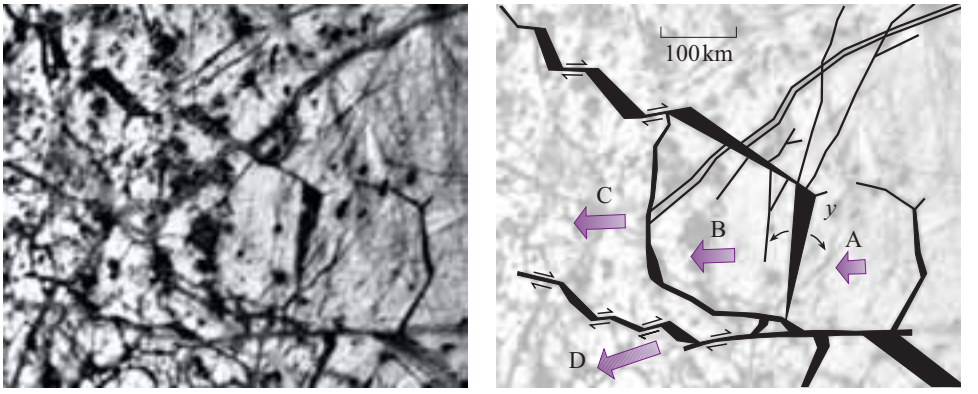
Look at the detailed image of Pwyll in Figure 4.17b. Does Pwyll have the three-dimensional shape that you would expect of a young crater?

Expert analysis shows that most of Pwyll's rim is less than 200 m high, and that (unusually for impact craters) its floor is hardly any lower than the terrain outside. Opinion is divided as to whether the impactor responsible for Pwyll actually penetrated right through the ice, but all are agreed that the crater shows the hallmarks of an impact into relatively thin (about 20 km in thickness) and weak ice.

Thus the paucity of large craters on Europa indicates that its surface is young, and the subdued cross-sectional shape of many of those craters that do occur suggests that the ice was relatively thin when they formed.

Fracturing and motion of the ice shell

If the rigid surface layer of Europa's ice is thin (or, at least, has been thin for some of the time), and overlies either water or some kind of weak and mushy ice as indicated by large craters such as Pwyll, then we might expect to find some evidence for fracturing and motion of the rigid ice shell. This is precisely what the pattern of dark bands such as those on Figure 4.16 appears to be showing us. An area from Figure 4.16 is enlarged in Figure 4.18, with an interpretation of how plates bounded by fractures in the rigid ice shell could have moved relative to one another.



(a) (b)

Figure 4.18 (a) An enlargement of the area that was outlined in yellow on Figure 4.16. The youngest dark bands (many of them wedge-shaped) can be seen to cross-cut and offset some of the older bands. (b) Sketch of the area covered by (a) showing how opening of the bands is consistent with shuffling and rotating neighbouring plates of ice. See text for discussion. The map has been simplified by omitting various younger blotches that appear on (a).

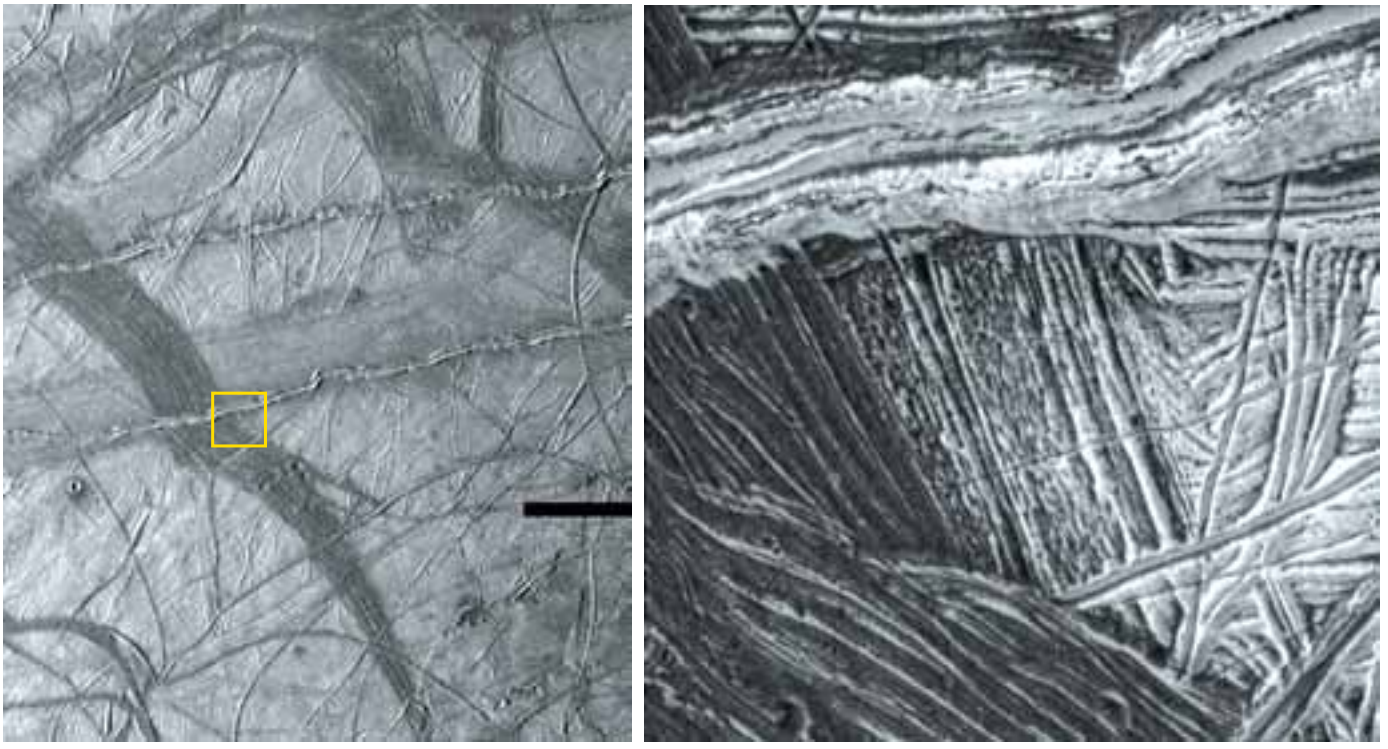
The arrows on Figure 4.18b suggest that the plates labelled A–D have all moved westwards relative to the ice at the right-hand (eastern) edge of the map. In addition, plate B has rotated about 5° anticlockwise relative to plate A (opening up the intervening wedge-shaped band that extends south from *y*); plate C has moved west relative to plate B and plate D has moved west relative to plate C.

It is tempting to make an analogy with plate tectonics on Earth, and to regard the stepped dark bands forming the north and south boundaries of plate C as lengths of spreading axis (or mid-ocean ridge) offset by transform faults. However, even if the interpretation in Figure 4.18b is correct, there are several important differences between plate tectonics on Europa and the Earth. First, Europa's jumble of overlapping dark bands (Figure 4.16) suggests that old spreading axes are abandoned and replaced by new ones after only a few tens of kilometres of spreading. However, on Earth most spreading axes last for tens to hundreds of millions of years, during which time they add hundreds or even thousands of kilometres of new lithosphere to the edges of the adjacent plates. On Earth, creation of new lithosphere at spreading axes is balanced globally by destruction of lithosphere at subduction zones.

On Earth, a subduction zone is where one lithospheric plate descends at an angle below another.

There is no analogue to terrestrial subduction zones on Europa, but it is obvious that if new areas of surface ice are being added to make the dark bands then other areas must be being destroyed at an equal rate. The processes operating on Europa to achieve such a balance remained a mystery until Galileo's more detailed images became available. You will examine this evidence soon, but first it is worth exploring the extra information that Galileo images can give about the dark bands themselves. Figure 4.19a is one such image. It shows that the pale areas between the dark bands that seemed relatively featureless at the resolution of the Voyager images can be seen at higher resolution to be criss-crossed by low ridges. At this level of detail, Europa's surface has been aptly described as looking like a ball of string. Furthermore, the 'ball of string' ridges also occur within the dark bands (running parallel with their edges). When we move up to even higher resolution, as in Figure 4.19b, the 'ball of string' ridges are even more obvious (and some can be seen to have central grooves running along them), whereas the distinction between dark bands and pale terrain has become hard to see.

It is uncertain exactly how the ridges on Europa have been built. Each is probably the result of some form of cryovolcanic eruption along a crack or fissure. If this is the case, the material erupted must have been in the form of mushy ice, or perhaps



(a) (b)

Figure 4.19 Galileo SSI images of part of Europa, near 16° S, 195° W. (a) A region, 150 km in width, recorded at 420 m per pixel. This is the area shown by the red outline on Figure 4.16 (whose shape is distorted in Figure 4.16 because of perspective). A prominent wedge-shaped dark band runs diagonally across the lower left of the image. It is cut by two narrow bright bands. The black bar at the right-hand side represents missing data. The outline indicates the area shown at higher resolution in (b). (b) Image, 20 km in width, recorded at 26 m per pixel. Apart from the bright band, which is the youngest feature shown on the image and overlies everything else, the whole surface (dark band and pale terrain alike) is seen to consist of a succession of cross-cutting 'ball of string' ridges.

a fountain-like spray of fragmented ice, analogous to a volcanic fissure eruption on Earth (Figure 4.20) and involving the escape of gaseous volatiles during eruption. Fortunately, the details of ridge-building are not important in order to understand the general surface history and its implications for ice thickness, which appear to be as follows:

- Each 'ball of string' ridge is symptomatic of a small amount of surface extension.
- The ridges occur in sets of up to about a dozen parallel ridges, and each set can usually be seen to be cut across by a younger set. There are at least four such sets within the portion of the dark band shown in Figure 4.19b. Although not quite parallel to each other, each set runs lengthways relative to the dark band, and would in total be responsible for the kind of spreading across a dark band indicated in Figure 4.18b.
- In the older pale terrain outside the dark band the ridge sets are oriented more variably, showing a long and complex history of surface creation.
- The dark bands are the youngest parts of the 'ball of string' surface, and evidently become paler as they age. (There are many ways in which this could happen. Some involve growth or fragmentation of ice crystals over time, others depend on chemical changes caused by long-term exposure to radiation.)

Figure 4.20 A basaltic fissure eruption on Earth. The rampart is built by congealed lava on either side of the fissure. This is a possible analogue for how the 'ball of string' ridges on Europa are constructed. Human figure in left foreground indicates the scale.

There are two things to add to finish the story of surface creation in the area covered by Figure 4.19. First, the bright bands cut across the 'ball of string' texture and so are clearly younger than it. These bands may be a slightly different kind of cryovolcanic feature – their feathery edges, seen at the highest resolution (Figure 4.19b), could represent debris shed downslope from a central high. Second, there are some very narrow grooves (barely visible in Figure 4.19b) that also cut both dark and pale 'ball of string' texture, one of which widens towards the east where it becomes an otherwise unremarkable contributor to the texture. Many features such as these are probably cracks where extension occurred without an accompanying eruption. Others are evidently the surface expressions of faults with sideways (instead of extensional) movement (as you will see shortly).

The dark bands and the intervening tracts of pale terrain were constructed by a long and complicated series of events, each of which was associated with spreading on a local scale.

More surface disruption

Now let's examine some detailed images of the region of Europa's northern hemisphere that was indicated on Figure 4.17. A medium resolution image is shown in Figure 4.21, and higher resolution images from within this area are shown in Figures 4.22–4.24.

QUESTION 4.6

Study Figure 4.21. How would you classify the majority of the surface in this region, including the part of it shown in more detail in Figure 4.22?

QUESTION 4.7

Look at the two features labelled A and B in Figure 4.22. A is a groove with a slightly raised rim on either side, running diagonally down to the right from the top of the image. B is a ridge, with a groove down its centre, running almost directly down from the top. Try to account for the relationship between these two features where they cross, and deduce which of these two features is the younger.

These images provide clear evidence that tectonics on Europa involve relative sideways movements as well as simple spreading apart of the surface.

If you are familiar with plate tectonics on Earth you will probably not be surprised by this. Now turn your attention to the other parts of Figure 4.21, notably those covered by Figures 4.23 and 4.24.

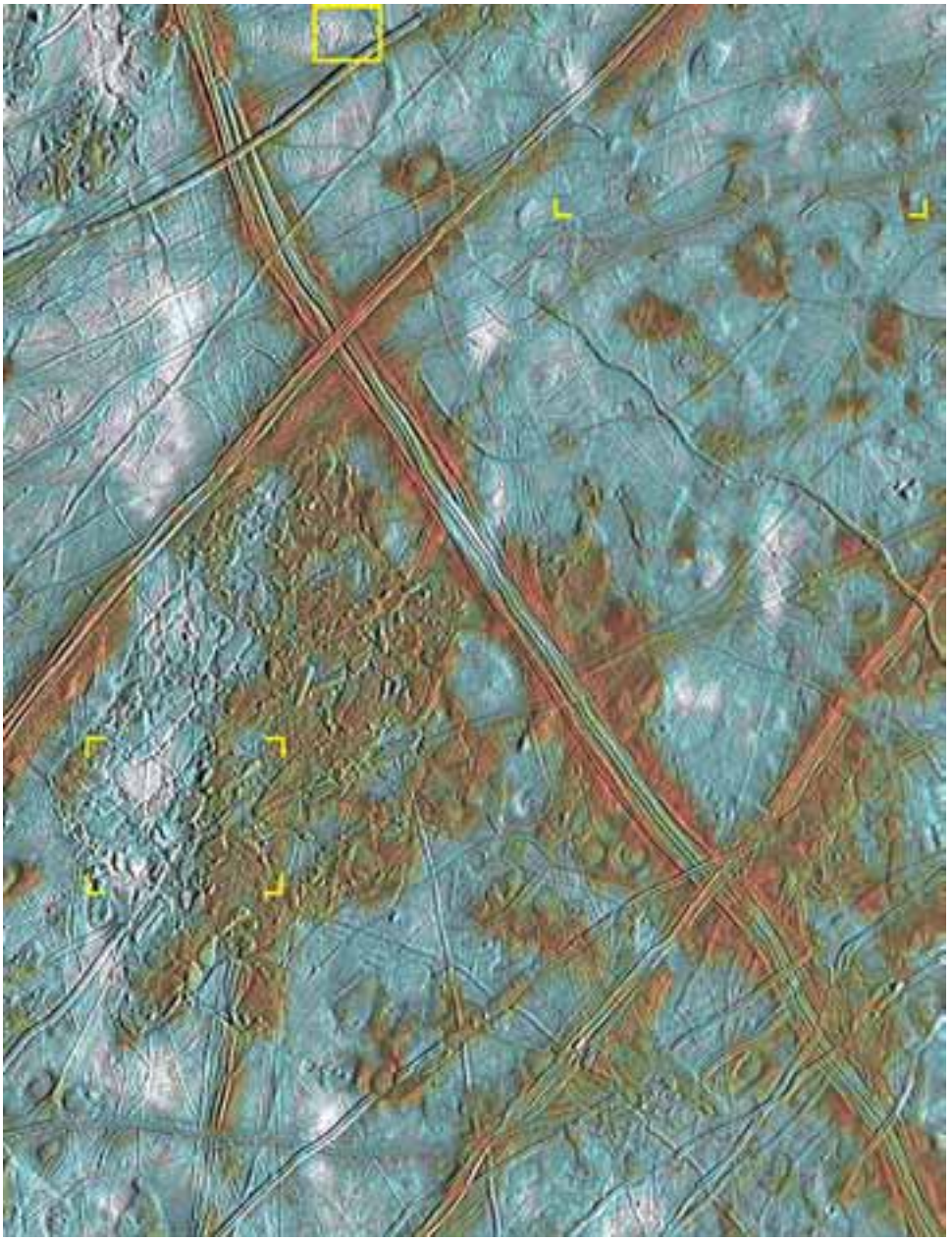


Figure 4.21 Galileo SSI image of part of Europa, 200 km in width, centred at 10° N, 270° W and recorded at 180 m per pixel. This image was made by combining near-infrared, green and violet images in red, green and blue, respectively, and then exaggerating the resulting colours. In this rendering, blue is the general colour of the icy surface, and ice-poor (probably salty) areas show up red. The white patches are thin sprinklings of ejecta from the crater Pwyll, which is 1000 km to the south. The yellow outline locates Figure 4.22. Other yellow marks indicate the lower corners of Figure 4.23 and all four corners of Figure 4.24.

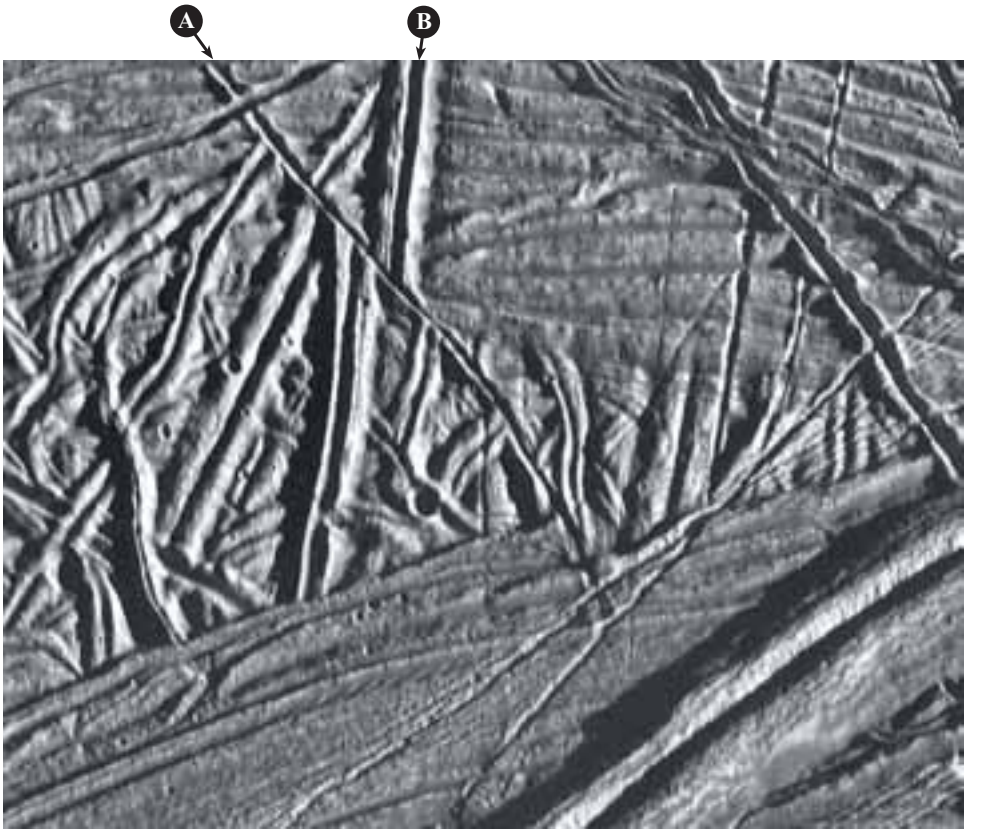


Figure 4.22 Galileo SSI image, 15 km in width, recorded at 20 m per pixel. Solar illumination is from the right. See Figure 4.21 for location. Letters A and B are referred to in Question 4.7.

QUESTION 4.8

What has happened to the 'ball of string' texture in (a) Figure 4.23, and (b) Figure 4.24?

The usual explanation for the dome features in regions such as Figure 4.23 is that they are places where warm buoyant material (which could be warm ice, slush or water) has risen from below. Where injected as an intrusion at shallow depth, the result is a subtle dome (e.g. the example in D4) over which the surface may have been sufficiently stretched to rupture. In more extreme cases the 'ball of string' surface appears to have melted, exposing the risen material, which is surrounded by cliffs that drop down to the new surface (e.g. the example in D/E–5/6). Elsewhere the risen material seems to have spread out across the top of the old 'ball of string' surface (e.g. the example in D/E–1/2). Regions such as the one covering Figure 4.24 are essentially just more extensive versions of the D/E–5/6 situation, and demonstrate the effects of heating events on a regional scale. There are several examples of this type of terrain on Europa, which is described formally as **chaos**. Conventionally, within a chaos region, the slabs of 'ball of string' surface are referred to as 'rafts' and the low-lying hummocky material in between is called 'matrix'.

The matrix is most simply interpreted as the (now refrozen) surface of an ocean that was exposed when the overlying ice was removed, presumably by melting caused by an injection of heat from below. Near the edges of chaos, rafts have broken away from the continuous ice sheet and drifted inwards by relatively small

Figure 4.23 Galileo SSI image, 80 km in width, recorded at 54 m per pixel. Solar illumination is from the right. See Figure 4.21 for location. The letters and numbers around the edge define 10 km × 10 km squares for reference in the answer to Question 4.8.

distances, and in many cases their original configurations can be deduced. Near the centres of chaos, rafts are less abundant and it is not usually possible to see how they once fitted together. The rafts are analogous to ice floes formed in the Earth's oceans when floating pack-ice breaks up in the spring. The even height of the cliff at the edge of most rafts shows that these rafts are lying horizontally. However, there is one raft, 5 km × 2 km in size, just to the northeast of the centre of Figure 4.24 with (to judge by the cliff's shadow) an exceptionally high cliff on its northwest side but no sign of a cliff on its southeast side. This raft is shown enlarged in Figure 4.25. It looks as though the raft has been tilted down towards the southeast. Some of the knobbly hills sticking up out of the matrix may be the corners of smaller or more steeply tilted rafts, the most obvious being a triangular hill with an exceptionally long shadow immediately to the south of the tilted raft in Figure 4.25.

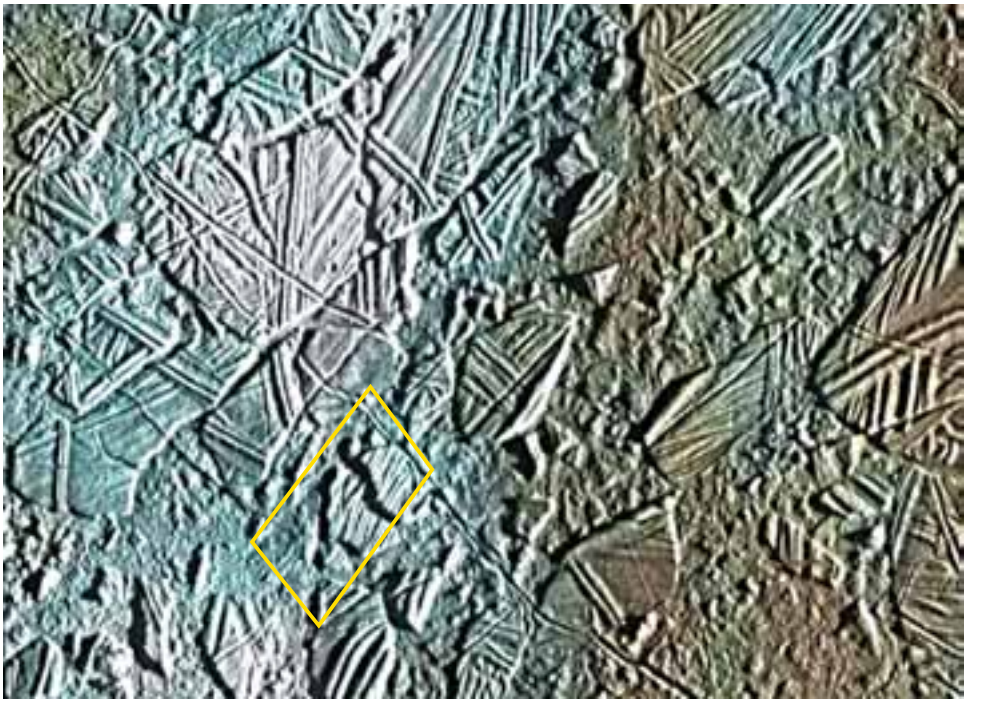


Figure 4.24 Galileo SSI image, 45 km in width, recorded at 54 m per pixel. Colours are constructed in the same way as in Figure 4.21. Solar illumination is from the right. See Figure 4.21 for location. The outline shows the area shown at very high resolution in Figure 4.26.

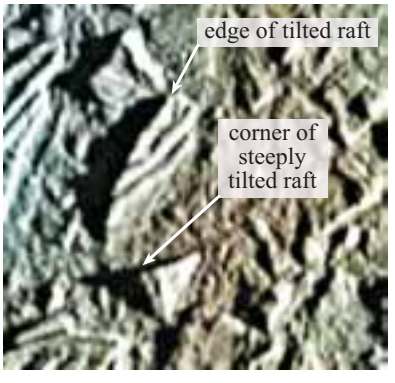


Figure 4.25 Enlargement of an area 8 km in width just northeast of the centre of Figure 4.24, showing a gently tilted raft and the corner of a more steeply tilted raft.

Chaos makes up about a third of the 9% of Europa's surface that has been imaged at adequate resolution (less than about 200 m per pixel) to make identification certain. On low-resolution images, both chaos and dome-disrupted regions such as Figure 4.23 appear as mottled terrain like the area in the northeast of Figure 4.16. The region covered by Figure 4.24 is named Conamara Chaos, and as you can see on Figure 4.21 it extends for about 100 km north to south and about 80 km east to west. The largest chaos region on Europa is more than a thousand kilometres across, and the small end of the size spectrum is exemplified by the resurfaced area at D/E–5/6 in Figure 4.23.

- If large chaos regions are places where 'ball of string' surface has been destroyed, what could be their significance for the global tectonics of Europa?
- Destruction of surface in chaos regions could balance the spreading implied by the creation of the 'ball-of-string' texture. We noted earlier that such spreading could not occur unless it was matched globally by the surface being destroyed at an equal rate. On Earth, this is achieved where plates are subducted deep into the mantle.

It is hard to imagine how we could prove that formation of chaos in one part of Europa is accompanied simultaneously by addition of new ridges and grooves to 'ball of string' textured regions elsewhere, unless we actually could see it happening. However, recognition of chaos does at least show how a balance could be achieved between the creation and destruction of surface.

So how old is Conamara Chaos?

Conamara Chaos is named after a region in the west of Galway, Ireland (usually spelt Connemara in English). The name derives from *Conmaicne mara*, meaning the seaside land of the descendants of Conmac. In Irish legend, Conmac was a son of Fergus Mòr, king of Ulster and Maedhbh, queen of Connacht.

■ Look carefully at Figure 4.24. Is the white ejecta from Pwyll visible on top of the matrix as well as on top of the rafts and, if so, what does this tell us about the relative ages of the chaos and Pwyll?

❑ Both rafts and matrix in the western part of Figure 4.24 are white, in contrast to redder surfaces in the east. This is because raft and matrix surfaces alike have been overlain by a sprinkling of ray ejecta from Pwyll. This is also apparent on Figure 4.21. The fact that the matrix has ejecta on top of it means that the chaos existed before Pwyll was formed. As noted earlier in this section, Pwyll is probably about 3 million years old, so this is the likely lower age limit for Conamara Chaos.

In addition to the white ray ejecta, there are quite a few craters less than 1 km in diameter in this region. These are more common within the rays, and so are almost certainly secondary craters produced by impact of the largest blocks of ejecta expelled from Pwyll. On Figure 4.24, these craters appear more common on the rafts than in the matrix. This difference could be apparent rather than real, because the jumbled surface texture of the matrix would make the craters difficult to see. However, even on the highest resolution images such as Figure 4.26 craters appear scarcer on the matrix. Some of the small craters on the rafts must pre-date the break-up into chaos, but if most of the small craters we see are Pwyll secondaries then some of those that formed on the matrix since its creation would seem to have been erased. One way this could have happened is if the matrix remained mobile and continued to deform for some millions of years *after* its surface froze, whereas the surfaces of the rafts were rigid for the whole time.

So, the evidence so far points to Conamara Chaos having formed (probably at least 3 million years ago) by melting of a patch of ice some tens of kilometres across, accompanied by break-up of the adjacent floating ice into rafts, some of which drifted inwards across the temporarily unroofed ocean. A skin of ice or slush would have rapidly covered any exposed water but, depending on the amount of heating from below, this skin could have remained sufficiently pliant to allow the rafts to plough through it for up to several million years. Before the matrix became fully rigid, ejecta from the Pwyll impact was distributed across the area, with the accompanying formation of secondary impact craters. Continuing deformation or local resurfacing of the matrix was sufficient to erase a significant proportion of the Pwyll secondary craters on the matrix.

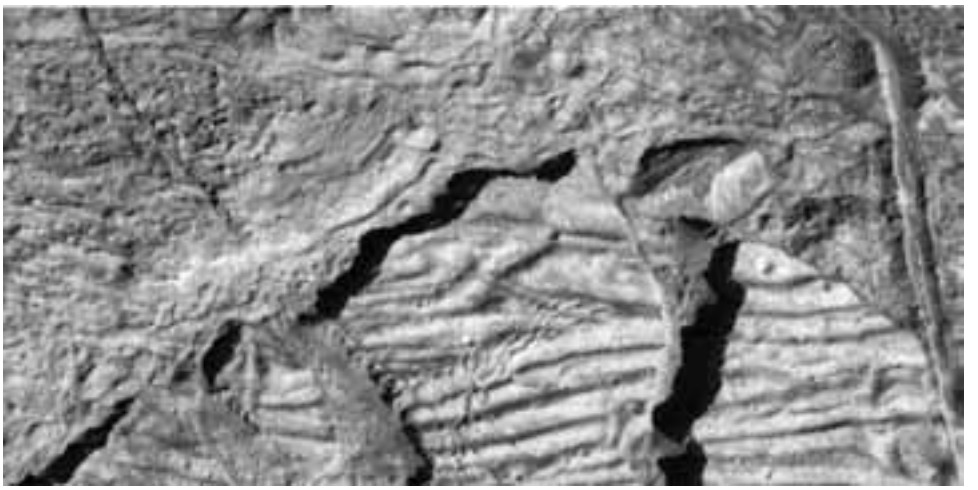


Figure 4.26 Galileo SSI image, 7 km in width, recorded at 9 m per pixel. Solar illumination is from the lower right. See Figure 4.24 for location.

But that is not quite the end of the story.

QUESTION 4.9

Examine Figure 4.24 again, and locate a groove that runs diagonally across the image from just below the northwest (top left) corner. Look closely at the units this groove cuts. Deduce the implications for the age of this groove and the nature of the material that it cuts.

You have now seen the last major piece of the puzzle. After its matrix has become sufficiently rigid, chaos on Europa begins to experience brittle fracturing, and new grooves form that look similar to some of those on ordinary 'ball of string' terrain. Perhaps, given sufficient time, rafts and matrix alike in a chaos region will become thoroughly overprinted by additional generations of ridges and grooves, and the rafts so split up by successive spreading increments across each crack that they lose their integrity. The entire area will take on the appearance of a ball of string – in fact, it will actually be 'ball of string' terrain. For all we know, areas such as those in Figures 4.19 and 4.22 could be former chaos of which no recognizable traces remain.

Chaos areas, and in particular the drifted rafts, are compelling evidence that, at least at the time of chaos formation, the 'ball of string' textured surface ice was floating on a liquid. This would not have to be an ocean of global extent, because the underlying liquid would not need to cover an area much wider than the overlying chaos.

In Section 4.2.4, we will argue that whether the ocean is global, local, permanent or ephemeral is of no great importance for the existence of life. However, first let's see if we can work out how thick the ice is.

4.2.3 How thick is Europa's ice?

You learned in Section 4.1.4 that geophysical data show the 'icy' outer part of Europa to be about 100 km thick, but that the information is inadequate to distinguish between the extreme possibilities of solid ice all the way down to the bedrock and a floating sheet of ice supported above a liquid ocean (Figure 4.13). The subdued topography of craters such as Pwyll and our interpretation of chaos regions both strongly suggest that the latter is more likely. We can determine how thick the ice was at the time of chaos formation, provided we are willing to take the present surface of the matrix to be at roughly the same height (relative to raft surfaces) as the surface of the ocean when it was exposed. The lack of any obvious disturbance of the matrix adjacent to the blocks even in the very high-resolution image in Figure 4.26 indicates that this is a reasonable assumption. If this is correct, then the height of a raft surface above the matrix carries important information.

- Look at the rafts in Figure 4.24. Do you get the impression that each raft has its surface at a different height above the matrix?

❑ With the exception of the tilted rafts, they all appear to be at about the same height.

This is just a crude visual impression. However, there are various ways to determine relative heights on spacecraft images. The best way is to use the stereoscopic information contained in two images of the same area taken from different perspectives. Unfortunately, Galileo did not obtain high-resolution stereoscopic images of Europa. Instead, we can measure the widths of the shadows cast by the rafts onto the matrix, and combine this information with knowledge of the angle of the Sun above the local horizon to estimate the height of the cliff. This shows that most of the cliffs at the edges of rafts in Figure 4.24 are about 100 m high.

■ Why would the surface of a raft (or the top of any object floating in a fluid) be higher than the surface of the matrix (or the fluid in which the object is floating)?

❑ The only simple explanation is that the rafts are less dense than the fluid in which they were floating.

This is certainly true of ice floating in the Earth's oceans, and gave rise to the metaphor 'only the tip of the iceberg', which refers to the small fraction of something that is apparent when most of it is hidden. On Europa, the height difference can tell us the total thickness of the rafts, if we know the densities of the raft and the ocean. The principle behind this is known to geologists and geophysicists as 'isostasy' (see Box 4.7).

Isostasy is really just another name for buoyancy.

BOX 4.7 THE THICKNESS OF A FLOATING RAFT

Figure 4.27 shows a tabular raft floating at equilibrium (i.e. at its position of neutral buoyancy) in a liquid. In this situation, the pressure at the base of the raft must be the same as the pressure in the liquid immediately adjacent to the base of the raft. The formula for pressure, P, at depth d beneath a substance of density ρ is given by:

$$P = \rho g d \qquad (4.2)$$

where g is the acceleration due to gravity. In the situation illustrated in Figure 4.27, identical pressures occur at the base of the raft, which occurs below a total raft thickness of $(h + w)$ and at a depth w in the liquid. The difference in any atmospheric pressure between the raft surface and the liquid surface is negligible, so we can ignore this and write:

$$P = \rho_1 g(h + w) = \rho_2 g w$$

As we are interested in determining the raft thickness, $(h + w)$, we can divide by g, to get:

$$\rho_1(h + w) = \rho_2 w \qquad (4.3)$$

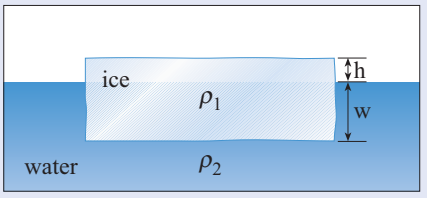


Figure 4.27 Cross-section of a raft (density ρ_1) floating in a liquid (density ρ_2). The height of the raft's surface above the liquid is h and the depth below the liquid surface to the base of the raft is w.

We do not actually know the density of the raft (impure ice) or of the liquid (likely to be a salt solution, rather than pure water). However, we can assume a reasonable range of values, given that we can be fairly sure that the raft is mostly H_2O ice and that the liquid is some kind of salty water. The density of water rich in dissolved sulfates of magnesium and sodium (for example of a composition close to that in Figure 4.15) would be about 1180 kg m^{-3}. Ice freezing from such a solution could have a density as high as 1126 kg m^{-3} if rich in these salts or as low as 927 kg m^{-3} if salt-free.

QUESTION 4.10

(a) Rearrange Equation 4.3 to find an expression for w.

(b) Use this rearranged equation to determine the maximum and minimum depths to the base of the rafts, and hence the raft thicknesses in Conamara Chaos, given that h is 100 m, ρ_1 is not less than 927 kg m^{-3} and not more than 1126 kg m^{-3}, and ρ_2 is 1180 kg m^{-3}.

If you were to assume pure ice floating in pure water, this method would give a raft thickness intermediate between the extremes you calculated in Question 4.10. Thus, the heights of the cliffs at the edges of rafts show with a fair degree of confidence that when the ice broke up to create the rafts its thickness was not less than a few hundred metres and not more than a few kilometres.

This is not necessarily the long-term ice thickness on Europa. Clearly, it is possible that the local heating event responsible for chaos generation might have melted quite a lot from the base of the continuous ice sheet before this finally broke up. On the other hand, the method we have used to calculate the thickness of the rafts relies on the ice of the re-frozen matrix being both thinner and weaker than the raft ice, at least until cooling-related thickening and ridge and groove development has turned the matrix into 'ball of string' terrain. So we can imagine regions of ice on Europa both thinner and thicker than the values you calculated in Question 4.10.

4.2.4 Heat and life

The weight of evidence in the case of Europa points strongly towards ice overlying salty water, at least within the past few millions years although not necessarily today. There are signs that localized heating episodes have melted and fractured the ice. The intensity of tidal heating has probably waxed and waned in step with fluctuations in the amount of forced eccentricity of Europa's orbit, but we can anticipate that conditions on Europa would have varied through a broadly similar range during much of the Solar System's lifetime. What, then, are the prospects for life on Europa?

Let's consider the surface ice first. You learned about cold-tolerant (psychrophilic) organisms in Chapter 2. On Earth, active microbial communities have been found within Antarctic sea ice at temperatures as low as –18 °C. Here, algae and other organisms survive by photosynthesis in summer that is possibly supplemented when there is less light available by metabolising dissolved organic matter, but these are probably survivor species that need liquid water for part of their life cycles.

■ Can you suggest four reasons why Europan surface ice is unlikely to be so hospitable for life as Antarctic sea ice?

❑ Firstly, Europa's surface temperature of –140 °C even at the equator is far lower than in Antarctic sea ice, and we know of no way for water-based metabolism to proceed in such cold conditions. Secondly, liquid water would occur here far less frequently than within the Antarctic ice shelf. Thirdly, Jupiter is 5.2 AU from the Sun, so (according to the inverse-square law) the sunlight available for photosynthesis on Europa is some 27 times weaker than on Earth. Fourthly, unless there is a thriving ecosystem elsewhere on Europa, there would be no dissolved organic matter food source to supplement the energy available from photosynthesis.

Thanks to the escape of tidal heat, the temperature within Europa's ice is likely to increase with depth. However, even on Earth the light intensity is too low for photosynthesis to continue more than about 20 m deep within the ice. This is only a tiny fraction of Europa's ice thickness. There could be no ice warmer than –20 °C at a shallow enough depth for photosynthesis, except within very young matrix ice of chaos regions, or in the walls of fissures for brief periods during fracturing or ridge building eruptions. It is faintly conceivable that primitive photosynthetic organisms may lie entombed and dormant within Europa's near-surface ice for periods of millions of years, and become active only during relatively brief episodes of local heating (full-blown chaos generation, or above warm dome-forming intrusions as in Figure 4.23, or within an active fissure). This would be a pretty marginal existence. It is perhaps to the energy and nutrients that could be provided by hydrothermal vents that we must appeal if we wish to find the basis of a robust and persistent ecology of the kind imagined by Arthur C. Clarke (Section 4.1.4).

Whether hydrothermal vents exist on Europa, and, if so, their abundance and their power, depend upon how deep within Europa the tidal heating occurs. This has not been determined, because it depends on unknown factors such as the strength and other properties of Europa's ice and rocky interior. At one extreme, virtually all the tidal energy could be dissipated within the icy shell (in which case chaos formation would be a result of direct heating of the ice). This would mean that the ocean was kept warm largely because of heat from above. Any hydrothermal vents on the ocean floor would be scarce and weak, and powered only by the feeble leakage of radiogenic heat from Europa's rocky interior. On the other hand, if tidal heating were concentrated in Europa's rocky part, flow of heat from the rock into the overlying ocean would be much stronger. As on Earth, ocean water would soak into the underlying hot rock, where it would become heated and react chemically, eventually escaping back into the ocean via hydrothermal vents. A static rocky substrate would not be very favourable for sustaining life because the ocean would deplete the available chemicals over million-year timescales. However, if tidal heating were sufficient to cause partial melting within Europa's rock, hydrothermal circulation would be especially strong over sites where igneous rock was being intruded at shallow depth, and strongest of all at any places where volcanic eruptions occurred onto the ocean floor. Moreover, the repeated arrival of new igneous rock at or a little way below the ocean floor would mean that the chemistry was continually renewed, so that some of the circulating water would always find something with which to react.

- Thinking back to what you learned in Chapters 1 and 2 about the relationship between life and hydrothermal vents, can you suggest why the presence of hydrothermal vents on Europa could be particularly important for the origin of life on Europa?
- Phylogenetic evidence, in particular the ribosomal RNA tree (Figure 1.37), suggests that thermophylic autotrophic microbes dependent on chemosynthesis are the last common ancestor for life on Earth. Therefore life on Earth may well have begun at hot vents. If it did, then it could perhaps have begun with equal ease at hot vents on Europa.

An ocean of global extent would not have been necessary for life to begin. Relatively small pools of water sandwiched between ice and hot rock would have been enough. However an ocean, or at least an extensive body of water, would certainly make it easier for life to survive. Life that was trapped in a single pocket of water would have no escape when the hydrothermal vent that had been feeding it cooled down and ceased to flow. It would have to survive in a frozen state until the unlikely eventuality of a new vent starting up nearby. However, an ocean, or at least an extensive seaway, would mean that organisms (including free-floating larval stages of any multicellular life) could drift from vent to vent, allowing species to survive – even though individual colonies would meet their demise with the extinction of their vent.

The primary producers at hot-vent ecosystems on Earth derive their energy from a redox (oxidation–reduction) reaction. Typically, they exploit a reaction whose equilibrium position depends on temperature. For example if a high temperature (such as where hot fluids react with rock during hydrothermal circulation) drives the reaction in one direction but a low temperature (where vent water mixes back with ocean water) tends to drive the reaction the other way, then an organism can extract energy by getting involved in this 'reverse' reaction. This is only effective when the low-temperature ('reverse') reaction is kinetically inhibited, which provides the opportunity for a biological catalyst to become involved.

A chemical reaction is 'kinetically inhibited' when there is a significant energy barrier to be overcome to enable the reaction to proceed.

An example of this in ocean-floor hydrothermal systems on Earth is the biological production of methane ('methanogenesis'). During hydrothermal alteration of newly created oceanic crust iron reacts with water. The iron is oxidized and the water reduced to hydrogen. Carbon is discharged in vent fluids as carbon dioxide, arising partly from oxidation of crustal and mantle carbon and partly from breakdown of carbonate rocks that have been drawn into the mantle at subduction zones. Thus, hot vent fluids are rich in carbon dioxide and hydrogen. In solution, these gases are related by the equilibrium reaction:

$$CO_2(aq) + 4H_2(aq) \rightleftharpoons CH_4(aq) + 2H_2O(l) \qquad (4.4)$$

When a chemical reaction is written this way, (aq) signifies something in aqueous solution, (l) signifies a liquid, (s) signifies a solid and (g) is a gas

At high temperatures the equilibrium lies well to the left, so that in a hot solution carbon dioxide and hydrogen are stable. At lower temperatures, including those in seawater, the equilibrium position lies well to the right, but in a lifeless ocean an energy barrier would inhibit the reaction from moving in this direction. However, with biological mediation most of the carbon dioxide and hydrogen can react to form methane and water as the temperature falls. This is the reaction that methanogenic bacteria exploit as their source of energy.

$$2CO_2(aq) + 6H_2(aq) \rightarrow (CH_2O)_n + CH_4(aq) + 3H_2O(l) \quad (4.5)$$

$(CH_2O)_n$ indicates carbohydrate in biological cell material, and the subscript n indicates that the real formula is more complicated than simply CH_2O.

In principle, this reaction could be used by Europan equivalents of methanogenic bacteria at hot vents. There are reasons, however, why this particular reaction may not be a viable source of biological energy on Europa. One reason is that without (so far as we know) subduction of oxidized species, Europa's hydrothermal fluids are likely to be considerably more reducing than the Earth's. This would lead to vent fluids being naturally rich in methane rather than carbon dioxide, which would therefore deprive methanogens of their energy source. Another reason is that high pressure drives the reaction in Equation 4.4 towards the right. Now do Question 4.11, which compares the pressure on the Earth's ocean floor with that at the floor of Europa's ocean.

QUESTION 4.11

The pressure on an ocean floor is given by the expression $P = \rho g d$, which you have already met in a slightly different context in Box 4.7, as Equation 4.2. For our current purpose, ρ is the average density of the overlying ocean, g is the acceleration due to gravity on the planetary body concerned, and d is the depth of the ocean. On Earth, we can take ρ to be 1030 kg m^{-3}, g to be 9.8 m s^{-2}, and d to be 3 km (the approximate depth of a mid-ocean ridge). On Europa, treating the ice thickness as negligible relative to the ocean thickness, we can take ρ to be 1180 kg m^{-3}, g to be 1.3 m s^{-2}, and d to be 100 km. Use these values to calculate the pressure at the exit of a hydrothermal vent on:

(a) the Earth's ocean floor

(b) Europa's ocean floor.

Thus, the pressure on Europa's ocean floor is about five times that at a mid-ocean ridge hydrothermal vent on Earth. This may not seem a big difference, and would be unlikely to have any adverse effect on, say, biological cell structure. However, it would affect the equilibrium in Equation 4.4, so that carbon tended to be outgassed as methane rather than carbon dioxide. The situation would be even less favourable for methanogenic life if Europa's subduction-deprived mantle is more reduced than the Earth's, because this would make the methane to carbon dioxide ratio very high in the first place. It would also mean that Europa is unlikely to provide favourable habitats for analogues of terrestrial SLiME (subsurface lithautotrophic microbial ecosystems) of the kind you read about in Section 2.6.

Perhaps, then, biological methanogenesis is not viable on Europa. In an extreme case, Europa's hydrothermal fluids could be so reducing that the only plausible oxidants that could provide an energy source for life would be oxidized metals, such as ferric iron (Fe^{3+}). A suitable reaction is represented by:

$$2Fe(OH)_3(aq) + H_2(aq) \rightleftharpoons 2FeO(s) + 4H_2O(l) \quad (4.6)$$

in which the iron in vent fluids is reduced by reaction with hydrogen. Alternative reducing agents could be hydrogen sulfide or even methane. In all these cases, biological organisms could feed off the energy released during reduction of Fe^{3+} to Fe^{2+}.

On the other hand, it is conceivable that Europa's ocean may actually be moderately oxidizing in character.

■ Can you recall from earlier in this chapter a mechanism whereby molecular oxygen is known to be generated on Europa?

❑ In Section 4.2.1 you learned how exposure to charged particles and solar ultraviolet radiation in the near-surface ice leads to radiolytic and photolytic breakdown of water molecules to produce oxygen and hydrogen.

The hydrogen escapes relatively easily to space, but much of the oxygen is held within the ice crystals. These processes are only effective in the upper few micrometres (μm) of the ice, but 'gardening' by micrometeorites and slightly larger impacts can be expected to mix the products to a depth of about 1 m in the regolith. We do not know how efficiently, if at all, such oxygen is eventually mixed into the ocean, but obviously this could occur from time to time when melting, especially during chaos formation, reaches the surface.

There is actually a radiolytic mechanism whereby oxygen could be generated from either ice or liquid water at *any* depth below the surface. This is because one of the common elements thought to be dissolved in the Europan ocean has a radioactive isotope.

■ Look back at Figure 4.15, and see if you can recognize which of these elements has a radioactive isotope.

❑ The element with a radioactive isotope is potassium.

The radioactive isotope is ^{40}K, which on Earth, and presumably Europa too, makes up about 0.012% of the total potassium today, and would have been about ten times more common shortly after Europa was formed. β-particles and γ-radiation are emitted by ^{40}K as it decays and both can radiolytically break water into hydrogen and oxygen, by means of the series of reactions indicated in Box 4.4.

This process could yield about 10^{10} moles of oxygen per year in Europa's ocean. Provided there is sufficient carbon available and a suitable reaction pathway, this would be enough to support about 10^7–10^9 kg yr^{-1} of biomass production. However, the limited availability of carbon in the right form and right place almost certainly means that the actual rate (if any) of biomass production in Europa's ocean is probably less than this. A likely value, allowing for a modest amount of hydrothermal energy, is about 10^5–10^6 kg yr^{-1}.

QUESTION 4.12

Rates of biomass production on the Earth today are about 5×10^{13} kg yr^{-1} by photosynthesis on land, a similar amount by marine photosynthesis (mainly by microscopic plankton), and about 10^{10} kg yr^{-1} by chemosynthesis at ocean-floor hydrothermal vents.

How do these rates of biomass production compare (by orders of magnitude) with the value proposed for Europa?

It is important to remember that the estimates for Europa are very uncertain, and could be underestimates by two or three orders of magnitude – or ridiculous overestimates if Europa supports no life at all. However, Europa could offer sites that are just as favourable for life to have originated as those on the early Earth, and equally hospitable for so-called extremophiles to flourish today, albeit in smaller quantities than on Earth. Europa's small mass (so small that it cannot hold onto an atmosphere) and its distance from the Sun beyond the 'habitable zone' have prevented it from developing a photosynthesis-dominated biosphere, but the viability of any chemosynthetically supported biosphere is independent of this. Thus, according to some assumptions at least, the prospects for life on Europa appear encouraging.

Whether any life has remained at the level of simple single-celled autotrophs or diversified into multicellular forms, and whether any heterotrophic organisms have evolved to prey on these (as imagined by Arthur C. Clarke) remains to be seen.

We will conclude our discussion of Europa with a brief look at plans to gather further data on this intriguing world.

4.2.5 How can we find out more about Europa?

The next mission to be launched to the outer Solar System is likely to be NASA's Jupiter Icy Moons Orbiter (JIMO). This is tentatively scheduled for launch in 2011 at the earliest. On arrival at Jupiter it would go into orbit first round Callisto, then Ganymede and finally Europa.

The main objectives of JIMO at Europa are to:

1 Determine the presence or absence of a subsurface ocean.

2 Characterize the three-dimensional distribution of any subsurface liquid water and its overlying ice layers.

3 Understand the formation of surface features, including sites of recent or current activity, and identify candidate sites for future lander missions.

QUESTION 4.13

JIMO is an ambitious mission with ambitious objectives. What techniques do you think a spaceprobe in orbit could use to meet these objectives?

The answer we gave to Question 4.13 covers all we expected you to come up with, but there are other techniques that are also likely to be useful. At the time of writing, the actual instrument package for JIMO has not been finalized. However, it is expected to include an imaging system with spectroscopic capability, a laser altimeter, and an ice-penetrating radar. The laser altimeter will map Europa's topography, and in particular it will determine the height of Europa's tidal bulges. The bulges should be only about 1 m high if the ice is solid throughout, but about 30 m high if there is 10 km of ice floating on water, so altimetry is a neat way of addressing the presence or absence of a subsurface ocean. The radar will be directed directly downwards with the intention of recording echoes from the ice–water interface. Unless the ice is particularly salty, which would tend to attenuate

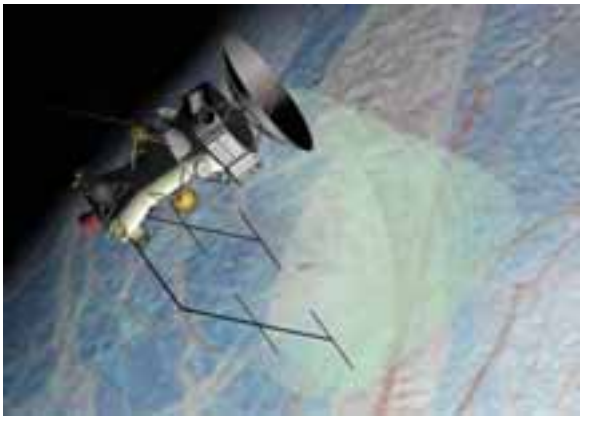


Figure 4.28 A Europa orbiter in action (not JIMO, but a simpler mission that has now been cancelled). The blue beam illuminating the surface is a schematic indication of the ice-penetrating radar beam, which is intended to map the ice thickness with a depth resolution of about 100 m.

the signal, the radar should detect the ice–water interface wherever it lies at less than about 10 km depth, as is likely to be the case in young chaos areas (Section 4.2.3). This, plus any further visual clues to ice thinness or recent activity from the imaging system, will be the main means of selecting landing sites for future lander missions.

You can keep track of plans, and eventually the mission as it proceeds, by selecting the Jupiter option at http://sse.jpl.nasa.gov/missions/index.cfm

An artist's impression of a Europa-orbiting mission in action is shown in Figure 4.28.

The next mission to Europa is likely to arrive several years after JIMO. The current ambition is to equip such a mission with a miniature robotic submarine (a 'hydrobot') capable of exploring the ocean to seek for signs of life. In order to deploy this, a way has to be found to make an access hole in the ice, which presumably must be done either by mechanical drilling or by using heat to melt a borehole. Even after landing on the thinnest ice, the technological challenges of making such a hole would be severe.

There is also another problem, which is the planetary protection issue (Chapter 3) of how to prevent contamination of Europa's biosphere with organisms inadvertently carried from Earth. It would be foolish to send a sophisticated suite of instruments to Europa unless we could be as certain as possible that any signs of biological activity were not attributable to microbes carried to Europa by the same mission or any previous spacecraft. Contamination of Europa's biosphere (or the accidental establishment of a biosphere where none had previously existed) would undermine any conclusions about the independent origin and evolution of life that could otherwise be drawn following the discovery and study of Europan life. Most investigators would recognize an ethical duty to safeguard the integrity of future studies of Europa's biosphere and to protect against potential harm to any Europan organisms. This duty is codified in legal form by the 1967 United Nations' *Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies*.

Very few terrestrial microbes would survive a journey to Europa, and of these only a tiny proportion would be likely to be able to feed and reproduce on Europa or in its ocean. However, just one viable organism delivered to the right (or wrong!) place that was then able to feed and multiply would do incalculable harm. With this in mind, a report on preventing biological contamination of Europa published in 2000 by the US National Academy of Sciences recommends that the bioload of any Europa-bound mission should be minimized by using levels of cleanliness during assembly and subsequent sterilization that are at least as stringent as those currently agreed for Mars missions.

Illuminating lessons about preparing to penetrate through a thick ice cover into a body of water may be learned from the case of Lake Vostok. This is a large lake that has been trapped beneath the Antarctic ice for possibly several million years, and is suspected of housing a sealed-in ecosystem. Exploration of Lake Vostok and the proper implementation of anti-contamination protocols are widely held to be realistic rehearsals for exploration of Europa's ocean, as discussed in Box 4.8.

BOX 4.8 LAKE VOSTOK – AN ICE CONUNDRUM

In 1974, Russian scientists began drilling deep into the ice at their Vostok research base, situated at the geomagnetic south pole in Antarctica. Samples of ice and the gases and other trace materials trapped within it provide a valuable and continuous record of climate changes and large volcanic eruptions during the past 400 000 years. Incidentally, viable micro-organisms were found entombed within the ancient ice too. It was not until 1994, by which time the borehole had reached a depth of about 3 km, that seismic and other studies revealed that the ice overlies the largest subglacial lake in the world, covering about 2×10^5 km^2, which is the same area as Lake Ontario. This is known as Lake Vostok (Figure 4.29).

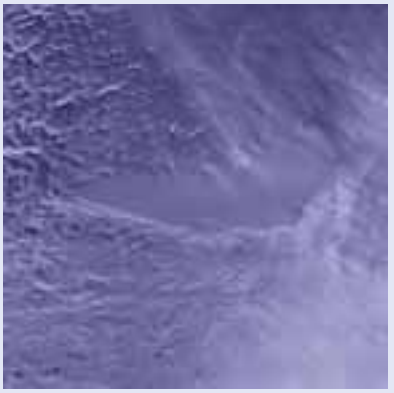


Figure 4.29 Satellite radar image showing the ice surface of part of Antarctica. Lake Vostok is the elongated flat area near the centre. Image is about 600 m across.

In places the water depth reaches about 1 km. The oldest ice overlying the water is less than a million years old, but the ice sheet as a whole is slowly flowing across the lake, so the lake itself may have been sealed off from the surface for as long as 14 million years. The lake is suspected of supporting its own ecosystem, subsisting either by a meagre rain of organic matter at places where the overlying ice melts or by chemical energy at suspected hot springs.

These realizations united scientists from many nations in plans to bore through the base of the ice in order to sample the lake water and deploy a probe into the lake. One method suggested to keep the hole sealed and to prevent contamination was that the hole should be drilled to within a few metres of the roof of the lake. A cylindrical probe would then be lowered to the base of the hole that could sterilize itself while waiting for the hole above to freeze over, and then melt its way down into the lake. It would pay out a tether behind itself as it travelled, which would act as a communications link to the surface (Figure 4.30).

However, two serious objections emerged that put at least a temporary halt to these schemes. First, the self-sterilization techniques for the probe were untested. Secondly, when the Russians had begun drilling back in the 1970s, they were anxious to stop the hole freezing shut behind the drill bit so they pumped a mixture of aviation fuel and antifreeze (Freon) into the hole. There is now 60 tonnes of this toxic chemical mix in the hole, and no one can be sure that none of this will leak into the lake if the hole is continued. Pressure from a coalition of environmental groups caused drilling to stop in 2001 (Figure 4.31), and plans for any kind of penetration into the lake were put on hold for maybe a decade.

Figure 4.30 Artist's impression of a probe released into Lake Vostok from the base of the borehole.

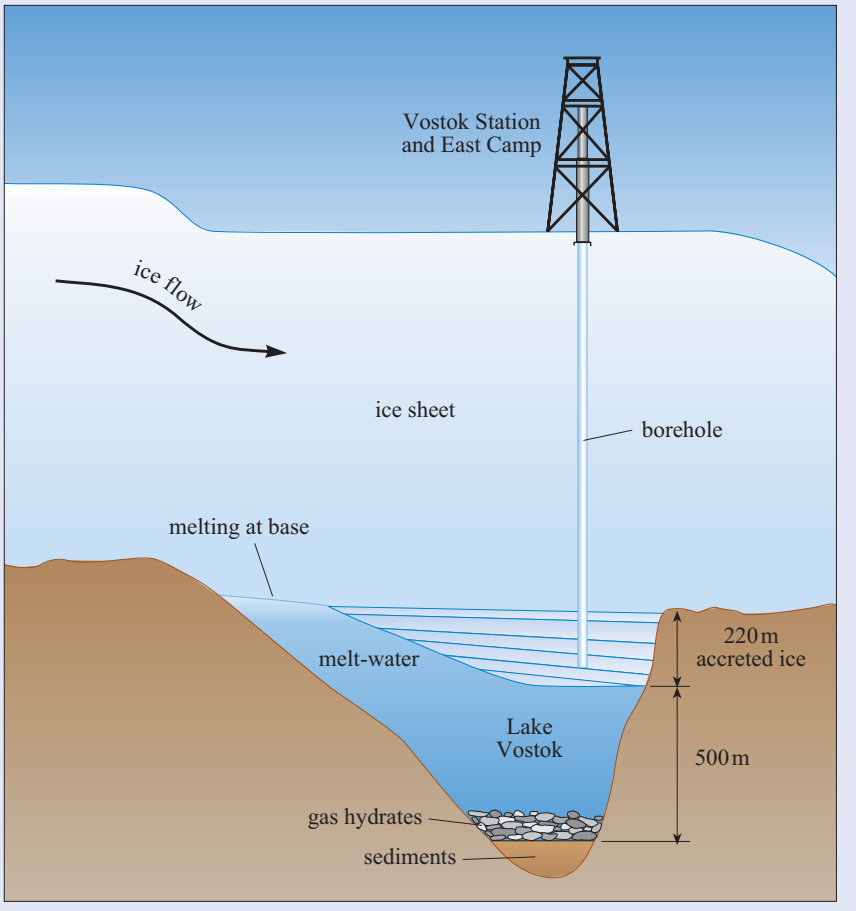


Figure 4.31 Schematic cross-section through Lake Vostok and the overlying ice (not to scale). The borehole stops less than 100 m before the roof of the lake. It has penetrated the full thickness of the ancient ice cap, and terminates within ice that has frozen more recently onto the roof.

4.3 Other icy bodies as abodes of life?

You have seen that Europa offers arguably the most promising habitat for present-day life in the Solar System, other than on the Earth itself. This is because ice or water overlying warm rock can lead to hydrothermal circulation, offering hot springs where life could originate in the first place and subsist on chemical energy thereafter. There would not actually have to be a global ocean below the ice, but this would help by allowing organisms to spread from one vent to another.

- Can you suggest places in the Solar System where conditions might now be, or might formerly have been, sufficiently similar to Europa for life to have got underway there too?
- Any of the tidally heated icy satellites would seem promising, especially those that were at any time heated sufficiently for a global ocean to form below the ice. Of those illustrated earlier in this chapter, Callisto and Rhea (Figure 4.10) do not seem promising, whereas Enceladus and Ariel (Figure 4.11b and d) would seem the likeliest.

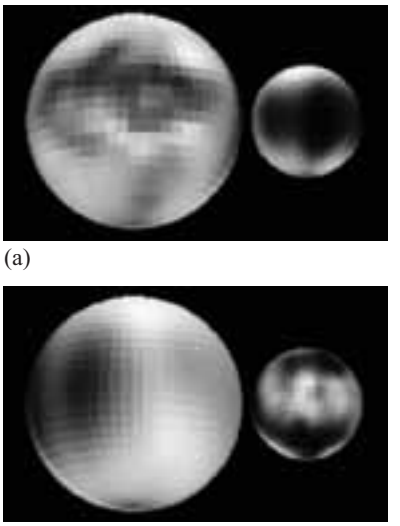
Enceladus is a particularly intriguing proposition. Voyager obtained useful images of less than half its surface. Although parts of Enceladus are fairly heavily cratered, other areas (e.g. the lower right of Figure 4.11b) show no craters at all even on the highest resolution (2 km per pixel) images. The Cassini mission (described in the next chapter) is scheduled for several close fly-bys of Enceladus during its 2004–2008 tour of the Saturn system. These new images may reveal when and how the youngest regions became resurfaced, help us to understand Encaladus's tidal heating history, and provide a basis for more informed speculation about its astrobiological potential. Look out for images and discussion at http://saturn.jpl.nasa.gov/science/index.cfm

The surface of Triton (Figure 4.6) shows great variety, with plenty of evidence of cryovolcanic resurfacing. We know from spectroscopic evidence that the surface ice is a mixture of nitrogen, methane, carbon dioxide, carbon monoxide and water, and it is likely that there is some ammonia too. This is probably a true differentiated crust in the geochemical sense, overlying a mantle that is richer in water-ice. Triton's bulk density suggests that a rocky core begins at a depth of about 350 km. There is a fair degree of superimposed impact cratering on all terrain types, so widespread cryovolcanism appears to have ceased – probably at least hundreds of millions of years ago. Apart from seasonal changes in the sizes of the polar caps of frozen nitrogen, and what appear to be solar-powered geysers rupturing the south polar cap, no current or recent activity has been identified. This is consistent with the lack of a known tidal heat source. However, in the aftermath of its capture by Neptune there would have been a period of probably about a billion years while tides acted to force Triton's orbit to become circular. This is probably when most of the cryovolcanism took place. During this period there could have been a Europa-like ocean below the ice with plenty of time for life to become established. If so, life could be clinging on thanks to feeble radiogenic heat – or perhaps future explorers will find nothing but the fossilized remains of an extinct biosphere.

Apart from Enceladus, the icy satellite that may be experiencing the greatest rate of tidal heating today is one that you probably did not consider at all. This is Charon, the single known satellite of Pluto (Table 4.1). We know much less about Charon than the other large icy satellites, because no spaceprobe has been there, so we have to rely on telescopic data, such as spectroscopic information and the albedo maps in Figure 4.32. Charon orbits in Pluto's equatorial plane, and their rotations are mutually tidally locked so that they permanently keep the same faces towards each other. There would seem little scope here for on-going tidal heating. However, Pluto's axis (and hence Charon's orbit) is tilted over at an angle of 119.6°. This leads to competing tidal pulls on Charon by Pluto and the Sun in such a configuration that, according to some models, there could be substantial tidal heating in Charon's interior today.

Pluto itself is probably not being heated tidally, but spectroscopy has revealed its surface to consist of at least as rich a cocktail of ices as on Triton. It is likely to be fully differentiated, especially if Charon owes its origin to a giant impact event similar to that which formed the Moon. An ocean, with life-bearing potential, could have persisted below the solid ice for a considerable period until most of the accretional heat from such a collision had leaked away. This heat would have been stoked up by tidal forces until Pluto's rotation became synchronous with Charon's orbital period. Speculation is likely to continue relatively unbounded until a spaceprobe visits Pluto and Charon. The first will probably be a much delayed

The tilt of a planet's axis is conventionally measured relative to the perpendicular to its orbital plane. Pluto's tilt of >90° signifies that its rotation is retrograde.



(a)

(b)

Figure 4.32 Maps of surface albedo patterns on Pluto (left) and Charon (right) calculated from variations in brightness as they rotate, and a series of mutual occultations that occurred during 1985–1990. (a) Charon-facing side of Pluto and anti-Pluto side of Charon; (b) anti-Charon side of Pluto and Pluto-facing side of Charon.

NASA mission currently called New Horizons. The earliest date this could now be launched is 2006, and its Pluto fly-by cannot happen until 2015 at the earliest.

In an icy satellite with no tidal heating, we would have to look to other sources of geothermal heat to power hydrothermal vents, such as radiogenic heating or heat left over from the accretion process. Any icy satellite with a differentiated structure must have experienced at least some water–rock geochemical interaction, but this may not necessarily have been strong enough or sufficiently prolonged to favour life.

Finally, what about Jupiter's outer two Galilean satellites?

- Does the evidence in Figures 4.10a and 4.14 make Callisto look like a favourable site for a Europa-style ocean?
- Figure 4.10a shows a uniformly heavily cratered and, therefore, ancient surface, and Figure 4.14 shows an interior that is only weakly differentiated. Both these factors suggest that an ocean is unlikely.

It is not thought likely that Callisto experienced significant tidal heating after its rotation became synchronous. Ganymede, however, may have been affected by tidal heating episodes (though not so strongly as Europa) when its degree of forced eccentricity fluctuated as a result of mutual orbital interaction with Europa and Callisto. It bears signs of this in the way that its ancient heavily cratered surface is transected by belts of younger terrain (Figure 4.33). However, even the youngest belts of cross-cutting terrain on Ganymede have numerous impact craters superimposed on them, and are likely to exceed a billion years in age.

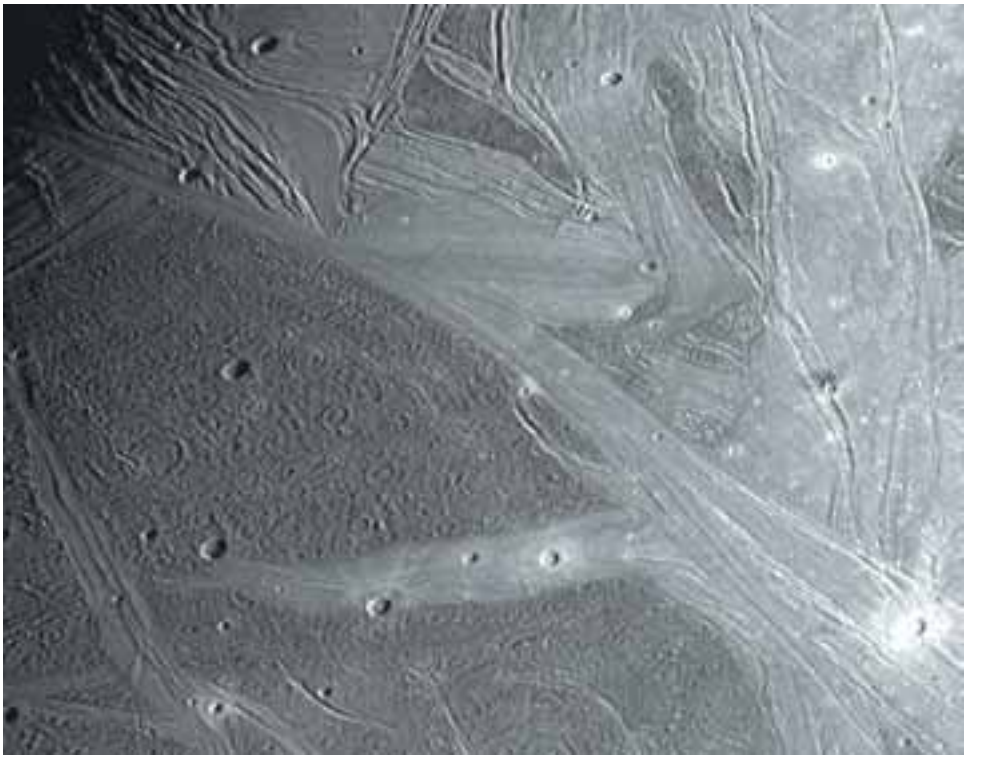


Figure 4.33 A Galileo SSI view, 600 km in width, of part of Ganymede. Several generations of ridged and grooved pale terrain cut across an older and more heavily cratered terrain.

The imaging data would seem to be giving us a clear story. However, the Galileo Orbiter measured magnetic fields apparently induced within both these satellites by their orbital passage through Jupiter's magnetosphere, which complicate the issue. In the case of Ganymede, the induced field could originate in an iron-rich core, but what we know of Callisto's internal density distribution shows fairly robustly that it can have no such core. That being so, the only reasonable explanation remaining seems to be that Callisto has an electrically conducting ocean at least 10 km thick and no more than 100 km deep. A similar ocean could also account for Ganymede's magnetic field.

The proposition that there are relatively shallow oceans beneath the surfaces of Ganymede and Callisto seems totally at odds with the ancient appearances of their surfaces, and allows us to end this chapter with a caution.

> Where there is an ocean there could be life, but we understand far too little about any of the icy satellites. Although there appear to be many reasons why some of them could harbour life, and that most could have done so at times in the past, it may be a long time before we know for sure.

Now test some of the knowledge and skills you have developed in this chapter by answering the following questions.

QUESTION 4.14

Examine again the groove in Figure 4.24 that you looked at in Question 4.9 (the same groove that you can see near the right-hand edge of Figure 4.26). On Figure 4.24, locate where the line of this groove passes between adjacent rafts about 5 km northwestward of the edge of the outline indicating Figure 4.26 (conveniently, 5 km is approximately the length of a short side of this outline).

(a) What evidence is there for the relative ages of this groove and the matrix between the rafts at this location?

(b) What are the implications for the time period over which the matrix was mobile in Conamara Chaos as a whole?

(c) How can we deduce that the ejecta from the Pwyll impact overlies the matrix at this location, and what does that tell you?

QUESTION 4.15

In Question 4.10b you calculated a raft thickness in Conamara Chaos based on the assumption that the heights of the cliffs at the raft edges were a result of rafts floating in quite a dense brine. However, perhaps at the time when the topography became 'frozen in' the rafts were actually floating in some kind of slush, which would be less dense than the brine. What is the implied raft thickness for a raft density of 1126 kg m^{-3} and a slush density 1140 kg m^{-3}?

QUESTION 4.16

Imagine it is the year 2100, and that the fifth of a series of probes into Europa's ocean has at last detected life in the form of micro-organisms that appear to be based on the same sort of DNA as on Earth. List the alternative implications for the establishment of life on Europa that could be drawn from this discovery, and how you would hope (eventually) to deduce the truth.

4.4 Summary of Chapter 4

- Many of the large icy bodies in the outer Solar System are internally differentiated. Thanks largely to tidal heating, some, especially Europa, are likely to have an ocean sandwiched between the icy exterior and the rocky core. Others may have had such an ocean in the past.
- Wherever water rests on warm rock, water must percolate into it and become heated. This will cause hydrothermal convection to begin. Hot, chemical-rich water will emerge at vents, where the resulting local chemical disequilibrium provides an opportunity for living organisms to extract energy by acting as mediators (biological catalysts) for redox reactions.
- If it is true that life on Earth originated at hydrothermal vents, then it is equally likely that life could have become established around similar vents at the 'ice'-rock interface on icy bodies.

CHAPTER 5
TITAN

5.1 Introduction

Saturn has at least 30 satellites, so you may wonder why we've devoted an entire chapter to one of them: Titan. Titan is unique in that it is the only satellite known to possess a dense atmosphere and, as you will see in Section 5.4 its surface may present us with a very exotic environment. Although it's not generally believed that Titan has ever harboured life, it does have a role to play in our understanding of the development of life. This is because the photochemical processes presently occurring in its atmosphere are believed to result in the formation of a wide range of organic molecules. In this chapter, you will look at our present knowledge of Titan, the theoretical models that explain some of these observations and finally the prospects for improving our knowledge of Titan in the near future.

5.2 Observations

Titan was discovered by the Dutch physicist and astronomer, Christiaan Huygens (1629–1695), pronounced 'hoi-gens' in English, (Figure 5.1) some 45 years after Galileo discovered the four large moons of Jupiter (the Galilean satellites). However, Titan was not named until the mid-1800s when British astronomer John Herschel (son of William Herschel who had himself discovered two satellites of Saturn) suggested that Saturn's satellites should be named after Saturn's brothers, collectively called the Titans, and Saturn's sisters, the Titanesses. These were the mythological giants who were believed to rule in the heavens before Jupiter conquered them. Because the satellite discovered by Huygens was so much larger than the rest, astronomers chose to name it Titan rather than naming it after one of the individual Titans.

Figure 5.1 Christiaan Huygens (1629–1695), sometimes spelled Christian Huyghens, Dutch mathematician, physicist and astronomer, who discovered Titan in 1655, using a telescope (built by himself and his brother Constantijn) of far better quality than those used by Galileo. Within a year of this discovery, he successfully explained the nature of Saturn's rings. Also amongst many other achievements in his career, he made fundamental contributions to the understanding of the nature of light and invented the pendulum clock. The ESA Probe that is due to land on Titan's surface as part of the Cassini mission has been named in his honour.

'Limb' is a term used by astronomers to describe the apparent edge of the Sun, Moon or any other celestial body with a detectable disc.

Not much more was learnt about Titan until the early years of the 20th century when the Spanish astronomer Comas Sola published some observations which mentioned, almost in passing, that he had detected the phenomenon of **limb darkening** when observing Titan, a phenomenon in which a planet or star appears darker at its limb than at its centre.

The exact cause is not important to us here, what is vital though is that this phenomenon requires the existence of an *atmosphere.* Limb darkening, therefore, provided the first indication that Titan possessed an atmosphere. The next major step forward came in the early 1940s when Gerard Kuiper (Figure 4.8), using a spectrometer on the new 82 inch (2.08 m) McDonald Observatory telescope in Texas, identified the characteristic signature of gaseous methane in the near-infrared light coming from Titan. This essentially confirmed the existence of an atmosphere around Titan. No other satellite of any of the planets has been found to have anything other than a minute trace of an atmosphere.

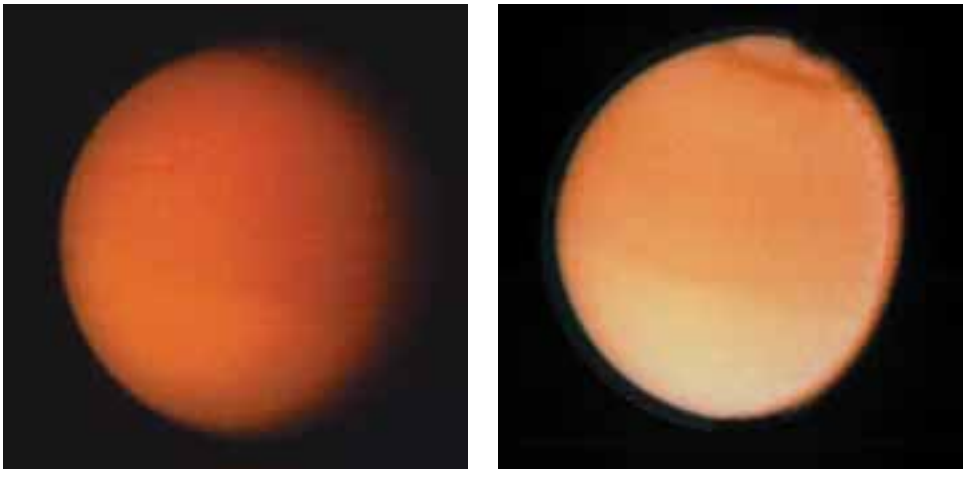
With improving telescopes and instruments, observations of Titan became more sophisticated over the following decades, but it was the space age that provided the 'quantum leap' in our understanding of Titan. The first spacecraft fly-by was by Pioneer 11 in September 1979 but with its relatively unsophisticated collection of instruments and a closest approach of some 363 000 km, little was learnt about Titan. It was in November 1980 when the Voyager 1 spacecraft flew by Titan at a distance of 4394 km that our knowledge of Titan increased enormously.

- What are some of the benefits of making observations of a planet or satellite from a fly-by spacecraft compared to ground-based Earth observations?

- The observations aren't constrained by limits imposed by the Earth's atmosphere (which means you can observe right across the infrared and ultraviolet parts of the electromagnetic spectrum where much of the information about the composition and structure of an atmosphere lies), the spatial resolution (i.e. smallest detail discernible) is much improved (compared with observing from the Earth) and the intensity of the emitted radiation (or any other phenomenon) is generally much higher because of the closeness of the observing point.

When the Voyager 1 instrument suite was turned towards Titan, several hundred images were taken with its on-board imaging system. A typical example is shown in Figure 5.2a. You might find this image slightly bland and disappointing – well so did the waiting scientists over 20 years ago! Indeed all the Voyager 1 images showed an almost featureless orange haze that covers all of Titan. However, images from Voyager 2, from a much greater distance, revealed more features. An example is shown in Figure 5.2b where you can see a faint dark band around the north pole and a slight contrast between the northern and southern hemispheres – the northern hemisphere was observed to be about 20% darker at blue (i.e. shorter) wavelengths.

Despite extensive image processing carried out on the images, there were no signs of any gaps in the haze to give even a fleeting view of the surface. Neither did they show any feature that could be described as a cloud, which could be tracked to give an indication of wind velocity in the atmosphere.



(a) (b)

Figure 5.2 (a) Voyager 1 and (b) Voyager 2 images of Titan acquired during the fly-bys in 1980 and 1981. The only discernible features are a dark polar 'hood' and a slight contrast between the hemispheres in (b).

However, the Voyager spacecraft carried an array of scientific instruments, including a set of spectrometers covering the infrared and ultraviolet parts of the electromagnetic spectrum**.** When the radiation from Titan was analysed with these instruments a set of complex spectra with a wealth of features was obtained, including spectra indicative of particular gases in Titan's atmosphere.

Figure 5.3 shows a typical spectrum from Titan obtained by the Voyager infrared spectrometer (IRIS – Infrared Radiometer Interferometer and Spectrometer). Each major feature is marked with an identification of the gas in the atmosphere that is responsible for emitting that particular wavelength. The height or intensity of each feature can be used to calculate the relative concentration of that particular gas in Titan's atmosphere. The calculated relative concentrations of all the detected gases from the Voyager spectrometers and later Earth-based telescopic studies are shown in Table 5.1.

■ What are the main similarities and differences between the compositions of Titan's atmosphere and Earth's atmosphere?

❑ The main similarity is that each atmosphere has nitrogen as its main constituent. A striking difference is that Titan's atmosphere is rich in hydrocarbons.

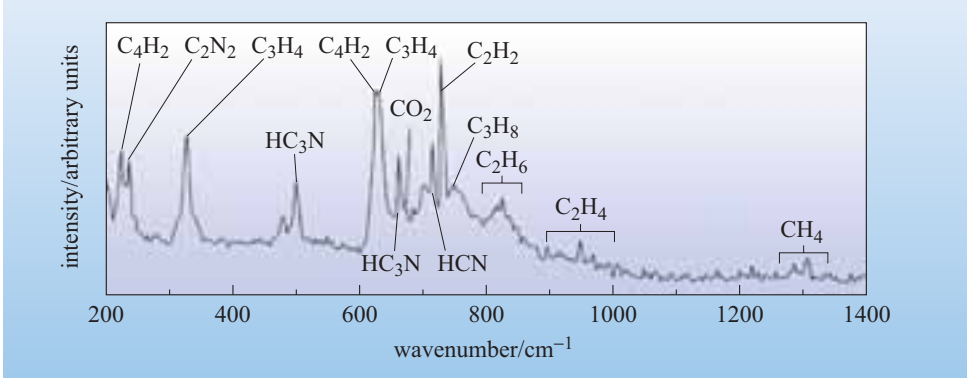


Figure 5.3 Voyager infrared spectrometer (IRIS) data from Titan's atmosphere.

Table 5.1 Composition of Titan's atmosphere. The relative abundance of each molecule is the fraction of the total number of molecules

Molecule	Relative abundance
N_2	0.97
CH_4 (methane)	3×10^{-2}
H_2	2×10^{-3}
CO (carbon monoxide)[a]	6×10^{-5}
C_2H_6 (ethane)	2×10^{-5}
C_2H_4 (ethene)	4×10^{-7}
C_2H_2 (ethyne)	2×10^{-6}
C_3H_8 (propane)	$(2 \text{ to } 4) \times 10^{-6}$
HCN (hydrogen cyanide)	2×10^{-7}
CH_3CCH (propyne)	3×10^{-8}
CHCCCH (butadiyne)	$(1 \text{ to } 10) \times 10^{-8}$
C_2N_2 (cyanogen)	$(1 \text{ to } 10) \times 10^{-8}$
HCCCN (cyanoethyne)	$(1 \text{ to } 10) \times 10^{-8}$
H_2O[b]	8×10^{-9}
CO_2	$(3 \text{ to } 7) \times 10^{-10}$

[a] Voyager did not detect carbon monoxide since IRIS did not cover the part of the spectrum where carbon monoxide might have been detected. It was discovered later by ground-based observations.

[b] Voyager determined a water concentration of less than of 11×10^{-9} (i.e. water was present at a concentration smaller than this value). The value in this table was determined at a later date by the Infrared Satellite Observatory (ISO).

The Voyager 1 observations were also able to provide some limited information on the variation of the relative concentration of gases with location in Titan's atmosphere. The 300 usable infrared spectra from Voyager 1 fell into 8 distinct latitudinal areas, from 60° S to 70° N, and therefore offered information on whether the minor constituents varied with location on Titan. The three most abundant hydrocarbons after methane, namely ethane, ethyne and propane, didn't show significant variations with latitude. They seemed to be homogeneously mixed in Titan's atmosphere from pole to pole. By contrast, ethene and propyne were found to increase by a factor of 10 towards the north pole, while the abundances seemed constant from mid-latitudes to the south pole. Other variations were also measured and some of these latitudinal variations are shown in Figure 5.4.

The latitudinal variations in the relative concentration of gases are only understood in the most general terms but are believed to be related to seasonal effects. In addition to these latitudinal measurements, a sequence of 30 infrared spectra taken towards Titan's north pole were apparently correlated with different altitude levels. This has enabled models of the vertical concentrations of some of Titan's gases detected near the north pole to be developed and these are shown in Figure 5.5. We will examine these in more detail in Section 5.3.1.

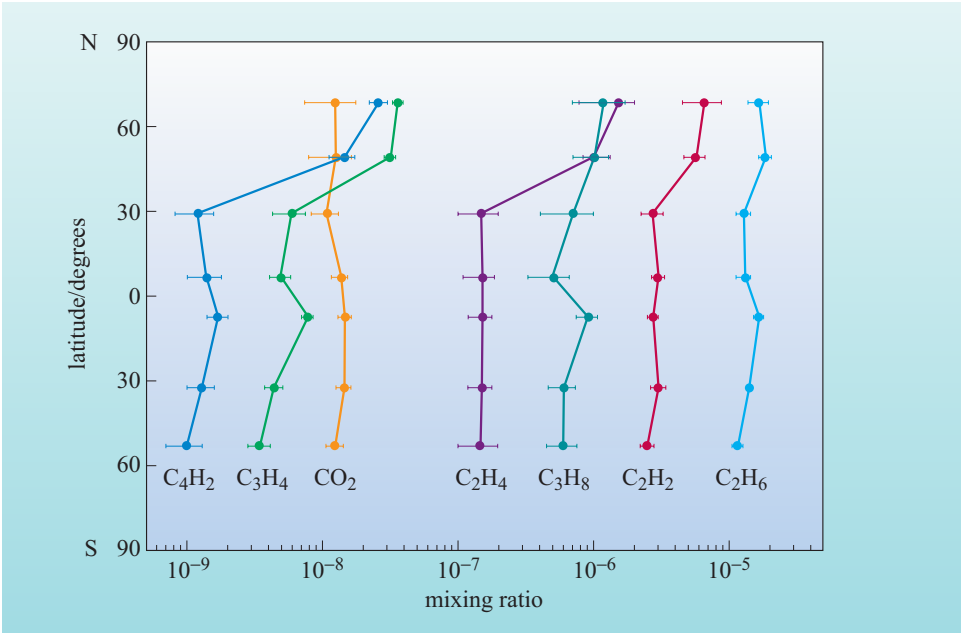


Figure 5.4 Compositional variability with latitude in Titan's stratosphere observed with Voyager 1.

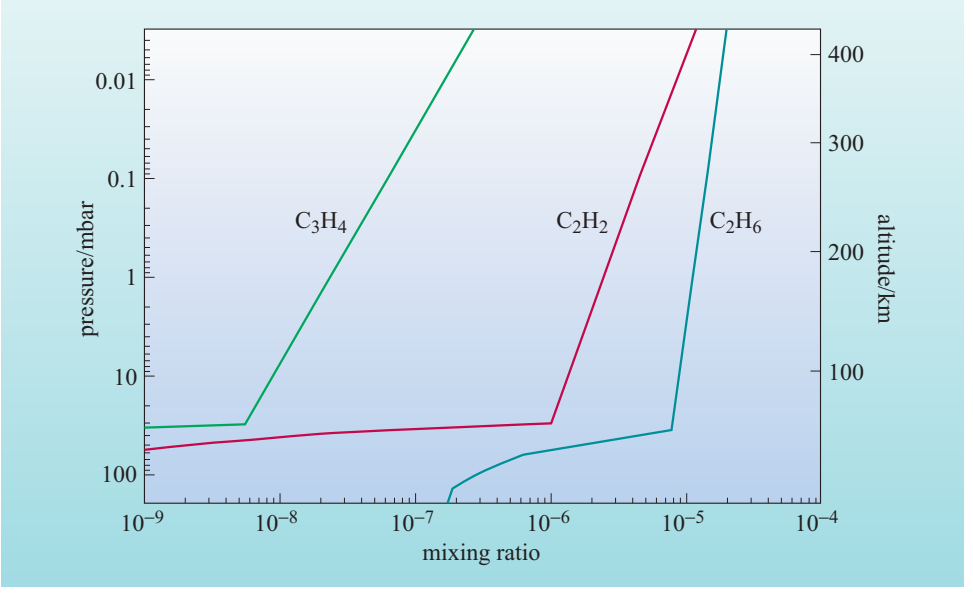


Figure 5.5 The modelled vertical concentration profiles of some of the gases detected near the north pole of Titan.

The Voyager observations of Titan's atmosphere caused enormous interest among planetary scientists. Like the Earth, its dominant species is molecular nitrogen, but the rich inventory of hydrocarbon gases raised special attention. As we shall see, their existence means that Titan's atmosphere is a vast chemical laboratory undergoing a whole range of complex reactions. In fact, Titan can be considered as a natural Solar System laboratory in which organic evolution is still occurring. It allows us to study organic chemical evolution under totally natural conditions over vast time-spans. Of particular interest is the relationship of Titan's atmospheric chemistry to prebiotic chemical pathways on other planets, including the early Earth.

Before considering aspects of the atmosphere in detail, we should consider some of the other data that Voyager was able to measure. The Voyager Radio Science System (RSS) used an onboard radio transmitter to send a signal back to the Earth. By arranging for the line-of-sight back to the Earth to pass through Titan's atmosphere (Figure 5.6), it was possible to measure various properties of the atmosphere such as temperature and pressure profiles (Figure 5.7). Measurements were possible down to the surface of Titan since the haze layer that prevented visible images of the surface from being obtained, did not adversely affect radio waves. The surface temperature is 94 K, which drops to 71 K at the tropopause level of about 45 km.

■ Should we be surprised by how cold Titan's surface temperature is?

❑ No – because of Titan's distance from the Sun, namely about 9.6 AU, the solar input to Titan is almost 100 (actually $9.6 \times 9.6 \approx 92$) times weaker than at the Earth. Therefore, unless there is another source of heat (i.e. internal), you would expect the temperature to be very low.

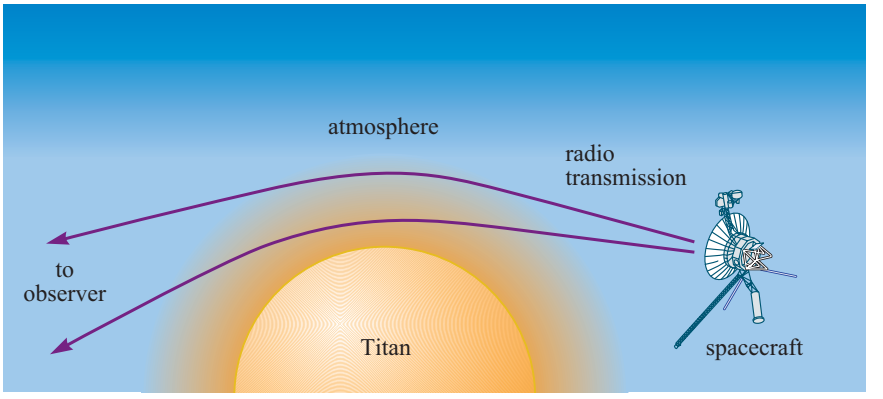


Figure 5.6 Schematic diagram showing the use of the Voyager RSS to determine atmospheric properties. When the path from the spacecraft to the Earth passes through an atmosphere, the signal is deflected, as shown. In addition, the strength of the signal is reduced and its polarization changed.

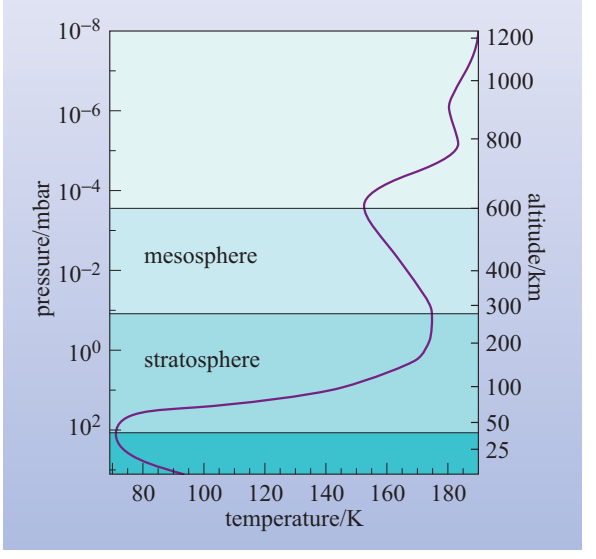


Figure 5.7 Temperature profile for Titan, derived from Voyager RSS data. Also derived from RSS data was the pressure profile that can be constructed by comparing the left-hand (pressure) and right-hand (altitude) axes. Also shown is the designation of various regions of the atmosphere, named by analogy with Earth, although 2 regions have been left blank because you will identify them in Question 5.1.

QUESTION 5.1

(a) Add the labels 'troposphere' and 'thermosphere' to the temperature profile for appropriate sections of the curve in Figure 5.7.

(b) By measuring from Figure 5.7, estimate the temperature gradient (or lapse rate) for the lowest regions of the atmosphere. How does this compare with the value measured by Voyager 1 and quoted in Section 5.4?

If you examine the pressure information in Figure 5.7, you will see that the surface pressure on Titan is about 1.5 bar, i.e. 50% greater than the surface pressure on Earth. Coupled with the fact that Titan's surface gravity is only about 15% of that on Earth, this means that we are dealing with a very substantial and dense atmosphere.

■ Can you recall a parameter that is useful for comparing the 'quantity of atmosphere' on a planetary body?

❑ The column mass, M_c (see Box 3.2) defined as:

$$M_c = P/g \qquad (5.1)$$

where P is the atmospheric pressure and g is the gravitational acceleration at the surface.

■ Using data from Table 5.2 and Equation 5.1, calculate the column mass of Titan's atmosphere. (*Hint*: Think carefully about units.)

❑ From Table 5.2, $P = 1496$ mbar ≈ 1.5 bar and $g = 1.35\ \mathrm{m\,s^{-2}}$. But we need to use the SI unit of pressure, namely the pascal (Pa), with 1 bar $= 10^5$ Pa.

So $M_c = 1.5 \times 10^5\ \mathrm{Pa}/1.35\ \mathrm{m\,s^{-2}} = 1.1 \times 10^5\ \mathrm{kg\,m^{-2}}$.

This figure should be compared with the atmospheric column mass value for Earth, namely $1.0 \times 10^4\ \mathrm{kg\,m^{-2}}$ thus confirming that Titan's atmosphere is more substantial than Earth's.

With the data obtained from the Voyager 1 encounter in November 1980, the Voyager 2 encounter some 9 months later (with a fly-by distance some 170 times greater than for Voyager 1, i.e. 663 385 km as opposed to 4394 km) and previous data, we are now in a position to examine some of Titan's basic parameters. These are shown in Table 5.2 and several points are worth noting. First, we see that Titan is larger than the terrestrial planet Mercury. It is the second largest planetary satellite, only just smaller than Ganymede. Titan, in common with most other planetary satellites, has its axial rotation period 'locked' to the orbital period. This is known as *synchronous rotation* and results from tidal forces generated when a major satellite orbits a planet (Box 4.2).

■ Can you think of another satellite that shows this effect?

❑ The Moon. This is why the same face is always turned towards the Earth.

For many years, Titan was believed to be the largest satellite before it was realized that the dimensions of the visible image included the haze layer above the solid surface.

Table 5.2 Titan's vital statistics.

Equatorial radius	(2575 ± 0.5) km
Mean density	1.88×10^3 kg m^{-3}
Distance from centre of Saturn	1.22×10^6 km
Mass	1.346×10^{23} kg
Surface gravity	1.35 m s^{-2}
Orbital period	15.95 days
Axial rotation period	15.95 days
Orbital eccentricity	0.0292
Surface temperature	(94.0 ± 0.7) K
Surface pressure	(1496 ± 20) mbar
Main atmospheric constituents	N_2, CH_4, H_2, CO
Mass of atmosphere	4×10^{17} kg

QUESTION 5.2

Using the data for the mass and radius of Titan in Table 5.2, confirm the value for the mean density of Titan. How does this value compare with other solid bodies in the Solar System and what does this suggest to you about the composition of Titan?

QUESTION 5.3

The surface gravity on a solid body can be calculated from the expression $g = GM/R^2$, where G is the gravitational constant, and M and R are the body's mass and radius respectively. Using the data for mass and radius in Table 5.2, confirm the value for surface gravity quoted in the same table.

5.3 Titan's atmosphere

The atmospheres of the giant planets are in constant turmoil. Recorded wind speeds can be very high, and there are vast storms. In chemical terms, however, it is safe to assume that below the cloud layers the various atoms and molecules are in chemical equilibrium. Indeed, the very turbulence of the atmosphere helps the gases to reach equilibrium by ensuring that they are well mixed.

■ What extra participant in atmospheric chemistry must we consider above the clouds that is less significant lower down?

❑ Sunlight.

Although in the dark outer reaches of the Solar System much less sunlight is received per m^2 than on Earth, the absorption of light still plays an important role in the atmospheric chemistry of Titan. This chemistry is rich and complex and, rather than trying to cover all the important reactions, we will consider two problems that arise from the observed concentrations of different molecules in Titan's atmosphere:

1 Carbon-containing molecules occur in Titan's atmosphere that would not be expected were the atmosphere in chemical equilibrium (Box 5.1).

2 The element nitrogen occurs on Titan as the diatomic nitrogen molecule, N_2, whereas it occurs as the compound ammonia, NH_3, in the atmospheres of the giant planets.

5.3.1 Hydrocarbons

As you saw in Table 5.1, the atmosphere of Titan contains a wide range of organic molecules, in particular hydrocarbons.

Modelling of Titan's atmosphere indicates that if it were in **chemical equilibrium**, the predominant hydrocarbon would be methane, CH_4, with only negligible amounts of the other hydrocarbons present (Box 5.1). So why do we observe hydrocarbons such as ethene, ethane, etc? These molecules are the result of photochemical reactions in the outer atmosphere. In those regions where solar radiation (particularly ultraviolet) penetrates, atmospheric composition is determined not by chemical equilibrium but by the interaction of molecules with radiation.

BOX 5.1 CHEMICAL EQUILIBRIUM

If we take a box containing a mixture of chemicals, say nitrogen, hydrogen and ammonia and leave it for a very long time, ensuring that the temperature and pressure remain constant and that no chemicals or radiation leave or enter the box, then the chemicals will reach equilibrium. At equilibrium, chemical reactions will be taking place but the total rate of production of each compound equals the total rate of its destruction so that the amounts of the various compounds present will stay the same. For any chemical reaction occurring, the relative amounts of the compounds involved that are present at equilibrium are given by the **equilibrium constant**, K, for that reaction. The value of K varies with the temperature, but does not depend on the total amount of chemicals present.

Let us take as an example of chemical equilibrium the reaction between carbon, C, and hydrogen, H_2, to form methane, CH_4:

$$C + 2H_2 = CH_4 \qquad (5.2)$$

We can put into our box any amount of carbon, hydrogen and/or methane. It does not matter whether we start with a mixture of carbon and hydrogen, or with methane, or with a mixture of hydrogen and methane, so long as both elements are present in sufficient amounts. After a very long time, we will have carbon, hydrogen and methane present in equilibrium amounts. If we measured the concentrations of the three compounds at equilibrium we could obtain the equilibrium constant, K. For the reaction in Equation 5.2, the equilibrium constant is given by

$$K = \frac{[CH_4]}{[C]\times[H_2]^2} \qquad (5.3)$$

where the square brackets [ ] denote concentrations. The value of K is such that if the concentration of hydrogen is very much higher than that of carbon then most of the carbon will be converted to methane. However, in the atmosphere of Titan, several molecules containing just carbon and hydrogen are observed. We have to include equilibrium constants for the reaction between carbon, C, and hydrogen, H_2, to form methane, CH_4, ethane, C_2H_6, ethene, C_2H_4, and ethyne, C_2H_2. Given the equilibrium constants we can calculate the relative amounts of hydrogen, methane, ethane, ethene and ethyne. But in addition we have to consider the equilibria between carbon, hydrogen and other elements such as nitrogen and oxygen to form compounds such as carbon monoxide, ammonia and water. All these equilibria are linked so that a large number of equations have to be solved to obtain the abundances. Luckily this is just the sort of problem that computers are good at.

From our knowledge of how chemicals react, we can choose for any planetary atmosphere a set of reactions involving the most likely molecules formed from the most abundant elements.

For example, assuming chemical equilibrium, models for the chemical composition of Titan's atmosphere predict fractional abundances of ethane, C_2H_6, and ethyne, C_2H_2, that are negligible. However, if the effect of ultraviolet radiation is included, then the models predict fractional abundances of 10^{-5} and between 10^{-8} and 10^{-6}, respectively. These figures agree reasonably well with the observed fractional abundances in Table 5.1.

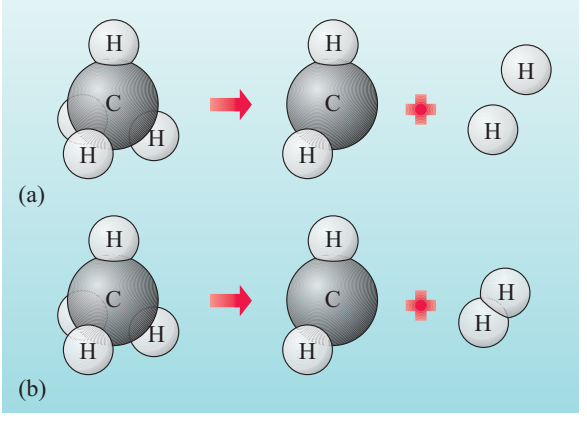


Figure 5.8 Schematic representation of the photodissociation of methane: (a) corresponds to Equation 5.4 and (b) corresponds to Equation 5.5.

How, then, can the observed variety of hydrocarbons be produced through the action of solar radiation on methane? The methane molecule has a central carbon atom joined to four hydrogen atoms, which can be thought of as lying at the corners of a tetrahedron (see Figure 5.8a). It can absorb ultraviolet radiation and break up or dissociate into smaller fragments in several ways, a process known as photodissociation. For example, one bond could acquire so much vibrational energy that the bond breaks and a hydrogen atom separates off. However, the C—H bonds do not necessarily vibrate in isolation, and two, three or four bonds can vibrate together. The most common photodissociations of methane in the atmospheres of Titan (and the giant planets) are those in which two hydrogens are lost. These are illustrated in Figure 5.8, and in Equations 5.4 and 5.5.

$$CH_4 + \text{photon} \longrightarrow CH_2 + H + H \qquad (5.4)$$

$$CH_4 + \text{photon} \longrightarrow CH_2 + H_2 \qquad (5.5)$$

A photon is a particle representing a basic unit (with a distinct energy) of visible light or other electromagnetic radiation.

In the atmospheres of the giant planets the most abundant molecule is H_2. This reacts with the carbon-containing product of Equations 5.4 and 5.5, methene, CH_2. The two main products of this reaction are methane and CH_3 (methyl). Since the methyl molecule is not bonded to four hydrogen atoms it has a spare electron in its outer shell (note the molecule still has an equal number of protons and electrons so it carries no charge). Were it to make a chemical bond this electron would be paired-off with an electron from another atom. However, since it is unpaired, this makes the methyl molecule highly reactive.

Molecules with no charge but which have an unpaired electron that can take part in forming chemical bonds are known as **radicals** and they are common products of photochemical reactions in which a bond is broken.

The formation of methane, of course, just takes us back to the starting material, however, the two methyl radicals will readily combine to form ethane, C_2H_6, as in Equation 5.6:

$$CH_3 + CH_3 + M \rightarrow C_2H_6 + M \qquad (5.6)$$

where M indicates a third molecule that does not take part chemically in the reaction but removes some of the energy of the reacting molecules.

A particularly interesting set of photochemical reactions occurs when methane loses three hydrogens to form the highly reactive radical CH (Figure 5.9). This happens about 8% of the time when methane absorbs ultraviolet radiation; it's an important reaction as it leads ultimately to the production of hydrocarbons containing long chains of carbon atoms.

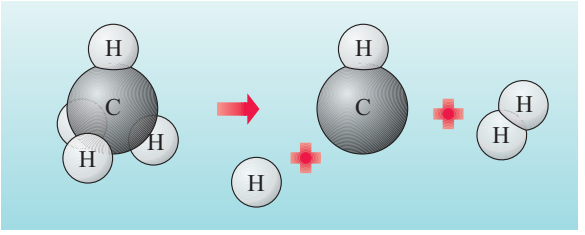


Figure 5.9 Schematic representation of the photodissociation of methane to form CH.

The CH radical produced by the loss of three hydrogen atoms reacts with methane to form a molecule in which two carbon atoms are bonded together, ethene, C_2H_4:

$$CH + CH_4 \longrightarrow H_2C{=}CH_2 + H \qquad (5.7)$$

The ethene produced readily absorbs ultraviolet radiation and loses hydrogen to form ethyne (HC≡CH) almost as soon as it is produced and so there is very little of it around. The observed abundance of ethene on Titan (and Jupiter) is lower than those of ethane ($H_3C{-}CH_3$) and ethyne.

C—C, C=C, and C≡C denote carbon atoms bound together by single, double and triple bonds respectively.

On Jupiter, ethyne is quite an abundant molecule in the atmosphere since when it is broken up by radiation, the products rapidly react with hydrogen to re-form ethyne:

$$HC{\equiv}CH + \text{photon} \longrightarrow HC{\equiv}C + H \qquad (5.8)$$

$$HC{\equiv}C + H_2 \longrightarrow HC{\equiv}CH + H \qquad (5.9)$$

■ Are these reactions (Equations 5.8 and 5.9) likely to happen on Titan?

❑ No. The major constituent of Titan's atmosphere is not hydrogen but the less reactive gas nitrogen.

Thus on Titan HC≡C does not re-form ethyne, instead it goes on to produce larger molecules. One of the simplest ways it can do this is illustrated by Equations 5.10 and 5.11.

$$HC{\equiv}C + HC{\equiv}CH \longrightarrow HC{\equiv}C{-}C{\equiv}CH + H \qquad (5.10)$$

$$HC{\equiv}C + HC{\equiv}C{-}C{\equiv}CH \longrightarrow HC{\equiv}C{-}C{\equiv}C{-}C{\equiv}CH + H \qquad (5.11)$$

Successive reactions of this sort can produce very long molecules with hundreds of linked carbon atoms.

■ Write an equation for the formation of the next molecule in the series after HC≡C—C≡C—C≡CH (i.e. one with eight carbon atoms).

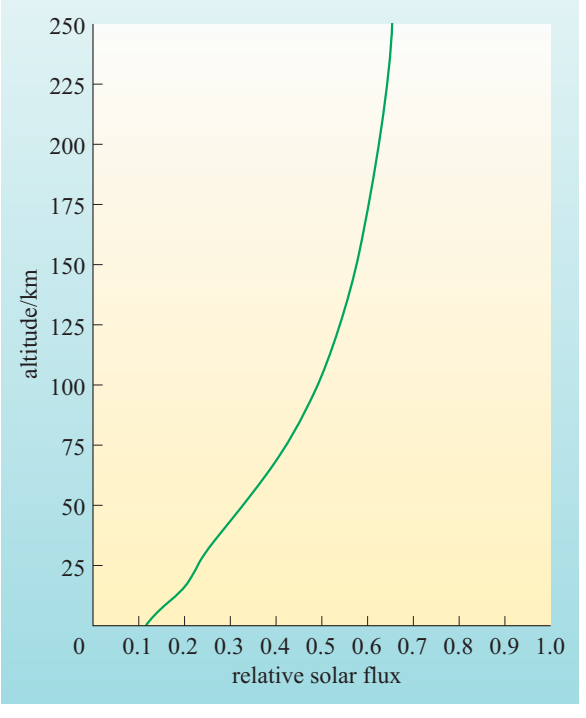
❑ $HC{\equiv}C + HC{\equiv}C{-}C{\equiv}C{-}C{\equiv}CH \longrightarrow HC{\equiv}C{-}C{\equiv}C{-}C{\equiv}C{-}C{\equiv}CH + H$ (5.12)

It is not known how far the production of carbon chains may have proceeded on Titan, but it is possible that some of the longer-chain molecules may be responsible for the haze observed in the atmosphere (see Section 5.3.3). The haze, therefore, is sometimes referred to as a photochemical haze or smog. However, the formation of very long chains may be inhibited by the presence of hydrogen atoms in the atmosphere, which would tend to promote the reverse reactions in Equations 5.10 and 5.11.

To obtain a detailed picture of the chemistry of Titan's atmosphere, we would need to consider all possible reactions. By building up our knowledge of which reactions are important, it is possible to construct a model of the major processes that take place. To do this, individual chemical reactions are studied in the laboratory to determine their rates and to deduce which are the important ones in the conditions that prevail in the particular atmosphere that is being modelled. In some cases, there are just a small number of reactions which dominate and which can describe fairly well the state of any particular atmosphere. In other cases, one has to consider a very large number of inter-related reactions in order to get a fair representation of the state of an atmosphere.

The photochemistry of methane is modified by the presence of large quantities of nitrogen on Titan (see Table 5.1). As nitrogen is relatively unreactive, it does not participate directly in Titan's atmospheric chemistry. However, N_2 molecules will dissociate under the action of solar photons, galactic cosmic rays and electrons from Saturn's magnetosphere to form atomic nitrogen, N. Sunlight is believed to be the dominant cause of the photodissociation of N_2 above an altitude of around 700 km in Titan's atmosphere. At lower altitudes dissociation by galactic cosmic rays is thought to be more significant with electrons from Saturn's magnetosphere also playing a role in the altitude range of 500 km to 750 km. The nitrogen atoms, however they are formed, do have an effect on the methane photochemistry already described as well as playing a role in the production of **nitriles**, a class of organic compound containing the group CN.

- Why should photodissociation be a more significant process above 700 km than below it?
- The strength of sunlight, which drives photochemical reactions, increases with greater height as solar radiation is absorbed in an atmosphere.



This is illustrated in Figure 5.10 which shows a prediction of how the intensity of solar radiation varies with height above Titan's surface. This suggests that the intensity near the surface is perhaps one-seventh of that at high altitude. This will clearly have a large effect on the rate of photochemical reactions – the rates are likely to be far higher in the stratosphere than at lower levels.

Following the acquisition of detailed compositional information on Titan's atmosphere by the Voyager fly-bys in the early 1980s, a series of increasingly sophisticated chemical models, based on the processes discussed above have been developed. To date, the most complete chemical model has considered a total of 122 reactions and 37 dissociation processes. These models have to take account of variations with height in an atmosphere. Factors that vary with height include the temperature and density, chemical abundances and the strength of sunlight. The most recent model for Titan's

Figure 5.10 The predicted variation of relative solar flux with altitude above Titan's surface. Relative solar flux is used to indicate that full allowance has been made at all altitudes for the amount of solar radiation that is absorbed, scattered and re-radiated.

atmosphere has had some success in describing the observed properties, however difficulties still remain that will require further observations and analyses.

The atmospheric models also provided some interesting insights into what has become known as the 'carbon monoxide problem'. Carbon monoxide, was not detected by either of the Voyagers. However, scientists had expected to find it since it is a common species and because it can be produced from the photolysis of CO_2 in the upper atmosphere. Carbon dioxide was detected by Voyager 1 in infrared spectra. Subsequently, CO was detected in 1983 in the near-infrared from ground-based observations. Carbon monoxide and carbon dioxide were then the only oxygen-bearing gases known in Titan's atmosphere. However, the observed abundances of CO and CO_2 imply that photodissociation of CO_2 is not the only process by which CO is produced in Titan's atmosphere. Carbon monoxide can also be produced by the reaction of the products resulting from the dissociation of water and methane molecules.

Photolysis or photodissociation refers to chemical reactions produced by exposure to light.

The water could be derived from icy particles from comets or from Saturn's ring system. The potential role of water in Titan's atmospheric chemistry was confirmed in 1998 when the European Space Agency's (ESA) Earth-orbiting Infrared Satellite Observatory (ISO) detected water. The superior spectral resolution of ISO's infrared spectrometer gave spectra with considerably more detail (Figure 5.14) than had been observed by Voyager. This enabled some of the fine spectral features characteristic of water to be detected. Even with the new data from ISO there still does not seem to be enough water in Titan's atmosphere to explain the observed amounts of CO and CO_2. This discrepancy will probably have to wait for its solution until the arrival of the Cassini–Huygens mission to the Saturnian system (Box 5.2).

BOX 5.2 THE CASSINI–HUYGENS MISSION

The Cassini–Huygens project is collaboration between The European Space Agency (ESA), NASA and the Italian Space Agency (ASI). The aim of the project is to place a spacecraft in orbit around Saturn and to deliver a probe to the surface of Titan. The former is the Cassini Orbiter provided by NASA and the latter is the Huygens Probe, the contribution from ESA. ASI is responsible for the spacecraft's 4 metre high-gain radio antenna and part of the communications system. The Orbiter is named after the Italian astronomer Jean-Dominique Cassini (Figure 5.11).

Cassini–Huygens was launched on 15 October 1997 by a Titan IVB/Centaur launch vehicle, and became the heaviest spacecraft (about 5630 kg at launch) to be launched towards the outer Solar System.

Figure 5.11 Jean-Dominique Cassini (born Giovanni Domenico) (1625–1712), the first of four generations of Italian astronomers who served as Director of the Paris Observatory. He discovered four Saturnian satellites (Iapetus, Rhea, Dione and Tethys) and in 1675 the distinct gap in Saturn's ring system, which now bears his name. The spacecraft (and mission) which will orbit the Saturnian system has been named in his honour.

Using the currently available propulsion technology, it isn't possible to launch a spacecraft of this mass directly to Saturn. Instead, a series of gravity assists had to be used. This technique, first employed by the Mariner 10 spacecraft at Venus in 1973 (which allowed it to reach Mercury), involves flying a spacecraft close to a planetary body in order for the spacecraft to gain energy. This requires the spacecraft to fly past the planet at just the right distance and angle. If successful, the technique reduces the launch energy requirements for a particular mission and enables otherwise impossible missions to be undertaken; it can also provide the means to carry a much larger payload than would otherwise be possible. In the case of Cassini–Huygens, four gravity assist manoeuvres have been used, at Venus (twice), Earth and Jupiter (Figure 5.12). These gravity assists have increased the spacecraft's velocity relative to the Sun by about 20 km s^{-1} and the spacecraft is due to reach Saturn in July 2004, releasing the Huygens Probe into Titan's atmosphere in December 2004.

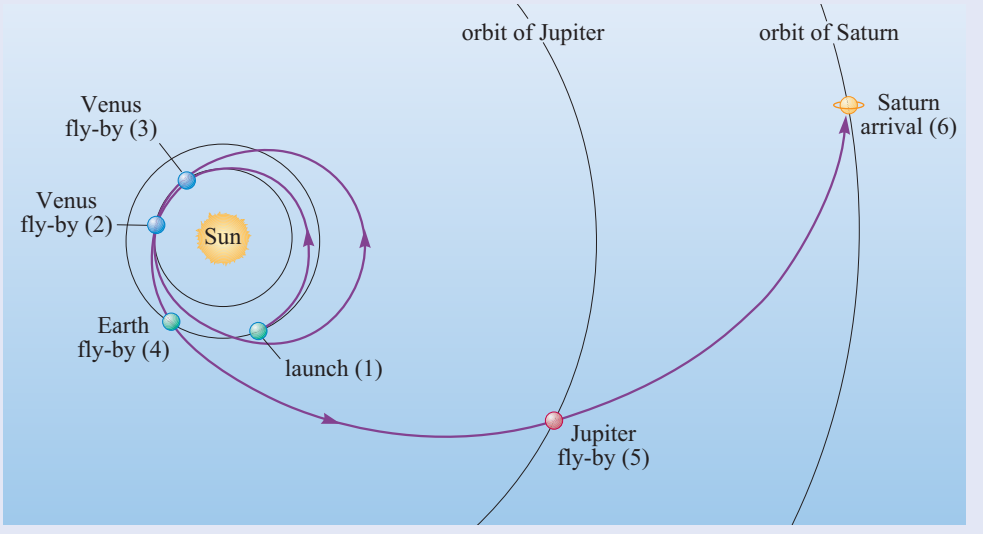


Figure 5.12 The Cassini–Huygens trajectory showing the four gravity assist manoeuvres.

Gravity assists are also known as planetary swing-bys or fly-bys.

Cassini–Huygens represents the second phase of the exploration of Saturn and its system. The first phase was characterized by Pioneer 11 and Voyagers 1 and 2, spacecraft which flew by Saturn giving us a snapshot of that environment. However, the Pioneer 11 and Voyager 2 fly-bys of Titan were relatively distant, limiting the resolution of the observations. In addition, any process or phenomenon that is time varying would probably have been missed. Cassini–Huygens, by remaining in orbit around Saturn and staying in close proximity to its targets will enable series of observations to be made and, based on the results, follow-up observations planned.

The Cassini–Huygens mission has five main scientific study objectives:

- Saturn itself,
- the extensive magnetosphere of Saturn,
- the Saturn ring system,
- the icy satellites,
- Titan.

These are addressed by a combination of the instrument package on the Cassini Orbiter, which comprises 12 scientific instruments, and the Huygens Probe with its payload complement of 6 instruments whose sole function is the investigation of Titan.

The Huygens Probe investigation of Titan has five aims that were agreed between NASA and ESA at the mission's planning stage. These are:

1 Determine the abundance of atmospheric constituents (including any noble gases), establish isotope ratios for abundant elements and constrain scenarios of the formation and evolution of Titan and its atmosphere.

2 Observe the vertical and horizontal distributions of trace gases; search for more complex organic molecules; investigate energy sources for atmospheric chemistry; model the photochemistry of the stratosphere and study the formation and composition of aerosols.

3 Measure the winds and the global temperature; investigate cloud physics, general circulation and seasonal effects in Titan's atmosphere and search for lightning discharges.

4 Determine the physical state, topography and composition of the surface; infer the internal structure of the satellite.

5 Investigate the upper atmosphere, its ionization and its role as a source of neutral and ionized material for the magnetosphere of Saturn.

The 318 kg Huygens Probe is shown in Figure 5.13. The front consists of a shield covered with thermal tiles to protect the probe from the heat generated during the high-speed entry into Titan's atmosphere. The back cover also provides thermal protection as well as housing the parachute compartment.

Figure 5.13 The Huygens Probe during final assembly. The 2.7 m front shield is clearly visible.

Both front shield and back cover are jettisoned during the upper part of the descent to leave an inner kernel containing the experiment platform (with an experiment payload mass of 48 kg) to descend to the surface (Table 5.6).

Table 5.6 The Huygens Probe scientific instruments.

Instrument	Acronym	Purpose
Aerosol Collector and Pyrolyser	ACP	Collects aerosol particles by deploying an extendable device into the airflow around the probe at two different altitudes. The collected particles are heated and the products will be passed to the GCMS (see below) for analysis.
Descent Imager and Spectral Radiometer	DISR	A collection of instruments to take both images and spectra of Titan's atmosphere and surface.
Doppler Wind Experiment	DWE	An experiment that uses equipment on both the Huygens Probe and Cassini Orbiter to provide information on the Probe's motion due to wind and turbulence.
Gas Chromatograph/ Mass Spectrometer	GCMS	Measures the chemical composition and determines the isotope ratios of the major gaseous constituents from 170 km altitude down to the surface.
Huygens Atmospheric Structure Instrument	HASI	Measures a wide range of physical properties of the atmosphere, including temperature and pressure profiles; wind speeds and turbulence; atmospheric conductivity; surface permittivity and radar reflectivity. It will also try to detect lightning.
Surface Science Package	SSP	Instruments, designed to study the surface in the region of the Probe landing site. Parameters to be measured include temperature, thermal conductivity, mechanical strength and the speed of sound. In the case of a liquid landing, liquid depth, density and surface wave properties will also be measured.

During Cassini's nominal four-year mission, it will carry out at least 40 fly-bys of Titan, mostly at altitudes of less than 2500 km. During these encounters, many of Cassini's instruments will be directed towards Titan. Of special note are various radar instruments. These include an altimeter capable of mapping the topography to a height precision of 150 m which will cover about 50% of the globe at 25 km spatial resolution by combining information from multiple fly-bys, and *synthetic aperture radar* that will produce radar images of the surface at 350 m to 1.7 km spatial resolution.

An important aspect of the photochemical models of Titan's atmosphere arose when the fate of some of the products of the various chemical pathways was considered. The models suggested that some of the products may well be solids or liquids. This raised the interesting possibility that the surface of Titan would be subjected to a steady rain or snowfall of a variety of organic materials including more complex macromolecular substances. Table 5.3 shows the main products of the photolysis of methane and carbon monoxide and their predicted downward fluxes. Of these, ethane is liquid at the temperature and pressure of Titan's surface while ethyne is solid. Even more intriguing is to consider what would be the accumulated effect of

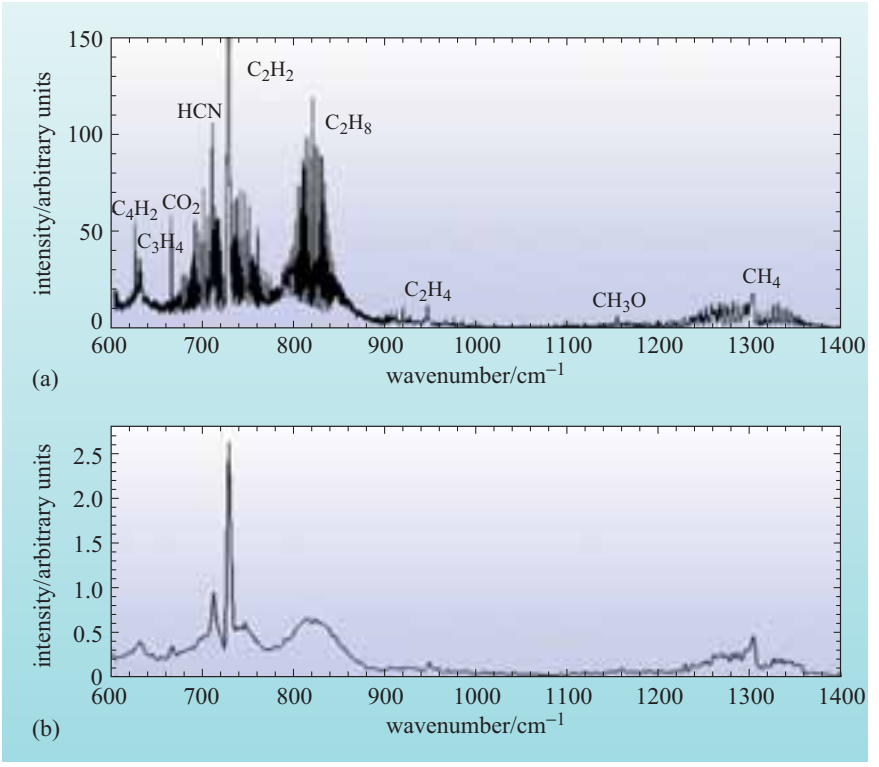


Figure 5.14 The same part of Titan's infrared spectrum observed by (a) ESA's ISO satellite and (b) Voyager infrared spectrometer. The superior spectral resolution of the former is shown by the far greater detail discernible.

these fluxes over the age of the Solar System. The last column in Table 5.3 shows the depth of each individual product that would accumulate over the age of the Solar System. This reveals that perhaps 600 m of liquid ethane might have collected on the surface. However, you should note that these depth estimates apply only on a 'billiard ball' type Titan; if there is any topography, the depth would be variable and the entire globe need not be submerged. We will return to these intriguing possibilities later when we look in detail at the nature of Titan's surface.

Table 5.3 Predicted downward fluxes of products of methane and carbon monoxide photolysis and their accumulated depths over the age of the Solar System.

Species	Flux (mols m^{-2} s^{-1})	Depth (km)
C_2H_6	5.8×10^{13}	0.6
C_2H_2	1.2×10^{13}	0.1
C_3H_8	1.4×10^{12}	0.02
CH_3C_2H	5.7×10^{11}	0.006
HCN	2.0×10^{12}	0.02
HC_3N	1.7×10^{11}	0.002
C_2N_2	6.0×10^{10}	0.001
CO_2	3×10^{9}	2×10^{-5}

5.3.2 The origin of nitrogen in Titan's atmosphere

You have already seen that Titan, like the Earth, has an atmosphere whose main constituent is nitrogen as N_2. Titan's size and rocky interior resemble the terrestrial planets more than the giant planets, but the types of molecule (apart from N_2) found in the atmosphere cause the atmospheric chemistry to resemble that on Jupiter rather than that on the Earth.

There are two possible explanations for the origin of a nitrogen atmosphere on Titan. The first is that when Titan formed, the very low temperature of that part of the solar nebula caused the nitrogen to become trapped as a **clathrate** in the icy layers as the satellite formed. A clathrate is produced when a substance, e.g. H_2O, forms with an open crystal structure that can admit small gas molecules, such as CO_2 or N_2, which then become trapped in the cavities. These small molecules can be held in the cavities until the crystalline substance is warmed or perhaps melted at which point they are released, for example by radiogenic or tidal heating. This model suggests that the nitrogen is present as N_2 because it was trapped as N_2 when the satellite was formed. Any ammonia present remained as a solid ice and was never released into the atmosphere.

A second model assumes that the nitrogen was initially present as ammonia, and that subsequent reactions converted this into nitrogen (Equation 5.13).

$$2NH_3 \longrightarrow N_2 + 3H_2 \qquad (5.13)$$

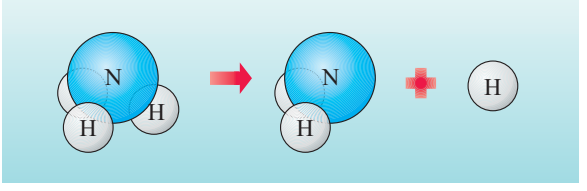


Figure 5.15 Schematic representation of the photodissociation of an ammonia molecule.

However, this reaction would be very slow at the temperatures on Titan and, if the atmosphere were at chemical equilibrium, there should be substantial amounts of ammonia present. One reason for the lack of ammonia in Titan's atmosphere becomes apparent when we consider the photodissociation of ammonia by sunlight. One way in which this can occur is illustrated in Figure 5.15. Here, the energy absorbed is channelled into one of the N—H bonds causing it to break.

The initial products of photodissociation are hydrogen atoms, which escape from Titan, and the radical NH_2, which rapidly reacts with another NH_2 radical to form the molecule **hydrazine**, N_2H_4, as in Equation 5.14:

$$H_2N + NH_2 + M \longrightarrow H_2NNH_2 + M \qquad (5.14)$$

in which M represents a molecule that participates in the reaction collision, but emerges unscathed chemically.

Generally when we write chemical formulae, we group together all like atoms as in N_2H_4, but sometimes it can make the structure of the molecule clearer if we write out the formula in its constituent parts. Writing N_2H_4 as H_2NNH_2 tells us for example that hydrazine consists of two NH_2 fragments joined by a bond between the two nitrogens. This sort of representation is particularly used for compounds of carbon.

Similar ammonia photochemistry occurs in the present atmosphere of Jupiter. However, there are two important differences between Jupiter's photochemistry and the proposed route to N_2 on Titan.

First, at the temperatures and pressures in the layers of Jupiter's atmosphere where photodissociation occurs, hydrazine condenses out to form a haze. This haze does not undergo any further chemical reactions. However, on Titan, it has been suggested that the hydrazine remained in the gas phase when Titan's atmosphere was being formed. For this to happen it would be necessary for the surface temperature of Titan to be some 50 K higher than it is at present. Hence the model must include some mechanism for raising the surface temperature.

Second, hydrogen is more readily lost from Titan than from Jupiter owing to the lower escape velocity on Titan. Consequently, hydrogen is more abundant on Jupiter

so that the NH_2 recombines with hydrogen, to re-form ammonia. However, on Titan, gaseous hydrazine will undergo a series of photodissociation reactions that eventually leads to the production of N_2H_2, a molecule that will break down to nitrogen and hydrogen:

$$N_2H_2 \longrightarrow N_2 + H_2 \qquad (5.15)$$

The hydrogen escapes into space, leaving nitrogen in the atmosphere.

Laboratory experiments to measure the efficiency of trapping gases in clathrates at around 75 K, indicate that argon and nitrogen (in the form of N_2) are trapped with about the same efficiency. So one way in which we might be able to determine the origin of Titan's nitrogen is to measure the abundance of argon. Voyager 1 was unable to detect argon in Titan's atmosphere since argon doesn't have a spectral line in the infrared part of the spectrum covered by the Voyager IRIS instrument. Any argon on Titan almost certainly came from the nebula out of which Saturn and Titan were formed. Thus, if clathrates on Titan trapped both argon and nitrogen from the nebula we would expect a certain ratio of Ar/N_2 (about 0.06), whereas if the nitrogen was produced from the photodissociation of ammonia, we would expect a much lower ratio. The Ar/N_2 ratio is therefore an important measurement for the Cassini–Huygens mission (Box 5.2).

5.3.3 Aerosols

The composition of the particles that make up Titan's haze has been the subject of a series of laboratory-based experiments. In these, a mixture of nitrogen and methane gas in a glass vessel is subjected to an electric discharge in order to simulate the interaction of sunlight, cosmic rays and electrons with Titan's atmosphere. Typically, after a period of several months, a thin layer of brown–orange 'goo' has formed on the inside of the reaction vessel. This so-called **tholin** (from the Greek word 'tholos' meaning mud), when subsequently analysed was found to contain over 75 different constituents, mainly hydrocarbons and nitriles. Although subjecting a gas mixture to an electrical discharge for a few weeks is not the same as irradiating Titan's atmosphere with UV photons and energetic particles for aeons, the results were, at least superficially, very encouraging and gave a qualitative explanation of the visual appearance of Titan (see Figure 5.2). In order to make a more quantitative comparison, it was necessary to compare some of the optical characteristics of the laboratory generated tholins with those of Titan's haze. This was done over wavelengths from ultraviolet through to infrared and showed reasonably good agreement between Titan haze and the laboratory generated particles suggesting that the proposed mechanism, namely the photochemical processing of a methane–nitrogen mix, is indeed the primary source of Titan's aerosols.

5.4 Modelling Titan's surface

Prior to the Voyager fly-bys in the early 1980s, there had not been a great deal of speculation concerning the nature of Titan's surface. However, data from Voyagers 1 and 2 enabled scientists to be far more quantitative about several aspects of Titan's atmosphere, and this had profound effects on our understanding of its surface. As you saw in Section 5.3.1, the abundance of gases in the atmosphere was determined to high degree of precision (Table 5.1). In addition, the radiation environment that Titan was subjected to was also determined quite accurately.

This fact was important because it enabled the rate of methane photodissociation (i.e. the rate at which methane is being destroyed) in Titan's atmosphere to be determined: the value turns out to be 4×10^{-12} kg m^{-2} s^{-1}.

■ Can you explain what the units for the rate of methane photodissociation (i.e. kg m^{-2} s^{-1}) mean?

❑ These units represent the mass of methane (i.e. kg) destroyed every second (i.e. s^{-1}) in a unit cross-section of the atmosphere (i.e. m^{-2}).

Since the methane in Titan's atmosphere is being destroyed by photodissociation then, unless it is replenished, it will eventually disappear. Scientists have estimated how long methane will remain before it's all lost by photolysis and get a figure of around 1 Ma. In astronomical terms this is a very short time and it presents us with a choice in interpreting this fact. We could say that we are seeing Titan at a rather special time in its evolution, namely the very small fraction of its entire lifetime when its atmosphere contains significant amounts of methane before it all disappears. However, astronomers don't feel comfortable with theories that suggest that we are in any sense in a privileged time or location. Therefore, accepting that methane is being lost at a high rate, it is preferable to suggest that there is a reservoir of methane that is able to replace continuously what is lost from the upper atmosphere. There are not too many choices for the location of this reservoir, the most obvious being on (or very near) the surface. The Voyager spacecraft were able to give us some information about this surface: they provided a fairly accurate measurement of both the temperature and pressure through the atmosphere right down to the surface. At the surface, the temperature was determined to be (94.0 ± 0.7) K and the pressure (1496 ± 20) mbar. With these data, as you saw in Chapter 3, we can determine in which phase (i.e. solid, liquid or gaseous) a substance exists through the use of a phase diagram (Box 3.3).

The conditions determined for the surface of Titan are very close to the **triple point** (see Box 3.3) of methane (90.7 K at 1.6 bar). This led to the remarkable suggestion that the reservoir of methane at the surface of Titan was in liquid form – namely a sea or ocean.

■ What other evidence have you met that suggests a liquid surface?

❑ A liquid surface was supported by one of the conclusions of the atmospheric models discussed in Section 5.3.1, namely that some of the products of the chemical processes occurring in the atmosphere were liquids.

The suggestion of an ocean was initially treated with scepticism, however the scientific community gradually realized that the arguments had a solid foundation and the idea of large bodies of surface liquid progressively took root. What was needed was a way to test this hypothesis. One test came from the temperature profiles through the atmosphere obtained by the Voyager spacecraft. If the atmosphere above a body of liquid is in equilibrium with the liquid, then it will be saturated with vapour and, consequently, the way in which the temperature varies with height should differ from an atmosphere that is not saturated. The temperature gradient for a saturated atmosphere is called the **wet adiabatic lapse rate**. An ocean, if it exists on Titan, is most likely composed of a mixture of methane and

ethane, with possibly some dissolved nitrogen also. The wet adiabatic lapse rate above such an ethane-rich ocean would have a predicted value of 1.4 K km^{-1}, consistent with the value observed from Voyager 1 temperature profiles of (1.38 ± 0.1) K km^{-1}.

Titan ocean models

Throughout the 1980s and 1990s, more work was done on possible ocean compositions and the solubility of various constituents in such an ocean. In addition, the data from several of the Voyager instruments were re-evaluated in the light of the improved understanding of Titan, resulting in a slightly different interpretation of the composition, temperature and pressure data. One result was the generation of a range of ocean models that were claimed to be consistent with the Titan atmospheric data. Properties of the two extreme ocean models are shown in Table 5.4.

Table 5.4 Composition and properties of the two extreme Titan ocean models.

	Ethane rich	Methane rich
Ethane and propane	90.9%	5%
Methane	7.3%	83.4%
Nitrogen	1.8%	6%
Argon	0	5.6%
Carbon monoxide	3.7×10^{-6}	9.2×10^{-6}
Hydrogen	9.0×10^{-7}	2.6×10^{-6}
Ocean temperature (K)	92.5	101
Surface air temperature (K)	93.1	100.6
Depth: Initial (km)	1.3	10.1
Depth: Current (km)	0.7	9.4
Duration of methane reservoir (Ma)	140	1000

The two extreme model oceans are a cold, ethane-rich ocean and a slightly warmer, methane-rich one. In the latter case, the greater methane inventory allows the ocean to resupply what is lost from the atmosphere by photolysis for a longer period (about 1 Ga as opposed to 140 Ma). The exact composition of the ocean would also have a significant effect on the quantity of heavier hydrocarbons that can be dissolved in it. The solid photolysis products collect in the ocean until they reach saturation and precipitate out, a process that is believed to happen quite rapidly, before accumulating on the ocean floor. For both of the models shown in Table 5.4, the main sediment is expected to be ethyne, which should accumulate to a depth of about 100 m. These models also suggest that the ocean may contain a mass of nitrogen comparable to that in the atmosphere – so it would be very significant in terms of the overall inventory balance on Titan.

■ How would you expect the amount of methane in the ocean to change over time?

❑ Unless there is another source of methane, one would expect methane to be gradually lost to the atmosphere and thus the ocean composition to evolve over time to become more ethane rich.

QUESTION 5.4

Assume that 50% of Titan's surface is covered with a methane–ethane–nitrogen ocean of average depth 0.5 km and density 0.66×10^3 kg m^{-3}. The ocean is made up of 70% methane, 25% ethane and 5% nitrogen *by mass*. Assuming that methane is lost from the atmosphere at a rate of 4×10^{-12} kg m^{-2} s^{-1}, calculate for how long the ocean can resupply the atmosphere. (State any assumptions you make.)

Winds and waves

There is little direct information on winds in Titan's atmosphere but indirect evidence suggests that they should exist. For example, Voyager 1 data showed that there is a temperature difference of some 15 K between the equator and latitudes 60°.

- Why should this fact imply the existence of atmospheric winds?
- When a significant temperature difference exists in a fluid, this tends to result in convection, which is the consequence of mass motion of the fluid, causing winds.

So if a significant body of liquid does exist on the surface of Titan, we can be reasonably confident that waves would be generated just as in terrestrial expanses of water. On Earth, the dominant force that restricts the growth of wind driven waves on an expanse of water is gravity. As Titan's gravity is only 15% of that on Earth, one might expect that waves on Titan's seas should grow to much greater heights under comparable conditions. In what was probably the first example of extraterrestrial oceanography, a group of scientists took the standard mathematical model which described the way in which waves are generated on Earth and changed all the input parameters such as gravity, liquid density and viscosity, to those which would be expected for oceans on Titan. The characteristics of the waves generated are shown in Table 5.5 with those under similar conditions for Earth shown for comparison.

Table 5.5 A comparison of wind-driven ocean waves for a fully developed sea[a] on Earth and Titan assuming a surface wind speed of 5 m s^{-1} which corresponds to a gentle to moderate breeze.

	Earth	Titan
Significant wave height[b] (m)	0.6	4.5
Wave speed (m s^{-1})	5.5	5.5
Wavelength (m)	11	105
Period (s)	3.5	11.5

[a] For a fully developed sea, there must be a stretch of open water of sufficient size in the direction that the wind is blowing so that the waves can build up to their maximum height. The length of this stretch of open water is called the 'fetch'. Typically on Earth a fully developed sea requires a fetch of 20 km for a wind speed of 1 m s^{-1} and as much as 200 km or more for a speed of 5 m s^{-1}.

[b] 'Significant wave height' is a term used by oceanographers to get around the fact that waves have a distribution of heights. It is close to the mean of the highest one-third of the waves present in a sea, and approximates visual estimates of wave height.

From Table 5.5, it is apparent that Titan's waves would be higher, more separated, of similar speed and thus less frequent when compared to waves on Earth. It may well transpire that the size of these waves leads to the demise of the Huygens Probe once it lands on the surface of Titan (Box 5.2). The Cassini–Huygens mission will have the capability to probe the surface, both directly and remotely, and will be able to tell us whether features such as wind driven ocean waves are a reality or simply the conjectures of theoretical planetary scientists.

Radar observations of Titan's surface

In 1990, the NASA communications dish at Goldstone, California was used to direct a radio signal at a wavelength of 3.5 cm. towards Titan while the Very Large Array telescope in New Mexico attempted to detect the echo or return signal. Although this was an extremely challenging task, they did manage to detect a reflected signal. This made Titan the most distant object from which a reflected radio signal had been received. Radiation at these wavelengths is relatively unaffected by the presence of haze or clouds in the atmosphere of Titan, yet the strength of the return signal is sensitive to the type of materials present on the surface from which it is reflected. The strength of the detected signal implied that the surface of Titan was reflecting about 10% of the incident radio signal. Note that because of the difficulty of the measurements, there is a fairly large uncertainty in this figure. This value was not consistent with Titan's surface being covered with global ocean of ethane–methane several hundred metres deep as this would be expected to reflect only about 2% of the incident radiation. Although this was not consistent with the idea of a global ocean, it was also at odds with Titan having an icy surface like the largest Jovian satellites, namely Europa, Ganymede and Callisto which reflect between 30% and 90% of the radiation incident upon them at radio wavelengths. Further radar observations have been carried out and these seem to support the earlier observations. However, there is some evidence that different regions of the surface have different radar characteristics, one area in particular appearing to be 'radar bright', implying that the surface of Titan is not uniform but varies in its properties. One interpretation is that the radar-bright region, which seems to cover a surface area approximately equal to that of Australia, is predominantly clean water-ice and the surrounding area is some form of hydrocarbon ocean. Overall, these observations show that Titan doesn't seem to match any other object observed at these wavelengths. In order to circumvent the apparent conflict between these measurements and the hypothesis of a deep global ocean, some theoreticians have tried to see if it is possible to modify this concept to make it consistent with the radar measurements. They have found that if the surface of the ocean were very frothy, perhaps as a result of breaking wind-driven waves as discussed earlier, or was 'dirty' as a result of floating solid products of atmospheric photolysis products, or even from dust from meteoritic impacts, then the radio reflectivity would be significantly higher and therefore perhaps consistent with the observations. This is sometimes referred to as the deep, dirty, frothy ocean model.

Titan's orbital eccentricity

Further insight on the possibility of Titan's oceans came from a theoretical consideration of Titan's orbital eccentricity (see Table 5.2) of 0.0292. A surface layer of liquid on Titan would manifest a tidal bulge just as occurs on Earth, but in the case of Titan this would be generated by Saturn's gravitational pull. This bulge generates tidal currents that act to dissipate Titan's orbital energy, causing the orbit to become circular over time. Calculations in the early 1980s suggested that if there

were an ocean of less than 400 m in depth, tidal friction would have caused the orbital eccentricity to have dissipated long ago and the orbit to have become circular. So their argument was that the ocean is either very deep (>400 m) or doesn't exist at all, although this wouldn't prevent the existence of small isolated lakes.

Surface imaging

When Titan is imaged at wavelengths other than those of visible light, in particular the infrared region, some interesting details begin to emerge. Figure 5.16 shows the variation of Titan's albedo with **wavenumber** in the near-infrared part of the electromagnetic spectrum.

Wavenumber = $1/\lambda$ where λ is the wavelength of the radiation.

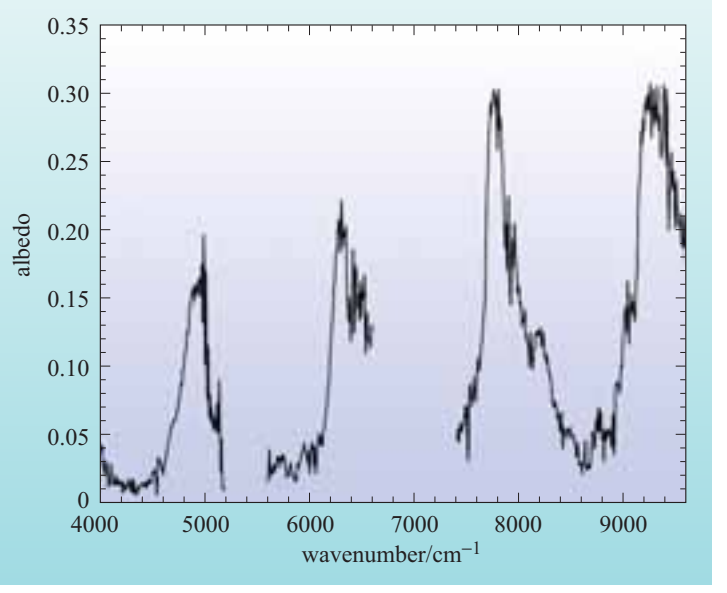


Figure 5.16 The variation of Titan's albedo with respect to wavenumber in the near-infrared part of the spectrum.

The troughs in Figure 5.16 have values close to zero (i.e. values less than 0.05). This means that at these wavelengths almost all the radiation incident on Titan is absorbed in the atmosphere. The peaks, which have values between 0.2 and 0.3, correspond to regions where a good proportion of the incident radiation is reflected back. So in these regions the atmosphere is at least partially transparent. These regions are termed *windows* because they enable us to 'see' through the atmosphere. This region of the spectrum has been investigated using ground-based telescopes and also by the Hubble Space Telescope (HST). The observations are generally in agreement and appear to show variations in the albedo during one axial rotation of Titan (i.e. over a period of 16 days). Significantly, these same variations have been observed over several different axial rotations.

■ Why is the consistency of the variations from different rotations significant?

❑ It implies that the variations are genuinely due to surface property differences because if they were due to cloud or haze variations, they wouldn't be expected to be long-lived (i.e. over many Titan rotations). However, if they were due to real surface variations, the opposite would be the case.

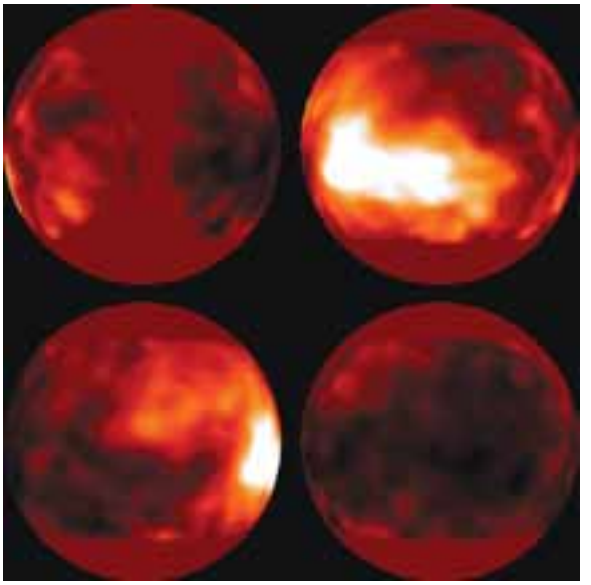


Figure 5.17 Four projections of Titan's surface (see text) determined from observations with the Hubble Space Telescope between 4 and 18 October 1994. The images were taken in one of the postulated atmospheric 'windows' in Titan's atmosphere in the near-infrared part of the spectrum. The top left image is of the area of Titan facing Saturn, and each subsequent image represents a region rotated by 90° from the previous one. The difference between the brightest and darkest regions represents only 10% of the total light collected but appears to be a real effect.

These observations have enabled the production of the first crude surface maps of Titan. Figure 5.17 shows four views of Titan from 14 images acquired by the HST in October 1994 in one of these near-infrared atmospheric windows. They represent views of different faces of Titan. The 'false colours' represent different amounts of near-infrared radiation reflected from different parts of the surface. There has been much speculation as to what physical reality they correspond to on the surface. We almost certainly won't know until the arrival of the Cassini–Huygens mission. One intriguing possibility is that the 'bright' regions represent a fairly clean icy surface, while the dominant 'dark' area corresponds to the hydrocarbon seas predicted by some.

5.5 Modelling Titan's interior

Without any direct contact with the surface of Titan, there has been no way of 'sounding' the interior of Titan. Indeed, there is only one strong constraint on the interior structure and composition of Titan.

- Can you suggest what this constraint is? (*Hint*: look at Table 5.2.)
- The mean density (in this case $1.88 \times 10^3\,\mathrm{kg\,m^{-3}}$) constrains the materials that can make up Titan's interior.

From solid bodies whose structure we understand a lot better, we know that most such bodies are made predominantly of icy and rocky material. To get an overall mean density of $1.88 \times 10^3\,\mathrm{kg\,m^{-3}}$, we would expect a dominance of icy rather than rocky material.

- If Titan is made up of such a mixture, what can you say about how this material would be distributed?
- We would expect the denser component to have differentiated to the centre (of perhaps a primitive molten body) with an overlaying layer of ice.

Any further progress, until we have more direct measurements from Titan, must rely on modelling of its structure and, in particular, on possible formation scenarios for Titan. There are several alternatives but the most favoured are a series of evolutionary models in which Titan formed inside a protoSaturn nebula, a flattened disc of gas, rich in methane and ammonia, and dust, encircling the protoplanet during its early stages of contraction. Titan would have been a hot object in the early stages after its formation, and more uniform in its composition. As it cooled, the heavy, rocky material would tend to fall to the centre, leaving the lighter materials, mostly water with an estimated 15% by weight of ammonia in solution, to form the outer core. Irradiation of the atmosphere and surface by solar ultraviolet radiation, cosmic rays and charged particles from Saturn's magnetosphere could have dissociated enough ammonia to form a thick nitrogen atmosphere, a precursor of the present situation.

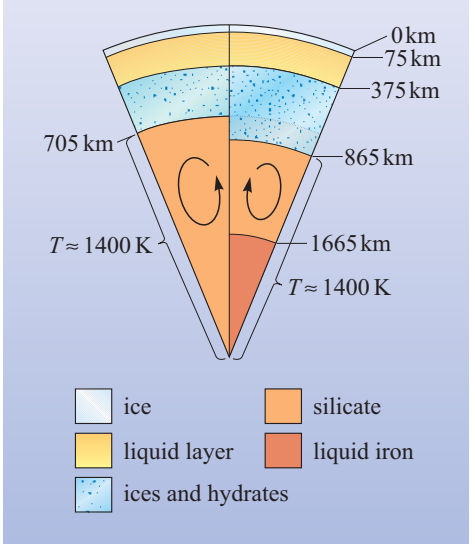


Figure 5.18 Model of a hypothesized Titan interior, with or without a liquid iron core. The location of a possible liquid layer is indicated; the arrows show convection in the icy layers. Distances shown are distances from the surface.

In more detailed interior models, the ice layers outside the rocky core have extents, locations and crystal structures that depend on the distribution of temperature and pressure. They also contain methane and ammonia clathrates, in which these molecules are trapped inside the water ice lattice, and perhaps other ices. The heavier core may itself be differentiated, with, for example, iron and sulfur compounds and other metals at the centre, overlain by the silicates and other rocky material. Figure 5.18 shows two of the possible interior models.

It seems possible, then, though highly speculative, that a liquid layer, perhaps as thick as 350 km, could exist some 75 km below the surface. The temperature in the liquid layer, if it is really present, is probably in the range 220 K to 250 K.

What of the composition of this subsurface ocean? Methane and nitrogen could be present. By analogy with Europa (see Chapter 4), it has been suggested that salty liquid water might be the main component.

- If the latter were the case, how might this be detectable by a fly-by spacecraft?

- Water, especially salty water, is electrically conducting. A conductor in motion in a magnetic field will produce a magnetic field of its own. Therefore, a conducting ocean passing through a giant planet's magnetic field will do just that. This can than be detected by a spacecraft-borne magnetometer, as has possibly occurred with the Galileo spacecraft at Europa (see Section 4.2).

Another possibility is that ammonia is the main constituent, while nitrogen–water, methane–water or ammonia–methane–water mixtures have also all been proposed. Unfortunately, the exact state of affairs is impossible to determine because of the lack of observational data. This is compounded by the lack of data on the properties of suitable high-pressure, low-temperature liquid mixtures. We have already seen how difficult it is to speculate concerning the nature of the possible subsurface liquid on Europa. But at least in the case of Europa, we have excellent and extensive data on the nature of its surface. By comparison, in the case of Titan, we know almost nothing.

QUESTION 5.5

Calculate (a) the closest and (b) the furthest possible distance from Earth to Saturn. How do you explain therefore that the Cassini spacecraft has covered a distance of about 3.2×10^9 km by the time it has reached Saturn?

QUESTION 5.6

Imagine that after the Huygens Probe arrives at Titan in 2005, the following data are obtained. In each case interpret these findings.

(a) The average abundance of argon in the atmosphere is found to be 0.05%.

(b) The Probe lands on a body of liquid in which the significant wave height is measured to be about 3 m.

(c) The strength of sunlight at the surface is found to be about one-thousandth of that at Earth's surface.

5.6 Summary of Chapter 5

- Titan is a planet-sized body with a thick, rich and complex atmosphere within which occurs a wide array of chemical reactions.
- Titan's surface is essentially unexplored by direct means but indirect evidence points to the possibility of extensive reservoirs of liquid methane and ethane, possibly even in the form of hydrocarbon seas or lakes.
- Knowledge of Titan's interior structure is very scant but there exists the possibility of a deep layer of liquid water below an ice crust. There is therefore a remote possibility of life but the lack of oxygen and low temperatures make this seem unlikely.
- Titan will be visited by the Cassini–Huygens space mission which aims to resolve many of the unknown observational features and thus give us an understanding of why Titan, alone among the planetary satellites, possesses a dense atmosphere.

ANSWERS AND COMMENTS

QUESTION 1.1

Noble gas elements are insignificant constituents of living organisms as they are unreactive and difficult to combine into molecules by biological activity. It is for this reason that they are absent from the 'humans' column in Table 1.1.

QUESTION 1.2

(a) **Table 1.7** Accretion rates on Earth today (completed).

Sources	Mass range /kg	Mass accretion rate (estimated) /10^6 kg yr^{-1}	Carbon %	Carbon accretion rate /10^6 kg yr^{-1}
meteoritic matter				
meteors (from comets)	10^{-17} to 10^{-1}	16.0	10.0	1.6
meteorites	10^{-2} to 10^{5}	0.058	1.3	7.5×10^{-4}
crater-forming bodies	10^{5} to 10^{15}	62.0	4.2	2.6
unmelted material contributing organic matter				
meteors (from comets)	10^{-15} to 10^{-9}	3.2	10.0	0.32
meteorites, non-carbonaceous	10^{-2} to 10^{5}	2.9×10^{-3}	0.1	2.9×10^{-6}
meteorites, carbonaceous	10^{-2} to 10^{5}	1.9×10^{-4}	2.5	4.7×10^{-6}

Note: carbon accretion rate = mass accretion rate $\times \dfrac{\text{carbon \%}}{100}$

(b) The greatest source of meteoritic carbon is crater-forming bodies. These objects are unlikely to arrive steadily over time.

(c) For organic matter the greatest source of carbon is meteors from comets. These objects will be arriving on the Earth relatively constantly.

(d) The accretion rate for total meteor carbon is much larger than that for meteor organic carbon. The carbon in meteors must be mainly inorganic.

(e) 42×10^6 kg of meteoritic carbon arrives in 10 years; 420×10^6 kg in 100 years and $420\,000 \times 10^6$ kg arrives in 100 000 years.

(f) 3.2×10^6 kg of meteoritic organic carbon arrives in 10 years; 32×10^6 kg in 100 years and $32\,000 \times 10^6$ kg in 100 000 years.

QUESTION 1.3

The carbon in the current biosphere is 6×10^{14} kg. At present-day rates, meteoritic materials would supply:

(a) a similar amount of carbon in

$$\frac{6 \times 10^{14} \text{ kg (biosphere carbon)}}{4.2 \times 10^{6} \text{ kg (meteoritic carbon yr}^{-1}\text{)}} = 1.43 \times 10^{8} \text{ yr.}$$

(b) a similar amount of organic carbon in

$$\frac{6 \times 10^{14} \text{ kg (biosphere carbon)}}{0.32 \times 10^{6} \text{ kg (meteoritic organic carbon yr}^{-1}\text{)}} = 1.87 \times 10^{9} \text{ yr.}$$

Rates of meteoritic infall would have been much higher earlier in the Earth's history.

QUESTION 2.1

The star is 10 000 times as luminous as the Sun. The effective temperature will therefore increase by the fourth root of 10 000 or 10. This implies an effective temperature of 2550 K. This is about 2300 °C.

QUESTION 2.2

To maintain the same temperature on Earth, we would have to move our planet out to a distance square root of 10 000 times larger, that is, to a distance of 100 AU from the Sun. This distance is about 2.5 times further out than the orbit of planet Pluto.

QUESTION 2.3

The reason that Mars is presently cold is that controlling the level of CO_2 in the atmosphere through the weathering of silicates (Box 2.2), and thereby regulating climate, requires both a substantial inventory of carbonate rocks and some mechanism for recycling them to CO_2. Even if carbonate rocks are abundant on the surface of Mars (in early 2003 they have not yet been identified spectroscopically by spacecraft or observed by landers), there is apparently no mechanism present to recycle them to CO_2. As you saw in Box 2.2, on Earth this process, termed decarbonation, occurs when oceanic carbonates are subducted, resulting in the decomposition of carbonate and the eventual release of CO_2 back to the atmosphere. Mars, being a small planet, has a cooler interior than Earth and shows no signs of global tectonic activity or recent volcanism.

QUESTION 2.4

The total mass of BIFs older than 2.5 Ga is 3.3×10^{16} kg. Of this mass, 30% is Fe_2O_3, thus $3.3 \times 10^{16} \times 0.3$ kg is Fe_2O_3, which is 9.9×10^{15} kg.

The relative atomic masses of oxygen and iron are 16 and 56, respectively, so the relative molecular mass of Fe_2O_3 is $(2 \times 56) + (3 \times 16) = 160$. Therefore the amount of oxygen incorporated in the BIF deposits is $(48/160) \times 9.9 \times 10^{15}$ kg = 2.97×10^{15} kg.

QUESTION 2.5

If we look at both the geological information you examined in this chapter and biochemical information you examined in Chapter 1, then several key points have emerged that suggest a possible scenario for the emergence of life on Earth:

The oldest rocks so far examined were apparently deposited under water, suggesting the presence of oceans. They contain sedimentary rocks, indicating that the processes of weathering and erosion must have been active. Thus the earliest geological record supports the idea that familiar geological and geochemical processes were operating extremely early in Earth history.

Models of planetary accretion, differentiation and mantle convection suggest that plate tectonics was operating on the early Earth and that up to five times more internal heat was being produced. The Earth's early atmosphere appears to have been composed primarily of CO_2, N_2 and water vapour.

Hydrothermal systems are a key environment that can provide energy to thermophilic and hyperthermophilic organisms that populate the deepest and shortest branches of the phylogenetic tree.

Thus, it seems increasingly difficult to find support for the idea you first met in Sections 1.5 and 1.8.2 that it was the input of external sources of energy into a reduced atmosphere that created a 'prebiotic soup' from which the first organism appeared as a heterotroph. Instead, we have a scenario for the emergence of life that includes internal forms of geochemical energy that result in the formation of environments in which autotrophic reductive metabolisms are nurtured.

QUESTION 2.6

It appears that there are significant ways in which conditions on the early Earth affected the potential for the emergence of life in hydrothermal systems: the higher magnitude of heat flow, a hydrosphere, and upper mantle and crustal rocks that could host hydrothermal systems. Taken together, these apparent differences between the present and the early Earth suggest that conditions in hydrothermal systems on the early Earth were more favourable to the emergence of life than they are at present.

QUESTION 2.7

Some extrapolations would seem to have a basis from the overall trends in evolution observed on Earth. If we assume that life elsewhere has a cellular basis, then increases in the size of organisms, their complexity and diversity from some initial starting point would seem likely to occur. However, we should not forget that for the first 3 Ga of life on Earth there were very few large species. As you'll see in Chapter 9, this has a significant effect on the probability of intelligent life elsewhere in the Universe.

QUESTION 2.8

Methanopyrus, and *Sulfolobus* occur close to the root of the archaea part of the tree, *Thermoplasma* occurs slightly further up the archaea lineage. The phylogenetic tree suggests that the last common ancestor may have been similar to heat-loving chemosynthetic organisms that populate hydrothermal vents today.

QUESTION 3.1

$g_E = GM_E/R_E^2$ and $g = GM/R^2$ where the subscript 'E' refers to values for Earth.

Therefore, $$\frac{g}{g_E} = \frac{M}{M_E}\left(\frac{R_E}{R}\right)^2 \frac{G}{G}$$

So G, the gravitational constant, cancels out and you are left with:

$$\frac{g}{g_{\mathrm{E}}} = \frac{M}{M_{\mathrm{E}}}\left(\frac{R_{\mathrm{E}}}{R}\right)^2$$
$$= (0.1) \times (2)^2$$
$$= 0.4$$

i.e. Mars's surface gravity is around 40% of Earth's.

QUESTION 3.2

Average conditions on the surface of Mars correspond to a pressure of 6.3 mbar and a temperature of around –60 °C. This corresponds to a point well to the left and slightly below the triple point, marked O in Figure 3.6, falling in the region marked 'ice'. Thus liquid water cannot exist under normal or average Mars conditions. However, a typical daily temperature range might be –100 °C to +15 °C. This range corresponds to a region that straddles the line OA in Figure 3.6 and passes into the region marked 'water vapour'. This explains why water-ice in the polar icecaps passes into the atmosphere during the warmer summer period. It's interesting to consider whether conditions on the surface of Mars can enter the region marked 'liquid' in Figure 3.6. To achieve this, we need to look for the possibility of excursions towards higher temperature and pressure. You saw in Section 3.2 that the average pressure is 6.3 mbar with a variation of 2.4 mbar due to seasonal factors. Furthermore, due to altitude variations over the surface, low-lying regions will experience higher pressures. Coupled with the fact that occasionally the temperature rises to maybe +20 °C, this means that sporadically conditions may fall in the region marked 'liquid' – however, this will be only be a temporary circumstance.

QUESTION 3.3

(a) The Viking biology experiments were clearly only deployed at the two Viking lander sites. There are many types of environment on Mars, some probably more favourable for extant life or relics of extinct life, that were not tested. In addition, soil samples were only collected from a short distance below the surface. In view of the oxidizing nature of the surface, we might have expected these samples to be sterile. Samples from greater depths may have produced different results. These are perhaps two of the most compelling arguments which could be used against the notion that the Viking biology experiments ruled out all possibilities of life on Mars.

(b) The results from the GEX, LR and PR experiments from the Martian samples were respectively that oxygen was emitted, labelled gas was emitted and that carbon was detected. These are superficially the same results that would be expected from terrestrial life samples (see Table 3.3). So without the control samples, which would have shown similar results, at least for the GEX and PR experiments, it might have been that the Viking biology experiment results would have been interpreted as indicative of the existence of life.

QUESTION 3.4

(a) In either direction, across or down the image, the resolution is given by the image size or scale divided by the number of pixels.

So, top to bottom, resolution = (4.5/512) km = 8.8 m.

And, side to side, resolution = (12.7/1024) km = 12.4 m.

(Note that it is possible to have different figures for resolution in different directions.)

(b) (i) The resolution in both directions is much smaller (and therefore better) than 500 m. So impact craters of this size could be distinguished (resolved).

(ii) Conversely, in this case, 1 m is below the figure for resolution in both directions. Therefore, 1 metre-scale boulders would not be resolved.

(c) In this case, the area covered by the image would be larger and thus the resolution would be greater (i.e. worse).

QUESTION 3.5

(a) The average speed depends on the molecular mass, m. More specifically, the speed varies as $(1/\text{molecular mass})^{1/2}$. So the more massive a molecule, the lower the average speed (as one would intuitively expect). So lighter molecules are more likely to have speeds above the escape velocity of a planetary body and therefore it will be harder to retain light gases (e.g. hydrogen) in an atmosphere.

(b) From Table 3.2, we see that the two most common constituents in the Martian atmosphere are CO_2 and N_2. From Appendix C, Table C1, we have the following values (Table 3.9) for the appropriate relative atomic masses:

Table 3.9 For Question 3.5(b).

Element	Relative atomic mass
C	12
O	16
N	14

So the relative atomic mass of CO_2 is 44 and of N_2 is 28. Therefore the ratio of the average speeds of these two species is:

Average speed $(CO_2/N_2) = (28/44)^{1/2} = 0.8$, i.e. the average speed of the CO_2 molecules is 80% of that of the N_2 molecules.

QUESTION 3.6

(a) Using appropriate values for masses and radii of Mars, Venus and the Moon, and substituting into Equation 3.8, we obtain the following values (Table 3.10) for the escape velocities:

Table 3.10 For Question 3.6(a).

	Escape velocity/km s^{-1}
Mars	5.0
Venus	10.4
Moon	2.4

So, in ascending order of the escape velocity, we have the Moon, Mars and Venus.

(b) However, this is not the only factor that dictates the likelihood of material from these bodies reaching the Earth. Other factors include:

1 The existence or not of an atmosphere on the relevant body. In the case of Venus, for example, the combination of the high escape velocity and the thick atmosphere means that some of the material ejected as a result of a surface impact will be vaporized during passage through the atmosphere of Venus.

2 Distance can also affect the likelihood of ejecta reaching the Earth.

3 Position in the Solar System. For example, proximity to the Sun (e.g. Mercury) or to a large planet such as Jupiter (e.g. as in the case of Io) can also adversely influence the chances of material reaching the Earth due to gravitational effects.

QUESTION 3.7

The answer is given in Table 3.11.

Table 3.11 Answer to Question 3.7.

	Category	Reasons
(i)	IV	Since comets are of interest for understanding the origins of life and contamination could jeopardize future experiments, Category IV is suggested. However, since there are many comets, it might be argued that a lower category (and thus lower level protection) is warranted.
(ii)	I	Mercury is not of direct interest for understanding the process of chemical evolution so Category I is appropriate.
(iii)	II	Since only a fly-by of Mars is planned, the concern here is primarily over unintentional impact which places it into Category II.
(iv)	IV	This category covers lander missions to targets of interest for understanding the origins of life and for which contamination could jeopardize future experiments. Mars is in this category.
(v)	V	All Earth-return missions are in Category V.

QUESTION 4.1

(a) There are various ways to work this out – here is ours. The value we are looking for is x, so we need to rearrange Equation 4.1 to isolate all the terms involving x on the same side. First, expand the bracket, to get:

$$\rho_{av} = x\rho_{dense} + \rho_{light} - x\rho_{light}$$

Next, subtract ρ_{light} from each side:

$$\rho_{av} - \rho_{light} = x\rho_{dense} - x\rho_{light}$$

Rearranging this equation:

$$\rho_{av} - \rho_{light} = x(\rho_{dense} - \rho_{light})$$

We can now divide both sides by $(\rho_{dense} - \rho_{light})$ to get:

$$\frac{(\rho_{av} - \rho_{light})}{(\rho_{dense} - \rho_{light})} = x$$

Now we can simply insert the density values we were given. Callisto's average density is ρ_{av}, ice density is ρ_{light} and rock density is ρ_{dense}, so:

$$x = \frac{(1.83 \times 10^3 \text{ kg m}^{-3}) - (0.95 \times 10^3 \text{ kg m}^{-3})}{(3.1 \times 10^3 \text{ kg m}^{-3}) - (0.95 \times 10^3 \text{ kg m}^{-3})}$$

$$x = \frac{0.88 \times 10^3 \text{ kg m}^{-3}}{2.15 \times 10^3 \text{ kg m}^{-3}} = 0.41$$

The fraction of Callisto's volume occupied by rock is about 0.41.

(b) One reason the value may be unreliable is that the densities used are for rock and ice at low pressure. In the interior of a large icy satellite the pressure might be high enough for self-compression to lead to significantly higher densities. Another reason is that the method assumes rock and ice only, and ignores the possibility that there could be an even denser component such as an iron-rich inner core.

QUESTION 4.2

Box 4.2 states that tidal force is inversely proportional to the cube of the orbital radius. Thus (tidal force on Europa)/(tidal force on Io) = (Io orbital radius)^3/(Europa orbital radius)^3 = $421.6^3/670.9^3$ = 0.249. Thus the tidal force on Europa is a quarter that on Io. (Note: the amount of tidal heating as a result of this force depends on other factors such as the amount of forced eccentricity and the body's internal properties.)

QUESTION 4.3

This is an exercise in reading values of a logarithmic scale. The concentration of Cl^- in terrestrial seawater is shown as 0.6 moles per litre. The concentration of Cl^- in Europa's ocean is shown as 0.02 moles per litre. The ratio between the two is 0.6/0.02 = 30. Thus the concentration of Cl^- in terrestrial seawater is thirty times that in Europa's ocean.

QUESTION 4.4

Answering this question was part of your learning process. Do not worry if you found yourself at a loss. However, we hope that after reading the answer you will be able to tackle a similar task better in the future.

(a) Most of the surface area appears fairly featureless and mid-grey. This is cut by a large number of linear features (bands), up to several tens of kilometres in width. Most of the bands are dark. Some consist of joined segments of straight lines and some are curved. There is one prominent curved bright band near the lower left. The surface pattern is different in the upper right (northeast), where the pattern of bands disappears and the surface takes on a mottled appearance. Topography becomes apparent only near the right hand (eastern) edge of the view, where the Sun was low in the sky. It is difficult to trace the dark bands into this region, but instead a series of curved ridges shows up.

(b) The dark bands must be younger than the pale (mid-grey) surfaces that they cut. The mottled terrain in the upper right is probably younger than most of the bands, because these disappear when they reach the mottled terrain. Some of the curved ridges in the lower right-hand corner appear to run over the bands, and so these curved ridges must also be younger.

QUESTION 4.5

Pwyll is circular in outline, which is to be expected, but its topography appears to be extremely subdued, even on this image that was recorded when the Sun was very low in the sky (to judge form the shadows in the surrounding area). The rim is very poorly expressed, and there is a cluster of central peaks rather than a single central peak such as you might expect in a crater of this size.

QUESTION 4.6

Although Figure 4.22 includes more variation in size of ridges and grooves than in the comparable sized area shown in Figure 4.19b, this and most of the area of Figure 4.21 has the basic 'ball of string' texture.

QUESTION 4.7

Feature A cuts across feature B, and so feature A must be the younger of the two. Moreover, the parts of feature B on either side of feature A are no longer aligned. They have been displaced to the right by nearly 1 km. The simplest explanation of this is that feature A is a fault with about 1 km of sideways movement across it (a geologist would describe it as a dextral (or right-lateral) strike-slip fault). You can get the same impression of displacement to the right where feature A offsets the edge of the relatively smooth surface in the lower third of the image. (Note that although A is younger than B, A is certainly not the youngest ridge or groove in this area: for example, an even younger groove cuts through A at right angles near the top of the image.)

QUESTION 4.8

(a) It is obvious that the 'ball of string' texture once covered the whole of the area shown in Figure 4.23. However, there are many patches about 10 km across where this texture can be seen in various degrees of disruption. For example:

(i) in square D4 the 'ball of string' surface has been warped upwards into a gentle dome, with a zig-zag fracture where its roof has been stretched apart;

(ii) in squares D/E–1/2 the 'ball of string' surface has been destroyed in a roughly rectangular area, except for a 4 km × 2 km fragment that survives near the southwest edge of the disrupted area. The surface of this whole disrupted area is domed upwards, but its edge must be lower than the surrounding terrain because it is surrounded by an inward-facing cliff;

(iii) in squares D/E–1/2 there is a dome that looks like a mushy extrusion across the original surface within which no identifiable traces of 'ball of string' texture remain.

Sites (i)–(iii) can be regarded as progressively more disrupted examples of 'ball of string texture'. A dome intermediate in character between those at sites (ii) and (iii) occurs at B1–B2. You may also be able to make out several more subtle domes within which the surface has not been fractured at all (in Figure 4.23 and nearby parts of Figure 4.21).

(Note: there are no patch-like depressions in this image; they are all domes. If you cannot perceive them as such, despite being told that the illumination is coming from the right, try rotating the page 90° anticlockwise, so that the illumination now comes from above. This is a more 'natural' illumination direction, and your brain may now be able to make better sense of the topography in the image.)

(b) Throughout Figure 4.24, the 'ball of string' surface has been fractured into slabs, which are bounded by cliffs and so stand higher than the intervening surface, which is occupied by hummocks a few hundred metres across. The slabs still retain their 'ball of string' texture, and by matching prominent ridges and grooves on adjacent slabs it is possible to see that the slabs in the northwest (top left) of the image have been jostled apart by distances of about 1 km. However, the further southeast you look in this image, the harder it is to identify matching slabs and the greater the proportion of new, low-lying, hummocky surface.

QUESTION 4.9

The groove cuts through rafts and matrix alike. Its appearance on the rafts is unremarkable, and it would be taken for just another element of each raft's 'ball of string' texture if we did not see it also cutting the matrix. Generally speaking, the groove's course is not deflected where it crosses from one surface type to another. This groove must have formed at a time when the matrix had become virtually as rigid as the rafts. It is seen cutting the matrix near the right-hand edge of Figure 4.26. There are at least two other grooves cutting the matrix in Figure 4.24. One runs parallel to the first groove, about 5 km to its southwest. The other is at right angles to the first groove, which cuts it about 5 km from the northwest corner of the image.

QUESTION 4.10

(a) Because it is w that we are trying to find, we need to get all the terms involving w into the same side of the equation.

Equation 4.3 can be expanded as:

$$\rho_1 h + \rho_1 w = \rho_2 w$$

Subtracting $\rho_1 w$ from both sides of this equation, we get:

$$\rho_1 h = \rho_2 w - \rho_1 w = w(\rho_2 - \rho_1)$$

And to find w we need to divide both sides by $(\rho_2 - \rho_1)$:

$$w = \frac{\rho_1 h}{(\rho_2 - \rho_1)}$$

(b) It might not be immediately obvious whether the maximum raft density will give the maximum or the minimum raft thickness, but it has to be one or the other. Inserting the value of $1126\,\text{kg m}^{-3}$ as ρ_1 in this equation and using 100 m as h and ρ_2 as $1180\,\text{kg m}^{-3}$, we get:

$$w = \frac{(1126\,\text{kg m}^{-3} \times 100\,\text{m})}{(1180\,\text{kg m}^{-3} - 1126\,\text{kg m}^{-3})} = \frac{(1126\,\text{kg m}^{-3} \times 100\,\text{m})}{54\,\text{kg m}^{-3}} = 2085\,\text{m}$$

The raft thickness is $(h + w)$, and so we need to add 100 m to this value, giving a raft thickness of 2185 m.

Inserting $927\,\text{kg m}^{-3}$ as ρ_1 in the same expression we get:

$$w = \frac{(927\,\text{kg m}^{-3} \times 100\,\text{m})}{(1180\,\text{kg m}^{-3} - 927\,\text{kg m}^{-3})} = \frac{927 \times 100\,\text{m}}{253} = 366\,\text{m}$$

and hence a raft thickness of 466 m.

The cliff height is certainly not known to three significant figures, so we should not quote these results to more than two significant figures. Thus, according to this method, the raft thickness is not less than about 470 m and not more than about 2200 m.

In fact, the less the density contrast between raft and fluid, the lower the height of the cliffs. If a raft has the same density as the fluid it barely floats at all. If a raft is very much less dense than the fluid, only a relatively small proportion of the raft's volume needs to be immersed in the fluid in order to displace an equivalent mass of fluid.

QUESTION 4.11

(a) Inserting the relevant values into Equation 4.2 (and remembering to convert from km to m), we get:

$P = 1030\,\text{kg m}^{-3} \times 9.8\,\text{m s}^{-2} \times 3000\,\text{m} = 3.0 \times 10^{7}\,\text{kg m s}^{-2}\,\text{m}^{-2} = 3.0 \times 10^{7}\,\text{Pa} = 30\,\text{MPa}.$

($\text{kg m s}^{-2}\,\text{m}^{-2}$ is force per unit area, which is pressure. The SI unit of pressure is the pascal, abbreviated Pa. Note that $\text{kg m s}^{-2}\,\text{m}^{-2}$ could be written as $\text{kg m}^{-1}\,\text{s}^{-2}$ but this would obscure the significance of kg m s^{-2} being the SI unit of force.)

(b) Similarly, inserting the relevant values into Equation 4.2 we get:

$P = 1180\,\text{kg m}^{-3} \times 1.3\,\text{m s}^{-2} \times 10^{5}\,\text{m} = 1.5 \times 10^{8}\ \text{kg m s}^{-2}\,\text{m}^{-2} = 1.5 \times 10^{8}\,\text{Pa} = 150\,\text{MPa}.$

QUESTION 4.12

Europa's annual biomass production is estimated to be at least eight orders of magnitude less than that of present-day Earth (a maximum of $10^{6}\,\text{yr}^{-1}$ on Europa versus a total of about $10^{14}\,\text{yr}^{-1}$ on Earth). Even if we compare only chemosynthetic biomass production, Europa is estimated to be at least ten thousand times (four orders of magnitude) less productive.

QUESTION 4.13

Perhaps the most obvious technique to use to find out more about Europa is extensive imaging of the surface, at high enough spatial resolution to identify chaos regions and with high enough spectral resolution to identify salts and other contaminants in the ice. You may also have thought of the use of a radar or laser altimeter to map the topography, and thereby contribute to Objective 3. Potentially, a radar instrument could also help significantly with Objectives 1 and 2, as discussed shortly in the text. Precise tracking of the orbiter's trajectory could give information about the details of Europa's gravity field, and hence its internal structure, which would also help with Objectives 1 and 2.

QUESTION 4.14

(a) The matrix between the rafts here is smooth and shows no sign of being cut by the groove. This means that the groove must be older than the matrix. (In case you found it hard to see the necessary detail on Figure 4.24, it is enlarged in Figure 4.34.) This is unlike what happens to the northwest and to the southeast, where the groove is clearly seen to cut the matrix (see Question 4.9).

Thus the matrix between the rafts here appears to be younger than the matrix elsewhere in Figure 4.24.

Figure 4.34 Enlargement of the area of interest for Question 4.14. The groove in question passes through A and B. The matrix-filled gap between rafts runs through C and D.

(b) If the matrix between the rafts is younger than the matrix elsewhere, the time sequence of events in the region as a whole must be: chaos formation, freezing and thickening of matrix until it becomes rigid enough for grooves to form on it, formation of this groove, local remobilization of matrix between the rafts erasing the groove at this location. We can conclude that this part of the matrix has been active over a protracted time period.

(c) The matrix between the rafts here is white, so Pwyll ejecta lies on top of it (see caption for Figure 4.21). Remobilization of the matrix would probably disrupt or destroy such a thin ejecta blanket, so probably the Pwyll impact post-dates this local remobilization event.

QUESTION 4.15

Inserting the new value of $1140\,\text{kg m}^{-3}$ for ρ_2 into the method used to answer Question 4.10b, we get:

$$w = \frac{(1126\,\text{kg m}^{-3} \times 100\,\text{m})}{(1140\,\text{kg m}^{-3} - 1126\,\text{kg m}^{-3})} = \frac{1126\,\text{kg m}^{-3} \times 100\,\text{m}}{14\,\text{kg m}^{-3}} = 8043\,\text{m}$$

The raft thickness is $(h + w)$, and so we need to add 100 m to this, giving a raft thickness of 8143 m, which we ought to quote to no more than two significant figures, i.e. 8100 m.

QUESTION 4.16

Alternative implications could be:

1 The site has been contaminated by viable organisms accidentally brought from Earth by an earlier probe.

2 Life is indigenous to Europa and arose there independently.

3 Life is indigenous to Europa, and both Earth and Europa were seeded from the same external (for example, cometary) source.

4 Life is indigenous to Europa, but arrived there as contamination via a meteorite from Earth (or Mars).

Implication 1 would be unlikely if the pre-launch cleaning and sterilization of all the previous probes was believed with confidence to be sufficiently stringent. However, it could only be ruled out by detailed genetic study of the 'Europan' micro-organisms to prove that they were not closely related to terrestrial species (hard to do using a

robotic probe) or if a sufficiently complex ecology (especially with multicellular organisms and heterotrophs preying on the autotrophs) were discovered that could not have had time to develop since the first possible contamination episode.

Implications 2–4 would all be taken as proof of extraterrestrial life, but detailed genetic studies would be necessary to try to establish which was correct. If Europan amino acids were discovered to have right-handed chirality in contrast to the left-handed chirality ubiquitous on Earth (Section 1.7), this would point towards implication 2. However, there is at least at 50:50 chance of left-handed chirality arising independently, so discovery of left-handed chirality on Europa would not help us to decide between any of these implications.

QUESTION 5.1

(a) Figure 5.19 shows the temperature profile for Titan with the labels 'troposphere' and 'thermosphere' added. The troposphere is the region near the surface where the temperature falls with altitude; the thermosphere is the high altitude region where the temperature increases with altitude. The curve has some similarity in shape to those for Mars and Earth.

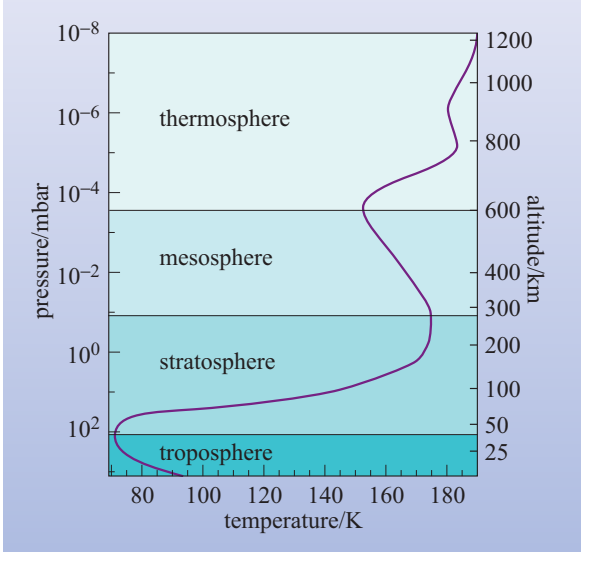


Figure 5.19 Temperature profile for Titan (as in Figure 5.7) with the troposphere and thermosphere labelled.

(b) By approximating the lower part of the curve in Figure 5.7 with a straight line, and extending it across the axes in the figure, one can estimate the slope, which is the lapse rate. This is found to be approximately 1 K km^{-1}. The value quoted in Section 5.4 for the lowest 10 km is (1.38 ± 0.1) K km^{-1}. Bearing in mind the coarseness of the plot, the agreement is as good as one might expect.

QUESTION 5.2

From Table 5.2, we have the following data for Titan:

$$\text{mass} = 1.346 \times 10^{23}\,\text{kg}, \quad \text{radius} = 2.575 \times 10^{6}\,\text{m}.$$

For a sphere, volume = $(4/3)\pi R^3$ and density ρ = mass/volume.

Therefore, Titan's density is given by:

$$\rho_{\text{Titan}} = \frac{1.346 \times 10^{23}\ \text{kg}}{\frac{4}{3}\pi(2.575 \times 10^{6}\ \text{m})^{3}}$$
$$= 1.9 \times 10^{3}\ \text{kg m}^{-3}$$

By examining Appendix A Tables A1 and A2, we see that the densities of most of the planetary satellites lie in the range (1 to 2) $\times$ 10^3 kg m^{-3}, while the terrestrial planets fall within the approximate range of (4 to 5.5) $\times$ 10^3 kg m^{-3}. Therefore Titan, with an average density of 1.9 $\times$ 10^3 kg m^{-3}, is consistent with the majority of the (icy) planetary satellites.

QUESTION 5.3

Substituting values from Table 5.2, we obtain:

$$g = \frac{GM}{R^2}$$
$$= \frac{(6.67 \times 10^{-11}\ \text{N m}^2\ \text{kg}^{-2}) \times (1.346 \times 10^{23}\ \text{kg})}{(2.575 \times 10^{6}\ \text{m})^2}$$
$$= 1.35\ \text{m s}^{-2}$$

This confirms the value for gravity quoted in Table 5.2.

QUESTION 5.4

The volume of the ocean is given by $(4\pi R^2) \times 0.5 \times D$ where D is the depth.

Mass of ocean is $(4\pi R^2) \times 0.5 \times D \times \rho$ where ρ is density.

Mass of methane in the ocean is $(4\pi R^2) \times 0.5 \times D \times \rho \times 0.7$.

Mass of methane lost per second from the atmosphere is 4×10^{-12} kg m^{-2} s^{-1} $\times$ $(4\pi R^2)$.

Therefore the time T for which the ocean can resupply the atmosphere with methane is:

$$T = \frac{4\pi R^2 \times 0.5 \times D \times \rho \times 0.7}{(4 \times 10^{-12}\ \text{kg m}^{-2}\ \text{s}^{-1}) \times 4\pi R^2}$$
$$= \frac{0.5 \times D \times \rho \times 0.7}{4 \times 10^{-12}\ \text{kg m}^{-2}\ \text{s}^{-1}}$$
$$= 5.8 \times 10^{16}\ \text{s}$$
$$\approx 2000\ \text{Ma}$$
$$\approx 2\ \text{Ga}$$

This figure is of the same order of magnitude as the age of the Solar System so suggests that this may be a plausible mechanism for maintaining methane in Titan's atmosphere.

This assumes that methane lost from the atmosphere to space is immediately replaced by methane from the ocean.

QUESTION 5.5

The closest and the furthest possible distances from Earth to Saturn will occur when Saturn and the Earth lie on the same line from the Sun, in the first case when they are both on the same side of the Sun, and in the other when they are on opposite sides of the Sun (these situations are technically known as conjunctions). From Appendix A, Table A1, we see that Earth's orbit is almost circular ($e = 0.017$) while for Saturn $e = 0.055$. At least for an approximate answer, we can assume that the orbits are circular and use the mean distance of each planet from the Sun, namely 1.0 AU and 9.5 AU.

(a) The closest distance will therefore be (9.5 – 1.0) AU = 8.5 AU and (b) the furthest distance will be (9.5 + 1.0) AU = 10.5 AU. These values correspond to 1.275×10^9 km and 1.575×10^9 km respectively.

A slightly more accurate result could be obtained by taking account of the fact that Saturn's orbital eccentricity means that its closest distance to the Sun (perihelion) is 9.0 AU and furthest distance 10.0 AU. Therefore the smallest possible separation between Earth and Saturn is (9.0 – 1.0) AU = 8.0 AU and largest separation is (10.0 + 1.0) AU = 11.0 AU. These values correspond to 1.2×10^9 km and 1.65×10^9 km respectively.

The reason that the Cassini spacecraft travels more than twice these distances is because it does not travel on a direct (ballistic) trajectory. Instead it has to use the gravity assist technique to deliver its payload to the Saturnian system. This involves a fairly complex trajectory (see Figure 5.12) with fly-bys of Venus, Earth and Jupiter and explains the distance travelled of 3.2×10^9 km.

QUESTION 5.6

(a) We saw in Section 5.3.2 that the abundance of argon is a good indicator of the origin of nitrogen on Titan. If the abundance is similar to what it had been in the nebula from which Titan's atmosphere formed, namely an Ar/N_2 ratio of about 0.06 (i.e. 6%), then it would be expected that Titan's nitrogen had been directly captured (into a clathrate) and gradually outgassed to form the present nitrogen in the atmosphere. On the other hand, if argon's atmosphere were very much less than this value, then Titan's nitrogen was expected to come from the photodissociation of ammonia. In this question, we are told that argon's abundance was found to be about 0.05% so it is the second explanation, namely an ammonia source for nitrogen, that is favoured.

(b) The fact that waves are measured on an open liquid suggests that they are wind driven. We see from Table 5.5 that for a surface wind speed of 5 m s^{-1}, waves of significant height of 4.5 m are expected on Titan. If we measure a value of about 3 m, then we might expect the surface wind (at the time of the measurement) to be slightly less than 5 m s^{-1}, namely about 3 to 4 m s^{-1}.

(c) The Earth's average distance from the Sun is 1.0 AU while for Saturn (and Titan) this distance is 9.5 AU. Therefore, if all other factors were equal, the strength of sunlight at Titan compared to its strength at Earth would be reduced by a factor of $(1.0/9.5)^2 \approx 0.01 = 1/100$. However, the measured factor is 1/1000 instead of 1/100, i.e. it is 10 times weaker than we would expect. This is almost certainly due to absorption by Titan's cloud and haze layers.

APPENDIX A USEFUL PLANETARY DATA

Table A1 Basic data on the planets (including the Moon).

	Mercury	Venus	Earth	Moon	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
Mass										
/10^{24} kg	0.330	4.87	5.97	0.074	0.642	1900	569	86.8	102	0.013
/Earth masses	0.055	0.815	1.00	0.012	0.107	318	95.2	14.4	17.1	0.002
Orbital semimajor axis[a]										
/10^6 km	57.91	108.2	149.6	149.6	227.9	778.4	1427	2871	4498	5906
/AU	0.39	0.72	1.00	1.00	1.52	5.20	9.54	19.19	30.07	39.48
Orbital eccentricity	0.206	0.007	0.017	0.055	0.093	0.048	0.054	0.047	0.009	0.249
Orbital inclination /degrees	7.0	3.4	0.0	5.2	1.9	1.3	2.5	0.8	1.8	17.1
Orbital period[b]	88.0 days	224.7 days	365.0 days	27.3 days	686.5 days	11.86 yr	29.42 yr	83.75 yr	163.7 yr	248.0 yr
Axial rotation period[b]/days	58.6	243	0.997	27.3	1.03	0.412	0.444	0.718	0.671	6.39
Axial inclination /degrees	0.1	177.3	23.5	6.7	25.2	3.1	26.7	97.9	29.6	119.6
Polar radius/km	2440	6052	6357	1738	3375	66 850	54 360	24 970	24 340	1137
Equatorial radius/km	2440	6052	6378	1738	3397	71 490	60 270	25 560	24 770	1137
Mean radius[c]/km	2440	6052	6371	1738	3390	69 910	58 230	25 360	24 620	1137
Density/10^3 kg m^{-3}	5.43	5.20	5.51	3.34	3.93	1.33	0.69	1.32	1.64	2.1
Surface gravity/m s^{-2}	3.7	8.9	9.8	1.6	3.7	23.1	9.0	8.7	11.1	0.7
Mean surface temperature/K	443	733	288	250	223					≥40
Effective cloud-top temperature/K						120	89	53	54	
Temperature at 1 bar pressure/K						165	135	75	70	
Rings	0	0	0	0	0	few	many	several	few	0
Satellites	0	0	1		2	≥39	≥30	≥21	≥8	1
Atmospheric surface pressure/bar[d]	≈10^{-15}	92.1	1.01	≈10^{-14}	6.3×10^{-3}					≈10^{-5}
Atmospheric surface density/kg m^{-3}	≈10^{-13}	67	1.293	≈10^{-13}	0.018					≈10^{-4}
Atmospheric column mass[e]/kg m^{-2}	≈10^{-11}	1.03×10^6	1.03×10^4	≈10^{-11}	1.69×10^2					≈1
Atmospheric main components (relatively minor components in parentheses)	O Na H_2 (He)	CO_2 (N_2)	N_2 O_2 (H_2O) (Ar)	Ar H_2 He Na	CO_2 (N_2) (Ar) (O_2)	H_2 He (CH_4)	H_2 He (CH_4)	H_2 He (CH_4)	H_2 He (CH_4)	N_2 (CH_4) (CO_2)

[a] Semimajor axis also represents the *mean* distance from the Sun.
[b] These are sidereal periods (i.e. referenced to the stars rather than to the Sun) quoted in Earth days or years.
[c] The mean radius is defined as the *volumetric* radius (i.e. the radius the body would have if it were a sphere of the same mass), and is calculated by $(R_e^2 R_p)^{1/3}$. The values quoted here for the gas giants are for the atmospheric layer where the pressure is equal to 1 bar (this also applies to the values for surface gravity).
[d] Although the SI unit for pressure is the pascal, we use bar here for simplicity and easy comparison, as the Earth's surface atmospheric pressure ≈1 bar. Note 1 bar = 10^5 Pa.
[e] Column mass is the mass of atmosphere situated above each unit area (1 m^2) of the planet's surface.

Table A2 Planetary satellites.

Planet	Satellite	Mean distance from planet/10^3 km	Orbital period/days	Mean radius[a]/km	Mass/10^{20} kg	Density/ 10^3 kg m^{-3}
Earth	Moon	384	27.3	1738	735	3.34
Mars	Phobos	9.4	0.32	11.1	0.00011	1.90
	Deimos	23.5	1.26	6.2	0.000018	1.76
Jupiter	Io	422	1.77	1821	893	3.53
	Europa	671	3.55	1565	480	2.99
	Ganymede	1070	7.15	2634	1482	1.94
	Callisto	1883	16.7	2403	1076	1.85
	≥56 others					
Saturn	Mimas	186	0.94	199	0.38	1.14
	Enceladus	238	1.37	249	0.73	1.21
	Tethys	295	1.89	530	6.2	1.00
	Dione	377	2.74	560	10.5	1.44
	Rhea	527	4.52	764	23.1	1.24
	Titan	1222	15.95	2575	1346	1.88
	Iapetus	3561	79.3	718	15.9	1.02
	≥24 others					
Uranus	Miranda	130	1.42	236	0.66	1.20
	Ariel	191	2.52	579	13.5	1.7
	Umbriel	266	4.14	585	11.7	1.4
	Titania	436	8.71	789	35.3	1.71
	Oberon	583	13.46	761	30.1	1.63
	≥16 others					
Neptune	Proteus	118	1.12	209	≈0.5?	≈1.2?
	Triton	355	5.88	1353	215	2.05
	Nereid	5513	360	170	≈0.3?	≈1.2?
	≥9 others					
Pluto	Charon	19.4	6.39	586	19.0	2.2

[a] The mean radius is defined as the *volumetric* radius (i.e. the radius the body would have if it were a sphere of the same mass).

Table A3 Asteroids that have been targets of spacecraft fly-bys or encounters.

Asteroid[a]	Spacecraft	Encounter date	Asteroid size/km	Mean radius[b]/km	Density/kg m^{-3}	Semimajor axis/AU
(951) Gaspra	Galileo	29 Oct 1991	19 × 12	6.1	2500 ±1000?	2.21
(243) Ida	Galileo	28 Aug 1993	58 × 23	15.8	2600 ±500	2.86
(253) Mathilde	NEAR	27 Jun 1997	59 × 47	26.4	1300 ±200	2.65
(9969) Braille	Deep Space 1	29 Jul 1999	2 × 1	0.7	not known	2.34
(433) Eros	NEAR	14 Feb 2000[c]	33 × 13	9.69	2670 ±30	1.46
(5535) Annefrank	Stardust	2 Nov 2002	8 × 4	2.5	not known	2.21

[a] Asteroids are initially numbered, and are then usually named also. We refer to them by (*number*) *name*.

[b] The mean radius is defined as the *volumetric* radius (i.e. the radius the body would have if it were a sphere of the same mass).

[c] NEAR went into orbit around Eros on this date. It remained there for a year and then landed on the surface of Eros on 12 Feb 2001.

Table A4 The largest known minor bodies in the Solar System.

Object	Semimajor axis/AU	Orbital period/yr	Orbital inclination	Orbital eccentricity	Mean radius[a]/km
Largest bodies in the asteroid belt:					
(1) Ceres	2.77	4.61	10.6°	0.079	457
(2) Pallas	2.77	4.61	34.8°	0.230	261
(4) Vesta	2.36	3.63	7.1°	0.090	250
(10) Hygiea	3.14	5.59	3.8°	0.121	215
(511) Davida	3.17	5.65	15.9°	0.180	163
Largest *known* (as of mid-2002) bodies in the Kuiper Belt (excluding Pluto):					
2002 LM_{60} ('Quaoar')	43.2	284	8.0°	0.036	650
2002 AW_{197}	47.5	327	24.3°	0.128	400–650?
(28978) Ixion	39.3	246	19.7°	0.245	400–650?
2002 TX_{300}	43.3	284	25.9°	0.121	350–600?
(20000) Varuna	43.3	285	17.1°	0.054	450

[a] The mean radius is defined as the *volumetric* radius (i.e. the radius the body would have if it were a sphere of the same mass).

Table A5 Some selected comets.

Comet[a]	Perihelion distance/AU	Semimajor axis/AU	Orbital period/yr	Eccentricity	Inclination	Velocity at perihelion/km s^{-1}
2P/Enke	0.338	2.22	3.30	0.847	11.8°	69.6
46P/Wirtanen	1.059	3.09	5.44	0.658	11.7°	37.3
81P/Wild 2	1.590	3.44	6.40	0.539	3.2°	29.3
26P/Grigg–Skjellerup	1.118	3.04	5.31	0.663	22.3°	36.0
55P/Tempel–Tuttle	0.977	10.3	33.2	0.906	162.5°	41.6
1P/Halley	0.587	17.9	76.0	0.967	162.2°	54.5
109P/Swift–Tuttle	0.958	26.3	135	0.964	113.4°	42.6
153P/Ikeya–Zhang	0.507	51.0	367	0.990	28.1°	59.0
Hale–Bopp	0.925	184	≈2500	0.995	89.4°	43.8
Hyakutake	0.230	1490	≈58000	0.9998	124.9°	87.8

[a] Well observed periodic comets (i.e. short-period comets) are numbered, somewhat like asteroids, and this is indicated by the designation *number* P/, for example 2P/Enke.

Table A6 Major annual meteor showers.

Date of maximum rate	Name of shower	Hourly meteor rate	Parent comet
3 Jan	Quadrantids	130	unknown
12 Aug	Perseids	80	Swift–Tuttle
21 Oct	Orionids	25	Halley
17 Nov	Leonids	25[a]	Tempel–Tuttle
13 Dec	Geminids	90	(3200) Phaethon[b]

[a] This rate is usually what is observed, but every 33 years or so, this shower can display much higher rates.

[b] When discovered, Phaethon was assumed to be an asteroid as no cometary coma was observed. However it is likely that some activity has been present in the past.

Table A7 Some notable Solar System exploration missions.

Mission	Launch	Description
Sputnik 1 (USSR)	4 Oct 1957	First Earth-orbiting satellite. Remained in orbit for 92 days.
Pioneer 4 (USA)	3 Mar 1959	4 Mar 1959: first lunar fly-by (within 60 000 km of Moon's surface).
Luna 2 (USSR)	12 Sep 1959	14 Sep 1959: first spacecraft to land (impact) on the Moon.
Venera 1 (USSR)	12 Feb 1961	19 May 1961: first Venus fly-by. (Contact lost before fly-by.)
Mars 1 (USSR)	1 Nov 1962	19 Jun 1963: first Mars fly-by. (Contact lost before fly-by.)
Venera 3 (USSR)	16 Nov 1965	1 Mar 1966: first spacecraft to land on Venus. (Contact lost before landing.)
Luna 9 (USSR)	31 Jan 1966	3 Feb 1966: First soft landing on the Moon. TV pictures returned to Earth.
Zond 5 (USSR)	14 Sep 1968	First spacecraft to orbit the Moon (18 Sep 1968) and return a payload safely to Earth (21 Sep 1968). Payload included turtles, flies, worms and plants.
Apollo 8 (USA)	21 Dec 1968	First manned mission to orbit the Moon (24 Dec 1968). Returned 27 Dec 1968.
Apollo 11 (USA)	16 July 1969	First manned landing on the Moon (20 July 1969). Crew: Neil Armstrong, Edwin 'Buzz' Aldrin, Michael Collins (orbiter). Returned 24 July 1969.
Apollo 12 (USA)	14 Nov 1969	Second manned landing on the Moon (19 Nov 1969). Crew: Charles Conrad, Alan Bean, Richard Gordon (orbiter). Returned 24 Nov 1969.
Apollo 13 (USA)	11 Apr 1970	Moon mission aborted after onboard explosion on 14 Apr 1970. Crew: James Lovell, Fred Haise, John Swigert (orbiter). Returned 17 Apr 1970.
Luna 16 (USSR)	12 Sep 1970	First robotic sample-return from the Moon. Returned approximately 100 g of lunar material.
Apollo 14 (USA)	31 Jan 1971	Third manned landing on the Moon (5 Feb 1971). Crew: Alan Shepard, Edgar Mitchell, Stuart Roosa (orbiter). Returned 9 Feb 1971.
Mars 3 (USSR)	28 May 1971	2 Dec 1971: first spacecraft to land on Mars. Soft landing. Images returned.
Apollo 15 (USA)	26 Jul 1971	Fourth manned landing on the Moon (30 Jul 1971). Crew: David Scott, James Irwin, Alfred Worden (orbiter). Returned 7 Aug 1971. First lunar rover used.
Pioneer 10 (USA)	3 Mar 1972	First outer Solar System mission. 3 Dec 1973: fly-by of Jupiter. Currently ≈80 AU from the Sun. Will reach the star Aldebaran in 2 million years!
Apollo 16 (USA)	16 Apr 1972	Fifth manned landing on the Moon (21 Apr 1972). Crew: John Young, Charles Duke, Thomas Mattingly (orbiter). Returned 27 Apr 1972.
Apollo 17 (USA)	7 Dec 1972	Sixth (and final) manned landing on the Moon (11 Dec 1972). Crew: Eugene Cernan, Harrison Schmitt, Ronald Evans (orbiter). Returned 19 Dec 1972.
Pioneer 11 (USA)	6 Apr 1973	4 Dec 1974: Jupiter fly-by. 1 Sep 1979: Saturn fly-by.
Skylab (USA)	14 May 1973	First manned orbiting 'space station'. Manned until 8 Feb 1974. Final usage of the Apollo Saturn V rocket.
Mariner 10 (USA)	3 Nov 1973	First (and only) spacecraft to go to Mecury. 5 Feb 1974: Venus fly-by. Mercury fly-bys on 29 Mar 1974, 21 Sep 1974 and 16 Mar 1975.
Viking 1 (USA)	20 Aug 1975	Mars orbiter and lander. 19 June 1976: reached Mars. 20 Jul 1976: lander touched down.

Table A7 continued.

Mission	Launch	Description
Viking 2 (USA)	4 Sept 1975	Mars orbiter and lander. 7 Aug 1976: reached Mars. 3 Sep 1976: lander touched down.
Voyager 2 (USA)	20 Aug 1977	First (only) spacecraft to undertake a tour of all the giant planets. 9 Jul 1979: Jupiter fly-by. 26 Aug 1981: Saturn fly-by. 24 Jan 1986: Uranus fly-by. 25 Aug 1989: Neptune fly-by.
Voyager 1 (USA)	5 Sep 1977	5 Mar 1979: Jupiter fly-by. 12 Nov 1980: Saturn fly-by.
ISEE-3/ICE (USA)	12 Aug 1978	11 Sep 1985: first spacecraft to 'distant fly-by' a comet (Giacobini–Zinner).
Venera 13 (USSR)	30 Oct 1981	1 Mar 1982: Venus landing. Returned colour images from the surface.
Giotto (ESA)	2 Jul 1985	13 Mar 1986: first close (600 km) fly-by of a cometary nucleus (comet Halley).
Magellan (USA)	4 May 1989	Venus orbit insertion 10 Aug 1990. Mapped Venus surface with radar (1990–1994).
Galileo (USA)	18 Oct 1989	First spacecraft to orbit one of the giant planets. 29 Oct 1991: fly-by of asteroid (951) Gaspra. 28 Aug 1993: fly-by of asteroid (243) Ida. 7 Dec 1995: Galileo reaches Jupiter and deployed probe enters the atmosphere of Jupiter. 21 Sept 2003: Galileo impacts Jupiter.
Ulysses (ESA)	6 Oct 1990	First spacecraft to leave the ecliptic plane and orbit around the Sun, passing over the north and south poles. 8 Feb 1992: Jupiter fly-by.
Near Earth Asteroid Rendezvous (NEAR) Mission (USA)	17 Feb 1996	First spacecraft to orbit and land on an asteroid. 27 Jun 1997: fly-by of asteroid (253) Mathilde. 14 Feb 2000: started orbiting near Earth asteroid, (433) Eros. 12 Feb 2001: spacecraft landed on Eros.
Mars Global Surveyor (USA)	7 Nov 1996	Highly successful Mars remote sensing mission. 12 Sep 1997: reached Mars. Mar 1999: began mapping planet.
Mars Pathfinder (USA)	4 Dec 1996	4 Jul 1997: landed on Mars. 6 Jul 1997: deployed the Sojourner rover.
Cassini–Huygens (USA + Europe)	15 Oct 1997	Mission to Saturn and Titan. 30 Dec 2000: Jupiter fly-by. 1 Jul 2004: Saturn orbit insertion. 14 Jan 2005: Huygens probe lands on Titan.
Deep Space 1 (USA)	24 Oct 1998	22 Sep 2001: close fly-by of comet Borrelly's nucleus. Images returned. 29 Jul 1999: fly-by of (9969) Braille.
Stardust (USA)	7 Feb 1999	Fly-by and cometary dust sample return mission to comet Wild 2. 2 Nov 2002: fly-by of asteroid (5535) Annefrank. 2 Jan 2004: fly-by of comet Wild 2. 15 Jan 2006: capsule carrying cometary dust lands on Earth for analysis.
2001 Mars Odyssey (USA)	7 Apr 2001	11 Jan 2002: entered Mars orbit. Acts as relay for 2003 rover missions.
Genesis (USA)	8 Aug 2001	Solar wind particle sample return mission. 3 Dec 2001: capture experiment deployed. Sep 2004: samples returned to Earth.
Rosetta (ESA)	2003	Comet orbiter and lander. Nominal mission plan: 10 Jul 2006: fly-by of asteroid (4979) Otawara. 24 Jul 2008: fly-by of asteroid (140) Siwa. 29 Nov 2011: orbit entry around comet Wirtanen. Sep 2012: lander deployed. (Note: exact mission plan may change.)
Mars Express (ESA) + Beagle 2	≈1 Jun 2003	Mars orbiter and lander. 26 Dec 2003: Mars Express enters Mars orbit, and the Beagle 2 spacecraft lands on the surface to look for isotope ratios indicative of life.

APPENDIX B SELECTED PHYSICAL CONSTANTS AND UNIT CONVERSIONS

Table B1 SI fundamental and derived units.

Quantity	Unit	Abbreviation	Equivalent units
mass	kilogram	kg	
length	metre	m	
time	second	s	
temperature	kelvin	K	
angle	radian	rad	
area	square metre	m^2	
volume	cubic metre	m^3	
speed, velocity	metre per second	$m\,s^{-1}$	
acceleration	metre per second squared	$m\,s^{-2}$	
density	kilogram per cubic metre	$kg\,m^{-3}$	
frequency	hertz	Hz	(cycles) s^{-1}
force	newton	N	$kg\,m\,s^{-2}$
pressure	pascal	Pa	$N\,m^{-2}$, $kg\,m^{-1}\,s^{-2}$
energy	joule	J	$kg\,m^2\,s^{-2}$
power	watt	W	$J\,s^{-1}$, $kg\,m^2\,s^{-3}$
specific heat capacity	joule per kilogram kelvin	$J\,kg^{-1}\,K^{-1}$	$m^2\,s^{-2}\,K^{-1}$
thermal conductivity	watt per metre kelvin	$W\,m^{-1}\,K^{-1}$	$m\,kg\,s^{-3}\,K^{-1}$

Table B2 Selected physical constants and preferred values.

Quantity	Symbol	Value
speed of light in a vacuum	c	$3.00 \times 10^8\,m\,s^{-1}$
Planck constant	h	$6.63 \times 10^{-34}\,J\,s$
Boltzmann constant	k	$1.38 \times 10^{-23}\,J\,K^{-1}$
gravitational constant	G	$6.67 \times 10^{-11}\,N\,m^2\,kg^{-2}$
Stefan–Boltzmann constant	σ	$5.67 \times 10^{-8}\,W\,m^2\,K^{-4}$
Avogadro constant	N_A	$6.02 \times 10^{23}\,mol^{-1}$
molar gas constant	R	$8.31\,J\,K^{-1}\,mol^{-1}$
charge of electron	e	1.60×10^{-19} C (negative charge)
mass of proton	m_p	1.67×10^{-27} kg
mass of electron	m_e	9.11×10^{-31} kg
Astronomical quantities:		
mass of the Sun	$M_\odot$	1.99×10^{30} kg
radius of the Sun	$R_\odot$	6.96×10^8 m
photospheric temperature of the Sun	$T_\odot$	5770 K
luminosity of the Sun	$L_\odot$	3.84×10^{26} W
astronomical unit	AU	1.50×10^{11} m

Table B3 Some useful conversions from alternative unit systems to SI units.

Quantity	Unit	SI equivalent
angle	1 degree	$(\pi/180)$ rad
pressure	1 bar	10^5 Pa
temperature	1 °C	1 K
energy	1 erg	10^{-7} J
	1 electron volt	1.60×10^{-19} J
	1 ton of TNT	4.18×10^9 J
length	1 foot	0.305 m
	1 mile	1.61×10^3 m
area	1 square inch	6.45×10^{-4} m^2
	1 square mile	2.59×10^6 m^2
mass	1 pound	0.454 kg
speed, velocity	1 mile per hour	0.447 m s^{-1}

Table B4 The Greek alphabet.

Name	Lower case	Upper case
Alpha	α	A
Beta (bee-ta)	β	B
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ε	E
Zeta (zee-ta)	ζ	Z
Eta (ee-ta)	η	H
Theta (thee-ta – 'th' as in theatre)	θ	Θ
Iota (eye-owe-ta)	ι	I
Kappa	κ	K
Lambda (lam-da)	λ	Λ
Mu (mew)	μ	M
Nu (new)	ν	N
Xi (cs-eye)	ξ	Ξ
Omicron	o	O
Pi (pie)	π	Π
Rho (roe)	ρ	P
Sigma	σ	Σ
Tau	τ	T
Upsilon	υ	Y
Phi (fie)	ϕ	Φ
Chi (kie)	χ	X
Psi (ps-eye)	ψ	Ψ
Omega (owe-me-ga)	ω	Ω

APPENDIX C THE ELEMENTS

Table C1 The elements and their abundances.

The relative atomic mass, A_r, is the average mass of the atoms of the element as it occurs on Earth. It is thus an average over all the isotopes of the element. The scale is fixed by giving the carbon isotope $^{12}_{6}C$ a relative atomic mass of 12.0. By convention, the Solar System abundance is normalized to 10^{12} atoms of hydrogen, whereas the CI chondrite abundance is normalized to 10^6 atoms of silicon. To directly compare chondrite abundance to Solar System abundance (by number), you would multiply chondrite abundance by 35.8.

Atomic number, Z	Name	Chemical symbol	Relative atomic mass, A_r	Solar System abundance		CI chondrite abundance by number
				by number	by mass	
1	hydrogen	H	1.01	1.0×10^{12}	1.0×10^{12}	2.79×10^{10}
2	helium	He	4.00	9.8×10^{10}	3.9×10^{11}	2.72×10^{9}
3	lithium	Li	6.94	2.0×10^{3}	1.4×10^{4}	57.1
4	beryllium	Be	9.01	26	2.4×10^{2}	0.73
5	boron	B	10.81	6.3×10^{2}	6.8×10^{3}	21.2
6	carbon	C	12.01	3.6×10^{8}	4.4×10^{9}	1.01×10^{7}
7	nitrogen	N	14.01	1.1×10^{8}	1.6×10^{9}	3.13×10^{6}
8	oxygen	O	16.00	8.5×10^{8}	1.4×10^{10}	2.38×10^{7}
9	fluorine	F	19.00	3.0×10^{4}	5.7×10^{5}	843
10	neon	Ne	20.18	1.2×10^{8}	2.5×10^{9}	3.44×10^{6}
11	sodium	Na	22.99	2.0×10^{6}	4.7×10^{7}	5.74×10^{4}
12	magnesium	Mg	24.31	3.8×10^{7}	9.2×10^{8}	1.074×10^{6}
13	aluminium	Al	26.98	3.0×10^{6}	8.1×10^{7}	8.49×10^{4}
14	silicon	Si	28.09	3.5×10^{7}	1.0×10^{9}	1.00×10^{6}
15	phosphorus	P	30.97	3.7×10^{5}	1.2×10^{7}	1.04×10^{4}
16	sulfur	S	32.07	1.9×10^{7}	6.0×10^{8}	5.15×10^{5}
17	chlorine	Cl	35.45	1.9×10^{5}	6.6×10^{6}	5240
18	argon	Ar	39.95	3.6×10^{6}	1.5×10^{8}	1.01×10^{5}
19	potassium	K	39.10	1.3×10^{5}	5.2×10^{6}	3770
20	calcium	Ca	40.08	2.2×10^{6}	8.8×10^{7}	6.11×10^{4}
21	scandium	Sc	44.96	1.2×10^{3}	5.5×10^{4}	34.2
22	titanium	Ti	47.88	8.5×10^{4}	4.1×10^{6}	2400
23	vanadium	V	50.94	1.0×10^{4}	5.3×10^{5}	293
24	chromium	Cr	52.00	4.8×10^{5}	2.5×10^{7}	1.35×10^{4}
25	manganese	Mn	54.94	3.4×10^{5}	1.9×10^{7}	9550
26	iron	Fe	55.85	3.2×10^{7}	1.8×10^{9}	9.00×10^{5}
27	cobalt	Co	58.93	8.1×10^{4}	4.8×10^{6}	2250
28	nickel	Ni	58.69	1.8×10^{6}	1.0×10^{8}	4.93×10^{4}
29	copper	Cu	63.55	1.9×10^{4}	1.2×10^{6}	522
30	zinc	Zn	65.39	4.5×10^{4}	2.9×10^{6}	1260
31	gallium	Ga	69.72	1.3×10^{3}	9.4×10^{4}	37.8
32	germanium	Ge	72.61	4.3×10^{3}	3.1×10^{5}	119

Atomic number, Z	Name	Chemical symbol	Relative atomic mass, A_r	Solar System abundance		CI chondrite abundance by number
				by number	by mass	
33	arsenic	As	74.92	2.3×10^2	1.8×10^4	6.56
34	selenium	Se	78.96	2.2×10^3	1.8×10^5	62.1
35	bromine	Br	79.90	4.3×10^2	3.4×10^4	11.8
36	krypton	Kr	83.80	1.7×10^3	1.4×10^5	45
37	rubidium	Rb	85.47	2.5×10^2	2.1×10^4	7.09
38	strontium	Sr	87.62	8.5×10^2	7.5×10^4	23.5
39	yttrium	Y	88.91	1.7×10^2	1.5×10^4	4.64
40	zirconium	Zr	91.22	4.1×10^2	3.7×10^4	11.4
41	niobium	Nb	92.91	25	2.3×10^3	0.698
42	molybdenum	Mo	95.94	91	8.7×10^3	2.55
43	technetium	Tc[a]	98.91	–[b]	–[b]	–[b]
44	ruthenium	Ru	101.07	66	6.8×10^3	1.86
45	rhodium	Rh	102.91	12	1.3×10^3	0.344
46	palladium	Pd	106.42	50	5.3×10^3	1.39
47	silver	Ag	107.87	17	1.9×10^3	0.486
48	cadmium	Cd	112.41	58	6.5×10^3	1.61
49	indium	In	114.82	6.6	7.6×10^2	0.184
50	tin	Sn	118.71	140	1.6×10^4	3.82
51	antimony	Sb	121.76	11	1.3×10^3	0.309
52	tellurium	Te	127.60	170	2.2×10^4	4.81
53	iodine	I	126.90	32	4.1×10^3	0.90
54	xenon	Xe	131.29	170	2.2×10^4	4.7
55	caesium	Cs	132.91	13	1.8×10^3	0.372
56	barium	Ba	137.33	160	2.2×10^4	4.49
57	lanthanum	La	138.91	16	2.2×10^3	0.4460
58	cerium	Ce	140.12	41	5.7×10^3	1.136
59	praseodymium	Pr	140.91	6.0	8.5×10^2	0.1669
60	neodymium	Nd	144.24	30	4.3×10^3	0.8279
61	promethium	Pm[a]	146.92	–[c]	–[c]	–[c]
62	samarium	Sm	150.36	9.3	1.4×10^3	0.2582
63	europium	Eu	151.96	3.5	5.3×10^2	0.0973
64	gadolinium	Gd	157.25	12	1.8×10^3	0.3300
65	terbium	Tb	158.93	2.1	3.4×10^2	0.0603
66	dysprosium	Dy	162.50	14	2.3×10^3	0.3942
67	holmium	Ho	164.93	3.2	5.2×10^2	0.0889
68	erbium	Er	167.26	8.9	1.5×10^3	0.2508
69	thulium	Tm	168.93	1.3	2.3×10^2	0.0378
70	ytterbium	Yb	170.04	8.9	1.5×10^3	0.2479
71	lutetium	Lu	174.97	1.3	2.3×10^2	0.0367

Atomic number, Z	Name	Chemical symbol	Relative atomic mass, A_r	Solar System abundance		CI chondrite abundance by number
				by number	by mass	
72	hafnium	Hf	178.49	5.3	9.6×10^2	0.154
73	tantalum	Ta	180.95	1.3	2.4×10^2	0.0207
74	tungsten	W	183.85	4.8	8.8×10^2	0.133
75	rhenium	Re	186.21	1.9	3.5×10^2	0.0517
76	osmium	Os	190.2	24	4.6×10^3	0.675
77	iridium	Ir	192.22	23	4.5×10^3	0.661
78	platinum	Pt	195.08	48	9.3×10^3	1.34
79	gold	Au	196.97	6.8	1.3×10^3	0.187
80	mercury	Hg	200.59	12	2.5×10^3	0.34
81	thallium	Tl	204.38	6.6	1.4×10^3	0.184
82	lead	Pb	207.2	110	2.3×10^4	3.15
83	bismuth	Bi	208.98	5.1	1.1×10^3	0.144
84	polonium	Po[a]	209.98	–[c]	–[c]	–[c]
85	astatine	At[a]	209.99	–[c]	–[c]	–[c]
86	radon	Rn[a]	222.02	–[c]	–[c]	–[c]
87	francium	Fr[a]	223.02	–[c]	–[c]	–[c]
88	radium	Ra[a]	226.03	–[c]	–[c]	–[c]
89	actinium	Ac[a]	227.03	–[c]	–[c]	–[c]
90	thorium	Th[a]	232.04	1.2	2.8×10^2	0.0335
91	protoactinium	Pa[a]	231.04	–[c]	–[c]	–[c]
92	uranium	U[a]	238.03	0.32	7.7×10^1	0.0090
93	neptunium	Np[a]	237.05	–[c]	–[c]	–[c]
94	plutonium	Pu[a]	239.05	–[c]	–[c]	–[c]
95	americium	Am[a]	241.06	–[c]	–[c]	–[c]
96	curium	Cm[a]	244.06	–[c]	–[c]	–[c]
97	berkelium	Bk[a]	249.08	–[c]	–[c]	–[c]
98	californium	Cf[a]	252.08	–[c]	–[c]	–[c]
99	einsteinium	Es[a]	252.08	–[c]	–[c]	–[c]
100	fermium	Fm[a]	257.10	–[c]	–[c]	–[c]
101	mendelevium	Md[a]	258.10	–[c]	–[c]	–[c]
102	nobelium	No[a]	259.10	–[c]	–[c]	–[c]
103	lawrencium	Lr[a]	262.11	–[c]	–[c]	–[c]

[a] No stable isotopes.

[b] Detected in spectra of some rare evolved stars.

[c] Too scarce to have been detected beyond the Earth (i.e. abundance value not well known).

ACKNOWLEDGEMENTS

Grateful acknowledgement is made to the following sources for permission to reproduce material in this book:

Cover photos: Background image: Juraj Toth (Comenius U. Bratislava), Modra Observatory; Thumbnail images: (from the left) first, second and fourth images NASA; third image © 1999 Photo Disc Inc.

Figure 1.1 Robert Thom; *Figure 1.2* © The Natural History Museum, London; *Figures 1.3–1.5 and 1.7a* Reprinted from *Origins of life on the Earth and in the Cosmos,* Zubay, G., Copyright © 2000, with permission of Elsevier Science; *Figure 1.10* 'The Pulse of Life', Lowenstein, J. M & Zihlman, L. in Gribben, J. (ed.), *A Brief History of Science,* Weidenfeld and Nicholson Ltd; *Figure 1.11* © Dan Sudia; *Figure 1.13* Courtesy of I. D. J. Burdett; *Figure 1.14* From *Biogenesis: Theories of Life's Origins* by N. Lahav, copyright © 1999 by Oxford University Press, Inc. Used by permission of Oxford University Press, Inc; *Figure 1.15* © European Space Agency; *Figure 1.16* Adapted from de Muizon et al., 1986 in Pendleton, Y. J. and Tielens, A. G. G. M. (1997) *From Stardust to Planetesimals,* Astronomical Society of the Pacific; *Figures 1.17 and 1.18* © NASA; *Figure 1.26* © Anglo-Australian Observatory, photography by David Malin; *Figure 1.30a* Sourced from www.angelfire.com; *Figure 1.30b* Sourced from University of Hamburg website, www.biologie.uni-hamburg.de; *Figure 1.31* Dr. David Deamer, UC Santa Cruz; *Figure 1.36* D. Thomson/GeoScience Features; *Figure 1.38* Lahav, N. (1999), *Biogengesis,* courtesy of Noam Lahav.

Figures 2.1, 2.3, 2.4, 2.23, 2.25: © NASA; *Figure 2.2* © NASA George Curruthers; *Figure 2.9* © The Natural History Museum; *Figure 2.15* Hammersley Iron Pty Ltd; *Figure 2.16* Andrew A. Knoll; *Figure 2.17* Professor J. W. Schopf; *Figure 2.18* Commonwealth Palaeontological Collections of the Australian Geological Survey; *Figure 2.20* Schidlowski, M. 'A 3,800-million-year isotopic record of life from carbon in sedimentary rocks', *Nature,* **333**, p.316, Macmillan Magazines. Reprinted with permission from the author; *Figures 22a and b* Simon Conway Morris, University of Cambridge; *Figure 22c and d* Peter Crimes, Liverpool University.

Figure 3.3a–c Mary Evans Picture Library; *Figure 3.4* http://www.nasm.si.edu/ceps/etp/mars/marsimg/mars_lowell.jpg. Copyright © Smithsonian, National Air and Space Museum; *Figures 3.5, 3.10, 3.11a–c, 3.12, 3.13, 3.18* NASA; *Figure 3.9* NASA/JPL/Malin Space Science Systems/USGS Flagstaff; *Figure 3.21a* Copyright © Proszynski I S-ka SA 1999–2001. Wszystkie prawa zastrzezone; *Figure 3.21b–e* Douglas A. Kurtze, North Dakota State University, Department of Physics; *Figure 3.22a and b* Everett Gibson (NASA/JSC); *Figure 3.23* Photograph courtesy of Stephen Hyde.

Figures 4.1, 4.2 and 4.8 © Science Photo Library; *Figure 4.3* © National Portrait Gallery; *Figures 4.4–4.6, 4.7, 4.10a, 4.11–4.14, 4.17, 4.19, 4.21–4.26, 4.28–4.29 and 4.33* © NASA; *Figure 4.10b* © Calvin J. Hamilton; *Figures 4.16 and 4.20* USGS/Cascades Volcano Observatory; *Figure 4.18* From *Satellites of the Outer Planets,* 2nd edition by David A. Rothery, copyright © 2000 by David A. Rothery. Used by permission of Oxford University Press, Inc; *Figure 4.30* © Rob Wood/Wood Ronsaville Harlin, Inc; *Figure 4.32* Image courtesy of Marc. W. Buie/Lowell Observatory.

INDEX

T

U

V

W