LIVING COMETS

Fred Hoyle
and
Chandra Wickramasinghe

UNIVERSITY COLLEGE CARDIFF PRESS
1985

First published in 1985 in Great Britain by
University College Cardiff Press
P.O. Box 78, Cardiff, CF1 1XL
United Kingdom

ISBN 0 906649 79 0

Printed in Great Britain by J.W. Arrowsmith Ltd., Bristol

CONTENTS

LIST OF FIGURES

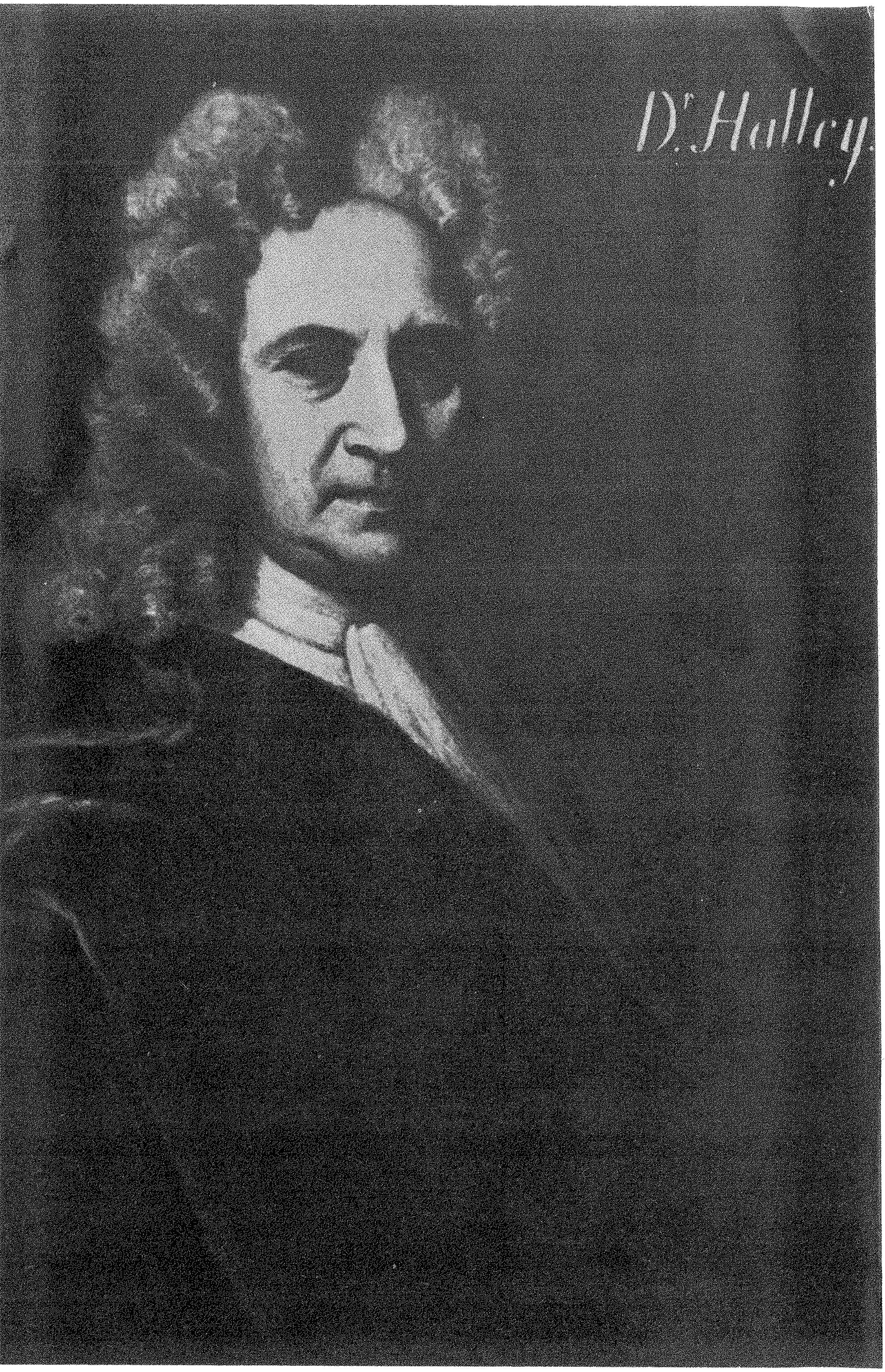
LIST OF TABLES

PHOTOGRAPHIC ILLUSTRATIONS PRECEDING CHAPTERS

Chapter

Acknowledgements

Photographs are reproduced by courtesy of NASA, the Munich Planetarium, the Lick Observatory, Prof. H. Pflug, Dr. R.B. Hoover, Dr. S. Al-Mufti and the Royal Astronomical Society.


Dr. Halley.

CHAPTER 1

COMET HALLEY IN ASTRONOMICAL HISTORY

Figure 1 shows what in the physical sciences we all learn at one quick gulp, that apart from small perturbations due to other planets, each planet P moves around the Sun in an elliptic orbit with the Sun at one of its foci. The geometrical equation of such an ellipse is:

$$r = \frac{a\,(1 - e^2)}{1 + e \cos \theta} \tag{1}$$

where the semi-major axis a is a size parameter and the eccentricity e is a shape parameter. For e small, the ellipse is like a circle but for e only a little less than unity the ellipse is highly elongated, with a pencil-like shape that is characteristic of the orbits of most comets.

The connection between geometry and physics comes when the parameters a, e are related to observable quantities, explicitly to the period T required by the planet to make a complete circuit of the ellipse, and to the speed of motion of the planet at particular points in the ellipse. T^2 is proportional to a^3, Kepler's third law, with the constant of proportionality related to the gravitational constant G and the mass M of the Sun by:

$$T^2 = \frac{4\,\pi^2\,a^3}{GM} \tag{2}$$

Thus the period depends only on the scale parameter a and is independent of the eccentricity e. The speed of motion at the perihelion point is $[GM(1+e)/a(1-e)]^{1/2}$, and at the aphelion point is $[GM(1-e)/a(1+e)]^{1/2}$. If from observations, we know either of these speeds, together with T, we therefore know both a and e, and hence the shape of the orbit. To know the complete orbit we need to specify further the plane of the orbit with respect to a background of distant objects (e.g. distant galaxies) and also to specify the direction of the apse line from perihelion to aphelion with respect to this background.

A further remarkable result one learns in the lightening rush of modern education is that if one multiplies r by the component of the velocity of P perpendicular to SP the result h, the angular momentum, is the same wherever P happens to be in the orbit,

$$h^2 = a\,(1 - e^2)\,GM. \qquad (3)$$

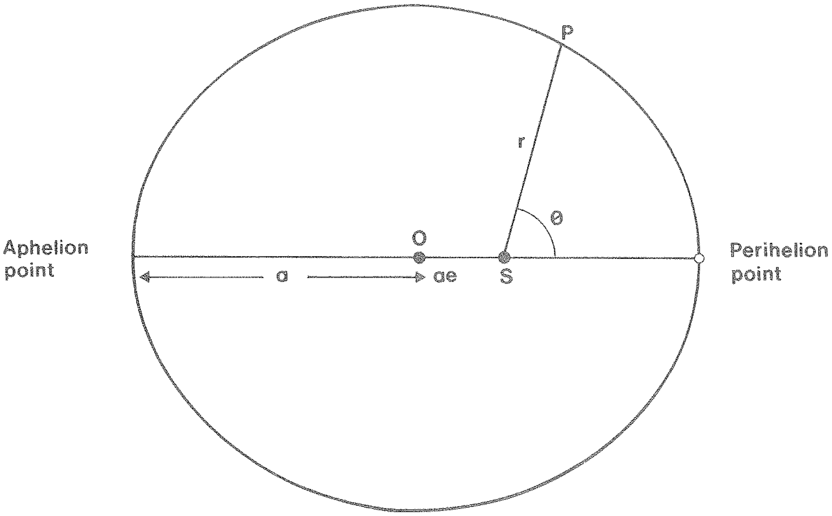


Fig. 1. The angle θ is measured from the perihelion point. The distance from the centre O to the focus S is ae, and the distance from S to P is r.

Since, as well as knowing the relations (1) to (3), we emerge today from a first-year university course in dynamics also knowing how (1) to (3) are to be derived from Newtonian theory, we have the perception that, if only we had lived five hundred years ago, what great people we would have been. The truth, however, is that (1) to (3) are not simple at all. Their acquisition took upwards of two millennia, with Comet Halley appearing as the last step in a truly great achievement. It was because of the role played by Comet Halley that we today accord special notice to this particular comet. Figure 2 shows the form of the basic data that ancient astronomers had to work from. Even recording the data in any accessible form must have been a really major problem, especially as writing materials in ancient times were scarce.

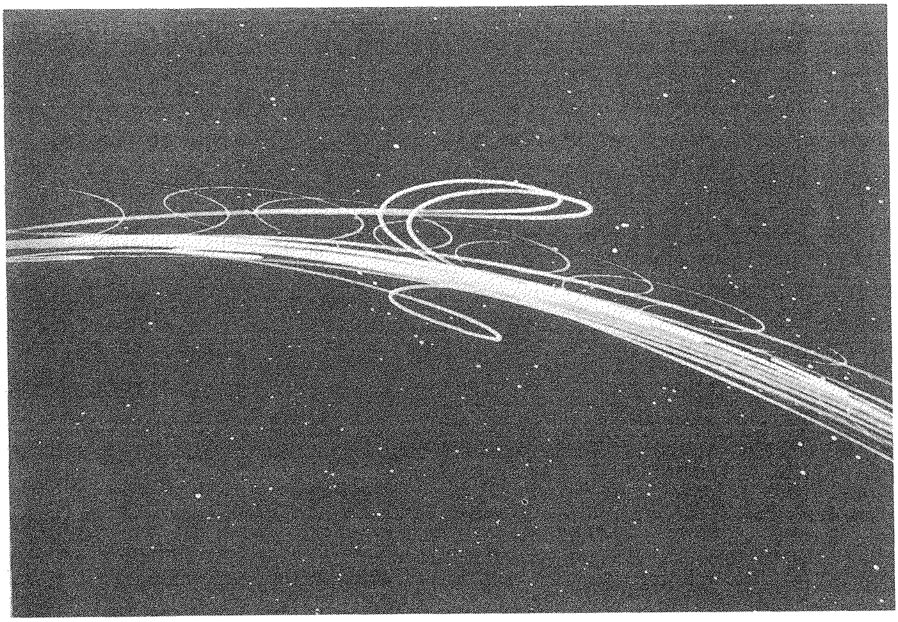


Fig. 2. This photograph was taken in the Munich planetarium, where the motions of the planets over a period of 17 years are simulated. It shows the apparent loops in the orbits of Mars, Venus, Mercury, Jupiter and Saturn.

The Greeks were already sophisticated geometers, as we know from Euclid's *Elements*. So it readily fell within their perceptions that planetary motions of the form shown in Figure 3 could explain the data in a semi-quantitative way, provided the rates of revolution of the Sun about the Earth, and of Mercury, Venus, and a typical outer planet O about the Sun were suitably adjusted. Mercury had to go around fastest, then Venus, then the Sun, then O. This would lead to advancing looped motions for Mercury and Venus, and to a nearly balanced forward and backward apparent motion for O, which were seen to be the main qualitative features of the observed motions.

A major stroke of perception, usually credited to Aristarchus of Samos, was to realize that the moving relationships in Figure 3, the kinematics of the situation, are the same as if the Earth also moved around the Sun, as in Figure 4, with the Earth's rate of revolution in Figure 4 the same as the Sun's rate in Figure 3.

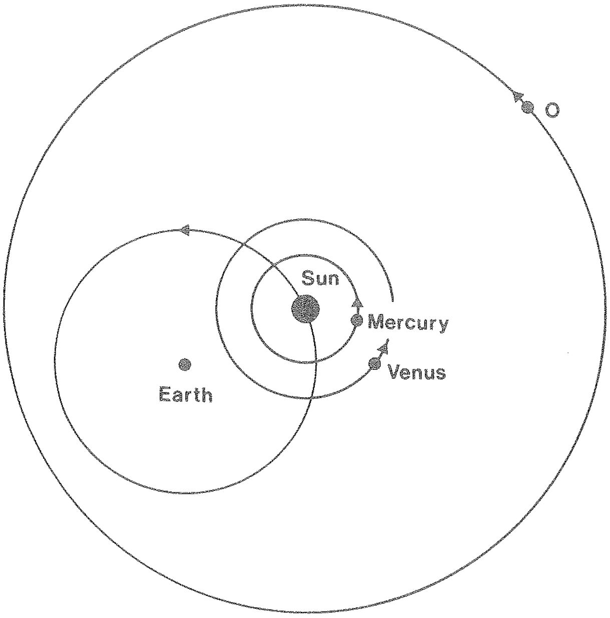


Fig. 3. An Earth-centred planetary geometry to account for the looped motions of Mercury and Venus.

Once it is realized that Figures 3 and 4 provide exactly the same geometrical relationship, the disposition of the human mind is to ask which one is 'true'? Nowadays, we know from the general theory of relativity that neither is true in any absolute sense. We also know, however, that to arrive at the general theory of relativity *via* Newton's theory, Figure 4 has to be preferred, since Figure 3 is an accelerated frame of reference in Newtonian terminology, an accelerated frame in which the inverse square law for gravitation cannot be used. No such considerations were available to the Greeks, who had no knowledge that a dynamical theory in this modern sense would ever be possible. The Greeks could therefore settle the biting question of which picture was 'true' only by prejudice. They chose the prejudice that because we do not feel any motion of the Earth, it must be Figure 3 with the Earth at rest that is 'true'.

It is interesting to notice the psychological trickery whereby Figure 4 was

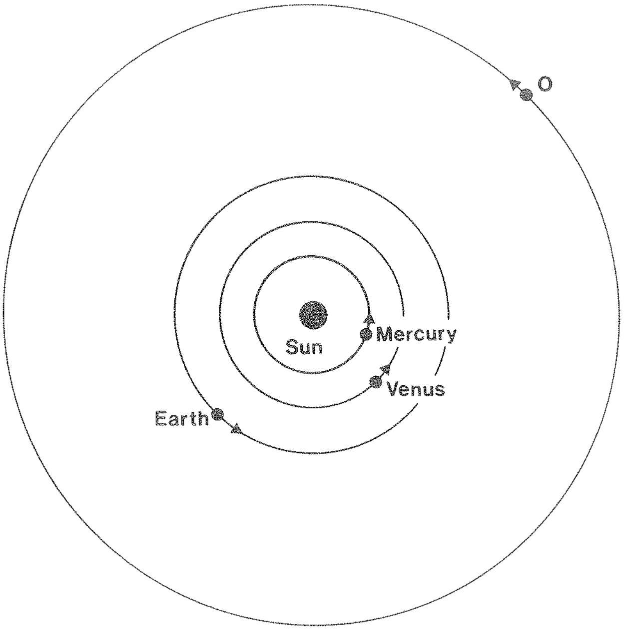


Fig. 4. A heliocentric planetary geometry.

denigrated in favour of Figure 3, especially as the same mode of trickery is employed throughout expensively funded modern science. While Figure 4 gives a simple semi-quantitative description of the facts it could be seen not to give a complete description, even as the facts were known in times B.C. The observations showed that the orbits of the different bodies could not be quite coplanar, and while this requirement applied to both ways of looking at the matter, it showed there were inevitable complexities, which served to take some of the shine off the simplicity argument in favour of Figure 4. The observations were accurate enough to prove that the orbits could not be circles. This was even the case for the Earth-Sun orbit, where the departure from circularity is small, with the eccentricity $e = 1.0167$. Even in ancient times, this small deviation from strict circularity of the Earth's orbit could be seen from the circumstance that the time involved from the spring equinox to the autumnal equinox is a little longer than the interval from the autumnal equinox to the spring equinox; there is about a three-day difference. For Mars especially with $e = 0.0933$, the discrepancy was very evident.

Whatever adjustment was made to the orbits to improve the relation of theory and observation could have been made equally to Figure 3 or to Figure 4. Logically speaking therefore, no way existed here for making a decision between the two figures, or so you might think. But this would be to reckon without the follow-my-leader syndrome on which nine-tenths of scientific belief rests. If Balldragon makes an admittedly significant contribution towards the solution of Problem A, then Balldragon's opinion on Problem B becomes widely accepted even though Balldragon has made no significant contribution at all to Problem B.

The simplest form of orbital construction due to Ptolemy is shown in Figure 5. With CP = a, AC = CS = ae, and with the line from A to the planet P turning at a suitable uniform rate, the motion of P can be shown to be the same as for the elliptic orbit of Figure 1, the same in so far as terms of the first order in the eccentricity e are concerned. This first order equivalence was sufficient to match theory and observation to within the observational accuracy of the planetary positions as they were known in ancient times. With the authority that thus fell on his shoulders, Ptolemy came down heavily in favour of Figure 3, and so did succeeding educated folk for the next millennium and a quarter.

It is in keeping with the modern breathless educational process that most of us emerge from it thinking Copernicus simply advocated a return to Aristarchus. He did nothing of the sort. Otherwise hardly a soul would have listened. Nor was it sufficient for Copernicus to point out that Ptolemy's construction when added to Figure 4 would explain the facts equally well. What Copernicus had to do was to find an orbital construction of his own, which while mathematically equivalent to Ptolemy's, looked very different to the eye, as in Figure 6.

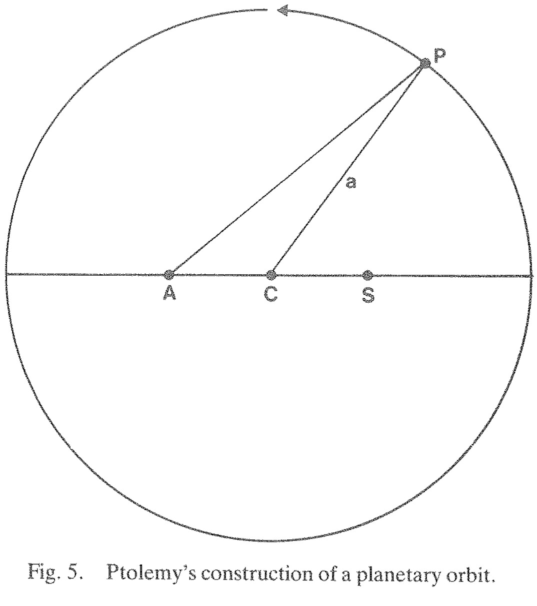


Fig. 5. Ptolemy's construction of a planetary orbit.

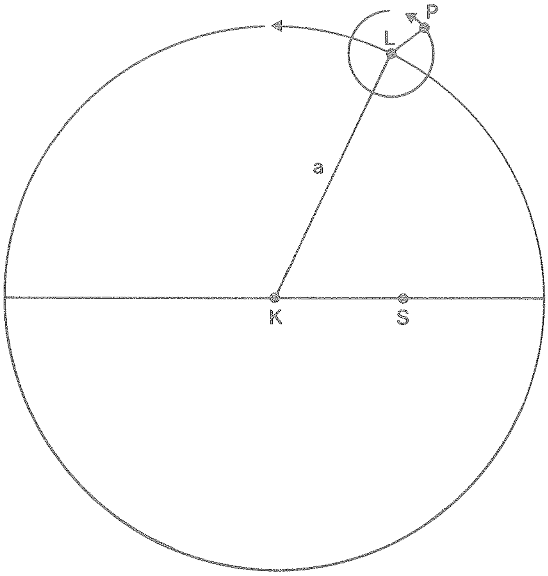


Fig. 6. Corresponding Copernican construction. Here L moves around K at the same rate as AP turns in Fig. 5. LP turns at twice the rate of LK.

Then with the authority that thus fell on his shoulders, and by arguing that Figure 6 was better than Figure 5 (because L turns uniformly about the centre K of the large circle whereas P in Figure 5 turns uniformly about the eccentric point A) Copernicus after many years of effort in finding Figure 6 at last felt able to recommend a preference for Figure 4, subject to the modifications of Figure 6. It is another of the innumerable mistakes inculcated by modern whizzbang education to believe that Copernicus had no sooner breathed the word than a sizeable number of people rushed to follow him. Nothing of the sort again. Copernicus' book, published in 1543, was just about kept alive in Germany, but it was not until ca.1600 that anything further really happened.

Progress depended on observations being made with improved accuracy, especially observations of Mars. Such observations were made by Tycho Brahe in the second half of the 16th century. Progress also depended on the observations, which were not openly published of course, coming into the hands of someone with the genius to use them. So a coincidence of fact and ability was needed, the facts of Tycho Brahe and the ability of Johannes Kepler. The coincidence was by no means accidental, for Kepler as a young man deliberately sought out Tycho and put up with almost endless insults from him in order to obtain the precious data, which he did on Tycho's death in 1601.

Scientific history tells us that, by working from Tycho's data, Kepler at last discovered that the planetary orbits are ellipses. History also tells us that the discovery was the overture to modern science. But when one asks exactly what Kepler did, confusion reigns. So here you have a situation that is represented to be perhaps the most critical step ever taken in science, without a clear explanation being given of what really happened.

Some years ago one of the present authors found himself in a position where conscience demanded that an effort be made to sort the position out. Finding difficulties in the most authoritative histories (*A History of Astronomy from Thales to Kepler*, J.L.E. Dreyer, Dover Publications Inc., 1953; *Kepler*, Max Casper, Abelard-Schuman, 1959), and finding out that other writers of the period had simply copied from these sources, he asked himself a non-historic question: Given the data available to Kepler, what could one actually do with it that would demonstrate the elliptic nature of the planetary orbits, explicitly the elliptic nature of the orbit of Mars? What was it possible to do with the data? With this question answered, it was then an easy matter to verify that the resulting method was at least consistent with the historic descriptions of what Kepler is said to have done.

It is a prerequisite that the period of revolution T of Mars around the Sun be known in advance with high accuracy. This was certainly so, because observational errors in the determination of T are not accumulated from one revolution to another. If the error for a single revolution was 1 per cent in T, the error for observations over ten revolutions would be 0.1 per cent in T, for a hundred revolutions would be 0.01 per cent in T and so on. This is because only two angular measurements are required, at the beginning and the end of the observations, irrespective of how many revolutions the observations may extend over. In effect, as the number of revolutions increases, the errors average to zero.

Figure 7 illustrates observations of Mars when the Earth is at points E_1, E_2, E_3, E_4, . . . of its orbit, these points being chosen from Tycho's observations so that the following conditions are satisfied:

(i) For one of the sets of observations Mars is in opposition, the point E_2 of Figure 7.
(ii) The time intervals between the observations for E_2 and those for each of the other points is an integral number of periods of revolution T of Mars.

Condition (ii) requires that the direction of the line from the Sun to Mars, SM be the same at the times when observations were made from E_1, from E_2, from E_3, from E_4, . . . Thus the direction of SM with respect to a fixed background of distant stars is the same and is known for every set of observations, since it is known for E_2. Since moreover the set of observations for E_1 also gave the directions of E_1M and E_1S, all the angles in the triangle SE_1M were known, and hence the ratio of the lengths SE_1 and SM could easily be calculated. It was similarly so for observations made at E_3, E_4, ... Hence the lengths of the radius vector from the Sun to the Earth were determined for a number of points on the Earth's orbit, determined in terms of the length of the radius vector from the Sun to Mars (as it was at the time of the opposition in question).

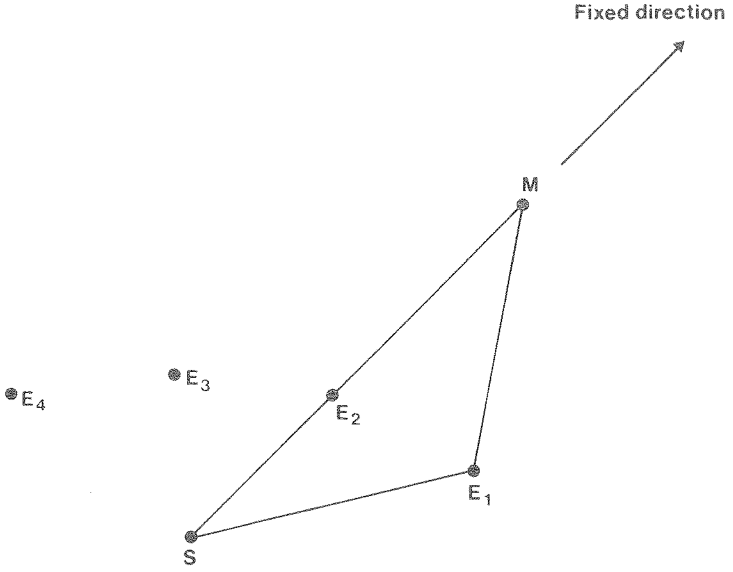


Fig. 7. Observations of Mars (M) from several positions E_1, E_2, E_3, E_4 of the Earth.

We are told in the histories that Kepler began by demonstrating, through the use of the oppositions of Mars, that the construction of Copernicus must be applied to the orbit of the Earth as well as to the other planets. (Copernicus had erred slightly by exempting the Earth from his construction.) This means Kepler first verified that the Earth's orbit is elliptical up to terms of the first order in the eccentricity e, which tells us that the data was probably good enough to work to the second order in e for Mars ($e = 0.0167$ for the Earth, $e = 0.0933$ for Mars). But how was the orbit of Mars to be obtained?

The answer to this crucial question was that Kepler had twelve cases like Figure 7 available, twelve oppositions of Mars, falling at different points on the orbit of Mars. Each case gave a ratio for the scale of the Earth's orbit to the heliocentric distance of Mars appropriate to the case in question. The heliocentric distances of Mars were all with respect to the same scale of the

Earth's orbit for positions on the Martian orbit where there had been oppositions and where the directions from the Sun to Mars were known with respect to distant stars. The final step, with the essential work now done, was to discover what kind of curve the points on the orbit of Mars served to define. The answer was an ellipse up to terms of the second order in the eccentricity e. Reference to histories now showed how confusion had been infused by a failure to explain clearly the essential part of the argument, and then by overdramatizing the comparatively easy last step, overdramatizing the final search for the curve defined by the calculated positions of Mars. That Kepler tried out this and that idea for the final curve is of little relevance compared with the basic idea summarized in Figure 7. The crowning error of the histories, however, has been to assert that Kepler determined the orbit of Mars to be an ellipse. He did nothing of the sort. Kepler showed the orbit to be the same as an ellipse up to terms of order e^2, but not to higher orders. Moreover, the demonstration was subject to an axiom which turned out to be substantially true, but which might not have been true. The axiom was that the orbits of the planets are closed curves. Without this axiom the method summarized in Figure 7 would not work. From Greek times onwards every curve had been drawn as closed, closed circles, so that perhaps the idea was so ingrown that everybody automatically accepted it.

The story as it has been usually told proceeds as follows. With the appreciation that a particle moving with speed v in a circular path of radius a experiences an outward acceleration v^2/a, it followed that in order to hold such a particle steadily in such a path there would need to be a balancing inwardly directed acceleration v^2/a. This was well understood ca. 1670 by Christiaan Huygens in Holland and by the Fellows of the Royal Society in London. Since a particle thus moving steadily completes the circle in time $T = 2\pi a/v$, one could thus say the needed inward acceleration was $4\pi^2 a/T^2$. If now one supposed planets to be held in their orbits by a radial force from the Sun, then in the approximation of a circular orbit the force would need to produce an inward acceleration $4\pi^2 a/T^2$. Next use Kepler's third law (deduced again from Tycho's observations) that T^2 is proporational to a^3 and the inward acceleration needs to be proportional to $1/a^2$, an inverse square law. So we are told that a perception arose that the Sun exerted an attractive force on a planetary body that was proportional to the inverse square of the distance of the body from the Sun, and in England Robert Hooke, Christopher Wren, and Edmund Halley are said to have gone around saying as much. Hooke, in particular, is known to have written to Newton in these terms.

In 1684, Edmund Halley visited Newton in Cambridge with the pertinent question: What would be the orbit of a planet moving under the attraction of a force from the Sun that varied according to the inverse square of the distance of the planet from the Sun? To this question, Newton is said to have answered that he had already calculated the matter,and that the orbit would be an ellipse. Overjoyed at this answer, Halley urged Newton to publish his solution at once. But it was not until two years later that Halley received Newton's encyclopaedic *Philosophiae Naturalis Principia Mathematica*, which went far beyond the promised answer to the question of 1684.

Before we come to some rather unlikely aspects of this story, let us write down the solution to Halley's question, in the form in which it is given nowadays, which is to say in Newtonian calculus instead of Newton's own more difficult geometrical demonstration. Start by taking a general function f(r) for the attractive acceleration produced on a small test body at distance r from the Sun. In polar co-ordinates the equations of motion are

$$\ddot{r} - r\,\dot{\theta}^2 = -\,f(r) \qquad (4)$$

$$r\,\ddot{\theta} + 2\,\dot{r}\,\dot{\theta} = 0 \qquad (5)$$

After multiplying by r, equation (5) integrates to give

$$r^2\,\dot{\theta} = \text{constant} = h, \text{ say.} \qquad (6)$$

Eliminating $\dot{\theta}$ between (4) and (6),

$$\ddot{r} - \frac{h^2}{r^3} = -\,f(r) \qquad (7)$$

Put $u = 1/r$ and note that $\dfrac{d}{dt} \equiv hu^2\,\dfrac{d}{d\theta}$

With these substitutions (7) is easily transformed to

$$\frac{d^2u}{d\theta^2} + u = \frac{f(u)}{h^2u^2}\,. \qquad (8)$$

Setting $f(u) = GMu^2$, as in the usual statement of the inverse square law, the right hand side of (8) is simply the constant GM/h^2, and the solution of (8) is then

$$u = \frac{GM}{h^2}(1 + e\cos\theta), \tag{9}$$

where e is a positive constant of integration, and a second constant of integration has been suppressed by measuring the angle θ to be such that $\theta = 0$ at the perihelion point. Putting $h^2 = a(1 - e^2)GM$ as in (3), (9) is the same as (1), namely the ellipse of Figure 1.

At this point one notices the way in which a theory can suddenly soar above the empirical facts on which it has been based. The result (9) holds regardless of the value of e, except that because the angle θ has by choice been taken as zero at the minimum value of r, e is required to be positive. Thus we can have $e > 1$, when (9) is the equation of a hyperbola, without the test particle needing to be bound to the Sun at all. Cases with e slightly less than unity give exceedingly elliptic orbits, with a ratio $(1 + e)/(1 - e)$ for the aphelion to perihelion distances, a result that was plainly of interest to the orbits of those comets that come close to the Sun at perihelion and then recede to much greater distances towards their aphelia.

Cometary orbits were only crudely documented in the 17th century, but a search by Halley of what was then available showed that three which had appeared in the years 1531, 1607 and 1682 had interestingly similar orbits. Could they really be the same body revolving round and round the Sun as the planets do? If so, Kepler's empirical basis for Newton's theory would be widened and deepened at a single stroke. When e is close to unity it has no meaning of course to speak of successive approximations in powers of e , as had been done for the planetary orbits up to terms of order e^2. For comets of high ellipticity all powers of e are in there together, and if the theory still remained valid the overwhelming implication would be that Kepler's identification of the planetary orbits with ellipses must still be correct at the higher powers e^3, e^4, ... The situation was that, if in the latter years of the 17th century, one believed in Kepler's perfect demonstration of elliptic orbits (the way dubious history has written it) then the comet which subsequently became known as Halley's comet was no more than a speck of icing on the Newtonian cake. But if one believed all one was then entitled to believe, that the observational demonstrations up to the 1680s had been only approximations for nearly circular orbits, the possibility of the apparitions of 1531, 1607, and 1682 being appearances of the same test body had all the elements of a critical experiment.

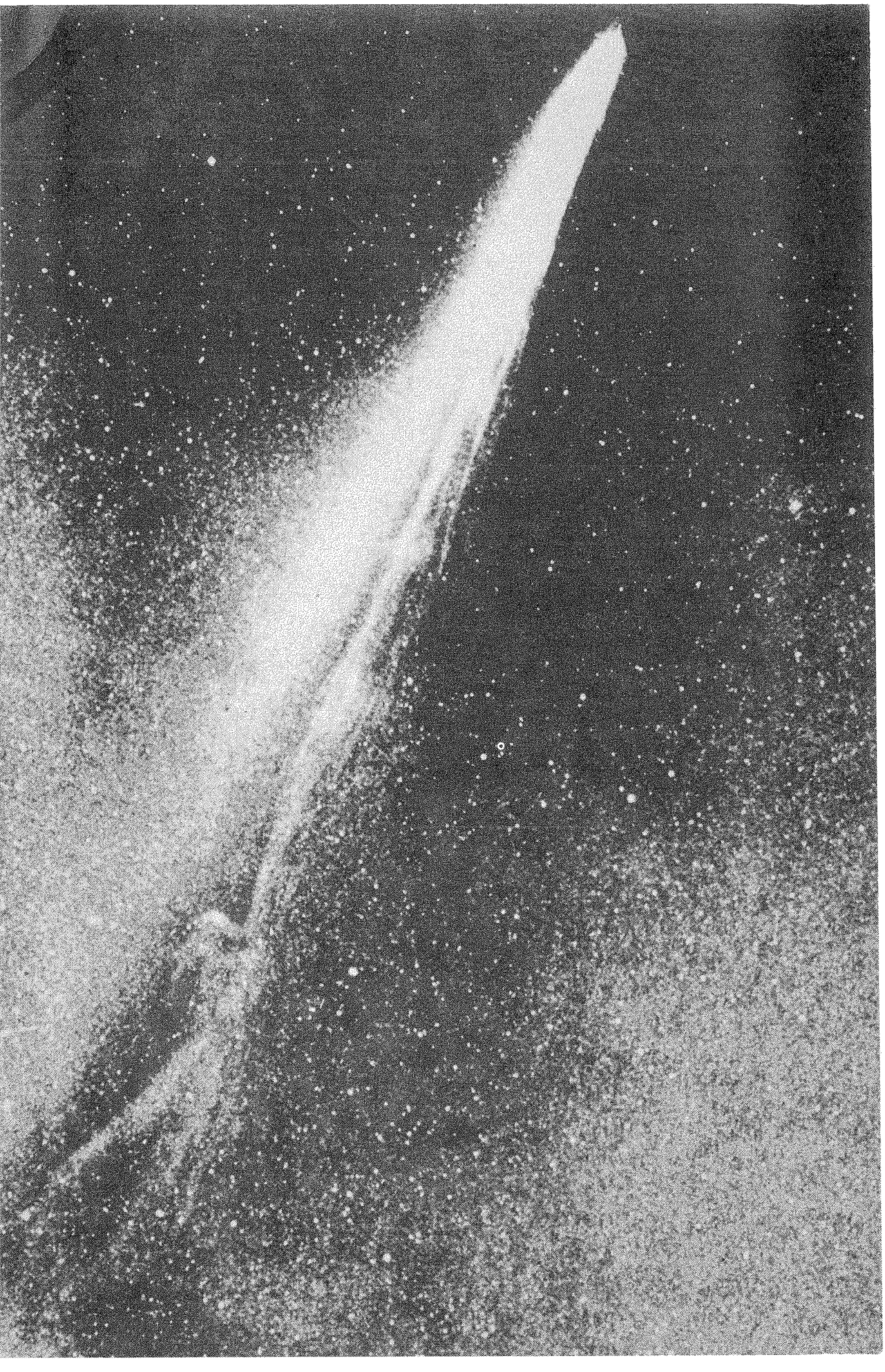
The time interval between the apparitions of 1531 and 1607 was close to 76 years, while the interval between the apparitions of 1607 and 1682 was close to 75 years. However, test bodies in highly elliptic orbits are subject to disturbances from the planets, especially Jupiter in this case, to an extent that is far greater than for test bodies like Mars on nearly circular orbits. An exact equality of the time intervals was not therefore to be expected. This complication prevented the matter from being driven home there and then. The way to proceed was to add a further interval of say 75½ years to the time when the comet of September 1682 was at its perihelion point. This put the next expected return to perihelion of the comet,if indeed the apparitions had all been of the same object, in the spring of 1758. As it turned out the comet, Halley's comet as it has become justly called, reached its perihelion point on 11 June, 1758. One can easily see why, following this triumphant return of 1758, mathematicians would have felt impelled to investigate the consequences of Newton's theory in the great detail which Newton himself actually gave in the *Principia*, the 'System of the World' as it became called. But why would Newton have felt impelled from 1684 to 1687 to make such an effort, an effort so intense as to bring him to the edge of breakdown? This is a question we do not recall seeing discussed. A skilled experimentalist like Newton could hardly have been deceived into thinking Kepler's result was perfect, and in any case Halley as an experienced astronomer would have known the limited extent to which the observations available to Kepler could be pushed. In the circumstances, one would have expected Newton to content himself by quickly giving the proof that Halley had asked for in 1684, and leaving it there for the moment.

An idealistic answer, which doubtless would have been acceptable to the 19th century, might be that Newton was an inspired genius who perceived the correct path ahead of its demonstration. This does not sound very convincing, however, since today we know that the general theory of relativity for weak gravitational fields, while approximately the same as Newton's theory, has an inverse cube dependence for the gravitational force that is added to a larger inverse square dependence. It was just good luck that gravitation in the solar system happens to be weak enough for this 'post-Newtonian' addendum to be omitted (except for its effect in turning the apse line of the orbit of Mercury). So far as anything known to Newton was concerned,the post-Newtonian addendum could have vitiated Kepler's ellipses at the higher approximations of e^3, e^4, ...

Another possibility is that, already in 1684, Newton had decided the cometary apparitions of 1531, 1607 and 1682 were from one and the same body, in which case the importance of Halley's comet to scientific history would be much greater than it is usually represented to be, for it would have played an important role in the appearance of the *Principia* itself.

A third possibility turns on equation (8), which shows that orbits are closed repeating paths only for the inverse square law. If Newton, like Kepler before him, believed the orbits to be closed, the situation would have seemed to him both elegant and decisive. This would of course put a judgement not based on experience in control of the situation, a judgement that turned out luckily only because the post-Newtonian approximation of the relativity theory happens to be so weak in the Solar System.

Even to this day gravitation remains an empirical phenomenon. Presumably the time will come when gravitation will appear as a consequence of particle physics at a deeper level than exists today. The expectation would be for the best of the empirical theories, the general theory of relativity, to appear as a statistical average from a much more complex situation. That such a statistical average should turn out in its weak field limit to lead to closed orbits for the two-body problem has remarkable implications, rather as the chemical properties of the carbon atom have remarkable implications. It is the closed orbits, or nearly closed orbits, which permit gravitational resonance phenomena to occur, and these – as we shall see in a later chapter – may well play a critical role in the evolution of the outer regions of the Solar System and in the origin of comets, so affecting the incidence of volatiles and life on to the Earth, adding this remarkable property to a rapidly growing list of so-called anthropic properties, without which we could not exist, in this case appearing at a very fundamental level.



CHAPTER 2

EVAPORATED MATERIAL FROM COMETS

Comets are generally regarded by astronomers as compact more-or-less spherical objects with diameters of order 10 km. Except possibly for a recent observation of Comet IRAS-Araki-Alcock (1983d) by M.S. Hanner, D.K. Aitken, R. Knacke, S. McCorkle, P.F. Roche and A.T. Tokunaga (JPL Cometary Science Team, Preprint Series No.49, 1984, *Icarus*, in press) where such a central object may have been directly observed (on the favourable occasion when this comet passed at the exceptionally small distance of only 4.7 million kilometres from the Earth) comets are seen because of gas and dust evaporated from them, the gas and dust being susceptible to observation either directly or spectroscopically.

For a comet that is approaching the Sun, the distance out to which it can be detected observationally depends in part on whether the comet is in a known orbit or not, and on how hard astronomers look for it. In view of the celebrity of Halley's comet, with orbit shown in Figure 8, an exceptional effort has been made to obtain early detection in advance of the perihelion passage of 1986, with the result on this special occasion that Comet Halley was detected about 4 years before perihelion passage. Detection beyond the orbit of Jupiter as in this special case would be unusual for hitherto unknown comets, however, which tend to be first discovered while at heliocentric distances of 2 or 3 astronomical units (1 AU is the mean distance of the Earth from the Sun, $\sim 1.496 \times 10^8$ km) which is to say commonly of the order of a year before perihelion passage. Some new comets escape detection, however, until they are within the Earth's orbit, and discoveries have even been made during a solar eclipse of comets with perihelion distances $q < 1$ AU.

Gas and dust evaporate most copiously while a comet is close to perihelion passage, so that comets tend to be more visible on their way out from the Sun than on their way in. Besides which, new comets being already known on their way out are then more closely observed and by more observers. Thus the portion of a comet's orbit after perihelion is usually better determined than the portion before perihelion. Comets of long period P (eccentricity e very close to unity) are found in some cases to be moving outwards from the Sun along hyperbolic orbits ($e > 1$). This raised the controversial question early in the century of whether comets might be visitors from interstellar space. What had to be done to decide this matter was to calculate the effects of planetary

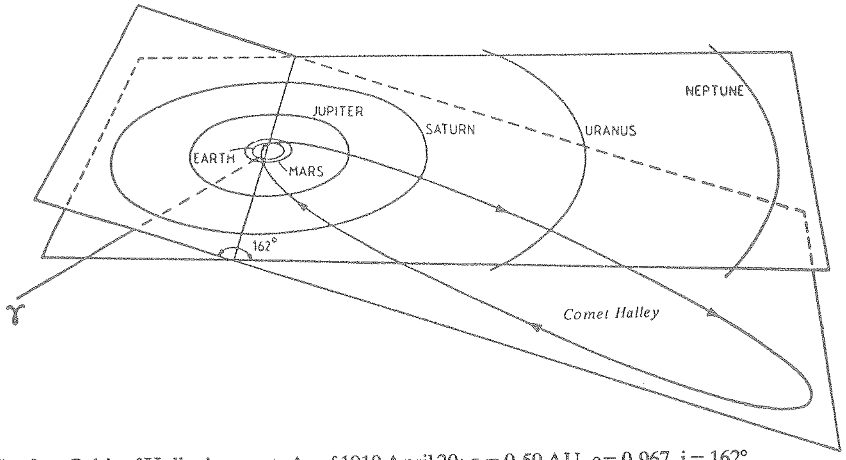


Fig. 8. Orbit of Halley's comet. As of 1910 April 20: $q = 0.59$ AU, $e = 0.967$, $i = 162°$, $\Omega = 58°$, $\omega = 112°$, $P = 76.1$ year, $a = 17.96$ AU, $v_p = 54$ km/s, $v_a = 0.9$ km/s.

perturbations, especially Jupiter, during the passage of such comets, because comets that come to the Sun along elliptic orbits could emerge with hyperbolic orbits due to perturbations. The problem was examined in great detail by E. Strömgren (*Pub.og. Meddel, Kobenh. Obs.*, No. 19, 1914), with the result that in all cases the incoming orbits had been elliptic ($e < 1$). Thus comets are Solar System objects.

In the absence of perturbations, the orbital motion of a comet is determined by equations (4), (5), which lead to (6) and (7). Putting $f(r) = GM/r^2$ in (7), multiplying by $\dot{r}$, and integrating gives

$$\frac{1}{2}\left(\dot{r}^2 + \frac{h^2}{r^2}\right) - \frac{GM}{r} = \text{constant}\,, \tag{10}$$

the well-known energy equation. When $e \simeq 1$, the constant in (10) is close to $-GM/2a$. This is because at aphelion $r \simeq 2a$ and $r^2 + h^2/r^2 \simeq 0$. Thus the energy equation for an unperturbed long-period comet is essentially

$$\frac{1}{2}\left(\dot{r}^2 + \frac{h^2}{r^2}\right) - \frac{GM}{r} = -\frac{GM}{2a}\,. \tag{11}$$

If now a comet experiences perturbations, while in the region of perihelion that changes its energy by amount ΔE, say, the energy equation as the comet returns to aphelion (with the perturbations having ceased) is

$$\frac{1}{2}\left(\dot{r}^2 + \frac{h^2}{r^2}\right) - \frac{GM}{r} = -\frac{GM}{2a} + \Delta E. \qquad (12)$$

Once again near aphelion, $\dot{r}^2 + h^2/r^2 \simeq 0$. Writing the new aphelion distance as $2(a + \Delta a)$, instead of the pre-perturbed value 2a, we have

$$\frac{\Delta a}{a\,(a + \Delta a)} = \frac{2\Delta E}{GM} . \qquad (13)$$

If the change a of the semi-major axis of the orbit is small compared to a itself, then to sufficient accuracy (13) can be written

$$\Delta\left(\frac{1}{a}\right) = -\frac{2\Delta E}{GM} . \qquad (14)$$

For comets that penetrate at perihelion inside the orbit of Jupiter, the magnitude of the right hand side of (14) is in general of order 5.10^{-4} AU^{-1}. Thus only if 1/a is appreciably larger than 5.10^{-4} AU^{-1} will Δ a be considerably less than a.

The inference from these simple considerations is that if 1/a is less than 5.10^{-4} AU^{-1}, i.e. $a > 2000$ AU, perturbations by Jupiter are capable of producing a gross change of the semi-major axis, even to the extent of changing an initially bound elliptic orbit into a hyperbolic orbit. These were the cases examined by E. Strömgren, which is to say comets with $a > 2000$ AU and period $P > \sim 10^5$ years. The present considerations show that the larger the initial semi-major axis, the greater the effect of perturbations (due mainly to Jupiter) can be on the distant regions of a comet's orbit.

A long-period comet with perihelion distance $q < 1$ AU typically evaporates $\sim 10^{13}$ gm of dust with particle sizes $\sim 10\ \mu$m (E.P. Ney, in *Comets*, ed. L.L. Wilkening, University of Arizona Press, p.323). The amount of gas is larger, $\sim 10^{14}$ gm. So long as the gas molecules remain neutral they continue to occupy a more or less spherical region around the nucleus, but when ionized by ultraviolet light from the Sun the gas becomes subject to non-gravitational forces that cause it to stream out of the coma in a direction away from the Sun. The non-gravitational forces arise from interaction with particle streams from the Sun, as was first suggested by L. Biermann, but probably not from direct particle-particle collisions as was originally thought. As M. Harwit and one of the present authors pointed out, the interaction comes from a magnetic field carried by the particle streams, which being unable to penetrate ionized gas of the coma tends to wrap itself around the coma, whence it squeezes out the ionized gas rather as toothpaste is squeezed out of a tube. The outcome is the gaseous component of the tail of the comet.

Small dust particles with dimensions ~ 10 μm and less are also subject to non-gravitational forces arising from solar radiation pressure, which causes them also to form a tail streaming out of the coma. Because the non-gravitational forces for gas and for dust particles are of different origins and nature, the gas tail and the dust tail are not coincident, although both point generally in directions away from the Sun. An example of the difference between the very irregular tail of the gas and the smooth tail of the dust is shown in Figure 9.

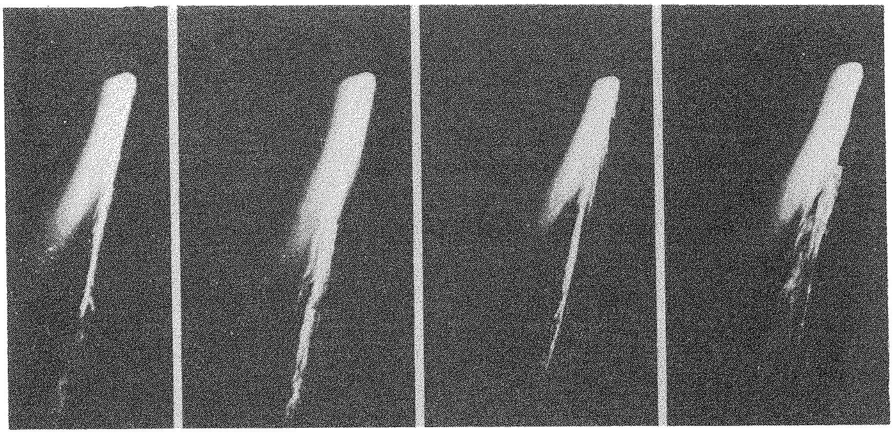


Fig. 9. The Comet Mřkos photographed on several days in 1957.

Figure 10 shows a schematic drawing of the coma with a shock front formed through the impact of a particle stream from the Sun, with typical molecules found in the coma, and with ionized molecules flowing out to add themselves to the gas tail. Table 1 gives a more complete listing of the molecules and elements which have been detected spectroscopically.

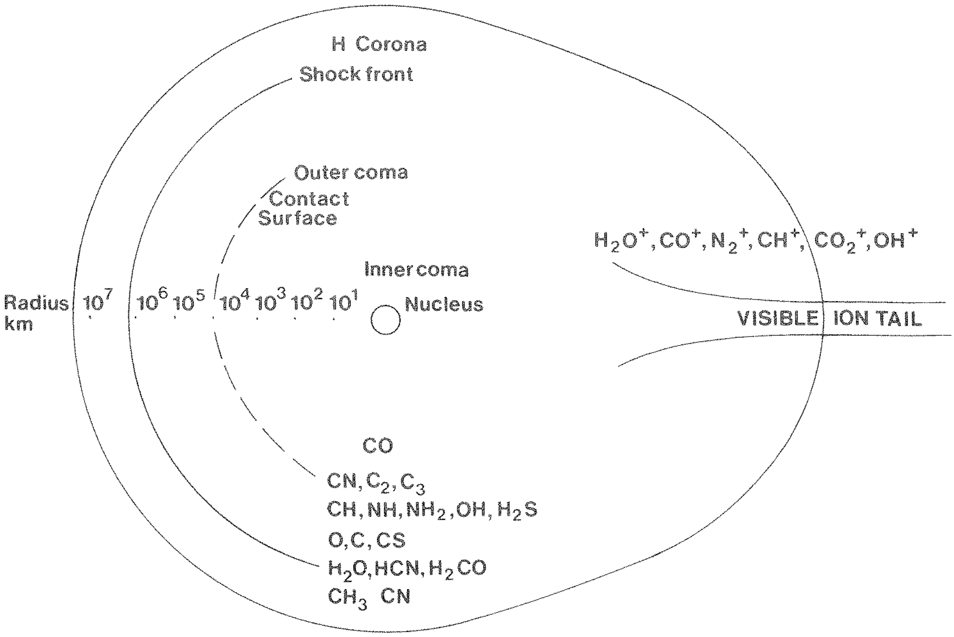


Fig. 10. Schematic principal gaseous features of a comet on a logarithmic scale of distance.

Table 1. *Atoms, ions, molecules, and molecular ions observed in comets*

Head	Tail
H, OH, H_2O, O, S C, C_2, C_3, CH, CN, CO, CS HCN, CH_3CN, NH, NH_3, Na, Fe, K, Ca, V, Cr, Mn, Co, Ni, Cu	CO^+, CO_2^+, H_2O^+, OH^+, CH^+, N_2^+, Ca^+, C^+, CN^+

For a generation at least it has been widely believed that comets are largely composed of water-ice. This is another example of the Balldragon syndrome mentioned in the previous chapter. Indeed we have still to find a single fact pointing significantly in favour of this point of view. The H_2O and H_2O^+ molecules present in comet comas and tails are detected only in weak bands.

The large quantities of H and OH could as well come from the breakdown of volatile organic molecules as from the dissociation of H_2O. Many comets evaporate orders of magnitude more rapidly than would a nuclear body composed of water-ice. To meet the need for rapid evaporation, comet nuclei were supposed to travel in their orbits surrounded by more or less permanent halos of water-ice grains, which having a larger surface area would be capable of evaporating more rapidly. Yet recent observations have shown no evidence for such halos (M.F.A. Hearn, E. Dwek, and A.T. Tokunaga, *Ap. J.*, *248*, 1981; L.E. Snyder, P.Palmer and C.W. Wade, *Ap.J.*, *269*, L21 1983). And a recent attempt to argue that grains evaporated from Comet Cernis gave a reflectivity at wavelengths near 3 μm that is characteristic of water-ice grains has been shown to be miscalculated (F. Hoyle, N.C. Wickramasinghe and M.K. Wallis, *Earth, Moon, and Planets*, in press 1985). These negative results for the water-ice hypothesis, together with the nature of the molecules actually detected (Table 1), suggest that the surface material of comets consists of organics of varying degrees of volatility.

We emphasize that this is not to say water is absent from the interiors of comets. What is being said is that the gas evaporated from comets, at any rate from those comets which evaporate rapidly and copiously, is not largely H_2O. As we shall see at a later stage, it is possible to conceive of a physical process occurring in the evolution of comets that would yield water in the interiors and volatile organics at their surfaces.

Most of the gas and dust evaporated from long-period comets eventually escapes out of the Solar System altogether to join the general interstellar medium. Organic molecules and radicals from comets must therefore contribute to the presence of such molecules in the interstellar gas. It is hard to say how significant the contribution could be, because a vast extrapolation from the estimate of $\sim 10^{14}$ gm per comet in the present day Solar System has to be made in order to arrive at a galaxy-wide quantitative estimate. For what it is worth, take one comet per annum per star in the galaxy, yielding collectively $\sim 10^{25}$ gm per annum for the whole galaxy. Over a time scale of 2×10^9 years, the cumulative effect would be $\sim 2 \times 10^{34}$ gm, which is a fraction $\sim 10^{-8}$ of the whole mass of interstellar gas, certainly a significant effect.

The possibility of a connection between organic molecules in comets and organic molecules in interstellar space has been noted by A.H. Delsemme, who has pointed out the suggestive similarities set out in Table 2. If there is a causal connection here one can ask for the sense of the causality. Is it comets $\rightarrow$ interstellar space in accordance with the discussion of the previous paragraph, or is it interstellar space $\rightarrow$ comets as one might have had at the time of origin of the Solar System? We think both, interstellar space $\rightleftharpoons$ comets. Organic molecules are constantly being disrupted by ultraviolet light, so that if such a

Table 2. *Chemical constituents of comets and the interstellar medium*

Products observed in comets	Possible precursors observed in interstellar space	Possible parent molecules observed in comets
H	Many observed molecules	Many observed molecules
O, OH, OH^+	$H_2O, H_2C{=}O, H_3C{-}OH$	H_2O, H_2O^+
CO^+	$CO, H_2C{=}O, H_3C{-}OH, NH{=}C{=}O$	
CH, CH^+	$H_2C{=}O, H_3C{-}OH, HC{\equiv}C{-}C{\equiv}N$	$H_3C{-}C{\equiv}N$
		$CH_3{-}C{\equiv}C{-}H$
C_2, C_3	$HC{\equiv}C{-}CH_3, HC{\equiv}C{-}C{\equiv}N$	$CH_3{-}C{\equiv}C{-}CN$
CN	$H{-}C{\equiv}N, H{-}N{=}C, HC{\equiv}C{-}C{\equiv}N$	$H{-}C{\equiv}N$
NH, NH_2	$NH_3, H_2N{-}CH{=}O, HN{-}C{=}O$	
N_2^+, CO_2^+	Other precursors or mechanisms involving collisions	

loop is to be kept going, somewhere in the loop there must be a corresponding input of organic molecules. We think this occurs in comets, which are thus the generators that keep the loop going. And since biological processes are overwhelmingly the most efficient way to convert inorganic materials to organics we think it is biological processes inside comets that really maintain the loop as an active on-going phenomenon.

For a comet of total mass $\sim 10^{18}$ gm a single perihelion passage in which $\sim 10^{14}$ gm material was evaporated would not be of much consequence to the comet as a whole. It would imply the evaporation of a surface layer only some 10 to 30 cm thick. However, such loss at many repeated perihelion passages must inevitably eat into the comet to a major degree. This happens with $\sim 10^4$ such passages, or if we suppose the evaporation rate falls off, say to $\sim 10^{13}$ gm as denudation proceeds, $\sim 10^5$ such passages.

The Earth is constantly embedded in a cloud of particles with dimensions $\sim 10\ \mu$m, a cloud that is flattened towards the plane of the ecliptic, a cloud which through the scattering of sunlight produces the so-called zodiacal light. From measurements of the brightness per unit area of the sky at varying angles from the Sun in directions along the plane of the ecliptic it is possible to make interesting inferences concerning the nature of the particles which produce this scattering. Such inferences are of considerable interest to the present discussion of comets because comets are the most likely source of origin for these zodiacal particles. The particles are unlikely to come from long-period comets, however, since material evaporated from long-period comets inevitably moves to great distances from the Sun, and as stated above most of it is eventually lost to interstellar space. Thus to explain the zodiacal cloud of particles we must look to short-period comets, there being a known class of about 50 such comets with periods less than 20 years.

The total mass of the zodiacal cloud of particles has been estimated to be ~ 10^{17} gm. This is for the portion of the cloud in the general region of the Earth's orbit. Because of the Poynting-Robertson effect which causes particles at ~ 1 AU with dimensions ~ 10 μm to spiral into the Sun on a time-scale of ~ 10^4 years, the cloud in the vicinity of the Earth must be replenished at a rate of ~ 10^{13} gm per annum. With two or three short-period comets passing through their perihelia each year, the average contribution in the form of dust particles required from the short-period comets would therefore be ~ 3×10^{12} gm per perihelion passage. While this would be a readily acceptable amount for long-period comets, it seems rather high for short-period comets such as P/Encke, which by now must have been considerably denuded. This raises the interesting possibility that one particular exceptional member of the group of short-period comets may make a disproportionately large contribution to the zodiacal cloud. This is the comet P/Schwassmann-Wachmann I, which quite atypically for a comet moves in an orbit of low eccentricity, $e = 0.132$, not much more elliptic than the orbit of Mars.

In the present century, P/Schwassmann-Wachmann I has undergone irregular major outbursts at intervals of ~ 25 years, roughly one outburst per revolution in its orbit. For such outbursts to be readily visible they must be large, since the semi-major axis for the orbit is 6.4 AU. Even at perihelion distance $q = 5.53$ AU, beyond the orbit of Jupiter, only a major outburst would be readily visible, so that P/Schwassmann-Wachmann I is evidently an unusually large as well as exceptional comet, whose outbursts appear to be self-driven rather than being due to a simple evaporation process. We shall return to this comet in a later chapter.

Figure 11 compares the observed brightness along the ecliptic at various elongations from the Sun of the zodiacal light with calculations for a spherical particle of refractive index $m = 1.60 - 0.1\,i$ and radius 15 μm, the observations being those of D.E. Blackwell and M.F. Ingham (*Mon.Not.R.astr.Soc.*, *122*, 133, 1961). The calculations and the observations have been normalized together at elongation = 10^8. The particle radii cannot be significantly smaller than 15 μm without spoiling the good fit of Figure 11, but they could be larger. What emphatically cannot be done is to remove the complex absorptive part of the refractive index, which we interpret as arising from a 5 to 10 per cent of free carbon within the particles (F. Hoyle and N.C. Wickramasinghe, *Astrophys.Sp.Sci.*, *104*, 223, 1984).

The circumstance that as well as scattering sunlight the zodiacal particles are also highly absorptive means that they must emit strongly in the infrared.

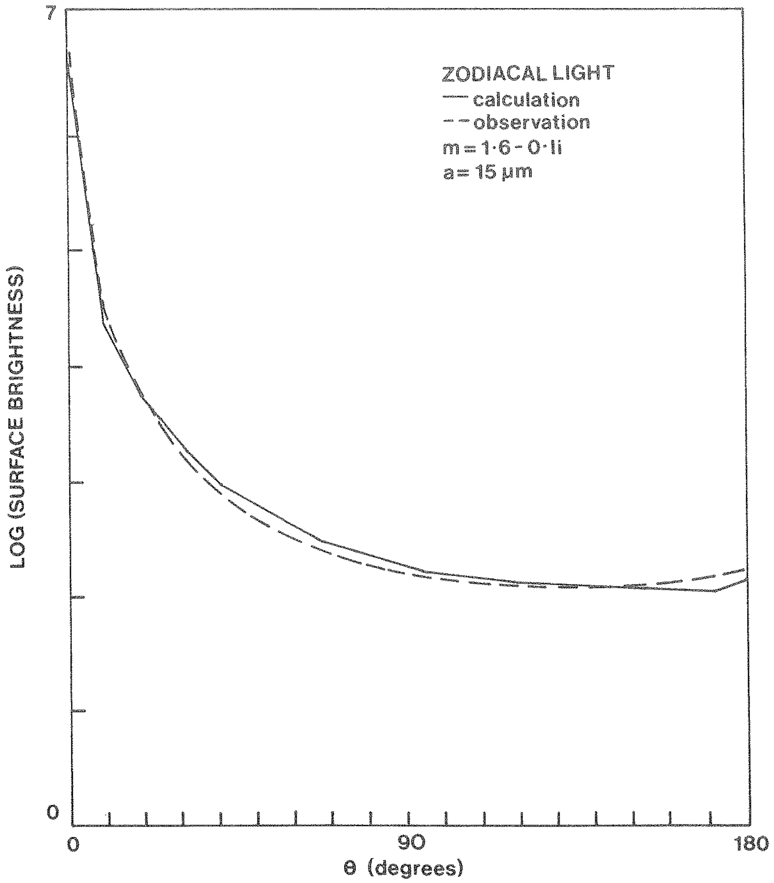


Fig. 11. Zodiacal light at 5400 Å for spheres (parameters in inset).

Calculations show that their infrared emissivity is almost independent of wavelength even to such long wavelengths as 100 μm. Figure 12 shows that this expectation is borne out by recent measurement with the IRAS satellite (M.G. Hauser, F.C. Gillett, F.J. Low, T.N. Gautier, C.A. Beichman, G. Neugebauer, H.H. Aumann, B. Baud, N. Boggess, J.P. Emerson, J.R. Houck, B.T. Soifer and R.G. Walker, *Ap.J.*, *278*, L15, 1984).

It is worth mentioning that interstellar molecular clouds contain particles with a remarkably large emissivity in the far infrared, particles with properties that need to be essentially the same as those inferred for the zodiacal cloud. The quantity of such particles is large, with minimum estimates ~ 10^6 solar masses for the whole of our galaxy, which is a fraction ~ 10^{-3} of all the

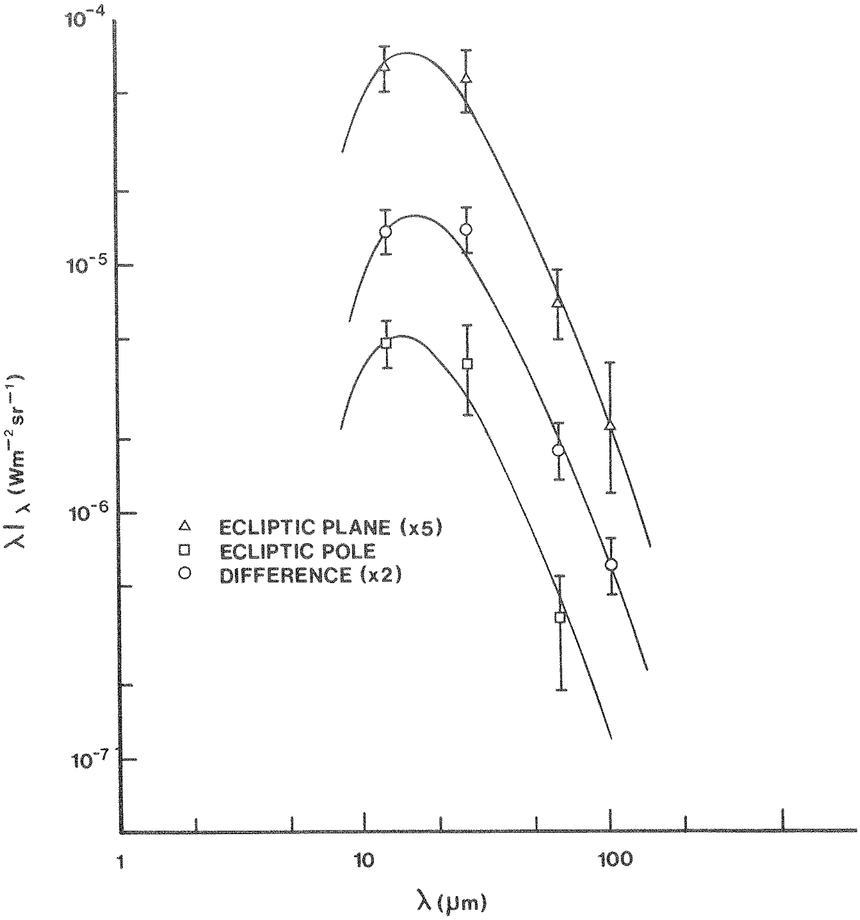


Fig. 12. Recent IRAS observations of infrared emission from the zodiacal cloud (points) compared with blackbody emission profiles (curves).

interstellar material. If these particles are cometary in origin, as the facts concerning their sizes and absorptivity suggest, then cometary activity must be far greater inside molecular clouds (probably occurring during star formation) than was estimated above. This adds considerable force to the above suggestion that the organic component of the interstellar clouds is cometary in origin, and provides for the effects of ultraviolet disruption which would otherwise be a problem at the rate of supply calculated above.

The calculated radius of 15 μm falls towards the upper end of the size range of the particles of extraterrestrial origin recovered by D.E. Brownlee from the high atmosphere of the Earth (in *Cosmic Dust*, ed. J.A. McDonnell, Wiley 1978, p.295). The latter are generally accepted to be from the zodiacal cloud, so that the zodiacal particles can actually be examined in the laboratory. They turn out to have a generally primordial chemical composition similar to the Sun, which would speak against the zodiacal cloud being of asteroidal origin as has sometimes been suggested, the point being that asteroids would be expected to have become chemically differentiated during the process whereby they condensed within the primordial solar nebula.

Brownlee remarks that the recovered particles are opaque black, slightly porous aggregates, with many of the larger ones having coatings with thickness $\sim 3 \times 10^{-6}$ cm of a low molecular weight carbonaceous material. Figure 13 shows a carbonaceous object recovered from one such particle, and we understand from private communication that this object is by no means unique. It is perhaps worth pointing out that there are interesting similarities between this object and a terrestrial object found in the Gunflint Chert of N.Minnesota, which has an age of about 2000 million years, and which is shown on the right of Figure 13. The object in the Gunflint Chert has been thought by some to be a bacterium.

There is a disposition to pass off evidence of biologically-suggestive morphological forms like that of Figure 13 as abiological products of the Fischer-Tropsch reaction, the basic form of which is $CO_2 + 4H_2 \rightarrow CH_4 + 2H_2O$. But the undoubted fact that even at high pressures for the reacting gases, and even with carefully prepared catalytic surfaces, the Fischer-Tropsch reaction has proved commercially unprofitable even with oil at $30 per barrel, should give one pause for thought. An analysis of the required catalytic action shows that delicate balances are required for intermediate reactions at the catalytic surfaces, which make it unlikely that the reaction could operate as a basis for the production of hydrocarbons in interplanetary space, where pressures would necessarily be much lower than have been used commercially.

Carbonaceous meteorites are very probably also cometary in origin. They have chemical compositons that are primordially solar like Brownlee's particles; their ages are somewhat greater than the age of the Earth; they are composed of small particles which have accumulated together in the manner of a sedimentary rock assembled under a compressive pressure of 10^7 to 10^8 dyne cm^{-2}, which are all properties that are hard to associate convincingly with an asteroidal or planetary mode of origin. The last two of the above-mentioned properties will be relevant when we come in a later chapter to consider the physical conditions that may well have obtained inside comets.

Carbonaceous chondrites have long been known to contain organic compounds, some so complex as to defy explicit analysis. Among identified

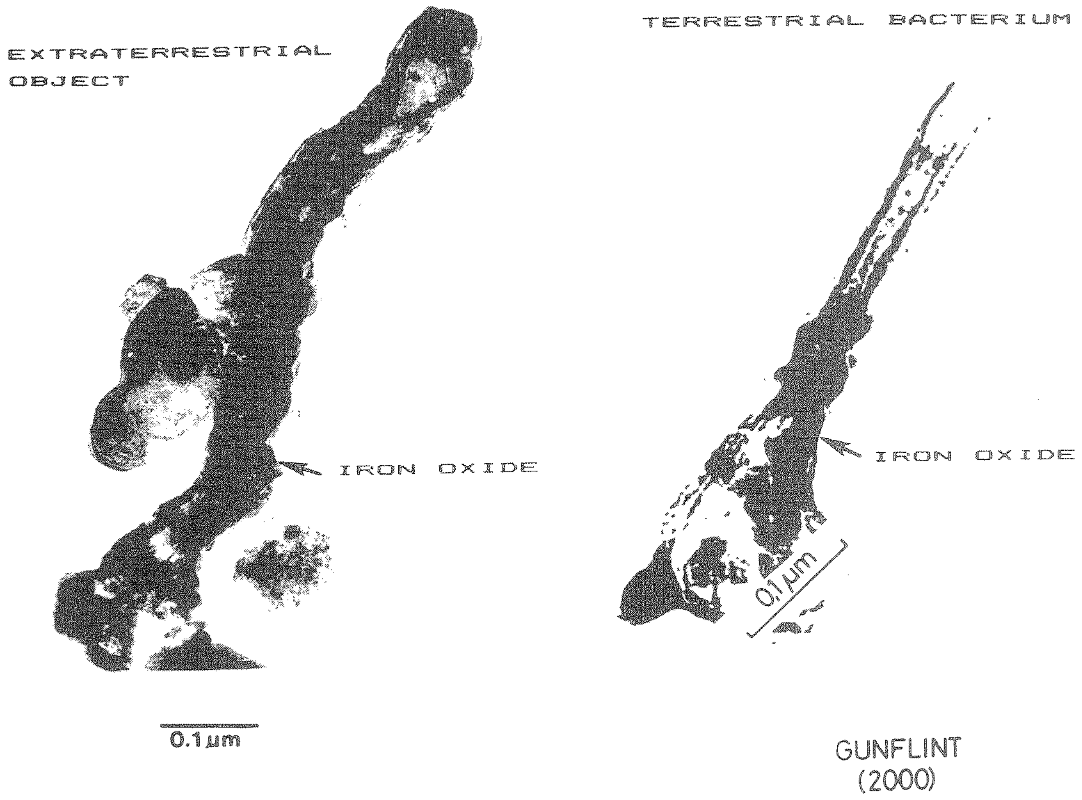
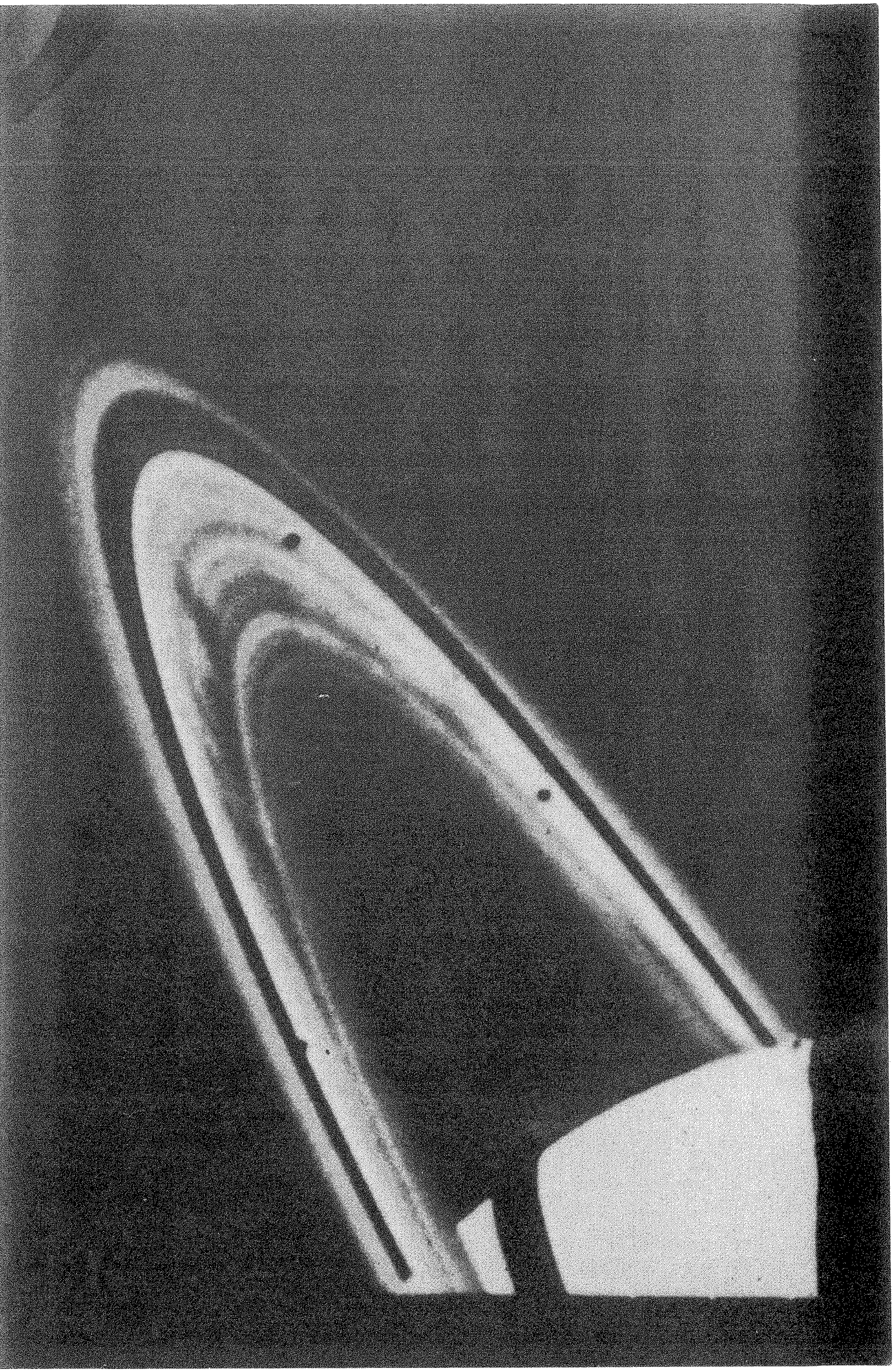


Fig. 13. Comparison of an object of organic composition in stratospheric dust with an object in the Gunflint chert generally believed to be a fossilised bacterium.

compounds are a number of amino acids of which the commonest are glutamine, aspartate, proline, leucine and alanine. For many years it was widely believed, either as a consequence of poor experiments and/or the Balldragon syndrome, that amino acids in carbonaceous meteorites are equal mixtures of the optically active L and D types, as is found to be the case in abiological systems, at any rate in abiological systems normally encountered in chemical laboratories. In biological systems, on the other hand, L types dominate heavily in proteins, but with L and D types both appearing in bacterial cell walls. Recent more sensitive experiments by M.E. Engel and B.Nagy (*Nature*, *296*, 837, 1982) have revealed the situation shown in Table 3, which is what would be expected for biological amino acids that have drifted modestly towards a racemic condition due to fossilization.

Table 3. *Murchison meteorite amino acid D/L values*

Extract	*Glu*	*Asp*	*Pro*	*Leu*	*Ala*
H_2O^1	0.322 (±0.044)	0.202 (±0.005)	0.342 (±0.065)	0.166 (±0.021)	0.682 (±0.062)
H_2O^2	0.30 (±0.02)	0.30 (±0.04)	0.30 (±0.02)	N.D.	0.60 (±0.03)
HCl^1	0.176 (±0.013)	0.126 (±0.004)	0.105 (±0.017)	0.029 (±0.002)	0.307 (±0.010)



CHAPTER 3

COMETS WITH PERIHELIA WITHIN THE ORBIT OF JUPITER

With few exceptions observed comets have perihelion distances $q < 5$AU, so penetrating within the orbit of Jupiter. We show in this chapter that all such comets must be transient objects with lifetimes that are less than 10^7 years, a result which in the next chapter we shall find points to the existence of a much larger reservoir of unseen comets having $q >> 5$ AU.

On each orbital revolution of a comet with $q < 5$ AU there is an energy interchange ΔE, as in equation (14) of Chapter 2, between the comet and Jupiter which on the average produces a change in the reciprocal $1/a$ of the semi-major axis of the cometary orbit that is of order 5×10^{-4} AU^{-1}, with the sign of ΔE depending on the details of the encounter of the comet with the Jupiter-Sun system.

The initial value one can attribute to the semi-major axis of a cometary orbit has an upper limit ~ 50,000 AU, because for appreciably larger values of a the cometary orbits would overlap the positions of the nearest stars and so be immediately unstable. A comet with $a = 50{,}000$ AU has aphelion distance $2a = 1.5 \times 10^{18}$ cm, about half the distance to the nearest star. The orbital period of such a comet is ~ 10^7 years, in which interval of time the comet must pass through the Jupiter-Sun system where it experiences an energy addition or subtraction according to the explicit dynamical details of the relation of the cometary motion to Jupiter's motion. If energy is added, the cometary orbit becomes hyperbolic and the comet is summarily ejected from the Solar System. If energy is subtracted, the semi-major axis is grossly changed from ~ 50,000 AU to ~ 2,000 AU in a single orbital passage, with the orbital period reduced drastically to 10^5 years. Dynamical evolution then proceeds with much greater rapidity.

Variations in the values of $|\Delta(1/a)|$ that arise from one orbital period to another are of less relevance to the present discussion than the distribution of additions (+) and subtractions (−) of energy, so that for simplicity in this discussion we assume $|\Delta(1/a)| = 5 \times 10^{-4}$ AU^{-1} in every orbital period. An appreciable systematic preponderance of + signs would lead to quick ejection from the Solar System, while an appreciable systematic preponderance of − signs would lead to a significant increase of $1/a$, say to 2×10^{-3} AU^{-1} in a time scale ~ 10^6 years. The orbital period would then be reduced an order of magnitude to ~ 10^4 years, and so the reduction in the semi-major axis would proceed even more rapidly. The reduction would therefore continue progressively, with the orbital period decreasing more and more rapidly, until the number of revolutions became large and the comet suffered denudation by evaporation effects, well within the time scale 10^7 years. This is on the assumption of a significant excess of energy subtractions.

The most favourable situation for retaining a comet at $a \simeq 2000$ AU would evidently be for there to be no systematic bias in the signs of $\Delta(1/a)$, which is to say for energy additions and subtractions to occur randomly. Even so, there would inevitably be changes of $1/a$ due to stochastic effects. Starting with $1/a = 5 \times 10^{-4}\ \mathrm{AU}^{-1}$, say, there are evidently equal probabilities at the next encounter of the comet with the Jupiter-Sun system that $1/a$ increases to $\sim 10^{-3}\ \mathrm{AU}^{-1}$, or that it decreases to zero, with the comet then being ejected from the Solar System. Thus in an ensemble of cases one-half will go to $1/a = 10^{-3}\ \mathrm{AU}^{-1}$ and one-half will be ejected. Then because additions and subtractions of energy are now taken to be equally probable, one-half of the cases going to $1/a = 10^{-3}\ \mathrm{AU}^{-1}$ will eventually proceed further to $2 \times 10^{-3}\ \mathrm{AU}^{-1}$, and one-half will return to zero, i.e. will be ejected from the Solar System. Continuing similarly, of the one-quarter of cases that go to $1/a = 2 \times 10^{-3}\ \mathrm{AU}^{-1}$, one-half will proceed still further to $1/a = 4 \times 10^{-3}\ \mathrm{AU}^{-1}$ and one-half will again return to zero and escape from the Solar System. Proceeding generally with this argument, the fraction of comets, all initially with $1/a = 5 \times 10^{-4}\ \mathrm{AU}^{-1}$ that have their semi-major axes shortened until $1/a = 5 \times 10^{-4}\,x\ \mathrm{AU}^{-1}$ ($x >> 1$) is simply $1/x$. Hence one per cent of cases have their semi-major axes shortened from the initial $a = 2000$ AU to 20 AU, the latter being comparable with the semi-major axis of Halley's comet. As always in stochastic processes, the number of orbital revolutions required to produce such a shortening goes as the square of the fluctuation, i.e. as $[(1/a)/|\Delta\ (1/a)|]^2$, which for $1/a = 5 \times 10^{-2}\ \mathrm{AU}^{-1}$, $|\Delta\ (1/a)| = 5 \times 10^{-4}\ \mathrm{AU}^{-1}$, is $\sim 10^4$ orbital revolutions. Since this number approaches the denudation limit for comets, $a \simeq 20$ AU approximates the largest practical fluctuation for an observed comet. This is due to stochastic effects involving randomly distributed steps $\pm 5 \times 10^{-4}\ \mathrm{AU}^{-1}$ at each orbital revolution. Whichever of the three possibilities discussed above arises in practice, the time-scale for a comet either to be ejected from the Solar System or for it to undergo so many orbital revolutions that it becomes greatly denuded is less than $\sim 10^7$ years (the time-scale in the stochastic case for 10^4 revolutions at $a = 20$ AU being only $\sim 10^6$ years). We are not aware of any theoretical discussion that would permit us to decide between the above possibilities. The approach of experienced investigators seems to be to examine computer simulation examples, dating from the pioneer work some twenty-five years ago of Lyttleton and Hammersley to a number of recent studies (c.f. E. Everhart in

Comets, ed. L.L. Wilkening, Univ. of Ariz. Press, p.659). In this rather indefinite situation it may be worth remarking that the actual data for cometary orbits suggests that the following is true: Energy additions and subtractions from cometary orbits with $q < 5$ AU occur in such a way that they yield a systematic negative effect proportional to $1/a$, viz.

$$\overline{\Delta E} = \text{constant} \, . \, \frac{1}{a} \, . \, \left| \Delta \left(\frac{1}{a} \right) \right| , \tag{15}$$

with the effect becoming stronger as the semi-major axis decreases. From (14) we then have per orbital revolution,

$$\overline{\Delta(1/a)} = \text{constant} \, . \, \frac{1}{a} \, |\Delta(1/a)| , \tag{16}$$

or with $|\Delta\,(1/a)|$ itself a constant in the present discussion the nett average increase of $1/a$ per revolution is proportional to $1/a$. Since the number of orbital revolutions in time dt is proportional to $dt/a^{3/2}$ this gives

$$\frac{\overline{d(1/a)}}{dt} = \frac{\text{constant}}{a^{5/2}} \tag{17}$$

For a steady mean flow of comets towards increasing $1/a$, the product of the right hand side of (17) and the cometary distribution function $\phi\,(1/a)$ say, must be constant, giving $\phi\,(1/a)$ proportional to $a^{5/2}$. In such a situation

$$\phi\left(\frac{1}{a}\right) d\left(\frac{1}{a}\right) = \text{constant} \, . \, a^{3/2} \, . \, \frac{da}{a} , \tag{18}$$

whence it follows that the steady-flow condition requires the number of comets per unit logarithmic interval in a to be proportional to $a^{3/2}$. Dividing this distribution function by the orbital period, we therefore see that the number of such comets observed near their perihelia should be a constant per unit logarithmic interval in the semi-major axis a, as it appears to be according to the data of Figure 14 (from L. Kresak, in *Comets*, ed. L.L. Wilkening, *loc.cit.* p.68).

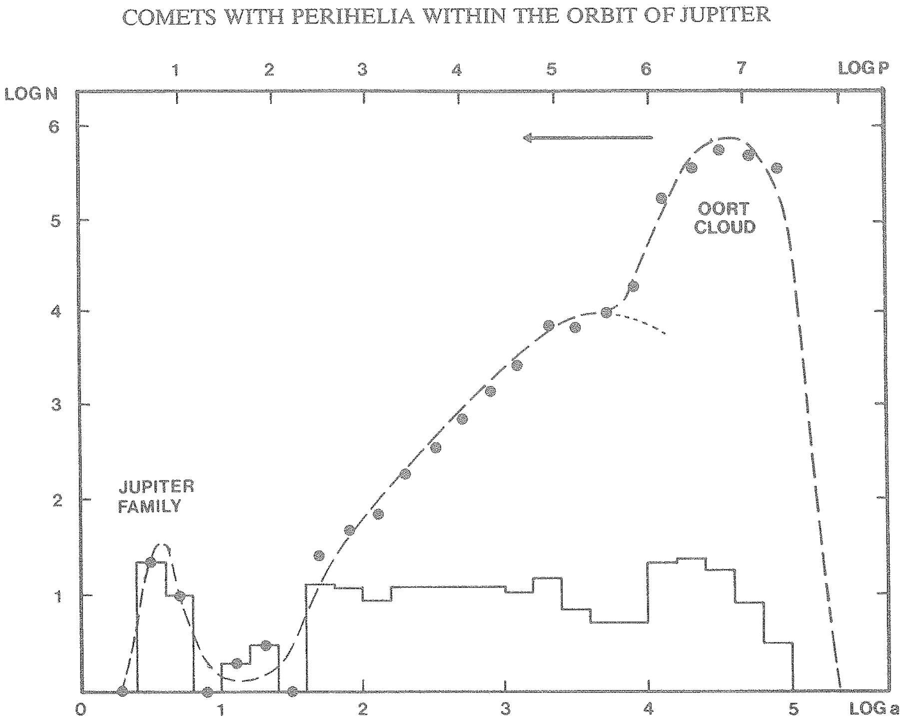


Fig. 14. The number of comets N as a function of their semi-major axis a (lower scale) or revolution period P (upper scale). The histogram shows the observed distribution of high-quality orbits corrected for the error dispersion. The dots and the curve show the values corrected for the probability of observation as a function of the revolution period.

We could attempt to derive a similar result for (1/a) within a purely stochastic model, by using a diffusion-type equation,

$$\frac{\lambda^2}{a^{3/2}} \frac{d\phi(1/a)}{d(1/a)} = \text{constant} \ , \tag{19}$$

using the same steady-flow condition as in (18). Here λ is a 'mean-free-path', which is to say λ is the magnitude of the ± variation in 1/a per orbital period. The orbital period being proportional to $a^{3/2}$, the ratio $\lambda / a^{3/2}$ in (19) can be seen to be the analogue of the velocity term in the usual form of the diffusion equation. Putting $\phi\,(1/a) \propto a^{5/2}$ in (19) leads to $\lambda \propto a^{-1}$, which may be compared with (14), viz

$$\Delta(1/a) = -\frac{2\Delta E}{GM} \qquad (14)$$

Since in a purely stochastic situation the ± variations of ΔE are constant in magnitude and therefore do not contain the required factor a^{-1} this approach fails, leaving only the systematic effect discussed above to explain the observations. The observed result that the histogram for log N displayed in terms of log a is of such nearly constant height seems so striking that the lack of a strong explanation for it is unsatisfactory. The problem evidently lies in the statistics of the three-body encounters of comets with the Jupiter-Sun system, which appear to contain remarkable properties that we do not recall seeing explained.

The gap in the histogram of Figure 14 from $\log a \simeq 0.8$ to $\log a \simeq 1.6$ also requires explanation. It is natural to argue that the histogram falls away at the smaller values of a because the number of orbital revolutions of the comets exceeds 10^4, causing such severe denudation that the residues are not visible to observers in their full number. But what then of the recovery of the histogram for short-period comets with $a < 10$ AU ($\log a < 1$ in Figure 14)? This 'Jupiter family' needs a further point added to the discussion if it is to be understood. A strong hint of what this further point may be comes from the circumstance that the inclinations i of the orbits of comets in the 'Jupiter family' are concentrated towards low values, with a strongly-marked peak at $i \simeq 8°$. Such low values ensure that cometary encounters with Jupiter can occur in a fraction of orbital revolutions at distances much smaller than the radius of Jupiter's orbit, which must in general yield shifts of $1/a$ that are much greater than the value 5×10^{-4} AU^{-1} assumed throughout the above discussion. Thus the orbit of a comet with $1/a$ initially small can evolve to $1/a$ large in a comparatively few close encounters, provided energy subtractions from the comet happen to dominate significantly over energy additions. The latter requirement is more likely to happen for a nearly random distribution of subtractions and additions when the number of encounters is small, as it will be for the close encounters considered above. In short, when numbers are small, fluctuations can be large.

The implication of this discussion is that, whereas the main part of the histogram of Figure 14 appears to be due to a slow persistent removal of energy from cometary orbits over many encounters with Jupiter at distances ~ 5 AU, the peak at $\log a < 1$ seems to arise from the stochastic effects which must necessarily occur for much closer encounters occurring in small numbers.

It will be recalled that the histogram of Figure 14 refers to observed cometary orbits, from which the cometary distribution function ϕ has been calculated to give the broken curve of Figure 14. From (16) our expectation was $\log \phi = (3/2) \log a + \text{constant}$, which is well borne out by Figure 14. This is for $1/a \geq \sim 5$

$\times 10^{-4}$ (AU)$^{-1}$, where the broken curve looks as if it is going to turn over and decline (corresponding to the fall of the histogram from log a = 3.3 to log a = 4). But what is then shown for log a > 4 is a curve rising to a high peak marked 'Oort Cloud' in Figure 14. This is a distribution of comets at a ≃ 50,000 AU that for the most part have not yet had a first encounter with the Jupiter-Sun system. The cloud is shown as comparatively numerous because of the much longer orbital period ~ 10^7 years at a ≃ 50,000 AU. Thus the cloud must be numerous enough to supply comets with log a < 3 at an adequate rate.

We come now to the striking inference that comets in the 'Oort Cloud' with q < 5 AU, which are the progenitors of the observed comets, cannot have existed in their present orbits for more than ~ 10^7 years. From a cosmogonic point of view the system is therefore very recent, and the next question one asks is evidently where did the 'Oort Cloud' come from? We begin to address this question in the next chapter.

CHAPTER 4

ENCOUNTERS WITH PASSING STARS AND GALACTIC CLOUDS

Stars which happen to penetrate the 'Oort Cloud' of comets will in essentially all cases pass through the cometary cloud at a speed that is large compared with typical cometary speeds, $\sim$ 10 km s^{-1} for a star and $(GM/r)^{1/2} \simeq 1/3$ km s^{-1} for the orbital speed of a comet at a heliocentric distance $r = 10{,}000$ AU. Thus an intruding star can be considered to 'race by' a comet before the latter has had time to change its position significantly relative to the Sun. For a closest approach at distance s of a star of mass M^* moving at relative speed V, the magnitude of the velocity perturbation induced by the star on the comet is therefore $\sim(GM^*/s^2).\ (s/V)$. If this perturbation is to be comparable with the orbital speed $(GM/r)^{1/2}$ of the comet, the distance s must be such that

$$\frac{s}{r} \approx \frac{M^*}{M}\left(\frac{GM}{r}\right)^{1/2} \cdot \frac{1}{V}. \tag{20}$$

With $M^*/M \simeq 1$, (20) gives $s/r \approx 1/30$, which implies a very close necessarily rare encounter. Excluding such infrequent cases, we might therefore be tempted to argue that passing stars have little effect on comets of the Oort Cloud, and so far as semi-major axes, and directions of the apse lines, are concerned this is true.

However, because cometary orbits are so highly elliptic, a little more care than the above is needed. The highly elliptic condition means that when distant from the Sun, as comets in the Oort Cloud are, the motions are dominantly radial. Thus it is the radial motion that is $\sim (GM/r)^{1/2}$, and to which the above argument applies. The transverse component of the velocity, on the other hand, is $(2GMq)^{1/2}/r$, where q is again the perihelion distance of the comet. A similar argument to the above applied specifically to the transverse component therefore gives

$$\frac{s}{r} \approx \frac{M^*}{M} \cdot \left(\frac{GM}{2q}\right)^{1/2} \cdot \frac{1}{V}. \tag{21}$$

Hence for the transverse component it is one-half of the speed of the comet at perihelion which has to be compared with the relative speed V of the passing star. Here q is of course the perihelion distance before the entry of the star.

In our quest for an understanding of the origin of comets we shall find that values of q in the range 20 to 30 AU are initially typical for the cometary cloud, giving $\sim (GM/q)^{1/2} \simeq 5$ km^{-1}, so that the right-hand side of (21) for V = 10 km s^{-1} is $\sim 1/3$, from which it follows that a star on entering the Oort Cloud can 'stir up' a fraction of the order of the square of the right-hand side of (21), i.e. about 10 per cent of all the comets in the cloud. This is in the sense that the small cometary transverse components of velocity become markedly changed, by amounts of the order of their original magnitudes. The important effect for these perturbed comets is to produce gross changes of their associated q values. Even if all initial q values were in the range 20 to 30 AU, some will now emerge with $q < 5$ AU, and so will form a group of comets of the kind that were considered in the previous chapter. Every penetration of a star into the Oort Cloud produces a new crop of comets of small perihelion distances, so the circumstance deduced in the previous chapter, that such a crop cannot persist for longer than 10^7 years, is no cosmogonic embarrassment provided:

(i) Stars enter the Oort Cloud at intervals less than 10^7 years.
(ii) By the term 'Oort Cloud' we imply a large reservoir of comets normally with q values in the range 20 to 30 AU, or perhaps for some of the comets even more distant perihelion distances than 30 AU.

As regards condition (i), write R for the radius of the Oort Cloud and n^* for the star density in the solar neighbourhood. Then $\pi R^2 n^* V$ is the rate at which stars enter the cloud. Putting R = 0.5 pc, $n^* = 0.1$ pc^{-3}, V = 10 km s^{-1} = 10^{-5} pc per year, and the rate is seen to be $\sim 10^{-6}$ per year, i.e. about one stellar encounter in 10^6 years, well within the range of condition (i). Indeed at any given epoch, for example the present 20th century, about ten distinct crops of comets of small perihelion distances arising from about ten distinct stellar encounters would be expected to be contributing to comets with $q < 5$ AU, which is to say contributing to the observed comets. Notice that in order to satisfy condition (i) the radius R of the Oort Cloud must necessarily be $> \sim 30{,}000$ AU. Otherwise the frequency of stellar encounters would become less than 10^{-7} per year, and there would be no steady supply of observable comets. We might still seek to argue that the present epoch is exceptional in the sense that we happen to live within $\sim 10^7$ years of a stellar encounter, even though the average encounter rate were less than 10^{-7} per year. There is, however, an observational means for excluding this non-uniform possibility, because in such a case the apse lines of all observed comets would point to the track of a single star, whereas with R $\simeq$ 0.5 pc and with ten encounters yielding ten distinct crops of observable comets, the distribution of directions of apse lines should be far more complex – essentially they should be 'all over the sky', which is the actual observed situation, as can be seen from Figure 15 (from L. Kresak, *loc. cit.*, p. 71).

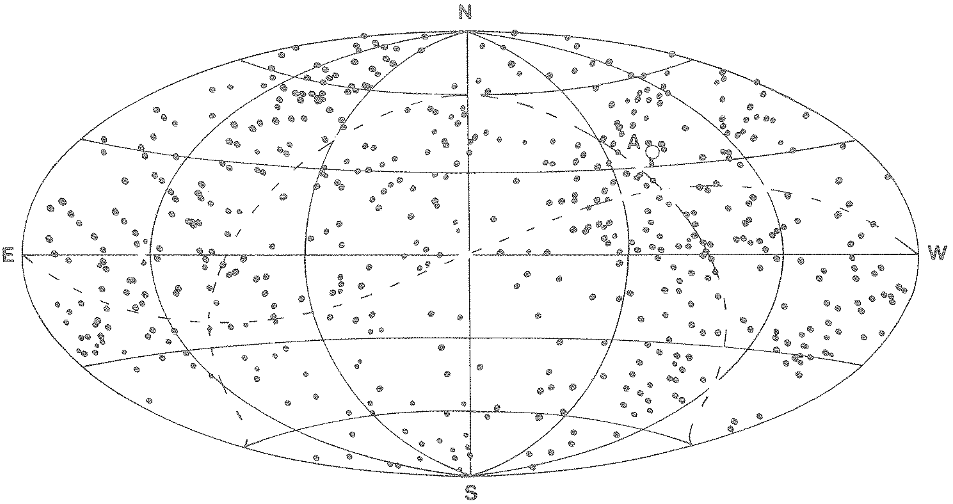


Fig. 15. Directions of apse lines in the sense from aphelion to perihelion of long-period comets plotted in an ecliptical equal area projection. North up, point or vernal equinox in the middle, ecliptical longitude increasing to the left. The two dashed curves show the positions of the equator (sinusoid-like) and of the galactic plane (horseshoe-like). Open circle, eight perihelia of the Kreutz groups; A, apex of the motion of the solar system with respect to its stellar environment (basic apex).

If we now could show:

(a) that during the formation of the Solar System an approximately spherically symmetric halo of comets became established around the Sun with a radial extension of ~ 0.5 pc, and

(b) the cloud of comets so generated has been stable over the lifetime of the Solar System,

we should be cosmogonically home and dry, although we should still need to investigate the detailed physical and chemical properties of individual comets. But unfortunately for one's peace of mind, both (a) and (b) cause trouble. W.M. Napier and M. Staniucha (*Mon. Not. R. astro. Soc.*, *198*, 723, 1982) have pointed out that a cloud of comets with R $\simeq$ 0.5 pc is likely to be unstable because of the passage from time to time of the Solar System close to a massive molecular cloud. Although these authors use computer simulation, the same result can be shown from simple considerations. During the lifetime of the Solar System it would be reasonable to suppose there has been several approaches to within ~ 3 pc of a molecular cloud of mass $M \simeq 10^4\,\mathrm{M}$, M being again the solar mass. The time required for such a passage at relative speed of 10 km s^{-1} would be $\sim 10^{13}$ sec., which is considerably less than the orbital periods of comets in the outer regions of the Oort Cloud. As before, therefore, the encounter is over and done with before such comets have time to move significantly in their orbits. The previous argument leading to (20) requires modification in this case, however, because it is not the whole acceleration of M that tends to strip away the comets from the Solar System, but only the differential acceleration between the comets and the Sun, which is $\sim GM\,\mathrm{r}/\mathrm{D}^3$, where D is the distance of approach of a comet to the molecular cloud and r is again the comet's heliocentric distance. The velocity perturbation generated in the comet's motion is thus $\sim (GM\,\mathrm{r}/\mathrm{D}^3).(\mathrm{D}/\mathrm{V})$, so that the condition for the comet to be held by the Sun against the tidally disruptive effect of the molecular cloud is

$$\frac{\mathrm{G}M}{\mathrm{D}^3}\,\mathrm{r}\cdot\frac{\mathrm{D}}{\mathrm{V}} < \sim \left(\frac{\mathrm{GM}}{\mathrm{r}}\right)^{1/2}, \tag{22}$$

i.e.

$$\frac{\mathrm{r}}{\mathrm{D}} < \sim \left[\mathrm{V}/\left(\frac{\mathrm{G}M}{\mathrm{D}}\right)^{1/2}\right]^{2/3} \cdot \left(\frac{\mathrm{M}}{M}\right)^{1/3}. \tag{23}$$

With D = 3 pc, $M = 10^4\mathrm{M}$, $\mathrm{V} = 10\ \mathrm{km\ s^{-1}}$, (22) gives $\mathrm{r} < 0.1\,\mathrm{D} \simeq 10^{18}$ cm. For such an encounter the outer regions of the Oort Cloud at $\mathrm{R} = 10^{18}$ cm just barely survive.

One can interpret this result as good or bad according to how one sees it. The result is good in the sense that it leads to an outer boundary for the Oort Cloud that fits expectation very well. It is also good in an important respect that will be raised again shortly. It can be judged bad, however, if one argues that the values chosen above for D and M are not a worst-case situation, but are rather typical choices of D, M for the score or so of encounters between the Solar System and molecular clouds that have occurred over the 4.6×10^9 years since the origin of the Solar System. The position is evidently more sensitive to the choice of D than of M, with extensive stripping occurring if an encounter should take place at, say, D = 1.5 pc. In this connection it is to be noted that the gravitational deflection of the whole Solar System by the molecular cloud reduces D by the factor $(1 + 2GM/DV^2)^{1/2}$ below what it would otherwise be. While this is not a significant effect for the values of M,D,V chosen above, any unusual combination of these parameters could make this factor relevant.

Napier and Staniucha conclude that stripping in the worst case must have been so severe that an Oort Cloud of comets with $R \simeq 10^{18}$ cm could not have persisted from the origin of the Solar System to the present day, so opening up the question of the origin of the comets to the suggestion that the Solar System may have captured them from interstellar space, possibly from the molecular clouds, a proposal that has subsequently been investigated by Napier and S.V.M. Clube. The difficulty with all such capture theories turns on the smallness of the solar accretion radius defined by $2GM/V^2$, which has only the value $\sim 3 \times 10^{14}$ cm for $V = 10$ km s^{-1}. Whatever explicit capture process one considers, the accretion radius always turns out to dominate the situation, at any rate in our experience, so that one is left with the requirements of an implausibly low encounter velocity V if capture at distances exceeding 10^{15} cm is required. Capture at 10^{18} cm appears to be out of the question.

More than thirty years ago R.A. Lyttleton suggested an interesting form of capture theory in which the Solar System acquires interstellar grains rather than whole comets. The grains become concentrated along an accretion axis which points from the Sun in a direction that is opposite to the motion of the Sun relative to the grains of the solar neighbourhood. Having become thus concentrated about a line, the grains were then taken to gather together into clumps which were supposed to constitute the comets as we see them. The theory had its outstanding success for the so-called Kreutz group of comets which have $q \simeq 0.008$, $P \simeq 758$ years, $a \simeq 83$ AU $\simeq 1.2 \times 10^{15}$ cm. An encounter speed $V \simeq 3$ km s^{-1} would give an accretion radius comparable to the aphelion distances of these particular comets, which encounter speed is not outside the range of possibility, although values of V in the range 10 to 20 km s^{-1} would be more typical.

The problem for Lyttleton's theory is that it leads to comets with short lifetimes (according to the arguments of Chapter 3) and so requires comets to be forming by accretion more or less all the time, and with values of V that are low, ~ 3 km s^{-1} or less, a situation which is excluded as an on-going situation.

The argument of Napier and Clube appears to us to be open to the criticism that it tacitly assumes the distribution of aphelion points in the Oort Cloud to lie in a spherical shell rather than in a whole sphere. Because comets spend most time near their aphelia, it would be possible to have a distribution of comets with aphelia in a shell from $r = 5 \times 10^{17}$ cm to $r = R = 1.5 \times 10^{18}$ cm, say, being stripped almost clean in an exceptional encounter of the Solar System with a molecular cloud. But if the aphelia initially are distributed throughout a sphere of radius R, those with heliocentric distances small enough to satisfy a condition of the form of (22) will survive the encounter, remaining as comets bound to the Sun. Of those which survive the encounter two situations can be distinguished:

(1) Aphelion points are appreciably shifted in position at the end of the encounter, with the associated comets still remaining a part of the Solar System.
(2) Aphelion points in a less distant inner region are little affected.

What happens it seems is that comets from situation (1) simply fill-out the more distant part of the Oort Cloud vacated by comets that are stripped. We cannot envisage a state of affairs in which some comets attain hyperbolic orbits with respect to the Sun without other comets attaining orbits which, while remaining elliptic, come very close to the hyperbolic condition. In other words, the energy distribution of the cometary orbits will be continuous after an encounter, so that if the distribution attains positive values (i.e. if some comets are stripped) others will inevitably have only slightly negative energy values and so will move after the encounter in a cloud of adequately large radius, $R \simeq 10^{18}$ cm for the cloud.

As well as maintaining the radius R at a large enough value for stellar encounters to generate frequent crops of comets with $q < 5$ AU, i.e. visible comets, large shifts of aphelion positions according to situation (1) deal with a critical cosmogonic problem that otherwise would probably be insuperable. All processes which seek to explain the origin of the comets as a part of the overall origin of the Solar System itself lead to orbits with apse lines that are approximately parallel to the plane of the ecliptic. Thus the initial distribution of aphelion points forms a disk, not a sphere. Encounters with molecular clouds, besides stripping comets from the edge of the disk, generate a distribution of aphelion points at large distance from the Sun that spreads from the initial disk into a more or less uniformly-filled spherical shell. Even to this day the inner part of the cometary cloud is probably more disk-like than spherical. Comets in the disk should on the cosmogonic basis discussed in the next chapter, have perihelion distances q in the range 20 to 30 AU and aphelion distances up to 30,000 AU, with the outermost more or less spherical shell then extending to 100,000 AU.

If we go back to Oort's original discussion of the origin of the cometary cloud (*Bull.Astron.Inst.Netherlands XI*, 408, 91-110.) it is easy to understand the cosmogonic issues involved here. Oort's early proposal was that comets originated in the explosion of a planet that once moved around the Sun in the outer part of the Solar System. Suppose the planet to have been in a circular orbit in the plane of the ecliptic. Write **V** for its velocity vector at the moment of explosion, with $|\mathbf{V}| = GM/r$, r being the radius of the orbit. Let the explosion cause cometary debris to be hurled out in all directions with a fixed speed v relative to the planet. Taking account of the direction in which a particular element of debris is ejected we can write $\mathbf{v}$ for its velocity vector relative to the centre of the planet at the moment of explosion. The kinetic energy per unit mass of the element is then $\frac{1}{2}(\mathbf{V} + \mathbf{v})^2$. According to the direction of v, the kinetic energy relative to the Sun evidently varies between an upper limit of $\frac{1}{2}(V + v)^2$ and a lower limit of $\frac{1}{2}(V - v)^2$. Ejecta with kinetic energy values exceeding $\frac{1}{2}V^2$ will be projected into orbits with aphelion distances exceeding the radius r of the orbit of the planet, indeed with aphelion distances increasing the more the kinetic energy moves towards its upper limit of $\frac{1}{2}(V + v)^2$. If the explosion were sufficiently violent for $v > (\sqrt{2} - 1)V$ ejecta with $\mathbf{v}$ substantially parallel to $\mathbf{V}$ would be expelled altogether from the Solar System, in which case there would be a cone with $\mathbf{V}$ as axis outside of which ejecta would be retained. In particular, ejecta having directions outside but close to the surface of this cone would be expelled to great distances from the Sun. Since the orbit of an individual element lies in a plane defined by $\mathbf{v}$ and by the direction from the planet to the Sun, it is clear that all the ejecta which come to be expelled far away from the Sun must necessarily move in a disk-like distribution, determined by the set of planes through the Sun and through the generators of the cone just described. Generally, the greater the distances to which ejecta are expelled from the Sun the flatter must be the distribution of their apse lines.

If now we return to Figure 15, it is apparent that the distribution of the apse lines of observed comets is anything but disk-like, and since observed comets are samples derived from the outer regions of the Oort Cloud, it follows that the outer regions of the whole Oort Cloud are at present anything but disk-like. Without the redistribution effect on aphelion positions produced by encounters with molecular clouds it would be hard, if not indeed impossible, to connect the initial cosmogonic situation described above with the present situation.

Attempts to discuss the origin of comets in a more sophisticated manner do not seem to lead to anything much different. As the outer planets of the Solar System formed, there must have been a stage where a number of bodies existed with velocities of escape from their surfaces exceeding 10 km s^{-1}, bodies that were probably in approximately circular orbits around the Sun. There must also have been a great amount of smaller debris, at least some of it moving in more disordered orbits. Consider a piece of debris approaching close to one of the larger bodies with initial relative velocity $\mathbf{v}_0$ and final velocity after the encounter $\mathbf{v}_1$, $|\mathbf{v}_1| = |\mathbf{v}_0| = v$ say. Writing $\mathbf{V}$ for the velocity vector of the larger body relative to the Sun, the initial kinetic energy per unit mass relative to the

Sun of the debris would be $\frac{1}{2}(\mathbf{V} + \mathbf{v}_0)^2$ and the final kinetic energy $\frac{1}{2}(\mathbf{V} + \mathbf{v}_1)^2$. As an example, suppose $\mathbf{v}_0$ to be at right angles to $\mathbf{V}$, and suppose $\mathbf{v}_1$ to be parallel to $\mathbf{V}$, corresponding to a situation in which the relative motion between the debris and the body happens to be turned through a right angle. Then the initial kinetic energy per unit mass of the debris is $\frac{1}{2}(V^2 + v^2)$ and the final kinetic energy is $\frac{1}{2}(V + v)^2$. As the random speed v relative to the larger body increases towards the orbital speed V of the latter, the jump in the kinetic energy of the debris thus occasioned simply by turning the direction of relative motion becomes large.

For $v = V/3$, $\frac{1}{2}(V^2 + v^2) = 5V^2/9$, while $\frac{1}{2}(V + v)^2 = 8V^2/9$. In effect, as the debris develops appreciable random motions, which at first are probably radially directed and so without great effect on the aphelion distances, sudden switches into transverse motions due to encounters with larger bodies produce almost impulsive jumps in the kinetic energy that are dynamically equivalent to the explosion of Oort's hypothetical planet, which is to say aphelion distances for pieces of debris, comets, can be greatly increased.

Likewise, perihelion distances of debris can be greatly decreased. Thus with $\mathbf{v}_0$ again directed perpendicular to $\mathbf{V}$ and $\mathbf{v}_1$ turned anti-parallel to $\mathbf{V}$, the kinetic energy per unit mass of this debris is reduced from $\frac{1}{2}(V^2 + v^2)$ to $\frac{1}{2}(V - v)^2$. For the case $v = \frac{1}{2}V$ and with the larger body in a circular orbit, the kinetic energy per unit mass of the debris would become $\frac{1}{8}V^2$, which is small enough for the perihelion distance of the debris to become less than the heliocentric distance of the larger body by the factor 7. In this way, debris involved in the condensation of the outer part of the Solar System could be brought to the inner regions, where it could maintain a severe bombardment of the Moon and inner planets. The Earth would very likely capture much of the material that was incident on it in this process. With C,N,O,H the main elements present in the outer planets, it appears probable that it was in this way that the Earth acquired its volatile materials, particularly the water of the terrestrial oceans.

CHAPTER 5

REFLECTIONS ON THE COSMOGONY OF THE SOLAR SYSTEM

There can be few other problems which have been extant so long, and yet which have led to no generally-agreed solution, as has the problem of the origin of the Solar System. Newton at first thought that by calculating the planetary motions backwards it might be possible to arrive at a situation which indicated what their genesis must have been. He came to realize, however, that comparatively small changes in the starting conditions for the planets could lead to big changes as the calculations proceeded backwards in time, and hence no firm reliance could be placed on the method. In effect, Newton had perceived the essence of what today would be called an irreversible process.

In the 18th century there were speculations of a dynamic, so-called catastrophic kind in which an interaction between the Sun and bodies from outside the Solar System were assumed to play a critical rôle. Then Laplace introduced the very different concept that planetary formation was an integral part of star formation. On Laplace's 'theory' (as it became called, although as Laplace had it the position was more a speculation than a theory) planetary systems would be essentially as common as star systems, whereas catastrophic theories made planet formation an exceptionally rare phenomenon.

Thereafter until the first third of the 20th century Laplace's view largely held the field. But in the 1920s and 1930s it was realized, notably by British mathematicians, that the Laplace theory was seriously flawed in respect of what was called the 'angular momentum problem'. This led in Britain at least to a vigorous revival of catastrophic theories. Following the 1939-45 war, however, there was a fad in astronomy for 'turbulence' as a solution to every problem within sight, and the claim was widely made and believed – with belief persisting to some extent to the present day – that turbulence could somehow overcome the difficulty which had been noticed in the Laplace theory. But the turbulence fad was merely an attempt by those who had studied aeronautics closely during the 1939-45 war to take over whatever other fields of science they could manage to infiltrate, with all the plaudits and research grants that accompany successful infiltration. This development was seriously counterproductive, with its ill-effects persisting to this day.

Another quite different idea arose in the late 1940s, which really was capable of overcoming the 'angular momentum problem' in the Laplacian form of theory. One of the present authors lived through these post-war developments, and a personal account follows of how the position changed quickly from ca. 1950 onwards, changed from obscurity into a considerable measure of clarity. The situation today remains essentially the same as it is described in this account by F.H. Before embarking on these reflections it should be added that in recent years a still different mode of attack on the problem has been adopted by some investigators. Whereas the theories up to the 1950s worked from known facts concerning the actual planets themselves, and so were in a considerable measure inductive, there have been almost wholly deductive attempts in recent years. The latter investigations work from what is believed to be true for star formation in general. But star formation is not one of the

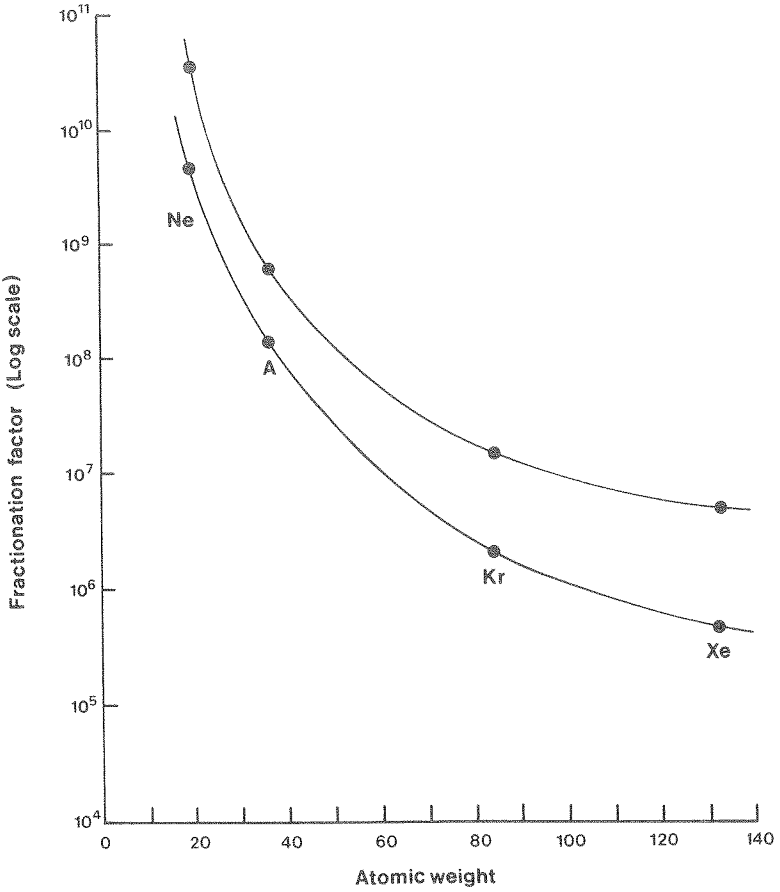


Fig. 16. Fractionation factors for noble gases.

better understood parts of astronomy, since investigations rest on uncertain premises, instead of on known facts as they were for earlier theories. In the authors' view this new procedure has not succeeded in throwing any worthwhile new light on the matter, which therefore rests today pretty much as it was in the 1950s, of which situation F.H. writes:

'I would like to begin, not at the beginning, but with ideas that came late for me, ideas due to Harrison Brown which appeared under the title "Rare Gases and the Formation of the Earth's Atmosphere", in the 50th year anniversary symposium held by the Yerkes Observatory (appearing in *The Atmosphere of the Earth and Plants*, ed. G.P. Kuiper, University of Chicago Press, 1949). Figure 16, which appeared as Figure 62 in Harrison Brown's article, contains crucial information relating to the problem of the origin of the Solar System of a kind I had not realized at all up to then, information that sets a constraint which any theory of the origin of the planets must satisfy as a high priority.

'The meaning of the ordinate in Figure 16 is that the rare gases He, Ne, Kr and Xe have been depleted on the Earth relative to Si by the "fractionation factor" plotted as ordinate, the two sets of points being calculated as maximum and minimum values for the amounts of rare gases. Minimum values were obtained from the concentrations of the rare gases in the terrestrial atmosphere, and maximum values obtained by assuming the rare gas content of the whole of the Earth's mantle to be given by the average amounts measured in the surface rocks. The case for the latter giving maximum values rests on the physical argument that it is harder for gases to remain occluded in rocks at very high pressures than in rocks at low pressure. Even if a critic were to contend that this argument may not be quite secure, it is hard to believe that the lower rocks of the mantle could contain amounts of gas that exceeded the amounts in the surface rocks by as much as a factor of $\sim 10^6$, the order of the fractionation factor for Xe.

'The atomic weights of Kr and Xe are so high that the fractionation factors for these heavier rare gases cannot be explained by simple thermal evaporation. Thus the protoplanet which eventually became the Earth did not have a solar mix of the elements, the protoplanet was not a glob of material of solar composition, as theories of the modern solar nebula variety would like it to have been. The Kr and Xe, and likely enough all elements existing in the gaseous phase, had already been separated-out when the protoplanet formed. As Harrison Brown said it (with original italics):

"In view of the large fractionation factors, it would appear that *during the process of Earth formation the mechanism was such as to prohibit the retention of an appreciable fraction of any substance that existed at that time primarily in the gaseous state*."

'To continue anecdotally for a while, it was almost a decade later that Harrison Brown organized a weekly seminar on cosmogonic problems. This was at the California Institute of Technology. When my own turn came in the series I gave the ideas which now follow and which had developed in my mind during the early 1950s. The date of those seminars could be fixed precisely because they coincided with an extended visit to Caltech by F. Houtermans, the father of the uranium-lead dating method. I remember Houtermans, partly because of the two seminars he gave, and partly because the presence of Houtermans, or even the mention of his name, always moved Walter Baade to laughter. Walter would tell how they had both been young-men-in-research at Göttingen, and how Houtermans had been a great fellow for burning the midnight oil. His ideas and suggestions became evermore imaginative, wild or preposterous (depending on how you saw them) as the late hours advanced into still later hours, to a point where the father of uranium-lead dating became widely known to his contemporaries as Eine-Kleine-Nachtmusik Houtermans. By 1955-56, however, Houtermans had become a thoroughly august figure, and it was in this spirit that we listened to him with due reverence and awe.

'As a young research student in the late 1930s I learned verbally of the problem of the origin of the Solar System from discussions with Ray Lyttleton and Harold Jeffreys, and I had some correspondence about it with James Jeans. So it was natural enough for me to begin as an adherent of the British catastrophic school with its emphasis on planetary dynamics, a school founded on the work of Jeans, Jeffreys and Lyttleton. Chemistry was remoter than the Moon, and considerations such as the one quoted above never entered our heads.

'Now just as Harrison Brown's discussion of the depletions of rare gases in the terrestrial atmosphere is an argument from chemistry with far-reaching implications, so there is a simple fact of planetary dynamics with far-reaching implications. Although the combined mass of the planets is only a little over 0.1 per cent of the mass of the Sun, the planets have combined angular momentum about the centre of the Solar System that is ~ 100 times greater than the angular momentum of the Sun. Gram for gram, planetary material has on the average $\sim 10^5$ times more angular momentum than solar material. This curious fact, having no explanation within Laplace's so-called theory, had led the Cambridge school of cosmogonists to investigate theories of a catastrophic kind: The close approach of another star to the Sun, or the close approach of another star to a one-time binary companion of the Sun. It was in this intellectual atmosphere that I began thinking about planets and their origin, and an early small contribution I made was to remark that the Sun would have been more likely to have lost an original binary companion through catastrophic mass loss from the companion than through a close encounter with a third star.

'In the 1940s I was doubtful of rival theories which sought to explain the origin of planets in terms of the turbulent development of a rotating solar nebula. Such a theory due in the mid-1940s to von Weizsaecker was widely and sympathetically received, but for me it was an unconvincing attempt to add flesh to the bones of the old Laplace theory. The reason for the sympathetic reception was that the von Weizsaecker theory fitted very well with the astronomical fashion of the day. The first fashion I encountered in astronomy was "radiation pressure". Back in the 1930s radiation pressure was supposed to be responsible for ninety-nine per cent of the mysteries in the universe. So long as you said a phenomenon was due to radiation pressure you could get away with a vague argument, without supplying details or working anything out in accordance with the more rigorous standards that were required when "radiation pressure" was not involved. Radiation pressure died as a fad, though not of course as a valid dominant process in a few cases, sometime before 1945. The running was then taken up by "turbulence", followed by "magnetic fields". The current fad is "black holes". "Relativistic beaming" is currently a subsidiary astronomical fad, and "gravitational lensing" threatened a year or two ago to become another subsidiary fad. Each fad has a component of truth in it, but a component with a scale that is modest compared with all the grandiose claims. The majority of my successes in astronomy are unpublished and unsung. They come from adopting the simple precept that whatever is attributed to the current fad is wrong. Experience shows that one is then eventually proved to be correct in better than nine cases out of ten, a success-rate superior to anything I have ever been able to achieve by even the most determined burning of midnight oil.

'Fads do have their uses, however, in an inverted kind of way, by forcing one to examine critically one's own antifad point of view. Thinking about the development of a rotating nebula of the kind that was being contemplated in the von Weizsaecker theory, it seemed to me that the effect of a turbulent transfer of angular momentum between different parts of the nebula would be, not to produce a star and planets, but to produce two stars, one of them – the central star as one might call it – two or three times more massive than the other. The likely outcome would be a close binary similar to a known class of binaries, the W Ursa Majoris stars. Both components of W Ursa Majoris binaries are in rapid rotation, a fact which then led me to wonder why the Sun was not in rapid rotation.

'I have never timed explicitly how long it takes for sudden shifts to take place in one's point of view; in my experience I would say no longer than a few seconds. What I suddenly saw in this instance was that the preoccupation of the Cambridge school with the large angular momenta of the planets had tended to conceal a significant and important question: Why was the angular momentum of the Sun itself so small? To this point, I would probably have answered that most dwarf main-sequence stars rotate slowly, so that as a member of its class the Sun was quite normal. It was rapid rotators like the W Ursa Majoris stars that were abnormal. This was a matter of empirical experience, I would have argued. But why then was empirical experience the way it was? Every condensation occurring within clouds of interstellar gas would be likely to possess sufficient initial angular momentum to produce a rapidly-rotating star – unless a large fraction of the initial angular momentum were lost in some way. These considerations created a dilemma which would be resolved only if some process of angular momentum loss existed for most dwarf stars.

'Since angular momentum is a conserved quantity, the angular momentum which had been lost by the Sun could not have been destroyed. It had to go somewhere, and evidently the next question was where? If it had been returned back to the interstellar gas from whence it came, well and good, the arguments of the Cambridge school could then survive largely unaffected. But what if most of the angular momentum of the Sun were passed to the material that eventually formed into the planets? One would then understand at a stroke why the planets have so much angular momentum and the Sun so little, and disastrously the Cambridge position would collapse without a shot being fired, for it had always been assumed implicitly by the catastrophists that no internal process of angular momentum transfer from Sun to planets could exist, a supposition which I felt to be correct so far as the turbulence of the von Weizsaecker theory was concerned. The crucial point was that to explain the facts (a slowly rotating Sun with planets orbiting the Sun at distances large compared with the solar radius) it was essential for the angular momentum to be transferred across more or less empty space, *not through* a *continuous medium* as was the case for turbulence. This was the nub of the matter, the empty space. How could angular momentum go across empty space? It was on this question that the fate of the catastrophic school rested.

'The years immediately following the Second World War was the time when Hannes Alfvén emphasized the importance of magnetic fields in astronomy. Alfvén had a magnetic theory of the origin of the planetary system, according to which neutral atoms falling towards the Sun became ionized by solar radiation, whereon they were subjected to the influence of the magnetic field of the Sun. The latter would cause the ions and the electrons to rotate with the same angular velocity as the Sun itself, imparting to the ionized atoms an angular momentum of an amount that was a factor ~d/R above the angular momentum of a similar atom within the Sun, where d was the distance from the Sun at which ionization took place with R the solar radius. Because values for d of order 10^3 R could be contemplated, the excess angular momentum acquired by ionized atoms at large d would be of planetary order. This was assuming the planets to have condensed from a multitude of such ionized atoms. It was here, however, that I disagreed with Alfvén. I well remember arguing with him at length one day in the 1940s, that one could not apply to matter in bulk (such as was needed to form the planets) considerations that were applicable only for single atoms. If matter in bulk fell into the weak solar magnetic field at d >> R it would simply continue to fall, pushing inwards the field, instead of acquiring significant angular momentum in the manner of a single atom.

'While I could not feel that Alfvén's argument in terms of single atoms had come to grips with the essentials of the problem, the introduction of a magnetic field certainly opened up a new train of ideas. Suppose one began with a rapidly-rotating protosun around which there was a torroidally-shaped disk of gas, with each element of the disk in nearly Keplerian motion around the protosun (leaving aside for the moment how such a situation had come about in the first place). One could conceive of gas in the disk as having initial distances d not much larger than R. Was there a process that could transfer angular momentum from protosun to torroidal disk, across the gap between them? One could see from the known facts of solar physics that the answer to this question was quite likely affirmative, for the following reason. From solar physics one knew there is indeed a magnetic field in the present-day Sun, with an intensity in sunspots at least rising to hundreds or even to thousands of gauss. From solar physics one also knew that flares near sunspots cause chromospheric material to be expelled more or less radially outward from the Sun at speeds of order 1000 km s^{-1}. Such expelled ionized material carries a magnetic field with it, which expands outwards into a loop or lobe, the magnetic lines of force remaining connected to roots beneath the solar photosphere. In such a situation the magnetic field acts to transfer angular momentum from the Sun to the material expelled by the flare, a process which tends to maintain the expelled material in co-rotation. If the field were strong enough in relation to the amount of the expelled material, the angular momentum so transferred could be very considerable on a gram for gram basis, essentially because in this form of the theory the intensity of the field falls as the inverse square of the distance d, rather than as the inverse cube for a dipole. Suppose now that such expelled flare material, instead of being ejected entirely from the Solar System, were to encounter a rotating disk of gas, and to become incorporated therein. The effect would be to transfer angular momentum from the Sun to the disk, just the process one was seeking. Angular momentum does not go by itself across the gap between Sun and disk. It is carried across the gap by the expelled material, which however remains comparatively small in its total amount.

'In order to store angular momentum from a protosun in rapid rotation initially, the torroidally-shaped protoplanetary disk of gas must move outwards to distances which can easily be shown to be ~ $R(M/m)^2$, where R is the radius of the protosun, M is its mass, and m is the mass of the planetary disk. Thus for $M/m >> 1$, the disk of gas must retreat to large values of d, assuming that it comes to acquire most of the initial angular momentum of the protosun. The outcome of the process, if maintained sufficiently, would therefore be a slowly-rotating central star accompanied by a torroidal-distribution at comparatively large d of rotating protoplanetary material, the kind of situation needed to explain the state of affairs in the Solar System. With M essentially fixed at the solar mass, the order of magnitude of d, and hence the order of magnitude of the radii of the orbits of the main planets of the System, turns on the disk mass m. To obtain a case like the Solar System m was required to be a few per cent of M.

'Consider now what happens during the process of angular momentum transfer as the distance of the protoplanetary gas increases from an initial value of order R to an eventual value of order $R(M/m)^2$. The temperature of the gas falls from an initial value of, say, 4000K, to a value of order 4000(m/M), which for $m = M/40$, say, is only 100K, far below the thermodynamic condensation points of refractory materials, and even below the condensation point of water. Although most of the protoplanetary material would remain gaseous, largely H_2 and He, condensable substances would form as solids suspended within the outward-moving gas. Refractory substances with thermodynamic condensation temperatures above 1000K would come out of the gas as solid grains and chunks while the gas was still comparatively close to the protostar, while volatile substances like water would only come out of the gas as the most distant regions of the Solar System were reached. In short, variations of thermodynamic condensation temperature between substances with differing degrees of refractivity and volatility are reflected in a segregation process from the gas.

'If all the solid particles were tiny grains that were carried along by the outward-moving gas, this segregation process with respect to condensation temperatures would have no significant eventual effect, because all the particles would then be carried to the outer regions of the system, where they would be mixed in proportions that were the same as in the original gaseous phase. But if some of the solids grew to veritable chunks, planetesimals, large enough to resist being carried along by the gas, refractory solids would be left behind in the inner regions of the system, with the gas continuing to acquire angular momentum from the central protostar, and so expanding more and more to large values of d. The segregation process from the gas would then be reflected in a system of planetesimals moving around the protostar in approximately Keplerian orbits, planetesimals that had been left behind by the outward-moving gas, left behind in an ordered sequence with respect to condensation temperatures, the highest temperatures nearest to the centre.

'Three important correspondences with our actual Solar System emerge immediately from these considerations. We see why the innermost planets are very largely composed of materials of the highest refractivity. We see why the masses of these planets are small compared with the masses of the outer planets, because the original gas contained only a small fraction of refractory substances. And so long as the time-scale for the planetesimals to aggregate into planets was longer than the time-scale for the process of angular momentum transfer, the gas would necessarily be gone before the newly-forming inner planets acquired gravitational fields of appreciable strength. So there would be an absence of gas as the planetesimals aggregated into planets, except for a tiny amount that occluded itself at temperatures ~1000K within the solid planetesimals. So we understand Harrison Brown's conclusion quoted above: "*...during the process of Earth formation the mechanism was such as to prohibit the retention of an appreciable fraction of any substance that existed at that time primarily in the gaseous state.*"

'Chemistry and dynamics thus joined together. This was the substance of my contribution in 1955-56 to Harrison Brown's weekly series at Caltech. The broad picture as I have now described it has survived essentially unchanged in my mind ever since that time. Such changes as there have been are mostly additions, some of which may be worth mentioning here.

'Let us return to the above discussion, at the point where it was assumed that the protosun came to be surrounded by a torroidally-shaped disk of gas, the disk having appreciably less mass than the protosun itself. How did such a situation arise one can ask? There is no difficulty in answering this question in general terms. The angular momentum acquired by a protostar from the interstellar cloud in which it condenses is likely to be amply large enough for it to spin-up during condensation to a point of rotational instability, a condition which arises when rotary forces at the star's equator become comparable with gravity. A disk of gas then emerges at the equator, and, provided the process of angular momentum begins immediately so as to relieve the incipient instability of the star, the mass m of the disk can remain small compared with M, as it is required to do in order to explain the situation in the Solar System. If, on the other hand, the process of angular momentum transfer worked too slowly to prevent m from becoming comparable with M, a double-star system like W Ursa Majoris, not a planetary system, would be the outcome. At first sight one might think from these straightforward considerations that scooping together all the present-day planets, conserving their angular momenta in the process and dumping them into the present-day Sun, would necessarily have the effect of making the Sun rotationally unstable. It comes as a minor shock to one's sense of logic that this is not the case – the Sun would then be spun-up to an equatorial rotational speed of about 100 km s^{-1}, which is not sufficient to cause instability, 300 to 400 km s^{-1} is required for that. However, one then notices that an important component of the original protoplanetary material is missing from the present-day planets. This is the H_2 and He gas which must have accompanied the material of the planets Uranus and Neptune. From their densities, about 1.6 g cm^{-3}, it is clear that these two planets can contain little H_2 and He, little compared with what must have originally been present. How much must have been present? Calculating on the basis that the masses of Uranus and Neptune are largely made-up from the next most abundant group of elements, namely C, N, and O, the present-day masses have to be multiplied by the abundance ratio by mass of H, He to C, N, O, which is a factor ~50 (assuming the original protoplanetary material to have been of normal cosmic composition). This leads to the surprise that the original quantity of gas in the region of Uranus and Neptune must have been ~1500 Earth masses, about four times greater than the combined masses of Jupiter and Saturn. With an extra factor of about 2 appearing in the angular momentum contribution, due to the larger radii of the orbits of Uranus and Neptune, the original angular momentum of the protoplanetary material had to be greater than the present-day angular momentum – mostly contributed today by Jupiter and Saturn – by a factor of order 10, amply sufficient if dumped into the Sun to endow the resulting body with rotational instability.

'Although one's sense of logic thus survives unimpaired, a new problem evidently arises, for how did essentially all of the H_2 and He which originally accompanied the material of Uranus and Neptune contrive to escape from the periphery of the Solar System? The answer is clear in principle, if not in detail. As the torroidal disk of gas moved outward to greater and greater distances from the protosun, a stage was eventually reached at which the holding power of solar gravity weakened sufficiently for H_2 and He, the gaseous particles of least mass, to evaporate thermally from the disk, back to interstellar space. The main uncertainty concerns the source of heating in the gas, for which there are several contenders; radiation from the protosun, the kinetic energy of a persistent wind of particles from the protosun, and a relative motion of the whole solar system with respect to gas of the cloud in which it condensed. It is still unclear which of these processes was the dominant one.

'It is also necessary to face up to an old problem which is still not solved in detail. Given that there is a swarm of planetesimals in nearly Keplerian orbits around the protosun, how do the planetesimals aggregate together into a number of full-blown planets? One can make quick progress in this problem if the orbits have negligible eccentricities to begin with, for then even the small mutual gravitational fields between the bodies are sufficient to bring planetesimals in adjacent orbits into a gentle juxtaposition with each other, permitting them to join together into larger and larger bodies, a process which can be shown to continue until sizes of the order of the satellites of planets are attained. Mutual gravitational forces between neighbouring bodies then become too weak to overcome the shearing effect of the gravitational field of the Sun.

'At this stage of the aggregation process one needs an opposite dynamical situation in which a quite large number of bodies of lunar dimensions, initially in nearly circular orbits, contrive to develop orbits of considerable eccentricity, so that the orbits of adjacent bodies then interlace each other, when collisions at random occur between them. Such collisions lead to an aggregation of the inner planets on a time-scale of ~10 million years, and to an aggregation of the outermost large planets on a time-scale of ~500 million years.

'The big question here is whether as a matter of celestial mechanics such a system of bodies really do evolve from nearly-circular to considerably-eccentric orbits. I have for long had the intention to try resolving this problem by a computer calculation, but a recollection of the subtleties of resonance effects, mentioned in Chapter 1 and from Newton's discussion of the motion of the Moon and clearly demonstrated in the work of Laplace, caused me to desist. So I have tended to make do with the old saw that whenever entropy can increase it will do so. Without inflexible intelligent control, physical systems always seem to develop towards increasing disorder, whether in classic thermodynamic situations or in one's personal filing system.'

We break the long quotation at this stage, since the breakpoint now reached coincides with the closing remarks of Chapter 4, where we had a profusion of bodies in the region of Uranus and Neptune. As a few of the bodies grew to protoplanetary sizes, gravitational encounters then served to spray smaller debris both outward into the region of the Oort Cloud of comets, and inward to the region of the Earth, with collisions between the Earth and the debris adding volatile materials such as the terrestrial ocean and atmosphere, which could not have been present on the Earth in the first phase. This was because of the high temperature refractory condition that at first obtained in the region of the inner planets. So by consistent argument we have come full circle, to the resolution of the quotation from Harrison Brown with which the present chapter began.

CHAPTER 6

ON A POSSIBLY FUNDAMENTAL PRINCIPLE IN CHEMISTRY AND ITS RELATION TO THE ORGANIC SOUP THEORY OF THE ORIGIN OF LIFE

There is generally-agreed evidence that bacterial life was present on the Earth as long as 3.5×10^9 years ago, and there is good evidence of life going back even to the oldest-known rocks, 3.8×10^9 years ago. At the latter date, a banded-iron formation was deposited in what today is Western Greenland. To produce such a formation containing high concentrations of Fe_2O_3 and Fe_3O_4 there must have been a source of oxygen that was not atmospheric, in order to oxidize the normal FeO and FeS present in terrestrial rocks. Atmospheric oxygen can be excluded, it may be noted, because atmospheric oxidation of FeO and FeS leads to the deposition of red-coloured sandstones without high concentrations of Fe_2O_3 or Fe_3O_4 being present in the sediments. The likely source of the needed oxygen was a by-product from such bacteria as *Pedomicrobium*, shown in Figure 17. Also shown in Figure 17 is an object discovered by Professor Hans Pflug within a sample of the Murchison meteorite. According to Pflug such objects are widely dispersed in the meteorite.

For those who believe that life originated on the Earth in a so-called organic soup, the slice of time available for the discovery of the genetic and enzymic complexity present in bacteria is reduced to a thin window. A truly immense stride in biochemical evolution would need to have been accomplished within the 800 million year time interval between the origin of the Earth 4.6×10^9 years ago and the date of the oldest presently- exposed rocks. If still older rocks were discovered, this 800 million year interval would almost surely be squeezed from below, since it would be unlikely for the origin of life to have coincided exactly with the oldest rock sequence that just happens to be exposed at the present day. And the 800 million year interval can certainly be squeezed from above, since the Earth at its formation was not in a condition to harbour life. The Earth had to acquire its volatile materials before there could be any possibility of life even existing, let alone originating here. The likely process whereby the Earth came to obtain its volatile materials has been described in

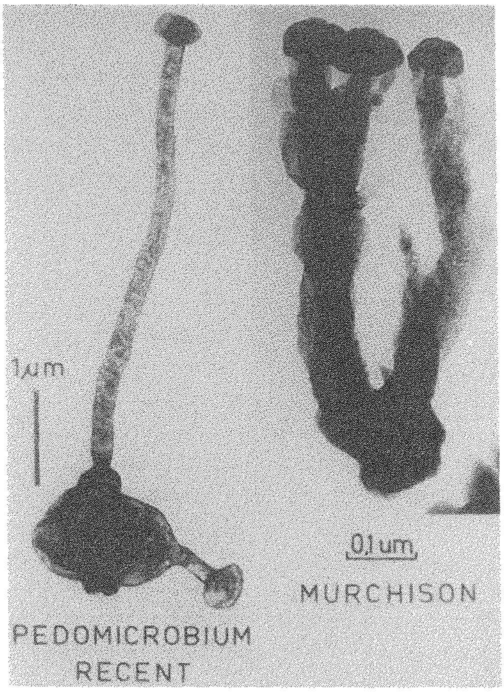


Fig. 17. Organic structures with flower-like heads found in the Murchison meteorite by Hans Pflug compared with a recent iron-oxidizing bacterium, *Pedomicrobium*.

Chapers 4 and 5, where it was seen to have arisen from the intricate dynamics of the condensation of the outer planets, especially of Uranus and Neptune, a process which has been estimated to have occupied the first ~500 million years in the history of the planetary system. These squeezing factors narrow the window for the hypothetical origin of life down to almost nothing at all.

Such unfavourable considerations have led members of the organic soup establishment to speak of 'instant life', implying that the astonishing biochemical complexity of life could arise all in a moment, just as it is supposed to have done by the creationists. That anything so absurd could be proposed at all is a consequence of the extent to which earlier absurdities of the organic soup establishment have been tolerated by a scientific world that has never been anxious, for the sake of its own peace of mind, to look into matters closely.

The amount of organic material present in the supposed terrestrial soup was miniscule compared with the amount which could have been present in the outer regions of the Solar System. If the hypothetical soup is taken to have contained as much organic material as there is currently in the Earth's biosphere the amount would have been $\sim 10^{18} - 10^{19}$ gm, a trivial amount compared to $10^{29} - 10^{30}$ gm for the outer regions of the Solar System. Small-scale debris in the latter could well have intercepted a considerable fraction of all sunlight, $\sim 4 \times 10^{33}$ erg s^{-1}, enormous compared to the amount of sunlight intercepted by the Earth, only $\sim 2 \times 10^{24}$ erg s^{-1}. Thus the availability of light for photosynthetic processes was also vastly larger in the outer regions, as was the availability of chemoautotrophic sources of energy and of radioactive energy. In relation to photosynthesis, it may be noted that maximum efficiency occurs at light intensities significantly lower than the flux of sunlight at the Earth's distance from the Sun, and that photosynthesis can continue down to very low light intensities.

Commonsense would therefore dictate that a scientific establishment which seeks passionately to avoid the conclusion that life is a large-scale cosmic phenomenon, and which must therefore maintain that life originated in the Solar System, would do far better to put its trust in processes occurring in the outer regions, the regions of Uranus and Neptune and of the genesis of the comets, than to rely on a harsh rocky Earth initially even without water and without atmosphere. Our impression is that the life-in-the-solar-system establishment would presently be glad to settle for this revised position, if only a way could be seen to avoid the issues still to be discussed in subsequent chapters, issues concerning the incidence of terrestrial diseases and of a driving evolutionary process from outside the Earth affecting the whole of terrestrial biology. These embarrassing concomitant matters are currently forcing the walking wounded of which this brigade is largely made-up to stay mewed in a siege condition, trying to maintain the old soup theory with their currently plaintive cry of 'instant life'. Otherwise sensible people think that our denunciations of the educational system are at best an amusing but cranky foible. This we believe to be wrong, for the educational system is responsible for producing a hugh miasma of erroneous belief against which even the most sceptical and watchful of us are nine-tenths powerless to resist. Notorious examples arise from time-to-time, but one has to suspect that even the most flagrant of them is but a tip of a very large iceberg. Indeed one has to suspect that, so far as science is concerned, education has reduced mankind in the year 1984 to a condition not too far removed from outright insanity, a suspicion that is all too frequently confirmed by weekly science magazines, with their perpetually hysterical atmosphere of breathtaking progress, but which actually consist overwhelmingly of eyewash.

Those organic soup specialists who have claimed to produce organics of interest to biology from initially inorganic materials, by irradiating the latter with high-grade sources of energy such as ultraviolet light and accelerated particle streams, have certainly contributed a great deal of eyewash, brazen eyewash if such can be contemplated. Brazen from the outset, because some of the claimed inorganic materials were almost surely not inorganic at all. Almost surely the CH_4 used in so-called prebiotic experiments was of biological origin. So are more exotic sources of carbon such as ethylene, and so quite likely is the principal source of nitrogen, NH_3. What should be considered so remarkable, a sane person can wonder, about producing substances of biological interest from what is already biological?

Doubtless the perpetrators of what, to put it plainly, has been a deceit would claim that CH_4 could have been obtained by wholly inorganic means. But could it? The circumstance that nobody connected with prebiological experiments, and nobody apparently among the throng which has accepted the claims of such experiments, has bothered to discuss this question, or even we suspect to notice that it exists, shows just how far the whole scientific community persistently deludes itself. Without the dinning thud of the educational process we doubt that such an extreme measure of self-deception would be possible. Somewhere, among people left to think quietly without being avalanched by the weekly science magazines, the point would long since have been noticed, and the organic soup theory would long ago have been consigned to where it belongs.

It is now some years since one of the authors arrived at a helpful proposition where unsolved problems are concerned, namely that it is useless to follow popular opinion, because if solutions to unsolved problems lay where popular opinion holds them to be they would have been found already. To this we now add the useful dictum that wherever a long-lasting confusion exists over the meaning of words something of significance lies waiting to be discovered. On this basis, the confusion over exactly what one means by ‘inorganic’ and ‘organic’ in chemistry deserves investigation.

Beginning with the Oxford Dictionary:

> *inorganic*
> Chem., of compounds not entering into composition of organized bodies; i. chemistry, that of mineral substances; not arising by natural growth; extraneous.
>
> *organic*
> Chem., of compounds existing as constituents of organized bodies, of hydrocarbons or their derivatives.

While these definitions touch our perceptions of what we mean by 'inorganic' and 'organic', they hardly approach the level of precision demanded in a scientific discussion.

Advanced texts usually take it for granted that we already know the meanings of the words. Thus *Advanced Inorganic Chemistry* (F.A.Cotton and G.Wilkinson, Interscience 1972) lives up to its title by plunging on the first page of the first chapter straight into group theory. As often in such situations an appeal to the *Encyclopedia Britannica* yields as clear an exposition as probably can be obtained. The general article on Chemistry (15th Edition) has the following on its first page:

> *Subdivisions of chemistry*. The field of chemistry encompasses the study of an uncounted and theoretically almost unlimited number of compounds. By the early 1970s there must have been more than 1,000,000 individuals working on chemical problems in independent, academic, industrial, and government laboratories, throughout the world for a myriad of personal, social, economic, and political reasons. In systematizing chemical knowledge and activities by grouping together related compounds, related systems, related methods, and related goals, a number of subdivisions of chemistry have developed: These subdivisions provide the basis of organization of academic curricula and literature and of bringing together scientists who share common interest. During the first half of the 20th century, undergraduate curricula were almost exclusively organized into courses in inorganic, analytical, organic, and physical chemistry and biochemistry, which were usually studied in that order. This organization of subject material is still apparent in many college catalogues, but it is difficult to defend, and, accordingly, many attempts are being made to organize academic programmes along other lines.
>
> *Organic and inorganic chemistry*. Organic chemistry and inorganic chemistry are subdivisions based upon the elements present in the compounds. Organic chemistry is the chemistry of carbon compounds, which, of course, also contain elements other than carbon, such as hydrogen, oxygen, sulphur, nitrogen, phosphorus, and chlorine. Inorganic chemistry encompasses all substances that are not organic. The separation of the study of carbon compounds from the rest of chemistry is defensible on the basis of the sheer numbers of carbon compounds that are of great interest and that not only have been but are still being intensively studied. The structure of the carbon atom is unique among atoms and makes possible this great array of compounds, which are stable under atmospheric conditions on Earth but are also sufficiently reactive to make possible a great variety of chemical changes.

The writer essentially reduces the division between inorganic and organic chemistry to a matter of arbitrary choice. One chooses 'organic' to be the chemistry of the compounds of carbon, and all other compounds are automatically 'inorganic' by definition. Without nuclear transmutations it would then be impossible to synthesize organic compounds from inorganic, for the evident reason that no inorganic material would contain carbon atoms. Substances like $CaCO_3$ would have been classed as 'organic', which of course would be against all precedent. This is not the way anybody thinks of CO, CO_2, or their associations in molecules like $CaCO_3$.

Whereas the writer in the *Encyclopedia Britannica* evidently defines the range of 'organic' substances too widely, the confinement of organics to hydrocarbons and their derivatives, as in the definition of the Oxford Dictionary, is too narrow. Amino acids are not hydrocarbons, nor is urea (H_2NCONH_2). The synthesis of urea from ammonium cyanate by Friedrich Woehler in 1828 is often mentioned at the very beginning of studies in organic chemistry, because it is supposed to have some deep significance. We have difficulty in understanding what the significance really was, although we have less difficulty in seeing what it was supposed to be. Urea is a biochemical, and what was claimed by people with a distaste for the old doctrine of vitalism was the synthesis of biomaterial from non-biological sources, a claim that was almost certainly false.

There is a great quantity of nitrogenous material in the soil, much of it in the form of ammonium salts of various kinds. While it would be difficult to prove with mathematical rigour that absolutely none of this material is of abiological origin, it is widely accepted that the bulk of nitrogenous material in the soil is the excretion products of denitrifying bacteria that operate to break down substances of biological origin. So if the ammonium cyanate in Woehler's synthesis came from a so-called natural deposit of some ammonium salt the chances are that the urea was obtained from material containing a biological product, in which case the logic of the anti-vitalist argument would be weakened to the point of nonsense.

It is clear from the example of the synthesis of urea that views on what constitutes 'inorganic' and 'organic' are by no means covered by unemotive definitions such as that attempted by the writer of the *Encyclopedia* article. The relation of chemistry to biology is evidently involved, with chemistry seeking to claim a status independent of biology. In the sense that, given the elements in atomic form, they could be assembled under controlled conditions into small quantities of any organics one cared to specify, chemistry does have a status independent of biology. But in the sense that only very small quantities may be attainable in practice the issue requires further consideration.

A DEFINITION OF INORGANIC MATERIAL

The elements are not available in atomic form in quantity except in mass flows from stars and in the fraction of preplanetary material that became heated to high temperature in the solar nebula. By inorganic material we shall mean those compounds which form when such heated gas cools at low pressure (ca. 10^{-6} bar in preplanetary material) together with those more complex compounds that can arise abiologically when planets themselves are condensed. For example, MgO and SiO_2 condensing in cooling gas at low pressure can form $MgSiO_3$ when brought together at much higher pressures during planet formation. This example cannot be extended automatically to include all minerals, however, because quite likely bacteria were involved in the formation of some minerals. Each case must be considered according to the facts.

Iron and others of the transition elements condense as metals at ~1500K. Should the iron condensates become aggregated into fair-sized blobs, the metal is preserved as cooling proceeds further, which is the likely route whereby the metallic core of the Earth orginated. As the temperature falls below 1000K, the thermodynamic balance for iron swings towards FeO and FeS, which happens for small particles and at the surfaces of large blobs. Other metals also emerge as oxides, Al_2O_3 for example. As the temperature falls below 1000K any remaining excess of oxygen goes to H_2O, while nitrogen emerges as N_2 unless it can be argued that catalysis causes a reduction to ammonia in the presence of an excess of H_2. Likewise carbon emerges as CO or CO_2 unless catalysis produces a reduction to CH_4 and other hydrocarbons. The free energy made available in the reaction $N_3 + 3H_2 \rightarrow 2NH_3$ is so little, however, only about 8kcal per mole of N_2, that the reaction does not go towards ammonia at a low pressure of $\sim 10^{-6}$ bar unless the temperature falls towards 100K. Because of the difficulty of maintaining suitable catalytic surfaces we take nitrogen to emerge in preplanetary material as N_2.

The free energy made available by the Fischer-Tropsch reaction with the basic form $CO + 3H_2 \rightleftharpoons CH_4 + H_2O$ is about 35kcal per mole of CO, which permits the reaction to swing from left to right at a temperature of ~500K or less. This is at the low pressure of the preplanetary gases. While such a temperature could in itself permit catalysis to occur, the reaction has proved difficult to operate even under controlled industrial conditions, with great care over the preparation of catalysts being found essential. The Fischer-Tropsch reaction was actually used in Germany in the extremis of the 1939-45 war for the production of synthetic oil. Commonsense suggests that if the reaction is difficult at pressures very much higher than in the solar nebula, and with carefully prepared catalysts, it was hardly likely to have been effective in preplanetary material. Otherwise the price of oil would never have soared to ca.$30 per barrel.

Taking the production of CH_4 and other hydrocarbons to be ineffective for the reasons just stated, the present definition of 'inorganic' compounds agrees with empirical practice to a remarkable degree, almost as if empirical practice had been based on considerations of star formation and planet formation. The notable exception is NH_3. We suspect that the the true inorganic form of nitrogen is N_2, not NH_3, which accords with the fact already noted above that the Earth possesses little, if any, inherent store of the ammonium radical outside of what has been produced by micro-organisms. The conclusion also has a correspondence with actual practice in industrial chemistry, where the production of NH_3 by the Haber-Bosch process has long been regarded as a watershed that leads to the production of organic materials that would otherwise not be accessible commercially. The ammonium situation is therefore critical and we shall discuss it in a little more detail in a moment.

A CONJECTURE

The way things happen to be in nature, it is not possible by purely chemical processes to pass in quantity from inorganic materials to organic materials.

The restriction to chemical processes requires atoms to be conserved. This restriction is related to an exclusion from the argument of biointellectual activities, as will be emphasized at the end of this chapter.

THE HABER–BOSCH PROCESS

This process well illustrates the problems one encounters in trying to find an example that disproves the above conjecture. The nitrogen used in the reaction $N_2 + 3H \rightarrow 2NH_3$ is inorganic, but what of the H_2? Early applications of the Haber process obtained H_2 by passing H_2O over hot coke, which of course was of biological origin. Later applications have obtained the H_2 from hydrocarbons of overtly biological origin. But H_2 can also be obtained from the electrolysis of H_2O, which is inorganic according to the above definition. So one asks whence comes the electricity used for electrolysis? If from coal or oil-fired power stations we are instantly back to biology. If from nuclear reactors then processes outside chemistry have intruded. If hydroelectric, the solar energy needed to lift water comes also from nuclear processes in the Sun. In all cases we have considered, the route towards disproving the conjecture turns out to be similarly blocked. Exceptions can be devised but only it seems for very small quantities of material.

Extraterrestrially, free hydrogen must be considered as an inorganic material. Even so, conditions once again conspire against the production of NH_3. Pressures in the preplanetary material are low. Temperatures in the atmospheres of planets such as Jupiter and Saturn are also low. Could one have the right kind of carefully-prepared catalyst, and could the catalyst be free-floating in a planetary atmosphere? In the absence of biology, conditions are not propitious. Micro-organisms, on the other hand, possess exceedingly efficient catalysts, they are small enough to float in gases and liquids, and they can function down to remarkably low temperatures. Biology is a wholly dominant catalytic agent, and in its presence there can be no effective competition from abiological processes.

WHY THE CONJECTURE COULD BE A CRUCIAL MATTER OF PRINCIPLE

As cooling of the preplanetary gases proceeded below 1000K, the ensemble of molecules would go more and more out of thermodynamic equilibrium. This was because the Boltzmann factors affecting equilibrium increase in importance as the temperature falls. At high temperatures thermodynamic equilibrium favours there being as many gas molecules as possible, which leads to inorganics like CO and N_2 being dominant. At sufficiently low temperatures, however, the positive free energy values obtainable in the formation of CH_4 and NH_3 favour the reduction of CO and N_2 in the presence of an excess of H_2. But according to the point of view developed above, the reactions $N_2 + 3H_2 \rightarrow 2NH_3$, $CO + 3H_2 \rightarrow CH_4 + H_2O$, do not go in a purely inorganic situation. When the temperature has fallen low enough for them to be thermodynamically preferred, their rates are then too slow unless highly specific catalytic surfaces are present, which seems unlikely to be maintained without the intervention of biointellectual processes. So cooling preplanetary gas at low pressure goes out of chemical balance; and it does so for some of its commonest molecules. According to our point of view it is this situation, applicable everywhere not just in the Solar System, which creates the primary niche for biology. Biology is 'nature's way' of moving much closer to thermodynamic equilibrium than would otherwise be possible. Stated the opposite way, if thermodynamic equilibrium could be reached abiologically, if inorganics could go to organics abiologically, if our conjecture were substantially untrue, biology would be short-circuited by the inorganic world and could not exist.

The physical raison d'être for biology is thus seen to be the widespread deviations from thermodynamic equilibrium that would otherwise exist on a universal scale, deviations which occur whenever hot material cools at protostellar pressures. Even if there were not many facts which show biology to be a universal phenomenon, the notion that the exquisitely complex enzymic systems of biology exist on the Earth alone, in order to cope with only a local departure from thermodynamic equilibrium, could be seen to be improbable if not indeed absurd.

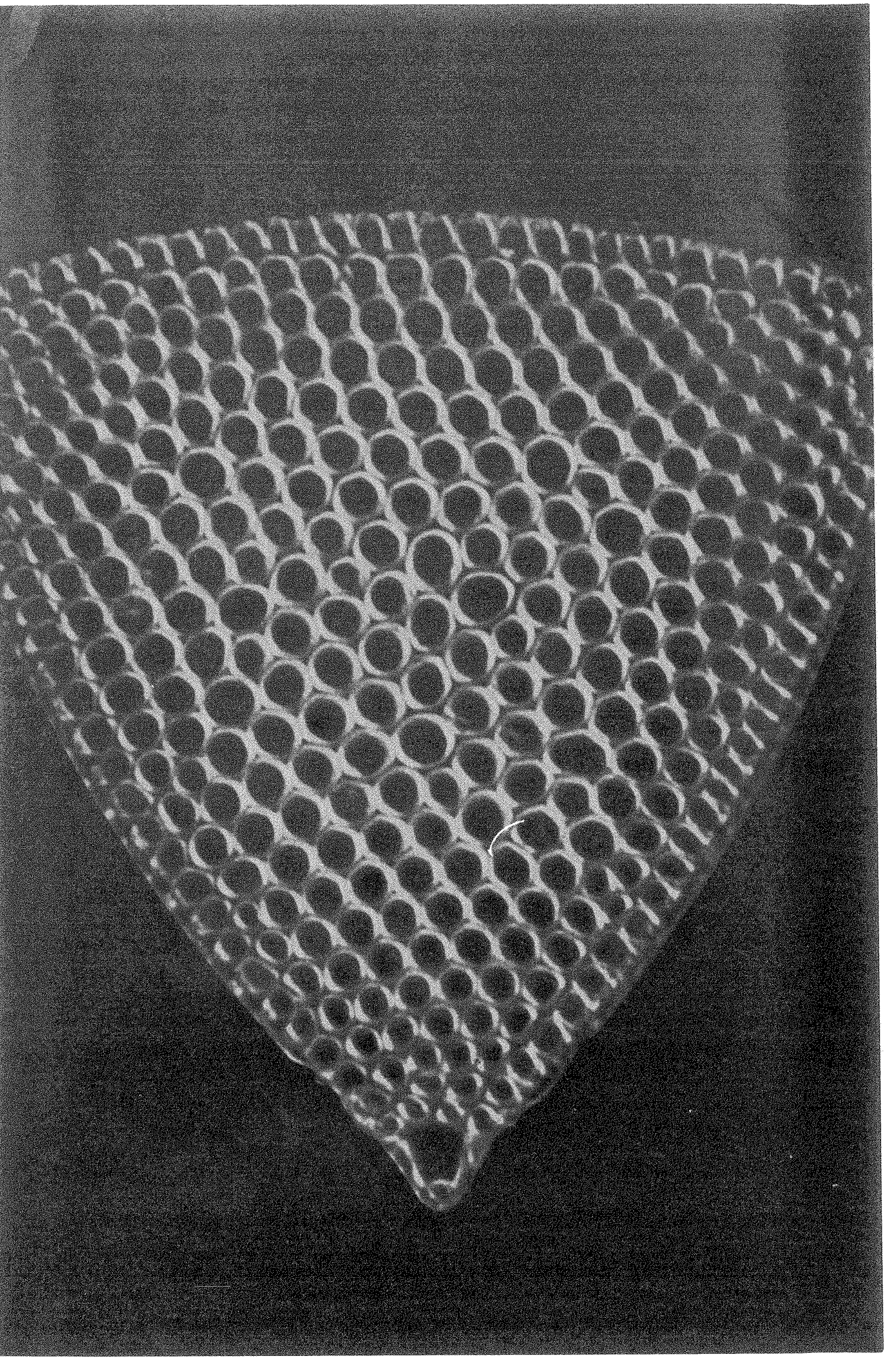
Anaerobic micro-organisms are likely to be found as chemoautotrophs living everywhere in deviations from thermodynamic equilibrium, provided the twenty or so chemical elements necessary for life are available in the localities in question, of which there are a number at the present day in our own Solar System. Of such possibilities, only the Moon (where the requisite elements are not all available) and Mars have so far been looked at, the latter only cursorily. Even so, many who have examined the details of the Viking missions to Mars have concluded that the missions were unfortunately terminated at the threshold of success.

All this has been subject to the chemical restriction that atoms be conserved. Although biology cannot it seems be by-passed in its control over thermodynamics in cooled material when atoms are conserved, biology can indeed be by-passed if this chemical condition is broken. Given nuclear power, H_2 becomes available terrestrially through electrolysis, when an appreciable route from the inorganic to the organic at last exists. Even so be it noted, biology has not been backward in incorporating the fruits of nuclear processes into its repertoire. More than 10^{17} grams of biomass are produced on the Earth through photosynthesis annually. It will be a long time before human technology produces 10^{11} tons of organics per year. Without photosynthesis, i.e. without nuclear input, biology is confined to comparatively low-grade chains for energy production. With photosynthesis, biology makes use of higher grade forms of respiration, as for instance glycolysis. It seems clear that biology has used the lower forms to step-up to the higher forms.

The advent of nuclear fission has had a profound effect on our human social systems. Perhaps the fear of people, as at Three Mile Island, of nuclear activities in general has something more in it than a straightforward fear of the explosive violence of nuclear weapons. In by-passing biology in a deep thermodynamic sense, nuclear power could in principle provide a mode of organized existence that was independent of biology. One can imagine a nuclear-powered robot society, computer-driven, existing independently of biology. Not tomorrow as in science-fiction. But why not in 50,000 years? With the discovery of nuclear power a crucial dividing line has been crossed. It is conceivable this crossing is vaguely perceived by people at large, and the perception of it may lie at the root of present unease, which may even be preprogrammed within us, as great abilities in mathematics and music appear to be preprogrammed. Preprogrammed by biology as a defence against a dangerous new rival.

The superpower confrontation is proceeding along what appears to be laid-out guidelines, with the two superpowers moving step-by-step rather as a couple executes a dance defined by chalk marks on a ballroom floor. The music grows louder, the beat more insistent, until the two are impelled together, whether they want to be so or not, to find themselves in a particle-antiparticle collision. Or like driving in a steadily thickening fog. Up to a point the driver remains in control. Then quite suddenly control becomes very difficult, and unless one somehow gets off the road a crash becomes inevitable. Mathematicians and physicists understand stability phenomena like this very well, and have no difficulty in perceiving that the progressive shortening of the time available for a response to be made to a first-strike nuclear attack is analogous to a thickening fog on the road. Perhaps the vehicles can be slowed and edged off the road before the fog becomes too thick. If not, a failure in the educational system to understand the broad principles of the interrelation of biology with physics and chemistry will in our view have played no small role in provoking the eventual disaster.

The last part of the above discussion may prove fanciful, and we hope it will. The concept of an entirely robot-controlled society may well also be fanciful. We have included it here so as not to exaggerate the strength of our chemical conjecture. If such a possibility were excluded, our conjecture could be upgraded, because devices such as nuclear reactors and turbines for hydroelectric power, being man-made, would then become biointellectual features that were excluded unnecessarily in the formulation of the conjecture. Including biointellectual activities alongside biochemical processes, the restriction on atoms being conserved may be abandoned, and we would then assert the strong proposition that it is impossible to pass in quantity from inorganic materials to organic materials except through the intervention of biology.



CHAPTER 7

BIOLOGICAL ACTIVITY IN THE EARLY SOLAR SYSTEM IN ITS OUTER REGIONS

However often one learns to accept the amazing ways in which biological systems make use, not only of their gross environment but of subtle aspects of physics and chemistry, one never becomes quite inured to new surprises. It was so with us on the day we learned in private communication that some species of bacteria can precipitate uranium salts from very weak solutions, 'they are practically uranium-eaters' our acquaintance told us. A possible answer to an unresolved conundrum over the Oklo reactor occurred to us thereupon to which we shall return in a moment.

Once again it was a surprise when we learned from Drs. R.B. Hoover and M.J. Hoover that some species of diatoms are able to concentrate other potentially fissile elements:

> '...it is established that many species are capable of thriving in environments containing extremely high concentrations of unusually lethal radioisotopes such as americium, plutonium, strontium, etc. Diatoms thrive in highly radioactive ponds, including the U-pond and the Z-trench at the Hanford facility, with the latter containing over 8kg of various radioisotopes of plutonium. Not only do diatoms live in this environment, but they seem to have a remarkable affinity for plutonium (c.f. R.M. Emery, D.C. Klopfer and W.C. Weimer, 1974, in Report prepared for the U.S. Atomic Energy Commission under Contract AT(45-1): 1830. BNHL-1867, p.44). The algae of these ponds, of which diatoms are by far the dominant form, concentrate ^{241}Am three millionfold, and certain isotopes of plutonium are accumulated to 400 million times the concentration in the surrounding water. The plant life in these radioactive ponds contains more than 95% of the total plutonium burden. Diatoms and *Potamogeton* alone contain more than 99% of this plutonium. In such an environment, diatoms grow in great abundance while continuously subjected to high levels of x-rays, gamma rays, alpha and beta particles'.

Deaths from leukaemia tend to show a peculiar very local clustering effect. We recall the example of a remote valley in New Zealand where over a time scale of a few years there were ~10 such deaths, a valley where there was no nuclear reactor. A cluster of six similar cases has recently come to public notice in the village of Seascale, closeby the Sellafield nuclear reactors of W. Cumbria. Somewhat naturally, the media have attributed the latter unfortunate deaths to the presence of the reactors, and as an outcome of media pressure the Ministry of the Environment of the British Government was led to set up a committee of enquiry into the matter. Members of the committee knew perfectly well from fully-attested statistics that neither the natural radioactive background nor the slight increment in the background caused by the reactors could explain the facts (except as a truly monstrous statistical fluctuation) and they also knew, which their critics apparently did not, of the existence of similar clusterings elsewhere, as in New Zealand where there had been no nuclear reactor. So the committee simply, and somewhat innocently, reported that the Sellafield reactors could not have been responsible for the six leukaemic deaths. Naturally London journalists writing for British weekly science magazines, who frequently meet together at various gatherings and perhaps over a pub lunch or two, and so tend to be of a single mind on such issues, had a field day over it. What the journalists could see with startling clarity, just like everybody else, was that while leukaemia cases might occur in small clusters it was apparently most peculiar that one such cluster should be found sitting almost on top of a considerable complex of reactors, where for one reason or another the management had not over past years been able to avoid the escape of small quantities of radionuclides into the local environment.

Diatoms do not rate highly in the educational system, and topics with low priorities receive scant attention in both books and in lecture rooms. Quite likely therefore, the facts concerning diatoms reported to us by Drs. R.B. and M.J. Hoover are unknown to the umpires of the weekly science magazines, to the media, and possibly to officialdom decked in its magisterial robes. Otherwise there would surely have been a rush to examine the water supplies at Seascale, not with a view to its dissolved contents, but with respect to micro-organisms suspended within it. Especially as the nearest Lakeland valley to this part of the Cumbrian coast is Eskdale, where granite rocks containing a high level of uranium (and so of the decay products of uranium, some alpha – active) outcrop the surface. If cultures in water pipes concentrate such products in the manner described above, with local populations imbibing the micro-organisms, perhaps with a further concentration occurring in the human body itself, the facts would become intelligible. One reason for the siting of the Sellafield reactors was related to the water supply from Eskdale and Ennerdale, it is ironic to notice, and this might be the connection the media have been seeking, an innocent connection that would not be much to their liking if it turned out to be true. Similar conditions obtaining elsewhere would of course produce the same effect, regardless of whether there were nuclear reactors in the districts in question.

Concerning the Oklo reactor, Dr. S.A. Durrani of the University of Birmingham wrote as follows:

> 'Nature, it would seem, had anticipated man by something like 1,800 million years in bringing about the first self-sustained nuclear chain reaction on the Earth. And, contrary to common belief, it was not in the squash court of the University of Chicago in December 1942, but in the wilds of what is today the Republic of Gabon at a place called Oklo that this fantastic phenomenon took place.
>
> 'The history of the discovery of the phenomenon, as it unfolded during the symposium, is briefly as follows. In June 1972 a team working under the direction of Dr. H.V. Bouzigues at the CEA service laboratory at Pierrelatte in France noticed a marked anomaly in the abundance of the uranium-235 isotope (0.7171 ±0.0010 in atomic per cent instead of the normal 0.7202 ±0.0006) during the certification of a secondary standard of UF_6 by the gas diffusion method. Later, much larger depletions of this isotope were discovered (down to 0.621%, and eventually to 0.296% U – 235) in uranium samples from this source, which was traced back to the Oklo deposit. First positive proof of the hypothesis that a natural chain reaction was responsible for the depletion of the fissile component was furnished by Mme. M. Neuilly and co-workers of CEA through the measurement of the ratios of fission-product rare earths detected in the ore by the spark source mass spectrometry technique. Two simultaneous submissions by the above two groups on September 25, 1972, to the Proceedings of the Academy of Sciences, Paris, announced the discovery and the proposed explanation of this remarkable phenomenon. It was pointed out that at the time of the reaction the natural abundance of the relatively fast-decaying ^{235}U isotope was more than 3%. This natural "enrichment", helped by the moderation of the fission neutrons by the water content of the soil which enhanced their fission efficiency, and possibly by the relative absence of neutron-absorbing elements in the surroundings, allowed a nuclear chain reaction to develop. It is perhaps worth mentioning that such a natural chain reaction had already been predicted, on theoretical grounds, by several scientists, notably by P.K. Kuroda as early as 1956. The scientific secretary of the symposium, Dr. R. Naudet of CEN, Saclay, has since late in 1972 been leading the "Franceville Project" established by the French CEA to investigate the phenomenon, and has done a great deal to promote its study internationally.'

The first announcement from the CEA laboratory provoked scepticism among nuclear physicists, because of the point alluded to briefly in the above quotation, the need for an absence of 'neutron-absorbing elements in the surroundings'. Very little in the way of elements such as cadmium or gadolinium would have poisoned the reactor, and the difficulty was to see how under aqueous conditions all such neutron poisons could have been conveniently absent. What happened subsequently was that French physicists gathered sufficient evidence concerning the presence of fission products at the site of the reactor to convince the sceptics. But without the problem of neutron poisons being cleared up satisfactorily.

The statement that some bacteria are 'practically uranium eaters' suggested both a possible cause of the Oklo phenomenon and a resolution of the neutron poison problem. Imagine bacteria in comparatively still water precipitating around themselves a high density coating of increasing thickness of some uranium salt, uranium oxide most likely, rather as bacteria precipitate calcarious material to produce stromatolites. The increasing coating would eventually cause the bacteria to sink to the bottom of the lake or pool in which they had been suspended. In the floor of the pool, suppose there to have been a bowl where more and more uranium-coated bacteria accumulated. A stage would be reached at which the growing colony went critical in the manner of a simple boiling water reactor using enriched uranium, the 'enrichment' for ^{235}U being per cent at the epoch of the Oklo reactor. The violent motion associated with boiling could scatter the bacteria in a timescale less than the interval $\sim$30 seconds required for the appearance of the main complement of delayed neutrons, thus maintaining stability should such a system threaten to become seriously supercritical.

All living systems produce a great measure of chemical segregation, accepting some elements and rigorously rejecting others. We have never heard of the elements gadolinium or cadmium being present in living organisms for example. In this way one could elegantly understand the absence of neutron poisons from a biological reactor, thereby overcoming the previously-mentioned difficulty which at first sight had erroneously seemed to obviate the French discovery.

If a so-called 'natural' reactor could arise 1800 million years ago, when the enrichment of ^{235}U was $\sim$3per cent, bioreactors could arise almost trivially one might suppose in the early days of the Solar System when the enrichment was $\sim$30per cent. This likelihood raises the possibility of an escape for life from the present-day straightjacket of temperature, the slim zone here on the Earth between being boiled alive (Venus) and being frozen silly (Mars). With a controlled heat source inside an adequately insulated body, life on the outside of the Solar System in its early history could have adjusted temperature conditions to suit itself.

The sites were planetesimals of various sizes, from a lunar scale down to a cometary scale, with liquid interiors generally at ~300K, surrounded by surface shells of frozen material having low heat conductivity. Superinsulators with porous structures have coefficients of heat conductivity $\sim 10^{-4}$ watt cm^{-1} K^{-1} (c.f. J.E. Parrott and A.D. Stukes, *Thermal Conductivity of Solids*, Pion Ltd., 1975, 143), a value that will be used in the following discussion.

The heat-release process was the one already described above, determined by the precipitation of the potentially fissile elements U, Th, by micro-organisms that subsequently sank towards the centres of the planetesimals where they contributed together to produce a critical reactor which stabilized itself by generating convective motions that mostly prevented the central concentration of fissile material from attaining a runaway supercritical condition (although a recollection of Oort's exploding planet flickers in one's mind at this point).

To estimate the potential amount of fissile material, it seems reasonable to suppose that breeding of ^{232}Th and ^{238}U to ^{233}U and ^{239}Pu could occur in a large measure – from the point of view of reactor technology this should have been 'easy' at a time when the ^{235}U enrichment was so high. Solar abundance tables by numbers of atoms give $(U + Th)/(C + N + O) \simeq 6 \times 10^{-9}$, which is a ratio by mass $\sim 10^{-7}$. This estimate rests on the amounts of U and Th actually found in meteorites, however, which raises the possibility that what has been measured for meteorites are low values subsequent to appreciable denudation by bacterial action. Calculations based on the so-called r-process for the primordial genesis of U, Th have run an order of magnitude higher than the measurements, and here we may well have the reason for this discrepancy – the meteoritic values are not primordial, thereby destroying a hallowed assumption of meteoritic chemists, which by a refusal to question it has attained the status of a religious dogma. If we take an intermediate position, with $(U + Th)/(C + N + O) = 3 \times 10^{-7}$ by mass, we shall not be far wrong.

The output of energy from the total fission of U + Th is $\sim 10^{18}$ erg g^{-1}. Hence with most of the mass of material being C, N, O, on the outside of the Solar System, the fission energy yield per gram from its content of U + Th would be $\sim 10^{18}. 3 \times 10^{-7} \simeq 3 \times 10^{11}$ erg. With $10^{29} - 10^{30}$ gm of C, N, O, the total energy available was therefore $\sim 10^{41}$ erg.

With the material having a density ~1 gm cm^{-3}, the mass of a body of radius R was $\sim 4\pi R^3/3$, and the total energy available for release inside the body $\sim 3 \times 10^{11} \times 4\pi R^3/3$ erg, with R in cm. Suppose such a body to have a surface shell of thickness 1 km through which the temperature fell from 300K on the inside to ~100K at the outer surface. With a heat conductivity of 10^{-4} watt cm^{-1} K^{-1} the heat loss through the shell would be $10^3 \times (4\pi R^2)\ (200/10^5)$ erg s^{-1}, with R again in cm. The heat availability is sufficient to make good this loss for a time T seconds given by

$$10^3 (4\pi R^2) (200/10^5)\ T = 10^{11} \times 4\pi R^3. \qquad (24)$$

Or with T in years and R now in kilometres,

$$T \simeq 2 \times 10^8 R \text{ years}. \qquad (25)$$

There would evidently be no difficulty for a body of lunar size, $R > 1000$ km, maintaining a liquid condition in its interior, and some comets might have been able to do so over at least the first 500 million years in the history of the Solar System. Excess energy output would simply lead to a thinner surface shell, while a reduction of output would thicken the shell, in effect with the shell thickness adjusting itself to the reactor output.

This solves the problem for the existence of chemoautotrophic biological systems under anaerobic conditions. If we reckon 3×10^{10} erg gm^{-1} as the average chemical energy available for chemoautotrophy, the total for the whole outer Solar System is $\sim 10^{40}$ erg, about an order of magnitude less than the radioactive energy, but still a very large amount. It is here that the considerations of the previous chapter become relevant, that there should be no way to unlock this great store of energy except through biology.

Unlocking the store of chemical energy degrades the material in a thermodynamic sense, which consideration raises a further critical question: Could there be any means of achieving photosynthesis and so avoiding the progressive degeneration due to chemoautotrophy? Very readily, provided fibre optics existed to channel light through the cold surface shell to the reservoir of warm liquid below. A little thought shows that such a possibility is not as fantastic as it might appear at first sight. Since biology has produced eyes with their ability to function over an exceedingly large light intensity range, eyes with sophisticated chromatic and spherical aberration corrections included, and with such acuity of focus that a bird can distinguish small scraps of food from unwanted debris at distances of several hundred metres, fibre optics should not have been any great obstacle. From a physical point of view, such a requirement amounts to combining translucence with low heat conductivity, and also with a high opacity in the infrared, conditions that together would permit the surface shell of material to act as a powerful greenhouse, thus easing the load on the internal heat production considered above, and extending the estimate (25) for the length of time T over which biological activity could continue.

A greenhouse effect could reduce very greatly the thickness of the required outer shell, making the penetration of visible light much less of a problem.

Light penetrates typically about 10 metres through many translucent materials, for example, through water-ice. For a fall of 200K through a 10 metre thickness of material with heat conductivity 10^{-4} watt cm^{-1} K^{-1} the heat flux is 2×10^2 erg cm^{-2} s^{-1}, which equals the flux of sunlight at a heliocentric distance of ~80 AU. This is for sunlight at normal incidence. On a rotating body the generally oblique incidence of sunlight (and no sunlight at all on the dark side) reduces the average flux by 4, so decreasing the corresponding calculated heliocentric distance by 2, from ~80 AU to ~40 AU, i.e. to the outskirts of the present-day distribution of planets. Hence there seems no reason in principle why a vast biological ensemble should not have persisted on the outside of the Solar System over an extended period of several hundred million years, and why it should not have done so in an ongoing replicative state. There seems no reason also why life forms on the Earth, especially among invertebrates, should not have been derived directly from this former condition, assuming a feasible form of transportation being available from the outer regions of the Solar System to the Earth. Comets perturbed by stellar encounters into orbits with perihelion distances $q < 1$ AU are the obvious candidates for such a means of transportation. There appears to be no reason either why bubbles of gas should not become established as vacuoles within the objects, so permitting subaerial biological forms to arise. If the present-day complexities of life can arise by evolution in a biosphere of only $\sim 10^{18} - 10^{19}$ gm, the possibilities for a supersystem with mass $\sim 10^{29} - 10^{30}$ gm would almost surely be immense, especially as collisional interchanges which must have taken place from time-to-time among the many objects would have permitted evolutionary steps to be widely shared among them.

On this view, comets are relics of a former large-scale biological environment existing in the outer regions of the Solar System. The total mass of the relics, say 10^{11} comets of individual masses $\sim 10^{18}$ gm, again enormously exceeds the terrestrial biosphere. The total cometary storage of biomaterial could be as high as $\sim 10^{29}$ gm, and it would be surprising if this large quantity of material had not dominated conditions at the terrestrial surface throughout the history of the Earth. We tend to think the opposite simply because the total mass of the Earth, $\sim 6 \times 10^{27}$ gm, is much greater than the mass of an individual comet. But the total mass of the Earth is an irrelevancy here. It is the mass of the terrestrial biosphere that in the present discussion really counts, and the biosphere matches only a single comet out of the $\sim 10^{10}$ comets which must have passed through the inner regions of the Solar System during the history of the Earth. The weighting factor in favour of comets controlling the evolutionary situation is evidently enormous.

We end this chapter by asking what happens should the nuclear engine inside a comet finally give out? With the internal heat source gone, and yet with heat losses continuing at the surface, the comet must eventually become cold and frozen throughout its interior. If water is an appreciable constituent, a liquid interior inside a solid shell could not freeze without dramatic events occurring, simply because of the volume expansion that water undergoes on freezing. First, as the engine gave out convection currents stirring the liquid would cease, and all solid particles hitherto suspended in the liquid would fall gently towards the centre. Among the particles could be small silicate grains together with other refractories as well as micro-organisms. Hence an aggregate of particles would be deposited by sedimentation, as the carbonaceous meteorites have been formed by sedimentation, and with an admixture of micro-organisms as Hans Pflug finds to be present in these meteorites. The sedimentation would proceed higgledy-piggledy, just as the small particles happen to settle out of the now-unstirred liquid.

Freezing goes progressively from the outside inwards. Water immediately inside the outer shell cannot freeze without space being created for it to squeeze into. However, unlike water freezing downwards in a lake, which can simply lift the surface skin of ice bodily in order to create the needed space, water inside a closed frozen shell cannot lift the shell without cracking it into two halves. This requires a pressure of the order of the tensile strength of hard-frozen ice to develop throughout the liquid interior, $\sim 3 \times 10^7$ dyne cm^{-2}, a pressure which then acts compressively on the central concentration of small particle sediments, just as the carbonaceous meteorites were acted on compressively by a pressure of this order.

If freezing were a discrete one-step affair, an entire cracking of the outer shell into two halves might happen, but with the freezing process occurring continuously, a steady pressure $\sim 3 \times 10^7$ dyne cm^{-2} would be maintained against the inner surface of the shell, and within small cracks as they opened up, probably in many places throughout the shell. The lowest density components of the liquid would be squeezed up into the cracks, and likely enough would eventually emerge at the outer surface of the shell. In such a continuous multicracking process the needed extra space to provide for the expansion of the water would be found through geyser-like spurts of liquid, up through newly-opened cracks, with the liquid welling out and eventually freezing on top of the shell, the needed space being thus found on the outside of the comet. Since there are many organic liquids with densities less than water it would be these that would pour out of the freezing comet in preference to water, so explaining the observed situation discussed in Chapter 2, where we saw that comets typically have highly volatile organic materials deposited on their outer surfaces.

The final picture to emerge of a frozen comet is not of a single fused solid ball, but of an exceedingly complex multicracked affair, with the whole comet internally stressed at $\sim 3 \times 10^7$ dyne cm^{-2}. The situation is analogous to a mass of coiled springs, all ready to go off at a touch, which happens whenever evaporation due to sunlight weakens particular holding points in the structure. Or like a group of drunks leaning on each other – take one out and repercussions are felt throughout the whole party. On this picture a comet would not be exactly the most restful place one might visit. Comets which approach close to the Sun often lose fair-sized chunks of themselves, which separate apart quite gently at speeds of no more than one or two metres per second. If comets were homogeneous solids, the tensile strength of the material would have to be less than 10^5 dyne cm^{-2} to permit this phenomenon (Z. Sekanina in *Comets*, ed. L.L. Wilkening, University of Arizona Press, p.251). Since no well-frozen solid material has a tensile strength remotely as low as this, we can conclude that either a comet is a multicracked ensemble, with bits of itself only very lightly attached to other bits, or the interior material is still liquid.

With R = 5 km, (25) gives $T = 10^9$ years. This estimate for the time scale over which a cometary nuclear engine could maintain a liquid condition in the interior is so close to the ages of comets that we might reasonably argue both ways, with smaller comets having undergone freezing, and with larger ones still maintaining liquid interiors, and perhaps still maintaining something of their original biological activity. We are tempted to associate P/Schwassmann-Wachmann I with this condition. The sporadic outbreaks of this comet can then be understood in terms of an accumulation of biochemically-produced gas within the interior, pockets of which break from time-to-time to the surface, expelling visible clouds of gas and particles, perhaps in a similar fashion to the generation of dust storms on Mars.

CHAPTER 8

THE TERRESTRIAL CONNECTION

The picture of the interior structure of a comet discussed in the preceding chapter taken together with Figure 18, which shows the orbits of the 'Jupiter family' of comets projected on to the plane of the ecliptic, gives a clear perception of the genesis of the zodiacal cloud of particles. As stated already in Chapter 2 the zodiacal particles are derived by evaporation from these short-period comets. Once in space and exposed to sunlight the particles spiral slowly towards the Sun due to the so-called Poynting-Robertson effect, which acts both to reduce

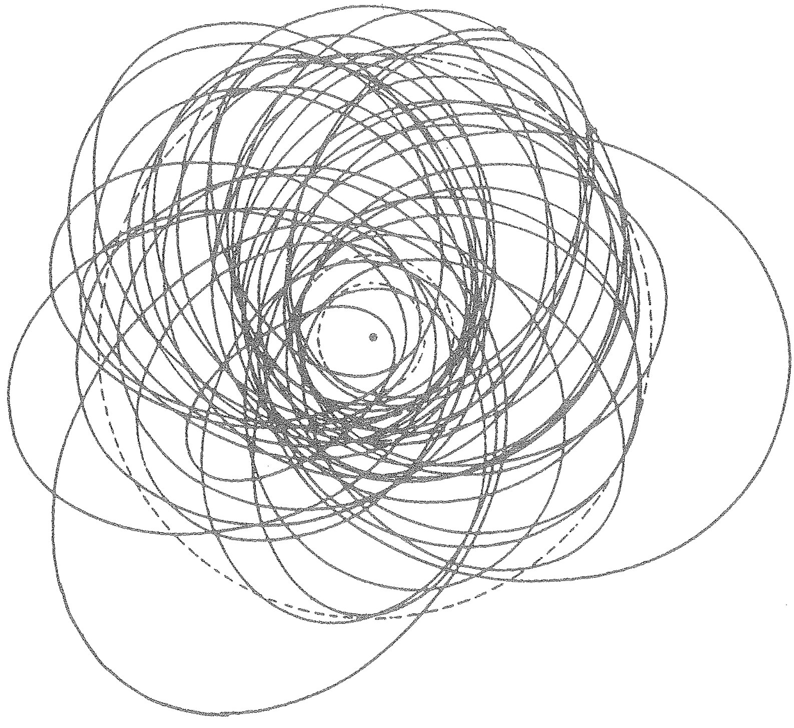


Fig. 18. The orbits of short-period comets lie mostly in the region between Mars and Jupiter. This is the graveyard of the comets. (From N.B. Richter, *The Nature of Comets*, Methuen, 1963).

the eccentricities and to reduce the semi-major axes of the orbits of the particles. The physical principle underlying the Poynting-Robertson effect can be understood from the case $e \approx 0$. With r the radius of the orbit, the speed of the particle is $(GM/r)^{1/2}$. The motion being transverse to the direction of the Sun there is an aberration effect which would cause an observer moving with the particle to see incident sunlight coming, not strictly from the radial direction, but at a small angle $(GM/rc^2)^{1/2}$ to the radial direction, where c is the speed of light. If for simplicity we consider the particle to absorb all the sunlight incident upon it, radiation pressure on the particle is $\pi a^2. L/4 \pi r^2 c$, where L is the solar luminosity and a is the radius of the particle. Still from the point of view of an observer moving with the particle the radiation pressure can be resolved into a main radial component and a small transverse component directed so as to oppose the motion of the particle, the latter being $(a^2 L/4r^2 c). (GM/c^2 r)^{1/2}$.

Radiation absorbed by the particle is re-radiated isotropically in the moving observer's frame of reference, and so contributes no force on the particle. Hence the observer finds the expression just given to be the transverse force, a result that can be transferred to the frame of a non-moving observer (only special theory is involved, the two frames *not* being instantaneously accelerated with respect to each other). Hence the non-moving observer sees the particle losing angular momentum mh at a rate given by multiplying the transverse force by r,

$$m \frac{dh}{dt} = - \frac{a^2 L}{4r^2 c} \cdot \left(\frac{GM}{c^2 r}\right)^{1/2} \cdot r. \tag{26}$$

Writing s for the density of the material, the mass m of the particle is $4 \pi a^3 s/3$, while the angular momentum h per unit mass is $(GM\ r)^{1/2}$. Inserting these expressions for m and h in (26) gives

$$\frac{dr^2}{dt} = \frac{-3}{4\pi} \frac{L}{asc^2} \tag{27}$$

from which we see that the radius of the orbit decreases significantly from an initial value r_0 in a time t given by,

$$t \simeq \frac{4\pi}{3} \frac{asc^2}{L} \cdot r_o^2 \simeq 7 \times 10^6\ asr_o^2 \text{ years} \tag{28}$$

with r_0 in AU, the particle radius a in cm, and s in gm cm^{-3}. For the particles of the zodiacal cloud with $a \simeq\simeq 1.5 \times 10^{-3}$ cm, $s \simeq 3$ gm cm^{-3}, the spiralling time from the Earth's distance $r_0 = 1$ AU is $\sim 3 \times 10^4$ years. Thus the zodiacal cloud in the vicinity of the Earth must be renewed by evaporation from short-period comets every ~ 30,000 years. Taking the local mass of the cloud to be 10^{17} gm, the requirement is for a renewal rate of $\sim 3 \times 10^{12}$ gm per year, the estimate used in Chapter 2.

Eventually all the volatile materials of the short-period comets will be evaporated, to leave only the inner sedimentary cores discussed in the previous chapter, the cores made up of small particles of all kinds compressed together at a pressure of $\sim 3 \times 10^7$ dyne cm^{-2}. Such non-evaporating residues have been plausibly identified with the so-called Apollo objects, of which about a thousand are estimated to be currently in orbits much like those of the short-period comets. No further evolution of these residues can arise except by collisions with other bodies, of which collisions with asteroidal bodies are the most likely. It will be recalled that the inclinations of the orbits of short-period comets are systematically low (and so would be the residues derived from them) which implies for orbits with aphelion distances of a few AU that considerable periods of time are spent in crossing the asteroidal belt, which has a range from about 2 to 3 AU. Collisions become more frequent as the pieces into which the residues are eventually fragmented become smaller and more numerous, so that in the end a fraction of the residues becomes a swarm of bodies with meteoritic sizes. The latter being small and being exposed to sunlight over long periods of time, often at perihelion distances less than 1 AU, micro-organisms within them tend to become cooked and fossilized.

Recently evaporated material from comets will not be in such a fossilized condition, however. When we remember the ability of micro-organisms to withstand the doses of hard radiation (the U-pond and Z-trench of the Hanford facility), when we note that even very small quantities of surrounding material provide an effective shield against solar ultraviolet light, when we also note that even less hardy species seem always to emerge from very heavy radiation doses with some fraction maintaining viability, it is apparent that the Earth is perpetually embedded in a halo of recently evaporated material in which micro-organisms can very well be potentially active, should they happen to enter the Earth's atmosphere as the Earth passes constantly through the halo in the course of its motion around the Sun. Add furthermore that viable specimens of the bacterium *Streptococcus mitis* were recovered from within a camera which had been left for two years on the lunar surface. The bacteria had therefore been exposed to very low pressure and to violent oscillations of temperature over an extended period. Micro-organisms are evidently very space-hardy.

It is known that every year the Earth acquires about 1000 tons of finely divided cometary material, which would amount to a lot of micro-organisms – if indeed the material consists of micro-organisms. Because the Earth and the acquired material are not precisely co-orbital the material necessarily moves very rapidly by ordinary standards with respect to the Earth. If the material were to impact a hard surface, for instance the surface of the Moon, it would instantly be destroyed as bugs are splattered on the windscreen of a fast-moving car. For the Earth, however, the diffuse gas of the high atmosphere provides a soft landing for particles as small as micro-organisms (but not for larger particles which become seriously heated, or even evaporated, as in the case of so-called shooting stars). The maximum size for the safe entry of an individual particle in the most favourable case is about one-tenth of a millimetre - bacteria have sizes typically a hundred times smaller than this, while viruses are more minuscule still. Thus the high atmosphere provides a soft landing for particles of the sizes of viruses and bacteria, a landing so soft that they would retain biological viability, subsequently falling gently down through the atmosphere as active agents.

There are many ways of testing this picture further, of which the most dramatic but expensive would be to visit a comet by space vehicle, to excavate a chunk of material and then to bring it back to Earth for chemical and biological analysis. But this experiment has of course already been done for us by the meteorites, which come to us at no cost at all. And an equally dramatic, far less expensive, much more extensive and subtle method, lies to hand in the medical sciences. This is because a fraction of incoming active agents could be pathogenic, with the ability to multiply themselves many billions of times in the cells of their victims, and so becoming visible even to the casual eye. Here we consider a few out of many such examples.

Whooping cough is an interesting case. It has for long been known that whooping cough occurs in cycles of about 3.5 years, which used to be explained on what was known as the 'density of susceptibles' theory. The idea was that, after children susceptible to the disease become exhausted by a particular epidemic, it then takes about three and a half years for new births to rebuild the density of susceptibles to the level at which a further epidemic will run. Thus the periodicity on this theory should have been a function of population density, with the shortest periods being found in inner city areas of very high density and either long periods or no periodicity at all in lightly populated country areas. But the periodicity was found to be everywhere the same, in town and country alike, and from one country to another. The facts were so obviously in sharp contradiction with the theory that in a rational society it would immediately have been forgotten and good riddance. Instead it has been propagated in the educational system from deluded teacher to credulous student over nearly a century. Figure 19 shows the record of whooping cough notifications for the period 1940-82. If this theory had been correct, the sudden

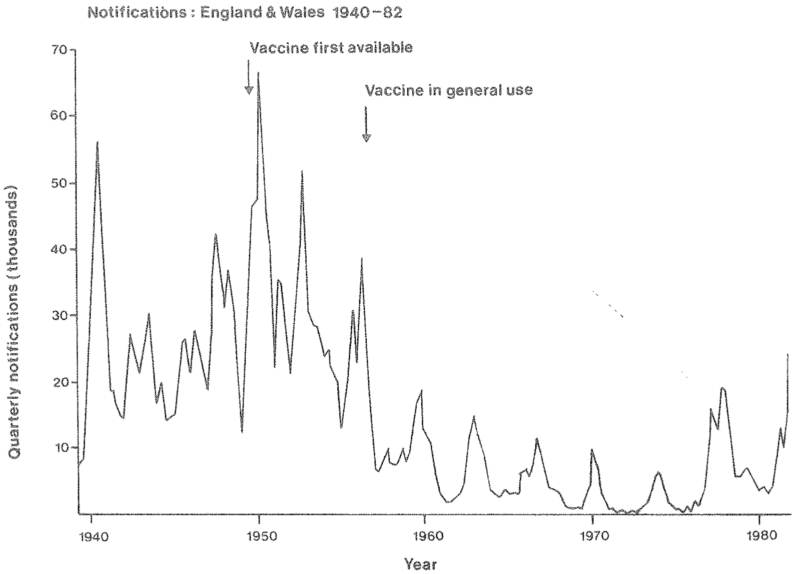


Fig. 19. Whooping cough notifications in England and Wales from 1940–82.

reduction in the density of susceptible children brought about in the 1950s by the introduction of an effective vaccine should have greatly disturbed the periodicity, or even destroyed it altogether. Yet the periodicity persisted exactly as before, but with the total number of cases greatly reduced. Although these further facts constitute as direct a disproof as one could wish for, the theory is still being canvassed. The data of Figure 19 is clear evidence of an externally forced periodicity, not an internally-generated effect at all, as would arise if the pertussis bacterium is incident on the Earth from outside at regular intervals. In such a situation scattering would be introduced in the occurrence of cases by the circumstance that particles with the sizes of such bacteria take about two years to descend the stratosphere under gravity. However, while such a blurring effect would tend to destroy strict regularity from one cycle to the next it could not change the long-term average of the periodicity. But in a moderately short run such as Figure 19 there could be uncertainties of ~ 1 year at each end of the record, so producing a modest error in the estimated mean period, which for the 12 cycles from the peak of 1942 to the peak of 1982 is ~ 3.3 years, essentially the same as the period of Comet Encke. The uncertainty shows, however, if the first ten years of the record are omitted. There are then 9 cycles from the peak of 1951 to the peak of 1982, for a mean period close to 3.5 years rather than 3.3 years.

A more far-reaching difficulty for making a precise identification with a particular comet comes from Figure 20, which shows that debris evaporated from Comet Encke does not move in exactly the same orbit as the comet itself – there is a spread around the comet orbit which produces a corresponding small spread in periodicity. Since the Earth can acquire debris from a range of such orbits, the times of acquisition of material cannot be rigorously periodic. Nevertheless, the general relation of Figure 18 to the period of P/Encke is interesting.

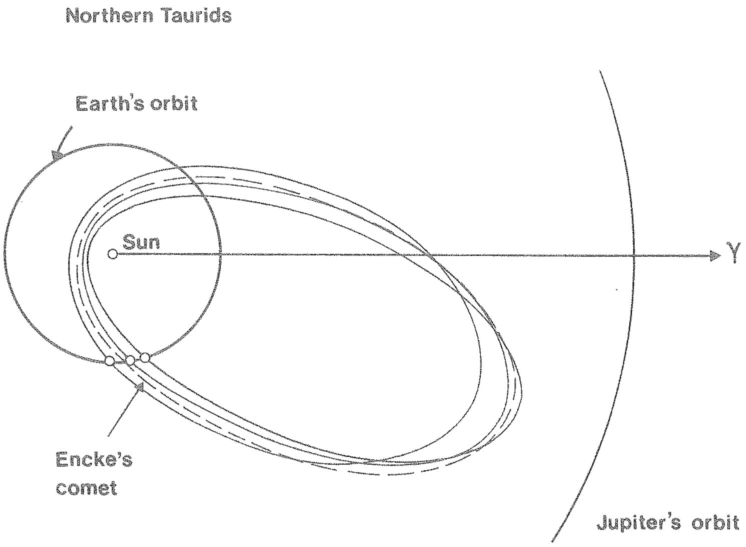


Fig. 20. The spread of orbits of the debris evaporated from Comet Encke.

There are a number of viral diseases which are specific to man, measles for example. Virologists are generally of the opinion that such specificity implies a long-term connection between our species and the viruses in question. Yet unless the viruses come from outside the Earth, there can have been no such connection, because such viruses as measles could not have maintained themselves at the low population densities of prehistoric times. One would need to suppose that all such viruses are of comparatively recent origin, a view that would be hard to support even if there was not direct evidence against it.

The heavily-forested regions with fast-flowing rivers in South America have long had a reputation among explorers for being one of the most difficult areas of the Earth to penetrate, especially as tribes in those regions were fiercely hostile to outsiders, and of course to each other. Such tribes scarcely numbered 500 persons and so were well below the level at which many viruses could be sustained indigenously. Yet when forest clearances in Surinam exposed the situation to outside investigation, by removing their hitherto almost impenetrable tree-cover, examples of deformities in the population due to poliomyelitis were found. The inference from this fact is that the polio virus found its way into the tribes by falling from the air.

There is a straightforward way to separate a disease caused by a pathogenic agent falling down through the atmosphere from one transmitted horizontally from person-to-person. The chance in the first case of contracting the disease should be much the same in town and country, whereas horizontal person-to-person transmission should run more strongly in heavily populated city areas. Figure 21 shows data collected by Dr. P. Jenkins, the Community Health

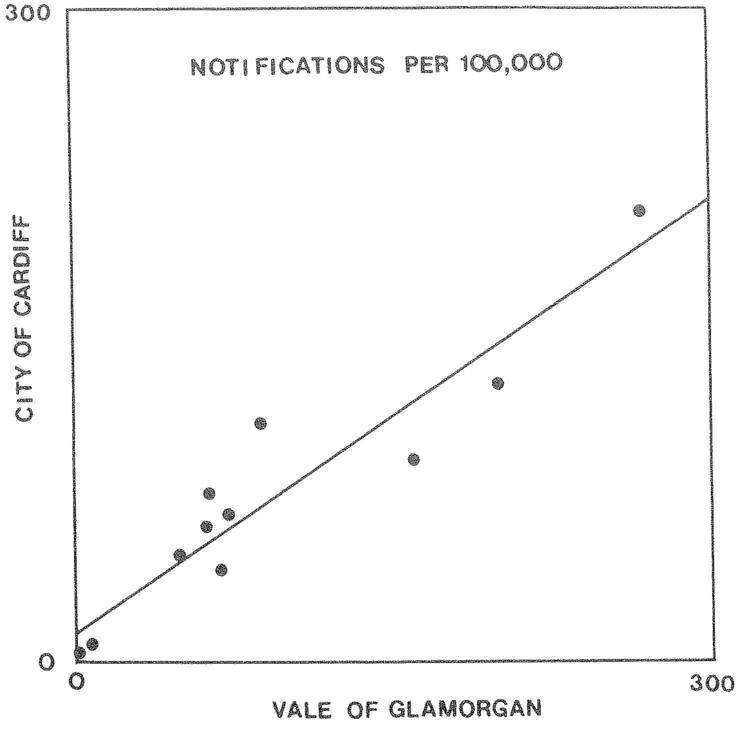


Fig. 21. Incidence of whooping cough, measles and infective jaundice in Cardiff city *vs* the Vale of Glamorgan.

Officer for the City of Cardiff, for so-called infectious jaundice, whooping cough and measles. Data from the heavily populated Cardiff city area is plotted against data from the Vale of Glamorgan, much of which is very rural. The bias, such as it is, goes just the wrong way for horizontal person-to-person transmission, for it is the lightly populated Vale of Glamorgan that on a normalized basis was the worse affected.

If the pathogens were falling from the air, however, local atmospheric turbulence due to the heat developed in the city area could very well have produced a measure of protective shielding, so explaining the apparently wrong-way-round bias of Figure 21. Some years ago we came to the conclusion that Influenza A does not spread much by horizontal transmission. If it did, worldwide pandemics would spread in these days of rapid travel far more quickly than they did in the 19th century. But data for the two major outbreaks of recent years, so-called Asian 'flu in 1957 and Hong Kong 'flu in 1967-68, shows just the same spreading speed reported in the 19th century (the spreading speed in 1967 in the USA was 6 weeks from California, where the first case occurred, to New York, and 12 weeks to Florida).

Because we became suspicious of arguments given in the literature which claimed to demonstrate the horizontal spread of influenza, we took the opportunity to investigate the 1977/78 pandemic in which the antigenic form of the influenza virus shifted remarkably from the H3N2 type to an old H1N1 type. Because pupils in schools uniformly had no established immunity to the latter, we chose to investigate epidemic outbreaks in a large number of fee-paying schools in England and Wales.

Much of what is often said turned out to be grossly wrong. It is widely stated for example in the literature that the horizontal transmission of influenza is proved by the very high attack rates which occur in institutions such as boarding schools. Figure 22 shows what happened in the approximate half of our sample where attacks had been appreciable - the other half did not experience noticeable outbreaks at all. The open circles in Figure 22 refer to attack rates in the range 0-25 per cent, the triangles to 25-50 per cent, the solid dots to 50-75 per cent, and the three cases with stars to 75-100 per cent. The supposed evidence in the literature had been artificially manufactured by reporting the latter three cases and leaving out the rest, a thoroughly dubious way of presenting the facts, but one quite typical of the methodology by which the educational system maintains its rapacious demands for taxpayers money.

Figure 23 shows the situation described by Dr. John Briscoe, the Medical Officer at Eton College. With several school houses having 4 standard deviations from the mean attack rate of 35 per cent for the school as a whole (1248 pupils), and one house with 6 standard deviations, it is certain from this one example alone that it is the *places* where people are located that decides the

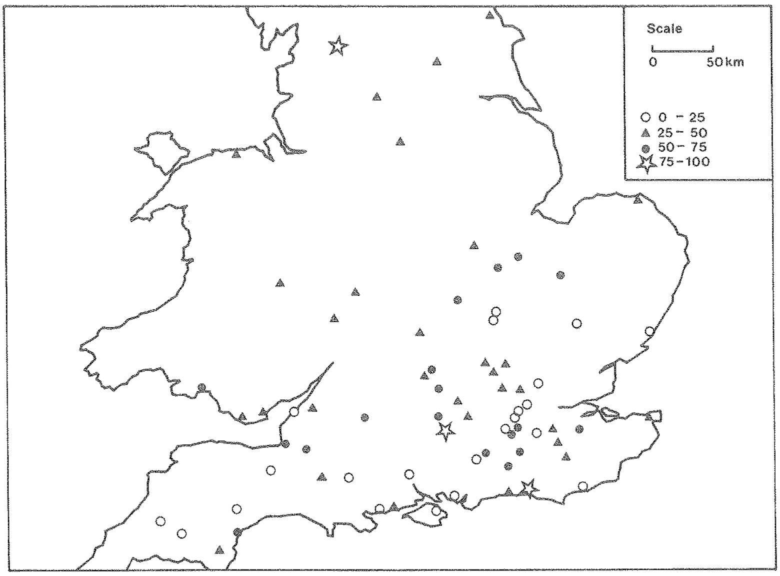


Fig. 22. Distribution of day-pupil attack-rates during the 1978 influenza epidemic in England and Wales.

chance of contacting influenza, rather than contacts with other people - pupils from the various houses at Eton had contact with each other in school classes, games, meals, etc., so that there was ample opportunity for pupils in College House (COLL) to have contracted the illness if it had gone horizontally, which evidently it did not.

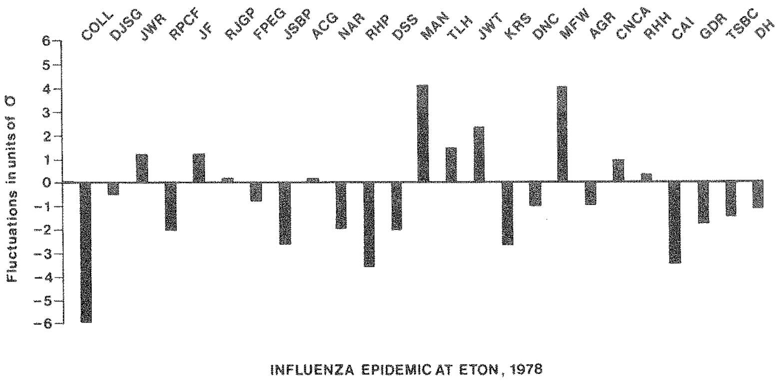


Fig. 23. Fluctuations from expected mean numbers of victims in several houses at Eton College. Fluctuations are expressed in terms of the sample deviation σ.

So let us think for a while how viral particles incident from space on the Earth's upper atmosphere might find their way into the human respiratory tract. While a bacterium is large enough to descend the stratosphere by itself under gravity, a viral particle is too small to do so. A viral particle must descend through a large scale downdraft of air, which happens most effectively over the winter months from December to March, which has been demonstrated by injecting a radioactive tracer in considerable quantity at the top of the stratosphere and by observing the time at which it appeared at an altitude ~ 20 km, which is to say towards the bottom of the stratosphere. Another way to put this same issue is that the heat engine in the high atmosphere works most effectively at the time of the year when the temperature difference between equator and pole is largest, from December to March for the northern hemisphere.

Figure 24 shows the almost clockwork-like regularity of attack of respiratory syncytial (RS) virus, a typical winter disease, coinciding with the atmospheric effects for the northern hemisphere.

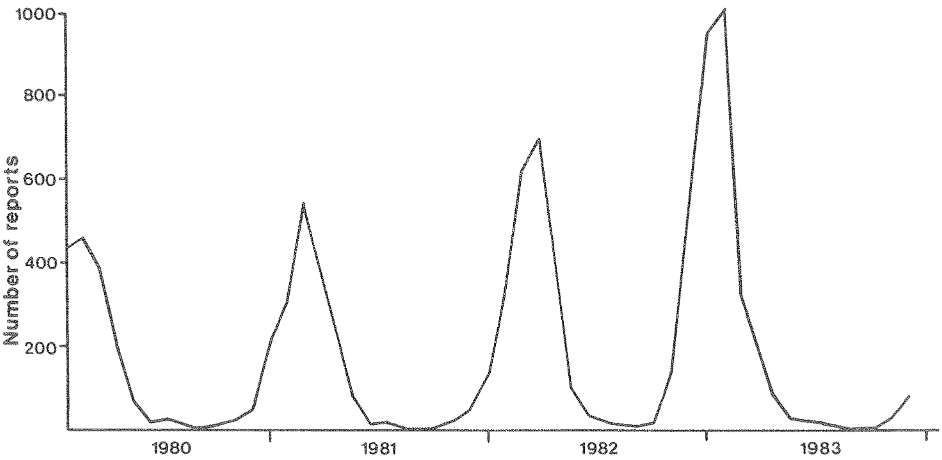


Fig. 24. Reports of the RS virus.

Figure 25 shows the situation for Influenza A, with monthly averages for a northern hemisphere site (Sweden), a tropical site (Sri Lanka) and a southern hemisphere site (Australia). To interpret Figure 25, the northern and southern temperate latitudes must alternate atmospherically, with December to March as being the months of incidence at ground level in the north and June to September the corresponding months in the south. Figure 25 shows just such an effect for the incidence of Influenza A, with a much weaker situation without discernable annual effect in the tropics.

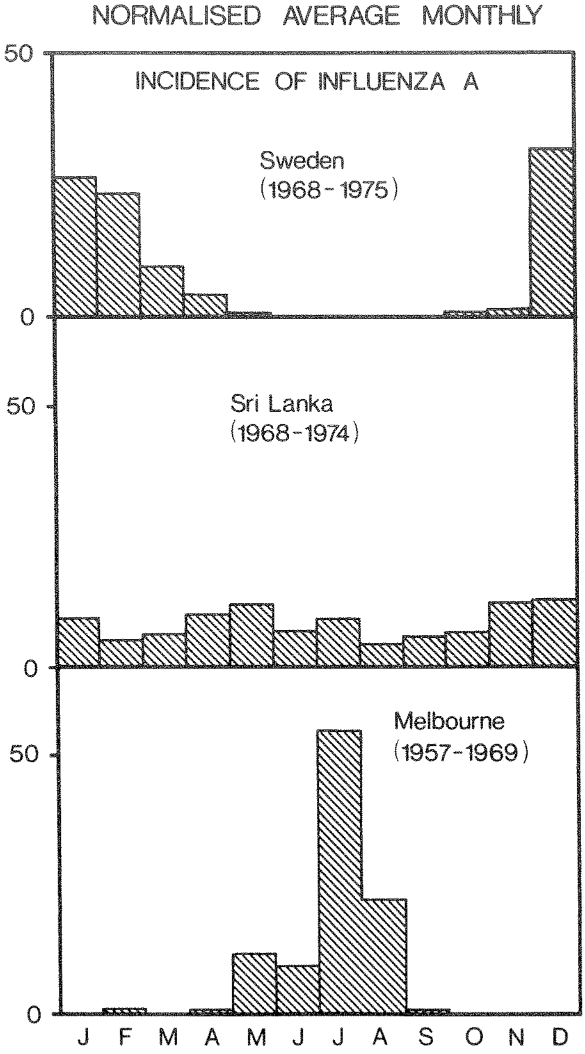


Fig. 25. Incidence of Influenza A in 3 geographically separated countries.

We have chosen to show the situation for Sweden rather than Britain, not because there is any important difference between Sweden and Britain, but to bring out the point that the simple physical cold of winter is not a relevant factor. Sweden has a really cold winter, whereas Australia has a clement winter that is not much cooler than a Swedish summer. If simple exposure to cold were important, the effect would long ago have been demonstrated under controlled conditions in the laboratory, which it has not been.

To understand how people actually catch these viral diseases consider next how, having descended the stratosphere, the viral particles come down through the lower atmosphere, the so-called troposphere. This lower descent happens much quicker than the stratospheric descent, because small particles fall through the lower atmosphere inside comparatively large water droplets. So it is inside raindrops that prospective pathogens descend finally to ground-level.

It is a matter of experience that we do not snuffle-up raindrops in ordinary rain. Rain which impacts the face tends to drip off the end of the nose instead of entering the respiratory tract. Hence viral particles in a shower of rain would pass us safely by on their way to the ground. But rain does not end because all the water has fallen out of the atmosphere. Rain ends because falling droplets evaporate before reaching the ground. Droplets evaporating immediately in front of one's face, releasing viral particles into the air would not be harmless, because the released particles could then be breathed into the respiratory tract. It is therefore the end of a shower of rain that is dangerous. You will rightly observe that such a condition must be a little unusual. Unusual and therefore highly irregular, and confined to small patches near ground-level, which is just what the influenza data for school houses shows to be the case.

An illusion was put about widely a few years ago that it would be impossible for viruses from space to mount pathogenic attacks on terrestrial plants and animals with the specificity that is actually observed, for example with measles being a specifically human disease. But this was merely an expression of opinion. Worse, it was untrue. Human viruses attack tissue-cell cultures of other primates, and most human viruses can be cultured even in chick embryo, a taxonomic class apart from humans.

The specificity does not come therefore from individual cells or from the viruses themselves, but from our immunity systems. This should really be no surprise because our immunity systems can be specific even to the extent of rejecting tissue from close relatives of our own species.

It appears likely that our very genetic make-up has an origin external to the Earth. As well as causing pathogenic attacks, viruses can simply add themselves to our chromosomes, placidly multiplying only as our cells divide. In this way we derive new genes as a matter of fact, not conjecture. Since it needs only an elementary mathematical calculation to show that genes are so astonishingly complex that they could never be produced by random internal shufflings of bases on our DNA, it is then but a short step to the realization that all of our genes are of external origin, added by viruses or by the even smaller fragments known as viroids.

Complete immunity to viral invasion can therefore be seen to be an impossibility for an evolving biological system. Our immunity systems have the task of admitting to the body any viruses which have a promise of utility. But then our immunity systems have the further task of defending the body when the experiment does not succeeed. Of course we suffer many defeats, many deaths due to disease, so that from an individual perspective the situation has unpleasant aspects to it. But the logical essence of biological evolution is that, because numbers are large,the biological ensemble as a whole can afford many failures in the course of finding one success.

The correct picture of biological evolution is that terrestrial plants and animals are aggregates of genetic components on which natural selection acts, determining which aggregates are viable in the terrestrial environment and which are not. By itself, natural selection can do nothing, however. One could select a lorry load of potatoes as much as one pleased – for size, shape, smoothness, and so on – but it will never become a lorry load of tomatoes. The crux of evolution is not natural selection, as 19th century biologists hoodwinked the world into believing, but the source of the genetic material upon which natural selection operates.

As soon as we say that all plants and animals are aggregates of the same cosmically derived gene pool, it follows that the same or similar genes can be present at widely different taxonomic levels. Approximately parallel lines of evolution then present no problem. One can readily understand otherwise extremely mysterious correspondences, as for instance why the colours of flowers and of insects often match each other quite perfectly, because the colours are produced by the same genes. Complex eyes have evolved three times during biological history – in the octopus and its relatives, in insects, and in fishes, reptiles and mammals. The three eyes do *not* come from a common ancestral eye but have appeared independently. Yet they operate in basically the same way – on the change of shape of the same retinol molecule.

So one can go on. Chemical substances extracted from plants often have an intimate relation to chemical processes in animals, as for instance in the drugs morphine, quinine and cocaine. Penicillin was originally extracted from a fungus. Close correspondences like these are inexplicable in terms of the conventional outlook in which the genes of such widely-separated species are required to have evolved independently of each other. The similarities are readily understood, however, by the same externally derived components being present in both plants and in ourselves.

The fossil evidence indicates quite strongly that the major steps in evolution occurred in rather the way research scientists use computers in tackling difficult problems. As new issues and difficulties crop up, bits and pieces are added to the program, which becomes more and more cluttered with minor added sub-routines and loops. Then a day comes when the human scientist has had enough of bits and pieces and patches, and decides to tidy up the program as a whole. If this is not what happened in biology it is difficult to understand the complete absence of *all* important connections in the fossil record.

Figure 26 shows what is known from the fossil record of the evolution of the various mammalian orders. Although the artist has tried to prejudice us by dottings and curvings into believing that somewhere back in the past the various orders were connected into a common stem, no connections have actually been found. Darwinists have always blamed the incompleteness of the fossil record, and they continue to do so. This seems more than a little dangerous. The intervals before present involved here are trivial from the point of view of modern geology. There are plenty of rock sequences from these comparatively recent geological periods exposed all over the world, and although particular sequences may be broken and incomplete it is hard to see why every rock sequence should be broken in just the same places.

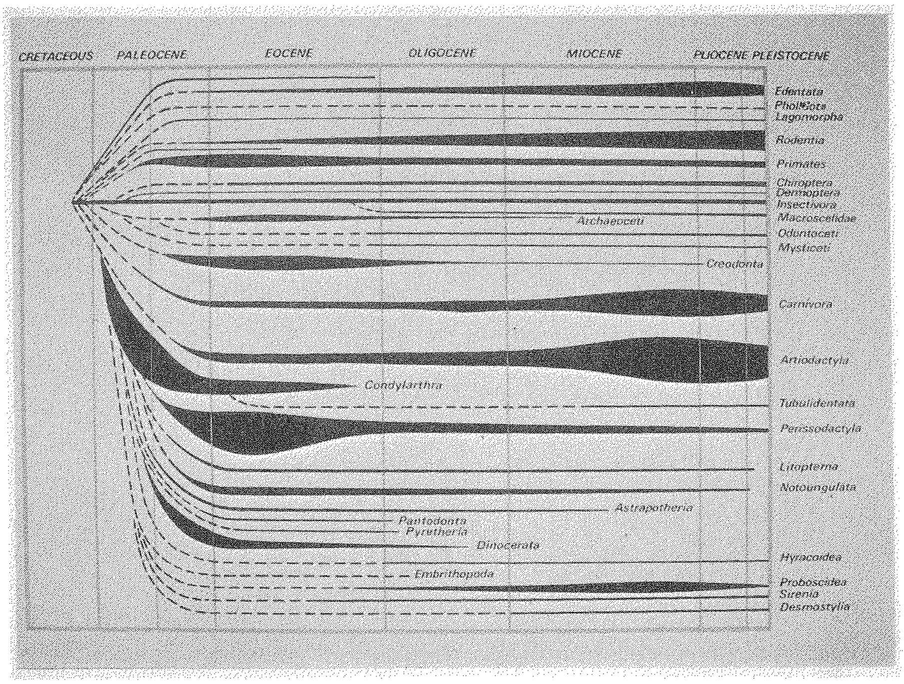


Fig. 26. Evolution of various mammalian orders.

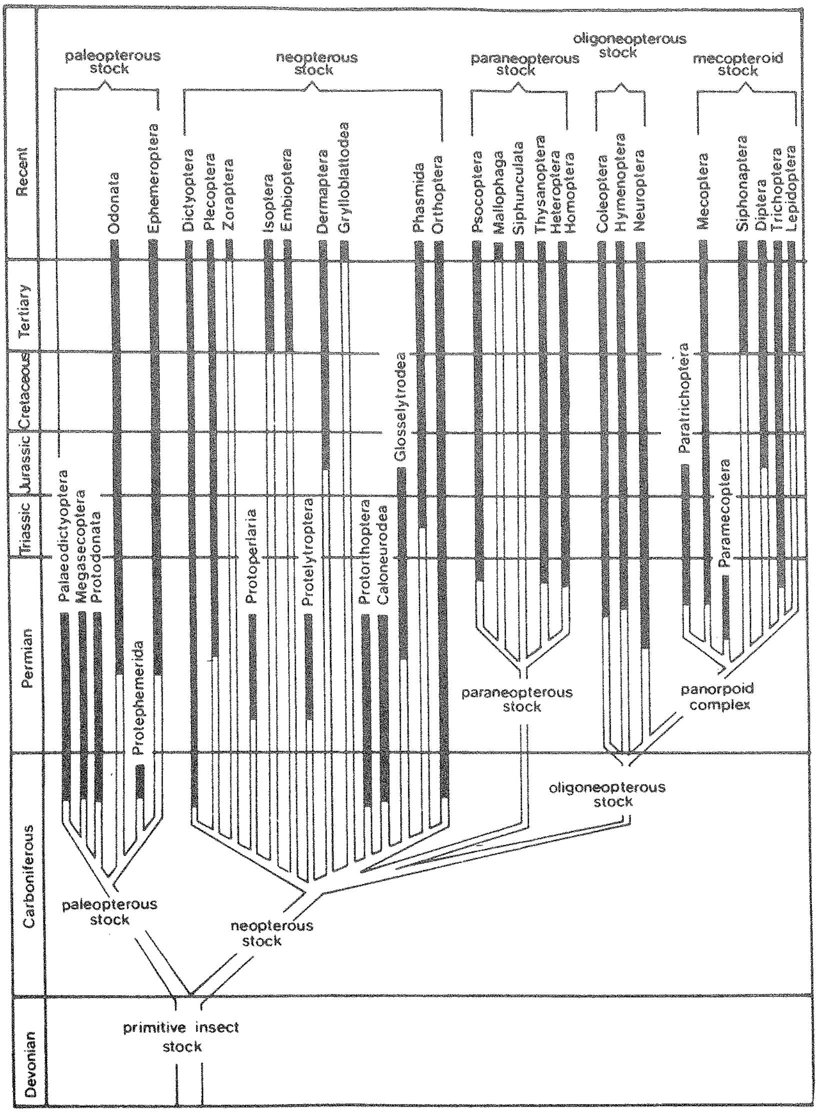


Fig. 27. Conjectured evolution of various orders of insects.

The unfilled bands of Figure 27 show the conjectured evolution of the various orders of insects, but once again the actual fossil evidence shown by the black lines have no connections at all to the supposed common stem. Figure 28 shows the situation at the higher taxonomic level of divisions and classes. Once again the artist by dottings and curvings had done his best to prejudice our minds, but once again there are no actual connections. Like so much else when

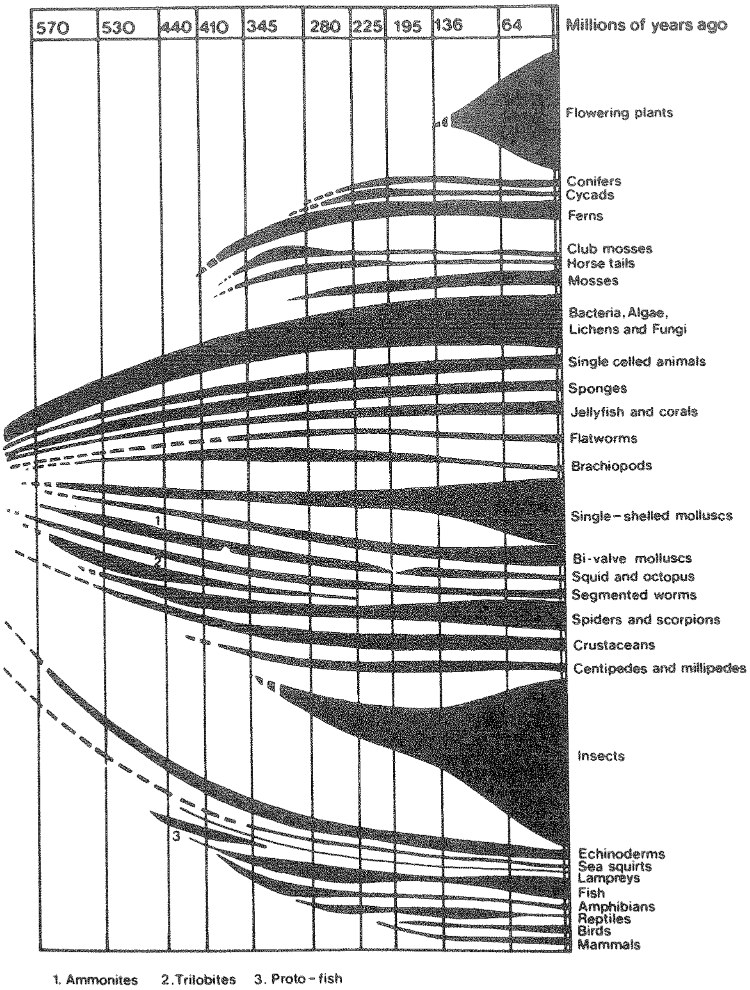
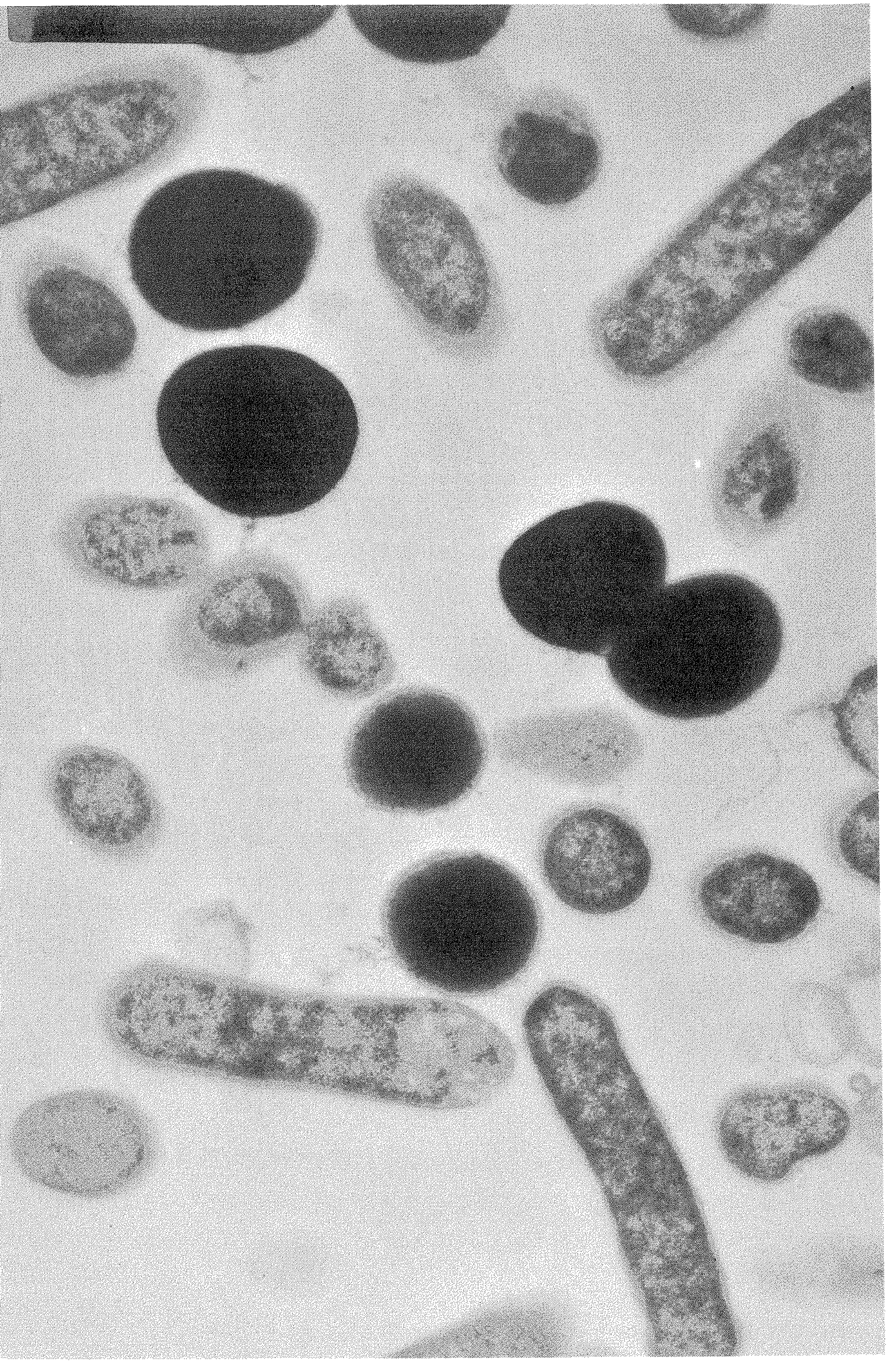


Fig. 28. Conjectured evolutionary connections between various taxonomic divisions and classes.

you look at the facts, this is not at all the way it should be according to Darwinists. It is not just that *some* of the important connections are missing. Essentially all the really important connections are missing. Nor is the issue one needing subtle debate, such as those who talk about punctuated equilibria would have us believe. The situation is gross and obvious, obvious to a point where it shouts aloud that the situation is not the way it is popularly supposed to be. Either we are looking at largely parallel lines of evolution, which is a possibility according to the point of view we have described, or evolution proceeds through major crises, major crises that are rapid and which we might describe as the successful attacks of disease. The facts permit nothing else.



CHAPTER 9

THE SOLAR SYSTEM CONNECTION

Figures 26, 27 and 28 raise interesting problems, quite apart from the difficulty which they set for 19th century theories of biological evolution. Whatever Darwinists may care to think, 19th century theory is not correct, and any normal person who has escaped unhinged from the educational system should be able to see at a glance from these figures that this is so.

Furthermore there are frequent cases, especially among invertebrates, of creatures which after beginning abruptly in the fossil record subsequently show little or no subsequent evolution. There are such cases among beetles, and among mayflies and dragonflies at the left of Figure 27. Among crustacea, unchanged fossil shrimps have been found in Sweden back to the Cambrian, near to the beginning of this class of creatures (which is characterized by encasement in chitinous material, as one finds in the shells of lobsters). Because of the readily preservable siliceous material in the membranes of diatoms, and because of the existence of diatoms in huge numbers, this particular phylum can be considered something of a test case. Diatoms appeared suddenly in the lower Cretaceous. From the first moment of their entry in the fossil record diatoms have shown all their marvellous intricacies of geometrical design, and they have remained essentially unchanged ever since. A similar situation has occurred in many other cases.

There are two ways in which one can attempt to understand the absences of connections in Figures 26, 27 and 28:

(1) Major evolutionary changes came in quantum jumps which happened so swiftly that they could not be captured in the slower-moving fossil record.

(2) While there has been an evolutionary tree, in the sense of the conjectural markings in these figures, evolution for the most part has not occurred on the Earth. Evolution on the Earth has been confined to taxonomic levels below that of 'order', which is to say to the outermost branches of the supposed evolutionary tree, more or less as actual trees show limited yearly changes only at their outermost branches.

To provide a satisfactory logical basis for quantum jumps it is essential to acquire new genetic possibilities in advance of the jumps themselves. A jump cannot occur unless the potential for the jump exists before it is actually made. In such a situation, the genetic potential cannot have experienced natural selection due to advantageous properties which a species acquires as a consequence of a jump, so that in a Darwinian sense the jumps are uncaused. Darwininsts have chosen persistently to ignore the evidence, preferring to cling to their prejudices than bow to the facts, thereby demonstrating ignorance of the first premise of all science. Others have attempted to argue that the genetic potential for quantum jumps is accumulated by random driftings of so-called neutral genes. If it were not for the immense complexities of gene structure revealed by modern microbiology, this might conceivably be a tenable position, but such are the complexities that acquiring significant gene structure by chance, and doing so repeatedly for a sequence of jumps, is entirely ruled out by probabilities that can be as small as 10^{-100} for a single gene alone.

To make quantum jumps possible, biological systems must acquire functioning genes from outside themselves, as for instance happens when genes are transferred artificially in the laboratory from one species to another. While it is possible to look on all biology, plants and animals alike, as an ensemble of genes that are shuffled around (though the agency of viruses or so-called plasmids) among its members in order to arrive at the most advantageous gene aggregrates – advantageous with respect to competition and survival – an attempt to make the ensemble self-contained runs into obvious difficulties, as for instance the origin of the genes in the first place.

Our solution to the problem of quantum jumps has already been given in the previous chapter, namely the incidence on to the Earth of functioning genes from cometary sources, whose contents reflect the earlier biological activity in the outer regions of the Solar System that was discussed in Chapter 7. Such an acquisition can happen through incidence on to the Earth of micro-organisms with sizes less than $\sim 1\ \mu\text{m}$, which can readily secure a soft landing in the high terrestrial atmosphere. When we turn to possibility (2) the situation in this respect is more difficult, however, since the embryos that would be required for already-established evolutionary lines to be acquired by the Earth are considerably larger than $\sim 1\ \mu\text{m}$, and so would not be expected to be incident here with a soft landing, except possibly in highly unusual conditions that require further discussion. Nevertheless, the evidence of sudden beginnings without appreciable subsequent evolution occurring is so strong for so many cases that possibility (2) should be seriously considered. In a former book (*Evolution from Space*, J.M. Dent, 1981, p.117) we discussed further evidence supporting possibility (2) for the particular case of insects.

The latter discussion was greeted with derision by many biologists, we think essentially because the educational process in biology consistently blurs the distinction between fact and belief. One friendly zoologist said to us in respect of Figures 26, 27 and 28, that students in biology are compelled to learn the hypothetical connections so thoroughly during their training that almost inevitably they come to perceive of the supposed connections as real. Our friend added that it came as a shock, even to him, to realize how little fact there really was in these figures, and how overwhelmingly the facts have been overridden by conjecture. Unless one has a clear understanding of exactly where fact ends and conjecture begins, it is well-nigh hopeless to attempt to think outside our educationally-imposed mental prisonhouse, in which those who instantly dismissed the concept of 'insects from space' as absurd were evidently inured without possibility of release.

Hypotheses cannot be dismissed as 'obviously absurd' on the grounds of belief, because in all ages hypotheses that were thought absurd have subsequently been found to be correct. Any claim that the present moment will turn out differently in this respect is likely to prove truly absurd. Even on purely *a priori* grounds we can see nothing unreasonable in the view that $10^{29} - 10^{30}$ gm of carbonaceous material in the outer regions of the Solar System gave rise to a major outburst of biological replication and evolution, which has been followed much later in time by the laying of certain of these evolutionary lines across the Earth.

Meteorites of appreciable size such as the Murchison meteorite reach ground-level with interiors that are comparatively cool, so that eggs and sperms of appreciable size could in principle reach the Earth within bodies of meteoritic size. Better still from the point of view of viability would be a larger body that fragmented into many smaller pieces on entering the terrestrial atmosphere. In this way one can conceive of embryos of considerable size reaching the Earth from comets, although the process could not occur on a day-to-day basis, but would have to relate to exceptional conditions, as for instance in epochs when the subset of comets with perihelion distances less than ~ 1 AU was unusually numerous.

The present-day subset of comets with $q < \sim 1$ AU has been derived, according to Chapter 4, from stars that entered the cometary cloud at heliocentric distances of $\sim 10^{18}$ cm, which happens at intervals of $\sim 10^{6}$ years. Such encounters with passing stars are sufficiently frequent to be considered the normal situation – we saw that about ten of them contributed to the present cometary subset. We also saw in Chapters 4 and 5 that the cometary cloud may well contain an inner region at heliocentric distances $< \sim 10^{17}$ cm that is comparatively disk-like, and in which the density of comets is far higher than in the outer more spherical regions of the Oort Cloud. Penetration of such an inner region by a passing star would evidently be a much rarer event. It would, however, be capable of generating a far greater crop of comets with $q < \sim 1$

AU, i.e. of comets whose evaporated products could reach the Earth. Closer stellar encounters of this kind would be expected to occur at intervals $>10^8$ years.

Several important irregular episodes which have occurred in the Earth's history can be attributed to stars which happened by chance to penetrate the cometary cloud in its denser inner regions at heliocentric distances $\sim 10^{17}$ cm : Episodes in which independent biological lines reached the Earth. Periods of heavy cratering of the Earth's surface. Periods in which exceptional numbers of species became extinct.

We believe a publication of ours contained the first suggestion in recent years that the major extinctions of the late-Cretaceous might have been associated with comets (Preprint: *Comets, Ice-Ages, and Ecological Catastrophes*, late 1977; and in *Astrophys. Space Sci.*, *53*, 1978, 523). Whereas we attributed the ecological disaster of the late-Cretaceous to a blocking of sunlight due to particles of cometary origin in the high terrestrial atmosphere, later publications which received a great deal of notice (particularly L.W. Alvarez, W. Alvarez, F. Asarco and H.V. Michel, *Science, 208*, 1980, 1095) attributed the extinctions to the violence of actual impact of a large bolide of cometary origin. In the past year or two there has been a trend away from the latter suggestion, since it is hard to believe that direct impacts, however locally disastrous, could produce a vast number of extinctions over the whole Earth. The current trend is to a blocking of sunlight, our original proposal. To the blocking of sunlight we would now add the incidence of an intense rain of biomaterial capable of causing diseases along the lines discussed in the preceding chapter. The rain of potentially pathogenic material would have continued for a time scale 10^6 years, so that the resulting extinctions would have occurred over an extended period rather than happening more or less instantaneously as would be the case in the bolide impact theory.

It is a matter of convention in science that one is not supposed to discuss the ethics of referencing, a convention which opens the door to rogues who become specialists at misreferencing, with many sly tricks to their discredit. The convention has also, in the present woefully degenerate condition of science, given rise to another convention which judges it to be right and proper to omit priorities deliberately in cases where the work of a scientist is judged to be generally heretical, even though the omitted priorities concern quite other matters. The attitude is that an heretical person must be given no credit for what he has admittedly done correctly, otherwise 'people' might even begin to take notice of his heresies. The weekly science magazine *Nature* is at the forefront of the application of this convention, even carrying matters as far as in-house attempts at the removal of references to authors of whom it disapproves, while continuing to print hypocritical editorials directed against the trend towards widespread plagiarism. The reader will look in vain for any reference to *Comets, Ice-Ages and Ecological Catastrophes* in the many letters and articles which *Nature* has published post 1978 on this subject.

If 19th century biological evolution is to have any hope of survival it must be largely in the environment discussed in Chapter 7, an environment which existed in the early days of the Solar System in its outer regions. The Darwinian theory is currently so hard-pressed that its adherents would quite likely be prepared to settle for such a transposition, provided we continue to speak about Darwinism applied to the Solar System as a whole. This is playing with words in order to preserve what has become a sociologically manic situation. Darwinism has always been presented as a terrestrial theory and there is a vast number of examples to prove it. Moreover, there is strong evidence to show that the outer part of the Solar System was not the beginning of biology. The trail leads much wider than that, to the galaxy at large, as we shall show in the next and final chapter and quite likely even beyond the galaxy to a universal cosmological scale.

There are some who are worried that widening the scope of biology has the effect of placing fundamental problems further off, both in space and time. So what is the point of it they ask? The point is that the facts point this way, and unless one follows the correct path fundamental problems will never be solved. Just the same plaintive cry could have been raised in the 17th century when it was first said that the stars are bodies like the Sun, and in our own century when it was said that spiral nebulae are bodies like the galaxy. The history of science shows all really fundamental issues do widen enormously in their scope whenever we shed our blinkers and permit ourselves to be led by the facts, painful though the process may be.

Another possibility for producing a great increase in the rate of addition of interplanetary material to the Earth's atmosphere would be the arrival among the short-period comets of a body of exceptional size and mass, say $> 10^{20}$ gm. From the discussions of Chapters 5 and 6 there seems no reason why such bodies should not be present in the cometary cloud, indeed bodies up to at least a lunar scale. Their comparative small number would explain why such larger bodies are not normally seen. It appears unlikely that the most massive comets to arrive at $q < 1$ AU over a period as long as $\sim 10^8$ years have been no greater than the masses which have arrived in only the last century or two.

CHAPTER 10

THE GALAXY CONNECTION

1. *THE EXTINCTION OF STARLIGHT AT VISUAL WAVELENGTHS*

Early measurements at visual wavelengths of the interstellar extinction curve by J. Stebbins *et al.*, (1939), established a broad result which has survived unchanged, namely that the amount of the extinction expressed on a logarithmic scale (or a magnitude scale) is approximately proportional to the reciprocal of the wavelength, $A_\lambda \propto 1/\lambda$. This result was refined in the 1960s by K.Nandy (1964), who showed that to a second order of approximation A_λ could be represented in a plot against $1/\lambda$ by two straight line segments, the two segments intersecting at about $1/\lambda = 2.4\ (\mu\text{m})^{-1}$, with the segment appropriate to blue wavelengths being somewhat shallower in slope than the segment appropriate to red wavelengths. Through our personal contacts with Dr. Nandy and with Dr. Reddish at the Edinburgh Royal Observatory, we were led to place considerable weight on this refinement, and it eventually led us to abandon a graphite-silicate model for the grains that we had proposed in 1968-69 (Hoyle and Wickramasinghe, (1968)). A revival of the graphite-silicate model by Mathis *et al.*, (1977), ignored this question, as will be shown in Section 3.

The absolute amount of the extinction is remarkably high, about 2 magnitudes at $1/\lambda = 1.8\ (\mu\text{m})^{-1}$ for a star at a distance of 1 kpc, a circumstance which forces the bulk of the grains to be composed of the commonest elements.

If abundances in the interstellar medium are taken to be approximately solar, silicate grains would be inadequate in abundance by a factor of about 3, even if their sizes were optimally chosen. Grains based on the C, N, O elements could be adequate, however, partly because the CNO group is cosmically more abundant than Mg, Al, Si, Fe, and partly because it can form solids of lower density. Yet with solar abundances even the C, N, O elements formed into solid grains of optimal sizes meet the extinction requirement with only a factor of about 2 in hand.

We do not know the abundances of the elements in the interstellar medium to be solar of course, and arguments by Clegg and Bell (1973) and by Pagel (1974) suggest that present day abundances might be higher than solar by a factor of about 1.7. By stretching this enhancement a little more, the silicate hypothesis could be saved on abundance grounds alone, but the detailed discrepancies given below in Section 3 would still remain.

2. *THE EXTINCTION PEAK NEAR 2200 Å*

Early satellite measurements of stellar extinction in the ultraviolet showed a maximum near 2200 Å, of amount about 6 magnitudes per kiloparsec of pathlength along the galactic plane, whereas about 3.5 to 4 mag per kpc would have been expected from an extrapolation of the data at visual wavelengths. Hence there is an excess extinction, known now to be largely due to absorption, centred near 2200 Å and with a width of a few hundred angstroms. Because it had been realized ahead of the ultraviolet observations actually being made that small graphite particles would produce strong absorption in the region of 2000 Å (Hoyle and Wickramasinghe (1963)) it seemed natural to suppose that the newly-discovered effect was indeed due to graphite, and this has remained the general view in spite of difficulties which became apparent in the late 1960s, namely constraints that are too restrictive on the sizes and shapes of the

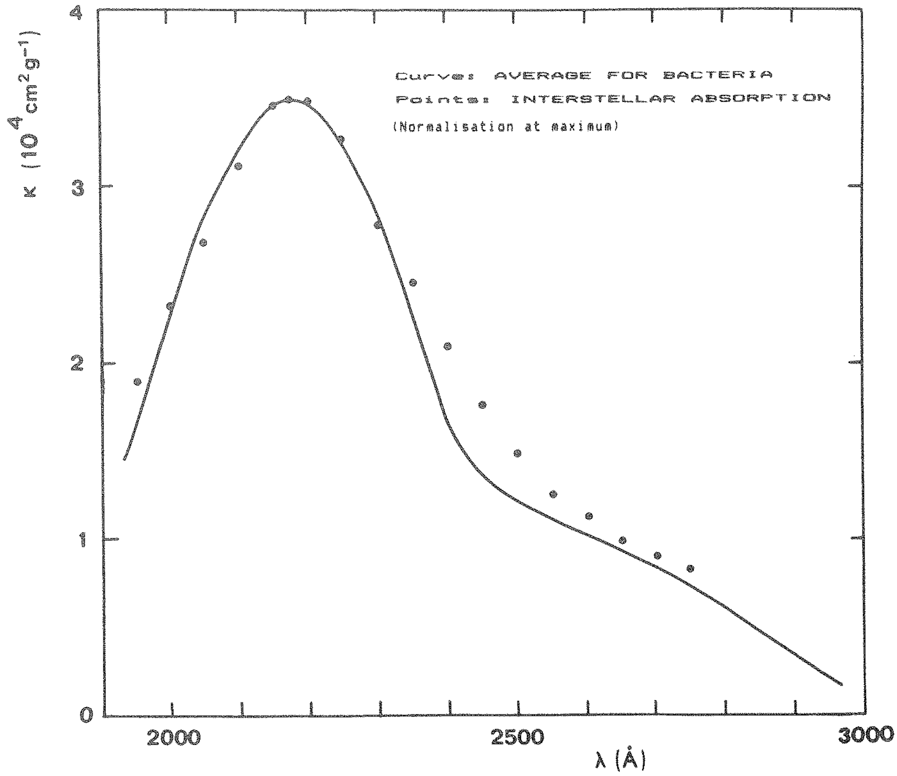


Fig. 29. Normalized interstellar absorption compared with average properties of micro-organisms.

particles – nearly precise spheres with radii almost precisely 0.02μm would be required.

Some years ago (Hoyle and Wickramasinghe (1979)) we showed that a mix of biomaterials produces absorption near 2200Å which is very like the measured interstellar absorption. Recently in collaboration with S. Al-Mufti we have shown further that this result is well-reproduced by bacteria

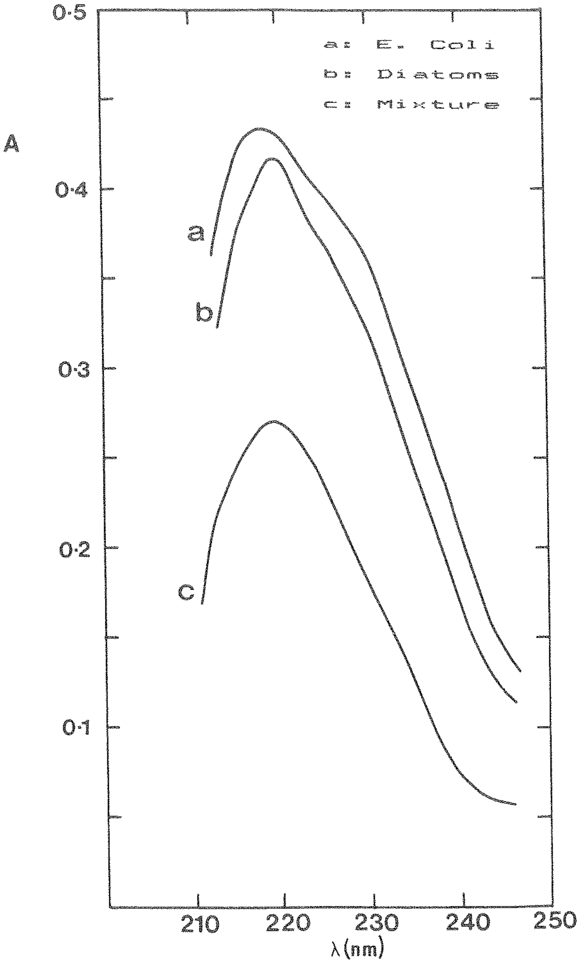


Fig. 30. The absorbance spectrum of dry micro-organisms in dimethylbutane. Curve (a) s for *E.coli*, (b) for diatoms, (c) for a mixture of (b) and (a) in the ratio 2:1. The mass of *E.coli* was measured accurately to correspond to 0.425 mg/15 ml.

themselves, as can be seen from Figure 29. Measured absorbences as functions of wavelength in a number of cases are shown in Figure 30 (where $A = \log_{10} I_0/I$, I being the measured intensity including absorption, and I_0 being without absorption in the comparison beam of the spectrometer). The peak absorption was measured to be ~ 35,000 cm^2 gm^{-1}, compared with an extinction due to scattering at 2200 Å that is estimated to be ~ 50,000 cm^2 gm^{-1}.

3. *QUANTITATIVE DIFFICULTIES WITH THE GRAPHITE-SILICATE HYPOTHESIS*

In a recent criticism of our point of view, it was said that the observed average extinction curve 'can be reproduced exactly in both amplitude and shape from a mixture of inorganic solids such as $MgSiO_3$ and graphite'. This is not true. If it were true, we would never have abandoned the graphite-silicate model which was our own suggestion (Hoyle and Wickramasinghe, 1969). The reference is to Mathis *et al.*, (1977), and so we will show here what really happens when one adopts the detailed model proposed by the latter authors. The model contains what from our point of view is a battery of *ad hoc* features, as well as making unacceptable adjustments to the optical constants of graphite. The best values for the optical constants of graphite available in the literature are those of Taft and Philipp (1965). Making adjustments to these values to suit oneself is in our opinion no game at all. The *ad hoc* features are:

(1) The mass ratio of graphite particles to silicate particles appears as a free parameter.

(2) The size distribution of both kinds of particle is taken without either observation or physical reason to be of the form

$$n(a)\,da = (\text{constant}) \cdot \frac{da}{a^{3\cdot 5}},$$

both being spheres with a the radial parameter.

(3) The end points of this distribution are also free parameters, the results being sensitive to the chosen upper limit of $a = 0.25\ \mu m$, especially for graphite.

Figure 31 shows calculated results for graphite particles alone (G) and for silicate particles along (S_1 being what we think the UV properties of silicates should be, S_2 being what Mathis *et al.*, think they should be). The ordinate scale of Figure 31 is chosen subject to the normalization condition that $\Delta m = 1.8$ at $\lambda^{-1} = 1.8\ (\mu m)^{-1}$.

Neither silicate particles alone nor graphite particles alone gives an extinction that matches the observations correctly. To obtain the classic $1/\lambda$ dependence it is necessary to combine the two kinds of particle, so that

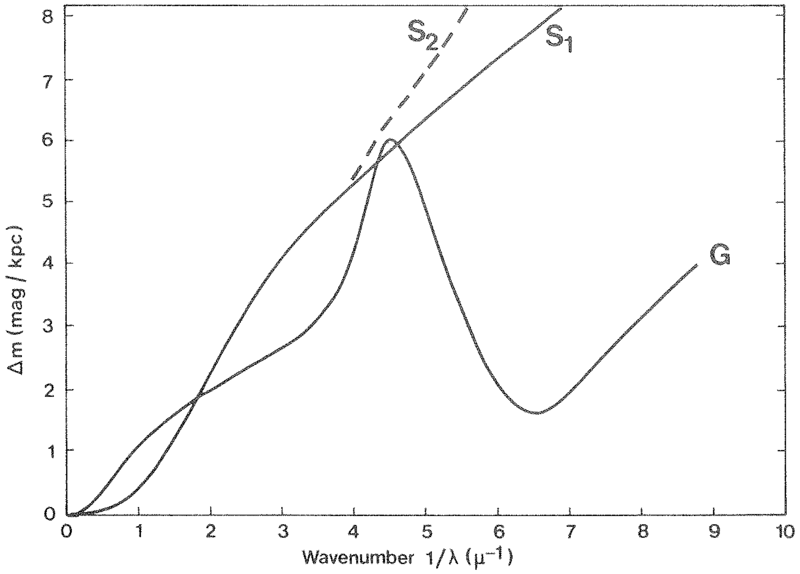


Fig. 31. Normalized extinction curves for Mathis *et al.*, distributions of silicates (S_1 ≡ model with m = 1.6; S_2 ≡ model with m(λ) given by measurements of Huffman and Stapp) and graphite.

obtaining the classic 1/λ dependence becomes an artefact of the free parameter appearing in *ad hoc* feature (1). Making the best weighting of the two kinds of particle (graphite to silicate in the ratio 1.8 at $\lambda^{-1} = 1.8\ \mu m^{-1}$) gives the result shown in Figure 32, GS_1 being the combination of G and S_1 and G S_2 the combination of G and S_2. The points marked in Figure 32 are the average observations which were claimed to be reproduced 'exactly' by the model. If the reader can be persuaded to consult Mathis *et al.* (1977) it will be found that the supposed 'exact' reproduction appears as Figure 4 of that paper, about which one can make the following comments:

(a) The scale is poor.
(b) Calculations and observations are inverted from the way they are usually presented.

It is usual to present observations as points and calculations as a curve, not the opposite way around.

Scale is important. Indeed, the standard of quantitative accuracy in any subject is shown by the scales on which curves are drawn. So far as the visual range is concerned, the scale of Figure 32, though better than that of Mathis *et al.*, is still too small to represent the standards that were set as long ago as the

1960s following the observations of Nandy. Figure 33 shows Nandy's observations for the Cygnus region compared with the calculated curve given by the many-parameter model of Mathis *et al.*, the situation being worse than Wickramasinghe and Nandy obtained in 1971.

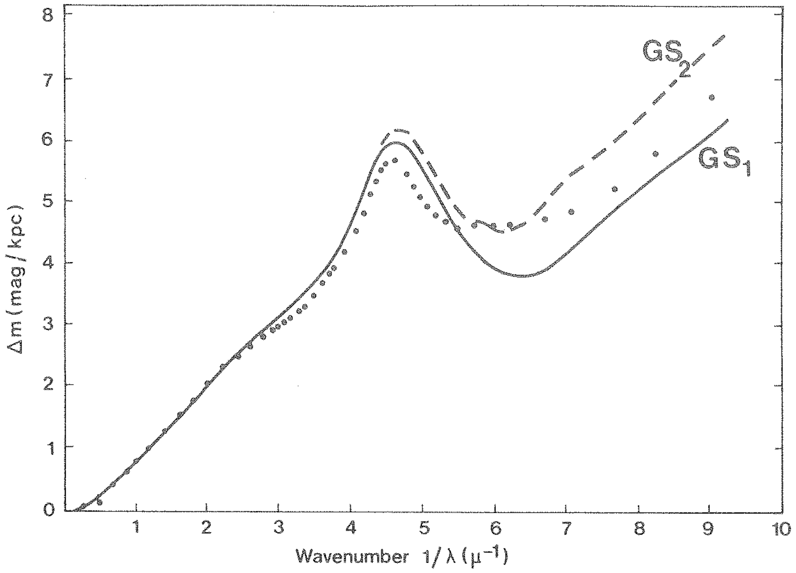


Fig. 32. Best fitting normalized extinction curves for graphite-silicate grain mixtures. GS_1 is for graphite and S_1 silicate (as above). Points are OAO data in the ultraviolet with Nandy's extinction data at optical wavelengths. Normalization is to $\Delta m = 1.8$ at $\lambda^{-1} = 1.8\,\mu m^{-1}$. The ratio of contribution of graphite to silicate extinction at $1.8\,\mu m^{-1}$ is 1.8:1.

4. *AN ALMOST EXACT DETERMINATION OF THE EXTINCTION CURVE IN THE VISUAL RANGE*

Figure 34 is a comparison of the values of Q_{ext} plotted against $2\pi a/\lambda = x$ for spheres of radius a. In both cases the particles are considered to contain weak absorbers (for λ in the visual range) the curves being obtained by Mie calculations for particles with refractive index 1.1-0.05 i in the upper panel and for 1.6-0.05 i in the lower panel.

Every particle in a distribution of sizes has a definite x at a particular chosen wavelength. So at a chosen wavelength the particles in a size distribution each have a distinct point on a curve calculated for its refractive index, the curve in the upper panel of Figure 34 if the refractive index were 1.1-0.05 i, and that in the lower panel if the refractive index were 1.6-0.05 i. When the wavelength λ changes the whole of such a distribution of points moves together up or down

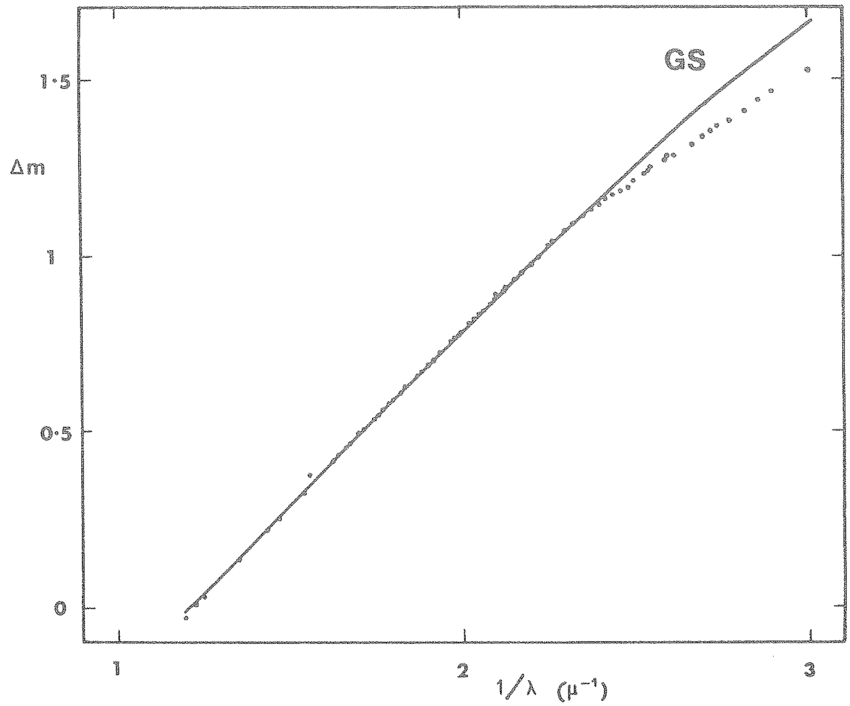


Fig. 33. Cygnus extinction curve compared to best fitting graphite-silicate model with Mathis *et al.*, distribution, and with m = 1.6 for silicates. Normalization is to Δm = 0 at $\lambda^{-1} = 1.22\ \mu m^{-1}$ Δm = 1 at $\lambda^{-1} = 2.22\ \mu m^{-1}$.

its appropriate curve, up for particles with diameters of ~ 1μm if the wavelength is reduced, down if the wavelength is increased. Should most of the points lie on the approximately linear part of the extinction curve, *and should the points remain on the linear part as the wavelength changes*, a 1/λ dependence of the extinction will evidently be obtained. Even if the particles are all closely of the same radius this situation is scarcely achievable for refractive index 1.6-0.05 i, because the curve is approximately linear only from x = 1.3 to x = 2.5. For refractive index 1.1-0.05 i on the other hand, the curve is approximately linear from x = 1 to values of x as large as 10, making it comparatively easy to achieve a 1/λ dependence. In this connection it is important to realize that the observational 1/λ law holds as a first approximation very widely in our galaxy, and even in the Large Magellanic Cloud where the 2200 Å feature discussed in Section 3 is distinctly weaker than in our galaxy. Thus the 1/λ law is essentially invariable, which points strongly to it being a consequence of the physical properties of the interstellar particles.

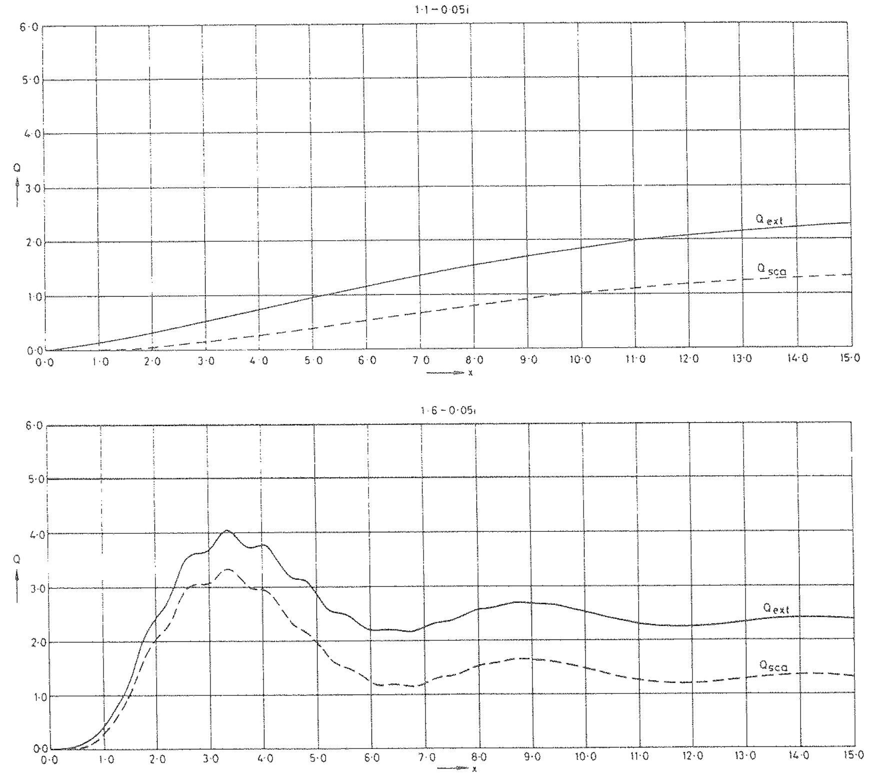


Fig. 34. Extinction and scattering efficiency factors as functions of $x = 2\pi a/\lambda$ for spheres of refractive index $m = 1.1 - 0.05\,i$, $m = 1.6 - 0.05\,i$.

Silicate particles have a real part of the refractive index in excess of 1.6, which is where the discrepancies of the graphite-silicate model shown by Figure 33 come from. A real part of the refractive index close to unity is needed to explain the $1/\lambda$ law convincingly, and also to give a chance of explaining the second-order deviation from the $1/\lambda$ law discovered by Dr. Nandy.

We have known this situation from the middle 1960s but were for long baffled by what kind of a particle could have a real part of the refractive index as low as 1.1. Of solid materials only hydrogen has an adequately small refractive index, which led us in the 1960s to explore the possibility that most of the interstellar grains might be solid hydrogen (Wickramasinghe and Reddish, (1968); Hoyle *et al.*, (1968)). The idea did not take root, however, because most astronomers considered that solid hydrogen is too volatile to exist at typical grain temperatures.

A small hollow particle behaves with respect to the scattering of light in very much the same way as a solid particle of the same average real part of the refractive index. That is to say, a hollow particle in which a fraction f of the particle is occupied by material with a real part of the refractive index n has an average refractive index $\tilde{n} = 1 + (n-1)f$, and would behave closely as a solid particle of refractive index $\tilde{n}$. The example $n= 1.5$, $f = 1/4$, gives $\tilde{n} = 1.125$, which is the kind of situation we are seeking. These values of n and f are appropriate for dry bacteria, so that if most of the interstellar particles were dry bacteria their extinction curve would be like the one in the upper panel of Figure 34. Bacteria often contain pigments which are weak absorbers, giving a small value of the complex part of the refractive index over the visual range of wavelengths, as was used in computing the curves of Figure 34.

It seems that to most people the idea of an important component of the interstellar grains being bacteria is so absurd as to be dismissed out-of-hand without bothering to consider the facts, as if one were born with *a priori* knowledge of what is true and false irrespective of observation and experiment. We ourselves were not blessed with this faculty and so we had perforce to proceed as the facts directed us. Noting that rod-shaped particles behave like spherical particles of the same diameter, we first compiled a size distribution for bacteria using a standard text (*Bergey's Manual of Determinative Bacteriology*). We were then able to relate this experimentally-determined size distribution to an extinction curve like the one at the top of Figure 34, doing so for various values of λ. Once a value of λ was assigned, the bacterial size distribution gave a set of values of x. The ordinate values corresponding to the set of x were then read-off from the curve and added together. Subject to multiplication by a normalization constant, this procedure gave the extinction values shown by the curve of Figure 35. The points of Figure 35 are Nandy's observations for the Cygnus region. Not only did the calculation give the classic $1/\lambda$ dependence in first order, but it gave Nandy's second-order deviation from strict linearity. It did so, moreover, without any *ad hoc* parameter-fixing being necessary. That commentators prefer Figure 33 with extensive parameter-fixing to Figure 35 without parameter-fixing passes our comprehension.

5. *ABSORPTION IN THE 3.4 μm WAVEBAND*

In the course of experiments carried out in the winter of 1980-81, it was found that all biological specimens we examined had much the same absorption in a wavelength range 3.4 ± 0.1 μm, as shown by the examples in the upper panel of Figure 36 (Hoyle *et al.*, (1982). It is fortunate that the real part of the refractive index of biological materials is very close to that of the KBr in which the materials are enclosed in order to make the infrared measurements, because this coincidence eliminates scattering as a source of extinction, permitting the

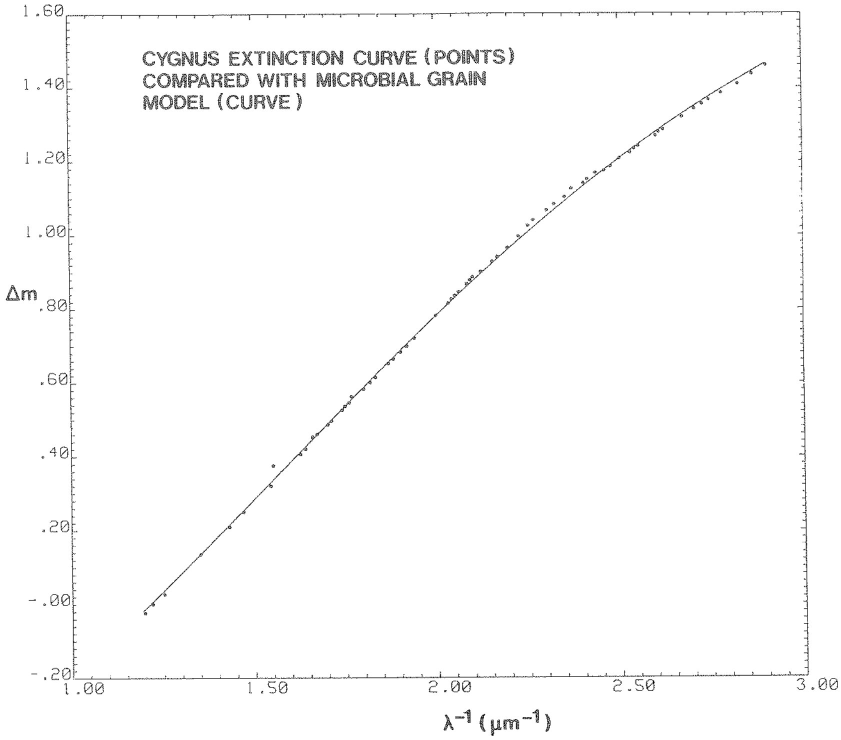


Fig. 35. Cygnus extinction curve (points) compared with microbial grain model (curve).

transmittance values which are measured to be interpreted as arising from absorption alone.

The lower panel of Figure 36 compares the extinction of bacteria heated to 350° C (B of the upper panel) with observations of the infrared source GC-IRS 7 (Allen and Wickramasinghe, (1981)). The observations were made several months after the laboratory experiments were done, and so were unknown to us at the time when the invariant situation at 3.4 ± 0.1 μm was discovered. The reason why the case of heated bacteria was used to make the comparison of the lower panel of Figure 36 was that specimens otherwise tend to pick-up some water owing to the humidity of the laboratory air, even after they have been carefully dried. Because of the immensely strong absorption of water between 3 and 3.1 μm only a very little unwanted water affects the measurements in this wavelength range.

If the disposition among astronomers had been to believe that interstellar grains are bacteria, the agreement between the experimental curve for bacteria

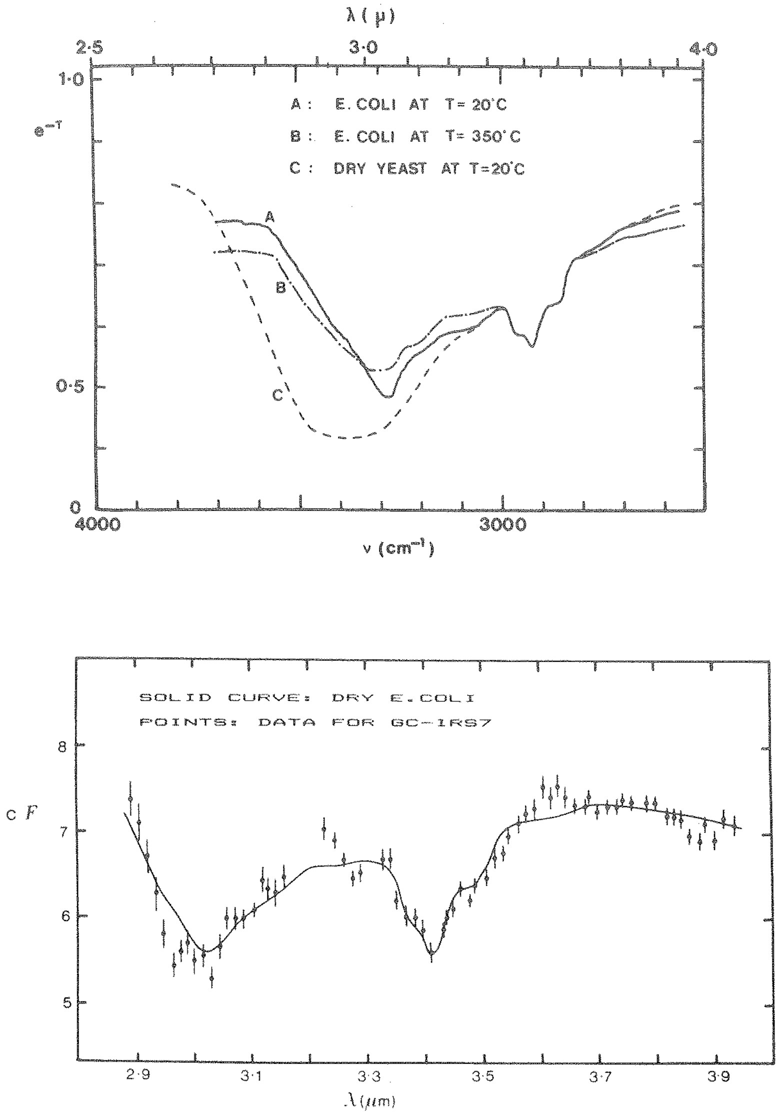


Fig. 36. Upper: The measured transmittance curves of micro-organisms. For *E.coli* a dry mass of 1.5 mg was dispersed in a KBr disk of radius 0.65 cm. The transmittance data for yeast was normalized to agree with the *E.coli* curves at $\lambda = 3.406\ \mu m$.
Lower: Comparison of the infrared data for GC – IRS 7 with predicted curve for dry *E.coli* for sample B in upper figure.

and the observational points for interstellar grains along a path going all the way to the galactic centre would have been regarded as decisive proof of their belief, especially as the laboratory experiments preceded the astronomical observations, making the situation predictive. But because astronomers, seemingly more than most folk, possess the *a priori* knowledge referred to above the following was said: We know *ex cathedra* that the interstellar grains are *not* bacteria. Therefore the agreement of Figure 36 must be fortuitous. Therefore some other material, some non-biological material, must be capable of producing essentially the same absorption as bacteria. Very well then, what is this other non-biological material? Having regard to abundance arguments which require that the CNO elements, possibly in combination with H, make a dominant contribution to the grains, the material must at least be organic. This is so because inorganic hydrides like H_2O, NH_3, do not fit the data at all, and because the absorption in the 3.4 ± 0.1 μm region demands C-H linkages. Since one of us was the first to propose and develop the idea that the interstellar grains might have an important organic component, this is at least in partial correspondence with our point of view (Wickramasinghe, (1974)).

So what non-biological material could produce an agreement with the observations of GC-IRS 7 to equal the bacteria of Figure 36? A number of proposals have been made that were in marked disagreement with the data, even in the 3.4 ± 0.1 μm band, and in no correspondence at all with the data outside this limited band. That these proposals were ever made was again due to the *a priori* knowledge possessed by astronomers and chemists which tells them that the bacterial explanation must be wrong.

So far as the 3.4 ± 0.1 μm band is concerned, but not outside this band, a recent suggestion of Duley and Williams (1983) should be noted. These authors use the experimental work of Watanabe *et al.*, (1982), who obtained hydrated amorphous carbon films through the decomposition of ethylene gas in a high temperature discharge. The films showed features at wavelengths of 3.38 μm, 3.41 μm, and 3.48 μm that are essentially identical to the wavelengths of features in the astronomical data. From the point of view of the critic this is the good side of the story. The bad side is that the detailed results of Watanabe *et al.* were dependent on the temperature of the discharge. Unlike the biological situation which was invariant, the hydrated-carbon films were considerably variable. Choosing the best of them, the fit to the observations for wavelengths from 3.3 to 3.5 μm is noticeably worse than the bacterial fit, and over the major part of the wider band shown in Figure 36 there is no approximation to a fit.

But why, we are compelled to wonder, is ethylene thought to be a plausible precursor material for a large fraction of the interstellar grains? Ethylene on the Earth comes from natural gas that is biological in its origin. We have no knowledge of the existence of ethylene in quantity except in this biological context, which is probably why the specific wavelengths obtained by Watanabe *et al.* agree so well with the astronomical wavelengths.

6. *THE 10 μm BAND*

The discovery (Woolf and Ney, (1969); Knacke *et al.*, (1969); Ney and Allen, (1969); Stein and Gillette, (1969)) of absorption and emission in a broad band centred around 10 μm has provided support for the graphite-silicate model discussed above in Section 3, the basis of the support being the moderately-strong oscillator strength due to the Si-O bond, which yields a mass absorption coefficient of about 6000 $cm^2\ gm^{-1}$ for silicates, an order of magnitude larger than the C-H bond near 3.4 μm which gives only about 600 $cm^2\ gm^{-1}$, but still considerably less than water-ice in the 3-3.1 μm region – the latter absorption is about 33000 $cm^2\ gm^{-1}$.

The simplest astronomical example of particles emitting in the 10 μm band comes from the Trapezium region of the Orion nebula. Most, but not all, astronomical infrared sources contain particles with nearly the same emission and absorption characteristics as the Trapezium region. A self-deceiving practice has grown up whereby 'silicates' and 'Trapezium material' have become identified together, so that for many people a 'silicate' is taken to possess the same infrared properties as Trapezium material, by divine command as it were. The truth is otherwise. A silicate is a silicate, with properties determinable in the laboratory. There are many forms of silicate minerals to be examined with a view to obtaining a correspondence with Trapezium material, if possible. As the outcome of 15 years of study, the best correspondences to have emerged are shown in the upper panel of Figure 37. This is worse, or at any rate not better, than one of us immediately obtained for an organic polymer as long ago as 1974 (Wickramasinghe, (1974)).

The ensemble of micro-organisms on Earth contains siliceous organisms as well as wholly carbonaceous ones, and indeed about a quarter of terrestrial oxygen production due to photosynthesis arises from siliceous organisms. We could see no better way of obtaining a suitable biological mix than simply taking a bucket of river water that was suitable in the sense the river supported fairly rich growths of micro-organisms. After culturing in the laboratory in order to secure an adequate concentration of organisms their infrared properties were measured, when the correspondence to the Trapezium nebula shown in the lower panel of Figure 37 emerged from the measurements (Hoyle *et al.*, (1982)).

For people blessed with profound *a priori* knowledge that there can be no connection between biology and astronomy, this was again so much water off a duck's back. But for us who perforce must work from facts it seemed impressive that, whereas the upper panel of Figure 37 took 15 years to achieve,

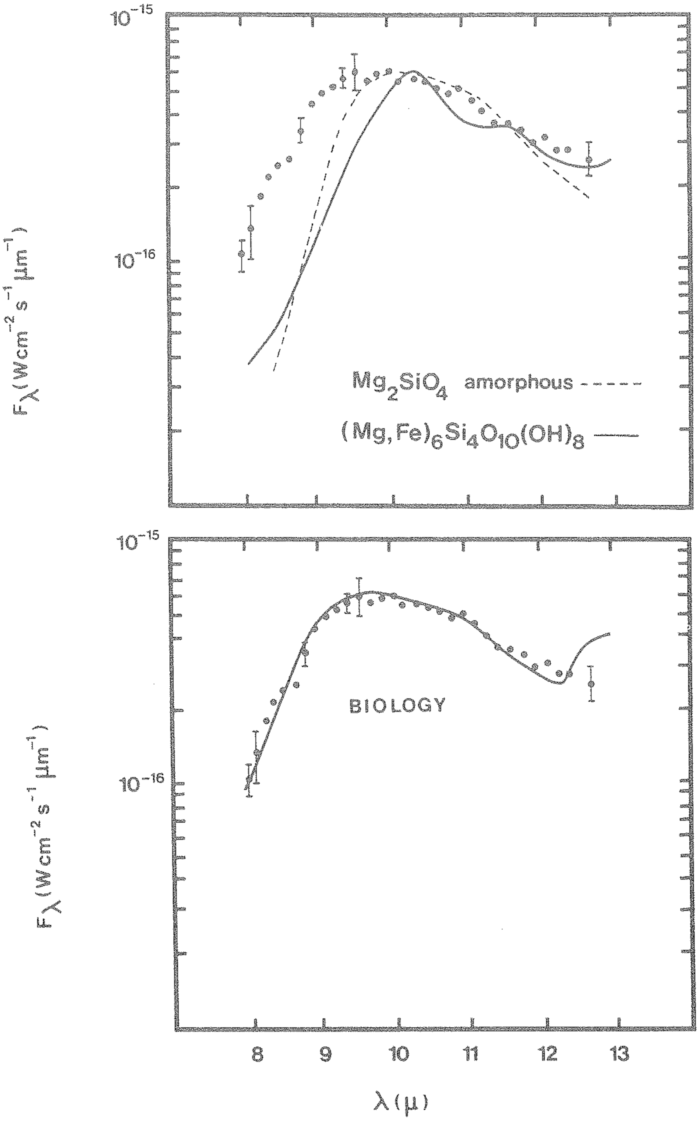


Fig. 37. Upper: Comparison between the flux measurements for the Trapezium nebula (points) with the prediction for 0.3 μm sized silicate grains (amorphous and hydrated) heated to 175K. Observations and calculations are normalized to correspond at maximum value of flux.

Lower: Comparison between the flux measurements for the Trapezium nebula (points) with the predicted curve for mixed diatom culture radiating at 175K. Observations and theory are normalized to correspond at maximum value.

or not to achieve according to how one looks at it, the lower panel of Figure 37 was obtained in less than a week. Once the idea was conceived the required correspondence came out immediately.

7. *CONCLUSIONS*

If the point of view outlined above were incorrect, we would have to suppose that an incorrect theory yields considerably better correspondences with observations ranging from the ultraviolet to the far infrared than does the correct theory. We would have an incorrect theory yielding very good correspondences quickly and without parameter-fixing, with the correct theory only yielding poor correspondences after many years of work and with much parameter-fixing. This makes no sense in logic or in experience. The basic issue goes beyond astronomy and biology as such. The issue is cultural. For reasons which require urgent examination, the educational process has conditioned us to believe with great passion that life is confined to the Earth. No facts have ever been obtained to support this position. Such support as there has been is confined to quizzical metaphysics, as is implied in the well-known question: 'Why aren't the others here?' This question is often developed with wild extrapolations, and with assumptions that are demonstrably incorrect, into the claim that man is a uniquely intelligent animal, a protestation which the poor quality of the argument itself contradicts. That this piffling stuff should be taken seriously is indicative of a sad state of affairs, indicative of an ennui which causes many scientists to hold strong opinions on matters which they have not calculated properly or even looked into with care, an ennui which causes them to become dominated by what one might call the hole-in-one syndrome. The question 'Why aren't the others here?' is an appeal to the hole-in-one syndrome, and so was the phosphorus argument discussed briefly below in a short Appendix. Golfers do occasionally manage to hole-in-one, but the probabilities such as they are favour those who spend many hours in mastering the details of their game. Similarly, one can only expect to arrive at a correct viewpoint in science through many hours of effort in which the facts are accorded primary status. To attempt to decide matters through *a priori* knowledge or cultural prejudice is an invitation to disaster, with as little chance of success as the inept golfer who swings wildly in an attempt for a hole-in- one. Those who dislike the contents of this chapter are advised to repeat the calculations for themselves, and the experiments if possible. Dislike is useless. Referring the matter to other supposed authorities is also useless. Only in one's own efforts, in the pondering of facts, can the truth be perceived.

APPENDIX

An issue has been raised concerning the availability of phosphorus. If the more numerous classes of interstellar particles are bacteria how much phosphorus is required to provide for their obligate needs? On the assumption the bacteria are all viable, the answer is about 2/3 percent of their dry weight, not the 2 to 3 percent found in bacteria that are multiplying very rapidly under conditions of luxuriant nutrient supply. The latter have many more ribosomal units than bacteria existing in nutrient-limiting circumstances, and it is the excessive ribosomal units which gobble up most of the phosphorus, not the obligate DNA and lipids.

Bacteria have sizes that fall on the linear part of the upper panel of Figure 34. Such sizes are optimal for the scattering of starlight, requiring a mass density that is a fraction s/300 of the average interstellar hydrogen density (in order to explain the amount of the scattering). Here s is the specific gravity of the particles, about 1.5 gm cm^{-3} for biomaterial. Thus the required bacterial density is 1/200 of the average hydrogen density.(It is worth noting that because s for silicates is about 3 gm cm^{-3}, the density requirement on silicates would be 1/100 of the hydrogen density). On the assumption that bacteria are viable, the obligate phosphorus requirement is therefore seen to be 1/30,000 of the hydrogen density. This is in terms of mass. In terms of numbers of atoms, phosphorus is required to be $(31 \times 30{,}000)^{-1}$ of the hydrogen number density, i.e. about 1 phosphorus atom to 10^6 hydrogen atoms. How does this compare with expectation?

In solar material the ratio of hydrogen atoms to phosphorus atoms is usually said to be about 3×10^6, but there is nothing sacrosanct about solar abundances. If heavy elements are enhanced above solar in the interstellar medium, say by a factor 1.7 (c.f. Clegg and Bell (1973); Pagel, (1974)), the margin between availability and requirement becomes negligible, especially as phosphorus abundances are determined from high-excitation lines in the solar spectrum for which oscillator strengths could easily be in error by the modest factor that is involved here. Hence there is no discrepancy, even if the bacteria are taken to be pristine. If the bacteria are partially degenerated, the lipids of the cell membranes could be broken down with their phosphorus returned to the external medium, which would reduce the phosphorus requirement by a factor of about 2, since a return to the medium would permit the same phosphorus to be used repeatedly. If the DNA were also broken down for the bulk of the bacteria, the possibility of phosphorus being used repeatedly reduces the requirement still further. Evidently then, there is a somewhat nice balance between availability and requirement, this being supportive rather than detrimental to the theory.

REFERENCES

Allen, D.A. and Wickramasinghe, D.T., 1981, *Nature*, *294*, 239

Bergey's Manual of Determinative Bacteriology, ed. Buchanan, R.E. and Gibbons, N.E., 1975, 8th edition, Williams and Wilkins Co.

Clegg, R.E.S. and Bell, R.A., 1973, *Mon.Not.R.astron.Soc.*, *163*, 13

Duley, W.W. and Williams, D.A., 1983, *Mon.Not.R.astron.Soc.*, *205*, 67

Hoyle, F., Wickramasinghe, N.C. and Reddish, V.C., 1968, *Nature*, *218*, 1124

Hoyle, F., Wickramasinghe, N.C., Al-Mufti, S. and Olavesen, A.H., 1982, *Astrophys.Space Sci.*, *83*, 405

Hoyle, F., Wickramasinghe, N.C. and Al-Mufti, S., 1982, *Astrophys.Space Sci.* *86*, 63

Hoyle, F., Wickramasinghe, N.C., Jansz, E.R. and Jayatissa, P.M., 1983, *Astrophys.Space Sci.*, *95*, 227

Hoyle, F. and Wickramasinghe, N.C., 1963, *Mon.Not.R.astron.Soc.*, *126*, 27

Hoyle, F. and Wickramasinghe, N.C., 1968, *Nature*, *223*, 459

Hoyle, F. and Wickramasinghe, N.C., 1979, *Astrophys.Space Sci.*, *65*, 241

Knacke, R.F., Gaustad, J.E., Gillett, F.C. and Stein, W.A., 1969, *Astrophys.J.*, *155*, L189

Mathis, J.S., Rumpl, W. and Nordsieck, K.H., 1977, *Astrophys.J.*, *217*, 425

Nandy, K., *Publs.R.Obs.Edinb.*, 1964, *3*, 142

Nandy, K., *Publs.R.Obs.Edinb.*, 1965, *5*, 13

Nandy, K., *Publs.R.Obs.Edinb.*, 1966, *5*, 233

Ney, E.P. and Allen, D.A., 1969, *Astrophys.J.*, *155*, L193

Pagel, B.E.J., 1974, NATO Conference, Cambridge, 23 July–9 August

Stebbins, J., Huffer, C.M. and Whitford, A.E., 1939, *Astrophys.J.*, *90*, 209

Stein, W.A. and Gillett, F.C., 1969, *Astrophys.J.*, *155*, L197

Taft, E.A. and Philipp, H.R., 1965, *Phys.Rev.*, *138*, 197

Watanabe, I., Hasegana, S. and Kurota, Y., 1982, *Japanese Journal of Applied Physics*, *21*, 856

REFERENCES

Wickramasinghe, N.C., 1974, *Nature, 252*, 462

Wickramasinghe, N.C. and Nandy, K., 1971, *Mon.Not.R.astron.Soc., 153*, 205

Wickramasinghe, N.C. and Reddish, V.C., 1968, *Nature, 218*, 661

Woolf, N.J. and Ney, E.P., 1969, *Astrophys.J., 155*, L181

INDEX

INDEX