Planetary Systems and the Origins of Life

Edited by
R.E. Pudritz, P. Higgs, J.R. Stone

Contents

1

Observations of Extrasolar Planetary Systems

Shay Zucker
Tel Aviv University

Abstract

Since the groundbreaking detection of the planet around 51 Pegasi, about 190 extrasolar planets have been detected so far. Most of those planets were detected through precision radial-velocity monitoring. This chapter reviews this technique, as well as other techniques, especially photometric transit detection. At the current stage the population of known extrasolar planets is large enough to make some preliminary statistical observations, such as those concerning mass distribution, orbital characteristics, multiplicity and host-star properties. The chapter reviews those observations, and concludes with the anticipated observations from space missions.

1.1 Introduction

A decade has passed since the first discovery of an extrasolar planet by Mayor & Queloz (1995) and its confirmation by Marcy et al. (1997). Since this groundbreaking discovery, about 190 planets have been found around nearby stars, as of May 2006 (e.g., Schneider 2006). Here we shall review the main methods astronomers use to detect extrasolar planets, and the data we can derive from those observations.

Aitken (1938) examined the observational problem of detecting extrasolar planets. He showed that their detection, either directly or indirectly, lay beyond the technical horizon of his era. The basic difficulty to directly detect planets is the brightness ratio between a typical planet, which shines mainly by reflecting the light of its host star, and the star itself. In the case of Jupiter and the Sun, this ratio is $2.5 \cdot 10^{-9}$. If we keep using Jupiter as a typical example, we expect a planet to orbit

at a distance of the order of 5 Astronomical Units (AU) from its host star. At a relatively small distance of 5 parsecs from the Solar System, this would mean an angular separation of 1 arc-second. Therefore, with present technology, it is extremely demanding to directly image any extrasolar planet inside the overpowering glare of its host star, particularly from the ground, where the Earth's atmosphere seriously affects the observations. Nevertheless, significance advances have been made in the fields of coronagraphy and adaptive optics and positive results are very likely in the coming few years.

The vast majority of the known extrasolar planets were detected by *indirect* means, which matured in 1995 to allow the detection of the giant planet around the star 51 Peg through the spectroscopic *radial-velocity* (RV) technique. RV monitoring is responsible for most of the known extrasolar planets. *Transit* detection, another indirect method, has gained importance in the last few years, already yielding a few detections. The next two sections will review those two methods, their advantages and their drawbacks. Section 1.4 will provide a brief account of the emerging properties of the extrasolar planet population. In Section 1.5 we will briefly review other possible methods of detection. Section 1.6 will conclude the chapter with some predictions about future observations from space.

1.2 Radial-velocity detections

Consider a star–planet system, where the planet's orbit is circular, for simplicity. By a simple application of Newton's laws, we can see that the star performs a reflex circular motion about the common center of mass of the star and planet, with the same period (P) as the planet. The radius of the star's orbit is then given by:

$$a_\star = a(\frac{M_\mathrm{p}}{M_\star}) \ , \tag{1.1}$$

where a is the radius of the planet orbit, and M_p and $M_\star$ are the planet mass and the star mass, respectively. The motion of the star results in the periodic perturbation of various observables that can be used to detect this motion. The RV technique focuses on the periodic perturbation of the line-of-sight component of the star's velocity.

Astronomers routinely measure RVs of objects ranging from Solar System minor planets to distant quasars. The basic tool to measure RV is the spectrograph, which disperses the light into its constituent

wavelengths, yielding the stellar *spectrum*. Stars like our Sun, the so-called *main-sequence* stars, have well known spectra. Small shifts in the wavelengths of the observed spectrum can tell us about the star's radial velocity through the *Doppler effect*. Thus, a Doppler shift of $\Delta\lambda$ in a feature of rest wavelength λ in the stellar spectrum corresponds to a radial velocity of:

$$v_{\rm r} = \frac{\Delta\lambda}{\lambda} c \ , \tag{1.2}$$

where c is the speed of light.

The most obvious parameters which characterize the periodic modulation of the radial velocity are the period – P, and the semi-amplitude –K (Fig. 1.1(a)). These two parameters are related to the planet mass via the general formula (e.g., Cumming et al. 1999):

$$K = (\frac{2\pi G}{P})^{\frac{1}{3}} \frac{M_{\rm p} \sin i}{(M_\star + M_{\rm p})^{\frac{2}{3}}} \frac{1}{\sqrt{1-e^2}} \ . \tag{1.3}$$

In this formula G is the universal gravitational constant, and e is the orbital eccentricity. The inclination of the orbital axis relative to the line of sight is denoted by i (Fig. 1.1(b)). In a circular orbit we can neglect e, and assuming that the planet mass is much smaller than the stellar mass, we can derive the empirical formula:

$$K = (\frac{P}{1\,{\rm day}})^{-\frac{1}{3}} (\frac{M_\star}{M_\odot})^{-\frac{2}{3}} (\frac{M_{\rm p} \sin i}{M_{\rm J}}) \ 203 \ {\rm m\,s^{-1}} \tag{1.4}$$

$M_{\rm J}$ denotes Jupiter mass, and $M_\odot$ stands for the Solar mass.

Close examination of Equation 1.4 reveals several important points. First, K has a weak inverse dependence on P, which means that the RV technique is biased towards detecting short-period planets. Second, the planet mass and the inclination appear only in the product $M_{\rm p} \sin i$, and therefore they cannot be derived separately using RV data alone. In principle, a planetary orbit observed edge-on (i close to 90°) will have exactly the same RV signature like a stellar orbit observed face-on (i close to 0). Statistics help to partly solve the conundrum, since values of $\sin i$ which are close to unity are much more probable than smaller values (e.g., Marcy & Butler 1998). In fact, for a randomly oriented set of orbits, the mean value of $\sin i$ is easily shown to be $4/\pi \approx 0.785$. Obviously, a better solution would be to seek independent information about the inclination.

Equation 1.4 shows the order of magnitude of the desired effect – tens or hundreds of meters per second. Detecting effects of this magnitude

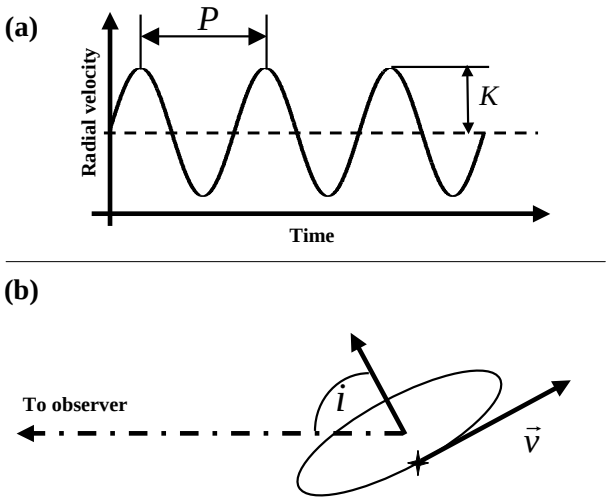


Fig. 1.1. (a) A schematic illustration of a periodic radial-velocity curve of a planetary orbit, showing the two quantities P (period) and K (semi-amplitude). (b) Visualization of the inclination angle (i) – the angle between the orbital axis and the line of sight.

requires precisions of the order of meters per second. Such precisions were almost impossible to achieve before the 1990s. Before that time, the only claim of a very low-mass companion detected via RV was of the companion of the star HD 114762. The semi-amplitude of the RV variation was about $600\,\mathrm{m\,s^{-1}}$, and the companion mass was found to be around $10\,M_\mathrm{J}$ (Latham et al. 1989; Mazeh et al. 1996). Although the existence of this object is well established, the question of its planetary nature is still debated. Alternatively, it could be a *brown dwarf* – an intermediary object between a planet and a star. The detection of smaller planet candidates had to await the development of instruments that could measure *precise* radial velocities.

Campbell & Walker (1979) were the first to obtain RVs of the required precision. They introduced an absoprtion cell containing hydrogen fluoride gas in the optical path of the stellar light in order to overcome systematic errors in the RVs, using the known spectrum of the gas for calibration. They carried out a pioneering survey of 16 stars over a period of six years, which yielded no detections, probably because of its small size (Campbell et al. 1988).

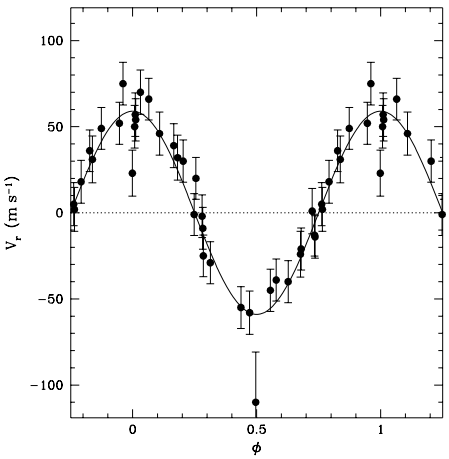


Fig. 1.2. The phase-folded RV curve of 51 Peg, from Mayor & Queloz (1995).

The first planet candidate detected using precise RV measurements was 51 Peg b. Mayor & Queloz (1995) used the fiber-fed ELODIE spectrograph in the Haute-Provence Observatory (Baranne et al. 1996), and obtained a RV curve of 51 Peg corresponding to a planet with a mass of $0.44\,M_{\rm J}$ and an orbital period of 4.23 days (Fig. 1.2). This short period means an orbital distance of 0.05 AU from the host star. The discovery was soon confirmed by Marcy et al. (1997), using the Hamilton echelle spectrograph at the Lick Observatory, with the iodine absorption cell technique (Butler et al. 1996). This proximity to the host star was a major surprise and it actually contradicted the previous theories about planetary system formation. This discovery, and many similar planets that followed (now nicknamed 'Hot Jupiters'), led to a revision of those theories, and to the development of the planetary migration paradigm (e.g., Lin et al. 1996). The current state of the formation and evolution theories is reviewed in Chapter 4 of this volume.

Since that first detection, several groups routinely perform RV measurements. The most prominent groups are the Geneva group, using fiber-fed spectrographs (Baranne et al. 1996), and the Berkeley group, using the iodine-cell technique (Butler et al. 1996).

1.3 Transit detections

We have seen in Section 1.2 that the basic drawback of the RV method is the lack of independent information about the orbital inclination, which leads to a fundamental uncertainty in the planet mass. Currently, the most successful means of obtaining this information is via the detection of planetary *transits*. In the Solar System, transits are a well known rare phenomenon where one of the inner planets (Mercury or Venus) passes in front of the Solar disk. The most recent Venus transit occurred on 8 June 2004, and attracted a considerable public attention and media coverage. Extrasolar transits occur when an extrasolar planet passes in front of its host-star disk. Obviously, we cannot observe extrasolar planetary transits in the same detail as transits in the Solar System. With current technology, the only observable effect would be a periodic dimming of the star light, because the planet obscures part of the star's surface. Thus, transits can be detected by *photometry*, i.e., monitoring the stellar light intensity.

The probability that a planetary orbit would be situated in such a geometric configuration to allow transits is not very high. For a circular orbit, simple geometrical considerations show that this probability is:

$$\mathcal{P} = \frac{R_\star + R_\mathrm{p}}{a} \tag{1.5}$$

where $R_\star$ and R_p are the radii of the star and the planet, respectively (e.g., Sackett 1999). For a typical hot Jupiter, this probability is about 10%.

The idea to use transits to detect extrasolar planets was first raised by Struve (1952), but the first extrasolar transit was observed only in 1999. Mazeh et al. (2000) detected a planet orbiting the star HD 209458, using 'traditional' RV methods. Soon after the RV detection, Charbonneau et al. (2000) and Henry et al. (2000) detected a periodical dimming of the light coming from HD 209458, at exactly the predicted orbital phase and with the same period as the RV variation, of 3.52 days. The light dimmed by about 1.5% for about 1.5 hour (Fig. 1.3). The two teams detected the transits using small and relatively cheap telescopes, demonstrating that it was realistic to achieve the required photometric precision by ground-based observations.

The depth of the transit (i.e., the amount by which the light intensity drops), depends on the fraction of the stellar disc obscured by the planet. Thus, assuming there is a reasonable estimate of the star's radius, we can use the depth to derive the planet radius. This is the first direct es-

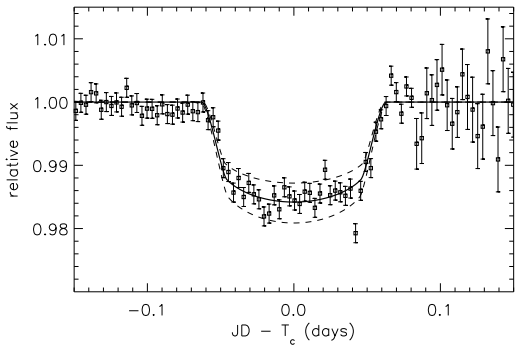


Fig. 1.3. The transit light curve of HD209458, from Charbonneau et al. (2000).

timate we have of a physical property of the planet itself. Obviously, the detection of transits immediately constrains also the orbital inclination (i) to values close to 90° (transits occur only when we observe the orbit edge-on or almost edge-on). Furthermore, the transit duration depends strongly on the orbital inclination, and we can use it to explicitly derive i (e.g., Sackett 1999). Thus, in combination with RV data, we can finally obtain a measurement of the planet mass – $M_{\rm p}$.

Brown et al. (2001) used the Hubble Space Telescope (HST) to obtain a very precise light curve of HD 209458. This light curve led to a very precise estimate of the planet radius – 1.347 Jupiter radii ($R_{\rm J}$). Using the inclination and the mass estimate from the RV orbit, the planet mean density could be derived – $0.35\,{\rm g\,cm^{-3}}$. The special circumstances of a transiting extrasolar planet were exploited by many more observations of HD 209458, with new clues about its atmosphere. Those observations are reviewed in Chapter 3 of this volume.

The successful observations of HD 209458 encouraged many teams to try to detect more transiting extrasolar planets. Currently about 25 surveys are being conducted by teams around the world (Horne 2006). Some of these surveys use small dedicated telescopes to monitor nearby stars (which are relatively bright), in large fields of view, like TrES (Alonso et al. 2004) or HATnet (Bakos et al. 2004). Other surveys, such as OGLE-III (Udalski et al. 2002) or STEPSS (Burke et al. 2004), focus on crowded fields like the Galactic Center or globular clusters, and monitor tens of thousands of stars.

So far, the only successful surveys were OGLE, with five confirmed planets, and TrES and XO, with one planet each. Their success highlights the difficulties such surveys face, and the problems in interpreting the observational data. The basic technical challenge is transit detection itself. The transits last only a small fraction of the time, and the drop in the stellar brightness is usually of the order of $1-2\%$ at most. The first obvious challenge is to reach a sufficient photometric precision. The next challenge is to obtain sufficient phase coverage, on the observational front, and efficient signal analysis algorithms, on the computational front.

The OGLE project has yielded so far 177 transiting planet candidates (Udalski et al. 2004). However, only five thus far were confirmed as planets. This is due to the fundamental problem in using photometry to detect giant planets. Since the transit light curve alone does not provide any information regarding the mass of the eclipsing companion, we have to rely on its inferred radius to deduce its nature. However, it is known (e.g., Chabrier & Baraffe 2000) that in the substellar mass regime, down to Jupiter mass, the radius depends extremely weakly on the mass. Therefore, even if we detect what seems to be a genuine transit light curve, the eclipsing object may still be a very low-mass star or a brown dwarf. The only way to determine its nature conclusively is through RV follow-up that would derive its mass. Thus, while only five of the OGLE candidates were shown to be planetary companions, many others were identified as stellar companions.

The proven non-planetary OGLE candidates demonstrate the diversity of the situations which can 'disguise' as planetary transits. OGLE-TR-122 is a perfect example of a low-mass star which eclipses its larger companion, like a planet would. Only RV follow-up determined its stellar nature (Pont et al. 2005). Another confusing configurations are 'grazing' eclipsing binary stars, where one star obscures only a tiny part of its companion, or 'blends', where the light of an eclipsing binary star is added to the light of a background star, effectively reducing the measured eclipse depth. In principle, these cases can be identified by close scrutiny of the light curve (Drake 2003; Seager & Mallén-Ornellas 2003), or using color information in addition to the light-curve shape (Tingley 2004). However, the most decisive identification is still through RVs.

Although transit detection alone is nut sufficient to count as planet detection, transit surveys still offer two important advantages over RV surveys. First, they allow the simultaneous study of much more stars – the crowded fields monitored by transit surveys contain thousands of

stars. Second, they broaden the range of stellar types which we examine for the existence of planets. While obtaining the required precision of RV measurements puts somewhat stringent constraints on the stellar spectrum, transit detection relies much less on the stellar type. Ths star simply has to be bright enough and maintain a stable enough brightness so that we can spot the minute periodical dimming caused by the transiting planet. Since stellar radii depend only weakly on the stellar mass, we should be able to detect planetary transits even around stars much hotter (and therefore more massive) than the Sun.

1.4 Properties of the extrasolar planets

As of May 2006, 193 extrasolar planets are known (Schneider 2006). This number, although not overwhelming, is already enough to make some preliminary statistical observations. Obviously, these findings have a significant effect on the development of theories concerning the formation and evolution of planets in general, and the Solar System in particular. We shall now review the most prominent features of the growing population of extrasolar planets.

1.4.1 Mass distribution

The definition of planets, especially a definition that would distinguish them from stars, was a central research theme since the very first detections. The most obvious criterion, which remains the most commonly used one, is simply the object mass. The boundary between stars and substellar objects, at $0.08\,M_\odot$, is already well known and physically understood (the mass above which hydrogen burning is ignited). A similar boundary was sought, that would apply for planets. This was set at the so-called 'deuterium burning limit', at $13\,M_\mathrm{J}$ (Burrows et al. 1997). This arbitrary limit is not related to the hypothesized formation mechanisms. The tail of the mass distribution of very low-mass companions suggests that objects with masses as large as $20\,M_\mathrm{J}$ exist. The question remains, whether the mass distribution of the detected planets indeed exhibits two distinct populations – planets and stars. The evidence is mounting that this is indeed the case. It seems that the mass regime between $20\,M_\mathrm{J}$ and $0.08\,M_\odot$ is underpopulated (Jorissen et al. 2001; Zucker & Mazeh 2001b), as can be seen in Fig. 1.4. This depletion was nickcnamed *the Brown-Dwarf Desert* (Marcy & Butler 2000; Halbwachs et al. 2000).

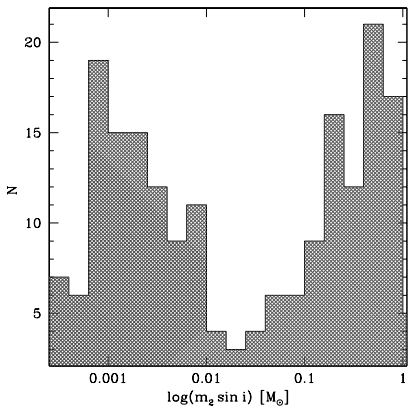


Fig. 1.4. The mass distribution of companions to solar-type stars. Note the dearth of planets at masses between 0.02 and 0.08 Solar masses – 'the Brown-Dwarf Desert'. From Udry et al. (2004).

1.4.2 Mass-period distribution

The existence of 'Hot Jupiters' implied the possible presence of massive planets close to their host stars. Since the RV technique is mostly sensitive to short periods and massive planets, we expect to find a dense population in this part of the mass-period diagram. However, this is not the case, as was shown by Zucker & Mazeh (2002), Udry et al. (2002) and Pätzold & Rauer (2002). This may hint that the planetary migration process is less effective for very massive planets (Nelson et al. 2000; Trilling et al. 2002). Alternatively, it could mean that at very short distances the planet 'spills over' part of its mass onto the host star (Trilling et al. 1998).

1.4.3 Orbital eccentricities

The second extrasolar planet detected, 70 Vir b, turned out to have a considerably eccentric orbit, with an eccentricity of 0.4 (Marcy & Butler 1996). HD 80606 b currently holds the eccentricity record, with an eccentricity of 0.93 (Naef et al. 2001). Such high orbital eccentricities were also a surprise. The matter in protoplanetary disks was assumed to orbit the central star in circular Keplerian orbits, and the planetary orbits were supposed to reflect this primordial feature by being relatively circular, like the orbits in the Solar System. In order to account for

the observed high eccentricities, several models were suggested, and it seems that no single model alone can explain all cases. Chapter 4 of this volume presents some processes which can take part in creating these eccentricities.

1.4.4 Host-star metallicity

The chemical composition of a planet-hosting star is related to that of the primordial molecular cloud where the star formed. The *metallicity* is a measure of the relative amount of elements heavier than hydrogen and helium in the stellar atmosphere. Santos et al. (2004) have shown that the frequency of planets is strongly related to the metallicity of the host star. This effect is much stronger than the selection effect caused by the presence of more metal lines in the stellar spectrum (e.g., Murray & Chaboyer 2002). This can plausibly be explained by the need to have enough solid material and ices in order to form planets in the protoplanetary disk, but a detailed explanation is still missing.

1.4.5 Planetary systems and planets in binary systems

The Solar System is known to include eight major planets. The presence of more than one planet around a host star is easily explained by the protoplanetary disk paradigm (see Chapter 1 of this volume). Thus, we expect extrasolar planets to appear in multiple planet systems as well. The first detection of an extrasolar planetary *system* was a system of three planets orbiting the star υ Andromedae (Butler et al. 1999). In such cases, the motion performed by the star is a combination of the various motions caused by all planets in the system (Fig. 1.5).

A related issue is the existence of planets in systems of binary stars. A considerable fraction of stars are actually binary stars, where two stars are gravitationally bound and orbit their barycenter. Currently, about twenty planets are known to orbit components of binary stars (Udry et al. 2004). The orbital characteristics of these planets seem to differ from those orbiting single stars, e.g., very massive planets can be found at very close proximity to the host star only in binary stellar systems (Zucker & Mazeh 2002).

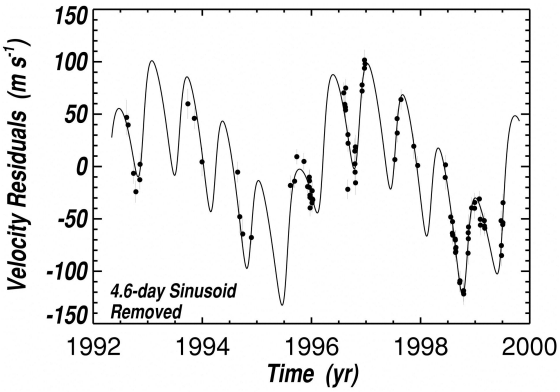


Fig. 1.5. The RV curve of υ Andromedae, after removing the motion caused by the shortest-period planet. Note the non-periodic nature of the motion, which is the sum of the motions caused by the two outer planets. From Butler et al. (1999).

1.4.6 Planetary radii

The transiting planets offer the possibility to study the radii of extrasolar planets. The number is still too small to draw very significant statistical conclusions, but it is worthwhile to examine the data. Figure 1.6 shows the mass-radius diagram of the ten transiting planets, together with Jupiter and Saturn for comparison, and representative isodensity lines. Early attempts to confront such diagrams with theoretical predictions seem promising (e.g., Guillot 2005; Laughlin et al. 2005). It seems that most hot Jupiters have sizes close to that of Jupiter itself, and therefore their structure is not significantly perturbed by their proximity to the star.

1.5 Other methods of detection

So far, the two techniques of RV and transits have contributed most of the observational knowledge on extrasolar planets around main-sequence stars. However, considerable efforts were put into exploring other techniques, mainly in order to have a better coverage of the various configurations in which planets can be found. In the following paragraphs we give a very brief account of four such techniques.

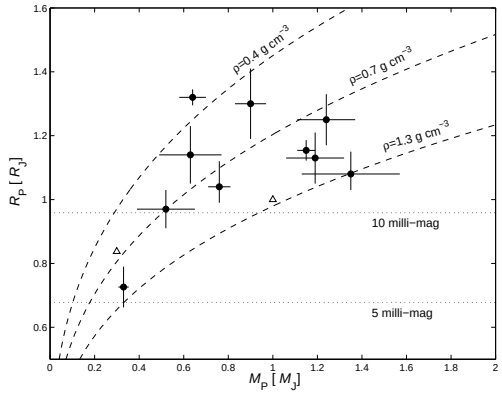


Fig. 1.6. Mass-radius relation for the known transiting planets. The triangles represent Jupiter and Saturn. The dashed lines are constant density lines, for densities of 0.4, 0.7 and 1.3 g cm^{-3}. The dotted line represent levels of photometric precision required for detecting planets orbiting Sun-like stars

1.5.1 Astrometry

In Sec. 1.2 we have shown that the star performs an orbital motion similar to that of the planet, but on a much smaller scale (Eq. 1.1). When we observe a star at a distance of d parsecs, we may be able to directly detect this orbital motion, scaled down, due to the distance, to a semi-major axis α:

$$\alpha = \frac{a_\star}{d} = \frac{a}{d}(\frac{M_\mathrm{p}}{M_\star}) \tag{1.6}$$

where α is measured in arc-seconds. The *astrometric signature*, α, is therefore proportional to both the planet mass and the orbital radius, unlike the RV semi-amplitude K, which is inversely related to the period (and therefore to the orbital radius). Astrometric techniques aim to detect planets by measuring this motion.

The first claim of an extrasolar planet detected by astrometry was published in 1969 by van de Kamp (1969). Van de Kamp made astrometric observations of Barnard's Star - one of the closest stars to the Solar System (at a distance of only 1.8 parsecs). He observed the star since 1938, and claimed that he had detected a planet around Barnard's Star, with a mass of 1.6 M_J and a period of 24 years. In the following years van de Kamp published a series of papers where he refined his findings and claimed to have detected not one, but two planets. Eventually,

van de Kamp's claims were refuted conclusively, e.g. by Benedict et al. (1999).

Astrometry does not suffer from the inclination ambiguity, and if an astrometric orbit is detected, the planet mass should be easily derived. The expected motion is extremely small, much below 1 milli-arcsecond, and detecting it from the ground requires developing new techniques to overcome the various noise sources. The most promising technique is interferometry, which was already proven to reach precisions of tens of micro-arcseconds in instruments in various stages of development, such as the Palomar Testbed Interferometer (Colavita & Shao 1994), the Keck Interferometer (Colavita et al. 1998), and ESO's VLTI (Mariotti et al. 1998). Several useful astrometric observations of known extrasolar planets were obtained using space-based observatories (Sec. 1.6).

1.5.2 Direct imaging

Various techniques aim at improving the chances to directly image extrasolar planets. These techniques include *coronagraphy*, which aims at suppressing the light from the star, *adaptive optics*, that try to compensate for atmospheric turbulence, and simply using longer wavelengths where the contrast between the star and the planet is more favorable. Combination of the two techniques seems a very promising path towards direct detection, e.g., the NICI instrument on the Gemini South telescope, which is currently being commissioned (Toomey & Ftaclas 2003).

Using NACO, an infrared adaptive optics instrument on the VLT, Chauvin et al. (2004) managed to image a giant planet orbiting the brown dwarf 2 M 1207 at an orbital distance of 55 AU. The planet is very faint, and its mass can only be inferred by comparing characteristics of its spectrum to models. Thus, Chauvin et al. estimate its mass at $5 \pm 1\,M_{\rm J}$. The large orbital distance implies a very long period, and therefore there is no dynamical measurement of the mass of 2 M 1207 b. This detection, being the first of its kind, like HD 209458, may open the gate to new kinds of observations which will be routinely performed on other planets that will be imaged directly.

1.5.3 Pulsar timing

Several planets were detected around pulsars a few years before the detection of 51 Peg b. In principle they were detected by radial velocities,

but those were not measured by usual spectroscopic means, but by the precise *timing* of the pulsar's pulses (Wolszczan & Frail 1992; Wolszczan 1994). Pulsars are neutron stars, i.e., they are no longer main-sequence stars but remnants of core-collapse supernova explosions. In light of the extremely violent process that forms pulsars, the existence of planets around them is very intriguing, but it is not the focus of our quest of seeking planets around main-sequence stars.

1.5.4 Gravitational Lensing

Due to a well-known phenomenon in General Relativity, when a faint relatively close star (the 'lens') passes in front of a very distant star (the 'source'), the light of the source undergoes a strong amplification. In case there is a planet orbiting the 'lens' star, it may be detected by its effect on the amplification curve (Sackett 1999). When only this magnification is observable, this phenomenon is commonly known as 'microlensing', to distinguish it from other lensing events which can be imaged in detail (e.g., lensing of galaxies by other galaxies). A few teams are routinely monitoring and collecting microlensing events, and there are already a few planet candidates detected this way (Bond et al. 2004), but follow-up studies of those candidates is not feasible because the lens stars are much too faint (we know they exist only by the lensing events). This renders the main contribution by these detections a statistical one, regarding the frequency of planets in the Galaxy.

1.6 Future prospects for space missions

In the very near future two space missions designed for detecting planetary transits are expected to be launched. Free from 'seeing' problems, caused by the atmospheric turbulences, those space telescopes are expected to monitor dense stellar fields in search of transiting planets. Those missions are the French-led COROT satellite, whose launch is scheduled for october 2006 (Bordé et al. 2003), and later the American satellite Kepler (Borucki et al. 2003) is expected to be launched around 2008. A technological precursor is the Canadian MOST satellite, which was designed for asteroseismology studies, but requires the same kind of photometric precision (Walker et al. 2003). Besides providing many detections of Jupiter-sized planets, these missions are also expected to detect earth-sized planet candidates. The main challenge ahead will be the RV follow-up, which will be extremely difficult for the small planets.

In the early 1990s, the astrometric satellite Hipparcos was launched by ESA. Hipparcos, operating for about three years, had an astrometric precision of the order of milli-arcseconds (ESA 1997). This was not enough for detecting the motion caused by any known extrasolar planet. It was enough, though, for putting upper limits on the planetary masses, thus proving the substellar nature of a few candidates (Pourbaix 2001; Pourbaix & Arenou 2001; Zucker & Mazeh 2001a). Two astrometric space missions are now planned that will hopefully reach the astrometric precision required for detection: the european Gaia mission is scheduled fo 2011 (Perryman et al. 2001) and the American Space Interferometry Mission (SIM) for 2015 (Shao 2004). Both missions are expected to reach astrometric precisions of the order of micro-arcseconds.

The prospects for the next decades are quite exciting – two missions are planned that will make use of innovative techniques to achieve the goal of imaging terrestrial planets! The European Space Agency is planning an infrared space interferometer – Darwin, that will comprise a flotilla of space telescopes (Fridlund 2004). The American counterpart by NASA is the TPF - Terrestrial Planet Finder (Beichman et al. 2004), which is now planned to include two separate missions – an interferometer (TPF-I) and a coronagraph (TPF-C). Both projects are extremely ambitious, both in terms of the technology required and in terms of their goal – to directly image terrestrial planets and study them for signs of Life.

Acknowledgments

I wish to thank the anonymous referee whose comments helped me improve this review. I am grateful to the Origins Institute at McMaster University, its director Ralph Pudritz, and its secretary Rosemary McNeice, for financial support and for organizing a productive and pleasant conference.

References

[1] Aitken, R.G. (1938). Is the solar system unique? *Astron. Soc. Pac. Leaflet*, **112**, 98–106.

[2] Alonso, R., Brown, T.M., Torres, G., Latham, D.W., Sozzetti, A., Mandushev, G., Belmonte, J.A., Charbonneau, D., Deeg, H.J., Dunham, E.W., O'Donovan, F.T. & Stefanik, R.P. (2004). TrES-1: the transiting planet of a bright K0V star. *Astrophys. J.*, **613**, L153–L156.

[3] Bakos, G., Noyes, R.W., Kovács, G., Stanek, K.Z., Sasselov, D.D. & Domsa,

I. (2004). Wide-field millimagnitude photometry with the HAT: a tool for extrasolar planet detection. *Publ. Astron. Soc. Pac.*, **116**, 266–277.

[4] Baranne, A., Queloz, D., Mayor, M., Adrianzyk, G., Knispel, G., Kohler, D., Lacroix, D., Meunier, J.-P., Rimbaud, G. & Vin, A. (1996). ELODIE: a spectrograph for accurate radial velocity measurements. *Astron. Astrophys. Sup.*, **119**, 373–390.

[5] Beichman, C., Gómez, G., Lo, M., Masdemont, J. & Romans, L. (2004). Searching for life with the Terrestrial Planet Finder: Lagrange point options for a formation flying interferometer. *Adv. Space. Res.*, **34**, 637–644.

[6] Benedict, G.F., McArthur, B., Chappell, D.W., Nelan, E., Jefferys, W.H., van Altena, W., Lee, J., Cornell, D., Shelus, P.J., Hemenway, P.D., Franz, O.G., Wasserman, L.H., Duncombe, R.L., Story, D., Whipple, A.L. & Fredrick, L.W. (1999). Interferometric astrometry of Proxima Centauri and Barnard's star using Hubble Space Telescope Fine Guidance Sensor 3: detection limits for substellar companions. *Astron. J.*, **118**, 1086–1100.

[7] Bond, I.A., Udalski, A., Jaroszyński, M., Rattenbury, N.J., Paczyński, B., Soszyński, I., Wyrzykowski, Ł., Szymański, M.K., Kubiak, M., Szewczyk, O., Żebruń, K., Pietrzyński, G., Abe, F., Bennett, D.P., Eguchi, S., Furuta, Y., Hearnshaw, J.B., Kamiya, K., Kilmartin, P.M., Kurata, Y., Masuda, K., Matsubara, Y., Muraki, Y., Noda, S., Okajima, K., Sako, T., Sekiguchi, T., Sullivan, D.J., Sumi, T., Tristram, P.J., Yanagisawa, T. & Yock, P.C.M. (2004). OGLE 2003-BLG-235/MOA 2003-BLG-53: a planetary microlensing event. *Astrophys. J.*, **606**, L155–L158.

[8] Bordé, P., Rouan, D. & Léger, A. (2003). Exoplanet detection capability of the COROT space mission. *Astron. Astrophys.*, **405**, 1137–1144.

[9] Borucki, W.J., Koch, D.G., Lissauer, J.J., Basri, G.B., Caldwell, J.F., Cochran, W.D., Dunham, E.W., Geary, J.C., Latham, D.W., Gilliland, R.L., Caldwell, D.A., Jenkins, J.M. & Kondo, Y. (2003). The Kepler mission: a wide-field photometer designed to determine the frequency of Earth-size planets around solar-like stars. *Soc. Photo-Opt. Instru.*, **4854**, 129–140.

[10] Brown, T.M., Charbonneau, D., Gilliland, R.L., Noyes, R.W. & Burrows, A. (2001). Hubble Space Telescope time-series photometry of the transiting planet of HD 209458. *Astrophys. J.*, **552**, 699–709.

[11] Burke, C.J., Gaudi, B.S., DePoy, D.L., Pogge, R.W. & Pinsonneault, M.H. (2004). Survey for transiting extrasolar planets in stellar systems I. fundamental parameters of the open cluster NGC 1245. *Astron. J.*, **127**, 2382–2397.

[12] Burrows, A., Marley, M., Hubbard, W.B., Lunine, J.I., Guillot, T., Saumon, D., Freedman, R., Sudarsky, D. & Sharp, C. (1997). A nongray theory of extrasolar giant planets and brown dwarfs. *Astrophys. J.*, **491**, 856–875.

[13] Butler, R.P., Marcy, G.W., Fischer, D.A., Brown, T.M., Contos, A.R., Korzennik, S.G., Nisenson, P. & Noyes, R.W. (1999). Evidence for multiple companions to upsilon Andromedae. *Astrophys. J.*, **526**, 916–927.

[14] Butler, R.P., Marcy, G.W., Williams, E., McCarthy, C. & Vogt, S.S. (1996). Attaining Doppler precision of $3\,\mathrm{m\,s^{-1}}$. *Publ. Astron. Soc. Pac.*, **108**, 500–509.

[15] Campbell, B. & Walker, G.A.H. (1979). Precision radial velocities with an absorption cell. *Publ. Astron. Soc. Pac.*, **91**, 540–545.

[16] Campbell, B., Walker, G.A.H. & Yang, S. (1988). A search for substellar

companions to solar-type stars. *Astrophys. J.*, **331**, 902–921.
[17] Chabrier, G. & Baraffe, I. (2000). Theory of low-mass stars and substellar objects. *Annu. Rev. Asron. Astr.*, **38**, 337–377.
[18] Charbonneau, D., Brown, T.M., Latham, D.W. & Mayor, M. (2000). Detection of planetary transits across a sun-like star. *Astrophys. J.*, **529**, L45–L48.
[19] Chauvin, G., Lagrange, A.-M., Dumas, C., Zuckerman, B., Mouillet, D., Song, I., Beuzit, J.-L. & Lowrance, P. (2004). A giant planet candidate near a young brown dwarf. Direct VLT/NACO observations using IR wavefront sensing. *Astron. Astrophys.*, **425**, L29–L32.
[20] Colavita, M.,M., Boden, A.F., Crawford, S.L., Meinel, A.B., Shao, M., Swanson, P.N., van Belle, G.T., Vasisht, G., Walker, J.M., Wallace, J.K. & Wizinowich, P.L. (1998). Keck interferometer. *Soc. Photo-Opt. Instru.*, **3350**, 776–784.
[21] Colavita, M.M. & Shao, M. (1994). Indirect planet detection with ground-based long-baseline interferometry. *Astrophys. Space. Sci*, **212**, 385–390.
[22] Cumming, A., Marcy, G.W. & Butler, R.P. (1999). The Lick planet search: detectability and mass thresholds. *Astrophys. J.*, **526**, 890–915.
[23] Drake, A.J. (2003). On the selection of photometric planetary transits. *Astrphys. J.*, **589**, 1020–1026.
[24] ESA (1997). *The Hipparcos and Tycho Catalogues*, ESA SP-1200. Noordwijk: ESA.
[25] Fridlund, C.V.M. (2004). The Darwin mission. *Adv. Space. Res.*, **34**, 613–617.
[26] Guillot, T. (2005). The interiors of giant planets: models and outstanding questions. *Annu. Rev. Earth. Pl. Sc.*, **33**, 493–530.
[27] Halbwachs, J.L., Arenou, F., Mayor, M., Udry, S. & Queloz, D. (2000). Exploring the brown dwarf desert with Hipparcos. *Astron. Astrophys.*, **355**, 581–594.
[28] Henry, G.W., Marcy, G.W., Butler, R.P. & Vogt, S.S. (2000). A transiting "51 Peg-like" planet. *Astrophys. J.*, **529**, L41–L44.
[29] Horne, K. (2006). http://star-www.st-and.ac.uk/~kdh1/transits/table.html
[30] Jorissen, A., Mayor, M. & Udry, S. (2001). The distribution of exoplanet masses. *Astron. Astrophys.*, **379**, 992–998.
[31] Latham, D.W., Mazeh, T., Stefanik, R.P., Mayor, M. & Burki, G. (1989). The unseen companion of HD 114762: a probable brown dwarf. *Nature*, **339**, 38–40.
[32] Laughlin, G., Aaron, W., Vanmunster, T., Bodenheimer, P., Fischer, D.A., Marcy, G.W., Butler, R.P. & Vogt, S.S. (2005). A comparison of observationally determined radii with theoretical radius predictions for short-period transiting extrasolar planets. *Astrophys. J.*, **621**, 1072–1078.
[33] Lin, D.N.C., Bodenheimer, P. & Richardson, D.C. (1996). Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature*, **380**, 606–607.
[34] Marcy, G.W. & Butler, R.P. (1996). A planetary companion to 70 Virginis. *Astrophys. J.*, **464**, L147-L151.
[35] Marcy, G.W. & Butler, R.P. (1998). Detection of extrasolar giant planets. *Annu. Rev. Asron. Astr.*, **36**, 57–98.
[36] Marcy, G.W. & Butler, R.P. (2000). Planets orbiting other suns. *Publ. Astron. Soc. Pac.*, **112**, 137–140.
[37] Marcy, G.W., Butler, R.P., Williams, E., Bildsten, L., Graham, J.R., Ghez,

A.M. & Jernigan, J.G. (1997). The planet around 51 Pegasi. *Astrophys. J.*, **481**, 926–935.

[38] Mariotti, J.-M., Denise, C., Derie, F., Ferrari, M., Glindemann, A., Koehler, B., Lévêque, S.A., Paresce, F., Schoeller, M., Tarenghi, M. & Verola, M. (1998). VLTI program: a status report. *Soc. Photo-Opt. Instru.*, **3350**, 800–806.

[39] Mayor, M. & Queloz, D. (1995). A Jupiter-mass companion to a solar-type star. *Nature*, **378**, 355–359.

[40] Mazeh, T., Latham, D.W. & Stefanik, R.P. (1996). Spectroscopic orbits for three binaries with low-mass companions and the distribution of secondary masses near the substellar limit. *Astrophys. J.*, **466**, 415–426.

[41] Mazeh, T., Naef, D., Torres, G., Latham, D.W., Mayor, M., Beuzit, J.-L., Brown, T.M., Buchhave, L., Burnet, M., Carney, B.W., Charbonneau, D., Drukier, G.A., Laird, J.B., Pepe, F., Perrier, C., Queloz, D., Santos, N.C., Sivan, J.-P., Udry, S. & Zucker, S. (2000). The spectroscopic orbit of the planetary companion transiting HD 209458. *Astrophys. J.*, **532**, L55–L58.

[42] Murray, N. & Chaboyer, B. (2002). Are stars with planets polluted? *Astrophys. J.*, **566**, 442–451.

[43] Naef, D., Latham, D.W., Mayor, M., Mazeh, T., Beuzit, J.-L., Drukier, G.A., Perrier-Bellet, C., Queloz, D., Sivan, J.-P., Torres, G., Udry, S. & Zucker, S. (2001). HD 80606 b, a planet on an extremely elongated orbit. *Astron. Astrophys.*, **375**, L27–L30.

[44] Nelson, R.P., Papaloizou, J.C.B., Masset, F. & Kley, W. (2000). The migration and growth of protoplanets in protostellar discs. *Mon. Not. R. Astron. Soc.*, **318**, 18–36.

[45] Pätzold, M. & Rauer, H. (2002). Where are the massive close-in extrasolar planets? *Astrophys. J.*, **568**, L117–L120.

[46] Perryman, M.A.C., de Boer, K.S., Gilmore, G., Høg, E., Lattanzi, M.G., Lindegren, L., Luri, X., Mignard, F. & Pace, O. (2001). Gaia: composition, formation and evolution of the Galaxy. *Astron. Astrophys.*, **369**, 339–363.

[47] Pont, F., Melo, C.H.F., Bouchy, F., Udry, S., Queloz, D., Mayor, M. & Santos, N.C. (2005). A planet-sized star around OGLE-TR-122. Accurate mass and radius near the hydrogen-burning limit. *Astron. Astrophys.*, **433**, L21–L24.

[48] Pourbaix, D. (2001). The Hipparcos observations and the mass of substellar objects. *Astron. Astrophys.*, **369**, L22–L25.

[49] Pourbaix, D. & Arenou, F. (2001). Screening the Hipparcos-based astrometric orbits of sub-stellar objects. *Astron. Astrophys.*, **372**, 935–944.

[50] Sackett, P.D. (1999). Searching for unseen planets via occultation and microlensing, in *Planets Outside the Solar System: Theory and Observations* (NATO-ASI), ed. J.-M. Mariotti & D. Alloin. Dordrecht: Kluwer, pp. 189–227.

[51] Santos, N.C., Israelian, G. & Mayor, M. (2004). Spectroscopic [Fe/H] for 98 extra-solar planet-host stars. *Astron. Astrophys.*, **415**, 1153–1166.

[52] Schneider, J. (2006). *The Extrasolar Planet Encyclopaedia* http://exoplanet.eu/index.php

[53] Seager, S. & Mallén-Ornellas, G. (2003). A unique solution of planet and star parameters from an extrasolar planet transit light curve. *Astrophys. J.*, **585**, 1038–1055.

[54] Shao, M. (2004). Science overview and status of the SIM project. *Soc. Photo-Opt. Instru.*, **5491**, 328–333.
[55] Struve, O. (1952). Proposal for a project of high-precision stellar radial velocity work. *Observatory*, **72**, 199–200.
[56] Tingley, B. (2004). Using color photometry to separate transiting exoplanets from false positives. *Astron. Astrophys.*, **425**, 1125–1131.
[57] Toomey, D.W. & Ftaclas, C. (2003). Near infrared coronagraphic imager for Gemini South. *Soc. Photo-Opt. Instru.*, **4841**, 889–900.
[58] Trilling, D.E., Benz, W., Guillot, T., Lunine, J.I., Hubbard, W.B. & Burrows, A. (1998). Orbital evolution and migration of giant planets: modeling extrasolar planets. *Astrophys. J.*, **500**, 428–439.
[59] Trilling, D.E., Lunine, J.I. & Benz, W. (2002). Orbital migration and the frequency of giant planet formation. *Astron. Astrophys.*, **394**, 241–251.
[60] Udalski, A., Paczyński, B., Żebruń, K., Szymański, M., Kubiak, M., Soszyński, I., Szewczyk, O., Wyrzykowski, Ł. & Pietrzyński, G. (2002). The Optical Gravitational Lensing Experiment. Search for planetary and low-luminosity object transits in the galactic disk. Results of 2001 campaign. *Acta Astron.*, **52**, 1–37.
[61] Udalski, A., Szymański, M., Kubiak, M., Pietrzyński, G., Soszyński, I., Żebruń, K., Szewczyk, O. & Wyrzykowski, Ł. (2004). The Optical Gravitational Lensing Experiment. Planetary and low-luminosity object transits in the fields of galactic disk. Results of the 2003 OGLE observing campaigns. *Acta Astron.*, **54**, 313–345.
[62] Udry, S., Eggenberger, A., Mayor, M., Mazeh, T. & Zucker, S. (2004). Planets in multiple-star systems: properties and detections, in *The Environment and Evolution of Double and Multiple Stars*, Proc. IAU Coll. 191, ed. C. Allen & C. Scarfe. *Rev. Mex. Astron. Astrof. Ser. Conf.*, **21**, pp. 207–214
[63] Udry, S., Mayor, M., Naef, D., Pepe, F., Queloz, D., Santos, N.C. & Burnet, M. (2002). The CORALIE survey for southern extra-solar planets. VIII. The very low-mass companions of HD 149137, HD 162020, HD 168433 and HD 202206: brown dwarfs or "superplanets"? *Astron. Astrophys.*, **390**, 267–279.
[64] van de Kamp, P. (1969). Parallax, proper motion, acceleration, and orbital motion of Barnard's Star. *Astron. J.*, **74**, 238–240.
[65] Walker, G., Matthews, J., Kuschnig, R., Johnson, R., Rucinski, S., Pazder, J., Burley, G., Walker, A., Skaret, K., Zee, R., Grocott, S., Carroll, K., Sinclair, P., Sturgeon, D. & Harron, J. (2003). The MOST asteroseismology mission: ultraprecise photometry from space. *Publ. Astron. Soc. Pac.*, **115**, 1023–1035.
[66] Wolszczan, A. (1994). Confirmation of Earth-mass planets orbiting the millisecond pulsar PSR B 1257+12. *Science*, **264**, 538–542.
[67] Wolszczan, A. & Frail, D.A. (1992). A planetary system around millisecond pulsar PSR 1257+12. *Nature*, **355**, 145–147.
[68] Zucker, S. & Mazeh, T. (2001a). Analysis of the Hipparcos observations of the extrasolar planets and the brown dwarf candidates. *Astrophys. J.*, **562**, 549–557.
[69] Zucker, S. & Mazeh, T. (2001b). Derivation of the mass distribution of extrasolar planets with MAXLIMA, a maximum likelihood algorithm. *Astrophys. J.*, **562**, 1038–1044.
[70] Zucker, S. & Mazeh, T. (2002). On the mass-period correlation of the

extrasolar planets. *Astrophys. J.*, **568**, L113–L116.

2

The Atmospheres of Extrasolar Planets

L. Jeremy Richardson
NASA Goddard Space Flight Center

Sara Seager
Carnegie Institution of Washington

2.1 Introduction

In this chapter we examine what can be learned about extrasolar planet atmospheres by concentrating on a class of planets that ***transit*** their parent stars. As discussed in the previous chapter, one way of detecting an extrasolar planet is by observing the drop in stellar intensity as the planet passes in front of the star. A transit represents a special case in which the geometry of the planetary system is such that the planet's orbit is nearly edge-on as seen from Earth. As we will explore, the transiting planets provide opportunities for detailed follow-up observations that allow physical characterization of extrasolar planets, probing their bulk compositions and atmospheres.

2.2 The Primary Eclipse

The vast majority of the currently-known extrasolar planets have been detected using the radial velocity technique.† As detailed in the previous chapter, the radial velocity method searches for periodic motion of a star caused by the gravitational pull of an orbiting companion. Figure 1.1 shows a sketch of a typical periodic radial velocity signal and the basic geometry of the planetary system. This method is sensitive only to movement of the star towards and away from the observer, that is, along the line of sight from the system to the observer on Earth. Thus, radial velocity observations provide only a determination of the ***minimum*** mass M of the planet, and the orbital inclination i of the system remains

† An up-to-date reference and catalog of all known extrasolar planets can be found at `http://vo.obspm.fr/exoplanetes/encyclo/encycl.html`

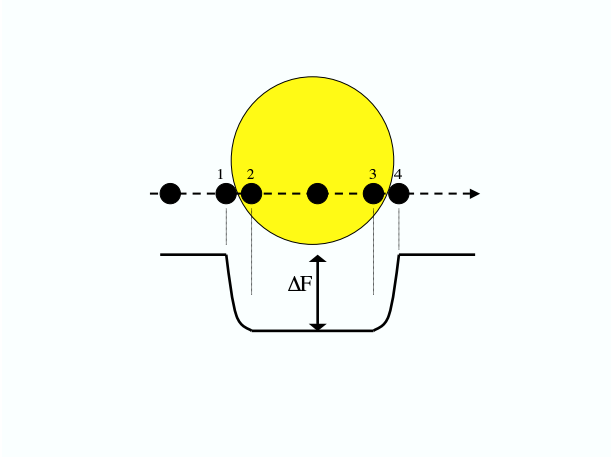


Fig. 2.1. Sketch showing a planet crossing the disk of its parent star. Transit light curve shown below.

unknown, as in

$$M = M_p \sin i, \tag{2.1}$$

where M_p is the ***true*** mass of the planet. (See Section 1.2 for further details.)

The ***primary eclipse***, or transit, occurs when the planet's orbit happens to be nearly edge-on as seen from Earth. This means that the planet periodically crosses in front of the star as it orbits, and we detect this as a decrease in the light from the star that occurs once per planet revolution, as indicated schematically in Figure 2.1. The dimming is typically a few percent or less for the currently known transiting planets. In this geometry, the orbital inclination is now known to be $\sim 90°$ (and can be determined precisely from the details of the transit light curve). We can therefore derive the true planetary mass, M_p, from Equation 2.1.

A number of other physical parameters of the planet and star can be derived from the shape of the light curve (115). The depth of the transit ΔF (i.e., the change in flux from outside transit to during transit, as shown in Figure 2.1) is directly proportional to the ratio of the area of

the planetary disk to the area of the stellar disk. That is,

$$\Delta F \equiv \frac{F_{\text{out of transit}} - F_{\text{transit}}}{F_{\text{out of transit}}} = \frac{A_p}{A_*} = \left(\frac{R_p}{R_*}\right)^2, \tag{2.2}$$

where F represents the total flux, A is the area of the disk (planet or star), and R is the radius (planet or star). With a sufficiently precise transit curve, it is possible to derive both the planetary and stellar radii simultaneously. With the planetary mass M_p and radius R_p, one can immediately calculate the average density of the planet from

$$\rho = \frac{M_p}{\frac{4}{3}\pi R_p^3}. \tag{2.3}$$

The discovery of transiting planets allowed a direct measurement of the true mass, radius, and density of planets outside the solar system for the first time. The planetary radius is key to determining the reflection and thermal emission of the planets from flux measurements. The density measurements derived from transit observations indicate that all but one of the transiting planets are hydrogen-helium gas giants, similar in bulk composition to Jupiter and Saturn in our own Solar System.

When the planet is in front of the star, the planet's atmosphere appears as an annulus surrounding the planetary disk, and some of the starlight passes through this annulus to the observer. The detection of starlight that has passed through the transiting planet's atmosphere in this manner is called ***transmission spectroscopy***. By measuring how much starlight is transmitted as a function of wavelength, we can learn about the atomic and molecular species present in the planet's atmosphere, providing a much greater wealth of information than simply the average density and bulk composition. We introduce the broad study of spectroscopy in Section 2.5 and discuss recent observations of transiting planets using transmission spectroscopy in Section 2.7.

2.3 The Secondary Eclipse

A planet that crosses in front of its parent star will disappear behind the star later in its orbit. This disappearance is called the ***secondary eclipse***. For a circular orbit, the secondary eclipse occurs exactly one-half of an orbital period after the primary eclipse. However, for a non-circular orbit, the secondary eclipse can occur earlier or later (depending on the eccentricity and the orientation of the orbit), and its duration can differ from that of the primary eclipse (82). In addition to clues about

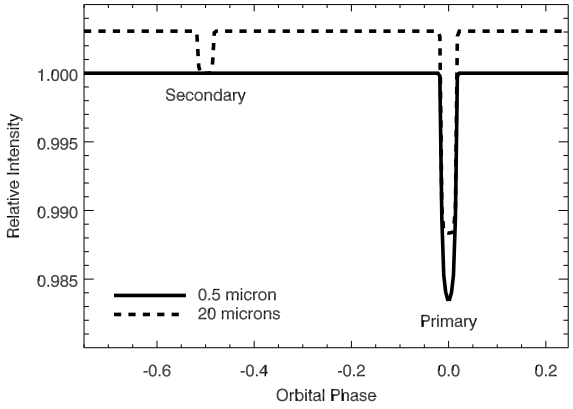


Fig. 2.2. Comparison of primary and secondary eclipses in the visible and infrared for the thermal emission of the planet HD 209458 b. These curves were calculated from a simple model that assumes the star and planet emit blackbody radiation only.

the eccentricity of the planet's orbit from the secondary eclipse timing and duration, the secondary eclipse yields information about the nature of the planet's atmosphere.

For example, in visible light, the secondary eclipse probes the amount of starlight reflected by the planet's atmosphere (called the ***albedo***). In the infrared, however, it measures the direct thermal emission (or intrinsic heat output) of the planet. In neither case does this imply imaging the planet; rather, the idea is to observe the total energy output of the system (star + planet) and attempt to detect a decrease as the planet is hidden from view.

Figure 2.2 illustrates this decrease in the total energy output of the system during secondary eclipse and shows that the thermal emission of the planet may detectable at infrared wavelengths using this technique. The basic situation is that the incident starlight (which peaks in the visible for a Sun-like star) is absorbed and reprocessed by the planet's atmosphere, and some of that radiation is later emitted at infrared wavelengths. The figure shows the thermal emission of the planet HD 209458 b relative to its parent star. This calculation assumes that both the star and planet emit only blackbody radiation (Equation 2.6), and it assumes that the planet emits uniformly in both hemispheres. In

the visible region (solid curve), the secondary eclipse is undetectable, both because the planet has virtually no emission at these wavelengths and the reflected light from the planet is $\ll 0.01\%$ of the stellar output. However, as the figure shows, the situation is quite different at 20 . The total intensity ***relative to the star*** is higher outside of the eclipse, because the planet has a small but measurable intrinsic energy output at this wavelength. The secondary eclipse appears as a dip of $\sim 0.3\%$ in the total intensity as the planet is hidden by the star.

The eclipse depths at visible and infrared wavelengths can be estimated with the following flux ratios. For reflected light,

$$\frac{F_p}{F_*} = A_g \left(\frac{R_p}{a} \right)^2 , \tag{2.4}$$

where A_g is the geometric albedo (the fraction of incident radiation scattered back into space when the planet is in full phase.), R_p is the planetary radius, and a is the orbital semi-major axis. For thermal emission,

$$\frac{F_p}{F_*} = \frac{T_p}{T_*} \left(\frac{R_p}{R_*} \right)^2 , \tag{2.5}$$

where T_p and T_* are the planet and star effective temperatures (see equation 2.19 for an estimate of T_p). Here we have used the approximation for the Wien tail of the blackbody flux whereby the flux ratio translates into a temperature ratio.

2.4 Characteristics of Known Transiting Planets

A total of ***ten*** transiting extrasolar planets have been discovered as of May 2006. Their physical characteristics are given in Table 2.1, and they are plotted in Figure 2.3. The upper panel (period vs. mass) illustrates the two groups of transiting planets. The 'hot Jupiters' (to the upper left of the plot) have masses smaller than that of Jupiter and orbital periods greater than ~ 2.5 days. This name is something of a misnomer, since the so-called hot Jupiters are quite different from our own Jupiter—because of the fact that they orbit at such small orbital distances, they are much hotter and therefore have different chemical species present in their atmospheres. The other group, often called the 'very hot Jupiters,' is characterized by planets that orbit much closer to their parent stars (with orbital periods less than 2.5 days) and are much more massive than Jupiter. These two dynamically distinct groups

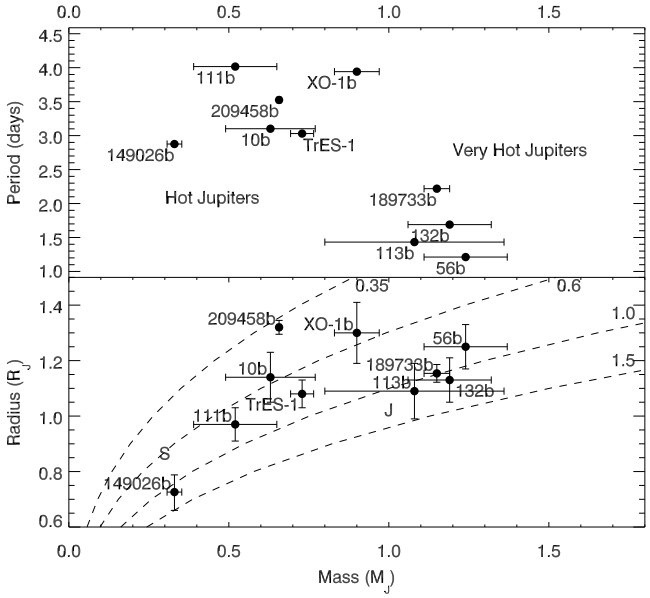


Fig. 2.3. The ten currently-known transiting planets, as a function of planetary mass. The upper panel shows the period vs. mass, and the lower panel shows radius vs. mass. Dashed curves indicate curves of constant density in g/cm^3. For comparison, Jupiter and Saturn are shown, marked by 'J' and 'S', respectively.

of planets may have different evolutionary histories, possibly resulting from different migration mechanisms (94), and thus could potentially have very different atmospheric properties. We now have one bright planet from each group—HD 209458 b and HD 198733 b—allowing us to compare observations of the two planets and gain insights into their atmospheric structure and evolutionary history.

The lower panel of Figure 2.3 shows the radius of each planet vs. mass. The dashed curves indicating constant density provide context for understanding the bulk composition of the planets. For example, most of the transiting planets are similar in average density to Jupiter ($\rho = 1.33$ g/cm^2) and Saturn ($\rho = 0.69$ g/cm^2). However, these 'close-in' extrasolar planets are likely to be quite different from our own Jupiter, due to the fact that they are much closer in to their stars. At orbital distances of $a < 0.1$ AU (by comparison, Mercury is at $a \sim 0.38$ AU), these planets are bombarded by radiation from their parent star and

Table 2.1. *Physical properties of transiting extrasolar planets.*

Planet	**Period (days)**	**Radius (R_J)**	**Mass (M_J)**	**T_* (K)**	**T_{eq} [1] (K)**
OGLE-TR-56b[2,3]	1.212	1.25 ± 0.08	1.24 ± 0.13	6119	1929
OGLE-TR-113b[2,4]	1.432	1.09 ± 0.10	1.08 ± 0.28	4804	1234
OGLE-TR-132b[5,6]	1.690	1.13 ± 0.08	1.19 ± 0.13	6411	1933
HD 189733 b[7,8]	2.219	1.154 ± 0.032	1.15 ± 0.04	5050	1096
HD 149026 b[9,10]	2.876	$0.726\ ^{+0.062}_{-0.066}$	0.33 ± 0.023	6147	1593
TrES-1[2,11,12]	3.030	1.08 ± 0.05	0.729 ± 0.036	5226	1059
OGLE-TR-10b[2,13,14]	3.101	1.14 ± 0.09	0.63 ± 0.14	6075	1402
HD 209458 b[2,15,16]	3.525	1.320 ± 0.025	0.657 ± 0.006	6117	1363
XO-1b[17]	3.942	1.30 ± 0.11	0.90 ± 0.07	5750	1148
OGLE-TR-111b[2,18]	4.016	0.97 ± 0.06	0.52 ± 0.13	5044	935

[1]Calculated from Equation 2.19 with $f = 1$ and $A_B = 0.3$; [2] (113); [3] (119); [4] (98); [5] (76); [6] (106); [7] (77); [8] (73); [9] (114); [10] (85); [11] (100); [12] (72); [13] (99); [14] (96); [15] (97); [16] (122); [17] (104); [18] (107);

are therefore expected to be hot ($T > 1000$ K). Jupiter, at 5 AU from the Sun, has a blackbody temperature of only 110 K. Because of the large temperature difference, we expect the atmospheric composition of the hot Jupiters to be significantly different from that of Jupiter. For example, at low temperatures ($T < 1000$ K), chemical equilibrium calculations show that carbon is mostly present in the form of CH_4, while at higher temperatures it appears as CO (78).

Finally, from Figure 2.3, we note that not all of the known transiting planets have densities similar to Jupiter and Saturn. The most extreme example is HD 209458 b, with an average density of ~ 0.35 g/cm^2. This planet was the first one found to exhibit a transit (83; 95), although it was originally detected using the radial velocity method (103). Since then, it has been observed extensively from the ground and from space, as we shall discuss further in Section 2.7. From the beginning, it was unclear why the planet appears to have a larger radius than is predicted by theory (79; 80), and this remains one of the unanswered questions in the field today (121).

2.5 Spectroscopy

We now turn from a general discussion of the transiting planets to the specific topic of spectroscopy and the radiative transfer equation. By studying the spectroscopy of extrasolar planets, we can gain key insights into the atmospheric composition, temperature, and structure of these planets. We begin this section by introducing the Planck blackbody law, which describes the thermal emission of an object in the absence of scattering or absorbing particles, and move to the radiative transfer equation, which does account for the effects of scattering and absorption. The radiative transfer equation governs the interaction of energy (in the form of emitted or absorbed radiation) with matter (in this case the particles that make up the planetary atmosphere). In Section 2.6 we give an overview of how models of planetary atmospheres are computed. Spectra derived from such models help interpret observational results and facilitate the design of future planet atmosphere detection instruments.

At the most basic level we can approximate the star and planet as blackbodies. In that case, where we ignore the details of the atmosphere, the emission is given by Planck's blackbody law:

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5(e^{hc/\lambda kT} - 1)}, \tag{2.6}$$

where k is Boltzmann's constant, T is the temperature of the blackbody, and h is Planck's constant. For a stellar or planetary atmosphere which contains a variety of different species that absorb, emit, and scatter radiation, however, the blackbody law is not sufficient to describe the resulting spectrum, and it becomes necessary to understand how matter interacts with the radiation.

Following convention, we begin by considering a pencil of radiation traveling through a medium. The energy in the beam is given by

$$dE_\nu = I_\nu(\mathbf{r}, \hat{\mathbf{n}}, t) \cos\theta \, dA \, d\Omega \, d\nu \, dt, \tag{2.7}$$

where I_ν is the monochromatic (spectral) ***radiance***—sometimes incorrectly called the intensity—which depends on the position vector $\mathbf{r}$ and the direction $\hat{\mathbf{n}}$, θ is the angle from the normal to the surface, and Ω is the solid angle in steradians.

Next, we explicitly describe how the radiation beam changes as it interacts with matter (of density ρ), traveling through a distance ds:

$$dI_\nu = -k_\nu \, \rho \, I_\nu \, ds + j_\nu \, \rho \, ds. \tag{2.8}$$

The first term on the right hand side represents the amount of radiation removed from the beam (extinction cross section k_ν) and the second term represents the amount of radiation added to the beam (emission cross section j_ν). Defining the source function S_ν as the ratio of the emission cross section to the extinction cross section, we have

$$\frac{dI_\nu}{\rho\, k_\nu\, ds} = -I_\nu + S_\nu. \tag{2.9}$$

This is the ***radiative transfer equation*** and it governs the fundamental physics at work in the atmosphere. The simplistic form of the radiative transfer equation hides its true complexity. The main problem lies in the nonlinearity of the equation. The solution of I_ν depends on j_ν, but if there is scattering in the atmosphere j_ν also depends on I_ν. A second problem lies in the definition of the source function $S_\nu = j_\nu/k_\nu$. The opacities that make up k_ν and j_ν can be composed of millions of lines for molecular species, and in the case of cloud opacities can involve a number of free parameters.

Finally, we can write the form of Equation 2.9 in a more conventional form by making a few more definitions. If we consider a plane-parallel atmosphere, we are interested only in radiation flowing in the vertical direction. Note that

$$ds = \cos\theta\, dz, \tag{2.10}$$

and we can define

$$\mu = \cos\theta \tag{2.11}$$

where θ is the angle measured from the vertical, or the zenith angle. We can now define the ***optical depth*** τ as

$$\tau_\nu(z) = -\int_z k_\nu(z)\, \rho(z)\, dz. \tag{2.12}$$

The minus sign appears because the optical depth is by convention measured from the top of the atmosphere increasing downward. Using the definition of the optical depth, we can rewrite Equation 2.9 as

$$\mu \frac{dI_\nu}{d\tau} = I_\nu - S_\nu. \tag{2.13}$$

The detailed solution of this equation is beyond the scope of this work, and we refer the interested reader to more comprehensive works that describe the solution and application of the radiative transfer equation (e.g., 105; 101; 111).

Fortunately, under specific assumptions, the solution to Equation 2.13

becomes simple. As discussed in Section 2.2, during transit the planet passes in front of the star, and some starlight passes through the annulus of the planetary atmosphere before reaching the observer. At visible wavelengths (where the thermal emission is negligible) the starlight is attenuated by the absorbing gases in the planet atmosphere. In this case, we take the emission, and thus the source function S_ν, to be zero, and Equation 2.13 reduces to

$$\mu \frac{dI_\nu}{d\tau} = I_\nu, \tag{2.14}$$

which can easily be integrated to obtain

$$I_\nu(z) \, = \, I_\nu(z{=}0)\, e^{-\tau_\nu(z)/\mu}. \tag{2.15}$$

This equation is known as Beer's Law or Lambert's Law (e.g., 101). It describes the dissipation of radiation as it travels through a medium. Because atoms and molecules absorb at specific wavelengths, the amount of starlight that is trasmitted through the planetary atmosphere changes with wavelength.

Another physical situation with a simple solution to the radiative transfer equation is the case of thermal emission and no scattering. This situation would hold at infrared wavelengths if clouds (i.e. scattering particles) were not present. In this case of thermal emission, the source function is simply the blackbody function:

$$S_\nu = B_\nu \tag{2.16}$$

The radiative transfer equation (2.13) then reduces to a linear form:

$$\mu \frac{dI_\nu}{d\tau} = I_\nu - B_\nu. \tag{2.17}$$

The solution is

$$I_\nu(z) \, = \int_0^\pi \frac{1}{\mu} \int_0^\infty B_\nu(\tau) \exp^{-\tau_\nu(z)/\mu} \, d\tau d\mu. \tag{2.18}$$

With a given vertical temperature and pressure profile, the opacities and hence B_ν can be computed, and the right hand side of the above equation is straightforward to integrate.

2.6 Model Atmospheres

A full model atmosphere computation is needed to understand the details of the planetary spectrum. Usually the models assume that the

planetary atmosphere is one-dimensional and plane-parallel (no curvature). The models produce the temperature and pressure as a function of altitude and the radiation field (that is, the emergent flux from the atmosphere) as a function of altitude and wavelength (see 117, and references therein). To derive these three quantities, three equations are solved: the radiative transfer equation (Equation 2.13), the equation of hydrostatic equilibrium, and the radiative and convective equilibrium. The boundary conditions are the incident stellar radiation at the top of the atmosphere and the interior energy (assumed) at the bottom of the atmosphere. With this type of calculation, only the planetary surface gravity and the incident stellar radiation are known with certainty. Although the physics governing the model is relatively simple, a number of assumptions are necessary in order for the calculation to proceed, including (117; 102):

- atmospheric chemistry (including elemental abundances, nonequilibrium chemistry, and photochemistry);
- cloud properties;
- atmospheric circulation;
- internal heat flow; and
- gaseous opacities.

Most published model results have typically used solar elemental abundances (i.e., having the same relative concentrations as the Sun.) Of course, this assumption is limited in that different stars will have relative abundances different from the Sun. Even the relative abundances of the elements in our Sun remain somewhat uncertain. More importantly, the solar system giant planets are enriched in carbon, and Jupiter and Saturn are also enriched in nitrogen relative to solar (see Marley *et al.* (102) and references therein). With the assumed elemental abundances, chemical equilibrium calculations determine the abundances of the different atomic, molecular, and liquid or solid species as a function of temperature and pressure. For example, given the elemental abundances of carbon and oxygen relative to hydrogen, the relative concentrations of methane (CH_4) and carbon monoxide (CO) can be computed. CH_4 and CO are particularly interesting molecules for the hot Jupiters (116) because it is unclear which one is the dominant form of carbon due to the uncertain temperatures and metallicities of the hot Jupiters. At higher temperatures and higher C to O ratios, we expect CO to be the dominant form of carbon, while at lower temperatures CH_4 is the dominant form of carbon.

With the computed abundances of chemical species, the opacities can be determined. The ***opacity*** represents the amount of radiation that a given species can absorb as a function of wavelength. The opacities of the expected chemical species in the model atmosphere play a pivotal role in determining the structure of the resulting spectrum. In particular, water, methane, ammonia (NH_3), sodium, and potassium all have significant spectral signatures for gas giant planets and are expected to be present in the atmospheres of these planets. Opacities are particularly sensitive to choices of metallicity, which species (atomic and molecular) are included, and whether equilibrium or non-equilibrium chemistry is considered. Absorption due to collisions between molecules (called Collision-Induced Absorption) also has a measurable effect, and modelers typically have to account for interactions between H_2–H_2 and H_2–He.

Cloud structure plays a critical role in controlling the resulting atmospheric spectra. Unfortunately, clouds are extremely difficult to model and represent one of the greatest uncertainties in the atmospheric models. The structure, height, and composition of the clouds depends on the local conditions in the atmosphere as well as the transport (horizontal and vertical) of the condensates present in the atmosphere. In "ad hoc" cloud models, the type of condensates, the degree of condensation, and the particle size distribution are all free parameters in defining the cloud structure. One-dimensional cloud models use cloud microphysics to compute these parameters (71; 88). All extrasolar planet atmosphere models currently in the literature further assume that the clouds are uniformly distributed over the entire planet.

Since the hot Jupiters are likely to be tidally locked (meaning the same hemisphere of the planet always faces the star), atmospheric circulation is key for redistributing absorbed stellar energy and determining the temperature gradients across the planet atmosphere. Atmospheric circulation models (e.g., 118; 86; 87) have not yet been coupled with radiative transfer models. In this absence, the atmospheric circulation has been parameterized by a parameter f: a value of $f = 1$ implies that the incident stellar radiation is emitted into 4π steradians (meaning the heat is evenly redistributed throughout the planet's atmosphere), while $f = 2$ implies that the incident stellar radiation is emitted into only 2π steradians (i.e., only the day side absorbs and emits the radiation, and there is no transport to the night side). This parameter is a way of quantifying the atmospheric dynamics, and it is used in the models to interpret the observed spectra (see Section 2.7). In model atmospheres

this factor f is used in reducing the incident stellar radiation. It is also used in estimating the equilibrium temperature $T_{\rm eq}$, defined as

$$T_{\rm eq} = T_* \sqrt{\frac{R_*}{2a}} \, [f(1 - A_{\rm B})]^{1/4}, \tag{2.19}$$

where T_* is the stellar temperature, R_* is the stellar radius, a is the orbital semi-major axis, and $A_{\rm B}$ is the (unknown) Bond albedo, which is the fraction of incident stellar radiation scattered back into space in all directions by the planet. This relation was used to derived the values listed in Table 2.1, assuming $f = 1$ and $A_{\rm B} = 0.3$.

We now turn to a discussion of the specific spectroscopic and photometric observations of extrasolar planets that have been conducted.

2.7 Observations

In this section, we summarize the important spectroscopic and photometric observations of transiting planets that have been conducted, during both primary and secondary eclipse. Most of these observations have been performed on the planet HD 209458 b, since it was detected first. We conclude by describing how the model calculations have helped to interpret these results.

As discussed in Section 2.2, the planetary spectrum can be probed during transit using a method called transmission spectroscopy. Although the planetary spectrum is ∼10,000 times fainter than that of the star, the differential nature of the measurement makes it possible to achieve this precision. Several detections and useful upper limits have been obtained on HD 209458 b:

- Sodium doublet detected (84);
- Hydrogen Lyman-α detected (120);
- Carbon monoxide upper limit (89).

The sodium detected was approximately a factor of three smaller than expected from simple models of the atmosphere, suggesting the presence of a high cloud that masks the true sodium abundance. The detection of the transit in H Lyman-α was huge—a 15% drop in stellar flux during transit, 10 times greater than the transit depth at visible wavelengths. This implies an extended atmosphere of 3 or 4 Jupiter radii, and suggests that the planet is losing mass over its lifetime. The CO non-detection further reinforces the notion of a high cloud in the planet's atmosphere.

The complimentary technique during secondary eclipse is called ***occultation spectroscopy***. Briefly, this involves taking spectra of the system when the planet is out of eclipse (when both the star and planet are visible) and comparing to spectra recorded when the planet is hidden during secondary eclipse. By carefully differencing these spectra, one can in principle derive the spectrum of the planet itself. Although this technique has not yet been successfully conducted on extrasolar planets, early attempts have yielded some useful information:

- Upper limit on emission near 2.2 (109);
- Upper limit on methane abundance (108).

Both of these limits were derived from ground-based observations, which are often limited by variations in the terrestrial atmosphere, making detection of spectral features difficult.

We now turn to photometric observations of the secondary eclipse that have occurred most recently. Although measurable, the effect due to the secondary eclipse is small, e.g., $\sim 0.3\%$ for HD 209458 b at 20 μm (see Figure 2.2), and decreasing for smaller wavelengths. NASA's Spitzer Space Telescope† is responsible for the first detection of a secondary eclipse of a transiting planet. Spitzer, with an 85-cm aperture, has three instruments on board that together perform photometry and spectroscopy at infrared wavelengths. In March 2005, two independent research groups announced detections of the secondary eclipse of two different planets using two Spitzer instruments. Observations of HD 209458 b with the Multiband Imagaing Photometer for Spitzer (MIPS) detected the secondary eclipse at 24 (90), while TrES-1 was observed in two wavelengths (4.5 and 8) with the Infrared Array Camera (IRAC) (85). These observations represent the first ***direct*** detection of an extrasolar planet. Most recently, the secondary eclipse of HD 189733 b was observed at 16 using the Infrared Spectrograph (IRS), although the observation was performed photometrically, not spectrscopically, using a detector that is normally used only to align the star on the slit (91).

The secondary eclipse detections provide a measurement of the ***brightness temperature*** of the planets, at the respective wavelengths. The brightness temperature is the blackbody temperature of an object at a particular wavelength; given the irradiance, the blackbody function (see Equation 2.6) can be inverted to solve for temperature. For HD 209458 b the brightness temperature at 24 is 1130 K, and for TrES-1 it is 1100 K

† `http://ssc.spitzer.caltech.edu/`

at 4.5 and 8 . HD 189733 b has a brightness temperature of 1117 K at 16 . Although models have predicted the effective temperature of the atmospheres of extrasolar planets, these are the first observational measurements of the temperature of an extrasolar planetary atmosphere.

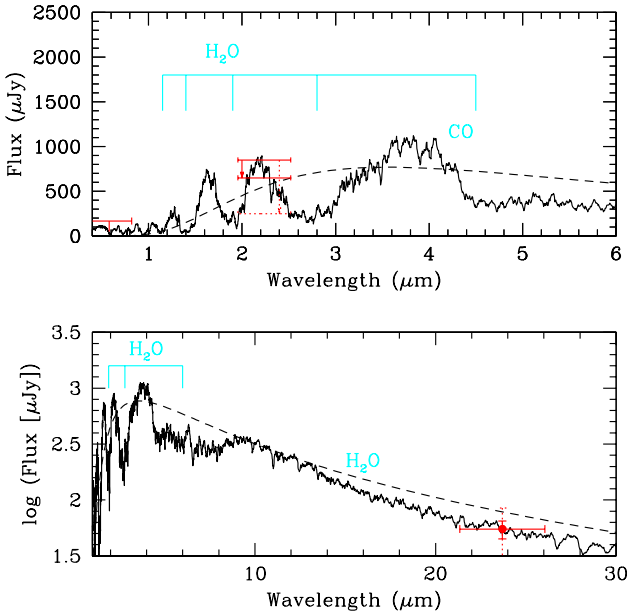
With the Spitzer photometry of several transiting planets, as well ground-based spectroscopic observations, we can now compare the observational results to theoretical calculations and begin to construct a comprehensive picture of the atmospheres of the transiting planets. In the wake of the three initial photometric detections of thermal emission from two extrasolar planets (90; 85), *four* theory papers (117; 74; 92; 81) appeared within a few months to explain the results! Some of these even have conflicting conclusions. One conclusion on which all of the explanations agree is that the planets are hot, as predicted. (We note that this was not a given; a planet with a high Bond albedo, for example, would reflect much of the incident stellar irradiation and therefore could be much cooler, as seen in Equation 2.19.)

The second point on which all modelers agree is that the TrES-1 data points at 4.5 and 8.0 are not consistent with the assumption of solar abundances, because the 8.0 flux is too high. Beyond these two conclusions, the interpretations diverge.

Seager *et al.* (117) conclude that a range of models remain consistent with the data. They include the 2.2 upper limit reported by Richardson *et al.* (109) (which has been largely ignored by modelers), as well as an upper limit on the albedo from the Canadian MOST satellite Rowe *et al.* (110), and are able to eliminate the models for HD 209458 b that are on the hot and cold ends of the plausible temperature range for the planet. Their work suggests that an intermediate value for f (see Equation 2.19) is most likely, indicating that the atmospheric circulation is somewhere between the two extremes (efficient redistribution vs. none at all). The interpretation by Seager *et al.* (117) and the observational results for HD 209458 b are shown in Figure 2.4.

Fortney *et al.* (92) show that standard models assuming solar abundances are consistent with HD 209458 b but only marginally consistent (within 2σ at 8.0) for TrES-1. For both planets, their best fit models assume that the incident stellar radiation is redistributed efficiently throughout the atmosphere (i.e., $f = 1$). On the other hand, Burrows *et al.* (81) conclude that the $f = 2$ case is more likely, indicating that the day side is significantly brighter in the infrared than the night side. They also infer the presence of CO and possibly H_2O. The resolution of this discrepancy awaits further Spitzer observations.

Fig. 2.4. Theoretical spectra of HD 209458 b with data points and upper limits. All models are for $f = 2$. The solid (black) curve is a cloud-free model with solar abundance, characterized by deep water vapor absorption features. The dotted (blue) curve is a cloudy model which due to the gray-like condensate opacities is like a blackbody. The dash-dot (green) curve is for C/O = 1.01 and all other elements in solar abundance. This model is characterized by strong CO and CH_4 absorption and weak H_2O absorption. From left to right the data points are MOST upper limit (110), a constraint on the H_2O band depth (108; 117), and the Spitzer/MIPS thermal emission point at 24 (90). The solid lines show 1 σ error bars or upper limits and the dashed lines show 3 σ values. Note the linear flux scale on the upper panel and the log flux scale on the lower panel.



Finally, with the recent Spitzer detection of HD 189733 b during secondary eclipse at 16 (91) and detections by IRAC and MIPS under analysis (D. Charbonneau, private communication), we have more data available for comparison to theoretical spectra. In addition, observations are being analyzed or planned to detect a mid-infrared emission spectrum of HD 209485 b and HD 189733 b, respectively, both using Spitzer/IRS. These observations would be the first observed emission spectra of an

extrasolar planet, and will advance our understanding beyond the few photometric data points we have now.

2.8 Future Missions

The spectroscopic and photometric observations of hot Jupiters have provided a wealth of information on their physical characteristics and led to insights about their atmospheric structure. What about extrasolar planets similar to the Earth? Although detection of such rocky planets remains just beyond the limits of current detection techniques, a few short-period planets with masses only 5–15 times that of the Earth (sometimes called "hot super-massive Earths") have been discovered (112; 75), pushing the detection limit to ever-smaller planets. We close this chapter with a brief discussion of how we might search for Earth-like planets around other stars and what future missions are being planned to tackle this fundamental question.

The goal of directly imaging an Earth-like planet is to search for ***biomarkers***, which are spectral features that can be used as diagnostics to search for the presence of life as we know it. The Earth has several such biomarkers that are indicative of habitability or life. Figure 2.5 shows two of these species: O_2 and its photolytic product O_3, two of the most reliable biomarker gas indicators of life. O_2 is highly reactive and therefore will remain in significant quantities in the atmosphere only if it is continually produced. There are no abiotic continuous sources of large quantities of O_2 and only rare false positives that in most cases could likely be ruled out by other planetary characteristics. N_2O is a second gas produced by life—albeit in small quantities—during microbial oxidation-reduction reactions. N_2O has a very weak spectroscopic signature.

In addition to atmospheric biomarkers, the Earth has one very strong and very intriguing biomarker on its surface: vegetation. The reflection spectrum of photosynthetic vegetation has a dramatic sudden rise in albedo around 750 nm by almost an order of magnitude! (This effect is not included in the model plotted in Figure 2.5.) Vegetation has evolved this strong reflection feature, known as the 'red edge,' as a cooling mechanism to prevent overheating which would cause chlorophyll to degrade. On Earth, this feature is likely reduced by a few percent due to clouds. A surface biomarker might be distinguished from an atmospheric signature by observing time variations; that is, as the continents, for example, rotate in and out of view, the spectral signal will change

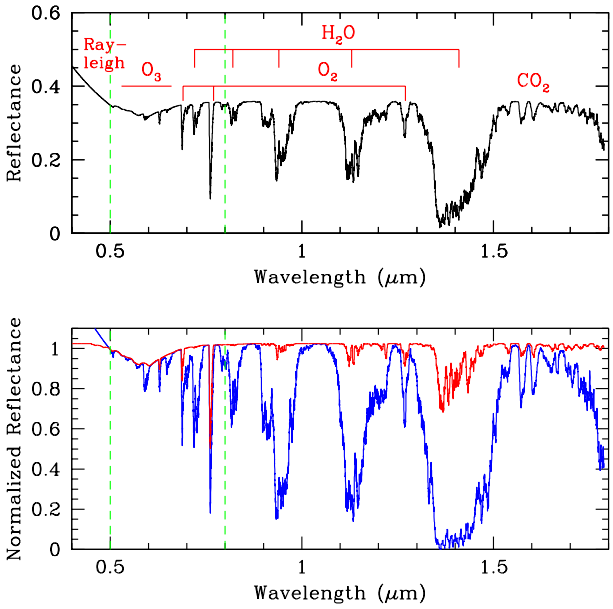


Fig. 2.5. Theoretical spectra of Earth. Upper panel: theoeretical model that matches the Woolf *et al.* (123) Earthshine data. The dashed vertical lines show the nominal wavelength range of TPF-C. Lower panel: normalized models of Earth showing effects of clouds. The top curve is for uniform high cloud coverage showing weaker water vapor features. The bottom curve shows the case with no clouds resulting in deep absorption features.

correspondingly. Other spectral features, although not biomarkers because they do not reveal direct information about life or habitability, can nonetheless provide significant information about the planet. These include CO_2 (which is indicative of a terrestrial atmosphere and has a very strong mid-infrared spectral feature) and CH_4 (which has both biotic and abiotic origins). The range of spectral features are needed to characterize Earth-like planet atmospheres.

The James Webb Space Telescope (JWST) (e.g., 93), tentatively scheduled for launch after 2013, will pick up where Spitzer leaves off, in terms of extrasolar planet characterization by primary and secondary eclipse studies. JWST is an infrared telescope with an aperture 6.5 m in diameter, representing a factor of ∼60 greater collecting area over Spitzer's 0.85 m diameter aperture. JWST will not only be able to detect thermal

emission spectra from hot Jupiters, but also may be able to see emission from hot, super-massive Earths. It may also be possible to perform transmission spectroscopy on such planets with JWST.

NASA's Terrestrial Planet Finder missions and ESA's Darwin mission seek to find and characterize Earth-like planets orbiting nearby stars. TPF is split into two separate missions, a visible coronograph (TPF-C) and an infrared nulling interferometer (TPF-I). Although scheduling and budgets for TPF are tentative, these missions would provide direct imaging of planets and thus low-resolution spectra of a wide variety of planet sizes and semi-major axes. One major goal of these missions would be to search the observed spectra for the biomarker features described above, in hopes of finding evidence for life on another world.

2.9 Summary

The transiting extrasolar planets have provided new opportunities to characterize the atmospheres and bulk compositions of worlds beyond our solar system. The geometry of these systems, in which the planet periodically crosses in front of its parent star (primary eclipse) and disappears behind the star (secondary eclipse), has allowed measurements of the true mass, radius, density, and (in a few cases) the brightness temperature of these planets for the first time. This chapter has also presented a brief overview of spectroscopy, summarized how model atmospheres are computed, and described the notable observations of transiting planets. Finally, this chapter has addressed the detection and characterization of Earth-like planets around other stars and summarizes a few missions being planned to accomplish this.

References

[71] Ackerman, A. S., & Marley, M. S. 2001. Precipitating Condensation Clouds in Substellar Atmospheres. , **556**(Aug.), 872–884.

[72] Alonso, R., Brown, T. M., Torres, G., Latham, D. W., Sozzetti, A., Mandushev, G., Belmonte, J. A., Charbonneau, D., Deeg, H. J., Dunham, E. W., O'Donovan, F. T., & Stefanik, R. P. 2004. TrES-1: The Transiting Planet of a Bright K0 V Star. , **613**(Oct.), L153–L156.

[73] Bakos, G. A., Pal, A., Latham, D. W., Noyes, R. W., & Stefanik, R. P. 2006. A stellar companion in the HD 189733 system with a known transiting extrasolar planet. *Arxiv astrophysics e-prints*, Feb.

[74] Barman, T. S., Hauschildt, P. H., & Allard, F. 2005. Phase-Dependent Properties of Extrasolar Planet Atmospheres. , **632**(Oct.), 1132–1139.

[75] Beaulieu, J.-P., Bennett, D. P., Fouqué, P., Williams, A., Dominik, M., Jorgensen, U. G., Kubas, D., Cassan, A., Coutures, C., Greenhill, J.,

Hill, K., Menzies, J., Sackett, P. D., Albrow, M., Brillant, S., Caldwell, J. A. R., Calitz, J. J., Cook, K. H., Corrales, E., Desort, M., Dieters, S., Dominis, D., Donatowicz, J., Hoffman, M., Kane, S., Marquette, J.-B., Martin, R., Meintjes, P., Pollard, K., Sahu, K., Vinter, C., Wambsganss, J., Woller, K., Horne, K., Steele, I., Bramich, D. M., Burgdorf, M., Snodgrass, C., Bode, M., Udalski, A., Szymański, M. K., Kubiak, M., Więckowski, T., Pietrzyński, G., Soszyński, I., Szewczyk, O., Wyrzykowski, Ł., Paczyński, B., Abe, F., Bond, I. A., Britton, T. R., Gilmore, A. C., Hearnshaw, J. B., Itow, Y., Kamiya, K., Kilmartin, P. M., Korpela, A. V., Masuda, K., Matsubara, Y., Motomura, M., Muraki, Y., Nakamura, S., Okada, C., Ohnishi, K., Rattenbury, N. J., Sako, T., Sato, S., Sasaki, M., Sekiguchi, T., Sullivan, D. J., Tristram, P. J., Yock, P. C. M., & Yoshioka, T. 2006. Discovery of a cool planet of 5.5 Earth masses through gravitational microlensing. , **439**(Jan.), 437–440.

[76] Bouchy, F., Pont, F., Santos, N. C., Melo, C., Mayor, M., Queloz, D., & Udry, S. 2004. Two new "very hot Jupiters" among the OGLE transiting candidates. , **421**(July), L13–L16.

[77] Bouchy, F., Udry, S., Mayor, M., Moutou, C., Pont, F., Iribarne, N., da Silva, R., Ilovaisky, S., Queloz, D., Santos, N. C., Ségransan, D., & Zucker, S. 2005. ELODIE metallicity-biased search for transiting Hot Jupiters. II. A very hot Jupiter transiting the bright K star HD 189733. , **444**(Dec.), L15–L19.

[78] Burrows, A., & Sharp, C. M. 1999. Chemical Equilibrium Abundances in Brown Dwarf and Extrasolar Giant Planet Atmospheres. , **512**(Feb.), 843–863.

[79] Burrows, A., Guillot, T., Hubbard, W. B., Marley, M. S., Saumon, D., Lunine, J. I., & Sudarsky, D. 2000. On the Radii of Close-in Giant Planets. , **534**(May), L97–LL100.

[80] Burrows, A., Sudarsky, D., & Hubbard, W. B. 2003. A Theory for the Radius of the Transiting Giant Planet HD 209458b. , **594**(Sept.), 545–551.

[81] Burrows, A., Hubeny, I., & Sudarsky, D. 2005. A Theoretical Interpretation of the Measurements of the Secondary Eclipses of TrES-1 and HD 209458b. , **625**(June), L135–L138.

[82] Charbonneau, D. 2003. HD 209458 and the Power of the Dark Side. *Pages 449–456 of: Asp conf. ser. 294: Scientific frontiers in research on extrasolar planets*.

[83] Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000. Detection of Planetary Transits Across a Sun-like Star. , **529**(Jan.), L45–L48.

[84] Charbonneau, D., Brown, T. M., Noyes, R. W., & Gilliland, R. L. 2002. Detection of an Extrasolar Planet Atmosphere. , **568**(Mar.), 377–384.

[85] Charbonneau, D., Allen, L. E., Megeath, S. T., Torres, G., Alonso, R., Brown, T. M., Gilliland, R. L., Latham, D. W., Mandushev, G., O'Donovan, F. T., & Sozzetti, A. 2005. Detection of Thermal Emission from an Extrasolar Planet. , **626**(June), 523–529.

[86] Cho, J. Y.-K., Menou, K., Hansen, B. M. S., & Seager, S. 2003. The Changing Face of the Extrasolar Giant Planet HD 209458b. , **587**(Apr.), L117–L120.

[87] Cooper, C. S., & Showman, A. P. 2005. Dynamic Meteorology at the Photosphere of HD 209458b. , **629**(Aug.), L45–L48.

[88] Cooper, C. S., Sudarsky, D., Milsom, J. A., Lunine, J. I., & Burrows, A. 2003. Modeling the Formation of Clouds in Brown Dwarf Atmospheres. , **586**(Apr.), 1320–1337.

[89] Deming, D., Brown, T. M., Charbonneau, D., Harrington, J., & Richardson, L. J. 2005a. A New Search for Carbon Monoxide Absorption in the Transmission Spectrum of the Extrasolar Planet HD 209458b. , **622**(Apr.), 1149–1159.

[90] Deming, D., Seager, S., Richardson, L. J., & Harrington, J. 2005b. Infrared radiation from an extrasolar planet. , **434**(Mar.), 740–743.

[91] Deming, D., Harrington, J., Seager, S., & Richardson, L. J. 2006. Strong Infrared Emission from the Extrasolar Planet HD189733b. *Arxiv astrophysics e-prints*, Feb.

[92] Fortney, J. J., Marley, M. S., Lodders, K., Saumon, D., & Freedman, R. 2005. Comparative Planetary Atmospheres: Models of TrES-1 and HD 209458b. , **627**(July), L69–L72.

[93] Gardner, J. P., Mather, J. C., Clampin, M., Doyon, R., Greenhouse, M. A., Hammel, H. B., Hutchings, J. B., Jakobsen, P., Lilly, S. J., Long, K. S., Lunine, J. I., McCaughrean, M. J., Mountain, M., Nella, J., Rieke, G. H., Rieke, M. J., Rix, H.-W., Smith, E. P., Sonneborn, G., Stiavelli, M., Stockman, H. S., Windhorst, R. A., & Wright, G. S. 2006. The James Webb Space Telescope. *Arxiv astrophysics e-prints*, June.

[94] Gaudi, B. S., Seager, S., & Mallen-Ornelas, G. 2005. On the Period Distribution of Close-in Extrasolar Giant Planets. , **623**(Apr.), 472–481.

[95] Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2000. A Transiting "51 Peg-like" Planet. , **529**(Jan.), L41–L44.

[96] Holman, M. J., Winn, J. N., Stanek, K. Z., Torres, G., Sasselov, D. D., Allen, R. L., & Fraser, W. 2005. High-precision Transit Photometry of OGLE-TR-10. *Arxiv astrophysics e-prints*, June.

[97] Knutson, H., Charbonneau, D., Noyes, R. W., Brown, T. M., & Gilliland, R. L. 2006. Using Stellar Limb-Darkening to Refine the Properties of HD 209458b. *Arxiv astrophysics e-prints*, Mar.

[98] Konacki, M., Torres, G., Sasselov, D. D., Pietrzyński, G., Udalski, A., Jha, S., Ruiz, M. T., Gieren, W., & Minniti, D. 2004. The Transiting Extrasolar Giant Planet around the Star OGLE-TR-113. , **609**(July), L37–L40.

[99] Konacki, M., Torres, G., Sasselov, D. D., & Jha, S. 2005. A Transiting Extrasolar Giant Planet around the Star OGLE-TR-10. , **624**(May), 372–377.

[100] Laughlin, G., Wolf, A., Vanmunster, T., Bodenheimer, P., Fischer, D., Marcy, G., Butler, P., & Vogt, S. 2005. A Comparison of Observationally Determined Radii with Theoretical Radius Predictions for Short-Period Transiting Extrasolar Planets. , **621**(Mar.), 1072–1078.

[101] Liou, K. N. 2002. *An Introduction to Atmospheric Radiation*. 2 edn. Academic Press.

[102] Marley, M. S., Fortney, J., Seager, S., & Barman, T. 2006. Atmospheres of Extrasolar Giant Planets. *Arxiv astrophysics e-prints*, Feb.

[103] Mazeh, T., Naef, D., Torres, G., Latham, D. W., Mayor, M., Beuzit, J., Brown, T. M., Buchhave, L., Burnet, M., Carney, B. W., Charbonneau, D., Drukier, G. A., Laird, J. B., Pepe, F., Perrier, C., Queloz, D., Santos, N. C., Sivan, J., Udry, S. ;., & Zucker, S. 2000. The Spectroscopic Orbit of the Planetary Companion Transiting HD 209458. , **532**(Mar.), L55–L58.

[104] McCullough, P. R., Stys, J. E., Valenti, J. A., Johns-Krull, C. M., Janes, K. A., Heasley, J. N., Bye, B. A., Dodd, C., Fleming, S. W., Pinnick, A., Bissinger, R., Gary, B. L., Howell, P. J., & Vanmunster, T. 2006. A Transiting Planet of a Sun-like Star. *Arxiv astrophysics e-prints*, May.

[105] Mihalas, D. 1970. *Stellar Atmospheres*. W. H. Freeman and Company.

[106] Moutou, C., Pont, F., Bouchy, F., & Mayor, M. 2004. Accurate radius and mass of the transiting exoplanet OGLE-TR-132b. , **424**(Sept.), L31–L34.

[107] Pont, F., Bouchy, F., Queloz, D., Santos, N. C., Melo, C., Mayor, M., & Udry, S. 2004. The "missing link": A 4-day period transiting exoplanet around OGLE-TR-111. , **426**(Oct.), L15–L18.

[108] Richardson, L. J., Deming, D., Wiedemann, G., Goukenleuque, C., Steyert, D., Harrington, J., & Esposito, L. W. 2003a. Infrared Observations during the Secondary Eclipse of HD 209458b. I. 3.6 Micron Occultation Spectroscopy Using the Very Large Telescope. , **584**(Feb.), 1053–1062.

[109] Richardson, L. J., Deming, D., & Seager, S. 2003b. Infrared Observations during the Secondary Eclipse of HD 209458b. II. Strong Limits on the Infrared Spectrum Near 2.2 Microns. , **597**(Nov.), 581.

[110] Rowe, J. F., Matthews, J. M., Seager, S., Kuschnig, R., Guenther, D. B., Moffat, A. F. J., Rucinski, S. M., Sasselov, D., Walker, G. A. H., & Weiss, W. W. 2005. MOST spacebased photometry of the transiting exoplanet system HD 209458: I. Albedo Measurements of an Extrasolar Planet. *American astronomical society meeting abstracts*, **207**(Dec.), –+.

[111] Salby, M. L. 1996. *Fundamentals of Atmospheric Physics*. Academic Press.

[112] Santos, N. C., Bouchy, F., Mayor, M., Pepe, F., Queloz, D., Udry, S., Lovis, C., Bazot, M., Benz, W., Bertaux, J.-L., Lo Curto, G., Delfosse, X., Mordasini, C., Naef, D., Sivan, J.-P., & Vauclair, S. 2004. The HARPS survey for southern extra-solar planets. II. A 14 Earth-masses exoplanet around μ Arae. , **426**(Oct.), L19–L23.

[113] Santos, N. C., Pont, F., Melo, C., Israelian, G., Bouchy, F., Mayor, M., Moutou, C., Queloz, D., Udry, S., & Guillot, T. 2006. High resolution spectroscopy of stars with transiting planets. The cases of OGLE-TR-10, 56, 111, 113, and TrES-1. , **450**(May), 825–831.

[114] Sato, B., Fischer, D. A., Henry, G. W., Laughlin, G., Butler, R. P., Marcy, G. W., Vogt, S. S., Bodenheimer, P., Ida, S., Toyota, E., Wolf, A., Valenti, J. A., Boyd, L. J., Johnson, J. A., Wright, J. T., Ammons, M., Robinson, S., Strader, J., McCarthy, C., Tah, K. L., & Minniti, D. 2005. The N2K Consortium. II. A Transiting Hot Saturn around HD 149026 with a Large Dense Core. , **633**(Nov.), 465–473.

[115] Seager, S., & Mallén-Ornelas, G. 2003. A Unique Solution of Planet and Star Parameters from an Extrasolar Planet Transit Light Curve. , **585**(Mar.), 1038–1055.

[116] Seager, S., Whitney, B. A., & Sasselov, D. D. 2000. Photometric Light Curves and Polarization of Close-in Extrasolar Giant Planets. , **540**(Sept.), 504–520.

[117] Seager, S., Richardson, L. J., Hansen, B. M. S., Menou, K., Cho, J. Y.-K., & Deming, D. 2005. On the Dayside Thermal Emission of Hot Jupiters. , **632**(Oct.), 1122–1131.

[118] Showman, A. P., & Guillot, T. 2002. Atmospheric circulation and tides of "51 Pegasus b-like" planets. , **385**(Apr.), 166–180.

[119] Torres, G., Konacki, M., Sasselov, D. D., & Jha, S. 2004. New Data and Improved Parameters for the Extrasolar Transiting Planet OGLE-TR-56b. , **609**(July), 1071–1075.

[120] Vidal-Madjar, A., des Etangs, A. L., Désert, J.-M., Ballester, G. E., Ferlet, R., Hébrard, G., & Mayor, M. 2003. An extended upper atmosphere around the extrasolar planet HD209458b. , **422**(Mar.), 143–146.

[121] Winn, J. N., & Holman, M. J. 2005. Obliquity Tides on Hot Jupiters. , **628**(Aug.), L159–L162.

[122] Winn, J. N., Noyes, R. W., Holman, M. J., Charbonneau, D., Ohta, Y., Taruya, A., Suto, Y., Narita, N., Turner, E. L., Johnson, J. A., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2005. Measurement of Spin-Orbit Alignment in an Extrasolar Planetary System. , **631**(Oct.), 1215–1226.

[123] Woolf, N. J., Smith, P. S., Traub, W. A., & Jucks, K. W. 2002. The Spectrum of Earthshine: A Pale Blue Dot Observed from the Ground. , **574**(July), 430–433.

3

Terrestrial Planet Formation

Edward W. Thommes
Canadian Institute for Theoretical Astrophysics

3.1 Introduction

The current count of detected exoplanets exceeds 180. All of these are giant planets: Gaseous bodies like Jupiter or Saturn with a couple of potentially Neptune-like examples recently added to the menagerie. However, impressive as the "king of the planets" and its kind are, it is a terrestrial planet—a modest rocky body three thousandths Jupiter's mass—which actually harbors the life in our Solar System. And as we comtemplate the possibility of life elsewhere in the Universe, it is inevitably upon terrestrial planets that we focus our attention. The problem is, apart from a pair of no doubt highly unusual bodies orbiting the pulsar PSR1257 + 12 (**?**), Mercury, Venus, Earth and Mars remain the only examples of terrestrial planets we know. It is thus especially important to attempt to understand the process of terrestrial planet formation, so that even in the absence of direct observations, we can assess the likelihood of other planets like our own existing around other stars. At the same time, such work allows us to plan ahead for future terrestrial planet-hunting missions like *KEPLER, COROT* and (if it is resurrected) the Terrestrial Planet Finder (*TPF*).

In this article we give an overview of the theory of terrestrial planet formation as it currently stands. We begin with the formation of the basic building blocks, planetesimals, in §3.2. In §3.3, we look at how protoplanets grow from the agglomeration of planetesimals. This brings us to the final stage of terrestrial planet formation in §3.4, wherein the protoplanets collide to produce bodies like Earth and Venus. We examine the "standard model" of how this stage plays out in §3.4.1. In §3.4.4, we look at what cosmochemistry can tell us about the timeline of the Earth's formation. In §3.4.2 and §3.4.3, we review some more re-

cent variations on the model which have attempted to better reproduce the present-day orbital properties of the terrestrial planets. In §3.5, we summarize the current state of the impact model for the origin of the Earth-Moon system. In §3.6, we consider terrestrial planet formation and evolution from an astrobiology perspective. We begin with the concept of the habitable zone in §3.6.1. In §3.6.2, we address the issue of water delivery to the terrestrial region. The Late Heavy Bombardment is discussed in §3.6.3. In §3.6.4, we look at possible effects on terrestrial planet habitability from the giant planets in a given system. A summary is given in §3.7.

3.2 The formation of planetesimals

In the first phase of terrestrial planet formation, solids condense out of the protostellar gas disk as dust grains, settle to the disk midplane and eventually form kilometer-sized planetesimals. The actual mechanism of planetesimal formation remains unclear. (**?**) put forward a model in which the dust layer becomes sufficiently thin that it undergoes gravitational instability, contracting (in the terrestrial region) into kilometer-sized clumps and thus producing planetesimals directly. However, (**?**) pointed out that turbulent stirring of the dust layer would likely prevent gravitational instability. The dust particles are on Keplerian orbits while the gas disk, being partly supported by its own internal pressure, revolves at slightly less than the Keplerian rate. The resulting vertical shear in velocity at the dust-gas interface induces turbulence in the dust layer, preventing it from becoming thin enough for gravitational instability to set in.

Because of this problem, the most widely held current view is that dust particles build planetesimals by pairwise accretion. Laboratory experiments suggest that growth of dust to cm-sized aggregates proceeds quite easily (**?** and references therein). At larger sizes, the process is less well understood. For one thing, it is unclear how readily larger bodies stick together. Turbulent eddies in the gas disk may play a role in bringing about local particle concentrations (e.g. **?**). However, orbital decay due to aerodynamic gas drag from the nebula seems to present a problem. Decay rates are peaked for meter-sized bodies; for typical nebula models, such bodies spiral from 1 AU into the central star on a timescale of order 100 years (**?**). Thus, it appears growth through this size regime must be extremely rapid in order to avoid the loss of

all solids. Another possibility is that fluctuations in nebular gas density, turbulent or otherwise, may locally stall the migration.

3.3 The growth of protoplanets

Once the solids in the protoplanetary disk are mostly locked up in km-size bodies, their interactions become dominated by gravity, and in a sense the dynamics become simpler. If the relative velocities between neighboring planetesimals are dispersion-dominated—i.e., the main contribution is the planetesimals' random velocities, rather than the Keplerian shear between neighboring orbits—their growth rate can be estimated using a simple "particle in a box" approach (**?** and references therein):

$$\frac{dM}{dt} = \rho_{\rm plsmls} \sigma v_{\rm rel}, \tag{3.1}$$

where $\rho_{\rm plsmls}$ is the density with which planetesimals are distributed in the planetesimal disk, σ is the collisional cross-section of a planetesimal, and $v_{\rm rel}$ is the average relative velocity between nearby planetesimals. The collisional cross-section of a planetesimal is given by

$$\sigma = \pi R^2 \left(1 + \frac{v_{\rm esc}^2}{v_{\rm rel}^2} \right) \equiv \pi r^2 f_g, \tag{3.2}$$

where R is the planetesimal's physical radius, $v_{\rm esc} = \sqrt{2GM/R}$ the escape velocity from its surface, and f_g is the enhancement of the cross-section by gravitational focusing. With a few substitutions and simplifications, we can gain more insight from Eq. 3.1. To begin with, we will assume that gravitational focusing is effective, so that $f_g \gg 1$. Also, $\rho_{\rm plsmls} = \Sigma_{plsmls}/2H_{\rm plsmls}$, where $\Sigma_{\rm plsmls}$ is the surface density of the planetesimal disk, and $H_{\rm plsmls}$ is the disk's scale height. If the planetesimals have a characteristic random velocity v, then the vertical component is $\approx v/\sqrt{3}$, and so $H \approx (rv)/(\sqrt{3} v_{\rm Kep})$. Finally, $v_{\rm Kep} \propto r^{-1/2}$, $r \propto M^{1/3}$ and $v_{\rm esc} \propto M^{1/3}$, and so the growth rate has the form

$$\frac{dM}{dt} \propto \frac{\Sigma_{\rm plsmls} M^{4/3}}{v^2 r^{3/2}}. \tag{3.3}$$

As shown by (**?**), planetesimal accretion is subject to positive feedback, which results in runaway growth: The largest bodies grow the fastest, rapidly detaching themselves from the size distribution of the planetesimals. The reason for this can be seen from Eq. 3.3: $dM/dt \propto$

$M^{4/3}$, and so the growth timescale is

$$t_{\rm grow} \equiv \frac{M}{dM/dt} \propto M^{-1/3}, \tag{3.4}$$

thus larger bodies grow faster than smaller ones.

(**?**) showed that runaway growth only operates temporarily; once the largest protoplanets are massive enough to dominate the gravitational stirring of the nearby planetesimals, the mode of accretion changes. From this point on, near a protoplanet of mass M, the planetesimal random velocity is approximately proportional to the surface escape speed from the protoplanet, and so $v \propto M^{1/3}$. Therefore, we now have

$$\frac{dM}{dt} \propto \frac{\Sigma_{\rm plsmls} M^{2/3}}{r^{3/2}} \tag{3.5}$$

which makes

$$t_{\rm grow} \propto M^{1/3}. \tag{3.6}$$

Thus growth becomes orderly; larger bodies grow more slowly. This mode of accretion is commonly called *oligarchic growth* (**?**), since a relatively small sub-population of large bodies dominates the dynamics of, and feeds off of, the surrounding planetesimal swarm. In the terrestrial region, the transition from runaway to oligarchic growth already happens when the largest protoplanets are still many orders of magnitude below an Earth mass (**?**). Thus, protoplanets spend almost all of their time growing oligarchically.

As long as nebular gas is present, the random velocity v of the planetesimals is set by two competing effects: Gravitational stirring acts to increase v, while aerodynamic gas drag acts to reduce it. One can estimate the equilibrium random velocity by equating the two rates (**?**). In this way, one can eliminate v and write the protoplanet growth rate in terms planetesimal and gas disk properties. Also, for simplicity we approximate the planetesimal population as having a single characteristic mass m. The result is (for details see **?**)

$$\frac{dM}{dt} \approx A \Sigma_{\rm plsmls} M^{2/3}, \tag{3.7}$$

where

$$A = 3.9 \frac{b^{5/2} C_D^{2/5} G^{1/2} M_*^{1/6} \rho_{\rm gas}^{2/5}}{\rho_{\rm plsml}^{4/15} \rho_M^{1/3} r^{1/10} m^{2/15}}. \tag{3.8}$$

C_D is a dimensionless drag coefficient ~ 1 for km-sized or larger planetesimals, M_* is the mass of the central star, $\rho_{\rm gas}$ is the density of the

gas disk, and ρ_M and ρ_m are the material densities of the protoplanet and the planetesimals, respectively. The parameter b is the spacing between adjacent protoplanets in units of their Hill radii. The Hill radius of a companion of mass M orbiting a body of mass M_* is defined as

$$r_H = \left(\frac{M}{3M_*}\right)^{1/3} r \tag{3.9}$$

and is a measure of body M's gravitational reach; essentially, it is the radius within which a third body (with mass $\ll M$) orbits M rather than independently orbiting M_*. An equilibrium between mutual gravitational scattering on the one hand and recircularization by dynamical friction on the other keeps $b \sim 10$ (**?**).

Protoplanet growth by sweep-up of planetesimals ceases when all planetesimals are gone. This leads to the notion of the *isolation mass*, which is the mass at which a protoplanet has consumed all planetesimals within an annulus of width br_H centered on its orbital radius. The isolation mass is given by

$$M_{\rm iso} = \frac{\left(2\pi b \Sigma_{\rm plsmls} r^2\right)^{3/2}}{\sqrt{3M_*}}. \tag{3.10}$$

In order to obtain values for protoplanet accretion times and final masses, a useful starting point is the "minimum-mass Solar Nebula (MMSN) model, (**?**), obtained by "smearing out" the refractory elements contained in the Solar System planets into a power-law planetesimal disk, then adding gas to obtain Solar abundance:

$$\Sigma_{\rm gas}^{\rm min} = 1700 \left(\frac{r}{1\,{\rm AU}}\right)^{-3/2} {\rm g\,cm^{-2}}, \tag{3.11}$$

$$\rho_{\rm gas}^{\rm min} \approx \Sigma_{\rm gas}/2H,\ H = 0.0472 \left(\frac{r}{1\,{\rm AU}}\right)^{5/4} {\rm AU} \tag{3.12}$$

and

$$\Sigma_{\rm solid}^{\rm min} = 7.1 F_{\rm SN} \left(\frac{r}{1\,{\rm AU}}\right)^{-3/2} {\rm g\,cm^{-2}} \tag{3.13}$$

where

$$F_{\rm SN} = \begin{cases} 1, & r < r_{\rm SN} \\ 4.2, & r > r_{\rm SN} \end{cases} \tag{3.14}$$

is the "snow line" solids enhancement factor: Beyond $r_{\rm SN}$ (=2.7 AU in the Hayashi model), water freezes out, thus adding to the surface density of solids.

Using Equation 3.10 with $b = 10$, the MMSN yields an isolation mass

of 0.07 $M_\oplus$ at 1 AU. If we assume a material density of 2 g/cm^{-3} for planetesimals and protoplanets, and a characteristic planetesimal size of ~ 1 km (with a corresponding mass of 10^{-12} $M_\oplus$), the growth timescale for an isolation-mass body at 1 AU is $t_{\rm grow} \equiv \frac{M}{dM/dt} \approx 6 \times 10^4$ yrs. Thus, the sweep-up of planetesimals into protoplanets proceeds very rapidly in the terrestrial region, concluding long before the nebular gas dissipates, which takes a few million years (**?**).

3.4 The growth of planets

With the protoplanets (or "oligarchs") now starved of planetesimals, the final stage of terrestrial planet formation requires them to accrete each other. For this to happen, their orbits have to cross. This is referred to as the giant impact (or chaotic) phase. We will begin with a description of how this unfolds in the "standard model", and then touch on some more recent work.

3.4.1 The standard model

Traditionally, N-body simulations have been the tool of choice for characterizing the giant-impact phase. The problem is particularly well suited to this approach, because once all the planetesimals have been locked up in perhaps tens to hundreds of protoplanets, N is no longer an intractibly large number. The first direct N-body calculations (**?** **?**) already revealed a number of key features of the giant-impact phase. They showed that an ensemble of lunar to martian size bodies, allowed to accrete one another in collisions, naturally formed Solar System-like terrestrial planets. They also demonstrated that this process is highly stochastic. Because only a few bodies remain at the end, in any single simulation the notion of a characteristic planet mass as a function of stellocentric radius becomes meaningless, and the approach we use in §3.3 can no longer be applied.

Taken together, the work of (**?**), (**?**), and (**?**) can be said to delineate the "standard model" of how the final stage of terrestrial planet formation plays out: Simulations begin with a population of $10^2 - 10^3$ Lunar- to Martian-sized protoplanets, the endproducts of the previous oligarchic growth phase. Different realizations of a particular initial condition typically vary only by the (randomly-generated) initial orbital phases of the protoplanets; the stochastic nature of these systems ensures

that this is enough to make the details of the evolution completely different from one simulation to the next. Nevertheless, averaging over enough simulations, a number of common features emerge: On a timescale of $\sim 1 - 3 \times 10^8$ years, these bodies collide with each other and accrete to form planetary systems which, on average, resemble the Solar System's terrestrial planets in mass, number, and orbital radius. Mercury and Mars, given their small size, seem to fit into the picture as essentially primordial oligarchs, which managed to come through the whole process without undergoing any major accretional collisions. The spins of the endproducts are dominated by the last few major impacts onto each, and to lowest order are randomly oriented. This contrasts with the the present-day Solar System, in which all terrestrial planet spins are approximately perpendicular to their orbital plane. However, the obliquities of the terrestrial planets have evolved significantly by processes such as tidal interactions and spin-orbit coupling, thus their spins are unlikely to be primordial. A more puzzling issue is that of the final orbital eccentricities and inclinations. Their time-averaged values are almost universally larger ($e, i 0.1$, for i in radians) than those of Earth and Venus today ($e, i \approx 0.03$) (**?** **?**)

3.4.2 Reducing the orbital eccentricities

Because numerical models of late-stage terrestrial planet formation are otherwise so successful, the problem of reproducing Earth and Venus's low e and i has been the subject of much effort; it really is a case of "so close and yet so far!" Considering that we are demanding that the terrestrial region evolve directly from wildly crossing orbits and giant impacts, to a very quiescent state of nearly circular, co-planar orbits, this failure is not really surprising. In the most general terms, to solve this problem an additional, dissipative physical process must be invoked, which will damp away the planets' noncircular motions. Perhaps the most obvious candidate is the planets' own leftover building material. If a significant mass in planetesimals is left in the system, this can exert *dynamical friction* on the planets (e.g. **?**): equipartition of the energy in noncircular motions makes larger bodies lose eccentricitiy, and smaller bodies gain it. The problem is, the timescale for sweeping up the planetesimals is, as we saw in §3.3, only $\sim 10^5$ years. Thus, a significant reservoir of planetesimals is unlikely to persist long enough to affect the dynamics after 10^8 years or more. Alternatively, a large number of small bodies,

right down to dust size, may be generated collisionally during the planet growth process (**?**).

As long as the nebular gas persists, it also acts as a source of damping for protoplanet eccentricities and inclinations. Though bodies larger than kilometer size are not strongly affected by aerodynamic gas drag, their gravitational interaction with the gas disk can have a significant effect on them. Each planet launches density waves into the interior and exterior part of the disk, exchanging energy and angular momentum with it. The net result is that planets migrate inward while having their eccentricities damped (**? ? ? ?**). A semianalytic estimate of the orbital and eccentricity decay timescales is (**?**)

$$t_m \equiv -\frac{J}{\dot{J}} = 3.5{\times}10^5 \left(\frac{F_s}{0.4H}\right)^{1.75} \left[\frac{1+\left(\frac{e_p a_p}{1.3H}\right)^5}{1-\left(\frac{e_p r_p}{1.1H}\right)^4}\right] \left(\frac{H/a_p}{0.07}\right)^2 \left(\frac{\Sigma_{\rm gas}}{588{\rm g\,cm^{-2}}}\right)^{-1} \left(\frac{M_p}{1M_\oplus}\right)^{-1} \left(\frac{a_p}{1AU}\right)^{-1/2} {\rm yrs} \tag{3.15}$$

and

$$t_e \equiv -\frac{e}{\dot{e}} = 2.5{\times}10^3 \left(\frac{F_s}{0.4H}\right)^{2.5} \left[1+\frac{1}{4}\left(\frac{e_p}{H/a_p}\right)^3\right] \left(\frac{H/a_p}{0.07}\right)^4 \left(\frac{\Sigma_{\rm gas}}{588{\rm g\,cm^{-2}}}\right)^{-1} \left(\frac{M_p}{1M_\oplus}\right)^{-1} \left(\frac{a_p}{1AU}\right)^{-1/2} {\rm yrs} \tag{3.16}$$

where J is the orbital angular momentum, a_p, e_p and M_p are the planet's semimajor axis and eccentricity, H is again the scale height of the gas disk, and F_S is a softening factor to account for the disk thickness. Thus an Earth-mass planet embedded in a MMSN gas disk suffers orbital decay on a timescale of $\sim 10^6$ years, and has its eccentricity damped on a timescale of $\sim 10^4$ years. Because of this difference in timescales, even when the gas disk has become so tenuous that only negligibly little planet migration is occurring, the planets can still be subject to significant eccentricity damping.

(**? ?**) performed simulations of this scenario. They found, first of all, that the giant-impact stage tends to be delayed until the gas is strongly depleted, such that $t_e 10^6$ yrs, approximately the crossing time for adjacent oligarchs. Therefore in this scenario, the terrestrial region is essentially frozen in the end state of oligarchic growth until the gas disk is tenuous enough that dynamical evolution can resume. Kominami & Ida demonstrated that low final eccentricities of Earth and Venus analogues can be generated in this way, but fine-tuning of the disk properties, specifically the gas dispersal time, is required. Even then, on average 6 to 7 bodies are left in the end, signficantly more than than the present-day number of Solar System terrestrial planets.

The fundamental problem is that making the oligarchs' orbits cross and damping their eccentricities are two conflicting objectives. Also, they found that including Jupiter and Saturn's gravitational perturbations on the terrestrial region tended to make the model work more poorly (in the sense that eccentricities and inclinations became higher), leading them to speculate that the terrestrial planets might have formed at a time when Jupiter and Saturn had not yet reached their final masses.

3.4.3 "Dynamical shakeup" as a pathway to the Solar System

(**?**) (see also **?**) developed a model for late-stage terrestrial planet formation which also hinges on the dissipating gas disk. However, they also included the effect of the gas disk on the precession of all the embedded orbits, with the result that things play out in a dramatically different way. This is due to the effect of a *secular resonance* of Jupiter. A detailed discussion of secular resonances is given in, e.g., (**?**). Here, we will restrict ourselves to a simplified, qualitative description: A secular resonance between two bodies occurs (approximately) when the precession frequencies of their orbits are equal. In our case, we are looking at the precession rate of terrestrial oligarchs relative to that of Jupiter. The former precess due to the gravitational influence of the gas disk, Jupiter, Saturn, and to a lesser extent, each other. Jupiter, meanwhile, is made to precess by the influence of Saturn and the gas disk, as well as a small contribution from the protoplanets themselves. When a given oligarch's precession is commensurate with that of Jupiter and a resonance occurs, a rapid exchange of angular momentum takes place between the two bodies. Being much more massive, Jupiter's orbit undergoes little change. The orbit of the small protoplanet, however, rapidly gains eccentricity. For a given gas disk density distribution, there is one particular heliocentric radius interior (and exterior, though this plays no role in the model) to Jupiter at which a body will be in secular resonance. As the gas disk dissipates (through a combination of accreting onto the star, photoevaporation, and stripping by stellar winds), its gravitational potential changes and so too do the precession rates of the embedded (proto-) planets and thus, the location of the interior secular resonance. The resonance begins near Jupiter in a massive gas disk, finally ending up—once the gas is all gone—at its present-day location at 0.6 AU, just inside the current orbit of Venus (**? ?**) Along the way it thus sweeps the asteroid belt and much of the terrestrial region and one after another the oligarchs have their eccentricity sharply raised, thus

crossing their neighbors' orbits. In this way, the giant impact phase is kick-started when there is still a significant amount of gas left in the disk. (**?**) perform numerical simulations of this process and find that accretion among oligarchs can finish rapidly in this way, leaving enough time for the last remnants of the gas disk to damp the eccentricities and inclinations of the end products. Another effect in evidence is a net inward migration of material during the resonance sweeping process, since bodies forced to high eccentricities in the presence of a damping force lose orbital energy as their eccentricity is damped again, thus spiraling inward (**?**). This offers a way to both clear the asteroid belt region (which is highly depleted relative to even the MMSN planetesimal disk) and deliver water-rich building material for the Earth (see §3.6.2 below).

3.4.4 Cosmochemistry

Parent-daughter pairs of radioactive isotopes can be exploited as chronometers to trace the timelines of processes in the early Solar System, potentially helping us to decide among the different formation models. Two such pairs are Hf-W and U-Pb. For both of these, the parent element is a lithophile that is retained in silicate reservoirs during planetary accretion, whereas the daughter element is segregated into the core. Since each of the large late-stage impacts would "reset the clock" by re-mixing core and mantle, measurements of relative abundances of these elements in the Earth can be used to estimate how long ago the giant-impact phase concluded. It is thought that the last impact was the Moon-forming one (see §3.5 below). Measurements of Hf-W have been used to derive a growth time of the Earth ranging from about 15 to 50 Myrs (**?** **?** **?**), intriguingly similar to the short timescales obtained by (**?**). However, the U-Pb chronometer implies a later formation time, $\sim 65 - 85$ Myrs (**?**). It may be that U-Pb traces the last stages of core segregation, while Hf-W traces the time for *most* of the core to finish forming (**?**). Given the difficult nature of these measurements, it should not be surprising that the different groups' results have still not completely been reconciled with each other. Future developments in this area may yet allow us to determine with some certainty whether the formation of the terrestrial planets was rapid or drawn-out.

3.5 The origin of the Earth-Moon system

The leading model for the formation of the Moon is that a Mars-mass impactor struck the proto-Earth, with the Moon accreting out of the resulting debris disk (**?** **?**). This scenario naturally results in a Moon depleted in volatiles and iron-poor (since the impact would have preferentially splashed out mantle material from the proto-Earth). The giant-impact model was explored further via a succession of smoothed-particle hydrodynamics (SPH) simulations, greatly increasing over time in resolution as computing power advanced (**?** **?** **?** **?**). As mentioned in §3.4.4, this impact is thought to have been the last major accretion event in the Earth's history.

The latest simulations do the best at reproducing the Earth-Moon system with an impactor just over 0.1 $M_\oplus$ striking proto-Earth at about a 45 degree angle. This ejects about a lunar mass of material exterior to Earth's Roche radius (the distance inside which a strengthless, self-gravitating body would be disrupted by Earth's tidal forces) at about 3 Earth radii ($R_\oplus$). The vast majority of this material comes from the impactor itself. Initially, 10 to 30% of it is in the form of silicate vapor. The orbital period at the Roche radius is only about 7 hours, thus subsequent evolution is very fast. (**?**) simulated the accretion of the Moon in the proto-lunar disk, showing that it takes between a month and a year to get to a single remaining lunar-mass body. The Moon's current orbital radius is ≈ 60 $R_\oplus$, thus its orbit has expanded significantly from the time it formed. This is effected by the Moon's tidal interaction with the Earth, which transfers angular momentum from the Earth's spin to the Moon's orbit. Early on, interaction with the last remnants of the protolunar disk likely also made a large contribution to the outward migration of the Moon (**?**)

3.6 Terrestrial planets and astrobiology

3.6.1 The habitable zone

The term "habitable zone" (HZ) goes back to (**?**). The criteria adopted for what constitutes a potentially life-supporting planet have varied widely. The most durable has proven to be simply the requirement that liquid water can exist on the planet surface (**?** **?**) . In fact, "Follow the water!" has become one of the guiding principles of astrobiology. (**?**) used a climate model to calculate this zone for stars later than F0, which have lifetimes exceeding 2 Gyrs, thus giving life ample time to

evolve. The inner edge of the HZ is set by water loss via photolysis and hydrogen escape; the outer edge is set by the formation of CO_2 clouds, which cool a planet by raising its albedo.

For G stars like the Sun, (**?**) conservatively estimate that the HZ stretches from 0.95 to 1.37 AU. F stars, being hotter, have a more distant HZ, while later type stars have a closer-in HZ (their Fig. 16). They point out that the logarithmic width of HZs is approximately constant across stellar types. This is significant because the spacing between adjacent planets ought to be set at least in part by their Hill radii; since $r_H \propto r$ (Eq. 3.9) this implies (for a given planetary mass) logarithmic spacing. Thus, the number of planets which fit into a habitable can be expected to be very roughly constant across stellar types. However, the closer-in HZ of later type stars does lead to a potential obstacle to life: For M type stars, planets in HZ will become tidally locked into synchronous rotation with the central star over a time less than the star's age. The resultant eternally-dark planetary hemisphere would likely be frozen and uninhabitable.

3.6.2 Water delivery

Water is the basis of life on Earth, but tracing the origin of the Earth's water is also central to understanding terrestrial planet formation. Most models of the protosolar nebula have too high a temperature at 1 AU for water to condense out of the gas (**?** and references therein). Hence, models of terrestrial planet formation generally invoke a way to deliver water from reservoirs at larger heliocentric distances. This idea is supported by the deuterium to hydrogen (D/H) ratio of the Earth's oceans, which is many times higher than expected in the protosolar nebula at 1 AU. Still uncertain is what fraction of such material came from the asteroid belt (e.g. **?**) versus the trans-Jupiter region (**? ? ? ?**), i.e. "wet asteroids" vs. comets. Observational evidence appears to favor the asteroid belt, since the D/H ratio of carbonaceous chondrites is closer to that of the Earth's oceans than is that of comets (**? ? ? ? ?**). One way to systematically deliver large amounts of icy material from beyond the snow line to the terrestrial region is via the late-stage formation mechanism outlined in §3.4.3.

3.6.3 The Late Heavy Bombardment

The cratering record on the Moon seems to suggest a spike in the impact rate about 700 Myrs after the formation of the Solar System (**?**). This cataclysmic event in our early history turns out to be a puzzle from the point of view of both celestial mechanics and the origin of life. Insofar as the latter is concerned, there exists some (controversial) carbon isotopic evidence for the existence of life 3.7 to 3.85 Gyrs ago (**?** and references therein), right around the time of the LHB. This is surprising since the LHB is likely to have sterilized the entire Earth (**?**). One explanation put forward by (**?**) is that around this time, life might simply have arisen repeatedly and on fairly short timescales, only to be continuously frustrated by catastrophic impacts. Thus as soon as the impacts ceased, life arose one last time and took hold. Another possibility is that life found a way to weather the impacts, either deep underground or in impact ejecta (**?**). Such ejecta may even have seeded life elsewhere in the Solar System, such as Mars—a sort of cosmic enactment of not having all the eggs in one basket.

From a dynamics point of view, the timing of the LHB has also been the cause of much study and speculation. As discussed in §3.3, the sweep-up of planetesimals in the terrestrial region proceeds on a timescale well below a million years. Where, then, do all the impactors for the LHB come from? Several different scenarios have been proposed. (**?**) suggested the breakup of an asteroid; however, to provide enough material for the LHB, the asteroid's size must have been an order of magnitude greater than Ceres. (**?**) showed that an extra planet beyond Mars, perhaps around 2 AU, could have been the culprit. Such a planet could have become unstable on a 700 Myr timescale, become eccentric, and crossing as-yet undepleted planetesimal reservoirs in the inner asteroid belt. These planetesimals would then have been perturbed onto orbits crossing the terrestrial region. The chief difficulty lies in preserving a significant planetesimal population between 2 AU and Jupiter for this length of time. Finally, (**?**) devised a model in which the LHB is triggered directly by Jupiter and Saturn. The giant planets almost certainly underwent a period of gradual, divergent migration as they exchanged angular momentum via the scattering of planetesimals (**? ?**). (**?**) posited an initially compact configuration for the giant planets, with Uranus and Neptune between 11 and 15 AU, and Jupiter and Saturn just inside their 1:2 mean-motion resonance. As planetesimals are scattered back and forth among the giant planets they migrate apart,

and at the moment when Jupiter and Saturn cross their 1:2 resonance divergently, both receive a kick in eccentricity, as previously shown by (**?**). This excitation, then, serves as the trigger for the LHB, delivering a large flux of planetesimals from the outer disk onto orbits that cross the terrestrial region. To a lesser extent, material remaining in the asteroid belt may also participate.

The trick is to make this event happen so late; (**?**) show that the timing of the migration can be controlled by the distance between the outermost giant planet and the inner edge of the planetesimal disk. Since many planetesimals will have already been accreted or scattered during the giant planets' formation, a distance of more than 1 AU between the giant planets and the inner disk edge, as in their best-fit model, is not implausible. However, although this model constitutes an elegant way to obtain an essentially arbitrarily long evolution followed by an abrupt shake-up, the initial locations of the giant planets need to be rather finely tuned, without any obvious cosmogonical justification.

3.6.4 More on the role of giant planets

In §3.4.3, we looked at one way in which the formation process of terrestrial planets in the Solar System may have been driven by the dynamical influence of Jupiter and Saturn. We now consider other, more general ways in which giant planets may affect terrestrial planets, with an emphasis on the issue of habitability.

While §3.4.3 describes a scenario in which the giant planets may help terrestrial planet formation along, it is just as possible for giant planets to hinder the formation of terrestrial planets. In a mature planetary system, the existence of a jovian planet in or near the HZ will make it unlikely terrestrial planets will have survived in the same region. However, even a stable HZ is no guarantee of a friendly environment for terrestrial planets. During the lifetime of the gas disk in the first few Myrs, substantial migration of planets, especially giant planets, is likely to occur due to gravitational planet-disk interactions (**? ? ?**). Numerous gas giants planets may in fact be lost during this time by migrating inside a disk gap ("Type II" migration) all the way into the central star (**?**). If gas giants form by core accretion (**?**), growing giant planet cores, ~ 10 $M_\oplus$ in mass, likely migrate even faster than gas giants, so that even a system in which no gas giant ever penetrates the terrestrial region may suffer an early period of being repeatedly transited by Neptune-sized bodies (**?**). Dynamical studies do suggest, though, that

even under such adverse conditions, some fraction of the (proto)planets in the terrestrial region can survive (**?** **?**).

(**?**) argued that the formation of gas giants ought to be unlikely around low-mass M dwarfs. They assumed that the disk mass scales with the stellar mass; this would make for a longer growth timescale and a lower final mass for planets grown about such stars ($M \leq 0.4\ M_{\odot}$). If gas giants grow by core accretion, this makes it less likely that cores of large enough mass to accrete the requisite massive gas envelope will grow within the time that the gas disk persists, 10 Myrs (GJ 876, M_{*}=0.3 $M_{\odot}$, orbited by two planets with $M \sin i = 0.6\ M_{\rm Jup}$ and $1.9\ M_{\rm Jup}$, is a notable counterexample). However, growing a terrestrial planet is a much less demanding proposition, and effectively the only time limit is the age of the system. Thus terrestrial planets ought to still be common around M dwarfs. In fact, with fewer gas giants to threaten their stability, it is conceivable that terrestrial planets are actually *more* common on average around lower-mass stars. Since M dwarfs are by far the most common stars, making up about 70% by number of the stars in the Galaxy, the total number of HZ-dwelling terrestrial planets is potentially very high.

Although a system with no giant planets at all may be a safer haven for terrestrial planet formation and survival, a lack of gas giants may negatively impact the habitability of the planets which do form, even those that end up square in their respective HZ. In the Solar System, Jupiter intercepts a large fraction of the comet flux which would otherwise cross the terrestrial region. It is thus often argued that without such a dynamical barrier, life would never have arisen and survived on our planet, and that a gas giant exterior to the HZ is an additional necessary condition for habitability. However, the issue is perhaps not quite as clear-cut as it appears at first glance; after all, Jupiter's perturbations also plays a large role in producing short-period comets from the Kuiper belt and long-period comets from the Oort cloud in the first place. In fact, outward scattering of planetesimals by Jupiter plays the main role in actually building the Oort cloud (**?**). Quantifying the difference in comet flux into the HZ of a mature planetary system with and without exterior gas giants is thus an important area for future work, and one which needs to be undertaken before we can properly assess the potential for low-mass stars to harbor life.

3.7 Summary

In our understanding of how terrestrial planets form, perhaps the most important aspect is that it seems to be *easy*. Numerical simulations have no problem producing such bodies in the inner 1 - 2 AU of a system. The sweep-up of planetesimals proceeds rapidly in this region, taking typically much less than 1 Myr. The simplest scenario, in which the resultant protoplanets then interact and collide in a gas-free environment, results in the formation of Earth-mass planets on a timescale of $\sim 10^8$ years for a "minimum-mass" distribution of solid material, i.e. one which begins with essentially just the mass of the present-day terrestrial planets. This is because planet formation in the inner few AU is a very efficient process; absent any systematic migration, almost all of the mass one puts in is built into planets. Thus terrestrial planet formation is fundamentally a very robust process, so much so that it may actually be difficult to *prevent* it from happening in any given system.

This being said, there are still a number of question marks about the details of how terrestrial planets grew in the Solar System and, by extension, how they form around other stars. The low eccentricities of Earth and Venus seem to require processes beyond the standard model. It is certainly possible that these low eccentricities will turn out to be a peculiarity of the Solar System; eccentricities < 0.1 are not in general necessary for long-term stability. Neither are they required for habitability. However, this "small detail" may in fact be pointing to a fundamentally different course of events in the last stage of formation, such as the one outlined in §3.4.3. Another puzzle is the nature of the Late Heavy Bombardment. What initiated it? Did life survive this event, or did it have to wait until the end of the LHB to arise? How likely are analogues to the LHB in other planetary systems? The probable ubiquity of terrestrial planet formation means that any given star ought to have a good chance of harboring one in its habitable zone. This prospect may be ruined if a giant planet exists close enough to the HZ to destabilize bodies within it. At the same time, though, the presence of (exterior) giant planets may be required in order to protect the HZ against comet bombardment. The habitable zone is defined by the presence of liquid surface water, but the source of that water is itself still not fully understood.

The discovery of the first extrasolar terrestrial planet will certainly be a momentous event. Until then, no amount of modeling can change the fact that our sample size of terrestrial planet systems is one. As

with the gas giants, surprises may await which overturn our currently-held picture of how such planets form. And of course, the search for exo-terrestrial planets is made especially exciting by the possibility that another life-bearing world is waiting to be discovered.

-1:2 scenario -Gladman idea for having stuff survive in ejecta

3.8 Other (where to put this?)

- M and L dwarfs (again): Laughlin & Adams say it's hard to form giant planets, but what the hell is GJ 876???
- from the "surviving Type I" paper: Migration may give us Neptune-mass "water worlds" in the HZ that originate from much larger radii
- more on migration: One might ask whether this can destroy terrestrial planet systems? some of this is in the PPV review; look there for references. Also mention that once the first giant planet has formed, it tends to act as a migration barrier (Thommes 2005)
- There's always that inconsistency between the mass needed in the terr. region and that needed for giant planets; perhaps we still are missing a part of the picture. [perhaps, also, this is an indication that some process, e.g. migration, removed a lot of the mass that was originally in the terrestrial region]

3.9 Summary

- terrestrial planets must be ubiquitous throughout the universe. They're dead-easy to form (how to get them completely circular is really just a finicky detail). The question to answer is, can a terrestrial-only system be a viable place for life to arise?
- exotic scenarios like life arising on a moon of a Jovian planet
- another exotic scenario: a "waterworld" brought in by migration

4

From Protoplanetary Disks to Prebiotic Amino Acids and the Origin of the Genetic Code

Paul Higgs & Ralph E. Pudritz
McMaster University

Abstract

Chapter forthcoming...

5

Emergent Phenomena in Biology: The Origin of Cellular Life

David Deamer
University of California, Santa Cruz

5.1 Introduction

The main themes of this chapter concern the phenomenon of emergence, the origin of life as a primary example of emergence, and how evolution begins with the inception of cellular life. The physical properties of certain molecular species are relevant to life's orogins, because these properties lead to the emergence of more complex structures by self-assembly. One such property is the capacity of amphiphilic molecules such as soap to form membranous boundary structures, familiar examples being soap bubbles and cell membranes. A second example is the chemical bonding that allows biopolymers such as nucleic acids and proteins to assemble into functional sequences. Self-assembly processes can produce complex supramolecular structures with certain properties of the living state. Such structures are able to capture energy available in the environment and initiate primitive reactions associated with metabolism, growth and replication. At some point approximately 4 billion years ago, cellular compartments appeared that contained macromolecular systems capable of catalyzed growth and replication. Because each cellular structure would be slightly different from all others, Darwinian evolution by natural selection could begin, with the primary selective factor being competition for energy and nutrients.

5.2 Defining Emergence

Researchers increasingly use the term emergence to describe processes by which more complex systems arise from seemingly simpler systems, typically in an unpredictable fashion. This usage is just the opposite of reductionism, the belief that any phenomenon can be explained by

understanding the parts of that system. Although reductionism has been a powerful tool in the sciences, the concept of emergence is also useful in referring to some of the most remarkable phenomena that we observe in nature and in the laboratory.

One example is the way that orderly arrangements of molecules appear when a relatively simple system of molecules undergoes a transition to a much more complex system. For instance, if we add soap molecules to water, at first there is nothing present except the expected clear solution of individual molecules dissolved in the water. At a certain concentration, additional molecules no longer dissolve, but instead begin to associate into small aggregates called micelles. As the concentration increases further, the micelles begin to grow into membranous layers that cause the originally clear solution to become turbid. Finally, if air is blown through a straw into the solution, much larger structures appear at the surface, the soap bubbles that are familiar to everyone.

Such emergent phenomena are called self-assembly processes, or sometimes self-organization. Typically the process is spontaneous, but in other instances an input of energy is required to drive self-assembly. One of the defining characteristics of an emergent property is the element of unpredictability. If we did not know by observation that soap molecules have properties of self-assembly, we could not have predicted that micelles would be produced simply by increasing the concentration of soap molecules in solution. Even knowing that micelles form, there is no physical theory available that can predict exactly what concentration of soap is required to form the micelles. And if soap bubbles were not such a common phenomenon, we would be astonished by their appearance when a soap solution is disturbed by addition of free energy in the form of an air stream.

5.3 Emergence of life: A very brief history

The origin of life was strongly dependent on physical conditions of the early Earth environment. Imagine that we could somehow travel back in time to the prebiotic Earth 3.9 billion years ago. The first thing we would notice is that it is very hot, with temperatures near the boiling point of water. Impacting objects the size of comets and asteroids add water, carbon dioxide, silicate minerals and small amounts of organic material to the planetary surface. They also produce craters on the moon that remain as a permanent record of the bombardment. The atmosphere has no oxygen, but instead is a mixture of carbon dioxide

and nitrogen. Land masses are present, but they are volcanic islands resembling Hawaii or Iceland, rather than continents. We are standing on one such island, on a beach composed of black lava rocks, with tide pools containing clear sea water. We can examine the sea water in the tide pools with a powerful microscope, but there is nothing to be seen, only a dilute solution of organic compounds and salts. If we could examine the mineral surfaces of the lava rocks, we would see that some of the organic compounds have formed a film adhering to the surface, while others assembled into aggregates that disperse in the sea water. Now imagine that we return 300 million years later. Not much has changed. The primary land masses are still volcanic in origin, although small continents are beginning to appear as well. The meteorite impacts have dwindled, and the global temperature has cooled to around 70 °C. We are surprised when we examine the tide pools, in which a turbidity has appeared that was not apparent earlier, and the mineral surfaces are coated with layers of pigmented films. When we examine the water and films with our microscope, we discover immense numbers of bacteria present in the layers. Life has begun. What happened in the intervening time that led to the origin of life? This is a fundamental question of biology, and the answer will change the way we think about ourselves, and our place in the universe, because if life could begin on the Earth, it could begin by similar processes on Earth-like planets circling other stars in our galaxy. The origin of life is the most extraordinary example of emergent phenomena, and the process by which life began must involve the same kinds of intermolecular forces and self-assembly processes that cause soap to form membranous vesicles and allow monomers like amino acids and nucleotides to form functional polymers. The origin of life must also have incorporated the reactions and products that are the result of energy flowing through a molecular system, thereby driving it toward ever more complex systems with emergent properties.

5.4 The first emergent phenomena: self-assembly processes on the early Earth

All cellular life today incorporates two processes we will refer to here as self-assembly and directed assembly. The latter process always involves the formation of covalent bonds by energy-dependent synthetic reactions, and requires that a coded sequence in one type of polymer in some way directs the sequence of monomer addition in a second polymeric species. The first process spontaneous self-assembly - occurs

when certain compounds interact through non-covalent hydrogen bonds, electrostatic forces and nonpolar interactions to form closed membrane-bounded microenvironments. The self-assembly of membranes is by far the simplest of the processes associated with the origin of cellular life. It occurs whenever lipid-like molecules are present in an aqueous environmnent, and does not require the covalent bond formation required for the synthesis of polymers such as proteins and nucleic acids. For this reason, it is reasonable to assume that membrane formation preceded the appearance of catalytic and replicating polymers. Boundary structures and the compartments they produce make energy available in the form of ion gradients, and can also mediate a selective inward transport of nutrients. Furthermore, membranous compartments in principle are capable of containing unique systems of macromolecules. If a replicating system of polymers could be encapsulated within a membrane-bounded compartment, the components of the system would share the same microenvironment, and the result would be a major step toward cellularity, speciation and true cellular function (Deamer and Oro, 1980; Morowitz et al. 1986; Morowitz, 1992; Dyson, 1995). We know very little about how this event might have occurred at the origin of cellular life, but recent advances have provided clues about possible sources of self-assembling molecules and processes by which large molecules can be captured in membrane-bounded microenvironments. Here we will describe the chemical and physical properties of such systems, and several experimental models that incorporate certain properties related to the origin of cellular life.

5.5 Sources of amphiphilic molecules

The most striking examples of self-assembling molecules are called amphiphiles, because they have both a hydrophilic (water loving) group and a hydrophobic (water hating) group on the same molecule. Amphiphilic molecules are among the simplest of life's molecular components, and are readily synthesized by non-biological processes. Virtually any hydrocarbon having ten or more carbons in its chain takes on amphiphilic properties if one end of the molecule incorporates a polar or ionic group. The simplest common amphiphiles are therefore molecules such as soaps, more technically referred to as monocarboxylic acids. A good example is decanoic acid, which is shown below with its ten carbon chain:

$$CH_3-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-CH_2-COOH$$

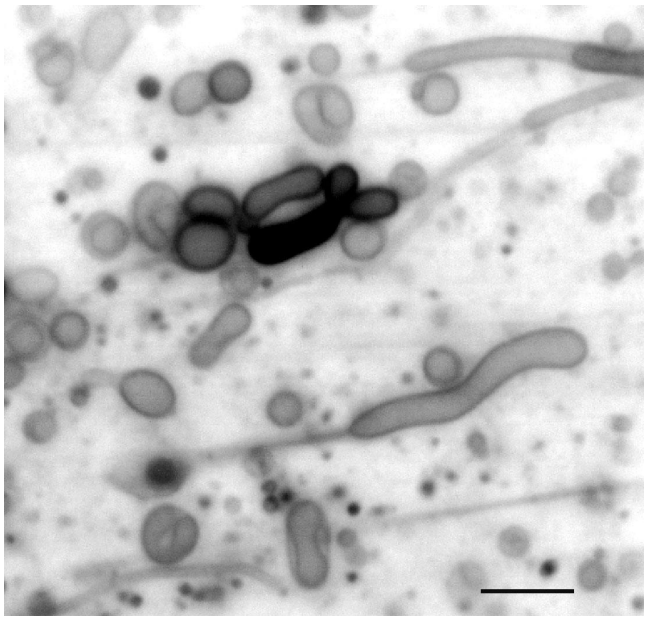


Fig. 5.1. Membranous vesicles are produced when decanoic acid is dispersed in aqueous salt solutions at neutral pH ranges.

McCollom et al. (1999) and Rushdi et al. (2002) demonstrated that a series of alkanoic acids and alcohols in this size range can be produced from very simple organic compounds by Fischer-Tropsch reactions that simulate geothermal conditions on the early Earth. We have found that such molecules readily form membranous vesicles, as shown in Figure 1 (Apel et al. 2002). This fact will become important later when we discuss the emergence of cellular life.

Two possible sources of organic compounds on a primitive planetary surface are delivery during late accretion, followed by chemical evolution, or synthesis by geochemical processes in the early atmosphere or hydrosphere. Earlier investigations focused on chemical synthesis of monomers common to the primary macromolecules involved in living systems, with the goal of determining whether it was possible that biologically relevant compounds were available on the primitive Earth. Most of these studies emphasized water-soluble compounds such as amino acids, nu-

cleobases and simple carbohydrates. The classic experiments of Miller and Urey (1953, 1959) showed that amino acids such as glycine and alanine could be obtained when a mixture of reduced gases was exposed to an electrical discharge. The mixture was assumed to be a simulation of the original terrestrial atmosphere which, by analogy with the outer planets, would have contained hydrogen, methane, ammonia and water vapor. At sufficiently high energy fluxes, such mixtures of reduced of gases generate hydrogen cyanide and formaldehyde, which in turn react produce amino acids, purines and a variety of simple sugars. The possibility that organic compounds could be synthesized under prebiotic conditions was given additional weight when it was convincingly shown that carbonaceous meteorites contained amino acids, hydrocarbons, and even traces of purines (Kvenholden 1970; Lawless and Yuen, 1979; Cronin et al. 1988). Such meteorites are produced by collisions in the asteroid belt between Mars and Jupiter. Asteroids range up to hundreds of kilometers in diameter, and are examples of planetesimals that happened to avoid accretion into the terrestrial planets. Asteroids and their meteoritic fragments therefore represent samples of the primitive solar system in which the products of prebiotic chemical reactions have been preserved for over 4.5 billion years. It was reasonable to assume that similar reactions and products were likely to have occurred on the Earth's surface. In the late 1970s it became increasingly clear that the archaean atmosphere was largely of volcanic origin and composed of carbon dioxide and nitrogen rather than the mixture of reducing gases assumed by the Miller-Urey model (Holland 1984; Kasting 1998). This is consistent with the fact that approximately 65 atmosphere equivalents of carbon dioxide are present in the Earth's crust as carbonate minerals, all of which must have passed through the atmosphere at some point as a gaseous component. Carbon dioxide does not support synthetic pathways leading to chemical monomers, so interest was drawn to the second potential source of organic material, extraterrestrial infall in the form of micrometeorites and comets. This was first proposed by Oro (1961) and Delsemme (1984) and more recently extended by Anders (1989) and Chyba and Sagan (1992). The total organic carbon added by extraterrestrial infall over 108 years of late accretion can be estimated to be in the rage of 1016 - 1018 kgs, which is several orders of magnitude greater than the 12.5 X 1014 kg total organic carbon in the biosphere (http://www.regensw.co.uk/technology/biomass-faq.asp). From such calculations it seems reasonable that extraterrestrial infall was a significant source of organic carbon in the prebiotic environment.

Even today meteorites and interplanetary dust particles (IDPs) deliver organic materials to the modern Earth at a rate of 106 kg/yr (Love and Brownlee, 1992; Maurette, 1998). The discovery of biologically relevant compounds in meteorites also indicated that organic synthesis can occur in the interstellar medium, which immediately leads to the question of sources and synthetic pathways. The most important biogenic elements (C, N, O, S, and P) form in the interiors of stars, then are ejected into the surrounding interstellar medium (ISM) at the end of the star's lifetime during red giant, nova, and supernova phases. Following ejection, much of this material becomes concentrated into dense molecular clouds from which new stars and planetary systems are formed (Ehrenfreund and Charnley, 2000; Sandford, 1996). At the low temperatures in these dark molecular clouds, mixtures of molecules condense to form ice mantles on the surfaces of dust grains where they can participate in additional chemical reactions. Comparison of infrared spectra of low temperature laboratory ices with absorption spectra of molecular clouds indicate that interstellar ices are mainly composed of H2O mixed with CO, CO2, CH3OH, NH3, and other components, the latter ingredients generally comprising 5 - 15 percent of the total. The ices are exposed to ionizing radiation in the form of cosmic rays (and secondary radiation generated by their interaction with matter), and UV photons from stars forming within the cloud. Laboratory experiments have shown that illuminating such ices with ultraviolet light leads to more complex molecular species (Greenberg 1993; Bernstein et al., 1995; Gerakines et al., 2000; Ehrenfreund et al., 2001). Hundreds of new compounds are synthesized, even though the starting ices contain only a few simple common interstellar molecules. Many of the compounds formed in these experiments are also present in meteorites, comets and interplanetary dust particles, and some are relevant to the origin of life, including amino acids (Munoz Caro et al., 2002; Bernstein et al., 2002, Bernstein et al., 2001), and amphiphilic material (Dworkin et al., 2001). Although it is now clear that organic molecules are synthesized in dense molecular clouds, the molecules must be delivered to habitable planetary surfaces if they are to take part in the origin of life. This requires that they survive the transition from the dense cloud into a protostellar nebula and subsequent incorporation into planetesimals, followed by delivery to a planetary surface. During the late bombardment period, which lasted until about 4 billion years ago, the amount of extraterrestrial organic material brought to the prebiotic Earth was likely to have been several orders of magnitude greater than current rates of infall (Chyba and

Sagan, 1992). Thus, the early Earth must have been seeded with organic matter created in the interstellar medium, protosolar nebula, and asteroidal/cometary parent bodies.

5.6 The emergence of primitive cells

Life on the Earth most likely arose from vast numbers of natural experiments in which various combinations of organic molecules were mixed and recombined to form complex interacting systems, then exposed to sources of energy such as light, heat, and oxidation-reduction potentials presented by donors and acceptors of electrons. This mixing and recombination probably did not occur in free solution, but rather in fluctuating environments at aqueous-mineral interfaces exposed to the atmosphere under conditions that would tend to concentrate the organic material so that reactions could occur. Through this process, incremental chemical evolution took place over a period of several hundred million years after the Earth had cooled sufficiently for water vapor to condense into oceans. At some point, membrane-bounded systems of molecules appeared that could grow and reproduce by using energy and nutrients from the environment. An observer seeing this end product would conclude that such systems were alive, but would be unable to pinpoint the exact time when the complex structures took on the property of life. Here we will assume that the structures described above would be recognizable cells, and that cellular compartments were required for life to begin. This assumption differs from the conjectures of Kauffman (1993) and Wchtershuser (1988) that life began as a series of reactions resembling metabolism. In their view, autocatalytic pathways were first established, perhaps on mineral surfaces. Over time the systems became increasingly complex to the point that self-reproducing polymers were synthesized, with cellular compartments appearing at a later stage. There is as yet little experimental evidence that allows a choice between the two perspectives of compartments first' or metabolism first'. However, because membrane structures readily self-assemble from amphiphilic compounds known to be present in mixtures of organic compounds available on the prebiotic Earth, it is highly likely that membranous compartments were among the first biologically relevant structures to appear prior to the origin of life. Evidence from phylogenetic analysis suggests that microorganisms resembling today's bacteria were the first form of cellular life. Traces of their existence can be found in the fossil record in Australian rocks at least 3.5 billion years old. Over the intervening years between life's

beginnings and now, evolution has produced bacteria which are more advanced than the first cellular life. The machinery of life has become so advanced that when researchers began subtracting genes in one of the simplest known bacterial species, they reached a limit of approximately 265-350 genes that appears to be absolute requirements for contemporary bacterial cells (Hutchinson et al. 1999). Yet life did not spring into existence with a full complement of 300+ genes, ribosomes, membrane transport systems, metabolism and the DNA –¿ RNA –¿ protein information transfer that dominates all life today. There must have been something simpler, a kind of scaffold life that was left behind in the evolutionary rubble. Can we reproduce that scaffold? One possible approach was suggested by the RNA World concept that arose from the discovery of ribozymes, which are RNA structures having enzyme-like catalytic activities. The idea was greatly strengthened when it was discovered that the catalytic core of ribosomes is not composed of protein at the active site, but instead is composed of RNA machinery. This remarkable finding offers convincing evidence that RNA likely came first, and then was overlaid by more complex and efficient protein machinery (Hoang et al., 2004). Another approach to discovering a scaffold is to incorporate one or a few genes into microscopic artificial vesicles to produce molecular systems that display certain features associated with life. The properties of such system may then provide clues to the process by which life began in a natural setting of the early Earth. What would such a system do? We can answer this question by listing the steps that would be required for a microorganism to emerge as the first cellular life on the early Earth:

- Boundary membranes self-assemble from amphiphilic molecules.
- Energy is captured either from light and a pigment system, or from chemical energy, or both.
- Ion concentration gradients are produced and maintained across the membrane by an energy-dependent process. The gradient is a source of energy to drive metabolism and synthetic reactions.
- The energy is coupled to the synthesis of activated monomers, which in turn are used to make polymers.
- Macromolecules are encapsulated, yet smaller nutrient molecules can cross the membrane barrier.
- The macromolecules grow by polymerizing the nutrient molecules, using the energy of metabolism.

- Macromolecular catalysts speed the metabolic reactions and growth processes.
- The macromolecular catalysts themselves are reproduced during growth.
- Information is contained in the sequence of monomers in one set of polymers, and this set is duplicated during growth. The information is used to direct the growth of catalytic polymers.
- The information is used to direct the growth of catalytic polymers.
- The membrane-bounded system of macromolecules can divide into smaller structures.
- Genetic information is passed between generations by duplicating the sequences and sharing them between daughter cells.
- Occasional errors (mutations) are made during replication or transmission of information so that a population of primitive living organisms can evolve through selection.

Looking down this list, one is struck by the complexity of even the simplest form of life. This is why it has been so difficult to define life in the usual sense of a definition, that is, boiled down to a few sentences in a dictionary. Life is a complex system that cannot be captured in a few sentences, so perhaps a list of its observed properties is the best we can ever hope to do. Given the list, one is also struck by the fact that all but one of the functions - self-reproducing polymers - have now been reconstituted individually in the laboratory. For instance, it was shown 40 years ago that lipid vesicles self-assemble into bilayer membranes that maintain ion gradients (Bangham et al 1965). If bacteriorhodopsin is in the bilayer, light energy can be captured in the form of a proton gradient (Oesterhelt and Stoeckenius, 1973) . If an ATP synthase is in the membrane, the photosynthetic system can make ATP by coupling the proton gradient to forming a pyrophosphate bond between ADP and phosphate (Racker and Stoeckenius, 1974) . Macromolecules such as proteins and nucleic acids can be easily encapsulated and can function in the vesicles as catalysts (Chakrabarti et al 1996) and as membrane transport agents. The system can grow by addition of lipids, and can even be made to divide by imposing shear forces, after which the vesicles grow again (Hanczyc et al. 2003). Macromolecules like RNA can grow by polymerization in the vesicle, driven by catalytic proteins (Monnard and Deamer, 2006). And finally, samples of cytoplasm from a living cell like E. coli are easily captured, including ribosomes. A micrograph of such vesicles is shown in Figure 2.

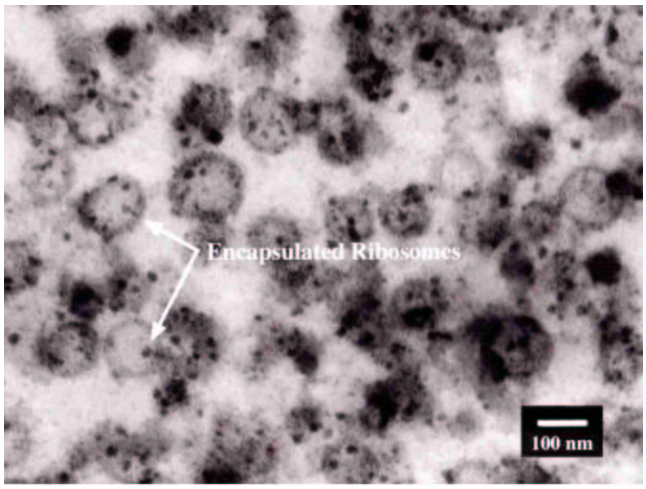


Fig. 5.2. Ribosomes from E. coli encapsulated in phospholipid vesicles. The vesicles were reconstituted from a detergent-lipid solution in the presence of cytoplasmic extracts from the bacteria, and 10 or so ribosomes were present in each vesicle. Micrograph courtesy of Z. Martinez.

The ability to capture structures as large as ribosomes has led to attempts to demonstrate translation in closed vesicles. This was first achieved by Yu et al. (2001) and later by Nomura et al. (2003) who captured samples of bacterial cytoplasm from E. coli in large liposomes. The samples included ribosomes, transfer RNAs and the hundred or so other components required for protein synthesis. The mRNA containing the gene for green fluorescent protein (GFP) was also included in the mix, which permitted facile detection of protein synthesis. The encapsulated translation systems worked, but only a few molecules of GFP were synthesized in each vesicle, because the only amino acids available were those captured within the vesicle. This limitation was resolved by Noireaux and Libchaber (2004), who included not only the mRNA for GFP in the mix, but also a second mRNA coding for the pore-forming protein alpha hemolysin. The hemolysin produced a channel in the lipid bilayer that allowed externally added nutrients in the form of amino acids and ATP to cross the membrane barrier and supply the translation process with energy and monomers. (Figure 3) The system worked well, and GFP synthesis continued for as long as four days.

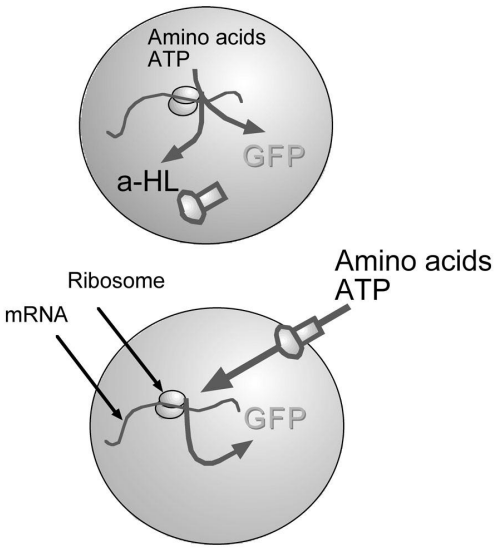


Fig. 5.3. Translation in a microenvironment. In the upper figure, amino acids and ATP encapsulated in the vesicle volume are used to make small amounts of green fluorescent protein (GFP) and alpha hemolysin (a-HL). The membrane prevents access to external amino acids and ATP, so the system will quickly exhaust the nutrients trapped in the vesicle. However, because the a-HL is a pore forming protein, it migrates to the membrane and assembles into a heptamer with a pore large enough to permit amino acids and ATP into the vesicle. Now the system can synthesize significant amounts of new protein, and the vesicle begins to glow green from the accumulation of GFP. Something similar must have happened on the pathway to the origin of cellular life so that the fist cells could have access to nutrients available in the environment.

These advances permit us to consider whether it might be possible to fabricate a kind of artificial life which is reconstituted from a complete system of components isolated from microorganisms. The system would by definition overcome the one limitation described above, the lack of a self-reproducing set of polymers, because everything in the system would grow and reproduce, including the catalytic macromolecules themselves and the lipid components of the boundary membrane. However, such a system requires that the genes for the translation system (ribosomes, tRNA, and so on) for DNA replication and transcription, and for lipid

synthesis must all be present in a strand of synthetic DNA. When one adds up the number of essential genes, total is over a hundred. This might at first seem like a daunting task, but in fact it is within the realm of possibility, in that the complete set of genes required to synthesize the polio virus was recently assembled in a strand of synthetic DNA (Cello et al. 2002). This thought experiment clearly points the limitations of our current understanding of the origin of cellular life. It does make clear that life could not begin as a complex molecular system of a hundred or more genes required to fabricate the simplest possible artificial cell that uses DNA, RNA and ribosomes for self-reproduction. Instead, as noted earlier, there must have been a kind of molecular scaffold that was much simpler, yet had the capacity to evolve into the living systems we observe today. Is there any hope that we might discover such a system? One possible lead is to find a ribozyme that can grow by polymerization, in which the ribozyme copies a sequence of bases in its own structure (Johnston et al. 2001). So far, the polymerization has only copied a string of 14 nucleotides, but this is a good start. If a ribozyme system can be found that catalyzes its own complete synthesis using genetic information encoded in its structure, it could rightly be claimed to have the essential properties that are lacking so far in artificial cell models: reproduction of the catalyst itself. Instead of the hundred or so genes necessary for the translation system described above, the number is reduced to just a few genes that allow the ribozyme to control its own replication, synthesize catalysts related to primitive metabolism, including synthesis of membrane components, and perhaps a pore-forming molecule that allows nutrient transport. Given such a ribozyme, it is not difficult to imagine its incorporation into a system of lipid vesicles that would have the basic properties of the living state.

5.7 Self-assembly processes in prebiotic organic mixtures

We can now return to considering process that were likely to occur on the prebiotic Earth, ultimately leading to the origin of life and initiating Darwinian evolution. We will first ask what physical properties are required if a molecule is to become incorporated into a stable cellular compartment. As discussed earlier, all membrane-forming molecules are amphiphiles, with a hydrophilic head and a hydrophobic tail on the same molecule. If amphiphilic molecules were present in the mixture of organic compounds available on the early Earth, it is not difficult to imagine that their self-assembly into molecular aggregates was

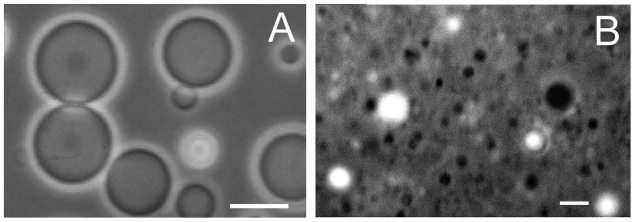


Fig. 5.4. Membranous compartments are produced by self-assembly of amhiphilic molecules extracted from the Murchison meteorite. The larger vesicles have a diameter of 20 30 micrometers.

a common process. Is this a plausible premise? In order to approach this question, we can assume that the mixture of organic compounds in carbonaceous meteorites such as the Murchison meteorite resembles components available on the early Earth through extraterrestrial infall. A series of organic acids represents the most abundant water-soluble fraction in carbonaceous meteorites (Cronin, 1988). We have extracted samples of the Murchison meteorite in an organic solvent commonly used to extract membrane lipids from biological sources (Deamer and Pashley, 1989). When this material was allowed to interact with aqueous phases, one class of compounds with acidic properties was clearly capable of forming membrane-bounded vesicles (Figure 4).

From these results, it is reasonable to conclude that a variety of simpler amphiphilic molecules were present on the early Earth that could participate in the formation of primitive membrane structures. Even if membranous vesicles were commonplace on the early Earth and had sufficient permeability to permit nutrient transport to occur, these structures would be virtually impermeable to larger polymeric molecules that were necessarily incorporated into molecular systems on the pathway to cellular life. The encapsulation of macromolecules in lipid vesicles has been demonstrated by hydration-dehydration cycles that simulate an evaporating lagoon (Shew and Deamer, 1993). Molecules as large as DNA can be captured by such processes. For instance, when a dispersion of DNA and fatty acid vesicles is dried, the vesicles fuse to form a multilamellar sandwich structure with DNA trapped between the layers. Upon rehydration, vesicles reform that contain highly concentrated DNA, a process that can be visualized by staining with a fluorescent dye (Figure 5).

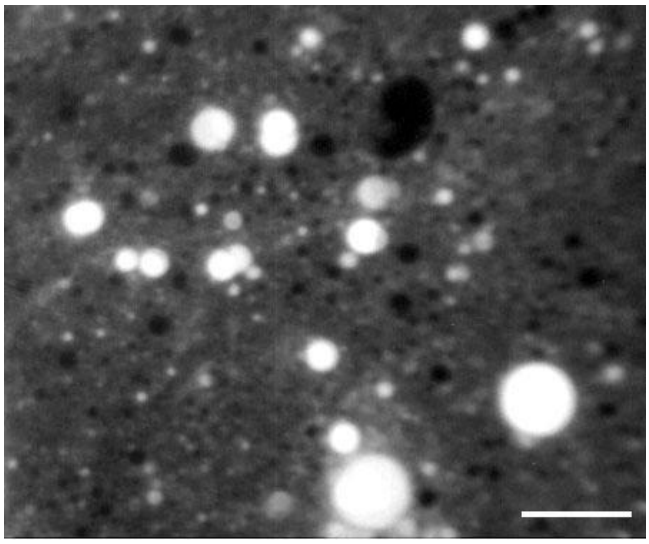


Fig. 5.5. Membranous compartments composed of very simple amphiphiles such as a fatty acid are capable of encapsulating macromolecules in stable vesicles referred to as protocells. In the image above, DNA molecules have been trapped in decanoic acid vesicles. The DNA is stained with acridine orange so that it can be visualized by fluorescence microscopy. The larger vesicles have a diameter of 20 30 micrometers

5.8 Emergence of membrane functions

Membranes have many functions in addition of acting as a container for the macromolecular polymers of life. Three primary membrane functions associated with a protocell would include selective inward transport of nutrients from the environment, capture of the energy available in light or oxidation-reduction potentials, and coupling of that energy to some form of energy currency such as ATP in order to drive polymer synthesis. The simplest of these functions is that of a permeability barrier which limits free diffusion of solutes between the cytoplasm and external environment. Although such barriers are essential for cellular life to exist, there must also be a mechanism by which selective permeation allows specific solutes to cross the membrane. In contemporary cells, such processes are carried out by transmembrane proteins that act as channels and transporters. Examples include the proteins that facilitate the transport of glucose and amino acids into the cell, channels that allow potassium and sodium ions to permeate the membrane, and active

transport of ions by enzymes that use ATP as an energy source. It seems unlikely that the first living cellular systems had evolved such highly specialized membrane transport systems, which brings up the question of how early cells overcame the membrane permeability barrier. One possibility is that simple diffusion across the bilayer may have been sufficient. To give a perspective on permeability and transport rates by diffusion, we can compare the fluxes of relatively permeable and relatively impermeable solutes across contemporary lipid bilayers. The measured permeability of lipid bilayers to small, uncharged molecules such as water, oxygen and carbon dioxide is greater than the permeability to ions by a factor of 109. Such values mean little by themselves, but make more sense when put in the context of time required for exchange across a bilayer: half the water in a liposome exchanges in milliseconds, while potassium ions have half-times of exchange measured in days. We can now consider some typical nutrient solutes like amino acids and phosphate. Such molecules are ionized, which means that they would not readily cross the permeability barrier of a lipid bilayer. Permeability coefficients of liposome membranes to phosphate, amino acids and nucleotides have been determined (Chakrabarti, 1994; Monnard 2002) and were found to be similar to ionic solutes such as sodium and chloride ions. From these figures one can estimate that if a primitive microorganism depended on passive transport of phosphate across a lipid bilayer composed of a typical phospholipid, it would require several years to accumulate phosphate sufficient to double its DNA content, or pass through one cell cycle. In contrast, a modern bacterial cell can reproduce in as short a time as 20 minutes. If solutes like amino acids and phosphate are so impermeant, how could primitive cells have had access to these essential nutrients? One clue may be that modern lipids are highly evolved products of several billion years of evolution, and typically contain hydrocarbon chains 16 to 18 carbons in length. These chains provide an interior oily portion of the lipid bilayer that represents a nearly impermeable barrier to the free diffusion of ions such as sodium and potassium. The reason is related to the common observation that oil and water don't mix. Using more technical language, ion permeation of the hydrophobic portion of a lipid bilayer faces a very high energy barrier which is associated with the difference in energy for an ion in water to dissolve in the oily interior of a lipid bilayer composed of hydrocarbon chains. However, recent studies have shown that permeability is strongly dependent on chain length (Paula et al. 1996). For instance, shortening phospholipid chains from 18 to 14 carbons increases perme-

ability to ions by nearly three orders of magnitude. The reason is that thinner membranes have increasing numbers of transient defects that open and close on nanosecond time scales, so that ionic solutes can get from one side of the membrane to the other without dissolving in the oily interior phase of the bilayer. Ionic solutes even as large as ATP can diffuse cross a bilayer composed of a 14 carbon phospholipid (Monnard and Deamer 2001). On the early Earth, shorter hydrocarbon chains would have been much more common than longer chain amphiphiles, suggesting that the first cell membranes were sufficiently leaky so that ionic and polar nutrients could enter, while still maintaining larger polymeric molecules in the encapsulated volume. There are several implications of this conjecture. First, if a living cell is to take up nutrients by diffusion rather than active transport, the nutrients must be at a reasonably high concentration, perhaps in the millimolar range. It is doubtful that such high concentrations would be available in the bulk phase medium of a lake or sea, which suggests that early life needed to be in an environment such a small pond undergoing periodic wet-dry cycles that would concentrate possible nutrients. On the other hand, that same cell also was likely to have metabolic waste products, so it could not be in a closed system or it would soon reach thermodynamic equilibrium in which the waste product accumulation inhibits forward reactions required for metabolism and synthesis. And finally, a leaky membrane may allow entry of nutrients, but it would also be useless for developing the ion gradients that are essential energy sources for all modern cells. All of these considerations should guide our thinking as we attempt to deduce conditions that would be conducive for the origin of cellular life.

5.9 Emergence of growth processes in primitive cells

Earlier reports, (Walde et al. 1994) showed that vesicles composed of oleic acid can grow and reproduce as oleoyl anhydride spontaneously hydrolyzed in the reaction mixture, thereby adding additional amphiphilic components (oleic acid) to the vesicle membranes. This approach has recently been extended by Hanczyc et al. (2003) who prepared myristoleic acid membranes under defined conditions of pH, temperature and ionic strength. The process by which the vesicles formed from micellar solutions required several hours, apparently with a rate limiting step related to the assembly of nuclei of bilayer structures. However, if a mineral surface in the form of clay particles was present, the surface in some way catalyzed vesicle formation, reducing the time required from

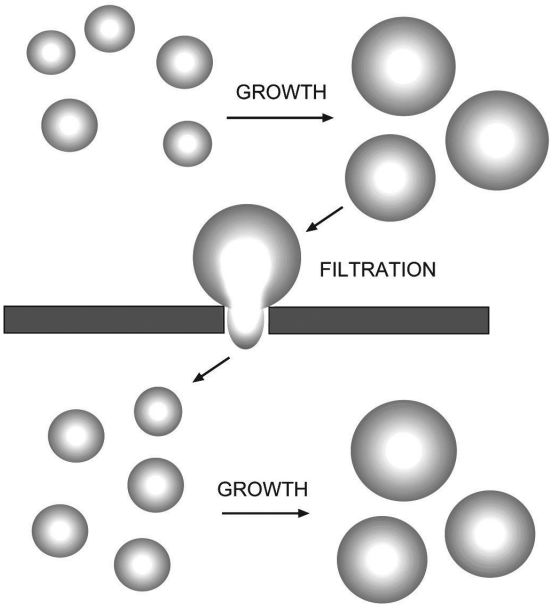


Fig. 5.6. Growth and division of lipid vesicles. (See text for details.)

hours to a few minutes. The clay particles were spontaneously encapsulated in the vesicles. The authors further found that RNA bound to the clay was encapsulated as well. In a second series of experiments, Hanczyc et al. (2004) showed that the myristoleic acid vesicles could be induced to grow by addition of fatty acid to the medium, presumably by incorporating fatty acid molecules into the membrane, rather than by fusion of vesicles. If the resulting suspension of large vesicles was then filtered through a polycarbonate filter having pores 0.2 micrometer in diameter, the larger vesicles underwent a kind of shear-induced division to produce smaller vesicles which could undergo further growth cycles (Figure 6). This remarkable series of experiments clearly demonstrated the relative simplicity by which complex system of lipid, genetic material and mineral catalysts can produce a model protocellular structure that can undergo a form of growth and division.

5.10 Environmental constraints on the first forms of life

Although self-assembly of amphiphilic molecules promotes the formation of complex molecular systems, the physical and chemical properties of an aqueous phase can significantly inhibit such processes, possibly constraining the environments in which cellular life first appeared. One such constraint is that temperature strongly influences the stability of vesicle membranes. It has been proposed that the last common ancestor, and even the first forms of life, were hyperthermophiles that developed in geothermal regions such as hydrothermal vents (Baross et al 1983) or deep subterranean hot aquifers (Pace, 1991). Such environments have the advantage of providing chemical energy in the form of redox potentials as well as abundant mineral surfaces to act as potential catalysts and adsorbants. However, because the intermolecular forces that stabilize self-assembled molecular systems are relatively weak, it is difficult to imagine how lipid bilayer membranes assembling from plausible prebiotic constituents would be stable under these conditions. All hyperthermophiles today have highly specialized lipid components, and it seems likely that these are the result of more recent adaptation than a molecular fossil of early life. A second concern is related to the ionic composition of a marine environment. The high salt concentration of the present ocean has the potential to exert significant osmotic pressure on any closed membrane system. All marine organisms today have highly evolved membrane transport systems that allow them to maintain osmotic equilibrium against substantial salt gradients across their membranes. Furthermore, the concentrations of divalent cations, in particular Mg2+ and Ca2+, were likely to exceed 10 mM in the early oceans, similar to their concentrations in sea water today. In the absence of oxygen, Fe2+ would also be present at millimolar concentrations., rather than the micromolar levels in contemporary sea water At concentrations in the millimolar range, all such divalent cations have a strong tendency to bind to the anionic head groups of amphiphilic molecules. This cause aggregation and precipitation, strongly inhibiting their ability to form stable membranes (Monnard et al. 2002). These considerations suggest that, from the perspective of membrane biophysics, the most plausible planetary environment for the origin of life would be at moderate temperature ranges (¡60 °C), and the ionic content would correspond to low ionic strength and pH values near neutrality (pH 5 - 8) with divalent cations at submillimolar concentrations. This suggestion is in marked contrast to the view that life most likely began in a marine environ-

ment, perhaps even the extreme environment of a hydrothermal vent. On argument favoring a marine site for life's beginning is that fresh water would be rare on the early Earth. Even with today's extensive continental crust, fresh water only represents 1

Acknowledgements

Portions of this chapter were adapted from a previous review: (Deamer et al. 2002).

References

[124] Baross, J.A. and S.E. Hoffman. (1983). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life, *Origins of Life* **15**, –327 - 235.

[127] Bernstein, M. P., Sandford, S. A., Allamandola, L. J., Chang, S., and Scharberg, M. A. (1995). Organic compounds produced by photolysis of realistic interstellar and cometary ice analogs containing methanol, *Astrophys* , –454, 327-344.

[127] Bernstein, M.P., Dworkin, J. P., Sandford, S. A. and Allamandola, L. J. (2001). Ultraviolet irradiation of naphthalene in H2O Ice: Implications for meteorites and biogenesis, *Meteoritics Planet* , –36, 351.

[127] Bernstein, M. P., Dworkin, J. P., Sandford, S. A., Cooper, G. W. and Allamandola, L. J. (2002). The formation of racemic amino acids by ultraviolet photolysis of interstellar ice analogs, *Nature* **416**, –401 - 403.

[521] Chyba, C.F. and Sagan, C. (1992). Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: An inventory for the origin of life, *Nature* **355**, –125-130.

[129] Cronin J.R., Pizzarello S. and Cruikshank, D.P. (1988). In Meteorites and the Early Solar System, Kerridge JF and Matthews MS, (eds, , – p. 819.

[130] Chakrabarti, A., Joyce, G.F., Breaker, R.R. and Deamer, D.W. (1994). RNA synthesis by a liposome-encapsulated polymerase, *J* , –39, 555-559.

[490] Deamer, D.W. and Oro, J. (1980). Role of lipids in prebiotic structures, *Biosystems* **12**, –167-75.

[134] Deamer, D.W. and G. L. Barchfeld. (1982). Encapsulation of macromolecules by lipid vesicles under simulated prebiotic conditions, *J* , –18, 203-206.

[134] Deamer, D.W, Dworkin, J. P., Sandford, S. A., Bernstein, M.P. and Allamandola, L. J. . (2002). The first cell membranes, *Astrobiology* , –2, 371 - 382.

[134] Deamer, D.W. and Pashley, R.M. (1989). Amphiphilic components of carbonaceous meteorites, *Orig* , –19, 21-33.

[135] Delsemme, A. (1984). The cometary connection with prebiotic chemistry, *Origins of Life* **14**, –51-60.

[136] Dworkin, J. P., Deamer, D. W., Sandford, S. A., and Allamandola, L. J. (2001). Self-assembling amphiphilic molecules: Synthesis in simulated interstellar/precometary ices, *Proc* , –98, 815-819.

[137] Dyson F. (1999). The Origins of Life, *Princeton University Press* , –

[477] Ehrenfreund, P. and Charnley, S. B. (2000). Organic molecules in the interstellar medium, comets, and meteorites: A voyage from dark clouds to the early Earth, *Ann* , –38, 427-483.

[477] Ehrenfreund, P., d'Hendecourt, L., Charnley, S. B., and Ruiterkamp, R. (2001). Energetic and thermal processing of interstellar ices, *J* , –106, 33291-33302.

[140] Gerakines, P. A., Moore, M. H., and Hudson, R. L. (2000). Energetic processing of laboratory ice analogs: UV photolysis versus ion bombardment, *J* , –106, 3338

[536] Greenberg, M., and Mendoza-Gomez, C.X. (1993). In: The Chemistry of Life's Origins, *Greenberg* , –1.

[143] Hanczyc M.M., Fujikawa S.M., and Szostak, J.W. (2003). Experimental models of primitive cellular compartments: encapsulation, growth, and division, *Science* **302**, –618 - 622.

[143] Hanczyc, M.M., and Szostak, J.W. (2004). Replicating vesicles as models of primitive cell growth and division, *Curr* , –28, 660 - 664.

[144] Holland, H.D. (1984). The Chemical Evolution of the Atmosphere and Oceans, *Princeton University Press* , –

[145] Hoang, L., Fredrick, K. and Noller H.F. (2004). Creating ribosomes with an all-RNA 30S subunit P site, *Proc Natl Acad Sci U S A* , –101, 12439-43. Hutchison, C., Peterson, S., Gill, S., Cline, R., White, O., Fraser, C., Smith, H. and Venter, J. 1999. Global transposon mutagenesis and a minimal Mycoplasma genome Science 286, 2165-2169. Ishikawa K, Sato K, Shima Y, Urabe I, Yomo T. (2004) Expression of a cascading genetic network within liposomes. FEBS Letters 576, 387.

[550] Kasting J.F. and Brown L.L. (1998). , *In* , –35.

[147] Kauffman, S. (1993). The Origin of Order, *Self* , – Oxford University Press, New York.

[148] Knauth, L.P. (2005). Temperature and salinity history of the Precambran ocean: implications for the course of microbial evolution, *Palaegeog* , –219, 53-69. Kvenvolden, K.A., Lawless, J.G., Pering, K., Peterson, E., Flores, J., Ponnamperuma, C., Kaplan, I.R. and Moore, C. 1970. Nature 28, 923.

[149] Lawless, J.G. and Yuen, G.U. (1979). Quantitation of monocarboxylic acids in the Murchison carbonaceous meteorite, *Nature* **282**, –396-398. Love, S. G., and Brownlee, D.E. (1993). A direct measurement of the terrestrial mass accretion rate of cosmic dust. Science 262, 550-553. Luisi, P.L. (1998) About various definitions of life. Orig Life Evol Biosph. 28:613-22.

[150] Maurette M. (1998). In: The Molecular Origins of Life, *A* , –147.

[151] McCollom, T.M., Ritter, G. and Simoneit, B.R.T. (1999). Lipid synthesis under hydrothermal conditions by Fischer-Tropsch-type reactions, *Orig* , –29,153-166.

[153] Miller, S. L. (1953). Production of amino acids under possible primitive Earth conditions, *Science* **117**, –528-529.

[153] Miller, S. L. and Urey, H. C. (1959). Organic compound synthesis on the primitive Earth, *Science* **130**, –245-251.

[157] Monnard, P.-A., Apel, C. L., Kanavarioti, A. and Deamer, D. W. (2002). Influence of ionic solutes on self-assembly and polymerization processes related to early forms of life: Implications for a prebiotic aqueous medium,

Astrobiology **2**, –139 - 152.
[157] Monnard, P.-A. and D. W. Deamer. (2001). Loading of DMPC based liposomes with nucleotide triphosphates by passive diffusion: A plausible model for nutrient uptake by the protocell, *Orig* , –31, 147-155.
[157] Monnard P-A. and Deamer, D.W. (2006). Models of primitive cellular life: polymerases and templates in liposomes, *Phil* , –
[157] Monnard, P.-A., Apel, C. L., Kanavariot,i A. and Deamer, D. W. (2002). Influence of ionic solutes on self-assembly and polymerization processes related to early forms of life: Implications for a prebiotic aqueous medium, *Astrobiology* **2**, –139 - 152.
[159] Morowitz, H.J. (1992). Beginnings of Cellular Life, *New Haven* , –
[159] Morowitz, H.J., Heinz, B. and Deamer D.W. (1988). The chemical logic of a minimum protocell, *Orig Life Evol Biosph* , –18, 281-7.
[160] Munoz-Caro, G.M., Meierhenrich, W.A., Schutte, W.A., Barbier, B., Arcones Segovia, A., Rosenbauer, W., Thriemann, H.P., Brack, A, and Greenberg, J.M. (2002). Amino acids from ultraviolet irradiation of interstellar ice analogues, *Nature* **416**, –403-406
[161] Noireaux V. and Libchaber A. (2004). A vesicle bioreactor as a step toward an artificial cell assembly, *Proc Natl Acad Sci U S A* , 2–139 - 152., 17669-74.
[162] Nomura, S., Tsumoto, K., Hamada, T., Akiyoshi, K., Nakatani, Y. and Yoshikawa, K. Gene expression within cell-sized lipid vesicles. (2003). Chem, *Biochem* , –4, 1172-1175.
[163] Oesterhelt, D. and Stoeckenius, W. (1973). Functions of a new photoreceptor membrane, , – Proc. Natl. Acad. Sci. U S A. 70, 2853-7
[164] Or, J. (1961). Comets and the formation of biochemical compounds on the primitive Earth, *Nature* **190**, –389-390.
[165] Pace, N.R. (1991). Origin of life - Facing up to the physical setting, *Cell* **65**, –531.
[166] Paula, S, Volkov, A.G., Van Hoek, A.N., Haines, T.H. and Deamer, D.W. (1996). Permeation of protons, potassium ions, and small polar molecules through phospholipid bilayers as a function of membrane thickness Biophys, *J* , –70, 339.
[167] Racker, E. and Stoeckenius W. (1974). Reconstitution of purple membrane vesicles catalyzing light-driven proton uptake and adenosine triphosphate formation, *J Biol Chem* , –249, 662-3.
[168] Rasmussen, S., Chen, L., Deamer, D., Krakauer, D.C., Packard, N.H., Stadelr, P.F., and Bedau, M.A. (2004). Evolution, *Transitions from nonliving to living matter* , –303, 963 - 966.
[169] Rushdi, A.I. and B. Simoneit. (2001). Lipid formation by aqueous Fischer-Tropsch type synthesis over a temperature range of 100 to 400o C, *Orig* , –31, 103 -118.
[170] Sandford, S. A., Bernstein, M. P., and Dworkin, J. P. (2001). Assessment of the interstellar processes leading to deuterium enrichment in meteoritic organics, *Meteoritics Planetary Sci* , –36, 1117-1133.
[171] Shew, R. and D.W. Deamer. (1985). A novel method for encapsulation of macromolecules in liposomes, *Biochim* , –816, 1-8.
[172] Singer, S.J. and Nicolson, G.L. (1972). The fluid mosaic model, *Science* **175**, –720 - 727.
[173] Szostak, J.W., Bartel, D.P. and Luisi, P.L. (2001). Synthesizing life, *Nature* **409**, –387-390. Wchtershuser, G. 1988 Before enzymes and tem-

plates: Theory of surface metabolism. Microbiol Rev. 52, 452-484.
[175] Walde. P., Wick, R., Fresta, M., Mangone, A. and Luisi, P.L. (1994). Autopoietic self-reproduction of fatty acid vesicles, *J* , –116, 11649.
[175] Walde, P., A. Goto, P.-A. Monnard, M. Wessicken and P.L. Luisi. (1994). Oparin's reactions revisited: Enzymatic synthesis of poly(adenylic acid) in micelles and self-reproducing vesicles, *J* , –116, 7541-7547.
[176] Yu, W., Sato, K., Wakabayashi, M., Nakaishi, T., K-Mitamura, E.P., Shima, Y., Urabe, I. and Yomo, T. (2001). J, *Biosci* , –92, 590.

6

Paper forthcoming...

Lynn Rothschild

7

Hyperthermophilic Life on Earth and on Mars?

Karl O. Stetter
Universitãt Regensburg

7.1 Introduction

Every living organism is adapted to a specific growth temperature. In the case of humans, this is 37 °C and an increase by 5 °C becomes fatal. In the world of microbes, the growth temperature range is much more diverse: heat lovers (thermophiles) grow optimally (fastest) at temperatures up to 65 °C (Brock, 1978; Castenholz, 1979). Since the time of Pasteur, it had been assumed generally that growing (vegetative) cells of bacteria were killed quickly by temperatures of 80 °C and above. The Pasteurization technology is based on this observation. In contrast, during the past few decades, hyperthermophiles (HT; Stetter, 1992) that exhibit unprecedented optimal growth temperatures in excess of 80 °C have been isolated (Stetter et al., 1981; Zillig et al., 1981; Stetter, 1982). HT turned out to be very common in hot terrestrial and submarine environments. In comparing the growth requirements of these present-day HT with the conditions on ancient Earth, similar organisms could or even should have existed already by early Archaean times. Propelled by impact energy, microbes could have spread in between the planets and moons of the early solar system. Is there any evidence for the existence of microbes at that time? Most likely, yes. But the recognition of ancient microfossils on the basis of morphology turned out to be difficult, leading to controversy. Nevertheless, there are chemical traces of life within rocks from Precambrian deep sea vents (Schopf et al., 1987; Brasier et al., 2002; van Zullen et al., 2002).

In my research, I search for HT, their environments, properties and modes of living. So far, I have gathered a collection of more than 1,500 strains of HT, among them are isolates that survive autoclaving at 121°C. Whether such organisms are growing slowly or just surviving

Fig. 7.1. Solfataric field at Kafla, Iceland.

at 121°C is unresolved (Blöchl et al., 1997; Kashefi and Lovley, 2003; Cowan, 2004).

7.2 Biotropes

HT form complex communities within water-containing geothermally and volcanically heated environments situated mainly along terrestrial and submarine tectonic spreading and subduction zones. Availability of liquid water is a fundamental prerequisite of life. At an increased boiling point of water (for example by elevated atmospheric, hydrostatic or osmotic pressure), several HT exhibit growth temperatures exceeding 100°C. Due to the presence of reducing gasses like H_2S and the low solubility of O_2 at high temperatures, the biotopes of HT essentially are oxygen-free (anaerobic). HT have been isolated from terrestrial and submarine environments.

7.2.1 Terrestrial biotopes

Natural terrestrial biotopes of HT are mainly hot springs and sulfur-containing solfataric fields (named after Solfatara, Italy), with a wide range of pH values (pH 0 9.0) and usually low salinity (0.1-0.5

7.2.2 Marine biotopes

Marine biotopes of HT consist of various hydrothermal systems situated at shallow to abyssal depths. Similar to ambient sea water, submarine hydrothermal systems usually contain high concentrations of NaCl and sulfate and exhibit a slightly acidic to alkaline pH (5 8.5). Otherwise, the major gasses and life-supporting mineral nutrients can be similar to those in terrestrial thermal areas. Shallow submarine hydrothermal systems are found in many parts of the world, mainly on beaches with active volcanism, like at Vulcano Island, Italy, with temperatures of 80 to 105 °C.

Most impressive are the deep sea "smoker" vents (Fig. 2), where mineral-laden hydrothermal fluids with temperatures up to approximately 400 °C escape into the cold (2.8°C), surrounding deep sea water and build up huge rock chimneys. Although these hot fluids are sterile, the surrounding porous smoker rock material appears to contain very steep temperature gradients, which provide zones of suitable growth temperatures for HT. Some smoker rocks are teeming with HT (for example, 10^8 cells of *Methanopyrus* per gram of rock inside a Mid Atlantic Snake Pit hot vent chimney). Deep sea vents are located along submarine tectonic fracture zones, (for example, the "TAG" and "Snake Pit" sites situated at the Mid Atlantic Ridge in a depth of approximately 4,000 m. Another type of submarine high temperature environment is provided by active sea mounts. Close to Tahiti, there is a huge abyssal volcano, Macdonald Seamount (28°58.7', 140°15.5' W), the summit of which is situated approximately 40 m below the sea surface. Samples taken during an active phase from the submarine eruption plume and rocks from the active crater contained high concentrations of viable HT (Huber et al., 1990).

7.3 Sampling and cultivation

Samples of HT–containing hot waters, soils, rock and sediments may serve as primary material to set-up enrichment cultures in the laboratory. Special care has to be taken to avoid contamination of the sample by oxygen. High temperatures are toxic to anaerobic HT. In contrast, at low temperatures (for example, 4 °C), in the presence of oxygen, anaerobic HT may survive for years. Transportation and shipping can be performed at ambient temperature. In the lab, anaerobic samples

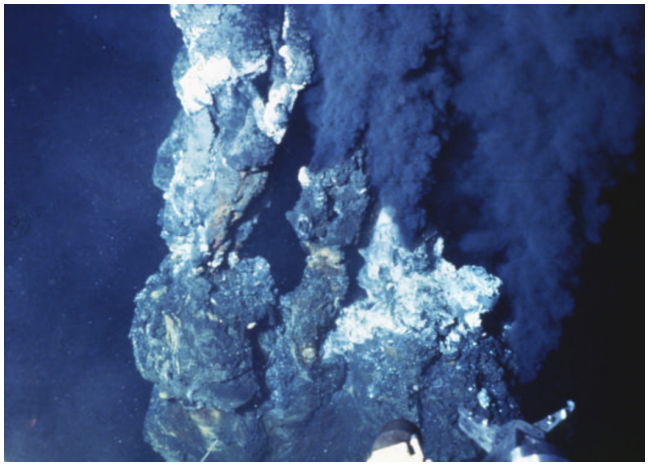


Fig. 7.2. Abyssal hot "black smoker" chimneys at the East Pacific Rise, 21°N. Depth : 2500 m, maximal fluid temperature: 365°C.

(in tightly stoppered 100 mL storage bottles and stored at 4 °C) can be used for successful enrichment cultures for at least 10 years.

Enrichment cultures can be obtained by simulating the varying geochemical and geophysical composition of environments. Various plausible electron donors and acceptors may be used under anaerobic, microaerophilic, or aerobic culture conditions. Depending on the (unknown) initial cell concentration and the doubling time of the organism, positive enrichment cultures of HT can be identified by microscopy within 17 days. For a deeper understanding of the organisms, the study of .pure cultures is required. Due to the high incubation temperatures, the traditional manner of cloning by plating does not perform well with HT (even by using heat- stable polymers like Gelrite). Therefore, we developed a new procedure for cloning single cells under the laser microscope, by employing optical tweezers (Ashkin and Dziedzic, 1987; Huber et al., 1995). Large cell masses are required for biochemical and biophysical investigations. For mass culturing of HT, a new type of high temperature fermentor was invented in collaboration with an engineering company (Fig. 3). Its steel casing is enamel-protected to resist the highly corrosive culture conditions. Sharp-edged parts like stirrers, gassing and sampling pipes and condensors are made of titanium. The

Fig. 7.3. Fermentation plant, University of Regensburg. Partial view,showing two 300 l fermentors and one 130 l fermentor for cultivation of HT.

cell yield of a 300 L fermentation may vary from approximately 3 g to 2 kg (wet weight), depending on the HT isolate.

7.4 Phylogenetic implications

Hot volcanic environments are among the oldest biotopes on Earth. What is known about the phylogenetic relationships among the organisms living there? Based on the pioneering work of Carl Woese, the small subunit ribosomal RNA (ss r RNA) is widely used in phylogenetic studies (Woese and Fox, 1977; Woese et al., 1990). It consists of approximately 1,500 nucleotides and is homologous to its eukaryotic counterpart. On the basis of sequence comparisons, a phylogenetic tree now is available (Fig 4). It shows a tripartite division of the living world into the bacterial, archaeal, and eucaryal domains. Within this tree, deep branches are evidence for early separation. The separation of the Bacteria from the stem shared by Archaea and Euckarya represents the deepest and earliest branching point. In contrast to the eucaryal domain, the bacterial and archaeal domains within the phylogenetic tree exhibit some extremely short and deep branches (short phylogenetic branches indicate a rather slow rate of evolution). Surprisingly, those are covered exclusively by hyperthermophiles, which, therefore form a cluster

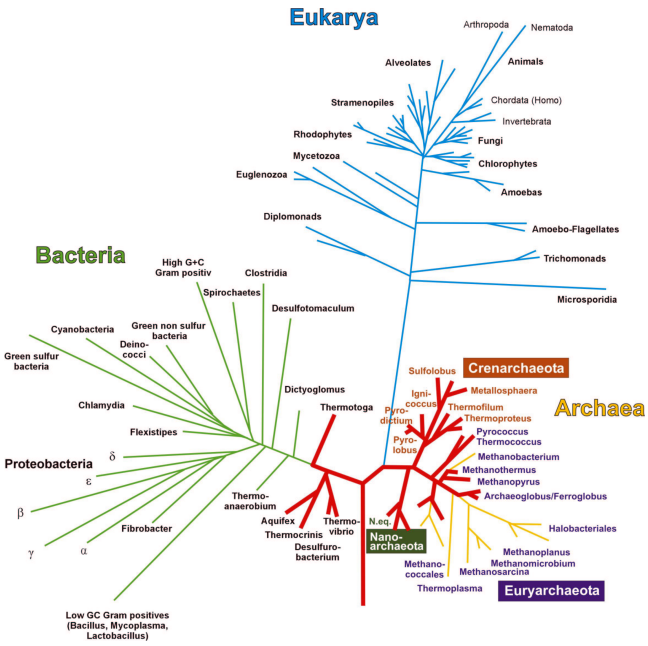


Fig. 7.4. Small subunit ribosomal RNA based Phylogenetic Tree. The bulky red lineages are representing HT.

around the phylogenetic root (Fig. 4, bulky lineages). The deepest and shortest phylogenetic branches are represented by the *Aquificales* and *Thermotogales* within the Bacteria and the *Nanoarchaeota*, *Pyrodictiaceae*, and *Methanopyraceae* within the Archaea. Long lineages represent mesophilic and moderately thermophilic Bacteria and Archaea (e.g. Gram-positives, Proteobacteria; *Halobacteriales*,; *Methanosarcinaceae*) indicating their ss rRNA had experienced a fast rate of evolution. Now, several total genome sequences are available. Phylogenetic trees based on genes involved in information management (for example, DNA replication, transcription and translation) parallel the ss-rRNA tree. Genes involved in metabolism, however, are prone to frequent lateral gene transfer, and a network rather than a tree mgiht reflect their phylogenetic relations (Doolittle, 1999).

To date, approximately 90 species of HT Archaea and Bacteria, which had been isolated from different terrestrial and marine thermal areas in the world are known. HT are very divergent, both in their phylogeny and

physiological properties and are grouped into 34 genera and 10 orders (Boone and Castenholz, 2001).

7.5 Physiologic properties

7.5.1 Energy sources

Most species of HT exhibit a chemolithoautotrophic mode of nutrition. Inorganic redox reactions serve as energy sources (chemolithotrophic), and CO_2 is the only carbon source required to build up organic cell material (autotrophic). Therefore, these organisms fix CO_2 by chemosynthesis and are designated chemolithoautotrophs. The energy-yielding reactions in chemolithoautotrophic HT are those involved in anaerobic and aerobic respiration (Fig. 5). Molecular hydrogen serves as an important electron donor. Other electron donors are sulfide, sulfur, and ferrous iron. As in mesophilic respiratory organisms, in some HT, oxygen may serve as an electron acceptor. In contrast, however, oxygen-respiring HT usually are microaerophilic and, therefore, grow only at reduced oxygen concentrations. Anaerobic respiration types are the nitrate-, sulfate-, sulfur- and carbon- dioxide types. While chemolithoautotrophic HT produce organic matter, there are some HT that depend on organic material as energy- and carbon sources. They are designated as heterotrophs. Several chemolithoautotrophic HT are opportunistic heterotrophs. These are able to use organic material alternatively to inorganic nutrients whenever they are available from the environment (e.g. by decaying cells). Heterotrophic HT gain energy either by aerobic or different types of anaerobic respiration, using organic material as electron donors, or by fermentation.

7.5.2 General physiologic properties

HT are adapted to environmental factors, including composition of minerals and gasses, pH, redox potential, salinity and temperature. Similar to mesophiles, they grow within a temperature range of approximately 25 to 30 ° C between the minimal and maximal growth temperature (Tab. 1). Fastest growth is obtained at their optimal growth temperature, which may be up to 106 ° C. Generally, HT do not propagate at 50 °C or below. Although unable to grow at low ambient temperatures, they are able to survive for years. Based on their simple growth requirements, HT could grow on any hot, wet place, even on other planets and

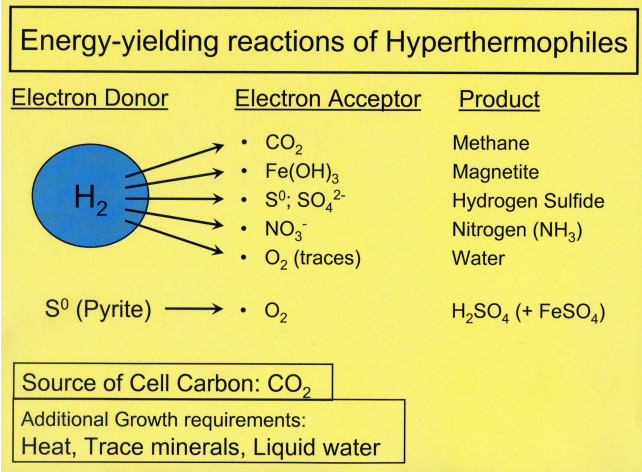


Fig. 7.5. Main energy-yielding reactions in chemolithoautotrophic HT (schematic drawing). Products are shown in red.

moons of our solar system, like Mars and Europa. To-day, the surface of Mars is too cold and contains no liquid water and, therefore is hostile to life as it is known on Earth. However, in a depth of a few kilometres below the permafrost layer, there may be hot liquid water and nutrients to support growth of HT. Life could have spread onto Mars via meteorites during the great bombardement, about 4 GA. At that time, the surface of Mars had been much hotter and contained liquid water, therefore being favourable to HT.

7.6 Examples of recent HT organisms

Within the bacterial domain, the deepest phylogenetic branch is represented by the HT *Aquifex* (Huber et al., 1992). Its type species, *Aquifex pyrophilus*, is a motile, rod-shaped chemolithoautotroph (Fig. 6). It is a facultative microaerophilic anaerobe. Under anaerobic conditions, *Aquifex pyrophilus* grows by nitrate reduction with H_2 and S^0 as electron donors. Alternatively, at very low oxygen concentrations (up to 0.5

From the walls of a black smoker at the Mid Atlantic Ridge, we had isolated the archaeon *Pyrolobus fumarii* (Blöchl et al., 1997) . Cells are lobed cocci, approximately 0.7 -2.5 m in diameter (Fig. 7). The species *Pyrolobus fumarii* is adapted optimally to temperatures of superheated

Table 7.1. *Growth conditions and morphology of hyperthermophiles*

Species	Min. Temp (°C)	Opt. Temp (°C)	Max. Temp (°C)	pH	Aeroboic (ae) vs. Anaer-obic (an)	Morphology
Sulfolobus acidocal-darius	60	75	85	1-5	ae	Lobed cocci
Metallosphaera sedula	50	75	80	1-4.5	ae	Cocci
Acidianus infernus	60	88	95	1.5-5	ae/an	Lobed cocci
Stygiolobus azoricus	57	80	89	1-5.5	an	Lobed cocci
Thermoproteus tenax	70	88	97	2.5-6	an	Regular rods
Pyrobaculum is-landicum	74	100	103	5-7	an	Regular rods
Pyrobaculum aerophilum	75	100	104	5.8-9	ae/an	Regular rods
Thermofilum pen-dens	70	88	95	4-6.5	an	Slender regular rods
Desulfurococcus mo-bilis	70	85	95	4.5-7	an	Cocci
Thermosphaera aggregans	67	85	90	5-7	an	Cocci in aggregates
Sulfophobococcus zilligii	70	85	95	6.5-8.5	an	Cocci
Staphylothermus marinus	65	92	98	4.5-8.5	an	Cocci in aggregates
Thermodiscus mar-itimus	75	88	98	5-7	an	Disks
Aeropyrum pernix	70	90	100	5-9	ae	Irregular cocci
Stetteria hy-drogenophila	70	95	102	4.5-7	an	Irregular disks
Ignicoccus islandicus	65	90	100	3.9-6.3	an	Irregular cocci
Pyrodictium occul-tum	82	105	110	5-7	an	Disks with cannulae
Hyperthermus butylicus	80	101	108	7	an	Lobed cocci
Pyrolobus fumarii	90	106	113	4.0-6.5	ae/an	Lobed cocci
Thermococcus celer	75	87	93	4-7	an	Cocci
Pyrococcus furiosus	70	100	105	5-9	an	Cocci
Archaeoglobus fulgidus	60	83	95	5.5-7.5	an	Irregular cocci
Ferroglobus placidus	65	85	95	6-8.5	an	Irregular cocci
Methanothermus so-ciabilis	65	88	97	5.5-7.5	an	Rods in clusters
Methanopyrus kand-leri	84	98	110	5.5-7	an	Rods in chains
Methanococcus igneus	45	88	91	5-7.5	an	Irregular cocci
Thermotoga mar-itima	55	80	90	5.5-9	an	Rods with sheath
Aquifex pyrophilus	67	85	95	5.4-7.5	ae	Rods

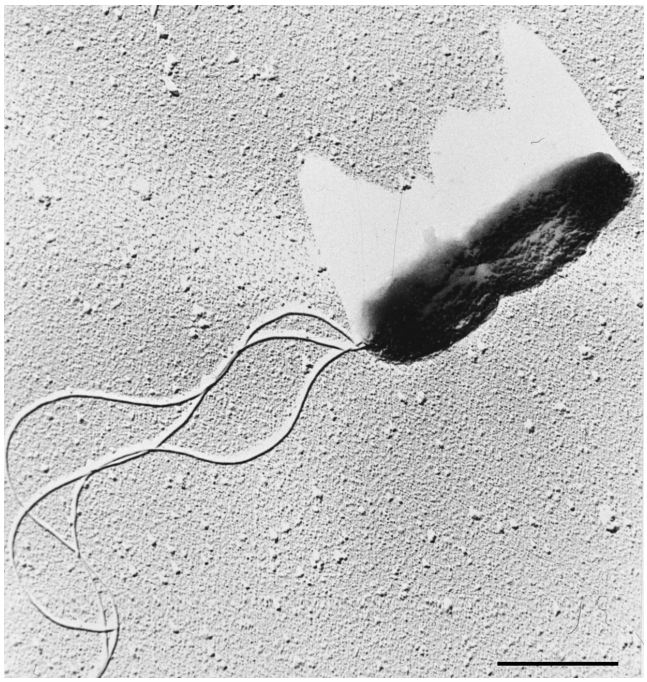


Fig. 7.6. *Aquifex pyrophilus*, dividing cell with a tuft of flagella. Pt - shadowing. Transmission electron micrograph. Scale bar, 1 μm.

water, exhibiting an optimal growth temperature of 106 ° C (Fig. 8) and an upper temperature border of growth at 113 ° C. It is so adapted to high temperatures that it is unable to grow below 90 °C. Cultures of *Pyrolobus fumarii*, similar to *Pyrodictium occultum*, are able to survive autoclaving for one hour at 121 ° C. A very closely related isolate (strain 121) exhibits the same optimal growth temperature (106 °C), but had been reported to grow slowly even at 121 °C (Kashefi and Lovley, 2003). However, this result has not been confirmed.

From a submarine hydrothermal system situated at the Kolbeinsey Ridge, North of Iceland, we were able to obtain our ultimate hyperthermophilic coccoid isolate, *Nanoarchaeum equitans*, which represents a novel kingdom of Archaea (Huber et al., 2002). With a cell diameter of only 400 nm, it is the smallest living organisms known. Cells grow attached to the surface of a specific crenarchaeal host, a new member of the genus *Ignicoccus*. (Fig. 9). Owing to their unusual 16S rRNA se-

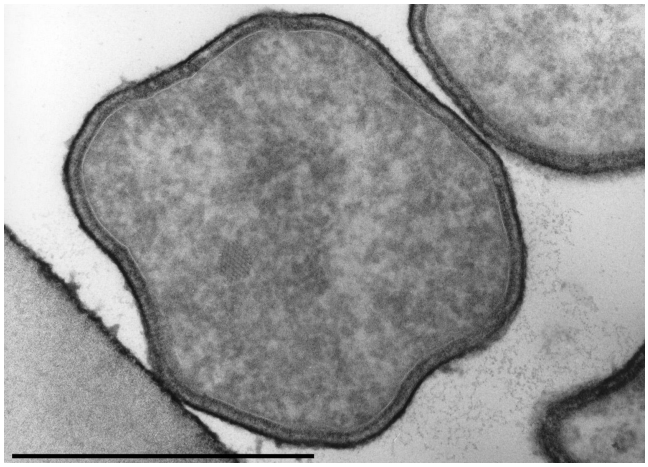


Fig. 7.7. *Pyrolobus fumarii*, lobed coccoid cell. Ultrathin section. Transmission electron micrograph. Scale bar, 0.5 μm.

quence, *Nanoarchaeum eqitans* cultures remained undetectable by commonly used "universal" primers in ecological studies involving the polymerase chain reaction. However, two different 16S rRNA genes could be detected in a coculture-derived DNA by Southern blot hybridization, by taking advantage of the generally high sequence homology of all 16S rRNA genes (about 75

References

[490] Ashkin, A., and Dziedzic, J. M. 1987. Optical trapping and manipulation of viruses and bacteria. *Science* 235, 1517-1520.

[490] Boone, D. R., and Castenholz, R. W. 2001. The Archaea and the deeply branching and phototrophic Bacteria. In Bergey's Manual of Systematic Bacteriology, Second Edition,ed Garrity, G. M.; Vol 1, pp. 169-387. Springer, New York Berlin Heidelberg.

[490] Blchl, E., R. Rachel, S. Burggraf, D. Hafenbradl, H.W. Jannasch, and K.O. Stetter. (1997). Pyrolobus fumarii, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113°C. *Extremophiles* 1, 14-21.

[490] Brasier, M. D., Green, O. R., Jephcoat, A. P., Kleppe, A. K., Van Kranendonk, M. J., Lindsay, J. F., Steele, A., Grassineau, N. V. 2002. Questioning the evidence for Earths oldest fossils. *Nature 416*, 76-81.

[490] Brock,T.D. 1978. Thermophilic Microorganisms and Life at High Temperatures.

[490] Castenholz, R.W. 1979. Evolution and ecology of thermophilic microor-

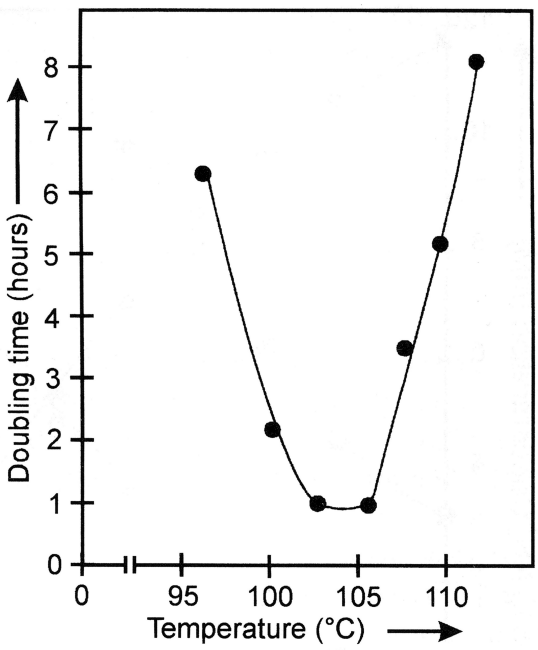


Fig. 7.8. *Pyrolobus fumarii*: Temperature-dependence of its doubling time. Optimal growth (approximately 50 minute doubling time) occurs between 103 and 106 °C.

ganisms. In:Strategies of Microbial Life in Extreme Environments, ed. M. Shilo, pp.373-392. Weinheim: Verlag Chemie.

[490] Cowan, D. A. 2004. The upper limit of life-how far can we go. *Trends Microbiol.* 12, 58-60.

[490] Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. *Science* 284, 2124-2129.

[490] Drobner, E., Huber, H., Wchtershuser, G., Rose, D., and Stetter, K. O. 1990. Pyrite formation linked with hydrogen evolution under anaerobic conditions. *Nature* 346, 742-744.

[490] Hohn, M.J., Hedlund, B.P., and Huber, H. 2002. Detection of 16S rDNA sequences representing the novel phylum Nanoarchaeota: indication for a broad distribution in high temperature. *Syst. Appl. Microbiol.*, 25: 551-554.

[490] Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., V.C. Wimmer, V. C., and K. O. Stetter. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417: 63-67.

[490] Huber, R, Stoffers, P., Cheminee, J. L., Richnow, H. H., and Stetter, K.

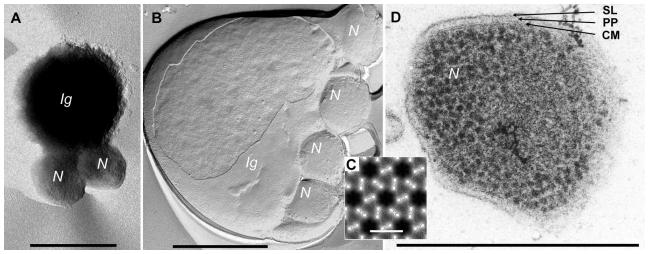


Fig. 7.9. *Nanoarchaeum equitans-Ignicoccus* sp.:Transmission electron micrographs. (A) Two cells of *N. equitans*, attached on the surface of the (central) *Ignicoccus* cell. Platinum shadowed. Scale bar, 1 m. (B) Freeze-etched cell of *Ignicoccus* (Ig) and attached cells of *N. equitans* (N) on the surface. Scale bar, 1 m. (C) Surface relief reconstruction of *N. equitans*. Dark: cavity; Bright: elevation. Scale bar, 15 nm. (D) Ultra thin section of a *N. equitans* cell. Single cell. CM: cytoplasmic membrane; PP periplasm; SL: S-layer. Scale bar: 0.5 m.

O. 1990. Hyperthermophilic archaebacteria within the crater and open-sea plume of erupting Macdonald Seamount. *Nature* 345, 179-181.

[490] Huber, R. Burggraf, S.,Mayer, T., Barns, S. M., Rossnagel, P., and Stetter, K. O. 1995. Isolation of a hyperthermophilic archaeum predicted by in situ RNA analysis. *Nature* 376, 57-58.

[490] Kandler, O. 1994. The early diversification of life. In Early Life on Earth, ed. S.Bengtson, pp. 152-160. Nobel Symposium No. 84. Columbia University Press, New York.

[490] Kashefi, K., and Lovley, D. R. 2003. Extending the upper temperature limit of life. *Science* 301, 934.

[490] Randau, L., Mnch, R., Hohn, M. J. Jahn, D., and Sll, D. 2005. Nanoarchaeum equitans creates functional t-RNAs from separate genes for their 5- and 3- halves. *Nature* 433, 537-541.

[490] Schopf, J. W., Packer, B. M. 1987. Early Archean (3.3 billion to 3.5 billion-year-old) microfossils from Warrawoona Group, Australia. *Science* 237, 70-73.

[490] Stetter K. O., Thomm, M., Winter, J., Wildgruber, G, Huber, H., Zillig, W., Janecovic, D., Knig, H., Palm, P., and Wunderl, S. 1981. Methanothermus fervidus, sp. nov., a novel extremely thermophilic methanogen isolated from an Icelandic hot spring. *Zbl. Bakt. Hyg., I. Abt. Orig. C2*, 166-178.

[490] Stetter, K. O. (1982). Ultrathin mycelia-forming organisms from submarine volcanic areas having an optimum growth temperature of 105°C. *Nature* 300: 258-260.

[490] Stetter, K. O. 1992. Life at the upper temperature border. In Frontiers of Life, ed. J. Tran Thanh Van, K. Tran Thanh Van, J.C. Mounolou, J. Schneider, and C. McKay, pp. 195-219. Gif-sur-Yvette: Editions Frontieres.

[490] Stetter,K. O., Huber, R., Blchl, E., Kurr, M., Eden, R. D., Fielder, M., Cash, H., and Vance, I. 1993. Hyperthermophilic archaea are thriving in

deep North Sea and Alaskan oil reservoirs. *Nature* 365 (1993) 743-745.

[490] van Zullen, M. A., Lepland, A., Arrhenius, G. 2002. Reassessing the evidence for the earliest traces of life. *Nature* 418, 627-630.

[490] Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M. D., Barnstead, M., Beeson, K. Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G. G., Simon, M., Sll, D., Stetter, K. O., Short, J.M., and Noordewier, M. (2003). The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci., U S A* 100:12984-12988.

[490] Woese C. R., and Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U S A* 74, 5088-5090.

[490] Woese C. R., Kandler,O., and M. L. Wheelis, M. L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. *Proc. Natl. Acad. Sci., U S A*, 87: 45764579.

[490] Zillig, W., Stetter, K. O., Schfer, W., Janekovic, D., Wunderl, S., Holz, I., and Palm, P. 1981. Thermoproteales: a novel type of extremely thermoacidophilic anaerobic archaebacteria isolated from Icelandic solfataras. *Zbl. Bakt. Hyg., I. Abt. Orig. C2*, 205-227.

8

Phylogenomics: how far back in the past can we go?

Henner Brinkmann, Denis Baurain & Hervé Philippe
Université de Montréal

8.1 Introduction

The remnants of ancient life on Earth are extremely rare and difficult to interpret, since they have been modified by billions of years of subsequent geological history. Therefore, any understanding of the evolution of Earth's biosphere will heavily rely on the study of extant organisms for which complete genome sequences are already available or will eventually be established. A central point is the phylogenetic inference of the Tree of Life from such genomic data. Serving as a unifying framework, this tree will allow to collect and structure all disparate and often incomplete pieces of information gathered by various disciplines. In the first half of this chapter, we will briefly explain the principles of phylogenetic inference and the major artefacts affecting phylogenetic reconstruction. Then we will introduce phylogenomics, starting with a theoretical presentation before illustrating the explained concepts through a case study centered on animal evolution. In the second half, we will review our present understanding of eukaryotic evolution and show how recent knowledge suggests that secondary simplification is an important mode of evolution that has been too long overlooked. We will also summarize the current views about the root and the shape of the Tree of Life. Finally, we will attempt to debunk two common hypotheses about the early evolution of Life, i.e. a cell fusion at the origin of eukaryotes and a hyperthermophilic origin of Life.

8.2 The principles of phylogenetic inference

8.2.1 The homology criterion

The reconstruction of the genealogical relationships of a set of species first requires the definition of a series of characters that are comparable across all studied organisms. This problem has been solved in 1843 by the palaeontologist Richard Owen based on the principle of connectivity previously established by the anatomist Etienne Geoffroy Saint-Hilaire: two organs, whatever their forms and functions, are homologous if and only if they are connected to the same structures. Owen simultaneously introduced an operational criterion and a theoretical distinction between mere analogy (similar function) and true homology (comparable structure across organisms, which will be explained a few years later by Charles Darwin as having been inherited from a common ancestor). Unfortunately, this important distinction is often ignored in modern Biology, a common function being generally deduced from a crude homology assessment. At the molecular level (genes and proteins), the establishment of homology that corresponds to the alignment of nucleotide or amino acid sequences relies on the very same principle: two positions are homologous if they are connected to similar neighbouring nucleotides or amino acids. The alignment of molecular sequences is a complex step for which numerous methods have been developed (for a review see (Wallace et al., 2005). In particular, the homology within poorly conserved regions can be extremely difficult to assess, even if the surrounding sequences are indisputably homologous. This explains why one can hardly decide whether the positions inside highly variable regions are indeed homologous, especially in the presence of sequence length heterogeneity. Because even the best tree reconstruction methods are sensitive to the "garbage in garbage out" principle, noisy areas can severely disturb the phylogenetic analysis when they are included in a dataset (Lake, 1991). Consequently, a conservative approach is to discard the regions of homologous sequences that are too divergent to be aligned with confidence (e.g. Castresana, 2000).

8.2.2 Cladistics and Maximum Parsimony

To explain how a phylogenetic tree can be inferred from a set of homologous characters, we will begin with the principles of cladistic analysis (Hennig, 1966) because they are easily understandable. The elementary task of phylogenetic inference is to group the species according to their

relatedness. This is achieved by identifying positions that have the same character state (e.g. the same amino acid) in a given group of species, but a different state in the remaining species. To illustrate our point, we will refer to Fig. 1, where the oldest events are on the left of the trees and the most recent on the right. Trees that depict the relative order of the speciation events are said to be rooted. In the following, we will assume that the character state of the root (i.e. the ancestral state on the far-left) is known. On the tree of Fig. 1a, a change of state is indicated on the internal branch of the group composed of taxa sp3 and sp4. This substitution actually occurred in the common ancestor shared by these species. Such a single change that characterises a monophyletic group is termed a synapomorphy (i.e. shared derived character state). All the trees that group the species sharing the same character state will require a single substitution (Fig. 1b) and will be preferred to the other trees that require two or more substitutions. The cladistic analysis is also known as Maximum Parsimony (MP) because it selects the tree(s) having the lowest number of substitutions (or steps). Problems crop up when the same character state is independently acquired by two unrelated lineages through convergence (Fig. 1c). In such case, the true phylogeny will require two substitutions, while erroneous trees grouping these unrelated lineages will require a single substitution (Fig. 1d) and would be preferred. However, since not all characters do support identical groups of species, cladistics tries to maximise the number of inferred synapomorphies by selecting the tree requiring the minimum number of steps, thereby following the principle of Occam's razor. The underlying hypothesis is that this approach should allow discovering all synapomorphies and all convergences. In practice, our assumption that the character state of the root is known is invalid. This issue is circumvented by applying the same parsimony criterion on unrooted trees. Then, the best tree is rooted using an outgroup, i.e. an externally recognized group of sequences that are a priori assumed to be located outside the group of interest. The rooting defines the polarisation of the evolutionary history, which allows a chronological interpretation of the emergence order of the species and the character states in the ingroup.

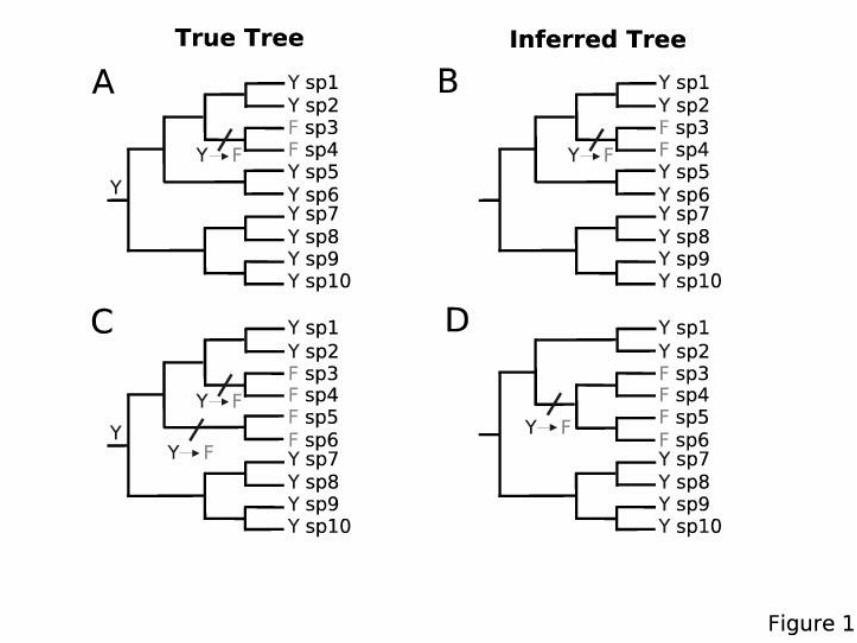


Fig. 8.1. **Principle of phylogenetic and non-phylogenetic signals.** The true evolutionary histories of an amino acid position are depicted in A and C, while the inferred histories are depicted in B and D. In A, a single substitution occurred and constitutes phylogenetic signal, i.e. a synapomorphy that allows to recover the correct grouping (B). In C, two substitutions occurred and the same amino acid was independently acquired by convergence. This constitutes a non-phylogenetic signal that prevents to recover the correct tree (D).

8.3 Artefacts affecting phylogenetic reconstruction

8.3.1 Multiple substitutions and character saturation

In theory, the faster a character evolves the more phylogenetic signal (i.e. synapomorphies) it will contain. However, as soon as a single character undergoes multiple substitutions, it becomes difficult to extract its phylogenetic signal, because closely related species will not necessarily share the same character state (Fig. 1c). Actually, for a fast evolving position, the number of substitutions required to support two or more conflicting topologies can be so similar that the reconstruction method is unable to choose among the different alternatives. In practice, it is virtually impossible to extract phylogenetic signal from very fast evolving

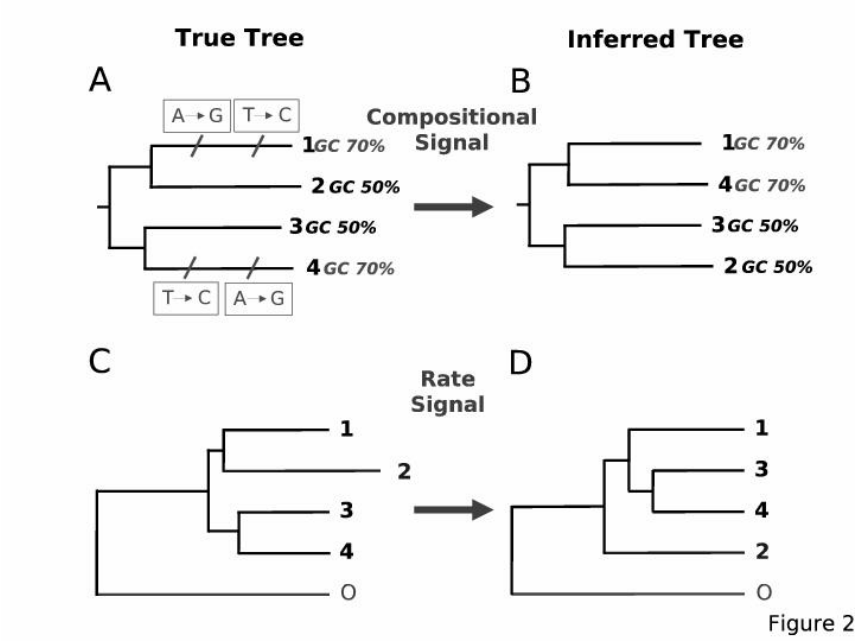


Fig. 8.2. Figure 1 caption goes here...

characters, which are said to be mutationally saturated. When the level of saturation is high, the inferred trees show a star-like topology, being totally unresolved because the phylogenetic signal that would allow defining the different groups could not be extracted.

8.3.2 Systematic biases and non-phylogenetic signals

Unfortunately, whenever many multiple substitutions accumulate in the data, the situation can be worse than a simple scenario based on the erosion of the phylogenetic signal. Indeed, in the presence of any kind of systematic bias, multiple substitutions will generate a non-phylogenetic signal potentially resulting in reconstruction artefacts, as we will illustrate in the case of the compositional bias of nucleotide sequences (Lockhart et al., 1992). Let us assume that the content of guanine and cytosine (G+C) in a group equals to 50

The variation of evolutionary rates across lineages is another important cause of non-phylogenetic signal due to convergent evolution. It

results in the long branch attraction (LBA) artefact in which the two longest branches of a tree tend to be grouped even when not closely related (Felsenstein, 1978). A particular albeit frequent case may arise whenever a distant outgroup is included to root the tree, i.e. to allow the polarisation of the characters under study. Indeed, if a species of the studied group evolves significantly faster than the others (Fig. 2c), it will artefactually emerge at the base of the group (Fig. 2d), because its long branch is attracted by the long branch of the distant outgroup. This phenomenon is wrongly suggesting an early divergence relatively to the remaining species.

Early on, rate and compositional biases have been identified as major sources of non-phylogenetic signals. An additional confounding bias that have recently attracted more attention is the heterotachous signal that is due to a variation of the evolutionary rate of a given position throughout time (Kolaczkowski and Thornton, 2004; Lockhart et al., 1996; Philippe and Germot, 2000) and remains to be thoroughly explored.

8.3.3 Gene duplication and Horizontal Gene Transfer

For the sake of simplicity, we have so far assumed that it is sufficient to analyse homologous positions for inferring phylogeny. Strictly speaking, this is correct as long as we are interested in gene phylogenies, but as soon as we want to deduce the species phylogenies, we must ensure that the genes under study are orthologous (Fitch, 1970). By definition, orthologous genes originate in speciation events, whereas paralogous genes originate in gene duplication events (Fig. 3a). Phylogenetic analyses of species must be imperatively based on orthologous sequences, because trees based on cryptic paralogs can be extremely misleading (Fig. 3b). Another problem is horizontal gene transfer (HGT), where a gene is transferred from a donor species to a possibly unrelated receiver species (Fig. 3c). This alien copy, often named xenolog, causes similar problems as paralogs. i.e. unrelated species are incorrectly grouped in phylogenetic trees (Fig. 3d).

Since gene (and even genome) duplications are frequent in eukaryotes, while HGTs are frequent in prokaryotes, the number of genes that are perfectly orthologous when considering all comparable extant organisms is probably close to zero. For instance, only 14 genes were found to exist in one and only one copy in 10 completely sequenced eukaryotic genomes (Philip et al., 2005). Fortunately, if a given gene has undergone a single recent duplication event, the latter will be easy to detect and the gene

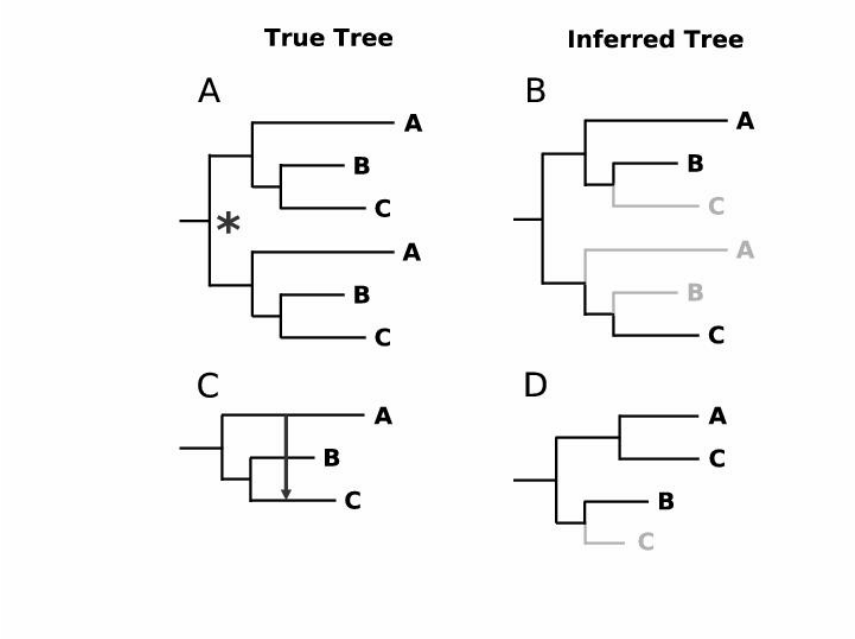


Fig. 8.3. **Orthology, paralogy and xenology and their consequence on phylogenetic inference.** A. At some point of its history, a gene is duplicated and gives rise to two paralogous copies. The duplication event is shown by a star. In the course of the subsequent speciation events, each copy evolves independently to generate a set of three orthologous genes. When a tree including both paralogs from each species (A, B and C) is inferred, the true species phylogeny is recovered for each paralog. B. In a tree inferred from different paralogs instead of orthologs, a wrong species phylogeny is recovered. The suboptimal gene sampling can be due to technical reasons (e.g. orthologous gene not yet sequenced) or to biological reasons (e.g. both copies have been differentially lost in the three lineages). C. True tree - During the evolution of lineages A, B and C, a gene is horizontally transferred from lineage A to lineage C. D. False tree - Because of the close similarity between the xenolog in C and the ortholog in A, a wrong organismal phylogeny grouping species A and C is recovered. As for paralogy above, the ortholog in C may be lacking for technical or biological reasons (e.g. the acquired xenolog has replaced the orthologous gene).

will still be usable (Philippe et al., 2005b) because it does not interfere with the species phylogeny. Although the same reasoning should hold for HGTs, it appears that the impact of transfer events on the phylogenetic structure is much more destructive (Philippe and Douady, 2003), even

when only a few HGTs have occurred. While some researchers suggested that because of transfer events species phylogenies can not exist (nor the representation as trees) (Doolittle, 1999), the current consensus is that a core of genes (a few hundreds within Bacteria) have experienced so few HGTs that they can be used to infer species phylogenies (for review see (Brown, 2003; Ochman et al., 2005; Philippe and Douady, 2003)).

8.4 Strengths and limitations of phylogenomics

8.4.1 Stochastic error and the need for more data

Phylogenies based on a small number of characters (both morphological and molecular) are sensitive to the stochastic (random or sampling) error. Consequently, the inferred trees are usually poorly resolved and often yield to conflicting results, though differences are seldom statistically significant. By considering more characters and/or characters with more substitutions the phylogenetic signal can be increased, since the stochastic error is due to the scarcity of substitutions that occurred along internal branches (the aforementioned synapomorphies). Starting from one or a few characters, as in the first classifications elaborated in the Middle Age, to tens or a few hundred characters, as in most recent studies based on morphological characters, the rule "the more characters the better" has always been applied. The advent of large scale sequencing allowed to gain about three orders of magnitude, resulting in an enormous improvement of the resolving power of phylogenetic inference. However, the switch from hundreds or a few thousand positions in single gene phylogenies (e.g. rRNA tree (Woese, 1987)) to hundreds of thousand positions in phylogenomic studies based on complete genomes is quite recent. Phylogenies with a high statistical support for most nodes have been recently obtained for various groups, such as mammals (Madsen et al., 2001; Murphy et al., 2001), angiosperms (Qiu et al., 1999), and eukaryotes (Rodrguez-Ezpeleta et al., 2005).

8.4.2 Systematic error and the need for better reconstruction methods

While the use of many characters drastically reduces the stochastic error, it does not necessarily constitute a solution to the problem of tree reconstruction artefacts (Philippe et al., 2005a). Indeed, the addition of more data increases both the phylogenetic and non-phylogenetic signals.

Therefore, in the presence of a systematic bias, the latter will eventually become predominant, especially when the phylogenetic signal is rather weak (Jeffroy et al., 2006). For example, a hyperthermophilic life style gives rise to a systematically biased composition of all proteins (Kreil and Ouzounis, 2001), thus potentially leading to an artefactual albeit highly supported tree. Obviously, there is an urgent need for designing reconstruction methods that are less sensitive to the systematic error induced by the use of large datasets. Hence, the research in this area is currently very active (Felsenstein, 2004). Although the previously presented MP method is intuitive, its improvement is difficult since it does not make any explicit assumptions about the underlying evolutionary process (see Steel and Penny, 2000). For example, the probability of a substitution event is implicitly assumed to be identical across all branches of the tree, whereas this assumption is clearly violated in cases where branch lengths are unequal, leading to LBA artefact. In contrast, probabilistic methods such as Maximum Likelihood (ML) and Bayesian Inference (BI) have been designed to take into account branch lengths and are therefore much less sensitive to the LBA artefact. More generally, in a probabilistic framework, the likelihood of a tree is computed using a model of sequence evolution able to handle numerous aspects of the underlying process of sequence evolution. This complex model enhances the extraction of the phylogenetic signal and greatly reduces the impact of non-phylogenetic signal because the probability that multiple substitutions have occurred is explicitly considered.

8.4.3 State of the art in evolutionary models

Currently, implemented models of sequence evolution have the following properties: (1) the various probabilities to substitute one character state by another are unequal (e.g. transitions are more frequent than transversions); (2) the stationary probabilities of the various character states can be unequal (e.g. A more frequent than T) and are generally estimated from observed frequencies; (3) evolutionary rates can be heterogeneous across sites (i.e. positions), this heterogeneity being usually modelled by a discrete gamma distribution (Yang, 1993). For nucleotides, the probabilities of the different substitutions are directly inferred from the data (GTR model (Lanave et al., 1984)), whereas for amino acids, values previously computed from large alignments of closely related sequences are preferred (e.g. WAG substitution matrix (Whelan and Goldman, 2001)). Taking into account these evolutionary hetero-

geneities efficiently improves the extraction of the phylogenetic signal, as exemplified by the fact that the introduction of the Gamma distribution is often associated with changes of the tree topology (Yang, 1996). Other evolutionary features that have been modelled include: heterogeneity of the G+C content through a non-stationary model (Foster, 2004; Galtier and Gouy, 1995; Yang and Roberts, 1995), heterotachy through a covarion model (Galtier, 2001; Huelsenbeck, 2002), and heterogeneous probabilities of substitution types across sites using mixture models (Lartillot and Philippe, 2004; Pagel and Meade, 2004). Although newer models are complex and probabilistic methods have been shown to be robust to model violations, reconstruction artefacts can still interfere with sequence-based phylogenomic analyses, as we will illustrate later in a case study of animal evolution.

8.4.4 Current limits of phylogenomic performance

Profiting from the synergistic effects of the massive increase of sequence data and the improvement of tree reconstruction methods, the resolution of the Tree of Life is rapidly progressing (Delsuc et al., 2005). However, a few important questions are expected to remain difficult to answer in the near future because the phylogenetic signal is either scarce or dominated by strong non-phylogenetic signals. Beside the aforementioned mutational saturation inherent to ancient events, the lack of phylogenetic signal can be due to the existence of short internal branches associated with numerous speciation events concentrated within a short time span (i.e. adaptive radiations). Furthermore, the resolution of ancient events is complicated by a dramatic reduction of the available data, because of the concomitant decrease in the number of orthologous genes (due to duplications and HGTs) and in the number of unambiguously aligned positions (due to the considerable sequence divergence).

8.4.5 Corroboration from non-sequence-based phylogenomic methods

Formally defined as the inference of phylogenies from complete genomes, phylogenomics is not limited to primary sequence data. Instead, its principles can also be applied to virtually any heritable genomic feature such as gene content, gene order or intron positions (see (Philippe et al., 2005a) for a recent review). Since the usual methods of tree reconstruction are used, strengths and limitations of phylogenies based on

these other types of characters are very similar to those based on primary sequences. However, because these various characters are largely independent, they provide a major source of corroboration, which is of primary importance to validate historical studies (Miyamoto and Fitch, 1995). Indeed, the fact that phylogenies inferred from different character types converge to the same tree topology is strongly suggesting that the correct organismal tree has been reconstructed. Although such integrated approaches have so far rarely been applied, the first studies indicate a good congruence for Bacteria and Metazoa (for recent reviews see Delsuc et al., 2005; Philippe et al., 2005b) . Nevertheless, if the same systematic bias (e.g. rate acceleration) simultaneously affects all genomic features, the same reconstruction artefacts are likely to occur.

8.4.6 Case study: Resolution of the metazoan evolution

Because of the general interest in animal evolution, we will present the resolution of this long-lasting problem as a case study to illustrate the theoretical concepts explained so far. Before 1997, metazoan taxonomy was essentially based on the presence or absence of true internal body cavities (coeloem) (Adoutte et al., 2000), with arthropods and vertebrates grouped among others into Coelomata to the exclusion of nematodes (Pseudocoelomata). Suspecting that the early emergence of the generally fast-evolving nematodes was the result of an LBA artefact (Philippe et al., 1994), Alguinaldo and coworkers (1997) sequenced the SSU rRNA gene from dozens of nematodes, until they identified one slowly-evolving species, Trichinella. By using Trichinella they were able to overcome the LBA artefact, revolutionising the picture of animal evolution by overruling the classical dichotomy between Coelomata and other animals. Instead, they found a new metazoan group named Ecdysozoa (Aguinaldo et al., 1997). Including among others arthropods, nematodes, tardigrades and onychophorans, these animals are characterised by a moult induced by a class of hormones known as the ecdysteroids. Nevertheless, several recent phylogenomic studies reject the Ecdysozoa hypothesis and find a significant support for the classical Coelomata hypothesis (Blair et al., 2002; Dopazo et al., 2004; Philip et al., 2005; Wolf et al., 2004), suggesting that the monophyly of Ecdysozoa is representing a rRNA-specific anomaly. These analyses use a large number of characters but only a very limited number of species, i.e. the few completely sequenced model organisms. Furthermore, only species

distantly related to animals (fungi, plants or apicomplexans) are used as outgroups, thus increasing the probability of an LBA artefact.

To address this question, we assembled our own data set both species- and gene-rich (49 species, 146 proteins and 35,371 amino acid positions) and our results strongly argue in favour of the group Ecdysozoa (Philippe et al., 2005b). In agreement with previous studies, when a poor species sampling is used, a strong support for Coelomata is recovered (Fig. 4a). In contrast, adding two choanoflagellates and a cnidarian (Hydra) has a dramatic effect since Ecdysozoa are now highly supported (Fig. 4b). This topological change is not surprising because the longest branch of the tree (leading to the distant outgroup) has been broken (Delsuc et al., 2005), what is known to reduce the impact of the LBA artefact (Hendy and Penny, 1989). However, it is worth noting that this result is only achieved through the use of ML. Indeed, even in the presence of a close outgroup, the use of the LBA-sensitive MP still results in a similar support for Ecdysozoa and the artefactual Coelomata, (data not shown).

Since reconstruction artefacts are primarily caused by multiple substitutions, eliminating the fastest-evolving characters should improve the quality of the phylogenetic inference, thereby reducing possible LBA artefacts. Actually, the removal of fast-evolving characters, performed using the SF method (Brinkmann and Philippe, 1999), produces exactly the same result as the addition of a close outgroup (Delsuc et al., 2005), i.e. a topological shift from Coelomata to Ecdysozoa (Fig. 5). Moreover, a similar result was independently obtained with a larger number of genes and two different data removal approaches (Dopazo and Dopazo, 2005; Philippe et al., 2005b). Therefore, all these analyses demonstrate that Ecdysozoa is a natural clade and that an LBA artefact is responsible for the high statistical support for Coelomata in species-poor phylogenomic studies.

While this case study shows that sophisticated reconstruction methods may still produce erroneous trees, there are major improvements due to the largely increased resolving power of phylogenomics. This is reflected in the fact that almost all other nodes in the metazoan phylogeny are well supported (Delsuc et al., 2006; Philippe et al., 2005b). In addition, reconstruction errors are drastically reduced as soon as a large number of species is considered, which is likely to become the rule in the next few years. Furthermore, starting from a large phylogenomic data set allows to remove highly saturated characters, which can be very useful

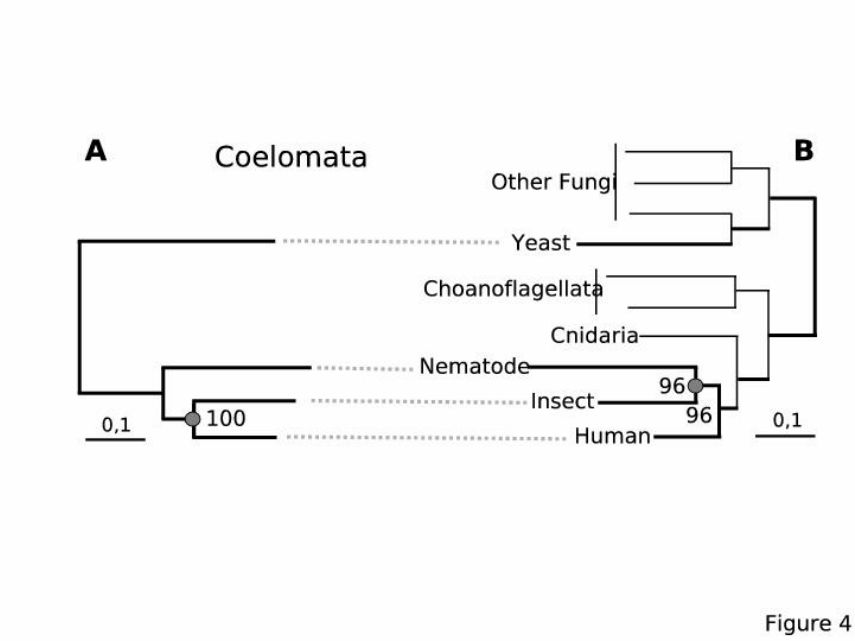


Fig. 8.4. Uejwkwkwk

in cases where a close outgroup is not available (Brinkmann et al., 2005; Burleigh and Mathews, 2004; Philippe et al., 2005b).

8.5 The importance of secondary simplification

8.5.1 The rise and fall of Archezoa

As for Metazoa, recent advances in phylogenetic analyses have deeply modified our view of the eukaryotic tree. Originally, both morphological (Cavalier-Smith, 1987) and molecular studies (mainly based on rRNA) (Sogin, 1991) suggested that several simple lineages (e.g. devoid of mitochondrion, Golgi apparatus or peroxysome) emerged in a stepwise fashion at the base of eukaryotes, followed by an unresolved multifurcation containing all complex, often multicellular, eukaryotes. The common interpretation of the classical rRNA tree is that these simple groups (e.g. microsporidia, diplomonads and trichomonads) are the direct descendants of genuinely primitive organisms and represent intermediate

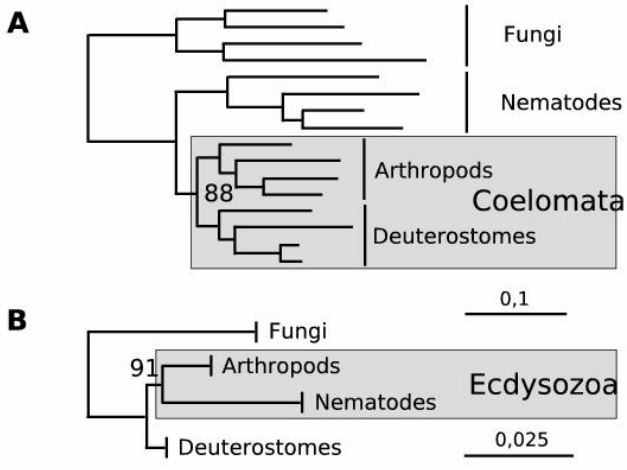


Figure 5

Fig. 8.5. **Avoiding the Long Branch Attraction artefact through the elimination of fast-evolving positions.** Phylogeny based on 146 genes inferred by ML method. In A, when all positions are considered, the fast evolving nematodes is artefactually attracted by the long branch of fungi. However, when only the slowest evolving positions are used, nematodes are correctly located as sister-group of arthropods (B). Redrawn from Delsuc et al. (2005).

stages in the progressive complexification of eukaryotic cells. Hence, they ware named Archezoa (Cavalier-Smith, 1987). This suggests that the endosymbiosis with an alpha-proteobacterium that gave rise to mitochondria (see below) had occurred relatively late in the course of eukaryotic evolution (Fig. 6A). Therefore, the study of eukaryotes that were supposed to have never possessed mitochondria was regarded to be of prime importance for the understanding of early eukaryotic evolution (Sogin, 1991). The observation that hundreds of genes had been transferred from the proto-mitochondrion to the nucleus (Lang et al., 1999) further underlines the dramatic modification of eukaryotic cells due to the mitochondrial endosymbiosis.

Unfortunately, a wealth of newly established data has been demon-

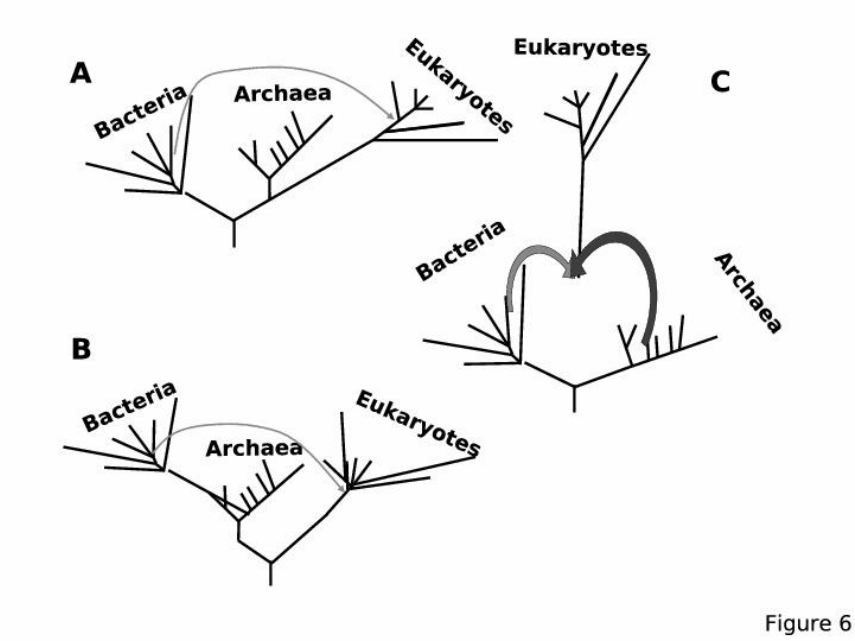


Fig. 8.6. **Most common views of the universal tree of life.** A Schematic representation of the Woese paradigm. The root is located in the bacterial branch and the mitochondrial endosymbiosis occurred relatively late during the evolution of eukaryotes. B Woese tree corrected for Long Branch Attraction artefacts. Archaea and Bacteria are sister-group, rendering prokaryotes monophyletic. The diversification of extant eukaryotes occurred relatively recently, after the mitochondrial endosymbiosis. C Genome fusion at the origin of eukaryotes.

strating more and more solidly that the Archezoa hypothesis is incorrect and that most likely none of the pre-mitochondriate eukaryotes had survived (Embley and Hirt, 1998; Philippe and Adoutte, 1998). First, the use of advanced reconstruction methods and/or of protein-encoding genes reveals that the early rRNA trees (Sogin, 1991) are severely biased by LBA artefacts due to non-phylogenetic rate signal (Edlind et al., 1996; Philippe, 2000; Silberman et al., 1999; Stiller and Hall, 1999). In fact, the lineages that emerge early in the rRNA tree are simply fast-evolving organisms that are erroneously attracted towards the base by the distant outgroup (Archaea). For example, microsporidia turn out to be highly derived fungi, while they were previously thought to be gen-

uinely "primitive" early eukaryotes (Keeling and Fast, 2002). Actually, the correct placement of microsporidia is difficult to recover (Brinkmann et al., 2005), because for most but not all of their proteins the non-phylogenetic signal (due to a high evolutionary rate) is stronger than the genuine phylogenetic signal.

8.5.2 Current view of mitochondrial origin and evolution

Since the new phylogenetic scheme places the former Archezoa back into the main eukaryotic radiation (known as the crown), it implies that the last common ancestor of extant eukaryotes (LCAEE) was much more complex than previously thought, i.e. the LCAEE had probably presented all features that are typical for the crown group. Therefore, numerous independent secondary simplifications must have occurred to generate simple extant eukaryotes from such a complex common ancestor. For instance, the LCAEE likely contained a large number of spliceosomal introns, which were massively and independently lost in most lineages (Roy and Gilbert, 2005). The most striking example are nevertheless mitochondria that were thought to have been independently lost many times. The discovery of genes of mitochondrial origin in the nucleus of supposedly amitochondriate eukaryotes first suggested their ancestral presence. Originally, it was thought that these genes had been transferred before the loss of mitochondria (Clark and Roger, 1995; Germot et al., 1997; Roger et al., 1998). However, and much more convincingly, remnants of mitochondria, i.e. small double-membrane-bound organelles containing nuclear-encoded mitochondrial proteins, were identified in all putatively "amitochondriate" organisms that were analysed for their presence (Bui et al., 1996; Tovar et al., 1999; Tovar et al., 2003; Williams et al., 2002). This indicates that the previously called "amitochondriate" organisms are simply lacking "aerobic respiration", the most prominent function of mitochondria. However, there are other functions performed by mitochondria that have persisted in all anaerobic eukaryotes, such as the synthesis of iron-sulphur clusters in diplomonads (Tovar et al., 2003).

8.5.3 Current view of eukaryotic evolution

Following the rejection of the Archezoa hypothesis, some progress was achieved in the resolution of deep eukaryotic phylogeny. Consequently, the division of all extant eukaryotes into six major groups was recently

proposed (Adl et al., 2005; Keeling et al., 2005; Patterson, 1999; Simpson and Roger, 2004). Some of these supergroups are solidly established, i.e. Opisthokonta (animals, choanoflagellates and Fungi) and Plantae (all three primary photosynthetic eukaryotes: green plants, red algae and glaucophytes). Moreover, there is accumulating evidence for both Amoebozoa (containing a large part of amoebas, e.g. Dictyostelium and Entamoeba) and Chromalveolata (alveolates, stramenoplies, haptophytes and cryptophytes); however, a final test based on the analysis of phylogenomic data set containing representatives from all major lineages from this supergroup is still missing. The support for the monophyly of the two remaining groups (Excavata and Rhizaria) is much more tenuous and definitively require much more sequence data in form of genome or EST projects. Finally, it is noteworthy that anaerobic/amitochondriate species appear to be found in most of the six supergroups.

8.5.4 Philosophical grounds for the rejection of secondary simplification

Despite the aforementioned progresses in eukaryotic phylogeny, the classical view, essentially rRNA-based, is still largely prevailing. Hence, we will try to explain why it is so difficult to obliterate, even in light of convincing evidence. The traditional taxonomy was influenced by the assumption that the evolution of Life was a steady rise to higher complexity, starting from "primitive" or "lower" (i.e. simple) organisms and ending with "evolved" or "higher" (i.e. complex) forms, especially humans. This conception is actually pre-darwinian and can be traced back to Aristotle's Scala Naturae, the great chain of being (Nee, 2005). As a result, organisms having an apparently simple morphology were naturally located at the base of the Tree of Life. In contrast, all recent molecular phylogenetic studies demonstrate that simplification constitutes a major evolutionary trend, encountered at all taxonomic levels. While this conclusion was already drawn more than 60 years ago by Andr Lwoff (Lwoff, 1943) its very slow acceptance by the scientific community was remarkably predicted and explained in the very same book (Lwoff, 1943). Briefly, the idea of complexification is tightly linked to the concept of progress through the implicit equation "progress = evolution towards more complexity". Consequently, the simplification process has always been affected by a strong negative connotation tending to its denial (Lwoff, 1943). From a sociological point of view, it should be noted that the rediscovery of simplification occurred concomitantly with se-

rious criticisms of the ideology of progress. Nevertheless, emphasising simplification does not deny that complexification did occur, but rather means that both processes should be taken into account for the reconstruction of the evolutionary past (Forterre and Philippe, 1999; Gould, 1996).

8.6 The Tree of Life

8.6.1 Solid facts and open questions

In their long quest towards the resolution of the Tree of Life, phylogeneticists generally agree that several of its major branches can be currently considered as reliably inferred. For example, the monophyly of two of the three domains of Life, Bacteria and Eukaryotes, is supported by numerous genes and does not seem to result from any known reconstruction artefact (Brown et al., 2001; Philippe and Forterre, 1999). Of course, the monophyly of a domain depends as a last resort on the location of the root of the Tree of Life because falling within a domain would render it paraphyletic rather than monophyletic (see below). In contrast, the monophyly of Archaea, which is often taken for granted (Woese et al., 1990), has never been significantly supported (Lake, 1988), even in multiple gene analyses (Brown et al., 2001; Daubin et al., 2002). Using a rate-invariant method applied to SSU-rRNA, Lake (Lake, 1988) proposed that Crenarchaeota are related to eukaryotes and Euryarchaeota to Bacteria, thus rendering Archaea paraphyletic whatever the location of the root. A subsequent analysis of both the SSU-rRNA and LSU-rRNA, as well as of the concatenation of the two largest subunits of the RNA polymerase, led to the same conclusion (Tourasse and Gouy, 1999). Hence, the monophyly of the archaeal domain is not yet established and deserves further studies.

The ancient events that are by far the most strongly supported are the endosymbiotic origins of the mitochondrion and the plastid, respectively from an ?-proteobacterium and a cyanobacterium. Moreover, within the three domains, the monophyly of the major phyla (e.g. Proteobacteria, Spirochaetes, Cyanobacteria, Crenarchaeota, Animals, Ciliates) is consistently recovered (Brochier et al., 2000; Daubin et al., 2002; Philippe et al., 2005b; Wolf et al., 2001).

As we have already pointed out, there nevertheless exist a few situations where the phylogenetic signal may be genuinely weak (or at least difficult to recover). These include (1) very rapid successions of specia-

tion events that are often characteristic for an adaptive radiation; (2) the presence of a very strong non-phylogenetic signal (e.g. the rate signal in microsporidia); or (3) the absence of closely related outgroup species due to the extinction of related groups (e.g. coelacanth or Amborella). The lack of close outgroups is particularly worrisome because it concerns various important groups, among which are angiosperms, mammals, birds, tetrapods, and even the three domains of Life: Archaea, Bacteria and Eukaryotes. Hence, in all these cases, there exists an increased risk of tree reconstruction artefacts due to the long branch of the outgroup attracting any fast-evolving ingroup like microsporidia or kinetoplastids (Brinkmann et al., 2005).

Therefore, the most important open questions concerning the Tree of Life are the reliable positioning of its root (addressed in the next section), followed by the relative branching order of the major phyla within each of the three domains. Not surprisingly, both kinds of issues are undisputedly affected by LBA artefacts (Brochier and Philippe, 2002; Lopez et al., 1999; Philippe et al., 2000) because of the great divergence separating such ancient groups.

8.6.2 The current paradigm: A bacterial rooting of the Tree of Life and a sister-group relationship between Archaea and Eukaryotes

Nowadays, a majority of researchers in the field regard as an established fact the specific relationship between Archaea and Eukaryotes that is deduced from a bacterial rooting of the Tree of Life. Originally proposed in 1989 (Gogarten et al., 1989; Iwabe et al., 1989) on the basis of anciently duplicated genes, this scenario was explicitly formulated the year after by Carl Woese (Woese et al., 1990). Because Woese's seminal paper led to the rapid and wide acceptance of these ideas, we will refer to the bacterial rooting of the Tree of Life and to the associated sister-group relationship between Archaea and Eukaryotes as Woese's paradigm. Until today, less than ten pairs of anciently duplicated genes, which already existed as two copies in the Last Universal Common Ancestor of extant Life (LUCA), have been identified. These universal paralogs can in principle be used to localize the root of the Tree of Life. Consistent with the paradigm, the analysis of all pairs but one (Lawson et al., 1996) by standard phylogenetic methods results in the bacterial rooting of the Tree of Life. However, there are serious reasons to assume that this rooting is actually due to an LBA artefact (Brinkmann and

Philippe, 1999; Forterre and Philippe, 1999; Lopez et al., 1999; Philippe and Forterre, 1999).

For nearly all molecular markers analysed for the rooting of the Tree of Life, the branches of Bacteria are the longest in each subtree, which potentially leads to their artefactual attraction by the very long branch of the outgroup (i.e. the paralogous group). Moreover, the long branch observed in Bacteria is not only true for anciently duplicated genes, but also for most genes involved in translation and transcription. Such a long branch likely stems from an acceleration of the evolutionary rate. A possible explanation for this phenomenon could be a simplification process during which Bacteria were radically streamlining the informational system inherited from a more complex common ancestor shared with Archaea and Eukaryotes. According to this view, the multi-subunit RNA polymerase and its TATA-box binding protein (the central element of the regulation of expression) found in Archaea and Eukaryotes would be ancestral, while the highly simplified bacterial polymerase (of the ?2??' type) and its single sigma factor derived (Brinkmann and Philippe, 1999). Because this drastic change has likely been associated with an accelerated evolutionary rate of Bacteria, the higher similarity between Archaea and Eukaryotes, especially observed for non-metabolic (informational) genes, is best explained as the result of the shared conservation of ancestral features (symplesiomophies), which are not informative in terms of phylogeny.

Furthermore, an analysis of the SRP54/SR? pair (involved in protein secretion), the most similar pair of anciently duplicated genes, focusing on the most slowly evolving positions recovers an eukaryotic rooting for the (SRP54) part of the tree (Brinkmann and Philippe, 1999). More precisely, in the SRP54 subtree, eukaryotes are basal, while prokaryotes form a monophyletic group composed of non-monophyletic Archaea and fast evolving Bacteria emerging from within the Archaea. The reduced distance of the outgroup coupled with the use of the slowly evolving positions greatly enhances the phylogenetic/non-phylogenetic signal ratio, as shown for animals (Fig. 4 and 5) and makes the eukaryotic rooting the most reliable hypothesis. Nevertheless, it should be noticed that this result is not statistically significant because too few positions remain available.

Certain authors interpret a shared insertion in the catalytic domain of the vacuolar ATPases (V-ATPases) from Archaea and Eukaryotes as a potential synapomorphy arguing for a sistergroup relationship of these two domains (Gogarten et al., 1989; Zhaxybayeva et al., 2005). How-

ever, the divergence observed between the V-ATPases and the bacterial F-ATPases is huge. Hence, there are at least three points to consider before drawing any conclusions based on V-ATPases. First, because producing a reliable alignment of the relevant part of the orthologs is far from obvious, we do not feel able to place the shared insertion in an alignment of the two paralogs (even more distant). Second, the ATPase family has three paralogs with a wide distribution within Bacteria and it is almost impossible to clearly establish which of the three bacterial copies is orthologous to the vacuolar ATPase (Philippe and Forterre, 1999). Third, in addition to most Archaea, there is a wide diversity of bacteria that actually possess both subunits of the V-ATPase, thus making questionable the scenario of an archaeal origin followed by a lateral gene transfer to Bacteria (Philippe and Forterre, 1999; Zhaxybayeva et al., 2005).

In conclusion, answering the question of the root of the Tree of Life and the related question of the basal groups within the three domains is of prime importance for our understanding of both LUCA and the early evolution of Life on Earth. Indeed, in the standard scenario derived from early studies (Burggraf et al., 1992; Gogarten et al., 1989; Iwabe et al., 1989), LUCA was a prokaryote-like hyperthermophilic organism (Fig. 6A), while in our alternative scenario based on refined phylogenetic analyses (Brinkmann and Philippe, 1999; Brochier et al., 2002), LUCA was an eukaryote-like mesophilic organism (Fig. 6B). Other important arguments in favour of an eukaryotic rooting of the Tree of Life are based on the RNA world hypothesis and are not detailed here (see Poole et al., 1999; Poole et al., 1998).

8.7 Frequent strong claims made with weak evidence in their favour

8.7.1 A genome fusion at the origin of eukaryotes

Starting a century ago with Mereschkowski's symbiogenesis (Mereschkowski, 1905), several genome fusion/or cell fusion scenarios to explain the origin of eukaryotes have been proposed (Lopez-Garcia and Moreira, 1999; Margulis, 1971). While these scenarios were originally grounded in Cell Biology, they gained acceptance when the rise of sequencing techniques suggested a chimerical nature for the eukaryotic cell, i.e. a greater similarity of metabolic genes to Bacteria and of informational genes to Archaea. The need to explain this chimerical nature was then triggering

the development of a multitude of additional fusion scenarios beginning in the late eighties (Zillig, 1987) and culminating in the nineties (Gupta and Golding, 1996; Martin and Mller, 1998; Moreira and Lpez-Garca, 1998). These scenarios can be separated into early and late variants depending on whether the fusion is supposed to have occurred before or after the divergence of the major prokaryotic phyla. Since they all assume that eukaryotes originated in a fusion of a bacterium and an archaeon, these scenarios actually imply that there were never any ancestral eukayotes prior to the fusion event, with Bacteria and Archaea being the only true ancestral lineages. Fusion scenarios are essentially compatible with Woese's paradigm (compare Fig. 6A and 6C). An additional yet obvious corollary of fusion scenarios is that they are supposed to have happened only once in the past and are heavily inspired by contemporary organisms that thrive in the same habitat. Finally, fusion scenarios most likely require partners lacking cell walls, which would otherwise hinder the fusion.

As already mentioned, the fundamental basis of these fusion scenarios is grounded in the observation that metabolic (operational) genes from eukaryotes are predominantly more similar to those of Bacteria, whereas genes of the genetic machinery (replication, transcription and translation) are more similar to those of Archaea. The similarity to bacterial metabolic genes can be straightforwardly explained by the massive gene transfer to the nucleus following both mitochondrial and plastid endosymbioses (Lang et al., 1999). In contrast, the similarity of the genetic machinery between Archaea and eukaryotes can be interpreted in two radically different ways. The standard interpretation is that this similarity reflects a genuine common ancestry (synapomorphies), while the alternative would be that it is due to sharing of ancestral states (symplesiomorphies), Bacteria being less similar because of an accelerated evolutionary rate. It is primordial to approximately know when this fusion event is proposed to have happened, because this will determine the expected relationships between nuclear genes of extant eukaryotes and genes found in the other two domains. Furthermore, one has to differentiate fusion hypotheses where the mitochondrial organelle has been simultaneously produced from those where the mitochondrial endosymbiosis is supposed to have happened later. For our purposes, we feel that a relatively cautious proposition should be the best starting point to verify deducible predictions. To our knowledge the most stringent formulation of a fusion scenario fitting the established scientific facts is the Hydrogen hypothesis (Martin and Mller, 1998). The elegance of

this scenario, which is essentially based on metabolic considerations, lies in the proposition that the bacterial partner (an alpha-protobacterium) will subsequently become the mitochondrion. In the related yet more complex Syntrophic hypothesis (Moreira and Lpez-Garca, 1998), the original bacterial partner is a delta-proteobacterium and the mitochondrial endosymbiosis takes place only afterwards. Finally, the archaeal partner postulated by both scenarios is a methanogenic archaeon.

If we take a closer look at the bacterial part of the postulated fusion, there are many eukaryotic genes that are specifically related to alphaproteobacteria, with most of them being encoded in the nucleus (Lang et al., 1999). This relationship is recovered despite the usually faster rate of eukaryotic sequences of mitochondrial origin relative to their bacterial counterparts. An even more specific and solid relationship is found for the eukaryotic genes of cyanobacterial origin that were introduced by the plastid endosymbiosis. In addition, many genes located in the eukaryotic nucleus that are similar to bacterial genes can not be associated with certainty to any of these two possibilities. This is not surprising because multiple causes may result in this situation. First, the most likely is that eukaryotic sequences are so divergent that the genuine phylogenetic signal has been erased. Second, the corresponding gene could have been lost in the original bacterial lineage. Third, some eukaryotic genes could have been laterally acquired from other Bacteria (Doolittle, 1998). While we could conclude that the observations are in excellent agreement with the predictions, we should keep in mind that we are actually looking at the evidence of the solidly established mitochondrial endosymbiosis. Indeed, the latter observations are theoretically indistinguishable from the predictions for the bacterial part of the Hydrogen hypothesis. Consequently, the only true phylogenetic test is to focus on the genetic machinery and to look for the sistergroup relationships between the genes of eukaryotes and those of methanogenic Archaea that are expected by both metabolic hypotheses (Martin and Mller, 1998; Moreira and Lpez-Garca, 1998).

While multiple examples exist, in which eukaryotic genes are more similar to the archaeal than to the bacterial counterpart, these usually involve distant relationships and do not point to any particular group of Archaea. This lack of specific affiliation is a major problem for the late fusion scenarios, because it suggests either that the sequences under study are too divergent or that the hypothetic methanogenic partner was not related to extant methanogens and has gone extinct. To further address these two possibilities we will give first some additional

information. In the analyses of the anciently duplicated genes that were used to infer the root of the Tree of Life, Archaea are usually the most slowly evolving group. Invariantly, any fusion scenario implies that the eukaryotic genes found in the nucleus of both bacterial and archaeal ancestry started to diverge from their prokaryotic counterparts at the same time. Therefore, it is difficult to explain the fact that the bacterial genes could be easily traced back to an alpha-proteobacteria, while the supposedly archaeal genes could not be affiliated to any extant methanogenic archaeon. In addition, the nuclear genes of archaeal ancestry have probably not been subject to the secondary simplification pressures that have been acting on mitochondrial genes of bacterial ancestry (and thus do not show an accelerated evolutionary rate). Taken together, these points rule out a large divergence as an explanation for the lack of a specific association between genes from eukaryotes and methanogenic Archaea comparable to the association observed between alpha-proteobacteria and mitochondrial genes. In order to examine the second possibility, i.e. that the hypothetic methanogenic partner was not related to extant methanogens and has gone extinct, it is worth noting that methanogenesis has most likely appeared only once within the domain Archaea (Slesarev et al., 2002). Moreover, methanogens are limited to a monophyletic group located in a derived position within the subdomain Euryarchaeota (Bapteste et al., 2005), which means that the ancestral metabolic type of this group was probably thermoacidophilic, as are the extant members of the subdomain Crenarchaeota. Finally, it seems that the split between these two archaeal subdomains is very deep (Brochier et al., 2005) and likely older than the origin of the late arising alpha-Proteobacteria, because the latter only evolved after the separation of the epsilon, the delta and the common ancestor of alpha and beta/gamma proteobacteria. Hence, the lack of any specific affinity to extant methanogenic Archaea is difficult to reconcile with the predictions of the two late fusion scenarios for the origin of eukaryotes. While the phylogenetic data are in favour of an extinct archaeal linaeage that would be basal to the known archaeal diversity, the assumption of contemporaneity inherent to any fusion scenario make this hypothetical archaeal partner incompatible with the time frame of the bacterial partner. Indeed, this would suggest that the primary separation into Cren- and Euryarchaeota did only occur after the diversification of alpha-proteobacteria (Cavalier-Smith, 2002). Therefore, sequence similarities between eukaryotes and Archaea would be better explained in

terms of common ancestry, as in Woese's original paradigm (Fig. 6A), or symplesiomorphies (Fig. 6B).

8.7.2 A hyperthermophilic origin of Life

Claims about a hyperthermophilic origin of Life are based on two different lines of arguments. The first involves the geological record, while the second results from early phylogenetic analyses of the SSU rRNA gene. In its distant past, Earth was very hot due to the frequent impacts of large meteorites that were generating enormous amounts of energy. According to most paleontologists, Life started very early in the history of our planet (about 3,800 mya) and only shortly after the end of the massive meteorite bombardment. Therefore, an ancestral adaptation to extremely thermophilic conditions would make sense, at least at first sight. Formally, an organism is recognised as a hyperthermophile when its optimal growth temperature is above 80C. Such a property is not known for any extant eukaryote. On the other hand, hyperthermophily in Bacteria is found in two groups, Thermotogales and Aquificales, that were among the most basal lineages in early rRNA trees. However, recent analyses suggest that these groups are probably related and not the most basal lineages in Bacteria (Brochier and Philippe, 2002; Daubin et al., 2002; but see Griffiths and Gupta, 2004). Actually, Thermotogales and Aquificales are only moderate hyperthermophiles, the most extreme hyperthermophiles being found within the Archaea. In that group, hyperthermophiles are broadly distributed, suggesting a hyperthermophilic origin of extant Archaea (Forterre et al., 2002). Interestingly, the reverse gyrase (involved in DNA supercoiling) found in Thermotogales and Aquificales appears to be of independent archaeal origin. This indicates that at least some of the sophisticated mechanisms required to thrive at high temperature first evolved in the common ancestor of Archaea to be only secondarily horizontally transferred to bacteria lineages adapting to a hyperthermophilic lifestyle (Forterre et al., 2000).

Nevertheless, the observation that there is no known hyperthermophilic eukaryote is often overlooked. The explanation most likely lies in the high instability of RNA at elevated temperatures, which implies that the half life of messenger RNA (mRNA) molecules is drastically reduced in a hyperthermophilic environment. While this is not a major issue for prokaryotes, which have a coupled transcription/translation machinery, this certainly penalises eukaryotes. In eukaryotes the premature mRNAs need to be processed (intron excision, addition of a cap and a

polyA-tail) before transfer across the nuclear envelope, in order to overcome the physical separation of transcription (nucleus) and translation (cytosol). These facts argue strongly against the possibility that hyperthermophilic eukaryotes did ever existed and limit possible scenarios for the relationship of the three domains of Life. Indeed, under a bacterial rooting, where eukaryotes derive late from Archaea, a hyperthermophilic origin of Life could be possible, whereas under an eukaryotic rooti ng, it would be excluded. However, the latter scenario is compatible with a secondary adaptation of the monophyletic prokaryotes to hyperthermophilic conditions, for example in the context of the Thermoreduction hypothesis (Forterre, 1995).

Whatever scenario of the origin and early evolution of Life is preferred, be it thermo-, meso- or even psychrophilic, there is an almost general consensus about the existence of an intermediate phase in the history of Earth known as the "RNA world". According to this convincingly substantiated theory (Jeffares et al., 1998; Poole et al., 1998), many functions today catalyzed by protein-based enzymes were originally performed by catalytic RNA molecules. Because of the thermolability of RNA, the RNA world hypothesis seems impossible to reconcile with the extreme thermal conditions associated with a hyperthermophilic origin of Life. Besides, current studies of thermophilic organisms and thermoadaptive mechanisms indicate that life at very high temperatures relies on the establishment and maintenance of complicated devices that are specific to present-day hyperthermophiles. Taken together, these observations support the idea of a secondary adaptation of already evolved and quite complex cells to thermo- and later hyperthermophilic conditions, rather than a hyperthermophilic origin of Life.

8.8 Conclusions

The advent of large-scale sequencing techniques has given rise to phylogenomics, a new discipline that attempts to infer phylogenetic relationships from complete genome data. Though not immune to the systematic biases present in some lineages, phylogenomics has largely benefited from the recent progress of probabilistic methods. Its resolving power has led to a drastic revision of the Tree of Life. One important lesson from our current understanding, especially in the eukaryotic domain, is that the equation 'simple equals primitive' rarely holds, i.e. many morphologically simple organisms have actually evolved from complex ancestors through a secondary simplification process. This does not mean that

the eukaryotic cell was complex from its very beginning, but rather that all extant eukaryotes can be traced back to an already highly evolved common ancestor, which has logically phased out its inferior competitors. Similarly, we have provided arguments in favour of a relatively sophisticated Last Universal Common Ancestor for all extant Life. In that context, prokaryotes would be the products of a streamlining process. While the emergence of the prokaryotic cell may reflect a secondary adaptation to thermophily (still largely present in Archaea), it is noteworthy that the RNA-world hypothesis rules out a hyperthermophilic origin of Life per se. Finally, there seems to be no stringent constraints on the early evolution of Life on Earth from a phylogenetic perspective. Hence, future research in Astrobiology should not rely on the biased picture of Evolution that was largely prevailing at the turn of the last century while looking for the manifestations of Life on other planets.

Acknowledgements

This work was supported by operating funds from Genome Qubec and NSERC. H.P. is members of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR), whom we thank for interaction support and is grateful to the Canada Research Chairs Program and the Canadian Foundation for Innovation (CFI) for salary and equipment support. D.B. is a postdoctoral researcher of the Fonds National de la Recherche Scientifique (FNRS, Belgium). D.B. is also gratefully indebted to the FNRS for the financial support of his stay at the University of Montral.

References

[177] Adl, S. M., A. G. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. Brugerolle, R. A. Fensome, S. Fredericq, T. Y. James, S. Karpov, P. Kugrens, J. Krug, C. E. Lane, L. A. Lewis, J. Lodge, D. H. Lynn, D. G. Mann, R. M. McCourt, L. Mendoza, O. Moestrup, S. E. Mozley-Standridge, T. A. Nerad, C. A. Shearer, A. V. Smirnov, F. W. Spiegel, and M. F. Taylor. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists, *J Eukaryot Microbiol* **52**, 399–451.

[178] Adoutte, A., G. Balavoine, N. Lartillot, O. Lespinet, B. Prud'homme, and R. de Rosa. (2000). The new animal phylogeny: reliability and implications, *Proc Natl Acad Sci U S A* **97**, 4453–6.

[179] Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals, *Nature* **387**, 489–93.

[180] Bapteste, E., C. Brochier, and Y. Boucher. (2005). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens, *Archaea* **1**, 353–63.

[181] Blair, J. E., K. Ikeo, T. Gojobori, and S. B. Hedges. (2002). The evolutionary position of nematodes, *BMC Evol* , –2:7.

[183] Brinkmann, H., M. Giezen, Y. Zhou, G. P. Raucourt, and H. Philippe. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics, *Syst Biol* **54**, 743–57.

[183] Brinkmann, H., and H. Philippe. (1999). Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies, *Mol* , –16:817-25.

[187] Brochier, C., E. Bapteste, D. Moreira, and H. Philippe. (2002). Eubacterial phylogeny based on translational apparatus proteins, *Trends Genet* , –18:1-5.

[187] Brochier, C., P. Forterre, and S. Gribaldo. (2005). An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences, *BMC Evol Biol* **5**, 36–

[187] Brochier, C., and H. Philippe. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria, *Nature* **417**, 244–

[187] Brochier, C., H. Philippe, and D. Moreira. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome, *Trends Genet* **16**, 529–533.

[189] Brown, J. R. (2003). Ancient horizontal gene transfer, *Nat* , –4:121-32.

[189] Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope. (2001). Universal trees based on large combined protein sequence data sets, *Nat* , –28:281-5.

[190] Bui, E. T., P. J. Bradley, and P. J. Johnson. (1996). A common evolutionary origin for mitochondria and hydrogenosomes, *Proc* , –93:9651-6.

[191] Burggraf, S., G. J. Olsen, K. O. Stetter, and C. R. Woese. (1992). A phylogenetic analysis of Aquifex pyrophilus, *Syst Appl Microbiol* **15**, 352–6.

[192] Burleigh, J. G., and S. Mathews. (2004). Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life, *Am* , –91:1599-1613.

[193] Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis, *Mol Biol Evol* **17**, 540–552.

[195] Cavalier-Smith, T. (1987). Eukaryotes with no mitochondria, *Nature* **326**, 332–333.

[195] Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification, *Int J Syst Evol Microbiol* **52**, 7–76.

[196] Clark, C. G., and A. J. Roger. (1995). Direct evidence for secondary loss of mitochondria in Entamoeba histolytica, *Proc* , –92:6518-21.

[299] Daubin, V., M. Gouy, and G. Perriere. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history, *Genome Res* , –12:1080-90.

[300] Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates, *Nature in press* , –

[300] Delsuc, F., H. Brinkmann, and H. Philippe. (2005). Phylogenomics and the reconstruction of the tree of life, *Nat* , –6:361-75.
[301] Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes, *Trends Genet* **14**, 307–11.
[301] Doolittle, W. F. (1999). Phylogenetic classification and the universal tree, *Science* **284**, 2124–9.
[203] Dopazo, H., and J. Dopazo. (2005). Genome-scale evidence of the nematode-arthropod clade, *Genome Biol* , –6:R41.
[203] Dopazo, H., J. Santoyo, and J. Dopazo. (2004). Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species, *Bioinformatics* **20**, –116-i121.
[204] Edlind, T. D., J. Li, G. S. Visvesvara, M. H. Vodkin, G. L. McLaughlin, and S. K. Katiyar. (1996). Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa, *Mol Phylogenet Evol* **5**, 359–367. Embley, T. M., and R. P. Hirt. 1998. Early branching eukaryotes? Curr Opin Genet Dev 8:624-629.
[206] Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading, *Syst* , –27:401-410.
[206] Felsenstein, J. (2004). Inferring phylogenies, *Sinauer Associates* , –
[207] Fitch, W. M. (1970). Distinguishing homologous from analogous proteins, *Syst* , –19:99-113.
[303] Forterre, P. (1995). Thermoreduction, a hypothesis for the origin of prokaryotes, *C R Acad Sci III* **318**, 415–22.
[303] Forterre, P., C. Bouthier De La Tour, H. Philippe, and M. Duguet. (2000). Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria, *Trends Genet* **16**, 152–4.
[303] Forterre, P., C. Brochier, and H. Philippe. (2002). Evolution of the Archaea, *Theor Popul Biol* **61**, 409–22. Forterre, P., and H. Philippe. 1999. Where is the root of the universal tree of life? BioEssays 21:871-879.
[211] Foster, P. G. (2004). Modeling compositional heterogeneity, *Syst Biol* **53**, 485–95.
[213] Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model, *Mol Biol Evol* **18**, 866–73.
[213] Galtier, N., and M. Gouy. (1995). Inferring phylogenies from DNA sequences of unequal base compositions, *Proc* , –92:11317-21.
[214] Germot, A., H. Philippe, and H. Le Guyader. (1997). Evidence for loss of mitochondria in Microsporidia from a mitochondrial- type HSP70 in Nosema locustae, *Mol Biochem Parasitol* **87**, 159–68.
[306] Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, and M. Yoshida. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes, *Proc* , –86:6661-5.
[216] Gould, S. J. (1996). Full House: The Spread of Excellence From Plato to Darwin, *Harmony Books* , –
[217] Griffiths, E., and R. S. Gupta. (2004). Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales, *Int Microbiol* **7**, 41–52.
[310] Gupta, R. S., and G. B. Golding. (1996). The origin of the eukaryotic cell, *Trends Biochem Sci* **21**, 166–71.
[219] Hendy, M. D., and D. Penny. (1989). A framework for the quantitative

study of evolutionary trees, *Syst* , –38:297-309.
[220] Hennig, W. (1966). Phylogenetic systematics, *University of Illinois Press* , –
[221] Huelsenbeck, J. P. (2002). Testing a covariotide model of DNA substitution, *Mol Biol Evol* **19**, 698–707.
[313] Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes, *Proc* , –86:9355-9.
[223] Jeffares, D. C., A. M. Poole, and D. Penny. (1998). Relics from the RNA world, *J Mol Evol* **46**, 18–36. Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics in press.
[225] Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray. (2005). The tree of eukaryotes, *Trends in Ecology* , –20:670-676.
[225] Keeling, P. J., and N. M. Fast. (2002). Microsporidia: biology and evolution of highly reduced intracellular parasites, *Annu Rev Microbiol* **56**, 93–116.
[226] Kolaczkowski, B., and J. W. Thornton. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous, *Nature* **431**, 980–4.
[227] Kreil, D. P., and C. A. Ouzounis. (2001). Identification of thermophilic species by the amino acid compositions deduced from their genomes, *Nucleic Acids Res* **29**, 1608–15.
[318] Lake, J. A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences, *Nature* **331**, 184–6.
[318] Lake, J. A. (1991). The order of sequence alignment can bias the selection of tree topology, *Mol* , –8:378-85.
[230] Lanave, C., G. Preparata, C. Saccone, and G. Serio. (1984). A new method for calculating evolutionary substitution rates, *J Mol Evol* **20**, 86–93.
[231] Lang, B. F., M. W. Gray, and G. Burger. (1999). Mitochondrial genome evolution and the origin of eukaryotes, *Annu* , –33:351-397.
[232] Lartillot, N., and H. Philippe. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process, *Mol* , –21:1095-1109.
[319] Lawson, F. S., R. L. Charlebois, and J. A. Dillon. (1996). Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life, *Mol* , –13:970-7.
[235] Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. (1992). Substitutional bias confounds inference of cyanelle origins from sequence data, *J* , –34:153-62.
[235] Lockhart, P. J., A. W. Larkum, M. Steel, P. J. Waddell, and D. Penny. (1996). Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis, *Proc* , –93:1930-4.
[322] Lopez, P., P. Forterre, and H. Philippe. (1999). The root of the tree of life in the light of the covarion model, *J* , –49:496-508.
[322] Lopez-Garcia, P., and D. Moreira. (1999). Metabolic symbiosis at the origin of eukaryotes, *Trends Biochem Sci* **24**, 88–93.
[238] Lwoff, A. (1943). L'volution physiologique, , –

[239] Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. (2001). Parallel adaptive radiations in two major clades of placental mammals, *Nature* **409**, 610–4.

[240] Margulis, L. (1971). Symbiosis and evolution, *Sci Am* **225**, 48–57.

[241] Martin, W., and M. Mller. (1998). The hydrogen hypothesis for the first eukaryote, *Nature* **392**, 37–41.

[242] Mereschkowski, C. (1905). Ueber Natur und Ursprung der Cromatophoren im Pflanzenreiche, *Biologisches Centralblatt* **25**, 593–604.

[243] Miyamoto, M. M., and W. M. Fitch. (1995). Testing species phylogenies and phylogenetic methods with congruence, *Syst* , –44:64-76.

[244] Moreira, D., and P. Lpez-Garca. (1998). Symbiosis between methanogenic archaea and delta-Proteobacteria as the origin of eukaryotes: The syntrophic hypothesis, *J Mol Evol* **47**, 517–530.

[245] Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. (2001). Molecular phylogenetics and the origins of placental mammals, *Nature* **409**, 614–8.

[246] Nee, S. (2005). The great chain of being, *Nature* **435**, 429–

[247] Ochman, H., E. Lerat, and V. Daubin. (2005). Examining bacterial species under the specter of gene transfer and exchange, *Proc Natl Acad Sci U S A* **102**, –1:6595-9.

[248] Pagel, M., and A. Meade. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data, *Syst Biol* **53**, 571–81.

[249] Patterson, D. J. (1999). The diversity of eukaryotes, *Am Nat* **154**, –96-S124.

[250] Philip, G. K., C. J. Creevey, and J. O. McInerney. (2005). The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa, *Mol* , –22:1175-84.

[330] Philippe, H. (2000). Long branch attraction and protist phylogeny, *Protist* **51**, 307–316.

[330] Philippe, H., and A. Adoutte. (1998). The molecular phylogeny of Eukaryota: solid facts and uncertainties, *Pages* **25**, 56– Philippe, H., A. Chenuil, and A. Adoutte. 1994. Can the cambrian explosion be inferred through molecular phylogeny ? Development 120:S15-S25. Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005a. Phylogenomics. Annu Rev Ecol Evol Syst 36:541-562.

[330] Philippe, H., and C. J. Douady. (2003). Horizontal gene transfer and phylogenetics, *Curr Opin Microbiol* **6**, 498–505.

[330] Philippe, H., and P. Forterre. (1999). The rooting of the universal tree of life is not reliable, *J* , –49:509-523.

[330] Philippe, H., and A. Germot. (2000). Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution, *Mol Biol Evol* **17**, 830–834. Philippe, H., N. Lartillot, and H. Brinkmann. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol 22:1246-53.

[330] Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller, and H. Le Guyader. (2000). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions,

Proc , –267:1213-1221.

[258] Poole, A., D. Jeffares, and D. Penny. (1999). Early evolution: prokaryotes, the new kids on the block, *Bioessays* **21**, 880–9.

[258] Poole, A. M., D. C. Jeffares, and D. Penny. (1998). The path from the RNA world, *J* , –46:1-17.

[259] Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes, *Nature* **402**, 404–7.

[260] Rodrguez-Ezpeleta, N., H. Brinkmann, S. Burey, B. Roure, G. Burger, W. Lffelhardt, H. Bohnert, H. Philippe, and B. Lang. (2005). Monophyly of primary photosynthetic eukaryotes: green plants, red algae and glaucophytes, *Current Biology* , –

[261] Roger, A. J., S. G. Svard, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin, and M. L. Sogin. (1998). A mitochondrial-like chaperonin 60 gene in Giardia lamblia: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria, *Proc* , –95:229-34.

[262] Roy, S. W., and W. Gilbert. (2005). The pattern of intron loss, *Proc Natl Acad Sci U S A* **102**, 713–8.

[263] Silberman, J. D., C. G. Clark, L. S. Diamond, and M. L. Sogin. (1999). Phylogeny of the genera Entamoeba and Endolimax as deduced from small- subunit ribosomal RNA sequences, *Mol Biol Evol* **16**, 1740–1751.

[264] Simpson, A. G., and A. J. Roger. (2004). The real 'kingdoms' of eukaryotes, *Curr Biol* **14**, –693-6.

[265] Slesarev, A. I., K. V. Mezhevaya, K. S. Makarova, N. N. Polushin, O. V. Shcherbinina, V. V. Shakhova, G. I. Belova, L. Aravind, D. A. Natale, I. B. Rogozin, R. L. Tatusov, Y. I. Wolf, K. O. Stetter, A. G. Malykh, E. V. Koonin, and S. A. Kozyavkin. (2002). The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens, *Proc Natl Acad Sci U S A* **99**, 4644–9.

[266] Sogin, M. L. (1991). Early evolution and the origin of eukaryotes, *Curr Opin Genet Dev* **1**, 457–463.

[267] Steel, M., and D. Penny. (2000). Parsimony, likelihood, and the role of models in molecular phylogenetics, *Mol* , –17:839-850.

[268] Stiller, J., and B. Hall. (1999). Long-branch attraction and the rDNA model of early eukaryotic evolution, *Mol Biol Evol* **16**, 1270–1279.

[269] Tourasse, N. J., and M. Gouy. (1999). Accounting for Evolutionary Rate Variation among Sequence Sites Consistently Changes Universal Phylogenies Deduced from rRNA and Protein-Coding Genes, *Mol Phylogenet Evol* **13**, 159–168.

[271] Tovar, J., A. Fischer, and C. G. Clark. (1999). The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica, *Mol Microbiol* **32**, 1013–1021.

[271] Tovar, J., G. Leon-Avila, L. B. Sanchez, R. Sutak, J. Tachezy, M. Van Der Giezen, M. Hernandez, M. Muller, and J. M. Lucocq. (2003). Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation, *Nature* **426**, 172–6.

[272] Wallace, I. M., G. Blackshields, and D. G. Higgins. (2005). Multiple sequence alignments, *Curr Opin Struct Biol* **15**, 261–6.

[273] Whelan, S., and N. Goldman. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-

likelihood approach, *Mol Biol Evol* **18**, 691–9.

[274] Williams, B. A., R. P. Hirt, J. M. Lucocq, and T. M. Embley. (2002). A mitochondrial remnant in the microsporidian Trachipleistophora hominis, *Nature* **418**, 865–9.

[345] Woese, C. R. (1987). Bacterial evolution, *Microbiol Rev* **51**, 221–71.

[345] Woese, C. R., O. Kandler, and M. L. Wheelis. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc* , –87:4576-9.

[278] Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades, *BMC Evol Biol* **1**, 8–

[278] Wolf, Y. I., I. B. Rogozin, and E. V. Koonin. (2004). Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis, *Genome Res* **14**, 29–36.

[281] Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites, *Mol Biol Evol* **10**, 1396–401.

[281] Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses, *Trends Ecol Evol* **11**, 367–370.

[281] Yang, Z., and D. Roberts. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life, *Mol Biol Evol* **12**, 451–8.

[351] Zhaxybayeva, O., P. Lapierre, and J. P. Gogarten. (2005). Ancient gene duplications and the root(s) of the tree of life, *Protoplasma* **227**, 53–64.

[283] Zillig, W. (1987). Eukaryotic traits in Archaebacteria, *Could the eukaryotic cytoplasm have arisen from archaebacterial origin* , –503:78-82.

9

Gene transfer, gene histories and the root of the tree of life

Olga Zhaxybayeva
Dalhousie University

J. Peter Gogarten
University of Connecticut

9.1 Introduction

Tracing organismal (species) histories on large evolutionary timescales remains a big challenge in evolutionary biology. Darwin metaphorically labeled these relationships the "Tree of Life", but in his notebook he expressed unhappiness with this label, because in the "Tree of Life" depicting species evolution only the tips of the branches are alive; this layer of living organisms rests on dead ancestors and extinct relatives. Darwin mused that therefore the term "Coral of Life" would be more appropriate (B26 in Darwin (1987)); however, this alternative label did not gain popularity. Different markers have been utilized to elucidate the relationship among different lineages: from morphological characters to complete genomes. Since complete genomes now are available for organisms from all three domains of life, it is possible to use large amounts of data to attempt deciphering the relationships between all known organisms.

Many comparative genome analyses have shown that different genes in genomes often have different evolutionary histories (e.g., Hilario and Gogarten 1993; Nesbø et al. 2001; Zhaxybayeva et al. 2004), which implies that the tree of life metaphor (and a bifurcating tree as a model for evolutionary relationships in general) might be no longer adequate (Doolittle 1999). The incongruence between gene histories can be attributed to many factors, one of which is horizontal gene transfer (HGT). Simulations based on coalescence have shown that HGT can affect not only the topology of an inferred phylogeny (and therefore inferences of last common ancestors), but also divergence times (Zhaxybayeva and Gogarten 2004; Zhaxybayeva et al. 2005). In this chapter we review

challenges in multi-gene analyses in light of HGT and discuss the consequences for inferring the position of the root of the tree/coral of life.

9.2 How to analyze multi-gene data?

In addition to different models and methods for sequence analyses (see chapter by Brinkmann et al. for an overview), many methodologies exist for the analysis of multiple genes. One approach is to combine individual genes into a single dataset to extract phylogenetic information that might be distributed over many gene families; this so-called "supermatrix approach" is often cited as a way to improve the resolution of the inferred phylogeny (Brown et al. 2001; Delsuc et al. 2005). In this approach, the individual gene alignments are concatenated into one super-alignment, if the individual gene histories are determined to have compatible (non-conflicting) phylogenetic histories. Then the super-alignment is analyzed as if it were one gene alignment, but with an advantage of containing a larger number of sites informative for phylogenetic analysis. A problem with direct concatenation is the selection of data to include. This selection is complicated by the fact that the absence of evidence for transfer cannot be taken as evidence for the absence of transfer. If one applies a stringent measure for the detection of conflict, nearly all genes agree with each other within the limits of confidence. The amount of conflict detected depends on the chosen limits of confidence and on the extent of taxon sampling (Snel et al. 2002; Daubin et al. 2003; Mirkin et al. 2003; Ge et al. 2005; Kunin et al. 2005). Testing the compatibility between different trees and the alignments from which these trees were derived using Shimodaira-Hasegawa (Shimodaira and Hasegawa 1999) or Approximately Unbiased (AU) test (Shimodaira 2002) has become the preferred tool to assess potential conflict between individual gene families (e.g., Lerat et al. 2003). In these tests, the fit of alternative topologies to an alignment is evaluated and the trees under which the data have a significantly worse probability are rejected and considered as incompatible with the data. (The probability of observing the data given a model of evolution and a phylogenetic tree is also known as the likelihood of a phylogenetic tree). However, the failure to reject a tree should not be mistaken as an evidence for congruence with a tree (Bapteste et al. 2004). A gene might indeed have evolved with a different history, and this history might be different from the consensus phylogeny, but the individual gene family contains too little phylogenetically useful information to make the likelihood of

the two phylogenies significantly different. This is analogous to the failure in detecting a significant correlation between fat intake and cancer (Prentice et al. 2006): it does not prove that the correlation does not exist; it only says that the correlation was not significant in the dataset, possibly because too small a sample was studied.

Another challenge to inferences based on concatenation may come from hidden, or unrecognized, paralogy in lineages that went through frequent gene duplication and aneupolyploidization (i.e., having multiple, albeit each incomplete, sets of chromosomes). Especially in animals with metameric organization (i.e., with the body divided into a number of similar segments), gene and genome duplications have long been postulated to have created the regulatory complexity necessary for the different bodyplans (e.g., Nam and Nei 2005). Multiple whole genome duplications were inferred for the early evolution of plants (Cui et al. 2006) and vertebrates (Escriva et al. 2002; Meyer and Van de Peer 2005). These gene duplications have led to gene families with often astounding diversity (e.g., Foth et al. 2006). However, it is not the complexity of the gene families in itself that generates problems in phylogenetic reconstruction; rather, the frequent loss of one or the other paralog (Hughes and Friedman 2004; Nam and Nei 2005) can lead to the inclusion of unrecognized paralogs in the datasets, with the result that some of the events in the genes' histories reflect gene duplication and not speciation events. For example, analyzing the homeobox gene superfamily in 11 genomes, Nam and Nei (2005) inferred 88 homeobox genes to have been present in the ancestor of bilateral animals. Thirty to forty of these were completely lost in one of the 11 species analyzed, many of the ones still represented underwent frequent gene duplications, especially in vertebrates, where more than 200 homeobox genes are found per haploid genome. In the study of four animal genomes, Hughes and Friedman (2004) observe massive (19-20% of gene families present in common ancestor) parallel loss in *Caenorhabditis elegans* and *Drosophila melanogaster*.

An alternative to the supermatrix approaches (Delsuc et al. 2005) is to analyze genes individually and to combine the resulting trees (or the bipartitions/embedded quartets constituting the trees) into a consensus signal (i.e., the phylogenetic signal supported by at least a plurality of genes) by using either supertree methods (Beiko et al. 2005), bipartition plotting (Zhaxybayeva et al. 2004) or quartet decomposition (Zhaxybayeva et al., accepted). As an example, Figure 1A shows bipartition analysis of 678 gene families present in 10 cyanobacterial genomes (see

Zhaxybayeva et al. (2004) for more details). A bipartition plot shows all bipartitions significantly supported by at least one gene family and allows us to extract quickly the plurality signal as well as gene families conflicting with it. Notably, for cyanobacteria, only three compatible bipartitions are supported by the plurality of genes, resulting in a not very resolved plurality consensus (Figure 1B). And even those three bipartitions are conflicted strongly by 13 gene families (i.e., these genes support a conflicting partition with >99% bootstrap support; see Figure 1 legend for the list of gene families). An advantage of the bipartition plotting approach is that only the signal retained in the plurality of datasets is used to synthesize the consensus; the disadvantage is that individual gene analyses often suffer from a lack of resolution due to an insufficient number of phylogenetically informative positions.

Regardless of the method to arrive at the consensus, the question still remains: what does the signal inferred from multi-gene data mean? Does it serve as a proxy for an organismal phylogeny? Does it sometimes reflect grouping by ecotypes? In the next section we explore these questions by considering an example of four marine cyanobacteria.

9.3 What does the plurality consensus represent?: Example of small marine cyanobacteria

At the time of the analysis, four genomes of small coccoid marine cyanobacteria of broadly similar lifestyles had been completely sequenced: *Prochlorococcus marinus* CCMP1375 (also known as SS120), *Prochlorococcus marinus* MED4, *Prochlorococcus marinus* MIT9313 and marine *Synechococcus* WH8102. Members of the *Prochlorococcus* genus have only been recently discovered due to their small size and anomalously low fluorescence (Chisholm et al. 1988). Marine *Synechococcus* and *Prochlorococcus* are proposed to diverge from a common phycobilisome-containing ancestor (Ting et al. 2002). (The phycobilisome is a light-harvesting complex associated with photosystem II and is used as a light-harvesting antenna in most cyanobacteria. Phylobilisomes, in contract to the light harvesting complexes of higher plants, are not embedded in the photosynthetic membrane, and they contain phycobilins as pigments, which give the cyanobacteria their typical blue-green color.) While marine *Synechococcus* still uses phycobilisomes as light-harvesting antennae (an ancestral trait), members of the *Prochlorococcus* genus lack phycobilisomes and utilize a different chlorophyl containing antenna complex (Pcb). *Prochlorococcus* also possess derivatives of chlorophyll a and b

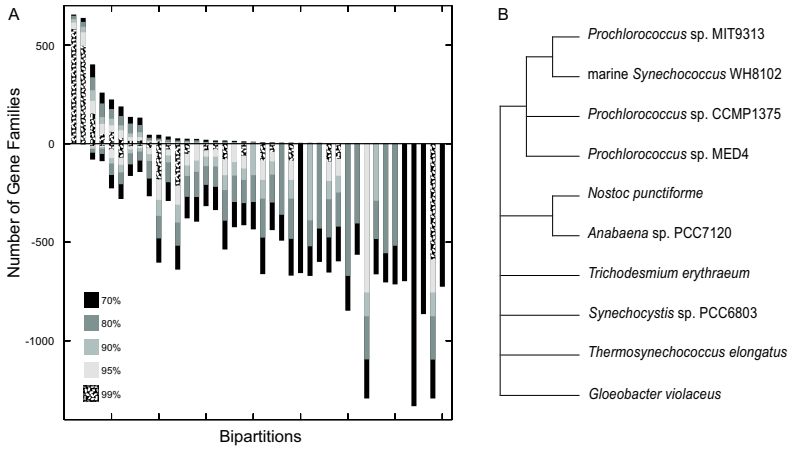


Fig. 9.1. Bipartition plot analysis of 678 gene families in 10 cyanobacterial genomes. An unrooted phylogenetic tree can be represented as a set of bipartitions (or splits). If a branch of a tree is removed, the tree "splits" into two sets of leaves. A bipartition of an unrooted phylogenetic tree is defined as a division of the tree into two mutually exclusive sets of leaves. (A) Plot of bipartitions with at least 70% bootstrap support. Each column represents the number of gene families that support (columns that are pointing upwards) or conflict (columns that are pointing downwards) a bipartition. The columns are sorted by number of gene families supporting each bipartition. The level of bootstrap support is coded by different shades of gray. For details of phylogenetic analyses see Zhaxybayeva et al. (2004). (B) Plurality consensus reconstructed from the three most supported partitions. The genes that are in conflict with the consensus at >99% bootstrap support encode ribulose bisphosphate carboxylase large subunit, cell division protein FtsH, translation initiation factor IF-2, ferredoxin, geranylgeranyl hydrogenase, amidophosphoribosyltransferase, photosystem II protein D2, photosystem II CP43 protein, photosystem II CP47 protein, photosystem I core protein A2, photosystem I core protein A1, photosystem II manganese-stabilizing protein and 5'-methylthioadenosine phosphorylase.

pigments that are unique to this genus (see Partensky et al. (1999) for a recent review). In addition, marine *Synechococcus* and *Prochlorococcus* are adapted to different ecological niches: marine *Synechococcus* is prevalent in coastal waters, while *Prochlorococcus* is ubiquitous in open subtropical and tropical ocean. Within *Prochlorococcus* marinus two "ecotypes" are differentiated: low-light adapted and high-light adapted types (Rocap et al. 2003). In the 16S rRNA tree, low-light adapted *Prochlorococcus* spp. form a paraphyletic clade with respect to high-light adapted *Prochlorococcus* spp. (Ting et al. 2002). In a recent

study, Beiko et al. (2005) report more than 250 HGT events among these marine cyanobacteria (Beiko et al. 2005). In their supertree as well as in our bipartition analyses (Zhaxybayeva et al. 2004) the marine cyanobacteria cluster into two clades: *P. marinus* CCMP1375 ('low-light' adapted) groups with *P. marinus* MED4 ('high-light'), and *P. marinus* MIT9313 ('low-light') groups with marine *Synechococcus* WH8102 (Beiko et al. (2005) and Figure 1B). Interestingly, the relationship among these four genomes as captured by the supertree and bipartition analyses neither supports the relationship inferred from phylogenetic analyses of 16S rRNA (e.g., Ting et al. 2002), nor the groupings based on proposed ecotypes (Ting et al. 2002; Rocap et al. 2003), nor the many derived characteristics shared between the three *Prochlorococcus* species. These results are confirmed through independent analyses of cyanobacteria utilizing embedded quartets (Zhaxybayeva et al., accepted). One explanation for the conflict between the genome consensus, and the many complex derived characters is rampant gene flow among these genomes, such that the plurality consensus no longer reflects the ecotype, physiology, and organismal history. In support for this hypothesis, cyanophages infecting marine cyanobacteria have been reported to contain genes important for photosynthesis (Mann et al. 2003; Lindell et al. 2004; Millard et al. 2004; Sullivan et al. 2005; Zeidner et al. 2005), and likely mediate transfer and recombination of these genes among the marine cyanobacteria (Zeidner et al. 2005).

Notably, such observation is not limited to prokaryotes. For example, in incipient species of Darwin's finches frequent introgression can make some individuals characterized as belonging to the same species by morphology and mating behavior genetically more similar to a sister species (Grant et al. 2004).

9.4 Where is the root of the "tree of life"?

In phylogenetic analysis using molecular sequences external information is required to identify the root of an inferred tree, i.e. the node from which all other nodes of the tree are descended. Usually, an outgroup is needed to root a tree, i.e. a taxon which is known to diverge earlier than the group of interest. However, this methodology becomes inapplicable to the problem of rooting the tree of life (since all organisms are part of the group of interest). One method to determine the root of the tree of life from molecular data is to use ancient duplication events. If a gene duplication event has occurred before the divergence of all extant

organisms, then the phylogenetic tree containing both gene duplicates will be symmetrical, and one set of duplicated genes can be used as an outgroup to the other. In the past, several pairs of anciently duplicated genes were detected and analyzed. The analyses produced a variety of controversial results as summarized in Table 1. Large-scale search of anciently duplicated genes did not bring any consensus (Zhaxybayeva et al. 2005), as locations of the root were observed in various places (Zhaxybayeva et al. 2005). If some groups of organisms evolve slower than others in molecular and physiological characteristics, then their phenotype might be considered as reflecting the phenotype of the most recent common ancestor (MRCA) (Woese 1987; Xue et al. 2003; Ciccarelli et al. 2006). However, these presumably primitive characteristics do not allow us to place the ancestor in tree of life, because they does not inform us about the relationship between the slow and fast evolving groups (Cejchan 2004).

Inspecting phylogenies based on molecular characters one often forgets that the deeper parts of the tree of life are formed by only those lineages that have extant representatives. However, these lineages were not the only ones present in earlier times. Many lineages went extinct, or at least their extant representatives have not been discovered. It is reasonable to assume that at the time of the organismal most recent common ancestor many other organisms were in existence as well (Figure 2). If the processes of speciation and extinction can be approximated by a random process, coalescence theory would be an appropriate tool to describe lineages through time plots (Zhaxybayeva and Gogarten 2004).

Due to vast time spans needed for the lineages to coalesce to their most recent common ancestor, the individual phylogenies from extant lineages usually have two longest branches leading to the MRCA, and therefore these phylogenetic analyses will be subject to potential long branch attraction (LBA) artifacts. Therefore, additional higher order shared derived characters (synapomorphies) are helpful to reinforce the inferred position of the root. Larger insertions/deletions or derived structural features are considered rare and mostly irreversible and may serve as synapomorphies (e.g., see Gupta (2001) for the application of presumably unique insertions/deletions to organismal classification). For example, the domain architecture differences between β and β' RNA polymerase subunits were used to polarize the tree of bacterial RNA polymerases (Iyer et al. 2004). In another study, the position of the root in the Glx-tRNA synthetase family tree was inferred from structural differences in anti-codon binding domains between GlnRS and GluRS

Table 9.1. *Different points of view on the location of the root of the tree of life.*

Location of the root	Phylogenetic Marker used	Citations
On the branch leading to the bacterial domain	V- and F-type ATPases catalytic and regulatory subunits	(Gogarten et al. 1989)
On the branch leading to the bacterial domain	translation elongation factors EF-tu/1 and EF-G/2	(Iwabe et al. 1989)
On the branch leading to the bacterial domain	Val/Ile amino-acyl tRNA synthetases	(Brown and Doolittle 1995)
On the branch leading to the bacterial domain	Tyr/Trp amino-acyl tRNA synthetases	(Brown et al. 1997)
On the branch leading to the bacterial domain	internal duplication in carbamoyl phosphate synthetase	(Lawson et al. 1996)
On the branch leading to the bacterial domain	components of signal recognition particle - the signal recognition particle SRP54 and its signal receptor alpha SRα	(Gribaldo and Cammarano 1998)
Inconclusive results: Archaea do not appear as a monophyletic group	components of signal recognition particle - the signal recognition particle SRP54 and its signal receptor alpha SRα	(Kollman and Doolittle 2000)
On the branch leading to the bacterial domain	aspartate and ornithine transcarbamoylases	(Labedan et al. 1999)
Inconclusive results: no statistical support for the best tree topology	histidine biosynthesis genes hisA/hisF	(Charlebois et al. 1997)
Within Gram-negative bacteria	Membrane architecture	(Cavalier-Smith 2002)
On the branch leading to the eukaryotic domain	translation elongation factor proteins, EF-1alpha and EF-2	(Forterre and Philippe 1999; Lopez et al. 1999)
On the branch leading to the bacterial domain, but the authors think it is an artifact due to LBA	elongation factors, ATPases, tRNA synthetases, carbamoyl phosphate synthetases, signal recognition particle proteins	(Philippe and Forterre 1999)
Between Archaea and Bacteria	structural features of Gln/Glu tRNA synthetases	(Siatecka et al. 1998)
Within Archaea	16S rRNA	(Lake 1988)
Within Archaea	transfer RNAs	(Xue et al. 2003)
Under aboriginal trifurcation	Various characteristics	(Woese et al. 1978)
Inconclusive results	RNA secondary structure	(Caetano-Anolles 2002)
Conceptual difficulties	-	(Bapteste and Brochier 2004)

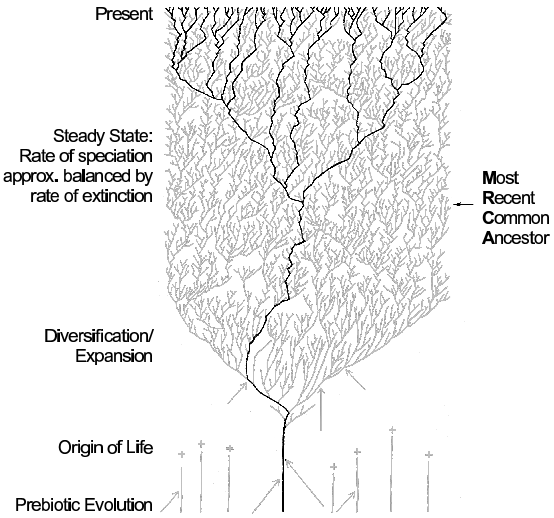


Fig. 9.2. Schematic depiction of a model for the Tree/Corral of Life highlighting the position of the most recent common ancestor. Extinct lineages are shown in gray. Extant lineages at the tip of the tree are traced back to their last common ancestors (in black).

(Siatecka et al. 1998). Below we examine in detail the use of additional characters in supporting the position of the root in the phylogenetic tree of ATPases.

9.5 Use of higher order characters: Example of ATPases

The use of ATPase catalytic and non-catalytic subunits to root the tree of life was originally introduced by Gogarten et al. (1989). This pair of anciently duplicated genes places the root on the branch leading to Bacteria with high confidence (see Figure 3). Either the catalytic or the non-catalytic subunits can be considered as ingroup, and the outgroup is provided by the paralogous subunits. The outgroup, a set of sequences rather divergent from the ingroup, joins the ingroup on the longest internal branch. While this placement of the root is recovered using different methods and evolutionary models (Gogarten et al. 1989; Iwabe et al.

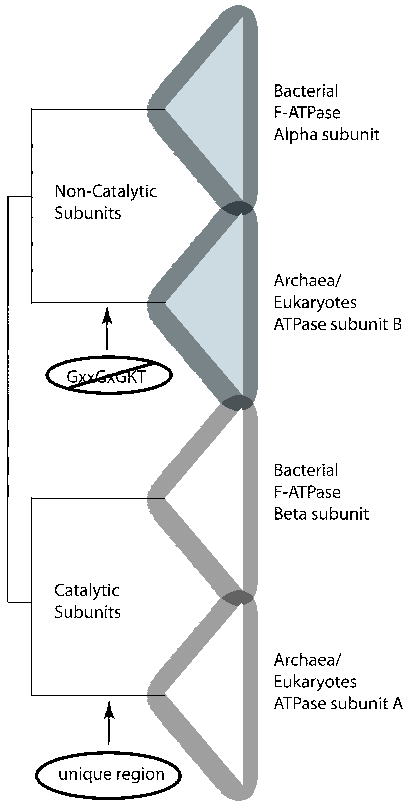


Fig. 9.3. Schematic tree showing the evolution of catalytic and non-catalytic subunits of ATPases (for detailed phylogeny see Fig. 2 in Zhaxybayeva et al. (2005)). Higher order characters are mapped to the branch leading to the clade where all the members of the clade possess the character. See text for more details.

1989), it also coincides with the place where the root would be located as the result of the LBA artifact (Philippe and Forterre 1999; Gribaldo and Philippe 2002). However, in case of the ATPases higher order characters exclude placing the outgroup within the archaeal/eukaryotic ATPases subunits (no higher order characters have been recognized for bacterial F-ATPases). The archaeal vacuolar ATPase non-catalytic subunits have lost the canonical Walker motif GxxGxGKT in their ATP binding pocket (Gogarten et al. 1989). This motif is present in the orthologous F-ATPase non-catalytic subunits as well as in all of the

ancient paralogs, including the paralogous Rho transcription termination factors (Richardson 2002) and ATPases involved in assembly of the bacterial flagella (Vogler et al. 1991). Similarly, the catalytic subunits of the archaeal and of the eukaryotic vacuolar type ATPase contain a faster evolving "non-homologous" region that has no counterpart in the catalytic F-ATPase subunits, nor is this region found in any of the non-catalytic subunits (Zimniak et al. 1988; Gogarten et al. 1989). The absence of the canonical Walker motif in the regulatory subunits and the presence of the non-homologous region in the catalytic subunits thus are shared derived characters of the vacuolar and archaeal ATPases that preclude moving the root of the ATPase phylogeny to a place within the clade constituted by the archaeal and eukaryotic ATPases.

9.6 Concluding Remarks

Consideration of gene transfer makes the analysis of species evolution among prokaryotes similar to population genetics. As is the case for within species analyses, gene trees for diverse groups of prokaryotes do not coalesce to the same ancestor, and the organisms that carried the molecular ancestors often lived at different times. For example, the human Y chromosome of all human males is traced back to a most recent common ancestor ("Adam") who lived approximately 50,000 years ago (Thomson et al. 2000; Underhill et al. 2000), while mitochondrial genes trace back to a most recent common ancestor ("Eve") who lived about 166,000-249,000 years ago (Cann et al. 1987; Vigilant et al. 1991). Many thousand years separate "Adam" and "Eve" from each other. The analogy to gene trees in recombining populations suggests that no single last common ancestor contained all the genes ancestral to the ones shared between the three domains of life. Each contemporary molecule traces back to an individual molecular most recent common ancestor, but these molecular ancestors were likely to be present in different organisms at different times. Therefore, even with more accurate phylogenetic reconstruction methods, one should not expect different molecular phylogenies to agree with one another on the placement of the most recent common ancestor of all living organisms. Adaptation of population genetics methodology may provide more fruitful results in studying early evolutionary events.

Acknowledgements

Olga Zhaxybayeva is supported through Canadian Institute of Health Research Postdoctoral Fellowship and is an honorary Killam Postdoctoral Fellow at Dalhousie University. This work was supported through the NASA Exobiology, NASA Applied Information Systems Research, and NSF Microbial Genetics (MCB-0237197) grants to JPG.

References

[284] Bapteste, E., Y. Boucher, J. Leigh, and W. F. Doolittle (2004). Phylogenetic reconstruction and lateral gene transfer, *Trends in Microbiology* **12**, 406–411.

[285] Bapteste, E., and C. Brochier (2004). On the conceptual difficulties in rooting the tree of life, *Trends in Microbiology* **12**, 9–13.

[286] Beiko, R.J., T. J. Harlow, and M. A. Ragan (2005). Highways of gene sharing in prokaryotes, *Proc Natl Acad Sci U S A* **102**, 14332–14337.

[287] Brown, J. R., W. F. Doolittle (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications, *Proc Natl Acad Sci U S A* **92**, 2441–2445.

[288] Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope (2001). Universal trees based on large combined protein sequence data sets, *Nat Genet* **28**, 281–285.

[289] Brown, J. R., F. T. Robb, R. Weiss, and W. F. Doolittle (1997). Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases, *J Mol Evol* **45**, 9–16.

[290] Caetano-Anolles, G. (2002). Evolved RNA secondary structure and the rooting of the universal tree of life, *J Mol Evol* **54**, 333–345.

[291] Cann, R. L., M. Stoneking, and A. C. Wilson (1987). Mitochondrial DNA and human evolution, *Nature* **325**, 31–36.

[292] Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification, *Int J Syst Evol Microbiol* **52**, 7–76.

[293] Cejchan, P. A. (2004). LUCA, or just a conserved Archaeon? comments on Xue et al., *Gene* **333**, 47–50.

[294] Charlebois, R. L., C. W. Sensen, W. F. Doolittle, and J. R. Brown (1997) Evolutionary analysis of the hisCGABdFDEHI gene cluster from the archaeon Sulfolobus solfataricus P2, *J Bacteriol* **179**, 4429–4432.

[295] Chisholm, S. W., R. J. Olson, E. R. Zettler, R. Goericke, J. B. Waterbury, and N. A. Welschmeyer (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone, *Nature* **334**, 340–343.

[296] Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork (2006). Toward automatic reconstruction of a highly resolved tree of life, *Science* **311**, 1283–1287.

[297] Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma, and C. W. dePamphilis (2006). Widespread genome duplications throughout the history of flowering plants, *Genome Res*, **16**, 738–749.

[298] Darwin, C. (1987). Charles Darwin's Notebooks, 1836-1844. Cambridge University Press.

[299] Daubin, V., N. A. Moran, and H. Ochman (2003). Phylogenetics and the cohesion of bacterial genomes, *Science* **301**, 829–832.

[300] Delsuc, F., H. Brinkmann, and H. Philippe (2005). Phylogenomics and the reconstruction of the tree of life, *Nat Rev Genet* **6**, 361–375.

[301] Doolittle, W. F. (1999). Phylogenetic classification and the universal tree, *Science* **284**, 2124–2129.

[302] Escriva, H., L. Manzon, J. Youson, and V. Laudet (2002). Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution, *Mol Biol Evol* **19**, 1440–1450.

[303] Forterre, P., and H. Philippe (1999). Where is the root of the universal tree of life? *Bioessays* **21**, 871–879.

[304] Foth, B. J., M. C. Goedecke, and D. Soldati (2006). New insights into myosin evolution and classification, *Proc Natl Acad Sci U S A* **103**, 3681–3686.

[305] Ge, F., L.-S. Wang, and J. Kim. (2005). The Cobweb of Life Revealed by Genome-Scale Estimates of Horizontal Gene Transfer. *PLoS Biology* **3**, e316.

[306] Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, and et al. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes, *Proc Natl Acad Sci USA* **86**, 6661–6665.

[307] Grant, P. R., B. R. Grant, J. A. Markert, L. F. Keller, and K. Petren. (2004). Convergent evolution of Darwin's finches caused by introgressive hybridization and selection, *Evolution Int J Org Evolution* **58**, 1588–1599.

[309] Gribaldo, S., and P. Cammarano. (1998). The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery, *Journal Of Molecular Evolution* **47**, 508–516.

[309] Gribaldo, S., and H. Philippe. (2002). Ancient Phylogenetic Relationships, *Theoretical Population Biology* **61**, 391–408.

[310] Gupta, R. S. (2001). The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int Microbiol* **4**, 187–202.

[311] Hilario, E., and J. P. Gogarten. (1993). Horizontal transfer of ATPase genes–the tree of life becomes a net of life, *Biosystems* **31**, 111–119.

[312] Hughes, A. L., and R. Friedman. (2004). Differential loss of ancestral gene families as a source of genomic divergence in animals, *Proc Biol Sci* **271 suppl 3**, S107–109.

[313] Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes, *Proc Natl Acad Sci U S A* **86**, 9355–9359.

[314] Iyer, L. M., E. V. Koonin, and L. Aravind. (2004). Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer, *Gene* **335**, 73–88.

[315] Kollman, J. M., and R. F. Doolittle. (2000). Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs, *J Mol Evol* **51**, 173–181.

[316] Kunin, V., L. Goldovsky, N. Darzentas, and C. A. Ouzounis. (2005). The net of life: reconstructing the microbial phylogenetic network, *Genome Res* **15**, 954–959.

[317] Labedan, B., A. Boyen, M. Baetens, D. Charlier, P. Chen, R. Cunin, V. Durbeco, N. Glansdorff, G. Herve, C. Legrain, Z. Liang, C. Purcarea, M. Roovers, R. Sanchez, T. L. Toong, M. Van de Casteele, F. van Vliet, Y. Xu, and Y. F. Zhang. (1999). The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor, *J Mol Evol* **49**, 461–473.

[318] Lake, J. A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences, *Nature* **331**, 184–186.

[319] Lawson, F. S., R. L. Charlebois, and J. A. Dillon. (1996). Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life, *Mol Biol Evol* **13**, 970–977.

[320] Lerat, E., V. Daubin, and N. A. Moran. (2003). From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria, *PLoS Biol* **1**, –19.

[321] Lindell, D., M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer, and S. W. Chisholm. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci U S A* **101**, 11013–11018.

[322] Lopez, P., P. Forterre, and H. Philippe. (1999). The root of the tree of life in the light of the covarion model, *J Mol Evol* **49**, 496–508.

[323] Mann, N. H., A. Cook, A. Millard, S. Bailey, and M. Clokie. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus, *Nature* **424**, 741.

[324] Meyer, A., and Y. Van de Peer. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD), *Bioessays* **27**, 937–945.

[325] Millard, A., M. R. Clokie, D. A. Shub, and N. H. Mann. (2004). Genetic organization of the psbAD region in phages infecting marine *Synechococcus* strains, *Proc Natl Acad Sci U S A* **101**, 11007–11012.

[326] Mirkin, B. G., T. I. Fenner, M. Y. Galperin, and E. V. Koonin. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evol Biol* **3**, 2–

[327] Nam, J., and M. Nei. (2005). Evolutionary change of the numbers of homeobox genes in bilateral animals, *Mol Biol Evol* **22**, 2386–2394.

[328] Nesbø, C. L., Y. Boucher, and W. F. Doolittle. (2001). Defining the core of nontransferable prokaryotic genes: the euryarchaeal core, *J Mol Evol* **53**, 340–350.

[329] Partensky, F., W. R. Hess, and D. Vaulot. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance, *Microbiol Mol Biol Rev* **63**, 106–127.

[330] Philippe, H., and P. Forterre. (1999). The rooting of the universal tree of life is not reliable, *J Mol Evol* **49**, 509–523.

[331] Prentice, R. L., B. Caan, R. T. Chlebowski, R. Patterson, L. H. Kuller, J. K. Ockene, K. L. Margolis, M. C. Limacher, J. E. Manson, L. M. Parker, E. Paskett, L. Phillips, J. Robbins, J. E. Rossouw, G. E. Sarto, J. M. Shikany, M. L. Stefanick, C. A. Thomson, L. Van Horn, M. Z. Vitolins, J. Wactawski-Wende, R. B. Wallace, S. Wassertheil-Smoller, E.

Whitlock, K. Yano, L. Adams-Campbell, G. L. Anderson, A. R. Assaf, S. A. A. Beresford, H. R. Black, R. L. Brunner, R. G. Brzyski, L. Ford, M. Gass, J. Hays, D. Heber, G. Heiss, S. L. Hendrix, J. Hsia, F. A. Hubbell, R. D. Jackson, K. C. Johnson, J. M. Kotchen, A. Z. LaCroix, D. S. Lane, R. D. Langer, N. L. Lasser, and M. M. Henderson. (2006). Low-Fat Dietary Pattern and Risk of Invasive Breast Cancer: The Women's Health Initiative Randomized Controlled Dietary Modification Trial, *JAMA* **295**, 629–642.

[332] Richardson, J. P. (2002). Rho-dependent termination and ATPases in transcript termination, *Biochimica et Biophysica Acta (BBA) Gene Structure and Expression* **1577**, 251–260.

[333] Rocap, G., F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser, and S. W. Chisholm. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation, *Nature* **424**, 1042–1047.

[335] Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection, *Syst Biol* **51**, 492–508.

[335] Shimodaira, H., and M. Hasegawa. (1999). Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* **16**:1114-1116.

[336] Siatecka, M., M. Rozek, J. Barciszewski, and M. Mirande. (1998). Modular evolution of the Glx-tRNA synthetase family–rooting of the evolutionary tree between the bacteria and archaea/eukarya branches, *Eur J Biochem* **256**, 80–87.

[337] Snel, B., P. Bork, and M. A. Huynen. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content, *Genome Res* **12**, 17–25.

[338] Sullivan, M. B., M. L. Coleman, P. Weigele, F. Rohwer, and S. W. Chisholm. (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3**:e144.

[584] Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data, *Proc Natl Acad Sci U S A* **97**, 7360–7365.

[340] Ting, C. S., G. Rocap, J. King, and S. W. Chisholm. (2002). Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies, *Trends Microbiol* **10**, 134–142.

[341] Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino, W. H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J. R. Kidd, S. Q. Mehdi, M. T. Seielstad, R. S. Wells, A. Piazza, R. W. Davis, M. W. Feldman, L. L. Cavalli-Sforza, and P. J. Oefner. (2000). Y chromosome sequence variation and the history of human populations, *Nat Genet* **26**, 358–361.

[342] Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. (1991). African populations and the evolution of human mitochondrial DNA, *Science* **253**, 1503–1507.

[343] Vogler, A. P., M. Homma, V. M. Irikura, and R. M. Macnab. (1991). Salmonella typhimurium mutants defective in flagellar filament regrowth and sequence similarity of FliI to F0F1, vacuolar, and archaebacterial ATPase subunits, *J Bacteriol* **173**, 3564–3572.

[345] Woese, C. R. (1987). Bacterial evolution, *Microbiol Rev* **51**, 221–271.
[345] Woese, C. R., L. J. Magrum, and G. E. Fox. (1978). Archaebacteria, *J Mol Evol* **11**, 245–251.
[346] Xue, H., K. L. Tong, C. Marck, H. Grosjean, and J. T. Wong. (2003). Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life, *Gene* **310**, 59–66.
[347] Zeidner, G., J. P. Bielawski, M. Shmoish, D. J. Scanlan, G. Sabehi, and O. Beja. (2005). Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates, *Environmental Microbiology* **7**:1505-1513.
[351] Zhaxybayeva, O., and J. P. Gogarten. (2004). Cladogenesis, coalescence and the evolution of the three domains of life, *Trends in Genetics* **20**, 182–187.
[351] Zhaxybayeva, O., P. Lapierre, and J. P. Gogarten. (2005). Ancient gene duplications and the root(s) of the tree of life, *Protoplasma* **227**, 53–64.
[351] Zhaxybayeva, O., P. Lapierre, and J. P. Gogarten. (2004). Genome mosaicism and organismal lineages, *Trends in Genetics* **20**, 254–260.
[351] Zhaxybayeva, O., J. P. Gogarten, R.L. Charlebois, W.F. Doolittle and R. T. Papke. (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events, *accepted*.
[352] Zimniak, L., P. Dittrich, J. P. Gogarten, H. Kibak, and L. Taiz. (1988). The cDNA sequence of the 69-kDa subunit of the carrot vacuolar H+-ATPase. Homology to the beta-chain of F0F1-ATPases, *J Biol Chem* **263**, 9102–9112.

10

Evolutionary Innovation versus Ecological Incumbency in the Early History of Life

Adolf Seilacher
Yale University

10.1 Abstract

As Darwinian evolution is a slow, but continuous process, one should expect that the fossil record reflects constant changes. This is not the case. At various levels one observes stasis (or incumbency) that must be rooted in ecosystems. The cascade mode of macroevolution is particularly evident in global mass extinctions – revolutions that wiped out ecologically ruling classes and gave previously subordinate ones a chance to rise to power.

The Cambrian explosion of metazoan phyla is no exception. It put an end to the Ediacaran world, which was (in the author's opinion) dominated by mat-forming prokaryotes and giant protozoans (Vendobionta and Xenophyophoria), while metazoans played a subordinate role for millions of years. Why didn't they take over right away? In the absence of macropredation, Ediacaran ecosystems could not be balanced by trophic relationships. Instead it is hypothesized that pathogens and endoparasites are the chief stabilizers, because they try to maintain the status quo for their own survival.

10.2 The Ediacaran World

The terminal period of the Proterozoic, called Ediacaran after a locality in the Flinders Ranges of South Australia (or Vendian in the Russian terminology), marks the first appearance of undoubted macrofossils. Because they are seemingly complex (and because they have been studied mainly by paleozoologists), all Ediacaran macrofossils were interpreted originally as early multicellular animals (metazoans). The title of Martin Glaessner's 1984 monumental book, " The Dawn of Animal Life",

expresses this view. The discovery of similar fossils in approximately 30 localities all over the world further contributed to the assumption that the Ediacaran Fauna represents simply a prelude to the Cambrian evolutionary explosion of metazoan phyla. Consequently, the Ediacaran period could be considered as the initial stage of the Paleozoic.

In the years following, this view has been challenged by the vendobiont hypothesis (Seilacher 1984, 1990, 1992). The challenge started with the observation that most of these organisms represent hydrostatic 'pneu' structures, whose various shapes were maintained by a quilted skin (as in an air mattress) and the internal pressure of the living content. Another feature shared by all vendobionts is the allometric growth of the quilt patterns. No matter whether the addition of new 'segments' continued throughout life (serial mode) or stopped at a certain point, followed by the expansion and secondary subdivision of established quilts (fractal mode), compartments never exceed a certain millimetric diameter. Such allometry is common in oversized unicellular organisms, such as certain algae (Acetabularia) and larger Foraminifera. It probably is controlled by the structure of the cells and metabolic diffusion distances (a globular cell the size of a walnut psychologically could not exist). Thus allometric compartmentalization of a multinucleate protoplasm by a chambered architecture may be considered as a strategy of unicellular (but multinucleate) organisms to reach large body sizes (i.e. as an alternative to multicellularity, which evolved independently in plants and animals). In conclusion, the quilted Ediacaran organisms are considered herein as giant unicells rather than multicellular animals (Seilacher et al., 2003)

Nevertheless, truly multicellular organisms have been found in Ediacaran biota. They are documented by (1) flattened macroalgae in contemporaneous black shales (Knoll, 2003), (2) phosphatized embryos (Donoghue and Dong, 2005), (3) shield-shaped sand sponges (Trilobozoa), (4) tubular shells (e.g. Cloudina), (5) worm burrows (Droser et al. 2005; Seilacher et al., 2005) and (6) the ventral death masks of a stem group mollusc (Kimberella), together with its radular scratches (Radulichnus). Except for the embryos and algae, these fossils are found in the same sandy facies that contain the vendobionts. Compared to them, however, the early metazoans remained rather small, much less numerous, and less diversified.

The title of a recent paper (Seilacher et al., 2003) expresses the revised scenario: "The Dawn of Animals in the Shadow Right of giant Protozoa". It also refers tacitly to a similar phenomenon in the evolution of mammals, who spent two thirds of their eologic history in the ecological

shadow of the dinosaurs. The present review focuses on this point: why didn't the evolutionary innovation lead immediately to ecological success? In other words (inspired by Chris McKay's talk at the conference): why didn't the Cambrian explosion of Metazoa happen earlier?

Before this question can be addressed, it is necessary to review the observational basis for the new scenario.

10.3 Preservational Context

The paleontological record is largely a matter of preservation. Dissociated bones and shells are the rule; soft-part impressions and articulated skeletons can be expected only in Konservat-Lagerstaetten (i.e. where hostile conditions near the bottom, mostly anoxia, fenced off scavengers and delayed decomposition). Rapid burial by event sedimentation also helps. As exemplified by the famous Burgess Shales (see contribution by Alexandra Pontefract and Jonathon Stone), host sediments usually were muddy, so that the fossils became flattened during compaction.

Ediacaran fossils do not fit this taphonomic scheme. They are found in sandstones, whose sedimentary structures reflect agitation by occasional storms or turbidity currents. Although Ediacarans were soft-bodied and had no stiff or mineralized skeletons, they left 3-dimensional impressions, as if cast with plaster of Paris. This paradox has been explained convincingly by Jim Gehling's (1999) "death mask" model: right after burial and death, the carcasses became coated by a bacterial film whose mineralization preserved their relief in three dimensions, but only on one side of the body. In the Flinders facies of S. Australia and the White Sea (Narbonne 2005), masks show the upper surface in vendobiont species that were mobile, felled, or attached flatly to the microbially bound bedding surface, while the lower side of the body is depicted in species that lived underneath the biomat, or in immersed anchors. In Newfoundland and England, organisms also are preserved on bedding planes; but due to the volcanic nature of the covering turbidities, felled as well as attached species preserve the *lower* surfaces of the fronds (but see Narbonne 2005).

Quite different is the Nama style of preservation in Namibia and other localities (Grazhdankin and Seilacher, 2002). There, similarly quilted organisms are 3-dimensionally preserved *within* flood-related sands (inundites). Their tops may be deformed by event erosion, but their convex bottoms never are overturned. This suggests that they lived within the sediment and were buried *in situ*.

10.4 Vendobionts as Giant Protozoans

As mentioned above, the view that the largest, most common, and most diversified organisms of the Ediacaran biota were giant protozoans (rather than metazoans) was based mainly on their allometric compartmentation. Partly overturned fronds show that quilting patterns were identical on upper and lower surfaces. The penetration of two infaunal *Pteridinium* specimens by another individual and their growth responses to this accident (Fig. 1) are hard to reconcile with metazoans. So are the resting tracks left behind by *Dickinsonia* and *Yorgia* (Fig. 2). They destroy the elephant-skin structure of the biomat and are too deep for having been produced simply by the weight of the organism; yet there are no scratches related to mechanical digging. Digestion of the biomat by microscopic pseudopodia best explains the morphology of these resting traces and the lack of a visible trackway connecting them.

Vendobionts, in the present definition sometimes resemble arthropods and other 'Articulata' by their segmentation. During ontogeny, minute new modules were introduced at the generative pole and grew larger by secondary expansion. Yet there are fundamental differences to segmented animals:

(1) Some vendobionts (e.g. the spindle-shaped form from Newfoundland; Fig. 3) had a generative pole at either end. In others, the 'head' segment could become secondarily generative after a traumatic accident (Fig. 1).

(2) Secondary growth concerned mainly the long axes of the segments, while their diameters remained fairly constant (sausage-shaped quilts).

(3) As there is no indication of molting, the flexible skin of these organisms must have been expandable. If the internal septa between quilts consisted of the same semi-rigid material, an arthropod-like ecdysis would have been impossible mechanically, particularly if quilts were further subdivided in a fractal mode.

(4) Vendobiont 'segments' never carry a functional earmark, not even the terminal module that would have contained the anus in a metazoan interpretation. Nor was there rigorously determined growth or any fixed countdown program.

(5) While segmental growth may have been controlled by Hox genes , there is no sign of a ParaHox cluster (Erwin, 2005) controling dorsoventral differentiation and the development of appendages.

There was, however, a major difficulty in considering the vendobionts as unicellulars: their quilts are much wider than the chamberlets of

large foraminifera and the cells of metazoans, whose diameters (rather than volumes) probably were restricted by metabolic activities (mainly diffusive) of the contained protoplasm. The solution to this dilemma came from *Xenophyophoria* that happen to survive on present deep-sea bottoms.

Some members of the unicellular, but multinucleate xenophyophores (*e.g. Stannophyllum*, Fig.4) resemble vendobionts by being foliate and consisting of negatively allometric, sausage-shaped chambers. These chambers, however, have agglutinated walls that can not expand secondarily, as did the quilts of vendobionts. Ediacaran xenophyophores (Fig. 4), which also include strings of globular chambers, lived in shallower waters than do modern ones and also embedded in microbial mats. This is why they are found *in situ* as positive hyporeliefs on bed soles, reminiscent of trace fossils. These protozoans also were compartmentalized allometrically. But, as in vendobionts, the chambers were too wide for unicellular standards. In xenophyophores, however, the reason for this dispropotion is known: rather than being filled by pure protoplasm, their chambers contain a softer *fill skeleton* (stercomare). Consisting of finer sediment particles taken up by the pseudopodia with the food, the function of the stercomare is to further subdivide the protoplasm into strands of permissible diameters (Tendal 1972). As observed in the Nama style of preservation (Seilacher et al. 2003) and in large *Charniodiscus* (Fig. 3), the chambers of Vendobionta appear to have been filled to about 50

In conclusion, it is reasonable to consider the Vendobionta as a coherent group of giant rhizopods that radiated into a variety of morphologies and life styles (Fig. 3). The epinym 'unicellular dinosaurs' refers to their unusual size, ecologic dominance and morphological disparity but also to their specialization, which made them more and more vulnerable towards global changes of any kind (Seilacher 1998). The coming of macropredation and the elimination of tough microbial mats by bioturbators were such global events

10.5 Kimberella as a stem-group Mollusc

As stated previously, the presence of truly multicellular animals in the Late Proterozoic is documented by various kinds of fossils. While 'worm' burrows and embryos are difficult to affiliate, *Kimberella* (originally described from South Australia as a cubozoan medusa) was clearly a mollusc-like creature. It is more common and better preserved in the White Sea region, Russia (Fig. 5). Although *Kimberella* certainly lived

above the sediment, death masks on sole faces from the White Sea preserve the ventral side. They provide the first earmark of a mollusc: the mucus secreted by the flat foot probably served as a matrix for the bacterial mask makers. Another clue is the morphology of this foot. It shows fine transverse wrinkles in the central area and a coarser crenulation along the circumference, related either to gills or to longitudinal and circular muscles that contracted upon death. One also recognizes the marginal impression of a limpet-like hood. As can be seen from occasional deformations, this hood was flexible and possibly covered by small sclerodermites, as in Cambrian halkieriids (Conway Morris and Peel, 1990). The ventral mask of *Kimberella* must have formed before the decay of the intestines, which caused an upward collapse in the center. All these features suggest an animal similar to Cambrian halkieriids or to modern polyplacophorans. However, it had no real shell and grazed algae not on rock surfaces, but on the biomats that were ubiquitous on Precambrian sea bottoms.

To these biomats we also owe a detailed record of feeding habits of *Kimbrella* specimens: radular scratch patterns (Radulichnus; Fig. 5). The radula is a feeding apparatus found in all modern mollusc classes except bivalves (where it probably has been lost with the transition to filter-feeding). The chitinous radula teeth scrape algal films from rock surfaces or an aquarium wall. On soft sediment, however, their characteristic scratch patterns become wiped out when the mollusc's foot crawls over them.

How could such scratches be preserved on Ediacaran soft bottoms? Precambrian sea bottoms were covered by biomats sufficiently tough to carry an animal the size of Kimberella specimens without leaving a trail. Radular teeth, on the other hand, penetrated deep enough to produce sharp undertraces at the mat's base, which is marked commonly by distinctive 'elephant-skin' structures (Fig. 6). Only the grazing activities of juveniles fail to be recorded, because their teeth did not penetrate deep enough to produce an undertrace.

In fact, Kimberella individuals would not have wiped out their own raspings anyway, due to a very unusual mode of grazing. Modern molluscs (and cows) move their heads to the left and right while grazing and crawl a step forward after every swing. Kimberella specimens, however, stayed in place and foraged with the mouth at the tip of a retractible trunk. Therefore the bipartite scratches are not arranged in continuous guided meanders, but in concentric arcs (Gehling, personal communication, 1995). As the width of the swing increased automatically the

wider the trunk extended, and the scratch field produced during each meal was conical, rather than a continuous band of meanders. Obviously, less energy had to be spent in stationary than in mobile grazing; but an extended trunk was probably too vulnerable after the onset of macropredation in the Cambrian ecologic revolution.

10.6 Worm Burrows

After seemingly complex trace fossils have been identified as either xenophyophores or pseudofossils, there still remains a fair number of distinctive "worm" burrows (Fig. 7). For example, *Nenoxites* specimens are found on top surfaces with an apparently pelleted wall, while Palaeophycus specimens commonly are preserved as a hypichnial furrow. The positive epirelief of Aulichnites specimens (Fig. 7) indicate active backfilling, but the median furrow and angular bending would better fit a short, possibly molluscan trace maker. Specimens of *Aulozoon* (Fig. 8), looking like a highly compressed sand sausage, likely are the backfilled burrows of flatworms (but see Droser et al., 2005).

Although being made by infaunal animals, these burrows penetrated no more than a few millimeters below the sediment-water interface. They can be interpreted adequately as undermat miners (Seilacher 1999), which fed on the decaying lower zone of the biomat. Deeper and more extensive sediment mixing (bioturbation) started only after the Cambrian substrate revolution (Bottjer et al. 2000), when many animals responded to the menace of macropredation by becoming infaunal.

Trace fossils also convey another message. As one knows from later examples, biomats provide an ideal substrate for tracks of arthropods, because their pointed leg tips (mineralized or not) penetrated the mat as easily as did radula teeth. Nevertheless, no such tracks have been found in the Precambrian.

While arthropods evidently were absent, other trace fossils (Bergaueria, Fig. 6) document the presence of actinia-like coelenterates. Unlike the resting traces of vendobionts (Fig. 2), they show concentric scratch patterns, which formed when the bottom parts of the hydraulic polyps actively burrowed for anchoring the body and withdrew it into the sediment when disturbed.

In conclusion, stem groups of various modern animal phyla certainly were present in the benthic communities of Ediacaran times and increasingly so towards the end of the period. Yet they remain rare and small compared to the associated vendobionts and xenophyophores. Neither

did the earliest animals evolve sophisticated search programs, nor have they ravaged the unprotected giant protists. Overall, the peaceful "Garden of Ediacara" (Mc Menamin 1986) appears to have been dominated by prokaryote biomats and giant protozoans. Even though the taxonomic composition of the biota may have changed (Narbonne 2005), the situation remained the same throughout Ediacaran times and in various environments. What made this strange world survive through tens of million years, in spite of the presence of metazoan animals?

10.7 Stability of Ecosystems

Even though Darwinian processes call for continuous change, the paleontological record tells us that macroevolution proceeded in cascades. From the level of local communities (Brett and Baird 1995) to that of global faunas (Sepkoski and Miller, 1985), one observes long periods of relative stasis interrupted by shorter intervals of rapid change (Gould 2002). The latter culminate in mass extinctions, of which the end of the Ediacaran world is a typical example. Their causation has been the matter of much debate; but it now appears that a variety of events may act as triggers, as long as they have a global effect. Another important factor is probably the readiness of the biosphere. Mass extinctions are typically preceded by greenhouse periods, in which warm climates allowed evolution to reach relatively high levels of niche partitioning and of morphological as well as behavioural specialization. This, in turn, led to an increased vulnerability in the face of environmental changes of any kind (Seilacher 1998). In a sense, the giant protozoans of Ediacaran times may be compared to Cretaceous ammonites, rudists and dinosaurs, although their demise and the "Cambrian Explosion" of metazoans probably was due to radical ecological changes (onset of macropredation; bioturbational substrate revolution) rather than an asteroid impact (see contribution by Alexandra Pontefract and Jonathon Stone).

The modalities of the *Cambrian Explosion*, however, do not explain why this turnover happened so late. Why doesn't a superior product lead to immediate success, as it should in a free market? Obviously the impediment relates to the ecological structure of the biosphere, rather than to the genome. One handy answer could be that metazoan newcomers found the best niches already occupied and had to wait for the demise of the occupants. But, firstly, niches are no independent entities; they can be defined only relative to an organism and its particular

requirements. In the present interpretation, the incumbent organisms were so different from later macrofaunas (for instance they obtained food and oxygen with microscopic pseudopodia spread over the whole body surface, in contrast to metazoan filter feeders) that niches could not be congruent. Secondly, niche diversification and trophic chains still were developed poorly in Ediacaran times, so that the subsequent metazoan radiation could fill an almost empty ecospace.

10.8 The Parasite Connection

For most of us, endoparasites and pathogens have a negative connotation and we think that the world would be better without them. In a less anthropocentric view, however, they appear to be essential for the long-term survival of all ecosystems. Like the arms race between predator and prey species at the trophic level, a constant race takes place between hosts and their parasites. But as we know, there is no ultimate winner, because most defence mechanisms (and drugs) work only at the level of the individual host and are eventually superseded by the genetic flexibility of rapidly reproducing parasites.

The positive effect of parasites derives from their vital interest in maintaining the status quo. To this end, all of them dampen fluctuations in population size caused by predator/prey cycles. In addition, heteroxenic parasites (protozoans and metazoans that reproduce asexually in the intermediate host and sexually in a quite different final one) balance the system between the two kinds of hosts by doing less harm to the weaker than to the stronger partner. For instance, in the case of the lion-gnu system (Seilacher et al. submitted; Fig, 8), the herbivores suffer little from their sarcocystid infection, except that the most loaded individuals become behaviourally conditioned for falling prey. In contrast, only one out of five lion cubs reaches the second year in its natural environment, largely due to the protozoan parasites.

Similarly, trematodes (liver worms) spare the intermediate host, a freshwater snail, by stopping their own reproduction when the host switches to hunger metabolism (Seilacher et al., submitted, and literature cited therein).

Another important feature is the host specifity of parasites. In the lion-gnu example, another sarcocystid species cohabiting in the same herbivore can not develop in the lion, but needs to reach a hyena sharing the same prey. This specifity, however, is at a high taxonomic level. The lions's parasite also may develop in any other member of the cat family

and the one of the hyena in any canid, because all constituent species share a similar histological and physiological outfit. In this way, higher taxonomic categories may become units of selection (for instance in mass extinctions). On the other hand, such flexibility makes parasites tolerant against minor changes of host species due to immigration, speciation, or extinction. Parasitic partnerships also may survive dramatic habitat changes of the host, as long as adequate transfer mechanisms can be established in the new environment. Thus, sarcocystids similar to the ones in the lion-gnu system went to sea with the whales in the Early Tertiary. Now baleens (derived from herbivorous ungulates) serve as their intermediate hosts, and the orcas (probably derived from terrestrial carnivores) as final hosts, while herrings may be the transport hosts back to baleens. The parasites' fidelity may last until a whole clade of hosts becomes eradicated in a mass extinction. On the other hand, their time-constant and well isolated environment allows endoparasites themselves to radiate in coevolution with their hosts, as shown by the large number of sibling species.

In conclusion, probably no animal (or plant) is without endoparasites and as the same organism hosts different kinds of them, parasites form networks glueing together established ecosystems.

Although we are far from understanding all these complex relationships (parasites that have no direct or indirect effect on humans remain underexplored), they possibly account for the stasis of ecosystems observed in the fossil record. Cases concerned range from the long-term identity of benthic communities ("coordinated stasis"; Brett and Baird 1995), to the evolution of "exclusive clubs" in ancient lakes and semi-restricted basins (snails in the Miocene Steinheim Lake and ceratites in the Triassic Muschelkalk Basin; Seilacher et al., submitted), and to the incumbency of ecological power structures at the largest scale.

Possibly the persistence of the Ediacaran biota has to do with such connections. Although the events in the Precambrian-Cambrian transition were different from the ones that led to the extinction of the dinosaurs, they resulted in the collapse of established ecosystems and allowed multicellular animals to become the rulers.

10.9 Conclusions

The Darwinian foundation of the evolutionary process is well established. The diversity of organisms, their geographic distribution, and molecular data are living testimonies. Yet, this theory fails to explain

long-term (macroevolutionary) patterns of stasis observed in the fossil record. Some of them have extrinsic causes. Periods of warmer greenhouse climates kindled evolutionary diversification and made the biosphere more vulnerable. Global catastrophies of various kinds then led to mass extinctions. But they did not happen at random intervals, because ecosystems need time to mature. The stasis observed must have its roots in interactions between species at the level of ecosystems. The processes involved are not recorded by fossils and they are difficult to approach experimentally because of the geologic time scales at which they operate.

Ediacaran biota are not only the earliest (and thereby strangest) ecosystems on earth that allow a paleobiological analysis. They also provide a prime example for the incumbency of such systems. The rulers in terms of numerical dominance and body sizes were not the members that we would, in retrospect, consider to have had the most-derived traits (the metazoans in this case), but highly developed ancient groups, such as the probably unicellular vendobionts. The Ediacaran also is unique by lacking macropredation, which modulated trophic chains in later times.

Today, endoparasites of viral, bacterial, protozoan and metazoan affiliations are present in virtually every organism. From the relatively few cycles that have been studied more extensively because of their epidemological and medical importance, we know that parasites have a major effect on the maintenance of ecological networks. Monoxenic pathogens dampen fluctuations in population size by series of endemic and epidemic states. Heteroxenic parasites, using more than one host, moreover control the balance between the two partners and thereby contribute in their own interest to the persistence of established ecosystems.

In the long-term view, we should better give up traditional anthropocentric preconceptions. Pathogens and endoparasites may be detrimental and deadly in the short term, but they contribute to the maintenance of the status quo. By the same token, endo-symbionts appear to be more advantageous; but they also make ecosystems more vulnerable against environmental changes, as can be observed in modern coral reefs.

There is no reason to believe that things were basically different in Ediacaran biota. They were incumbent despite the absence of macropredation. If life would start anew or on another planet, the same kind of interactions would probably evolve: competitional change at the visible scale versus conservational players acting behind the ecological scene,

just as mutating selfish genes (Dawkins 1999) behind the Darwinian one. The problem is to scientifically prove or falsify such a scenario.

Acknowledgements

Thanks to Peter Wenk for opening my eyes towards parasites and the organizers of the symposium for inviting me and making me think about life beyond our own planet. Leo Hickey and Krister Smith (Yale University) critically read earlier versions and my wife Edith pointed out major inconsistencies. Roger Thomas (Lancester) pointed to the analogy with selfish genes.

References

[490] Bottjer, D.J., Hagadorn, J. W. and Dornbos, S.Q., 2000. The Cambrian Substrate Revolution. *GSA Today*, 10: 1-9

[490] Brett, C. E., Baird, G. C. 1995. Coordinated stasis and evolutionary ecology of Silurian to Middle Devonian faunas in the Appalachian Basin. In: Erwin, H. E. and Anstey, R. L. (eds.), New Approaches to Speciation in the Fossil Record. Columbia Univ. Press: 285-315.

[490] Conway Morris, S. and Peel, J.S., 1990. Articulated halkieriids from the Lower Cambrian of north Greenland. *Nature* 345: 802-805

[490] Dawkins, R. 1999: The selfish gene. Oxford University Press, 352 pp.

[490] Droser , L. M., Gehling, J. G. and Jensen, S., 200?. Ediacaran Trace Fossils: True and False. Special Publication, Peabody Museum of Natural History, Yale University: 125-138

[490] Donoghue; P.C. J. and Dong, X., 2005. Embryos and Ancestors. In: D. Briggs (ed.) Evolving Form and Function: Fossils and Development. Special Publication, Peabody Museum of Natural History, Yale University: 81-99

[490] Erwin, H. D. 2005. The Origin of Animal Body Plans. In: D. Briggs (ed.) Evolving Form and Function: Fossils and Development. Special Publication, Peabody Museum of Natural History, Yale University: 67- 80.

[490] Gehling, J.G. 1999. Microbial mats in terminal Proterozoic siliciclasics: Ediacaran death masks. *Palaios* 14: 40-57

[490] Gehling, J.G.Droser, M.L., Jensen, S.R. and Runnegar, B.N. 2005. Ediacara Organisms: Relating Form to Function. In: D. Briggs (ed.) Evolving Form and Function: Fossils and Development. Special Publication, Peabody Museum of Natural History, Yale University:43-66

[490] Glaessner, M. 1984. The Dawn of Animal Life: A biohistoric study. Cambridge University Press, 370 pp.

[490] Gould, S.J. 2002. The structure of evolutionary theory. Belknap Press; London, 1433 pp.

[490] Grazhdankin, D. and Seilacher A. 2002. Underground Vendobionta from Namibia. *Palaeontology* 45, 57-78.

[490] Knoll, A.H. 2003. Life on a young planet. The first three billion years of Evolution on Earth. Princeton, N.J.: Princeton University Press: 287pp.

[490] McMenamin, M.A.S. 1986. The garden of Ediacara. *Palaios*, 1: 178-182.
[490] Narbonne, G.M., 2005. The Ediacaran biota: Neoproterozoic origin of animals and their ecosystem. *Ann Rev. Earth Planet. Sci.* 33: 13.1-13.22
[490] Seilacher, A. 1984. Late Precambrian and early Cambrian Metazoa: preservational or real extinctions? -In: H.D.Holland A.F.Trendall (eds.), Patterns of Change in Earth Evolution. Dahlem Konferenzen: 159-168, Springer Verlag, Heidelberg.
[490] Seilacher, A. 1992. Vendobionta and Psammocorallia: lost constructions of Precambrian evolution. *J.Geological Soc.London*, 149:609-613
[490] Seilacher, A. 1998. Patterns of Macroevolution: how to be prepared for extinction - Comptes rendus de l'cademie des Sciences, Sciences de la terre et des planetes, 327:431-440.
[490] Seilacher, A., 1999. Biomat-related lifestyles in the Precambrian. *Palaios*, 14: 86-93.
[490] Seilacher, A., Grazhdankin, D. and Legouta, A. 2003. Ediacaran Biota: The dawn of animal life in the shadow of giant protists. *Paleontological Research* 7: 43-54
[490] Seilacher, A., Wenk, P. and Reif, W.E. (submitted). The parasite connection in Ecosystems and Macroevolution. Naturwissenschaften.
[490] Sepkoski, J. J. Jr Miller, A.I. 1985 Evolutionary faunas and the distribution of Paleozoic benthic communities. In: Valentine, J.W. (ed.) Phanerozoic diversity patterns: Profiles in macroevolution., Princeton University Press :153-190
[490] Tendal, O. S., 1972. A monograph of the Xenophyophoria (Rhizopoda, Protozoa). *Galathea Report*, 12: 7-99.

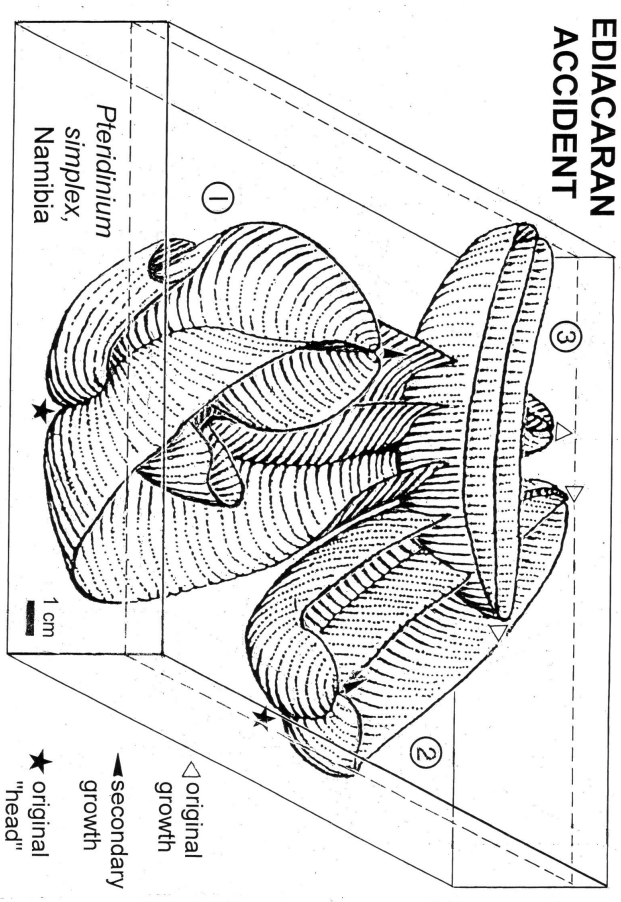


Fig. 10.1. Infaunal vendobionts, such as Pteridinium, were immobile. When a newcomer (nr. 3) grew through resident individuals, the latter responded by growing upwards at the wrong ("head") end. Nr. 1 first broadened its lateral vanes in the proximal part. As it turned over to a higher level, the vanes swopped functions. The left lateral vane retained its position, but after having grown a pronounced fold it reversed dorsoventrality; the right vane turned smoothly into the new median vane; and the original median vane became the new right vane, again with a fold. Nr. 2 had only started to do the same before the whole colony became terminally smothered. As such overfolding by growth is rather common, it may have also been induced by sedimentation alone. (Modified from Grazhdankin and Seilacher, 2002)

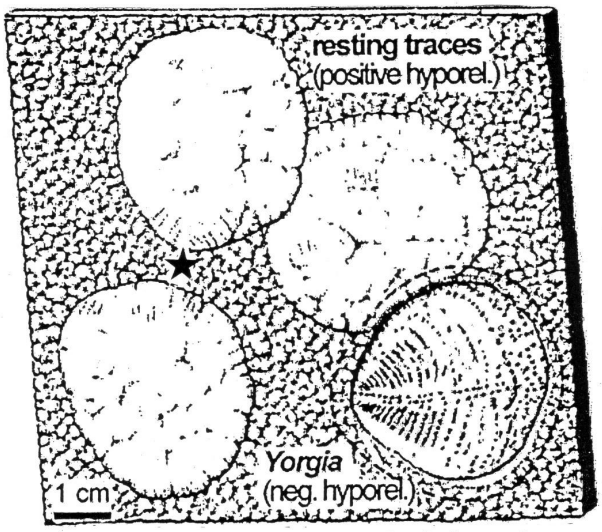


Fig. 10.2. Some prostrate vendobionts were able to creep over the biomat (elephantskin structure, leaving behind series of resting traces but no connecting trail. Their bumpy surface without scratch marks suggests a digestive mode of burrowing. Occasionally (star), the trace maker impressed its quilts before leaving. Note that the traces and their axes are not aligned. They also have larger diameters (broken line) than the negative hyporelief mask at the last station, suggesting contraction of the organism while feeding. (After Fedonkin, 2003).

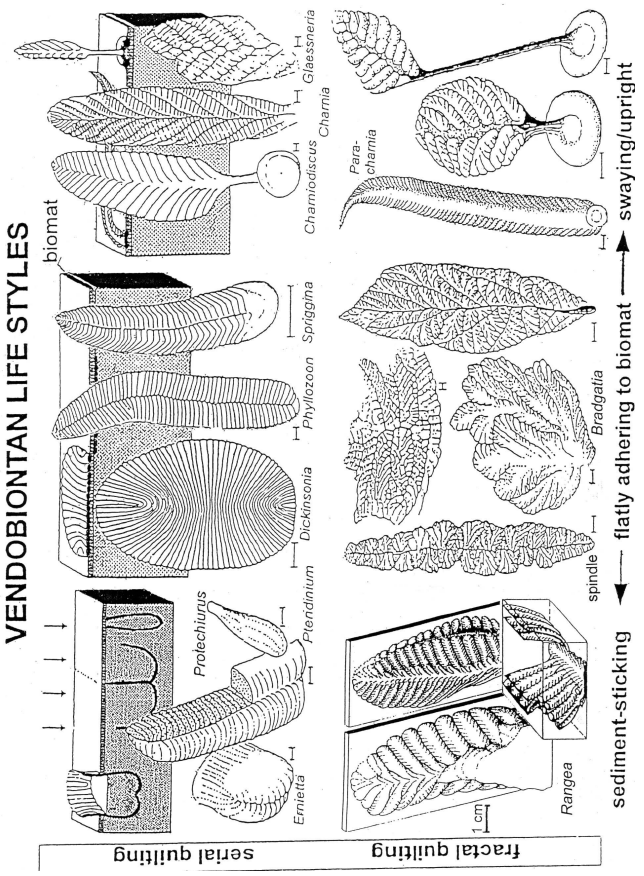


Fig. 10.3. While sharing a quilted foliate construction, vendobionts became large and radiated into a variety of shapes and life styles relative to the ubiquitous Precambrian biomats. Note that Dickinsonia (as Yorgia, Fig. 2) was mobile on tops of the mat, while Phyllozoon probably lived in place below it (Fig. 7). Among the anchored forms, Paracharnia was built for swaying in a current, while a stiff stem allowed other forms to stand upright in quieter waters. Note also that in Rangea the opposite branch of each quilt pair bent alternatingly into one of the two inner vanes and thereby avoided space problems. (Modified from Seilacher, 2003).

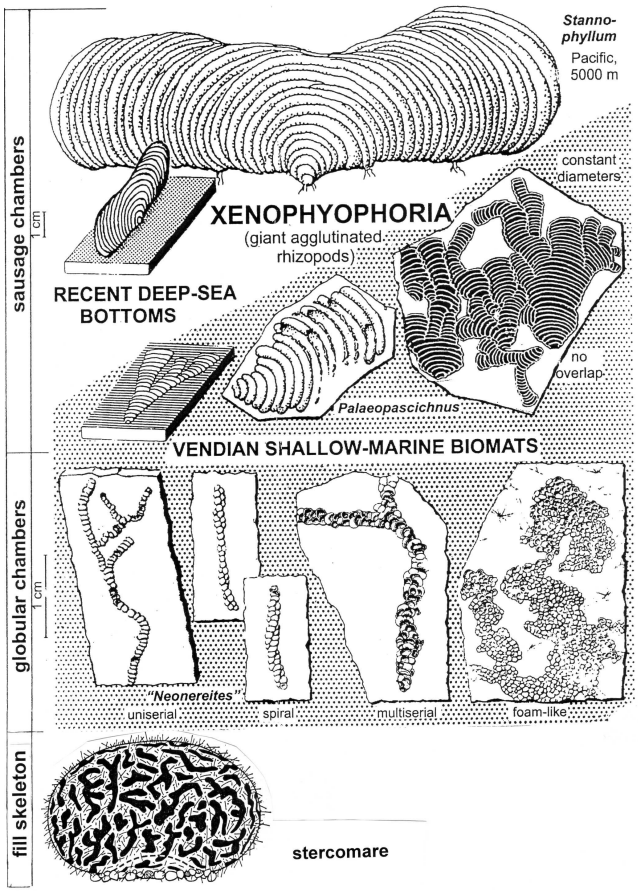


Fig. 10.4. Because they survive in the present deep sea, xenophyophores are known to be giant rhizopods. Their chambers resemble vendobiontan quilts by sizes, shapes, and allometric diameters, but walls are agglutinated and non-expandable. As a silty fill skeleton (stercomare) further subdivides the protoplasm, chambers can be wider than in foraminiferan shells. Shallow-marine Ediacaran representatives, long held for trace fossils, lived probably embedded in biomats. (Modified from Seilacher et al., 2003; enlarged cross section from Tendal, 1972).

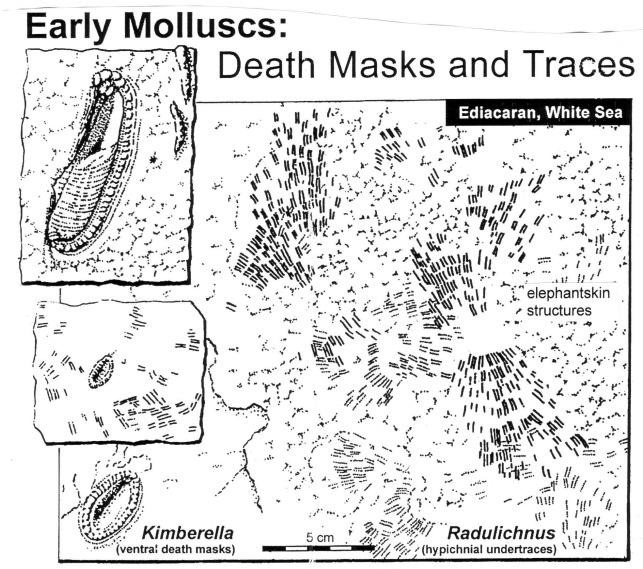


Fig. 10.5. Ventral death masks and associated radula scratches (Radulichnus) identify Kimberella as a stem-group mollusc. Grazing traces of juveniles are not preserved, because they did not reach the sole of the biomat. Kimberella also differed from later chitons and gastropods by grazing with an expandable proboscis in a stationary mode rather than in continuous meanders.(From Seilacher et al., 2003).

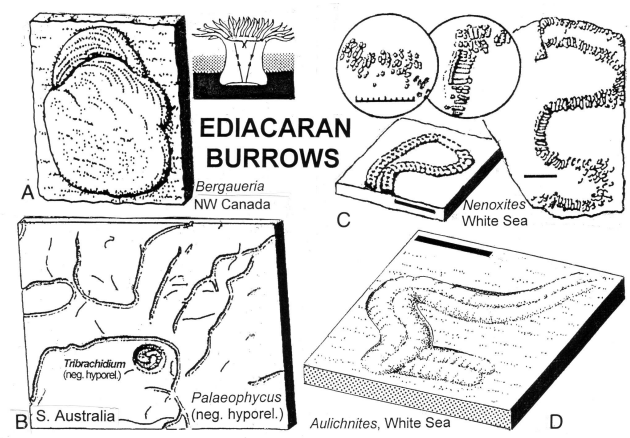


Fig. 10.6. In Ediacaran times, metazoan burrowers did not penetrate more than a few millimeters into the sediment. A. In actinian resting traces (Bergaueria sucta), slight lateral movements are expressed by concentric lines. B. Palaeophycus sp. records the horizontal movement of a worm-like undermat miner that avoided a possible sand sponge (Tribrachidium) . C. The "worm" burrow Nenoxites appears to have been lined with fecal pellets. D. Aulichnites is probably the backstuffed burrow of a slug-like creature. (A.-B. from Seilacher et al., 2003; C-D after Fedonkin).

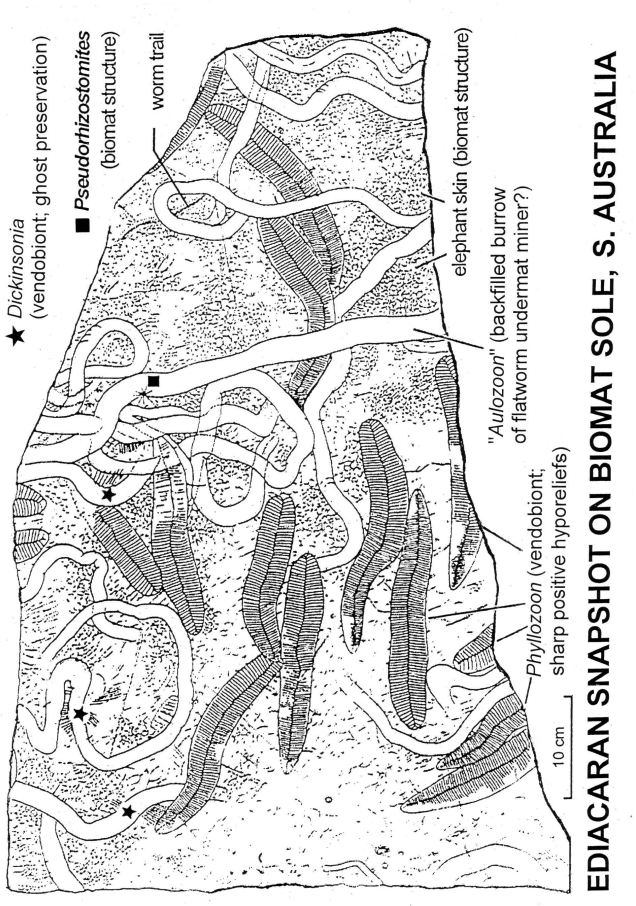


Fig. 10.7. Inverse relief, sharp contours, and the tendency to hug each other with little overcrossing suggest that the uniformily sized individuals of Phyllozoon are preserved in situ (but see Gehling et al., 2005, p.52) below the biomat, while the ghost impressions of Dickinsonia are probably pressed through from above the mat. In contrast, the flat sand sausages ("Aulozoon") intercross themselves, but respect Phyllozoon. (After Seilacher et al., 2003; see Gehling et al, 2005, fig.5 for drawing drawing of complete slab).

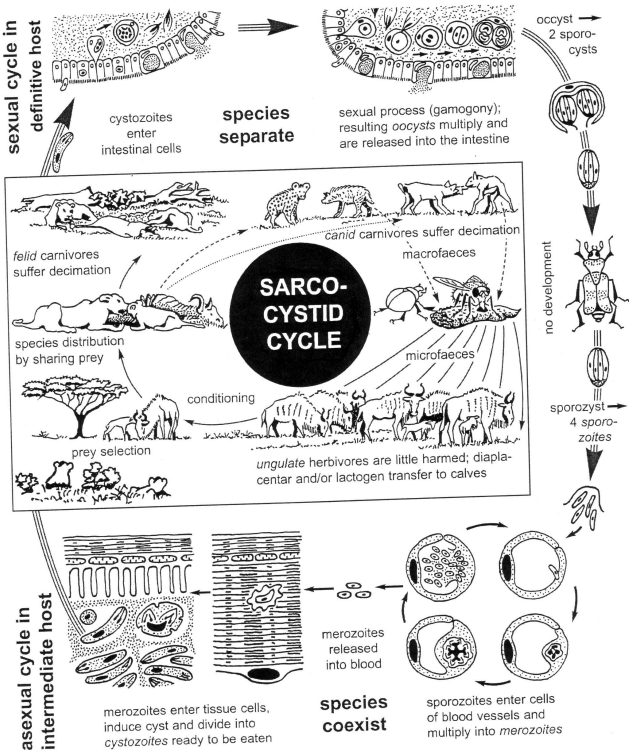


Fig. 10.8. Like other heteroxenic parasites, sarcocystids balance host populations by damaging the stronger carnivores more than the herbivores. (From Seilacher et al., submitted. Drawings by P. Wenk).

11

Title/paper forthcoming...

Jonathon R. Stone
McMaster University

12

The Search for Life on Mars

Chris P. McKay
NASA Ames Research Centre

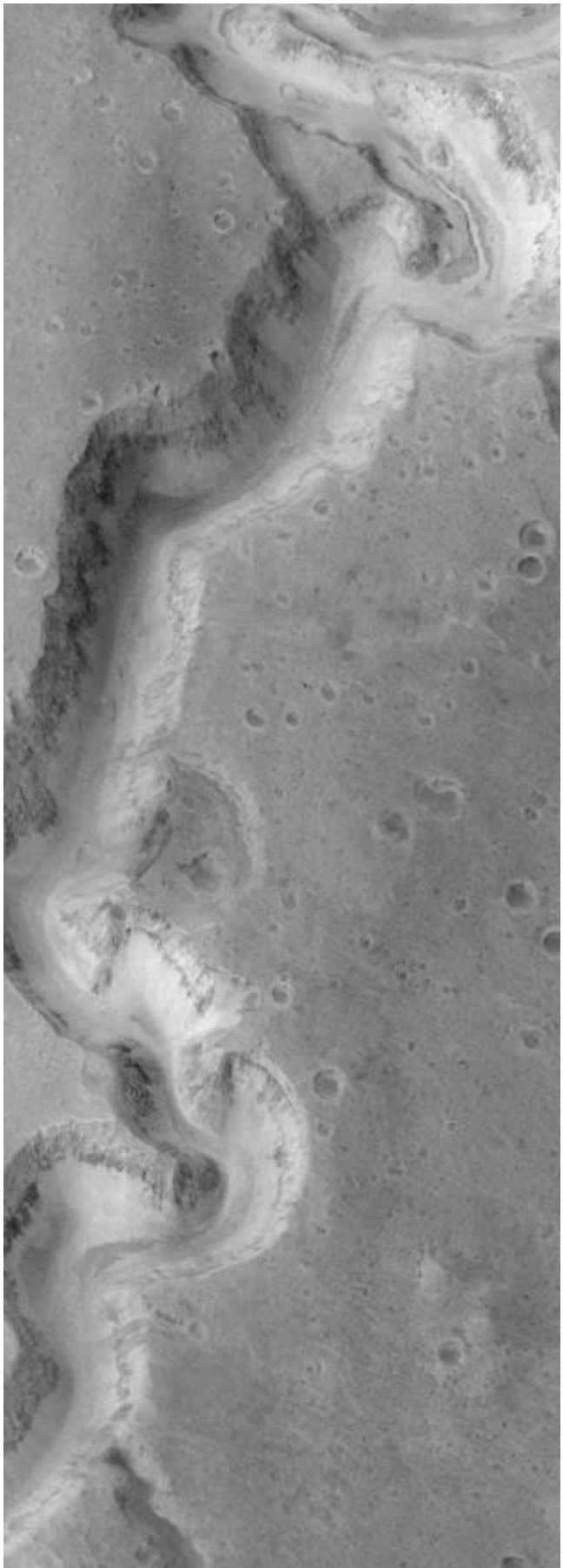
Abstract

The early environment of Mars had liquid water and may have had life. A key question is determining if that life shares a common ancestor with life on Earth. Samples from ancient ice-rich permafrost in the south polar regions of Mars may be the best chance to find biological material that would allow such a determination.

12.1 Introduction

Mars is the world that has generated the most interest in life beyond the Earth. There are three reasons why Mars is the prime target for a search for signs of life. First, there is direct evidence that Mars had liquid water on its surface in the past, and there is the possibility that there is liquid water in the subsurface at the present time. Second Mars has an atmosphere, albeit a thin one, that contains CO_2 and N_2. Third, conditions on Mars are cold and dry and thus are favorable for the preservation of evidence of organic remains of life.

Mars may be cold and dry today but there is compelling evidence that earlier in its history Mars did have liquid water. This evidence comes primarily from the images taken from orbitial spacecraft. Figure 1 from Malin and Carr (1999) shows an image of a canyon on Mars and represents probably the best evidence for extended and repeated, if not continuous, flow of liquid water on Mars. Water is the common ecological requirement for life on Earth. No organisms are known that can grow or reproduce without liquid water. Thus, the evidence that sometime in its early history Mars had liquid water is the primary motivation for the search for evidence of life (McKay, 1997).



The search for life beyond the Earth is one of the main goals of Astrobiology. If life is or was present on Mars it would be important to understand the relationship of Martian life to Earth life. It is possible that Martian life and Earth life are related — part of the same tree of life. This could have resulted either from an exchange of life from one of these worlds to the other via meteorites or by the seeding of both worlds by infalling material carrying life. This later concept, known as panspermia, has been the focus of renewed interest is recent years primarily as an explanation for the origin of life on Earth very soon after the end of the impact bombardment.

It would be more interesting scientifically and philosophically if Martian life were not related to Earth life but represented a second genesis of life (McKay, 1997). This case is interesting scientifically because we would then have, for the first time, an alternative biochemistry to compare with terrestrial biochemistry. In addition the fact that life arose independently twice in our Solar System would be persuasive evidence that life is common in the universe.

Thus the full astrobiological investigation of Mars goes beyond just a search for signs of life. It is also a study of the nature of that life and its genetic relationship to Earth life. To study the nature of Martian life requires that we access biological material on Mars — organism either living or dead. In this paper I discuss the implications of this requirement to access Martian biology and suggest approaches and locations on Mars that may be fruitful in this respect.

12.2 Mars Today and the Viking Search for Life

The surface of Mars today is cold and dry. The surface environmental conditions are summarized in Table 1 and compared to Earth. From the perspective of biology the most important feature of the martian atmosphere is the low pressure. The time and spatial averaged pressure on Mars is about 0.6 kPa (Haberle et al., 2001). The triple point of water is 0.61 kPa and water does not exist as a liquid when the total pressure is below this level. The low pressure on Mars results in the absence of water as liquid on Mars. For pressures slightly above the triple point the liquid is only marginally stable since the boiling point and freezing point nearly coincide. At a surface pressure of 1 kPa water will boil at a temperature of 7°C. Kahn (1985) has shown that if liquid water were present on the surface of Mars it would lose heat rapidly by evaporation, even if it was at 0°C, since it would be close to its boiling point. This loss

of heat would cause the water to freeze since it is so close to the freezing point. The amount of energy input required to maintain the liquid state is larger than the solar constant on Mars for pressures less than about 1kPa and becomes extremely large as the pressure approaches 0.61 kPa.

Parameter	Mars	Earth
Surface pressure	0.5 - 1 kPa	101.3 kPa
Average Temperature	-60°C	+15°C
Temperature Range	-120°C to +25°C	-80°C to +50°C
Composition	95% CO_2	78% N_2
	2.7% N_2	21% O_2
	1.6% Ar	1% Ar
Indident Solar Radiation	149 W m^{-2}	344 W m^{-2}
Surface Gravity	3.73 m s^{-2}	9.80 m s^{-2}
Solar Day	24h 39m 35.238s (1 "sol")	24h
Sidereal Year	687 days, 668.6 sols	365.26 days
Obliquity of axis	25 deg	23.5 deg
Eccentricity	0.0934	0.0167
Mean Distance to Sun	1.52 AU (2.28×10^8km)	1 AU (1.49×10^8 km)

Recent experiments by Sears and Moore (2005) with liquid water exposed to a CO_2 atmosphere at 0.7 kPa and 0°C in a large environmental chamber imply an evaporation rate for water on Mars under these conditions of 0.73 ± 0.14 mm/h — in agreement with the value calculated by assuming that evaporation depends on diffusion and buoyancy. This corresponds to a latent heat flux of 60 W m^{-2}, which is significant compared to the average solar constant at the top of the martian atmosphere of 150 W m^{-2}.

At the Viking 2 landing site (48°N) frost was observed on the water surface in spring (e.g., Hart and Jakosky, 1986). The considerations above suggest that this water would turn to vapor as conditions warmed with little or no transient liquid water phase. However, even a brief transient liquid water film might be important in terms of chemical oxidation reactions even if it is not adequate for biological function.

The only spacecraft to Mars that included a search for life or organic material were the two Viking Landers in 1976. Each lander contained three biology experiments and a device to detect and characterize organic material (a combination gas chromatograph mass spectrometer).

The three Viking biology experiments were: 1) the pyrolytic release experiment (PR) which sought to detect the ability of microorganisms

in the soil to consume CO_2 using light (Horowitz and Hobby, 1977); 2) the gas exchange experiment (GEx) which searched for gases released by microorganisms when organic nutrients added to the soil (Oyama and Berdahl, 1977); and 3) the labeled release experiment (LR) which sought to detect the release of CO_2 from microorganisms when radioactively labeled organic nutrients were added to the sample (Levin and Straat, 1977).

Both the GEx and the LR experiment gave interesting results. When moisture was added to the soil in the GEX experiment, O_2 was released. The release was rapid and occurred when the soil was exposed just to water vapor. The response persisted even if the soil was heated to sterilization levels (Oyama and Berdahl, 1977). The LR experiment indicated the release of CO_2 from the added organics. This response was attenuated with heating and eliminated when the soil temperature was raised to sterilization levels (Levin and Straat, 1977). The release of O_2 in the GEX experiment was not indicative of a biological response. The LR results, however, were precisely what would be expected if microorganisms were present in the martian soil.

Nonetheless, a biological interpretation of the LR results is inconsistent with the results of the GCMS. The GCMS did not detect organics in the soil samples at the level of one part per billion (Biemann et al., 1977; Biemann, 1979). One ppb of organic material would represent more than 10^6 cells existing alone in the soil (Klein, 1978; 1979). However it is not likely that there are microorganisms in the soil on Mars without associated extracellular organic matterial. The lack of detection of organics is the main reason for the prevailing view that non-biological factors were the cause of the reactivity of the martian soil. In the soils of the extreme arid core of the Atacama desert — the most extreme desert location on Earth — even soils with undetectable microbial concentrations have detectable levels of organic material (Navarro-Gonzalez et al. 2003).

It is important to note that all three of the Viking biology experiments were based on providing martian organism conditions and media for growth. That is, they were incubation experiments, similar to culture experiments. It is now known, but was not known at the time of the Viking missions, that culture experiments fail to detect over 90% of soil microorganisms on Earth (e.g., Kirk et al. 2004). The fact that most soils on Earth will grow up in a culture media is due to the incredible diversity of life in those soils not to the robustness of culturing as a way to detect life. We also now know that there are soils on Earth

such as the Atacama Desert, as discussed above (Navarro- Gonzalez et al. 2003), where there are low levels of bacteria but nothing grows in any known culture media. Non-culture dependent methods have been developed since the time of Viking and these have found widespread use in environmental microbiology however these methods are keyed to specific features of terrestrial life, e.g., DNA, ATP, which may not be present in martian life if it represents a second genesis. Nonetheless, future biology experiments on Mars must make use of these culture independent methods.

12.3 Search for Second Genesis

As mentioned above the fundamental question about life on Mars is whether or not it represents a second genesis of life. To determine this we must access intact Martian microbes. Four possible sources have been considered: 1) dormant life at the surface, 2) subsurface ecosystems, 3) organisms preserved in salt (by analogy with salt and amber on Earth), and 4) organisms preserved in permafrost. We consider each of these.

12.3.1 Surface life

There are two possible occurrences of surface life on Mars. The first is the based on the biological interpretation of the Viking Labeled Release Experiment. If the LR results are assumed to be due to biology then dormant martian microbes are widespread in the surface soils of Mars. If the LR results are due to chemical agents as is widely believed then no dormant microbes are present.

A more plausible model for surface life on Mars related to episodic liquid water formed in the north polar regions during favorable orbital conditions. As first pointed out by Murray et al. (1973) the conditions in the polar regions of Mars change dramatically in response to changes in the parameters of Mars' orbit. A recent analysis by Laskar et al. (2002) show that of particular interest over the past 10 Myr are periodic changes in the obliquity and timing of perihelion. Figure 2 adapted from Laskar et al. (2002) show the obliquity, eccentricity, and summer equionox insulation at the north pole over the past 10 Myr. It is useful to consider three epochs in this time history; the last 0.5 Myr, from 0.5 to 6 Myr ago, and 6 to 10 Myr ago.

In the first epoch (the last 0.5 Myr) the obliquity only slightly varies and changes in polar conditions are dominated by the relative phase of

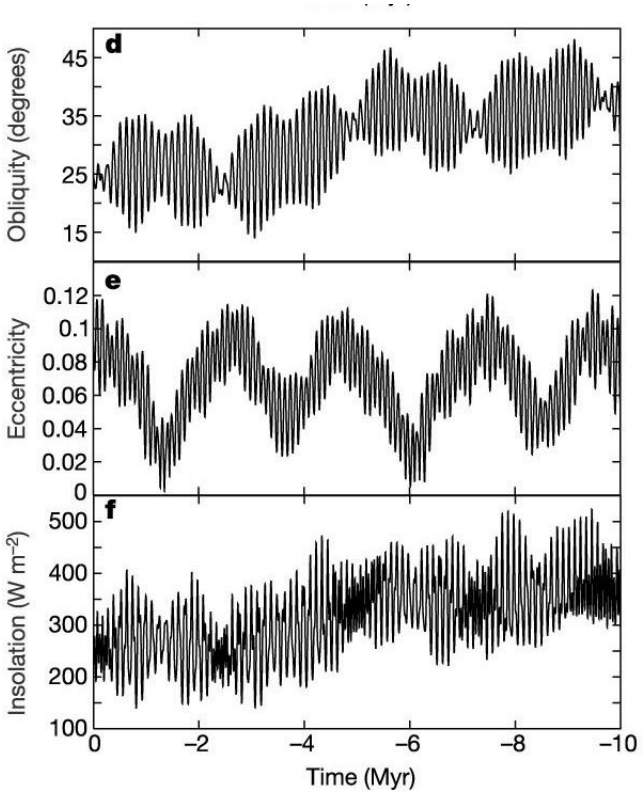


Fig. 12.2. Obliquity cycles and insolation at the North Pole of Mars over the past 10 Myr (adapted from Laskar et al. 2002).

perihelion and equinox. During this epoch, Mars' eccentricity remains high (∼0.1) and therefore the solar flux at perihelion is 49% more than at aphelion. Today perihelion (Ls=251) almost coincides with summer solstice (Ls=270) and the southern summer sun is therefore stronger than the northern summer sun. Procession of the orbit reverses this situation in ∼50,000 years. The effect is not symmetrical due to the fact that the north polar regions are at low elevation higher pressure and therefore as discussed below the formation of liquid water by the melting of ground ice is possible. This is not possible in the southern polar regions which are at much higher elevation. Thus ∼50,000 years

ago conditions in the north polar region of Mars were different than today, the sunlight reaching the polar regions was about 49% stronger.

The second epoch to consider in the orbital history of Mars is the period from about 0.5 Myr ago to 6 Myr ago. Laskar et al. (2002) show that during this period the obliquity of Mars varies considerable over the range 15° to 35° with the average value approximately equal to the value today of 25°. During this epoch there are even larger changes in polar summer sunlight. At the highest obliquity and at high eccentricity, the summer sun can be twice brighter than the value at the present time.

The third epoch to consider begins about 6 Myr ago when the obliquity assumes a larger average value (35°) with excursions as high as 45° (Laskar et al. 2002). During this epoch the maximum summer sun in the north polar regions can be 2.5 times the present value.

Time Period	Summer Insolation	Depth of ice exchange
Martian year	200 W m^{-2}	0.1 m
0.5 Myr	300 W m^{-2}	<0.5 m
0.5 to 6 Myr	400 W m^{-2}	<1 m
6 to 10 Myr	500 W m^{-2}	1 – 2 m

Jakosky et al. (2003) discuss the potential habitability of Mars' polar regions as a function of obliquity. They conclude that temperatures of ice covered by a dust layer can become high enough (-20°C) that liquid brines solution form and microbial activity is possible. Rivkina et al. (2001) have shown that microorganism can function in ice-soil mixtures at temperatures as low as -20°C.

Costard et al. (2002) computed peak temperatures for different obliquities for varying surface properties and slopes. They found that peak temperatures are > 0°C at the highest obliquities, and that temperatures above -20°C occur for an obliquity as low as 45° (Costard et al., 2002). They suggested this as a possible cause of the gullies observed by Malin and Edgett (2000).

Environmental conditions on the surface of Mars today are inauspicious for the survival and growth of even the hardiest terrestrial life forms. The Antarctic cyrptoendolithic microbial ecosystems and snow algae found in alpine and polar snowpacks are probably the best candidates for martian surface life (McKay, 1993). Both ecosystems can grow in environments where the mean air temperatures are below freezing but the temperatures in the substrate (the sandstone rock and the snowpack, respectively) must be at or about melting and liquid water

must be present for growth. There are no plausible models for the growth of these systems anywhere on Mars at the present time.

1. Pressure above triple point (610 Pa)
2. Ice near the surface
3. High insolation during summer

However conditions at the northern polar regions are the closest to habitability in several respects (Table 3). First the low elevation of the northern plains results in atmospheric pressures that are above the triple point of liquid water. Indeed the pressure measured by the Viking 2 lander at 48°N never fell below 0.7 kPa. As Lobitz et al. (2001) and Haberle et al. (2001) show, the northern plains of Mars are the main location on the planet where liquid water could be present and stable against boiling due to low pressure. If surface insolation increased in the northern polar regions the surface ice would melt to form liquid. In contrast in the southern polar regions the warmed ice would sublime due to the low pressure. The second factor that favors habitability in the polar regions is the presence of ice near the surface. The third factor is due to the nature of the polar seasons. While orbital average conditions at the polar regions can be quite cold, the summer sun never sets. Indeed for the range of obliquities considered here the polar regions receive more sunlight per day at solstice than anywhere else on the planet. The polar summer solstice is an energy rich period and can cause strong seasonally dependent melting. This is observed in the polar regions of Earth. The effect is even stronger as the obliquity increases from its present value of 25° to 45°. Although at the present time liquid water in not expected in the northern latitudes on average, Hecht (2002) has shown that under favorable conditions of solar exposure melting of ice can form liquid at the present time. This is consistent with the results of Clow et al. (1987) for the melting of a dusty snowpack. In both cases increased solar heating compared to the present average conditions is required. As the obliquity increases the average conditions are closer to melting and the variations required to produce liquid water are reduced. Thus liquid water should become more plentiful.

12.3.2 Subsurface Life

Although current conditions on Mars suggest there is little chance for life on the surface, there is interest in the possibility of subsurface life on Mars (Boston et al., 1992). Liquid water could be provided by the heat of

geothermal or volcanic activity melting permafrost or other subsurface water sources. Gases from volcanic activity deep in the planet could provide reducing power (as CH_4, H_2, or H_2S) percolating up from below and enabling the development of a microbial community based upon chemolithoautotrophy, especially methanogens that use H_2 and CO_2 in the production of CH_4. Stevens and McKinley (1995) and Chapelle et al (2001) have reported on a microbial ecosystem deep within basaltic sediments on Earth that are based on methanogens and are completely independent of the surface biosphere. In this terrestrial system H_2 comes from weathering reactions between water and basaltic rocks. With a source of hot water, all the ingredients for this subsurface habitat are present on Mars; CO_2 comprises the bulk of the martian atmosphere and basaltic rocks are abundant. Lin et al. (2005) have reported on a subsurface microbial ecosystem also based on methanogens but with the H_2 produced by radioactive decay.

The possibility of subsurface life on Mars today depends on the existence of hydrothermal systems. While, it certainly seems that volcanic activity on Mars has diminished over geological time, carter ages (Hartmann et al., 1999) and the age of the youngest Martion meteorite of 160 Myr ago (ref??) indicates volcanism on Mars as recently as that time. Volcanic activity by itself does not provide a suitable habitat for life — liquid water presumably derived from the melting of ground ice is also required. It is likely that, any volcanic source in the equatorial region would have depleted any initial reservoir of ground ice and there would be no mechanism for renewal. Closer to the poles ground ice is stable (Fanale and Cannon, 1974; Squyres and Carr, 1986; Feldman et al., 2002). It is conceivable that a geothermal heat source could result in cycling of water through the cryosphere (Squyres et al., 1987; Clifford, 1993). The heat source would be melting and drawing in water from any underlying reservoir of groundwater or ice that might exist.

The D/H measurements of water in the SNC meteorites shows that it has an enrichment of D about equal to that in the present martian atmosphere (Watson et al., 1994). Assuming that this enhancement is due atmospheric escape then this similarity suggests that there was an exchange between that atmosphere and the rocks from which the SNC meteorites derived. Probably this exchange involved hydrothermal groundwater systems driven by volcanism or impact events (Gulick and Baker, 1989).

Such hypothetical ecosystems are neither supported, nor excluded, by current observations of Mars. Tests for such a subsurface system involve

locating active geothermal areas associated with ground ice or detecting trace quantities of reduced atmospheric gases that would leak from such a system. The reports of possible CH_4 in the martian atmosphere could be an indication of such subsurface hydrothermal activity and possibly biology (Formisano et al., 2004; Krasnopolsky et al. 2004).

12.3.3 Preserved in Salt

There are reports of organisms preserved in a viable state over geological time in dehydrating substances such as amber and salt. The oldest well-established preservation of life is found in amber. Cano and Borucki (1994) reported (but not as yet independently confirmed) that bacteria can be preserved in amber for 25 million years. However amber is a product of trees and thus is unlikely on Mars. Microorganisms are also found in ancient salt. Vreeland et al. (2000) demonstrated retrieval of organisms from salt that is 250 Myr old. However it is not clear that the organisms are as old as the salt. The main difficulty is because salt, unlike amber, is not impermeable. Small drops of water can migrate through salt in the presence of a temperature gradient leaving the crystal structure of the salt intact with no apparent trace of their movement. Thus, it remains uncertain that the organisms found in the 250 Myr old salt on Earth are not more recent contaminants.

For Mars the issue of contamination is not so critical. If the desire is to obtain a specimen of Maritian life, it does not matter if it is an ancient organism or a geologically recent contamination. Thus the preservation potential of salt deposits should be considered. However, as yet there are no locations on Mars where large salt deposits, similar to salt domes on Earth, are known to exist.

12.3.4 Permafrost

The microbiology of permafrost locations on Earth have been investigated and it has been shown that viable microorganism can be recovered from Siberian permafrost that is ∼3.5 Myr old (Gilichinsky, 1992). New work in Beacon Valley, Antarctic indicates the presence of recoverable microorganisms in ice that is thought to be 8 Myr old.

On Mars there may be extensive permafrost that dates back 3 to 4 Gyr. Recent data from the Mars Oddessy spacecraft have confirmed the suggestion that the polar regions of Mars are rich in ground ice (Feldmann et al., 2002). The south polar regions, but not the polar

cap deposits themselves, are of particular interest because this region contains ancient cratered terrain presumably dating back to the end of the heavy bombardment, 3.8 Gyr ago. The actual polar cap deposits are probably much younger. One region of particular interest, centered on 80°S, 180°W, is shown in Figure 3 from Smith and McKay (2005). Here the terrain is heavily cratered, there is ground ice present and furthermore there is strong crustal magnetism in the surface materials. The presence of strong crustal magnetism confirms the antiquity of these terrains and suggests that they have been relatively unaltered since their initial deposition. This location may represent the site of the oldest, coldest, undisturbed permafrost on Mars. Martian microorganisms may be trapped and preserved in this permafrost (Smith and McKay, 2005) .

12.4 Detecting a Second Genesis on Mars

There are several ways to search for life. First we can be searching for life as a collective general phenomenon. However life might also be a single isolated organism. And that organism might be dead. Finally, signs of life may be fossils, artifacts or other inorganic structures. In the search for life on Mars, any of these would be of interest. Definitions of life typically focus on the nature of the collective phenomenon. In general, such definitions are not useful in an operational search for life on other worlds. The one exception is the proposal by Chao (2000) to modify the Viking Labelled Release (LR) experiment to allow for the detection of organisms that improve their capacity to utilize the provided nutrients. This would in principle provide a direct detection of Darwinian evolution and could unambiguously distinguish between biological metabolism and chemical reaction. Chao (2000) argues that Darwinian evolution is the fundamental property of life and other observables associated with life result from evolutionary selection. His method for searching for evolution would be practical if the right medium can be selected to promote the growth of alien microbes. Unfortunately, we now know that only a fraction of microorganisms from an environmental sample grow in culture.

Of course growth experiments of any kind do not detect dead organisms. Yet the remains of dead organisms are potentially important evidence of life on another planet. And so are fossils. However, there is an important distinction between dead organisms and fossils. A fossil is evidence of past life but it does not reveal anything about the bio-

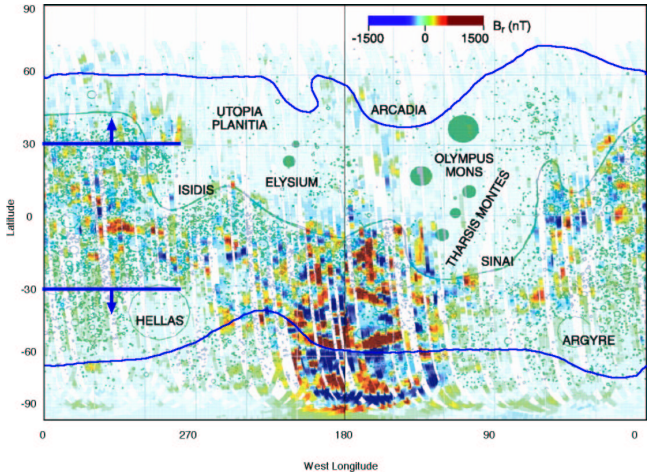


Fig. 12.3. Crustal magnetism, crater distribution and ground ice on Mars. Each green dot represents a crater with diameter greater than 15 km. The boundary between the smooth northern plains and the cratered southern highlands is shown with a green line. The crustal magnetism is shown as as red for postitive and blue for negative. Full scale is 1500 nT. The typical strength of Earth's magnetic field at the surface is 50,000 nT. The solid blue lines show the extent of near surface ground ice as determined by Odyssey mission. Ground ice is present near the surface polarward of these lines. Crater morphology indicates deep ground ice poleward of 30o (Squyres and Carr, 1986), shown here by dark blue lines and arrows. The region between 60 and 80oS at 180oW is heavily cratered, preserves crustal magnetism, and has ground ice present. This is our suggested target site for drilling. This figure is adapted from Acuña et al. (1999), based on the crater distribution in Barlow (1997). The distribution of near surface ground ice is from Feldman et al. (2002). Figure from Smith and McKay (2005).

chemical or genetic nature of that life. If we are searching for a second example of life then we need to be able to compare the nature of that life to Earth life. For this an organism is needed, either dead or alive, but a fossil is not sufficient.

As discussed above, a promising target for the search for biological remains of past life is focused on the subsurface. The deep permafrost on Mars may hold remnants of past life (Smith and McKay, 2005). The organisms in the ground ice are likely to be dead from accumulated radiation dose but their organic remains could be analyzed and compared to the biochemistry of Earth life.

I have argued previously (McKay, 2004) that one way to determine if a

collection of organic material is of biological origin, is to look for a selective pattern of organic molecules similar to, but not necessarily identical with, the selective pattern of biochemistry in life on Earth. Pace (2001) has argued that life everywhere will be life as we know it. He contends that the biochemical system used by life on Earth is the optimal one and therefore evolutionary pressure will cause life everywhere to adopt this same biochemical system. It is instructive to consider this argument in the context of a conceptual organic phase space. If we imagine all possible organic molecules as the dimensions of a phase space, then any possible arrangement of organic molecules is a point in that phase space. We can define biochemistries as those points in phase space that allow for life. The biochemistry of Earth life — life as we know it — represents one point in the organic phase space. We know that this one point represents a viable biochemistry. Pace's (2001) contention that biochemistry is universal is equivalent to stating that in the region of phase space of all possible biochemistries there is only one optimum biochemistry and thus any initial set of biochemical reactions comprising a system of living organisms will move toward that optimum as a result of selective pressure. If Pace (2001) is correct then the only variation between life forms that we can expect is that associated with chirality. As far as is known, the left and right forms of chiral organic molecules (such as amino acids and sugars) have no differences in their biochemical function. Life is possible that is exactly similar in all biochemical respects to life on Earth except that it has right instead of left amino acids in its proteins and left instead of right sugars in its polysaccharides.

The question of the number of possible biochemistries consistent with life is an empirical one and can only be answered by observations of other life forms on other worlds, or by the construction of other life forms in the laboratory. The observation or construction of even one radically alien life form would suffice to show that biochemistry as we know it is not universal.

Life as Lego: The pattern of biochemistry of Earth life follows what I have called the "Lego principle" (McKay, 2004). This is the unremarkable observation that life on Earth uses a small set of molecules to construct the diverse structures that it needs. This is similar to the children's play blocks known as Legos in which a few different units repeated over and over again are used to construct complex structures. The biological polymers that construct life on Earth are the proteins, the nucleic acids, and the polysaccharides. These are built from repeated units of the 20 left-handed amino acids, the 5 nucleotide bases, and the right-

handed sugars. The use of only certain basic molecules allows life to be more efficient and selective. Evolutionary selection on life anywhere is likely to result in the same selective use of a restricted set of organic molecules. As discussed above, I believe it is premature to conclude that all life anywhere will use the same set of basic biomolecules. Thus I suggest that life will always use some basic set but it may not be the same basic set used by life on Earth. This characteristic biogenic pattern of organic molecules would persist even after the organism is dead. Given our present state of understanding of biochemistry, we are not able to propose alternative and different biochemical systems that could be the basis for life, but that may reflect a failure of our understanding and imagination rather than a restriction on the possibilities for alien life.

A sample from the deep permafrost in the southern hemisphere of Mars could be analyzed for organic material with a fairly simple detection system. If organic material was detected then it would be of interest to characterize any patterns in that organic material that would indicate a "Lego principle" pattern. Clearly one such pattern is the identical pattern of all Earth life; 20 amino acids, the five nucleotide bases, A, T, C, G, and U etc. However, more interesting would be a clear pattern different from the pattern known from Earth life. Figure 4 shows a schematic diagram of how a biological pattern would be different from a non-biological pattern.

Implementing this search in practical terms in near term missions will require a sophisticated ability to separate and characterize organic molecules. Currently the instrument best suited for this task is a GCMS with solvent extraction. However, new methods of fluorescence and Raman spectroscopy could provide similar information and may have a role in future mission applications.

12.5 Conclusions

The search for evidence of past life on Mars is motivated by the direct evidence that Mars once had liquid water. Studies of life on Earth strongly indicate liquid water as the essential requirement for life. Mars presents a challenge and an opportunity. The challenge is to explore a distant planet with a complex history first with robotic probes and eventually with human explorers. The opportunity is to learn about the nature of life, to search for a possible second type of life in our own solar system and thereby begin to understand the profound philosophical and

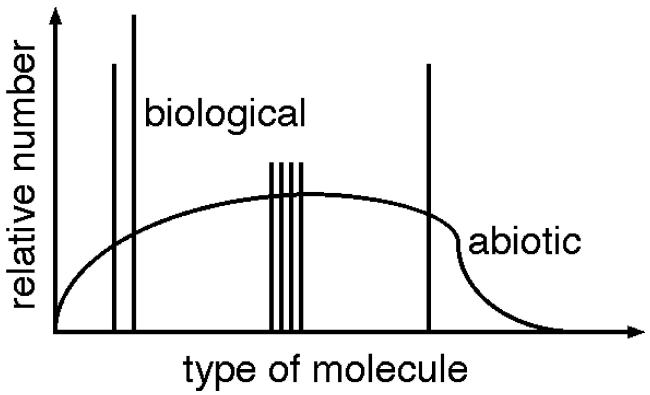


Fig. 12.4. Comparison of Biogenic with Nonbiogenic Distributions of Organic Material Nonbiological processes produce smooth distributions of organic material, illustrated here by the curve. Biology, in contrast selects and uses only a few distinct molecules, shown here as spikes (e.g., the 20 left handed amino acids on Earth). Analysis of a sample of organic material from Mars or Europa may indicate a biological origin if it shows such selectivity. Figure from McKay (2004).

scientific issues related to life in the universe. It is a search worth the best efforts of the planetary science and space engineering community.

Acuña, M. H., J. E. P. Connerney, N. F. Ness, R. P. Lin, D. Mitchell, C. W. Carlson, J. McFadden, K. A. Anderson, H. Reme, C. Mazelle, D. Vignes, P. Wasilewski, and P. Cloutier, Global Distribution of Crustal Magnetism Discovered by the Mars Global Surveyor MAG/ER Experiment, Science, 284, 790-793, 1999.

Barlow, N. 1997. Mars: impact craters, In: Shirley, J.H., Fairbridge, R.W. (Eds.), Encyclopedia of Planetary Sciences, Chapman and Hall, London.

Biemann, K. 1979. The implications and limitations of the findings of the Viking Organic Analysis Experiment, J. Mole. Evol 14, 65-70.

Biemann, K., J. Oro, P. Toulmin, III, L.E. Orgel, A.O. Nier, D.M. Anderson, P.G. Simmonds, D. Flory, A.V. Diaz, D.R. Rushneck, J.E. Biller, and A.L. LaFleur 1977. The search for organic substances and inorganic volatile compounds in the surface of Mars. J. Geophs. Res. 82 4641-4658.

Boston, P.J., M.V. Ivanov, and C.P. McKay, On the possibility of chemosynthetic ecosystems in subsurface habitats on Mars, Icarus, 95, 300-308, 1992.

Cano R.J., and M.K. Borucki, Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science 268, 1060-1064, 1995.

Chao, L. (2000) The Meaning of Life. Bioscience 50, 245-250.

Chapelle, F.H., K. O'Neill, P.M. Bradley, B.A. Methe, S.A. Ciufo, L.L. Knobel, and D.R. Lovley, A hydrogen-based subsurface microbial community dominated by methanogens, Nature 415, 312-315, 2002.

Clifford, S.M., A model for the hydrologic and climatic behavior of water on Mars. J. Geophys. Res., 98, 10,973-11,016, 1993.

Clow, G. D. 1987. Generation of liquid water on Mars through the melting of a dusty snowpack, Icarus 72, 95-127.

Costard, F., F. Forget, N. Mangold, J.P. Peulvast (2002) Formation of Recent Martian Debris Flows by Melting of Near-Surface Ground Ice at High Obliquity. Science, 295: 110-113.

Fanale, F.P. and W.A. Cannon, Exchange of adsorbed H2O and CO2 between the regolith and atmosphere of Mars caused by changes in surface insolation, J. Geophys. Res. 79, 3397-3402, 1974.

Feldman, W.C., W.V. Boynton, R.L. Tokar, T.H. Prettyman, O. Gasnault, S.W. Squyres, R.C. Elphic, D.J. Lawrence, S.L. Lawson, S. Maurice, G.W. McKinney, K.R. Moore, R.C. Reedy, Global distribution of neutrons from Mars: Results from Mars Odyssey, Science, 297, 75-78, 2002.

Formisano, V., S. Atreya, T. Encrenaz, N. Ignatiev, and M. Guiranna, Detection of methane in the atmosphere of Mars, Science 306, 1758-1761, 2004.

Gilichinsky, D.A., E.A. Vorobyova, L.G. Erokhina, D.G. Fyordorov-Dayvdov, and N.R. Chaikovskaya, Long-term preservation of microbial ecosystems in permafrost, Adv. Space Res., 12(4), 255-263, 1992.

Gulick, V. C., and V. R. Baker 1989. Fluvial valleys and martian paleoclimates, Nature 341, 514-516.

Haberle, R.M., C.P. McKay, J. Schaeffer, N.A. Cabrol, E.A. Grin, A.P. Zent, and R. Quinn, On the possibility of liquid water on present-day Mars, J. Geophys. Res., 106, 23317-23326, 2001.

Hart, H., and Jakosky, B.M. 1986. Composition and stability of the condensate observed at the Viking Lander 2 site on Mars. Icarus 66:134-142.

Hartman, W.K, M. Malin, A. McEwen, M. Carr, L. Soberblom, P. Thomas, E. Danielson, P James, and J. Veverka (1999) Evidence for recent volcanism on Mars from crater counts. Nature 397, 586-589.

Hecht, M.H., Metastability of liquid water on Mars, Icarus 156, 373-386, 2002.

Horowitz, N.H. and G.L. Hobby (1977). Viking on Mars: The carbon assimilation experiments. J. Geophys. Res. 82, 4659-4662.

Jakosky B.M., K.H. Nealson, C. Bakermans, R.E. Ley and M.T. Mellon 2003. Subfreezing Activity of Microorganisms and the Potential Habitability of Mars' Polar Regions. Astrobiology 3, 343-350.

Kahn, R., The evolution of CO_2 on Mars, Icarus 62, 175-190, 1985.

Kirk Kirk, J.L., Beaudette, L.A., Hart, M., Moutoglis, P., Klironomos, J.N., Lee, H., Trevors, J.T. 2004, "Methods of studying soil microbial diversity", Journal of Microbiological Methods, 58, 169-188.

Klein, H. P., The Viking biological experiments on Mars, Icarus 34, 666-674, 1978.

Klein, H.P. 1979. The Viking Mission and the search for life on Mars, Rev. Geophys. and Space Phys. 17, 1655-1662.

Krasnopolsky, V. A., J. P. Maillard, and T. C. Owen, Detection of methane in the martian atmosphere: evidence for life?, Icarus 172, 537-547, 2004.

Laskar, J., Levrard, B., Mustard, J. 2002. Orbital forcing of the martian polar Orbital forcing of the martian polar deposits. Nature, 419, 375-377.

Levin, G.V. and P.A. Straat (1977). Recent results from the Viking Labeled Release Experiment on Mars. J. Geophys. Res. 82, 4663-4667.

Lobitz, B., B.L. Wood, M.A. Averner, and C.P. McKay, Use of spacecraft data to derive regions on Mars where liquid water would be stable. Proceed. Nat. Acad. Sci., 98, 2132-2137, 2001.

Lin, L.-H., J. A. Hall, J. Lippmann, J. A. Ward, B. Sherwood Lollar, and T. C. Onstott. 2005. Radiolytic H2 in the continental crust: Nuclear power for deep subsurface microbial communities. Geochem. Geophys. Geosys. 6:Q07003. [Online.] doi:10.1029/2004GC000907.

Malin, M.C. and M.H. Carr 1999, Groundwater formation of martian valleys. Nature 397, 560-561.

Malin, M.C. and K.S. Edgett Sedimentary Rocks of Early Mars, 2000. Science, 290, 1927-1937.

McKay, C.P., Relevance of Antarctic microbial ecosystems to exobiology. in "Antarctic Microbiology", ed. E. Imre Friedmann, Wiley-Liss, New York, 593-601, 1993.

McKay, C.P., The search for life on Mars, Origins Life Evol. Biosph., 27, 263-289, 1997.

McKay, C.P., The search for a second genesis of life in our Solar System,

in First Steps in the Origin of Life in the Universe, ed. J. Chela-Flores, pp 269-277, 2001.

McKay, C.P. (2004) What is life and how do we search for it on other worlds. PLoS Biol. 2, 1260-1263.

Murray,B.C., Ward, W.R. and Yeung,S.C. 1973. Periodic insolation variations on Mars. Science 180, 638-640.

Navarro-Gonzalez, R., F.A. Rainey, P. Molina, D.R. Bagaley, B.J. Hollen, J. de la Rosa, A.M. Small, R.C. Quinn, F.J. Grunthaner, L. Ceceres, B. Gomez-Silva, and C.P. McKay. Mars-like soils in the Atacama Desert, Chile and the dry limit of microbial life, Science, 302, 1018-1021, 2003.

Oyama, V.I. and B.J. Berdahl (1977). The Viking gas exchange experiment results from Chryse and Utopia surface samples. J. Geophys. Res. 82, 4669-4676.

Pace N., The universal nature of biochemistry. Proc. Natl. Acad. Sci. USA 98, 805- 808, 2001.

Rivkina, E.M., E.I. Friedmann, C.P. McKay, and D.A. Gilichinsky (2000) Metabolic activity of permafrost bacteria below the freezing point, Appl. Environ. Microbio. 66, 3230-3233.

Sears, D.W.G. and S.R. Moore 2005. On laboratory simulation and the evaporation rate of water on Mars. Geophys. Res. Lett. 32, L16202, doi:10.1029/2005GL023443.

Smith, H.D. and McKay, C.P. (2005) Drilling in ancient permafrost on Mars for evidence of a second genesis of life. Planet. Space Sci., 53, 1302-1308.

Squyres, S.W., D.E. Wilhelms, and A.C. Moosman, Large-scale volcano-ground ice interactions on Mars, Icarus, 70, 385-408, 1987.

Squyres, S.W., and M.H. Carr 1986. Geomorphic evidence for the distribution of ground ice on Mars. Science 231, 249-252.

Stevens, T.O. and J.P. McKinley, Lithoautotrophic microbial ecosystems in deep basalt aquifers, Science 270, 450-454, 1995.

Vreeland, R.H., W.D. Rosenzweig and D.W. Powers, Isolation of a 250 million year old bacterium from primary salt crystals, Nature 407, 897-900, 2000.

Watson, L.L., I.D. Hutcheon, S. Epstein, and E.M. Stolper, Water on Mars: Clues from deuterium/hydrogen and water contents of hydrous phases in SNC meteorites. Science, 265, 86-90, 1994.

13

Life in the Dark Dune Spots of Mars: a testable hypothesis

Eörs Szathmary & Tibor Ganti
Collegium Budapest

Tamas Pocs
Eszterházy Károly College

Andras Horvath
Konkoly Observatory

Akos Kereszturi, Szaniszlo Berzci & Andras Sik
Eötvös University

13.1 Introduction

The aim of this chapter is to present one of the very rare exobiological hypotheses. The main thesis is that there could be life in the Dark Dune Spots (DDSs) of the Southern polar region of Mars, between -60 and -80 degrees latitude. The spots have a characteristic annual morphological cycle and it is at least suspected that liquid water forms in them every year. We propose that a consortium of simple organisms (similar to bacteria) comes to life each year, driven by sunlight absorbed by the photosynthetic members of the consortium. A crucial feature of the proposed habitat is that life processes take place only under the cover of water ice/frost/snow. By the time this frost disappears from the dunes, the putative microbes, named Mars Surface Organisms (MSOs) must revert to a dormant state. The hypothesis has been worked out in considerable detail, it has not been convincingly refuted so far, and it is certainly testable by available scientific methods. We survey some of the history, the logical thread, the testable predictions of, and the main challenges to the DDS-MSO hypothesis.

13.2 History

The spots in question were observed on images made by the Mars Orbiter Camera (MOC) onboard the Mars Global Surveyor (MGS) spacecraft

between 1998 and 1999 (images are credited to NASA/JPL/Malin Space Science Systems). These features appear in the Southern and Northern polar regions of the planet in the spring, ranging in diameter from a few dozen to a few hundred meters. Malin and Edgett published their first observation on the Internet in 1999, and then in print (Malin & Edgett, 2000). One of us (A.H.) began to analyze the images in the summer of 2000, based on images from the South that were clear and freely downloadable from the Internet ($http://www.msss.com/moc\_gallery/$). It became clear that one group of these spots were strictly localized to dark dunes (DDs), clearly distinguishable from the usual rusty terrain of Mars Thus we coined the term Dark Dune Spots (DDSs; Fig. 13.1). Nobody can deny that the DDSs are striking and that they call for an explanation. As we shall see, the first explanation (Edgett & Malin, 2000), based on simple frosting and defrosting of the dry ice (carbon dioxide) cover simply does not work. Noting that the majority of the spots are circular, the suspicion arose that some biological activity may also be involved in spot formation. Indeed, without giving the scale and the source, several people thought they were looking at an image of a Petri dish with some bacterial culture. However, one cannot base a hypothesis on an analogy that is so much out of place and scale. Nevertheless, years of work have led to a detailed hypothesis involving biological phenomena that is consistent with all of the observed features. Conversely, we are not aware of any abiotic hypothesis that could explain the full set of observations. This of course does not mean that the biological hypothesis is right. Nevertheless we feel encouraged by the fact that since 2001, when the biological hypothesis was published (Horváth et al. 2001) all new observations and data have made the hypothesis more plausible rather than to the contrary. The main point is that the hypothesis is testable and we are sure to know the answer in the foreseeable future, if astrobiological activity continues.

Perhaps the most important intellectual precedent of our hypothesis is the suggestion by Lynn Rothschild (1995), putting forward cryptic photosynthetic microbial mats as a potential Martian way of life. These mats can be found on the Earth, and the community is under permanent dust cover. She pointed out that such a cover would be very important on Mars as protection from an aggressive environment. Although we have missed this important precedent, now we happily incorporate it into the full hypothesis. Needless to say, some think that proposing extant life on Mars is an extraordinary claim, requiring extraordinary evidence. We shall come back to this issue in the Discussion. First, the

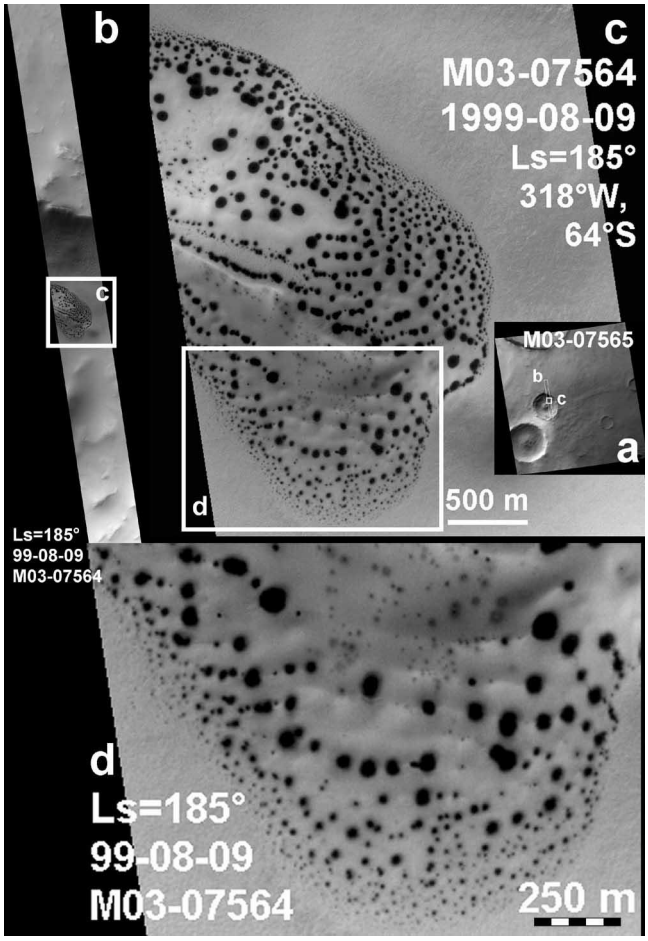


Fig. 13.1. Dark Dune Spots on a Southern dune covered with white frost, photographed by the Mars Orbiter Camera onboard MGS at Ls=185 (1999.08.09., image no. M03-07564, M03-07565). Subset images: a) the observed crater, b) the MOC image strip , c) the dark dune d) the magnified image of the spots with high resolution

claim may not be so extraordinary after all, but this rank is in the eye of the beholder anyway. Second, we do not have extraordinary evidence yet, but we see clearly how such evidence can be obtained.

13.3 Basic facts and considerations about DDSs

Two kinds of frost/snow cover can be distinguished in the polar region: the permanent ice cap and the seasonal cover. Several spot-like features, categorized as spots, fans, blotches and halos (Christensen et al. 2005) can be distinguished, none of which is fully explained. DDSs can be rather well told apart from the rest by the following features: 1. low albedo, i.e. they are darker than the surrounding frost cover; 2. they occur in a belt in the Southern polar region between 60 and 80 degrees latitude; 3. they are located predominantly on dark dunes inside craters; 4. they have an internal structure consisting of a dark core and surrounding lighter ring; 5. their diameter is between 5 and 200 m; 6. they show up at the end of local winter, they grow in size and disappear with the frost (seasons on Mars are indicated with solar longitude (Ls)); 7. they reappear annually at the same sites with a frequency between 50 and 65% (Fig. 13.2).

As mentioned above, the first explanation given for DDS formation was defrosting (sublimation) of dry ice (Malin & Edgett, 2000). As we shall see below, this explanation is far from sufficient, but of course sublimation plays a subsidiary role since the dry ice cover is removed by this process. Kieffer (2003) gave a more involved explanation, according to which carbon dioxide gas is formed between the soil and the dry ice due to the absorption of sunlight, which then erupts to the surface. As a result the frost diminishes, and the gas vent carries with itself fine dust that falls back on the top of the frost. We agree that this is good candidate explanation for the formation of the "fans" that have only a limited similarity to the DDSs. Fans are fan-shaped spots that predominantly show up between the DDSs-fields and the permanent ice cap. As we shall see from the fact presented below, this explanation does not seem to be valid for the DDSs. In contrast we suggest that liquid water is an indispensable component of DDS formation.

Throughout the years (Gánti et al. 2003) we have established the following facts, strongly favoring the role of liquid water:

(i) Spots start developing between the ice cover and the soil and they continue to do so until the ice cover disappears in the summer. Sublimation of dry ice occurs only at the surface; hence it cannot explain processes beyond the ice.

(ii) Spots appear only on the dark dunes, quite often exactly marking the edge of the dune field. This implies that somehow the dune material affects the formation, development, and maintenance of

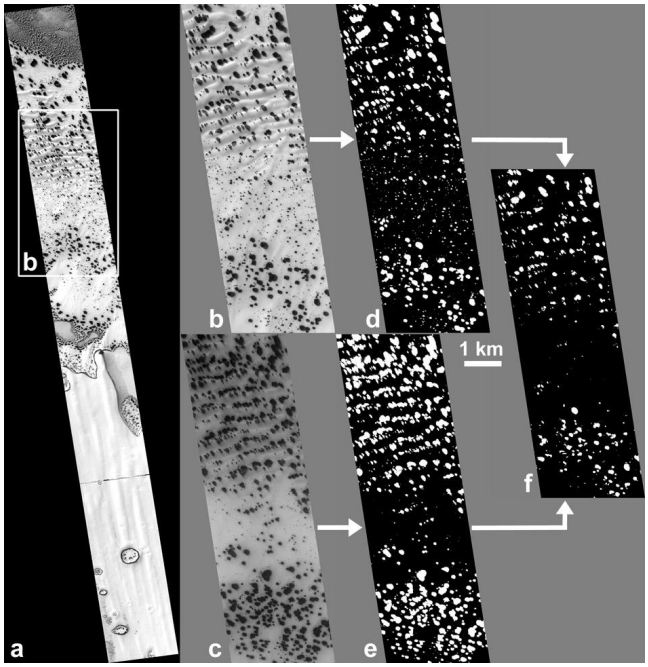


Fig. 13.2. Recurrence of Dark Dune Spots in 1999 and 2001. a) an MGS MOC narrow-angle image (M07-02775) of the study area; b) part of this area in 1999; c) the same area as on b) one Martian year later in 2001 (E07-00101); d) and e) show only the DDSs with white color were present in the images b) and c); while in f) the overlapping area of DDSs occured both in 1999 and 2001 is visible

the spots. Formation of the frost can of course be influenced by surface texture and temperature, but sublimation necessarily proceeds at the surface of the ice. Of course the color of the soil surface can affect sublimation by heat absorption and conduction, but, in agreement with localization and spatial recurrence of the spots, this effect should be additional to the mere dark surface of the dunes.

(iii) On horizontal planes the spots are practically circular (Figs 13.1, 13.7), implying some isotropic formative cause. Sublimation depends on the degree of insolation, the wind currents, etc. Hence the location and shape of the spots would be much more erratic if caused by sublimation only. In contrast, the almost perfect cir-

cular shape, the localization, and the growth of the spots imply a mechanism spreading radially from a centre. This observation is consistent with the in situ formation of liquid water.

(iv) Importantly on slopes the spots become elongated and assume an ellipsoid shape, of which the big axis is always parallel with the gradient (Fig. 13.3). Though slope winds are supposed to be frequent on Mars (Magalhaes & Gierasch, 1982; Lee et al., 1982), we observed that there are cases where the diffuse fan formed by the wind does not overlap with the darker part of the flow-like structures originating from the DDSs (these seepages will be discussed below). This further suggests that some fluid phase is causal to the formation of the spots: gravitation cannot effect sublimation on this spatial scale. Under the given environmental conditions this could be liquid water (Reiss & Jaumann, 2002).

(v) On steeper slopes "flows" (called seepages) originate from the elongated spots, and point downwards. This by itself implies that some liquid phase is moving downhill (Fig. 13.4).

(vi) The soil of the dunes is covered by white frost over the winter. By early summer this cover sublimates and the dark material of the dune is exposed. However, some light grey spots can be clearly discerned against this dark surface on the summer images. The distribution of summer grey spots coincides with that of the earlier DDSs (Fig. 13.5). This grey remnant on the surface is again evidence against a mere sublimation explanation.

(vii) Finally, the fact that the spots appear late winter and develop in the spring implies that their formation needs sunlight. In summary, DDS formation below the ice cover is triggered by the sun.

These facts and arguments clearly show that sublimation alone or gas vents are an insufficient explanation. Especially the evidence for liquid phase calls for a different mechanism. What may cause the spots then? What allows the formation of water below the frost so that it remains liquid for months?

We think that our DDS-MSO hypothesis is a good candidate. We assume that in the past Mars was much more hospitable to life than now (Squires et al. 2004), and a biota did build up on the planet. Drastically changing surface conditions have wiped out most part of this ancient living world, but some creatures could have survived provided they adapted to harsh, but still annually recurrent living conditions. Between two favorable periods, these creatures survived by evolved strategies of drying

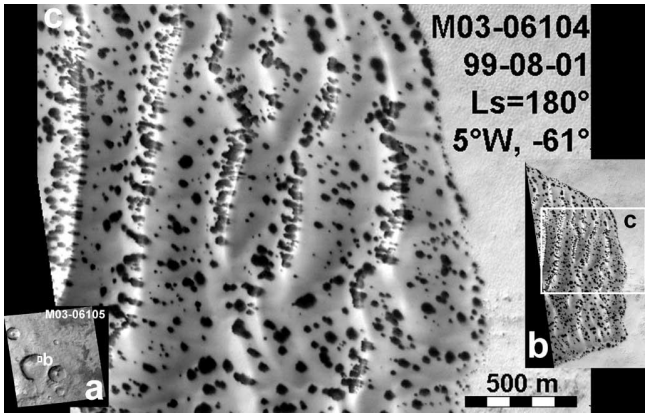


Fig. 13.3. Elongated DDSs: a) the crater, b) the edge of the dunefield, c) an enlarged part of the dunefield. The individual dunes follow a nearly vertical stripe pattern. The Sun illuminates from the left and as a result the slopes tilted to the left are brighter while the opposite side is darker. The slopes facing left are also substantially steeper, and as a result of the higher tilt several spots are elongated there. The darker slopes facing right are longer and milder. Their tilt is so low that the spots located on them are nearly circular rather than elongated. (M03-06104, M03-06105, 1999.08.01., Ls=180.3°, 5.35°W 60.77°S)

out and freezing. If indeed such organisms survived, there was a strong selection pressure to evolve photosynthetic pigments with very efficient light harvesting. Such organisms could then reactivate themselves each year provided they are sheltered by a layer of water frost/ice/snow above them. MSOs could thus melt the water ice around them, and provide the liquid water for themselves.

Thus the hypothetical lifecycle of the MSOs can be visualized (Fig. 13.6) as follows:

(i) In the winter MSOs lay dormant: there are no active life processes. A layer of (according to laser MOLA measurements 0.2 to 1 m thick, Aharonson et al. 2004) frost covers them. Above ground and the dormant MSOs there is a layer of water snow/ice, above this there is a thicker layer of dry ice.

(ii) Late winter spot formation begins by the absorption of photosynthetic MSOs. Cells get reactivated and a liquid water lens is formed about the center of the colony. This central region appears as a grey spot on the images (Fig. 6). To be sure, sublimation on

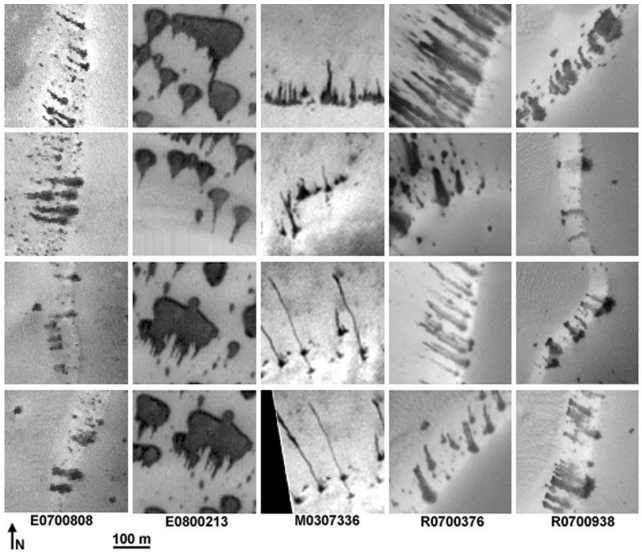


Fig. 13.4. Examples for seepage structures originating from DDSs. Images are 300x300 meter sized, north is up. Parameters of the original images for each columns: E07-00808, 69.17°S 150.80°W, 2001-08-13, Ls=213.51°; E08-00213, 70.63°S 16.99°W, 2001-09-03, Ls=226.63°; M03-07336, 69.49°S 17.41°W, 1999-08-07, Ls=183.82°; R07-00376, 69.18°S 150.98°W, 2003-07-06, Ls=216.29°; R07-00938, 69.18°S 150.81°W, 2003-07-13, Ls=220.74°.

the surface of the dry ice is also facilitated from heat produced at the bottom. It is important that a thin layer of water vapor is expected to form between the liquid water and the water ice, decreasing the heat conductivity and help to maintain the liquid phase.

(iii) The center of spots loses the water/carbon dioxide cover first. This will appear as the black core on the images, surrounded by a still radially extending grey ring. We propose that in the center the organisms have already reverted to their dormant state, whereas in the grey ring they are still active.

(iv) Finally, by mid-summer all the frost is gone and all MSOs form a dormant surface/subsurface colony that is visible as the summer grey spot.

If the proposed lifecycle is valid, it naturally explains the observed phenomena detailed above. It becomes natural why the spots are initi-

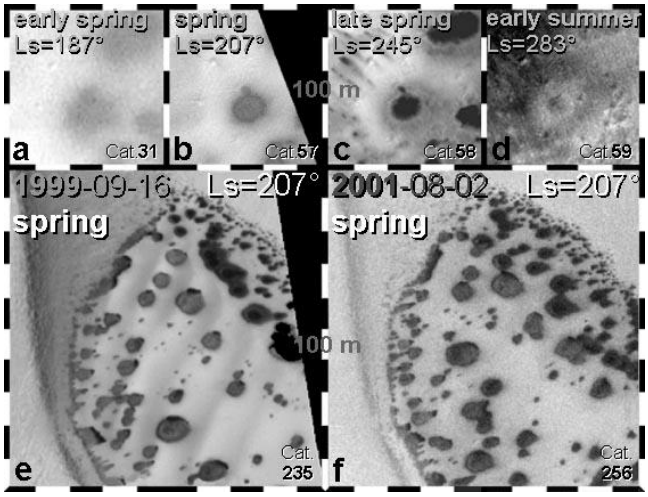


Fig. 13.5. Seasonal changes of DDSs from early spring to early summer according to the progress of seasons. a) Ls=187° (early spring), M04-00678; b) Ls=207°, M07-02824; c) Ls=245° (middle spring), M09-03813; d) Ls=283°, (early summer) M11-02076. e) and f) subset images show the annual recurrence of DDSs on the same dune field from two different years: e) Ls=207°, 1999.09.16., M07-02775; f) Ls=207°, 2001.08.02., E07-00101

ated below rather than above the frost. The formation of liquid water explains the pattern of growth on flat surfaces and slopes, and it also explains the seepages that originate from DDSs on steep slopes. It then becomes natural, as Clow (1997) suggested, that liquid water will flow each year below the water ice cover. This would be impossible with a bare surface or one covered with dry ice only.

A natural outcome of the hypothesis is that MSOs can live only where there is at least a seasonal water frost cover which is thick enough for protection against cold, evaporation and UV irradiation, yet it is thin enough for photosynthesis. Closer to the Pole it will become colder, the frost will be thicker and insolation will be less favorable.

A fact in favor of our hypothesis is that the dunes are the first to catch frost in the autumn and the last to be become defrosted in the summer. This extends the period during which a water ice/snow cover is present. This is, we believe, is an explanation why the DDSs stick to the dunes: the dune surface by virtue of its texture attracts water frost early in the autumn. The fact that surfaces catch frost with varying efficiency is well

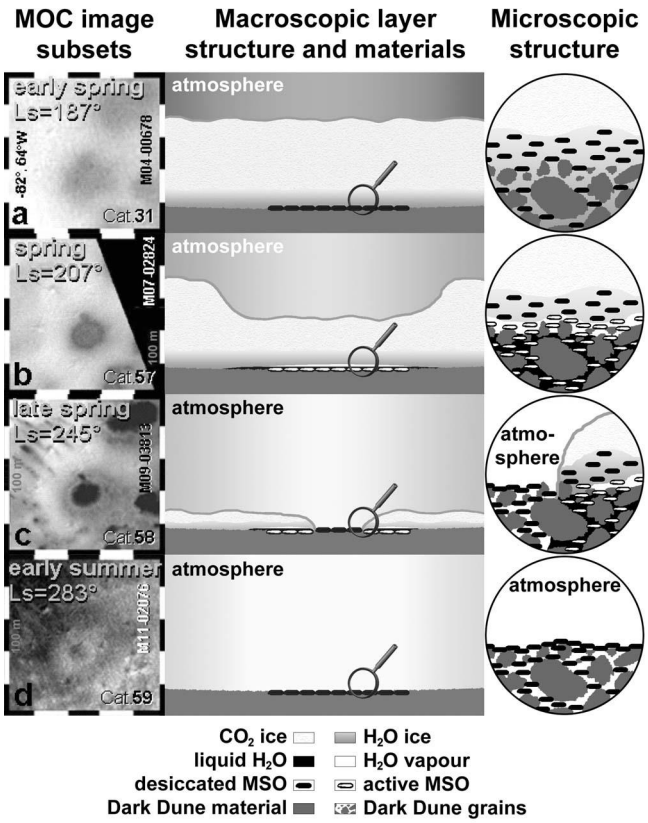


Fig. 13.6. Hypothetical life-cycle of Mars Surface Organism in Dark Dune Spots, MOC photographs (left, M04-00678, M07-02824, M09-03813, M11-02076), hypothetical cross section (middle, right) during a) early spring, b) spring, c) late spring, d) summer

known here on Earth: vegetation, for example, is a particularly efficient frost attractor.

A further consideration favoring dunes over the red regolith may have to do with their different chemical composition. Red color comes from oxidized iron, predominantly hematite (Fe_2O_3). After the Viking missions it has become clear that UV and hematite together lead to the formation of aggressive oxidants (e.g. Möhlmann, 2004), which destroy organic matter efficiently. The color of the dunes is dark blue/dark vi-

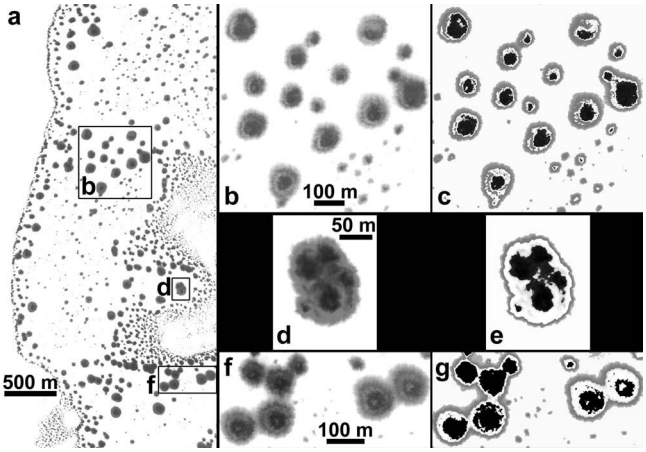


Fig. 13.7. Size and inner structure of Dark Dune Spots on a Martian intracrater dune field at 65°S and 15.55°W with enhanced contrast. The left image is an MGS MOC narrow-angle photo (M07-00853) acquired on 1999-09-05, Ls=200.6°. Subset images show the structural units of DDSs

olet, and they are assumed to consist of basaltic sand (Herkenhoff & Vasavada, 1999). If so, then the dune surface will be a milder environment for life than the usual red surface.

Finally, the DDS hypothesis can explain the fine structure of the spots that is observed in some of the best images (Fig. 13.7). There is a very thin whitish ring between the black center and the grey ring. If the model in Fig. 13.6 holds, then there is liquid water under the grey ring, which will leak towards the black center. But there is no protecting ice shield any more, so water will quickly evaporate, and it will also cool quickly. Thus some re-frosting of liquid water is expected where the black interior and the grey ring meet.

This hypothesis may be considered nice but for the moment it looks very qualitative. The devil is hiding in the details. Therefore, in the next section we consider the challenges to the hypothesis.

13.4 Challenges and answers

There are three major challenges to the hypothesis: (1) if there is photosynthesis, what is the hydrogen donor? If it is oxygen, where is it? (2) How is it possible to resist the harsh UV irradiation on the surface? (3)

It may just be too cold even under the frost. We consider these problems in turn.

13.4.1 The nature of photosynthesis and the microbial consortium

Central to our hypothesis is the assumption that primary productivity of this ecosystem rests on photosynthesis. Since water plays a central role also, it is natural to assume that the source of reducing power is the photolysis of water. Such a process happens in cyanobacteria, for example, and oxygen is produced as a by-product. Where is the oxygen then? This objection could be very serious, since according to Lovelock (1979), the best astronomical sign for life is an atmosphere out of equilibrium. On this basis some would say our hypothesis cannot hold. But this is not a valid objection, for a number of reasons. (i) We are dealing not what people call 'abundant life'; rather, we are postulating life in 'pockets' (refugia), meaning that the present biomass is tiny. (ii) The Martian atmosphere is out of equilibrium. There is some oxygen (Krasnopolsky et al. 1996.) and some methane (Krasnopolsky et al. 2004; Formisano et al. 2004) also. The source of the latter is unknown, and the former is constantly being produced by the UV lysis of water. This fact leads to the conclusion that there must be a great oxygen sink in the Martian surface (Kolb et al. 2002.). Thus the small amount of oxygen produced by such a small biota could remain unnoticed. (iii) We are postulating not only photosynthetic, but also heterotrophic microbes. Thus oxygen could easily be recycled by respiration. The reason oxygen was allowed to accumulate in the terrestrial atmosphere is that a large amount of organic material was buried by tectonic processes. Without such a process oxygen would not have accumulated here either. Since there is no plate tectonics on Mars (Carr, 1981), and there is the aforementioned large oxygen sink, it becomes understandable why the concentration of oxygen in the air is so low.

It may just be that the presence of methane can also be explained at least in part by the action of some MSOs. It is known that in some hydrothermal systems on Earth photosynthetic organisms and methanogens occur close to each other, where methane is the end product of decomposition of organic matter from the main algal-bacterial mat (Ward, 1978). DDs are not a hydrothermal system, however. The finding of Tung et al. (2005) offers another solution: they found methanogenic archaebacteria at a depth of 3 km below the Greenland ice sheet that op-

erate at a survival metabolism level of -9°C. The same authors calculated that on Mars a habitat of around 0°C could host enough methanogens to account for the amount of methane if a 10-m-thick layer existed with a density of around 1 cell/ml. An alternative could be production in the spots. The rate of loss from the Martian atmosphere is 270 tons per year (Krasnopolsky et al. 2004), which must be offset by production. If there are 10 million active spots, then each must produce 27 g methane per year, and with an average area of 2500 m^2 per spot, this amounts to 10.8 mg/m^2/year of production, a modest value. Kral et al. (2004) showed that methanogens can grow on a Mars soil simulant when supplied with carbon dioxide, molecular hydrogen, and varying amounts of water. They say that "Currently, the surface of Mars is probably too cold, too dry, and too oxidizing for life, as we know it, to exist" (p. 615). If the DDS-MSO hypothesis is correct, this does not hold. Our new methanogenic MSO hypothesis rests on the assumption that the structure of the mat allows the activity of photosynthesizers and methanogens as well, and the source of hydrogen would be organic material from primary production.

13.4.2 The menace of UV irradiation

As mentioned above, the surface of the dunes may be a more hospitable environment because of the reduced level of aggressive oxidizing chemical species. Yet the UV at wavelengths shorter than 290 nm is a serious challenge. This radiation without a special protective mechanism is sterilizing for terrestrial organisms. Various strategies exist for microbes to escape from this threat: (1) protection by the ice cover; (2) protection by screening compounds and by dead cells; (3) protection by a surface layer of sand/dust; and (4) escape into the soil when appropriate. We consider these options in turn.

(i) Protection by ice layers. This case has thoroughly been analyzed by Córdoba-Jabonero et al. (2005); in this discussion we closely follow their treatment. Fortunately, the high latitude by itself reduces surface radiation levels due to lower solar zenith angles. Too much light is bad because of UV and too little light is bad for photosynthesis. Based on terrestrial organisms, Córdoba-Jabonero et al. (2005) have defined a Radiative Habitable Zone (RHZ) and considered its applicability to the Martian polar environments. Although ionizing radiation also reaches the surface, tolerance of

this hazard can be taken from granted. It was concluded that the CO2 ice cover alone is an insufficient shield against UV radiation, since its thickness does not exceed 1 m. (Their model is valid only for CO2 ice and not for CO2 "snow". Unfortunately the microphysical structure of the condensed CO2 is unknown, just like the probable difference between the UV-shielding capacity of CO2 snow, relative to CO2 ice.) In contrast, a few centimeters of water snow would be sufficient to provide protection to the MSOs at a level that would be tolerable for terrestrial microbes also. The snowy condition appears crucial: if we replace it by water ice, then an ice sheet about 1 m thick is required. Sadly, we do not know anything about the nature or thickness of water frost precipitating on the dunes in autumn: altimetry provides the thickness of all layers together. Whereas a water frost sheet of several cm to at most 1 or 2 dm is not excluded, other means to reduce the UV hazard must be presented.

(ii) UV protection by pigments is an obvious solution. On the Earth cyanobacteria have special pigments (like scytonemin and gloeocapsin) in their mucilaginous sheath that shield the cells from UV (Garcia-Pichel & Bebout, 1996; Garcia-Pichel & Castenholz, 1994; Garcia-Pichel, Sherry & Castenholz, 1992). In addition, the dead cells can leave this sheath behind, which remains there as a passive protection for the other cells. The ratio of light-harvesting and light-screening pigments is regulated as a function of PAR, UV and temperature in the mat-forming cyanobacterium Phormidium murrayi (Roos & Vincent, 1998). It would strongly support our hypothesis, if the French operated OMEGA reflectance spectrometer could detect UV screening pigments in the area of DDSes freshly opened up from the snow/ice cover.

(iii) Also, it is conceivable that if we look at the case more closely we are dealing with sub-surface organisms, albeit at the mm scale. Lynn Rothschild (1995) analyzed an interesting ecosystem called a "cryptic" microbial mat. Photosynthetic organisms live there under the protection of sand and gravel. The color of the algae becomes visible if one scrapes away the surface sand. As Rothschild emphasized, this lifestyle is ideal for protection against UV and any surface oxidants, thus it offers a good model for a possible Martial lifestyle. Experiments on dried monolayers of the desiccation-tolerant, cyanobacterium *Chroococcidiopsis* sp.

029 showed that even 1 mm of Martian-like soil provides efficient protection against a simulated UV flux (Cockell et al. 2005).

(iv) The case of the terrestrial analogues is presented in the next section; here it suffices to say that one type of the so-called cryptobiotic crust consists of cyanobacteria that take shelter under the surface under harsh conditions and return to it when conditions improve. This lifestyle is a combination between a surface-bound classical microbial mat lifestyle and the near-surface cryptic mat lifestyle.

We believe that these mechanisms taken together are more than enough to yield the sufficient protection mechanism against UV radiation. Now we turn to the challenge of temperature.

13.4.3 The plausibility of liquid water and the challenge of low temperature

Our hypothesis is qualitative at present, so a physical model of heat balance will be mandatory in the future. The main question is whether liquid water can indeed form in the DDSs, as suggested by the morphology of their development; and whether temperatures permit a *sustainable population dynamics* of the mat. The latter is by no means an automatic consequence of the former.

Surface temperatures above the frost depend on the season. By summer they can rise above 0 °C (Supulver et al. 2001; Reiss & Jaumann, 2002) which is obviously permissive, but there is a problem late winter and early spring, when they can be around -60 and -70 °C. Is it credible that liquid water can form below the frost, where the insulating layer is at most 1 m of water and dry ice altogether? The surface of dunes and, if we are correct, the MSOs absorb the sunlight. Temperatures are bound to rise and the solid state greenhouse effect (Matson & Brown, 1989) helps the process. The temperature at which liquid water appears depends strongly on the salinity of the soil. If the dunes undergo yearly wetting and drying, as we suggest, then the surface must be salty. Melting temperatures of salty solutions can decrease to below -50 °C depending on their composition (Mellon & Phillips, 2001; Kargel & Marion, 2004). Of course we know nothing about the degree of salinity on the dunes, and in any case we consider it unlikely that the effect could be so strong as to keep a macroscopic body of fluid water liquid at -50 °C. In contrast, an effect down to -30 °C would be more realistic

and could help the proposed lifecycle. Intense microbial activity was discovered in the Discovery deep hypersaline basin of the Mediterranean Sea where the water is almost saturated (5 M) with MgCl2 (van der Wielen et al. 2005).

It is important to note that for the initiation of the lifecycle a macroscopic body of liquid water is not necessary. We know that Antarctic lichens start photosynthesizing at -18 °C below the snow because enough molecules of water leave the surface of the snow cavity and enter the cells at this low temperature (Kappen et al. 1996). Also, liquid water at the micro- and nanoscale is markedly different from macroscopic water. Thin films of liquid water exist well below subzero temperatures and can be in contact with individual soil grains and cells. Adsorption water can exist on Mars in liquid form down to at least -40 °C (Möhlmann, 2004, 2005). Price (2000) discusses how organisms continue to have access to water and nutrients even well below the freezing point of pure water, by utilizing aqueous veins at triple junctions of ice grains. So, MSOs may not ever be in a completely desiccated medium.

It is not unlikely that photosynthetic organisms can actively contribute to heating by their light harvesting apparatus. The light phase of photosynthesis is not dependent on temperature. Thus at low temperatures more light is absorbed than can be consumed by the dark photosynthetic reactions (including carbon fixation). Photons may thus be converted to heat by photosystem II in cyanobacteria (Morgan-Kiss et al. 2006). We predict that in MSOs analogous processes have been selected as powerful adaptations to heat up in late winter.

We conclude that the appearance of a sufficient quantity of liquid water to kick-start metabolism late winter in the MSOs is not unrealistic, but we also acknowledge that the issue is very quantitative and proper modeling is needed to settle the issue in the future. For the sake of the argument in the sequel we assume that liquid water in some available form is not a problem.

Price and Sowers (2004) presented a thoughtful analysis of temperature dependence of microbial metabolism. They differentiate between growth (through reproduction), maintenance (active metabolism and maintenance of ion levels, etc. with no net growth), and mere survival (when macromolecular damage is repaired only). For the same organism decreasing temperature pushes the cell in the subsequent domain. Low temperature has of course several adverse effects. A general consequence is the slow-down of metabolism due to Arrhenius' law: other things being equal, rates of chemical reactions decrease exponentially

with decreasing temperature (Vincent, 1998). There is an advantage from this to psychrophilic organisms that is commonly not realized: the rates of adverse reactions also go down with temperature. Therefore the most serious type of effect is one that does not decrease, or even increases with decreasing temperature. Membrane properties, formation of ice crystals and radiation damage belong to this category. Generally, recurrent freezing-thawing cycles are expected to be very adverse. They apply to MSOs in the first part of their growth period.

Regarding metabolism Price and Sowers (2004) conclude, based on experimental evidence and inference, that there is "no evidence of a threshold or cutoff metabolic rate at temperatures down to -40°C" (p. 4636). This is very important because (if they exist) MSOs must be at least as good as terrestrial organisms. The boundary temperatures between survival, maintenance and growth are a matter of species-specific adaptations.

Liquid water and energy input should allow metabolism but low temperatures are not favorable for the net growth of the mat, as analyzed by Friedmann et al. (1993). Long-term productivity of cryptoendolithic microbial communities in Antarctica is very low (net ecosystem productivity is as low as 3 mg C/m2/year). However, one part of this may be the stress of frequent dehydration-rehydration in a desert that would not apply to MSOs.

This is clearly an issue where detailed physical modeling and adequate Mars chamber experiments must be carried out in the future.

13.5 Partial analogues on Earth

It is good if an exobiological hypothesis can be supported by evidence from analogous organisms and lifecycles from Earth. Yet any such analogy is bound to be partial, considering the differences in past and present environmental conditions of the different planets and the differences in the paths that evolution may have taken on the two planets. Clearly, a dark dune type of environment is nowhere to be found on Earth. Extremophiles from various extreme habitats serve as a worst-case approach to Martian hypothetical organisms like the MSOs. They present a worst case in the sense that if organisms have evolved and are still alive on Mars, they are expected to have better adapted to Martian conditions than terrestrial organisms.

A possible objection to this argument is that terrestrial extremophiles may have discovered the specific boundaries of life using nucleic acids

and proteins as informational molecules. If the two living worlds are related (which is not unlikely) then terrestrial extremophiles almost exactly delineate what would be possible on Mars in terms of occupying extreme habitats. This counter-argument misses one important point about evolution, however. In the case of Earth these extreme habitats are rare and the population numbers of the occupying organisms are low. In other words, only a small fraction of the total living population has been exposed to this extreme form of directional selection. In contrast, on Mars environmental deterioration of living conditions has been widespread. If we accept that life did exist on Mars more than 3 billion years ago, then it could have attained population numbers similar to those on the ancient Earth. It follows that when conditions deteriorated, a large population was exposed to directional selection. Favorable mutations are more common in a large than in a small population (see Maynard Smith, 1998 for background reading on evolutionary genetics). Consequently, extremophiles can be expected to be more 'extreme' in their tolerance on Mars than on our home planet, at least in certain ecological dimensions. Even in case of the earth, would anybody have guessed that an organism like Deinococcus (Englander et al. 2004; Cox & Battista, 2005) could exist? We should expect many more such striking cases on Mars if there is life there at all.

Another factor is time. Nadeau et al. (2001) determined the phylogenetic relations of psychrophilic oscillatorians (cyanobacteria). They diverged from their relatives about 20 million years ago when Antarctica cooled down. Most other psychrophiles are much younger than this (Morgan-Kiss et al. 2006). In contrast, Martian organisms must have been under directional selection for billions of years.

This general argument does not tell us anything about the real tolerance of Martian organisms, however. All the calculations on UV, salinity and temperature tolerance in the previous sections have been based on terrestrial organisms since we do not have anything better at hand. Bearing this in mind we now turn to some partial terrestrial analogues to the hypothetical MSOs.

Salt tolerance is important because of the supposed salinity and temperatures of the DDs. There are some deep hypersaline basins of the Eastern Mediterranean Sea, where a rich biota is found in almost saturated salt solution (van der Wielen et al. 2005). Note that life was thought impossible under such conditions before. Recently in the Bannock basin a very rich prokaryotic community was discovered in the oxic-anoxic transition zone, where a chemocline with varying conditions

is found (Daffonchio et al. 2006). A much steeper chemocline could exist in the upper few centimeters of the DDs.

Perhaps it is the cyanobacteria that come closest to the MSOs in their properties. They are everywhere in extreme habitats, including hot and cold deserts. Apparently they can survive for tens of thousands of years in the permafrost (Vishnivetskaya et al., 2001). In northeastern Siberia the oldest viable cells date back to 2-3 million years (Gilichinsky et al. 1995). Several species are multiple extremophiles, for example desiccation and UV tolerant of the same time. The expression of the protein WspA in Nostoc is modulated by UV irradiation and desiccation, and it modulates the three-dimensional extracellular matrix by binding to the UV-absorbing pigment complexes mycosporine and scytonemin (Wright et al. 2005).

A special case of microbial mats is the so-called cryptobiotic crust or biological soil crust (Belnap & Lange, 2003). They are found in arid environments on the soil or rock surface. Remarkably, their assimilation intensity rivals that of higher plants in the same region. Some cyanobacteria are extreme halophiles as well (e.g. Garcia-Pichel, Nubel & Muyzer, 1998). Large masses of them can occur in soda lakes of saturated sodium carbonate solution, such as African Magadi Lake in the Rift Valley or in the shotts of Sahara desert. This is important, because as explained above, we can expect considerable salt deposition on the dark dunes that (according to our hypothesis) undergo a regular wetting-drying cycle. Efficient Na^+/H^+ antiporters in the membrane help the life of the halotolerant (up to 3.0 M NaCl) cyanobacterium Aphanothece halophytica (Wutipraditkul et al. 2005).

Chroococcidiopis is not only desiccation and UV tolerant, but somewhat similar to Deinococcus, it is also able to repair extensive DNA damage following ionizing radiation, which ability is linked to its desiccation tolerance (Billi et al. 2000). In contrast, the genome of Nostoc commune is protected against oxidation and damage by a different strategy after decades of desiccation, aided by the non-reducing disaccharide trehalose (Shirkey et al. 2003).

We find remarkable survival strategies in various forms of the cryptobiotic crust (Pócs et al. 2004). According to our own observations we find pigment-rich species in the upper layer of deserts, which shield the less UV-tolerant species below them. An even more exciting case is that of the Microcoleus species, which during the dry period leave their mucilaginous sheath and glide down in the soil a few millimeters. There

they develop a new sheath and with the advent of the rainy season they glide back to the surface again.

The main objection to these examples is that ice and a prolonged low temperature do not play a role in them. For such a comparison we must look for different analogues. Rivkina et al. (2000) analyzed metabolism of permafrost bacteria below the freezing point. They show that these (non-photosynthetic) bacteria are active down to -18 °C, where their minimum doubling time is ca. 160 days. Moreover, metabolic activity is directly proportional to the thickness of the liquid, adsorbed water layer.

A remarkable case of permanently ice-enclosed consortium of partly photosynthetic bacteria was described by Priscu et al. (1998). Organisms become active under sunlight in water inclusions around aeolian-derived sediments; photosynthesis, nitrogen fixation, and decomposition occur in parallel. The main driving process is photosynthesis at a depth of a few meters. MSOs would be better off because they would be in direct contact with underlying soil. Under the ice cover (about 5 m thick) there is liquid water where a psychrophilic phytoplankton can be found, living at less than 1% of normal PAR (photosynthetically active radiation; Morgan-Kiss et al. 2006).

13.6 Discussion and outlook

According to astronomer the late Carl Sagan, "Extraordinary claims require extraordinary evidence". The DDS-MSO hypothesis is sometimes refuted on the grounds that it is an extraordinary claim without extraordinary evidence. This objection may not hold, for a number of reasons.

First, 'extraordinariness' is in the eye of the beholder. In the particular case we often find that in the end the opponents think that life itself is unlikely, and they do not really believe in extraterrestrial life, past or present. Since we lack an accepted scenario for the origin of life, this stance is legitimate but it is good to be aware of it. Second, the claim is not that we take it as fact that there is life on the Martian dunes, but we present a testable hypothesis for a phenomenon. We agree that in a way confirmatory evidence would be extraordinary. What could such evidence be?

Of course the best would be to go there and have a look. This presents substantial financial and engineering challenges, but ultimately-should other evidence prompt it-it must be done. The second best option is

to receive a weird spectroscopic signal from the DDSs. Spectrometry should operate in a broad spectrum and with good spatial resolution. This not yet granted but there are ideas for example to identify scytonemin, chlorophyll, phycocyanin based on autofluoresce (Schoen & Dickensheets, 2000) with spectroscopic instruments (Wynn-Williams et al. 2002), or with the combination of in-situ Fourier transform micro-, near-infrared and a visible spectrometers (Hand et al. 2005), or in situ analysis of hopanoids as long-lived bacterial cell wall products and photosynthetic pigments (Ellery & Wynn-Williams, 2003).

The most recent American mission of Mars Reconnaisance Orbiter could provide us with very suggestive optical imagery at a resolution of up to 25 cm with HiRSE and 6.55 nm/channel spectral resolution between 370 and 3920 nm with CRISM (Bergstrom, Delamere & McEwen, 2004). Although in biology morphological evidence is often misleading, new images could help a lot in guiding future research.

There is the possibility of building appropriate chambers and test various elements of the DDS-MSO hypothesis in them. Despite ambitious attempts, we have to say that the chambers known to us are inappropriate for this purpose: something (either the UV, or the temperature, or the frost, or the gas) is always missing. We encourage well-equipped laboratories to test our hypothesis.

Finally, a shrewd objection to the DDS story is the following. If the hypothesis assumes that we are looking at the refugial remnants of a flourishing past Martian biota, then it is very surprising that they remained alive just long enough so as to allow for us to find them. This is indeed somewhat striking, but we know other similar examples. Geographical, linguistic and genetic distances still show very good correlations on the Earth and they have been the subject of exciting studies (e.g. Cavalli-Sforza, 1997). It has also been pointed out that they have been analyzed in the 'last hour' at many sites, because increased mobility will soon undermine the pattern. "Chance favors the prepared mind" (Louis Pasteur).

Acknowledgements

This work was supported by the ESA ECS-project No. 98004.

References

[402] Aharonson, O., Zuber, M.T., Smith, D.E., Neumann, G.A., Feldman & W.C., Prettyman, T.H. (2004). Depth, distribution, and density of CO2 deposition on Mars. *J. Gephys. Res.* 109, E5, CiteID E05004.

[403] Belnap, J. & Lange, O.L. (2003). *Biological soil crusts: Structure, function, and management.* Revised 2nd printing. Springer, Berlin Heidelberg, 503.

[404] Bergstrom, J.W., Delamere, W.A. & McEwen, A. (2004). MRO high resolution imaging science experiment (HiRISE): instrument test, calibration and operating constraints, *55th International Astronautic Federation Congress* IAC-04-Q.3.b.02

[405] Billi, D., Friedmann, E.I., Hofer, K.G., Caiola, M.G. & Ocampo-Friedmann, R. (2000). Ionizing-radiation resistance in the desiccation-tolerant cyanobacterium *Chroococcidiopsis*. *Appl. Env. Microbiol.* 66, 1489-1492.

[406] Cavalli-Sforza, L.L. (1997). Genes, peoples, and languages. *Proc. Natl Acad. Sci. USA* 94, 7719-7724.

[518] Carr, M.H. (1981). *The Surface of Mars.* Yale University Press, New Haven.

[408] Christensen, P.R., Kieffer, H.H. & Titus, T.N. (2005). Infrared and Visible Observations of South Polar Spots and Fans. *American Geophysical Union*, Fall Meeting 2005, #P23C-04.

[] Clow, G.D. (1987). Generation of liquid water on Mars through the melting of a dusty snowpack. *Icarus* 72, 95-127.

[409] Cockell, C. S., Schuerger, C., Billi, D., Friedmann, E. I. & Panitz, C. (2005). Effects of a simulated martian UV flux on the cyanobacterium *Chroococcidiopsis* sp. 029. *Astrobiology* 5, 127-140.

[] Córdoba-Jabonero, C., Zorzano, M.-P., Selsis, F., Patel, M.R. & Cockell, C. S. (2005). Radiative habitable zones in martian polar environments. *Icarus* 175, 360-371.

[411] Cox, M.M. & Battista, J. R. (2005). *Deinococcus radiodurans* - the consummate survivor' *Nat. Rev. Microbiol.* 3, 882-892.

[412] Daffonchio, D., Borin, S., Brusa, T., Brusetti, L., van der Wielen, P.W., Bolhuis, H., Yakimov, M.M., D'Auria, G., Giuliano, L., Marty, D., Tamburini, C., McGenity, T.J., Hallsworth, J.E., Sass, A.M., Timmis, K.N., Tselepides, A., de Lange, G.J., Hubner, A., Thomson, J., Varnavas, S.P., Gasparoni, F., Gerber, H.W., Malinverno, E., Corselli, C., Garcin, J., McKew, B., Golyshin, P.N., Lampadariou, N., Polymenakou, P., Calore, D., Cenedese, S., Zanon, F. & Hoog, S.; Biodeep Scientific Party (2006). Stratified prokaryote network in the oxic-anoxic transition of a deep-sea halocline. *Nature* 440, 203-207.

[413] Ellery, A. & Wynn-Williams, D. (2003). Why Raman Spectroscopy on Mars? - A Case of the Right Tool for the Right Job. *Astrobiology*, 3, 565-579.

[414] Englander, J., Klein, E., Brumfeld, V., Sharma, A. K., Doherty, A. J. & Minsky, A. (2004). DNA toroids: framework for DNA repair in Deinococcus radiodurans and in germinating bacterial spores. *J. Bacteriol.* 186, 5973-5977.

[415] Formisano, V., Atreya, S., Encrenaz, T., Ignatiev, N. & Giuranna, M. (2004). Detection of Methane in the Atmosphere of Mars. *Science*, 306, 1758-1761.

[416] Friedmann, E.I., Kappen, L., Meyer, M.A. & Nienow, J.A. (1993). Long-term productivity in the cryptoendolithic microbial community of the Ross Desert, Antarctica. *Microb. Ecol.* 25, 51-69.

[] Gánti, T., Horváth, A., Bérczi, Sz., Gesztesi, A., & Szathmáry, E. (2003). Dark Dune Spots: Possible biomarkers on Mars? *Orig, Life Evol. Biosph.* 33, 515-557.

[422] Garcia-Pichel, F. & Bebout, B. M. (1996). The penetration of UV radiation into shallow water sediments: high exposure for photosynthetic communities. *Mar. Ecol. Progress Ser.* 131, 257-261.

[422] Garcia-Pichel, F. & Castenholz, R. W. (1994). On the significance of solar ultraviolet radiation for the ecology of microbial mats. In Stal, L. J., Camuette, P. (eds): *Microbial mats. Structure, Development and Environmental Significance.* 77-84, Springer, Heidelberg.

[422] Garcia-Pichel, F., Nubel, U. & Muyzer, G. (1998). The phylogeny of unicellular, extremely halotolerant cyanobacteria. *Arch. Microbiol.* 169, 469-482.

[422] Garcia-Pichel, F., Sherry, N.D. & Castenholz, R.W. (1992). Evidence for a UV sunscreen role of the extracellular pigment scytonemin in the terrestrial cyanobacterium *Chlorogloeopsis* spp. *Photochem. Photobiol.* 56, 17-23.

[423] Gilichinsky, D.A., Wagener, S. & Vishnivetskaya, T.A. (1995). Permafrost microbiology. *Permafrost and Periglacial Processes.* 6, 281-291.

[424] Hand, K.P., Calrson, R., Sun, H., Anderson, M., Wynn, W. & Levy, R. (2005). Waves of the Future (for Mars): In-Situ Mid-infrared, Near-infrared, and Visible Spectroscopic Analysis of Antarctic Cryptoendolithic Communities. *American Geophysical Union*, Fall Meeting 2005, #P51D-0959.

[425] Herkenhoff, K.E. & Vasavada, A.R. (1999). Dark material in the polar layered deposits and dunes on Mars. *Journal of Geophysical Research*, Volume 104, E7, 16487-16500.

[] Horváth, A., Gánti, T., Gesztesi, A., Bérczi, Sz. & Szathmáry, E. (2001). Probable evidences of recent biological activity on Mars: appearance and growing of dark dune spots in the South Polar Region. *37th Lunar and Planetary Science Conference* #1543.

[427] Kappen, L., Schroeter, B., Scheidegger, C., Sommerkorn, M. & Hestmark, G. (1996). Cold resistance and metabolic activity of lichens below 0C. *Adv. Space Res.* 18, 119-128.

[549] Kargel, J.S. & Marion, G.M. (2004). Mars as a Salt-, Acid-, and Gas-Hydrate World. *35th Lunar and Planetary Science Conference*, #1965.

[429] Kieffer, H.H. (2003). Behaviour of CO_2 on Mars: real zoo. *Sixth International Conference on Mars*, #3158.

[430] Kolb, C., Lammer, H., Abart, R., Ellery, A., Edwards, H.G.M., Cockell, C. S. & Patel, M.R. (2002). The Martian oxygen surface sink and its implications for the oxidant extinction depth. In *Proceedings of the First European Workshop on Exo-Astrobiology, Graz, Austria.* 181-184. Ed.: Huguette Lacoste. ESA SP-518, Noordwijk, Netherlands: ESA Publications Division.

[432] Krasnopolsky, V.A., Maillard, J.P. & Owen, T.C. (2004). Detection of methane in the martian atmosphere: Evidence for life? *Icarus* 172, 537-547,

[432] Krasnopolsky, V.A., Mumma, M.J., Bjoraker, G.L. & Jennings, D.E.

(1996). Oxygen and Carbon Isotope Ratios in Martian Carbon Dioxide: Measurements and Implications for Atmospheric Evolution. *Icarus*, 124, 553-568.

[433] Kral, T.A., Bekkum, C. R. & McKay, C.P. (2004). Growth of methanogens on a Mars soil simulant. *Orig. Life Evol. Biosph.* 34, 615-626.

[434] Lee, S.W., Thomas, P.C. & Veverka, J. (1982). Wind streaks in Tharsis and Elysium - Implications for sediment transport by slope winds. *J. Geophys. Res.* 87, 10025-10041.

[435] Lovelock, J. (1979). *Gaia, A new look at life on planet Earth.* Oxford University Press.

[436] Magalhaes, J. & Gierasch, P. (1982). A model of Martian slope winds - Implications for eolian transport. *J. Geophys. Res.* 87, 9975-9984.

[437] Malin, M.C. & Edgett, K.S. (2000). Frosting and Defrosting of Martian Polar Dunes, *31st Annual Lunar and Planetary Science Conference*, #1056.

[] Malin Space Science Systems, Mars Orbiter Camera Image Gallery, Images at $http://www.msss.com/moc_gallery$

[438] Matson, D.L. & Brown, R.H. (1989). Solid-state greenhouses and their implication for icy satellites. *Icarus*, 77, 67-81.

[439] Maynard Smith, J. (1998). *Evolutionary Genetics*, Oxford Univ. Press.

[440] Mellon, M.T. & Phillips, R.J. (2001). Recent gullies on Mars and the source of liquid water. *Journal of Geophysical Research*, 106, 1-15.

[] Möhlmann, D. (2004). Water in the upper martian surface at mid- and low-latitudes: presence, state, and consequences. *Icarus*, 168, 318-323.

[] Möhlmann, D. (2005). Adsorption Water-Related Potential Chemical and Biological Processes in the Upper Martian Surface, *Astrobiology*, 5, 770-777.

[443] Morgan-Kiss, R.M. , Priscu, J.C., Pocock, T., Gudynaite-Savitch, L. & Huner, N.P. (2006). Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. *Microbiol Mol Biol Rev.* 70, 222-252.

[444] Nadeau, T.L., Milbrandt, E.C. & Castenholz, R.W. (2001). Evolutionary relationships of cultivated Antarctic oscillatorians (cyanobacteria). *J. Phycol.* 37, 650-654.

[] Pócs, T., Horváth, A., Gánti, T., Bérczi, Sz. & Szathmáry, E. (2004). Possible cripto-biotic crust on Mars? ESA SP-545, 265-266.

[448] Price, P.B. & Sowers, T. (2004). Temperature dependence of metabolic rates for microbial growth, maintenance, and survival. *Proc. Natl Acad. Sci. USA* 101, 4631-4636.

[448] Price, P.B. (2000). A habitat for psychrophiles in deep Antarctic ice. *Proc. Natl Acad. Sci. USA* 97, 1247-1251.

[569] Priscu, J.C., Fritsen, C.H., Adams, E.E., Giovannoni, S.J., Paerl, H.W., McKay C. P., Doran, P.T., Gordon, D.A., Lanoil, B.D. & Pinckney, J.L. (1998). Perennial Antarctic lake ice: an oasis for life in a polar desert. *Science* 280, 2095-2098.

[450] Reiss, D. & Jaumann, R. (2002). Spring defrosting in the Russel crater dune field - recent surface runoff within the last martian year. 33th Lunar and Planetary Science Conference, #2013

[451] Rivkina, E.M., Friedmann, E.I., McKay, C.P. & Gilichinsky, D.A. (2000). Metabolic activity of permafrost bacteria below the freezing point. *Appl.*

Env. Microbiol. 66, 3230-3233.

[452] Roos, J.C. & W.F. Vincent, W.F. (1998). Temperature dependence on UV radiation effects on Antarctic cyanobacteria, *J. Phycol.* 34, 118-125.

[453] Rothschild, L.J. (1995). A "cryptic" microbial mat: a new model ecosystem for extant life on Mars. *Adv. Space Res.* 15, 223-228.

[454] Schoen, C.H. & Dickensheets, D.L. (2000). Tools for Robotic In Situ Optical Microscopy and Raman Spectroscopy on Mars, *Concepts and Approaches for Mars Exploration*, 275.

[455] Sheehan, W. (1996). *The Planet Mars: A History of Observation and Discovery.* The University of Arizona Press, Tucson.

[456] Shirkey, B., McMaster, N.J., Smith, S.C., Wright, D.J., Rodriguez, H., Jaruga, P., Birincioglu, M., Helm, R.F. & Potts, M. (2003). Genomic DNA of Nostoc commune (Cyanobacteria) becomes covalently modified during long-term (decades) desiccation but is protected from oxidative damage and degradation. *Nucleic Acids Res.* 31, 2995-3005.

[583] Squyres, S.W., Grotzinger, J.P., Arvidson, R.E., Bell, J.F., Calvin, W., Christensen, P. R., Clark, B.C., Crisp, J.A. , Farrand, W.H., Herkenhoff, K E., Johnson, J.R., Klingelhöfer, G., Knoll, A.H., McLennan, S.M., McSween, H.Y. Jr., Morris, R. V., Rice, J.W. Jr., Rieder, R. & Soderblom, L.A. (2004). In Situ Evidence for an Ancient Aqueous Environment at Meridiani Planum, Mars. *Science* 306, 1709 - 1714

[458] Supulver, K.D., Edgett, K.S. & Malin, M.C. (2001). Seasonal changes in frost cover in the Martian south polar region: Mars Global Surveyor MOC and TES monitoring the Richardson crater dune field. *32th Lunar and Planetary Science Conference*, #1966.

[459] Tung, H.C., Bramall, N.E. & Price, P.B. (2005). Microbial origin of excess methane in glacial ice and implications for life on Mars. *Proc. Natl Acad. Sci. USA.* 102, 18292-18296.

[460] van der Wielen, P.W., Bolhuis, H., Borin, S., Daffonchio, D., Corselli, C., Giuliano, L., D'Auria, G., de Lange, G.J., Huebner, A., Varnavas, S.P., Thomson, J., Tamburini, C., Marty, D., McGenity, T.J., Timmis, K.N. & BioDeep Scientific Party (2005). The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science.* 307, 121-123.

[461] Vincent, W.F. (1988). *Microbial ecosystems of Antarctica.* Cambridge Univ. Press.

[] Vishnivetskaya, T.A., Rokhina, L.G., Spirina, E.V., Shatilovich, A.V., Vorobyova, E.A. & Gilichinsky, D.A. (2001). Ancient viable phototrops within the permafrost. *Nova Hedwigia*, Beiheft 123, 427-441.

[593] Ward, D.M. (1978). Thermophilic methanogenesis in a hot-spring algal-bacterial mat (71 to 30°C). *Appl. Env. Microbiol.* 35, 1019-1026.

[463] Wynn-Williams, D.D., Edwards, H.G.M., Newton, E.M. & Holder, J.M. (2002). Pigmentation as a survival strategy for ancient and modern photosynthetic microbes under high ultraviolet stress on planetary surfaces. *International Journal of Astrobiology* 1, 39-49.

[464] Wright, D.J., Smith, S.C., Joardar, V., Scherer, S., Jervis, J., Warren, A., Helm, R.F. & Potts, M. (2005). UV irradiation and desiccation modulate the three-dimensional extracellular matrix of *Nostoc commune* (Cyanobacteria). *J. Biol. Chem.* 280, 40271-4081.

[465] Wutipraditkul, N., Waditee, R., Incharoensakdi, A., Hibino, T., Tanaka,Y., Nakamura, T., Shikata, M., Takabe, T. & Takabe, T. (2005). Halotolerant cyanobacterium Aphanothece halophytica contains NapA-

Type Na+/H+ antiporters with novel ion specificity that are involved in salt tolerance at alkaline pH. *Appl. Env. Microbiol.* 71, 4176-4184.

14

Titan: A New Astrobiological Vision of Titan from the Cassini-Huygens Data

François Raulin
Universités Paris 7 et Paris 12

Abstract

Titan, the largest satellite of Saturn and the only satellite in the solar system having a dense atmosphere, is one of the key planetary bodies for astrobiological studies, due to: i) its many analogies with planet Earth, in spite of much lower temperatures; ii) the presence of well observed active organic chemistry, involving several of the key compounds of prebiotic chemistry; and iii) the potential presence of a water ocean in its internal structure. Since the insertion of the Cassini spacecraft in the Saturn system, on July 1rst, 2004, the Cassini-Huygens NASA-ESA mission has already started to provide a tremendous amount of scientific data of paramount importance, in particular from an astrobiology point of view. On the 14th of January 2005, the Huygens probe entered the atmosphere of Titan. Thanks to its six scientific instruments, it was able to carry out a detailed in situ analysis of this environment during the 2.5 hours of descent, and for more than one hour after it landed, safely, on Titan's surface. All instruments have provided precious data on the atmosphere and surface of Titan. Some of the still very preliminary data are described, and the astrobiological consequences of these new data are presented and discussed.

14.1 Introduction

The Earth is certainly, so far, the most interesting planetary body for astrobiology since it is still the only one where we are sure that life is present. However, there are many other bodies of astrobiological interest in the solar system. There are planetary bodies where extraterrestrial life (extinct or extant) may be present, and which thus would offer the

possibility of discovering a second genesis, the nature and properties of these extraterrestrial living systems, and the environmental conditions which allowed its development and persistence. Mars and Europa seem to be the best place for such a quest. On the other hand, there are planetary bodies where a complex organic chemistry is going on. The study of such chemistry can help us to better understand the general chemical evolution in the universe and more precisely the prebiotic chemical evolution on the primitive Earth. Comets are probably the best example, specially considering that their organic content may have been also involved in the prebiotic chemistry on the primitive Earth.

Titan, the largest satellite of Saturn may cover these two complementary aspects and is thus an interesting body for astrobiological research. Moreover, with an environment very rich in organics, it is one of the best targets to look for prebiotic chemistry at a full planetary scale. This is particularly important when considering that Titan's environment presents many analogies with the Earth. Studying Titan today may give us information on the conditions and processes which occurred on the Earth four billion years ago. In addition, models of the internal structure of Titan strongly suggest the presence of a large permanent subsurface water ocean, and the potential for extant life.

Since the Voyager flybys of Titan in the early 1980s our knowledge of this exotic place, the only satellite of the solar system having a dense atmosphere, has indeed been drastically improved. The vertical atmospheric structure has been determined, and the primary chemical composition, trace compounds, and especially organic constituents described. Several other atmospheric species have also been identified later on by ground based observation and by ISO. Other ground based and Hubble observations have also allowed a first mapping of the surface, showing a heterogeneous milieu. However, at the beginning of the millennium, many questions still remained concerning Titan and its astrobiological aspects. What is the origin of its dense atmosphere? What is the source of methane? How complex is the organic chemistry? What is the chemical composition of the aerosols which are clearly present in the atmosphere (and even mask the surface in the visible wavelengths)? What is the chemical composition of Titan's surface? What is the nature of the various potential couplings between the gas phase the aerosol phase and the surface and their role in the chemical evolution of the satellite and its organic chemistry? How close are the analogies between Titan and the primitive Earth? Is there life on Titan?

The NASA-ESA Cassini-Huygens mission was designed to explore the

Saturn system in great detail, with a particular focus on Titan, and to bring answers to these questions. Indeed, since the successful Saturn orbital insertion of Cassini on July 1st 2004, and the release of the Huygens probe in Titan's atmosphere on January 14th, 2005 (Lebreton et al, 2005) many new data have already been obtained which are essential for our vision and understanding of Titan's astrobiological characteristics.

This paper reviews three main aspects of Titan with astrobiological importance, on the basis of these new data provided by Cassini-Huygens (Table 1), and complemented by theoretical modeling and laboratory experimental studies.

14.2 Analogies between Titan and the Earth

With a diameter of more than 5100 km, Titan is the largest moon of Saturn and the second largest moon of the solar system. It is also the only one to have a dense atmosphere. This atmosphere, clearly evidenced by the presence of haze layers (Fig. 1) extends up to approximately 1500 km (Fulchignoni et al, 2005). Like Earth, Titan's atmosphere is mainly composed of dinitrogen (= molecular nitrogen), N_2. The other main constituents are methane, CH_4, with a mole fraction of about 0.016 to 0.02 in the stratosphere, as measured by the Composite InfraRed Spectrometer (CIRS) instrument on Cassini (Flasar et al, 2005) and the Gas Chromatograph-Mass Spectrometer (GC-MS) on Huygens (Niemann et al, 2005) and dihydrogen (= molecular hydrogen,) H_2, with a mole faction of the order of 0.001. With a surface temperature of approximately 94 K, and a surface pressure of 1.5 bar, Titan's atmosphere is nearly five times denser than the Earth's. Despite of these differences between Titan and the Earth, there are several analogies that can be drawn between the two planetary bodies.

The first resemblances concern the vertical atmospheric structure (see Table 2). Although Titan is much colder, with a troposphere (94- 70 K), a tropopause (70.4 K) and a stratosphere (70-175 K) its atmosphere presents a similar complex structure to that of the Earth and also includes, as recently evidenced by Cassini-Huygens, a mesosphere and a thermosphere. Because of a much higher density in the case of Titan, the mesosphere extends to altitudes higher than 400 km (instead of only 100 km for the Earth), but the shape looks very much the same.

These analogies are linked to the presence in both atmospheres of greenhouse gases and antigreenhouse elements. Methane has strong absorption bands in the medium and far infrared regions corresponding to

Table 14.1. *Cassini -Huygens Science Instruments and IDS's and the potential astrobiological return of their investigation.*

Cassini Instruments and Interdisciplinary Programs	**P.I., Team Leader or IDS**	**Country**	**Astrobiological Return**
Optical Remote Sensing Instruments			
Composite Infrared Spectrometer (CIRS)	V. Kunde / M. Flasar	USA	+++
Imaging Science Subsystem	C. Porco	USA	+++
Ultraviolet Imaging Spectrograph (UVIS)	L. Esposito	USA	++
Visual and I.R. Mapping Spectrometer	R. Brown	USA	++
Fields Particles and Waves Instruments			
Cassini Plasma Spectrometer	D. Young	USA	+
Cosmic Dust Analysis	E. Grün	Germany	+
Ion and Neutral Mass Spectrometer	H. Waite	USA	+++
Magnetometer	D. Southwood / M. Dougherty	U.K.	
Magnetospheric Imaging Instrument	S. Krimigis	USA	
Radio and Plasma Wave Spectrometer	D. Gurnett	USA	
Microwave Remote Sensing			
Cassini Radar	C. Elachi	USA	+++
Radio Science Subsystem	A. Kliore	USA	++
Interdisciplinary Scientists			
Magnetosphere and Plasma	M. Blanc	France	+
Rings and Dust	J.N. Cuzzi	USA	+
Magnetosphere and Plasma	T.I. Gombosi	USA	+
Atmospheres	T. Owen	USA	+++
Satellites and Asteroids	L.A. Soderblom	USA	+
Aeronomy and Solar Wind Interaction	D.F. Strobel	USA	++
Huygens Instruments and Interdisciplinary Programs	**P.I. or IDS**	**Country**	**Exobiological Return**
Gas Chromatograph-Mass Spectrometer	H. Niemann	USA	+++
Aerosol Collector & Pyrolyser	G. Isra¨l	France	+++
Huygens Atmospheric Structure Instrument	M. Fulchignoni	Italy	++
Descent Imager/Spectral Radiometer	M. Tomasko	USA	+++
Doppler Wind Experiment	M. Bird	Germany	+
Surface Science Package	J. Zarnecki	U.K.	+++
Interdisciplinary Scientists			
Aeronomy	D. Gautier	France	++
Atmosphere/Surface Interactions	J.I. Lunine	USA	++
Chemistry and Exobiology	F. Raulin	France	+++

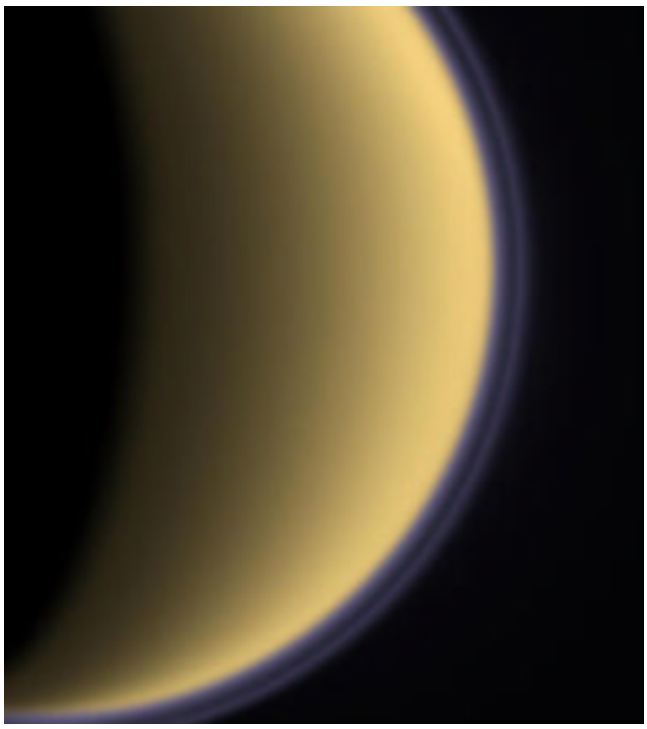


Fig. 14.1. This colorized image taken with the Cassini spacecraft narrow-angle camera shows the complex structure of Titan's atmospheric haze layers. It was taken on July 3, 2004, at a distance of about 789,000 kilometers from Titan. Image Credit: NASA/JPL/Space Science Institute

the maximum of the infrared emission spectrum of Titan and is transparent in the near UV and visible spectral regions. It thus can be a very efficient greenhouse gas in Titan's atmosphere. H_2, which is also absorbing in the far IR (through bimolecular interaction its dimmers) plays a similar role. In the pressure-temperature conditions of Titan's atmosphere, CH_4 can condense but not H_2. Thus, on Titan, CH_4 and H_2 are equivalent respectively to terrestrial condensable H_2O and non-condensable CO_2. In addition, the haze particles and clouds in Titan's atmosphere play an antigreenhouse effect similar to that of the terrestrial atmospheric aerosols and clouds (McKay et al, 1991).

Indeed, methane on Titan seems to play the role of water on the Earth, with a complex cycle which still has to be understood. Although

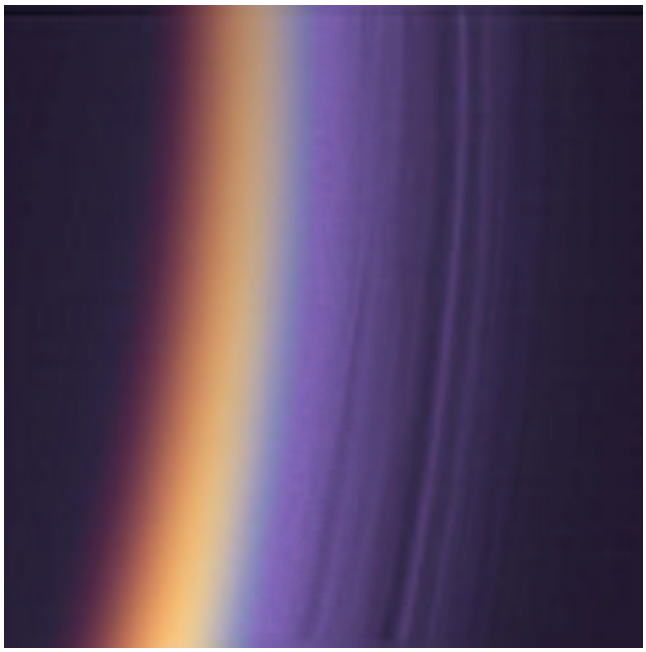


Fig. 14.2. This ultraviolet image of Titan's night side limb, also taken by the narrow-angle camera, shows many fine haze layers extending several hundred kilometers above the surface. Image Credit: NASA/JPL/Space Science Institute

the possibility that Titan is covered with hydrocarbon oceans (Lunine, 1993) is now ruled out (West et al, 2005), it is still possible that Titan's surface include lakes of methane and ethane, although they have not yet been detected by Cassini. Nevertheless, so far the Imaging Science Subsystem (ISS) camera on Cassini has detected dark surface features near the south pole (Figure 2) which could be such liquid bodies. Moreover, the Descent Imager/Spectral Radiometer (DISR) instrument on Huygens has provided pictures of Titan's surface which clearly shows dentritic structures (Figure 3) which look like a fluvial net in a relatively young terrain (fresh crater impacts), strongly suggesting recent liquid flow on the surface of Titan (Tomasko et al, 2005). In addition, GC-MS data show that methane mole fraction increases in the low troposphere (up to 0.05) and reaches the saturation level at approximately 8 km altitude, allowing the possible formation of clouds and rain (Niemann

Table 14.2. *Main Characteristics of Titan (including the HASI-Huygens data)*

Surface radius	2.575 km
Surface gravity	1.35 m s^{-2} (0.14 Earth's value)
Mean volumic mass	1.88 kg dm^{-3} (0.34 Earths value)
Distance from Saturn	20 Saturn radius (1.2 x 10^6 km)
Orbit period around Saturn	16 days
Orbit period around Sun	30 years

Atmospheric data

	Altitude (km)	Temperature (K)	Pressure (mbar)
Surface	0	93.7	1470
Tropopause	42	70.4	135
Stratopause	250	187	1.5 x 10^{-1}
Mesopause	490	152	2 x 10^{-3}

et al, 2005). Furthermore, the Gas Chromatograph-Mass Spectrometer (GC-MS) analyses recorded a 50 percent increase in the methane mole fraction at Titan's surface, suggesting the presence of condensed methane on the surface near the lander.

Other observations from the Cassini instruments clearly show the presence of various surface features of different origins indicative of volcanic, tectonic, sedimentological meteorological processes, as we find on Earth (Figure 4).

Analogies between Titan and the Earth have even been pushed further by comparing Titan's winter polar atmosphere and the terrestrial Antarctic ozone hole (Flasar et al, 2005), although they implied different chemisty.

Another important comparison concerns the noble gas composition and the origin of the atmosphere. The Ion Neutral Mass Spectrometer (INMS) on Cassini and GC-MS on Huygens have detected argon in the atmosphere. Similarly to the Earth atmosphere, the most abundant argon isotope is ^{40}Ar, which comes from the radioactive decay of ^{40}K. Its stratospheric mole fraction is about 4x10^{-5}, as measured by GC-MS (Niemann et al 2005). The abundance of primordial argon (^{36}Ar) is about 200 times smaller. Moreover, the other primordial noble gases have a mixing ratio smaller than 10 ppb. This strongly suggests that Titan's atmosphere, like the Earth, is a secondary atmosphere produced by

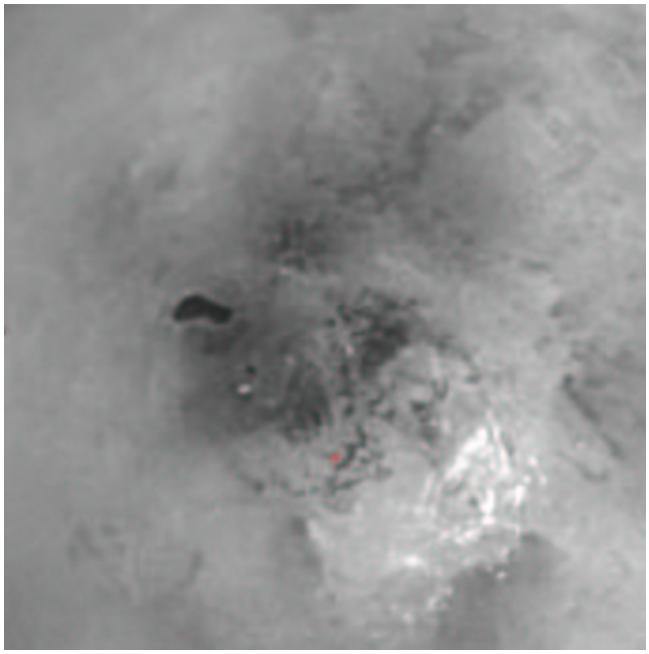


Fig. 14.3. A picture of Titan taken near the south pole by the ISS narrow angle camera on Cassini shows, in addition to bright clouds, the presence of a dark feature on the upper left which could be a liquid hydrocarbon lake. The cross in the middle of the picture marks the pole. Image Credit: NASA/JPL/Space Science Institute

the degassing of trapped gases. Since N_2 cannot be efficiently trapped in the icy planetesimals which accreted and formed Titan, contrary to NH_3, this also indicates that its primordial atmosphere was initially made of NH_3. Ammonia was then transformed into N_2 by photolysis and/or impact driven chemical processes (Owen, 2000; Gautier and Owen, 2002). The $^{14}N/^{15}N$ ratio measured in the atmosphere by INMS and GC-MS (183 in the stratosphere) is 1.5 times less than the primordial N and indicates that several times the present mass of the atmosphere was probably lost during the history of the satellite (Niemann et al, 2005). Since such evolution may also imply methane transformation into organics, this may be also the indication of large deposits of organics on Titan's surface.

Analogies can also be made between the organic chemistry which is very active now on Titan and the prebiotic chemistry which was active

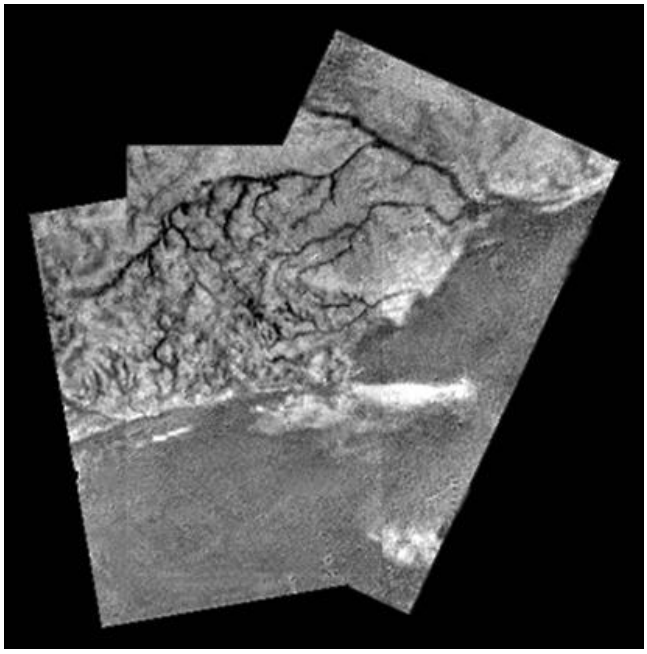


Fig. 14.4. Channel networks, highlands and dark-bright interface seen by the DISR instrument on Huygens at 6.5 km altitude. Credit: ESA/NASA/JPL/University of Arizona

on the primitive Earth. In spite of the absence of permanent bodies of liquid water on Titan's surface, both chemistries are similar. Several of the organic processes which are occurring today on Titan imply some of the organic compounds which are considered as key molecules in the terrestrial prebiotic chemistry, such as hydrogen cyanide (HCN), cyanoacetylene (HC_3N) and cyanogen (C_2N_2). In fact, with several

Indeed, a complex organic chemistry seems to be present in the three components of what one can call, always by analogy with our planet, the "geofluids" of Titan: air (gas atmosphere), aerosols (solid atmosphere) and surface (oceans).

14.3 A complex prebiotic-like chemistry

In the atmosphere of Titan, CH_4 chemistry is coupled with N_2 chemistry producing the formation of many organics hydrocarbons and N-

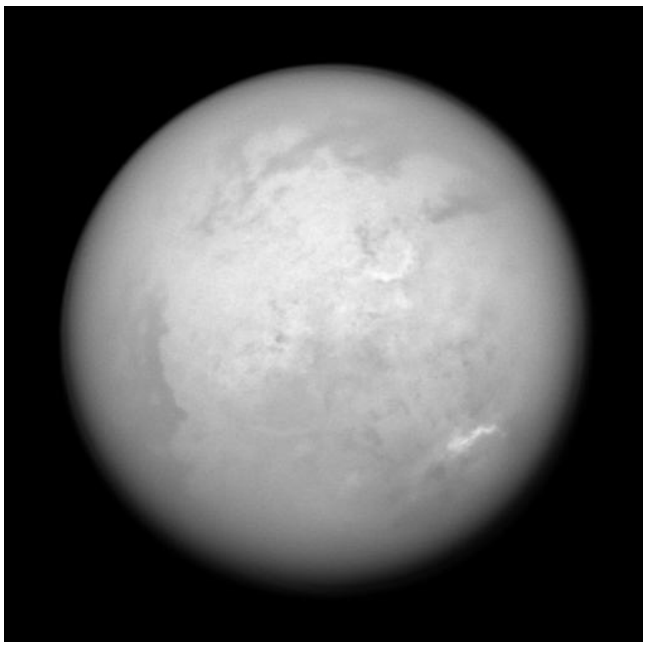


Fig. 14.5. Titan, seen by Cassini spacecraft narrow-angle camera shows a very diversified surface, with bright (like the so-called Xanadu region in the middle of the picture) and darker areas. Image Credit: NASA/JPL/Space Science Institute

containing organic compounds - in gas and particulate phase. Those are hydrocarbons, nitriles and complex refractory organics. Several photochemical models describing the chemical and physical pathways involved in the chemical evolution of the atmosphere of Titan and estimating the resulting vertical concentration profiles of the different involved molecules have been published for the last 20 years. For a review, see the most recent publications and the included references (Lebonnois et al, 2001; Wilson and Atreya, 2004; Hebrard et al, 2005).

The whole chemistry starts with the dissociation of N_2 and CH_4 through electron and photon impacts. The primary processes allow the formation of C_2H_2 and HCN in the high atmosphere. These molecules play a key role in the general chemical scheme: once they are formed, they diffuse down to the lower levels where they allow the formation of higher hydrocarbons and nitriles. Additional CH_4 dissociation proba-

bly also occurs in the low stratosphere through photocatalytic processes involving C_2H_2 and polyynes.

Another approach to the study of organic chemistry on Titan, very complementary of photochemical modelling, is to develop simulation experiments in the laboratory. These experiments seem to well mimick the real processes since recent experiments, carried out in particular at LISA, produce all the gas phase organic species already detected in Titan's atmosphere, within the right orders of magnitude of relative concentration for most of them. Such observation demonstrates the validity of these recent experimental simulations. The experiments also produce many other organics which can be assumed to be also present in Titan's atmosphere. Thus, simulation experiments appear as a very useful guide for further searches (both by remote sensing in situ observations). The gas phase but also the aerosol phases are concerned by such an extrapolation.

In the gas phase, more than 150 different organic molecules have been detected in the simulation experiments (Coll et al, 1998, 1999a). These global simulations of Titan's atmospheric chemistry use an open reactor flown by a low pressure N_2-CH_4 gas mixture. The energy source is a cold plasma discharge producing mid-energy electrons (around 1-10 eV). The gas phase end products (molecules) are analyzed by IRFTS (InfraRed Fourier Transform Spectroscopy) and GC-MS (Gas Chromatography and Mass Spectrometry) techniques; the transient species (radicals and ions) are determined by on line UV-visible spectroscopy. The evolution of the system is also theoretically described using coupled physical and chemical (ions and neutrals) models. The identified organic products are mainly hydrocarbons and nitriles. The absence at a detectable level of molecules carrying amino groups, like amines, with the exception of ammonia, must be highlighted. These experiments have allowed the detection of all gaseous organic species observed on Titan, including C_4N_2 (Coll et al. 1999b). Among the other organics formed in these experiments and not yet detected in Titan's atmosphere, one should note the presence of polyynes (C_4H_2, C_6H_2, C_8H_2) and probably cyanopolyyne HC_4-CN. These compounds are also included in photochemical models of Titan's atmosphere, where they could play a key role in the chemical schemes allowing the transition from the gas phase products to the aerosols. Also of astrobiological interest is the formation of organic compounds with asymmetric carbon such as

CH_3 C*H (C_2H_5) $CH{=}CH_2$ and CH_3 C*H (CN) - $CH{=}CH_2$

Recent experiments on N_2-CH_4 mixtures including CO at the 100 ppm

level (Bernard et al, 2003; Coll et al, 2003) show the incorporation of O atoms in the produced organics, with an increasing diversity of the products (more than 200 were identified). The main O-containing organic compound is neither formaldehyde nor methanol, as expected from theoretical models (both thermodynamic and kinetic), but oxirane (also named ethylene oxide), $(CH_2)_2O$. Oxirane thus appears as a good candidate to search for in Titan's atmosphere. These studies also show the formation of ammonia at noticeable concentration, opening new avenues in the chemical schemes of Titan's atmosphere.

Simulation experiments also produce solid organics, as mentioned above, usually named tholins (Sagan and Khare, 1979). These Titan tholins are supposed to be laboratory analogues of Titan's aerosols these tiny solid particles which are present in Titan's atmosphere and mask the surface of the satellite in the visible. They have been extensively studied since the first work by Sagan and Khare more than 20 years ago (Khare et al, 1984; 1986 and refs. included). These laboratory analogues show very different properties depending on the experimental conditions (Cruikshank et al, 2005). For instance, the average C/N ratio of the product varies between less than 1 to more than 11, in the published reports. More recently, dedicated experimental protocols allowing a simulation closer to the real conditions have been developed at LISA using low pressure and low temperature (Coll et al, 1998; 1999a) and recovering the laboratory tholins without oxygen contamination (from the air of the laboratory) in a glove box purged with pure N_2. Representative laboratory analogues of Titan's aerosols have thus been obtained and their complex refractive indices have been determined (Ramirez et al, 2002), with for the first time - error bars. These data can be seen as a new point of reference to modelers who compute the properties of Titan's aerosols. Systematic studies have been carried out on the influence of the pressure of the starting gas mixture on the elemental composition of the tholins. They show that two different chemical-physical regimes are involved in the processes, depending on the pressure, with a transition pressure around 1 mbar (Bernard et al, 2002; Imanaka et al, 2004).

The molecular composition of the Titan tholins is still poorly known. Several possibilities have been considered such as HCN polymers or oligomers, HCN-C_2H_2 co-oligomers, HC_3N polymers, HC_3N-HCN co-oligomers (Tran et al, 2003 and refs. included). However it is well established that they are made of macromolecules of largely irregular structure. Gel filtration chromatography of the water soluble fraction of Titan tholins shows an average molecular mass of about 500 to 1000

Dalton ((McDonald et al., 1994). Information on the chemical groups included in their structure has been obtained from their IR and UV spectra and from analysis by pyrolysis-GC-MS techniques (Ehrenfreund et al., 1995; Coll et al, 1998; Imanaka et al,, 2004; and refs. included). The data shows the presence of aliphatic benzenic hydrocarbon groups, of CN, NH_2 and C=NH groups. Direct analysis by chemical derivatization techniques before and after hydrolysis allowed the identification of amino-acid or their precursors (Khare et al,, 1986). Their optical properties have been determined (Khare et al, 1984; McKay, 1996; Ramirez et al, 2002; Tran et al, 2003; Imanaka et al., 2004), because of their importance for retrieving observational data related to Titan. Finally, it is obviously of astrobiological interest to mention that Stoker et al.(1990) demonstrated the nutritious properties of Titan tholins for micro-organisms.

Nevertheless, there is still a need for better experimental laboratory simulations to well mimic the chemical evolution of the atmosphere, including the dissociation of dinitrogen by electron impact with energies close to the case of Titan's atmosphere, and the dissociation of methane through photolysis processes. Such an experiment is currently under development at LISA, with the SETUP (Simulation Exprimentale et Thorique Utile la Plantologie) programme which, in a dedicated low temperature flow reactor, couples N_2 dissociation by electron and CH_4 photodissociation by 2-photon (248 nm) laser irradiation, and theoretical studies, in order to improve the chemical schemes. The preliminary results demonstrate the dissociation of methane through the 2-photon process (Romanzin et al, 2005).

Several organic compounds have already been detected in Titan's stratosphere (Table 3). The list includes hydrocarbons (both with saturated and unsaturated chains) and nitrogen-containing organic compounds, exclusively nitriles, as expected from laboratory simulation experiments. Most of these detections were performed by Voyager observations, at the exception of the C_2 hydrocarbons which were observed before, acetonitrile which was detected by ground observation in the millimetre wavelength and water and benzene which were tentatively detected by ISO, the ESA Infrared Space Observatory. Since the Cassini arrival in the Saturn system, the presence of water and benzene has been unambiguously confirmed by CIRS. In addition, the direct analysis of the ionosphere by INMS during the low altitude Cassini fly-bys of Titan shows the presence of many organic species at detectable levels (Fig. 5), in spite of the very high altitude (1100-1300 km).

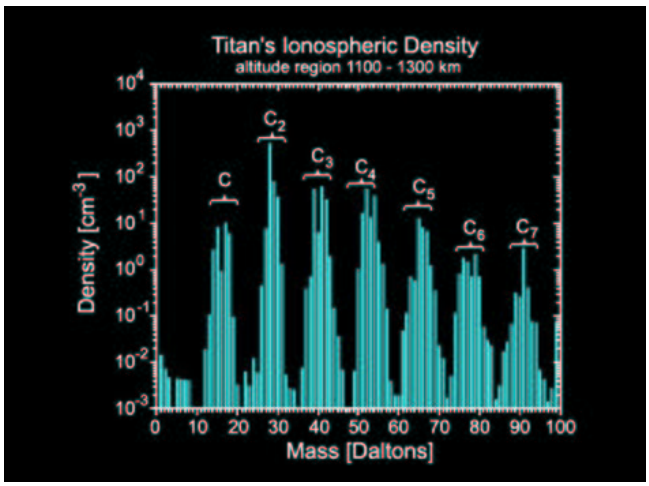


Fig. 14.6. mass spectrum of Titan's ionosphere near 1,200 altitude. The spectrum shows signature of organic compounds including up to 7 carbon atoms.

Surprisingly, GC-MS on board Huygens has not detected a large variety of volatile organic compounds in the low atmosphere. The mass spectra collected during the descent show that the medium and low stratosphere and the troposphere are poor in volatile organic species, at the exception of methane. Condensation of these species on the aerosol particles is a probable explanation for these atmospheric characteristics (Niemann et al, 2005). These particles, for which no direct data on the chemical composition were available before, have been analyzed by the Aerosol Collector and Pyrolyser (ACP) instrument. ACP was designed to collect the aerosols during the descent of the Huygens probe on a filter in two different regions of the atmosphere. Then the filter was heated in a closed oven at different temperatures and the produced gases were analysed by the GC-MS instrument. The results show that the aerosol particles are made of refractory organics non volatile compounds which are only vaporized after degradation at high temperature. The ACP data also show that these organics release HCN and NH_3 during pyrolysis (Israel et al, 2005). This strongly supports the tholin hypothesis: from these new and first *in situ* measurement data it seems very likely that the aerosol particles are made of a refractory organic nucleus, covered with condensed volatile compounds (figure 6). The nature of the pyrolysates provides information on the molecular structure of the re-

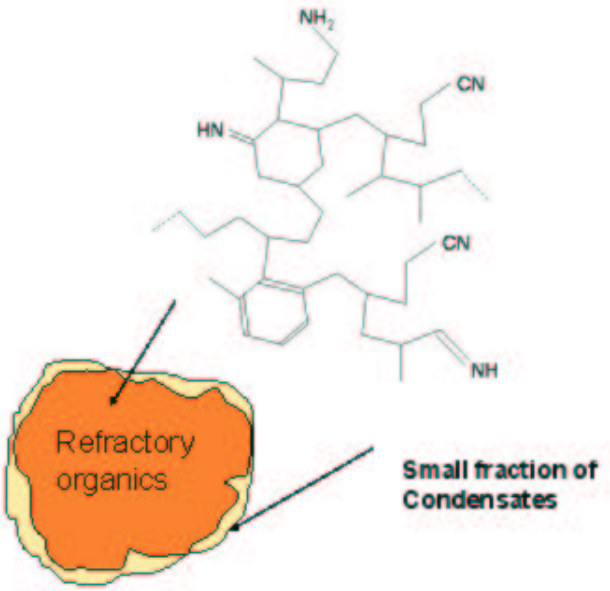


Fig. 14.7. Model of the chemical composition of Titan's aerosol from the Huygens-ACP dat

fractory complex organics: it indicates the potential presence of nitrile groups (-CN), amino groups (-NH_2, -NH- and -N¡) and /or imino groups (-C=N-).

Furthermore comparison of the data obtained for the first (mainly stratospheric particles) and second (mid troposphere) samplings indicate that the aerosol composition is homogeneous (Israel et al, 2005). This also fits with some of the data obtained by the Descent Imager and Radial Spectrometer, DISR, relative to the aerosol particle which indicates a relatively constant size distribution of the particles with altitude (with a mean dimension of the order of one micron).

These particles sediment down to the surface where they likely form a deposit of complex refractory organics and frozen volatile. DISR collected the infrared reflectance spectra of the surface with the help of a lamp, illuminating the surface before the Huygens probe touched down. The retrieving of these infrared data show the presence of water ice, but no clear evidence so far of tholins. The presence of water ice is also

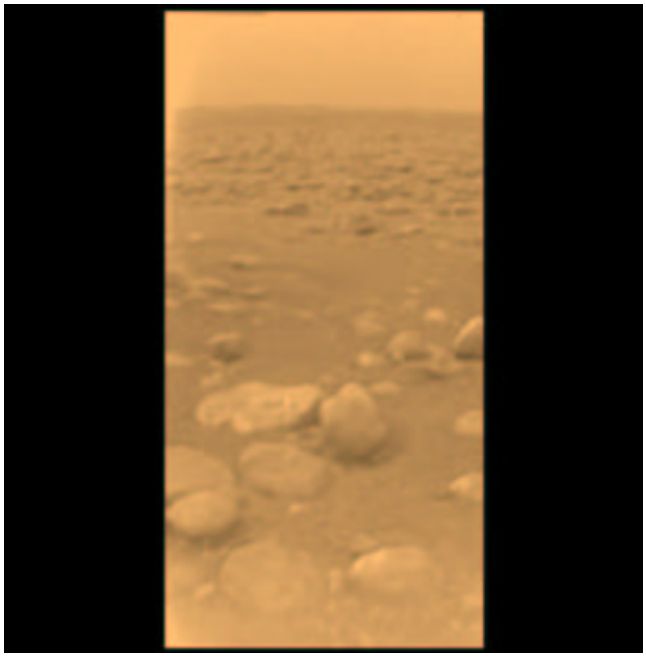


Fig. 14.8. The surface of Titan as seen by the Huygens DISR camera. Credit: ESA/NASA/JPL/University of Arizona

suggested by the data of the SSP instrument (Zarnecki et al, 2005). Its accelerometer measurements can be interpreted as the presence of small water ice pebbles on the surface where Huygens has landed, in agreement with the DISR surface pictures (Figure 7). On the other hand, GC-MS was able to analyse the atmosphere near the surface for more than one hour after the touch down. The corresponding mass spectra show the clear signature of many organics, including cyanogen, C_3 and C_4 hydrocarbons and benzene:, indicating that the surface is much richer in volatile organics than the low stratosphere and the troposphere (Niemann et al, 2005). These observations are in agreement with the hypothesis that in the low atmosphere of Titan, most of the organic compounds are in the condensed phase.

Thus, altogether, these new data show the diversity of the locations where organic chemistry is taking place on Titan. Surprisingly the high atmosphere looks very active, with neutral and ion organic processes; the high stratosphere, where many organic compounds have already been

detected before Cassini and since Cassini arrived in the Saturn system, also shows an active organic chemistry in the gas phase. In the lower atmosphere this chemistry seems mainly concentrated in the condensed phase. Titan's surface is probably covered with frozen volatile organics together with refractory, tholin-like organic materials.

Irradiating effects of cosmic rays reaching Titan's surface may induce additional organic syntheses, particularly if these materials are partly dissolved in some small liquid bodies made of low molecular weight hydrocarbons (mainly methane and ethane). This could indeed allow the additional formation of reactive compounds such as azides as well as the polymerization of HCN (Raulin et al, 1995). Moreover, the interface between the liquid phase and the solid deposits at the surface may include sites of catalytic activity favourable to these additional chemical reactions.

In spite of the surface temperatures, even the presence of liquid water is not excluded. Cometary impacts on Titan may melt surface water ice, offering possible episodes as long as about 1000 years of liquid water (Artemieva and Lunine, 2003). This provides conditions for short terrestrial-like prebiotic syntheses at relatively low temperatures. Low temperatures reduce the rate constants of prebiotic chemical reactions, but may increase the concentration of reacting organics by eutectic effect which increases the rate of the reaction. In addition, the possible presence of a water-ammonia ocean in the depths of Titan, as expected from models of its internal structure (Tobie et al, 2005, and refs. included), may also provide an efficient way to convert simple organics into complex molecules, and to reprocess chondritic organic matter into prebiotic compounds. These processes may have very efficiently occurred at the beginning of Titan's history (with even the possibility of the water-ammonia ocean exposed to the surface) allowing a CHNO prebiotic chemistry evolving to compounds of terrestrial biological interest.

Even if these liquid water scenarios are false, the possibility of a pseudo biochemistry, evolving in the absence of a noticeable amount of O atoms cannot be ruled out, with a N-chemistry, based on ammono analogues replacing the O-chemistry (Raulin and Owen, 2002). Such alternatives of terrestrial biochemistry where, in particular the water solvent could be replaced by ammonia or other N-compounds, have also been recently re-examined by Benner (2002) and by Schulze-Makuch and Irwin (2004).

14.4 Life on Titan?

Several ways can thus be considered in Titan's environment to drive chemistry to prebiotic chemistry and even to biotic systems on Titan. But if life emerged on Titan, are Titan's conditions compatible with the sustaining of life? The surface is too cold and not energetic enough to provide the right conditions. However, the (hypothetical) subsurface oceans may be suitable for life. Fortes (2000)has shown that there are no insurmountable obstacles. With a possible temperature of this ocean as high as about 260 K and the possible occurrence of cryovolcanic hotspots (where volcanic activity increases the temperature) allowing 300 K, the temperature conditions in Titan's subsurface oceans could allow the development of living systems. Even at depth of 200 km, the expected pressure of about 5 kbar is not incompatible with life, as shown by terrestrial examples. The expected pH of an aqueous medium made of 15 percent by weight of NH_3 is equivalent to a pH of 11.5. Some bacteria can grow on Earth at pH 12. Even the limited energy resources do not exclude the sustaining of life. Interestingly, the situation seems similar to that of Europa. However, it must be pointed out that Europas ocean may be much closer to the surface (see chapter by Greenberg). Moreover, Europa's ocean is more likely to be in direct contact with a silicates floor than Titan, and more likely to include hydrothermal vents that Titan.

Taking into account only the potential radiogenic heat flow ($5x10^{11}$ W) and assuming that 1As expected, no sign of macroscopic life has been detected by Huygens when approaching the surface of after it landed. This can be concluded in particular from the many pictures taken by DISR of the same location on Titan during more than one hour after landing. But this does not exclude the possibility of the presence of a microscopic life. The metabolic activity of the corresponding biota, even if it is localized far from the surface, in the deep internal structure of Titan, may produced chemical species which diffuse through the ice mantle covering the hypothetical internal ocean and feed the atmosphere. It has even been speculated in several publications that the methane we see in the atmosphere today is the product of biological activity (Simakov, 2001). If this was the case, the atmospheric methane would be notably enriched in light carbon. Indeed, on Earth, biological processes induce an isotopic fragmentation producing an enrichment in 12C: $^{12}C/^{13}C$ increases from 89 (the reference value, in the Belemnite of the Pee Dee Formation) to about 91-94 depending on the biosynthe-

sis processes. The $^{12}C/^{13}C$ ratio in atmospheric methane on Titan, as determined by the GC-MS instrument on Huygens is 82 (Niemann et al, 2005). Although we do not have a reference for $^{12}C/^{13}C$ on Titan, this low value suggests that the origin of methane is likely to be abiotic.

14.5 Conclusions

Although exotic life, like methanogenic life in liquid methane cannot be fully ruled out (McKay and Smith, 2005), the presence of extent or extinct life on Titan seems very unlikely. Nevertheless, with the new observational data provided by the Cassini-Huygens mission, the largest satellite of Saturn looks more than ever as a very interesting object for astrobiology. The several analogies of this exotic and cold planetary body with the Earth and the complex organic chemical processes which are going on now on Titan provide a fantastic means to better understand the prebiotic processes which are not reachable anymore on the Earth, at the scale and within the whole complexity of a planetary environment.

The origin and cycle of methane on Titan illustrate the whole complexity of the Titan's system. Methane may be stored in large amount in the interior of the satellite, under the form of clathrates (methane hydrates) trapped during the formation of the satellite from the Saturnian subnebula where it was formed by Fisher-Tropsch processes (Sekine et al, 2005). It may also be produced through high pressure processes, like serpentinization allowing the formation of H_2 by reaction of H_2O with ultramafic rocks, or by cometary impact (Kress and McKay, 2004). Interestingly, those processes have rarely been considered in the case of the primitive Earth, although they may have contributed to a possible reducing character of the primordial atmosphere of our planet, as mentioned earlier in this chapter. This is an example of how Titan's study is indeed providing new insights into terrestrial chemical evolution.

In Titan's atmosphere, methane is photolysed by solar UV, producing mainly ethane and tholins-like organic matter. The resulting life time of methane in Titan's atmosphere is relatively short (about 10 to 30 myr). Thus methane stored in Titan's interior may be continuously replenishing the atmosphere, through degassing induced by cryovolcanism. This volcanic activity at low temperature, where terrestrial lava is replaced by water, has been clearly evidenced from the first images of Titan's surface provided by the Visual and Infrared Mapping Spectrometer (VIMS), ISS and Radar instruments on Cassini (Sotin et al, 2005).

Fig. 14.9. One of the largest crater impact (about 80 km diameter) observed on Titan's surface by two Cassini instruments: VIMS infrared (left) and false-color image (right) and radar image (center) and false-color image (right). The faint halo, slightly bluer and darker than the surrounding parts, is probably somewhat different in composition. Since it is made of material excavated when the crater was formed, this indicates that the composition of Titan's upper crust varies with depth, Credit: NASA/JPL/University of Arizona

It may also be released episodically to the atmosphere, as recently suggested by Tobie et al (2006). In any case, the methane cycle should result in the accumulation of large amounts or complex organics on the surface and large amounts of ethane, which mixed with the dissolved atmospheric methane should form liquid bodies on the surface of in the near sub-surface of the satellite. It is possible that the dark feature seen in Figure 2 is one of these expected liquid bodies.

The Cassini-Huygens mission is far from complete. It will continue its systematic exploration of the Saturnian system up to 2008, and probably 2011 if the extended mission is accepted. Numerous data of paramount importance for astrobiology are still expected from several of its instruments (Table 1). The CIRS spectrometer should be able to detect new organic species in the atmosphere during the future limb observation of Titan, specially at the pole. ISS and VIMS should provide a detailed picture of Titan's surface revealing the complexity but also the physical and chemical nature of this surface and its diversity, as recently evidenced by the discovery of a 5-micron bright spot (Barnes et al, 2005). Radar observation will also continue the systematic coverage of Titan's surface which shows contrasted regions of smooth and rough areas, suggesting a possible shoreline. The coupled observation of the same regions by these instruments, as already performed (Fig. 8) with will surely bring essential information to our understanding of this new, exotic and astonishing worlds.

Acknowledgements

The author wish to thanks the Cassini-Huygens teams, and particularly the Huygens project scientist, Jean-Pierre Lebreton, his colleague Olivier Witasse, and the PIs of the Huygens instruments, specially Guy Israel and Hasso Niemann, for making available several of the data used in this paper. The author also wishes to thank his colleagues of the LISA-GPCOS team, especially Patrice Coll, Marie-Claire Gazeau, Eric Hebrard and Mai-Julie Nguyen for useful discussions during the preparation of the manuscript. Final thanks go to the European Space Agency, ESA, and the French Space Agency, CNES, for financial support.

References

[466] Abbas,O., and Schulze-Makuch, D. (2002). *ESA SP* **518**, -345-348.
[467] Artemieva, N., and Lunine, J. (2003). *Icarus* **164**, -471-480.
[468] Barnes et al (2005). *Science* **310**, -92-95.
[469] Benner, S. (2002). *Planning session for NRC Committee on the Origins and Evolution of Life* , (http://www7..nationalacademies.org/ssb/weirdlife.html).
[471] Bernard, J.-M., Coll, P., and Raulin, F. (2002). *Proc. 2d European Workshop on Exo-/Astro-Biology, ESA SP* **518**, -623-625.
[471] Bernard, J.-M., Coll, P., Coustenis, A., and Raulin, F. (2003). *Planet. Space Sci.* **51**, -1003-1011.
[475] Coll, P., Bernard, J.-M. , Navarro-Gonzlez, R., and Raulin, F. (2003). *Astrophys. J.* **598**, -700-703.
[475] Coll, P., Coscia, D., Gazeau, M.-C., and Raulin, F. (1998). *Origins of Life and Evol. Biosph.* **28**, -195-213.
[475] Coll, P.et al (1999a). Experimental laboratory simulation of Titan's atmosphere: aerosols and gas phase *Planet. Space Sci.* **47**, -1331-1340.
[475] Coll, P., Guillemin, J.C., Gazeau, M.C., and Raulin, F. (1999b). *Planet. Space Sci.* **47**, -1433-1440.
[476] Cruikshank, D.P., Imanaka, H., and Dalle Ore, C.M. (2005). *Adv. Space Res.* **36**, -178-183.
[477] Ehrenfreund, P.et al (1995). *Adv. Space Res.* **15**, -335-342.
[478] Feng T., Owen B. T., Pavlov, A.A, and De Sterck, H. (2005). *Science* **308**, -1014-1017.
[479] Flasar, F.M. et al, (2005). *Science* **308**, -975-978.
[480] Fortes, A.D. (2000). *Icarus* **146**, -444-452.
[481] Fulchignoni, M.et al (2005). *Nature* **438**, -785-791.
[482] Gautier, D., and Owen, T. (2002). *Space Science Review* **104**, -347-376.
[483] Hebrard, E., Benilan, Y., and Raulin, F. (2005). *Adv. Space Res.* **36**, -268-273.
[484] Imanaka, H. et al. (2004). *Icarus* **168**, -344366.
[485] Israel, G. et al (2005). *Nature* **438**, -796-799.
[487] Khare, B.N., Sagan, C., Arakawa, E.T., Suits, F., Callicott, T.A., and Williams, M.W. (1984). *Icarus* **60**, -127-137.
[487] Khare, B.N., Sagan, C., Ogino, H., Nagy, B., Er, C., Schram, K.H., and Arakawa, E.T. (1986). *Icarus* **68**, -176-184.

[488] Kress, M.E., and McKay, C. P. (2004). *Icarus* **168**, -475-483
[489] Lebonnois, S., Toublanc, D., Hourdin, F., Rannou, P. (2001). *Icarus* **152**, -384-406.
[490] Lebreton, J.-P. et al (2005). *Nature* **438**, -758-764
[555] Lunine, J. I. (1993). *Rev. Geophys.* **31**, -133-149.
[492] McDonald, G.D., Thompson, W.R., Heinrich, M., Khare, B.N., and Sagan, C. (1994). *Icarus* **108**, -137-145.
[495] McKay, C.P.: (1996). *Planet. Space Sci.* **44**, -741-747.
[495] McKay, C.P., Pollack, J.B. and Courtin, R. (1991). *Science* **253**, -1118-1121.
[495] McKay, C.P., and Smith, H.D. (2005). *Icarus* **178**, -274276.
[496] Niemann, H.B. et al (2005). *Nature* **438**, -779-784.
[497] Owen, T. (2000). *Planet. Space Sci.* **48**, -747-52.
[498] Ramirez, S.I., Coll, P., Da Silva, A., Navarro-Gonzalez, R., Lafait, J., and Raulin, F. (2002). *Icarus* **156**, -515-530.
[500] Raulin, F., Bruston, P., Paillous, P., and Sternberg, R. (1995). *Adv. Space Res.* **15**, -321-333.
[500] Raulin, F., and Owen, T. (2002). *Space Science Review* **104**, -379-395.
[501] Romanzin, C. et al (2005). *Adv. Space Res.* **36**, -258-267.
[502] Sagan, C., and Khare, B.N. (1979). *Nature* **277**, -102-107.
[503] Schulze-Makuch, D., and Irwin L. N. (2004). *Life in the Universe, Expectations and constraints* -Springer
[504] Sekine, Y. Sugita, S., Shido, T., Yamamoto, T., Iwasawa, Y., Kadono, T.,and Matsui, T. (2005). *Icarus* **178**, -154-164.
[505] Simakov, M.B. (2001). *ESA SP* **496**, -211-214.
[506] Sotin, C. et al (2005). *Nature* **435**, -786-789.
[507] Stoker, C.R., Boston, P. J., Mancinelli, R. L., Segal, W., Khare, B. N., and Sagan, C. (1990). *Icarus* **85**, -241-256.
[509] Tobie, G., Grasset, O., Lunine, J.I., Mocquet, A., and Sotin, C. (2005). Titan's internal structure inferred from a coupled thermal-orbital model, *Icarus* **175**, -496-502.
[509] Tobie, G., Lunine, J.I., and Sotin, C. (2006). *Nature* **440**, -61-64.
[510] Tomasko M.G., et al (2005). *Nature* **438**, -765-778.
[511] Tran, B.N., Joseph, J. C., Ferris, J. P., Persans, P. D., and Chera, J. J. (2003). *Icarus* **165**, -379-390.
[512] West, R. A., Brown, M. E., Salinas, S. V., Bouchez, A. H., and Roe H. G., (2005). *Nature* **436**, -670-672.
[513] Wilson, E. H. and Atreya, S.K. (2004). *J. Geophys. Res.* **109**,(E6), E06002.
[514] Zarnecki, J.C. et al (2005). *Nature* **438**, -792-795

15

Europa, the Ocean Moon: Tides, permeable ice, and life

Richard Greenberg
University of Arizona

15.1 Introduction: Life beyond the Habitable Zone

As the horizons of the field of Astrobiology have extended to the rapidly growing population of known extra-solar planetary systems, the concept of a habitable zone around each star has been of particular interest. With each discovery, the question is raised about whether the new planet is in the so-called *habitable zone*, or whether hypothetical orbits within that zone would be stable. In this context, the *habitable zone* is commonly defined as a range of distances from the star where temperatures would be in a range for life to be viable and for water to exist near the surface in a phase able to support living organisms (e.g., Kasting et al. 1993, Menou and Tabachnik 2003, Raymond and Barnes 2005). This restrictive definition can contribute to a relatively pessimistic view of the potential for extra-terrestrial life (e.g. Ward and Brownlee 2000). It has even been used to advocate planetary exploration objectives, purportedly motivated by the search for life, that are restricted to the terrestrial planets in our solar system (e.g. Hubbard 2005).

Unless corrected, this widespread semantic usage of the term *habitable zone* would leave out one of the most likely places for exploration to reveal extraterrestrial life: Jupiter's moon Europa. The analyses cited above would neglect tidal friction, which is known in the case of Jupiter's Galilean satellites to provide significant heating, comparable in fact to the solar heating in the habitable (and inhabited) terrestrial-planet zone. (Ward and Brownlee (2005) implicitly acknowledge this broader range of habitability in using an image of Conamara Chaos on Europa as the frontispiece of their book, while still tending to define the habitable zone in the more restricted sense.) Of course, heat alone may not provide energy accessible to organisms, but the heat can promote a physical

setting that facilitates derivation of metabolic energy from chemical or solar (or stellar) sources.

While the observational astrophysical community is not yet ready to worry about satellites, and some political interests might prefer to focus on the Moon and Mars in our own solar system, a true description of the habitable zone of any planetary system should include any appropriate satellite systems whose heat source may be relatively independent of a nearby star.

This chapter reviews what is known about the physical character of Europa's potential biosphere, specifically the 100-km-deep liquid water ocean, and the thin layer of ice that lies above it. The global ocean just below the ice likely has received substances released from the deep silicate interior, while the surface ice has been bombarded by cometary and asteroidal material, as well as impacts from the swarm of circum-Jovian particles. The dynamic ice crust controls the interface of these endogenic and exogenic substances, maintaining disequilibrium conditions but allowing interaction over various spatial and temporal scales, as shown in this chapter.

The permeability of the ice layer is thus critical to providing a setting that supports life, both in the ocean and within the crust itself. For a complete discussion of the physical character and likely processes operating on Europa, and their implications for life, see Greenberg (2005). Another recent review from a biological perspective is by Lipps and Rieboldt (2005).

An aspect of Europa's physical character is that the environment at a given location in the ice evidently changes on various time-scales. Short term change provides mixing and disequilibrium conditions that could supply any individual organisms with daily needs; stability over thousands of years could help an ecosystem to flourish; and changes over longer time-scales (still very rapid compared with usual geological change) would force adaptation, perhaps serving as a driver for biological evolution.

In order to help visualize how the physical character of Europa might support life, a hypothetical ecosystem within a crack in the ice is described in Section V. But first we review the appearance of Europa and interpret what is seen in the context of the effects of tides.

15.2 The surface of Europa

15.2.1 Global Scale

Europa's surface is composed predominantly of water ice according to reflectance spectra (Pilcher et al., 1972), and thus the moon would appear to the human eye as a nearly uniform white sphere, 1565 km in radius. The gravitational figure measured by spacecraft indicates that an outer layer as thick as 150 km has the density of H_2O (Anderson et al., 1998). Much of that thick layer is probably liquid water according to estimates of tidal heating (Section III.C.1 below), while the surface is frozen due to radiative cooling.

The surface does contain more than pure water ice. Even global-scale pictures (with resolution $>$ 10 km per pixel) taken with Voyager and Galileo spacecraft cameras show (when the contrast is enhanced as in Fig. 1) orange-brown markings of still-unidentified substances, which likely include hydrated salts (McCord et al., 1998a) and sulfur compounds (Carlson et al., 1999). The patterns include splotches ranging from tens to 100s of kilometers across and a network of narrow lines. When the splotches and lines are observed at high resolution (as discussed below), they prove to represent the two major resurfacing processes on Europa: formation of chaotic terrain and tectonics, respectively.

The resurfacing has been rapid and relatively recent as evidenced by the paucity of craters. Apparently, impacts have had a minimal role in shaping the current surface. In Fig. 1, for example, only crater Pwyll, with its extensive white rays, is readily evident at this scale (lower left). The paucity of craters tells us that the current surface must be very young. It is continually reprocessed at such a great rate that most of the observable terrain, structures and materials have probably been in place $<$ 50 Myr (Zahnle et al. 2003).

Even from the global appearance we see indications that the surface has been reworked by stress and heat, forming the tectonic and chaotic terrain, respectively. This chapter shows how these processes modified the surface and have provided, in diverse ways, access to the ocean. The relatively dark, orange-brown material that marks sites of these processes is indistinguishable whether it appears along a tectonic lineament or around a patch of chaotic terrain. It may represent concentration of impurities by thermal effects due to the near exposure of warm liquid or be the most recently exposed oceanic substances. All this activity is rapid, recent, and thus likely on-going.

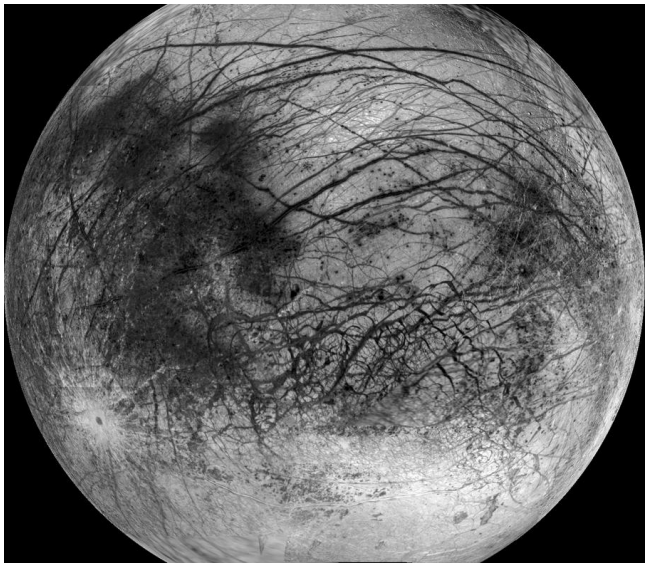


Fig. 15.1. A full-disk view of Europa shows large scale lineaments indicative of tectonic cracks, dark splotches marking thermally created chaotic terrain. Craters are rare, although a prominent rayed crater, Pwyll, is evident to the left. The equator runs across the center here. The dark wedge-shaped and tightly curved features just south of the equator are mostly dilational bands.

15.2.2 Tectonics

Ridges and cracks In regional-scale images(1 km per pixel), many of the global-scale dark lines resolve into double lines (Fig. 2). Counting the brighter zone between the dark lines, Voyager scientists named these features "triple bands" (Lucchitta and Soderblom, 1982). Fig. 2 shows an X-shaped intersection of global-scale triple bands that is located to the left side of Fig. 1. In both figures, rays from crater Pwyll 1000 km to the south are visible across this region. Fig. 2 has nearly vertical illumination aligned with the camera's point of view (i.e. low phase angle), so the visible features represent albedo and color variations. The same area viewed at 200 m resolution and with more oblique illumination (Fig. 3), reveals the morphology of the large-scale lineaments, showing them to be complexes of ridges. Each ridge complex consists of sets of ridges, roughly parallel, although crossing or intertwined in various places.

How does this structural form (in Fig. 3) relate to the appearance at lower resolution and vertical illumination in Fig. 2? As shown in Fig.

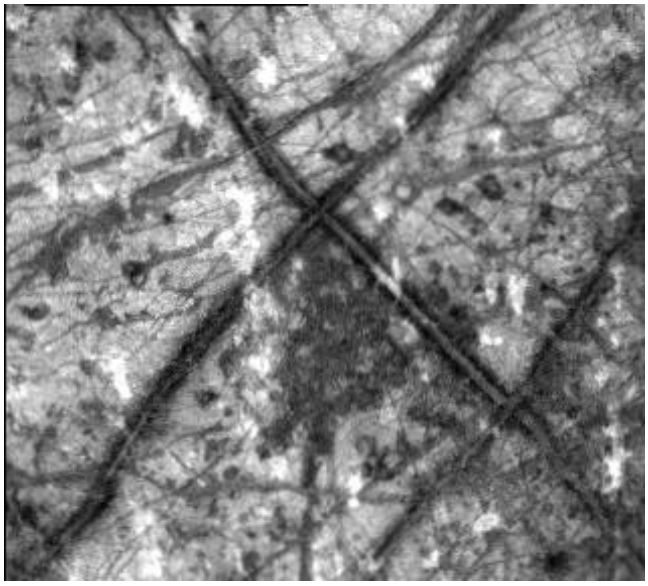


Fig. 15.2. A higher resolution view of the prominent intersection of lineaments at the left of Fig. 1 shows that the lines resolve into double dark lines known as Triple bands. The dark splotch just south of the intersection is Conamara Chaos, which is about 80 km across. The white streaks are rays from crater Pwyll, 1000 km to the south.

3, along either side of the major ridge complexes the adjacent terrain is slightly darker than average. This subtle, diffuse darkening is what shows as the thin dark lineaments on global scale images (e.g. Fig. 1) or as the double dark components of the "triple bands" at kilometer resolution (e.g. Fig. 2). The ridge systems, which are structurally the most significant characteristic of the global-scale lineaments, lie between the darken margins and at lower resolution can only be seen as the relatively bright center line of the triple bands.

The dark coloration on the margins, which on global and regional scale images were most prominent, prove to be quite subtle in the higher resolution images with oblique illumination (e.g. Fig. 3). In general, the dark areas do not correlate with any single type of morphological structure, although they do tend to be somewhat smoother than average. The association with ridge complexes is very important from the perspective of planetary geochemistry: The ridges probably mark cracks along which

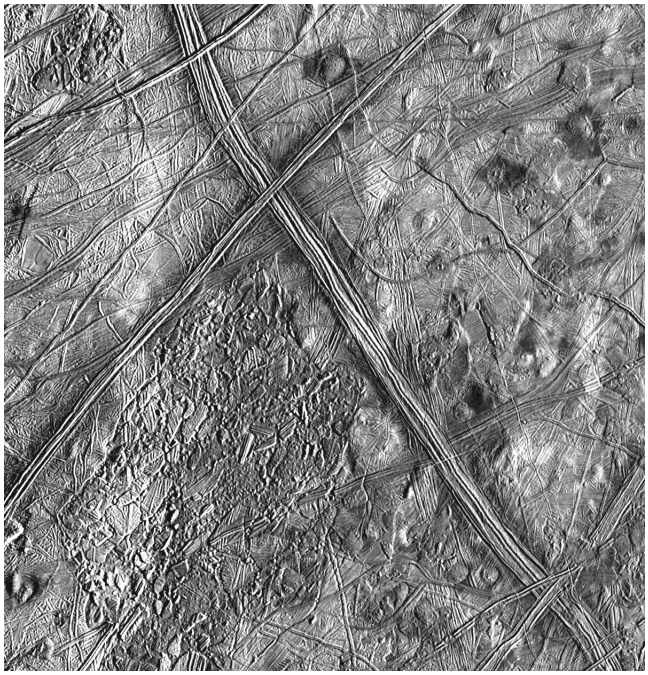


Fig. 15.3. The area shown in Fig. 2 is here shown with oblique illumination and resolution of 200 m, showing more morphological detail. The lineaments are complexes of double ridges. The double dark lines in Fig. 2 prove to be only the faint darkening alongside the ridge complexes. Conamara is the archetype of chaotic terrain: rafts of older surface, displaced in a lumpy matrix.

oceanic substances have been able to reach the surface; thus, these margins are the first of several examples of darkening associated with ocean water near or at the surface.

Examination of the ridge complexes that make up the global lineaments (e.g. in Fig. 3) shows them to be composed of multiple sets of double ridges. In fact double ridges are ubiquitous on Europan tectonic terrain. While ridges appear in several varieties (Greenberg et al. 1998), the common denominator seems to be the simple double-ridge system. Numerous examples, besides those that make up the global-scale lineaments, can be seen in Fig. 3. Fig. 4 shows a close up of terrain (just north of that shown in Fig. 3) densely packed with double ridges. Fig. 5 shows double ridges in the highest resolution image ever taken of Eu-

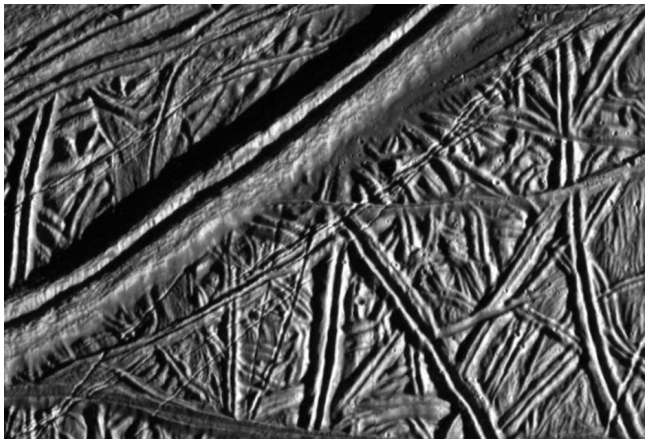


Fig. 15.4. Densely ridged terrain at the top of Fig. 3, here at much higher resolution. The large double ridge is about 2 km across.

ropa, with a road-cut-like cross section that offers a view of the interiors of some ridges.

As ridges formed across the surface, they covered over the preexisting terrain. A single ridge pair can often cover over 1000km^2. In some places, ridges have crossed over ridges repeatedly in the same area creating densely ridged terrain like that shown in Fig 4.

Such areas often appear smooth in images with resolution inadequate to resolve the individual ridges. Geologists have mapped such smooth-seeming terrains as "background plains" (e.g. Prockter et al., 1999), promoting an assumption that this is the oldest type of terrain on Europa (Greeley et al. 2000). That terminology might suggest an initial slate-cleaning event, but such "plains" are really only an artifact of resolution. There is no evidence that resurfacing has ever been other than a continual process nor that there has been any change in styles of resurfacing over time. Looking back in time, there is a geological horizon, but that limit probably represents the oldest features that are still recognizable from a continual history of resurfacing that started even earlier. There is no evidence that the time horizon represents a single slate-cleaning event.

There is considerable evidence regarding the mechanism that creates the double ridges. The global scale lineaments correlate reasonably well with theoretical tidal stress patterns (Section III.B.1), indicating that

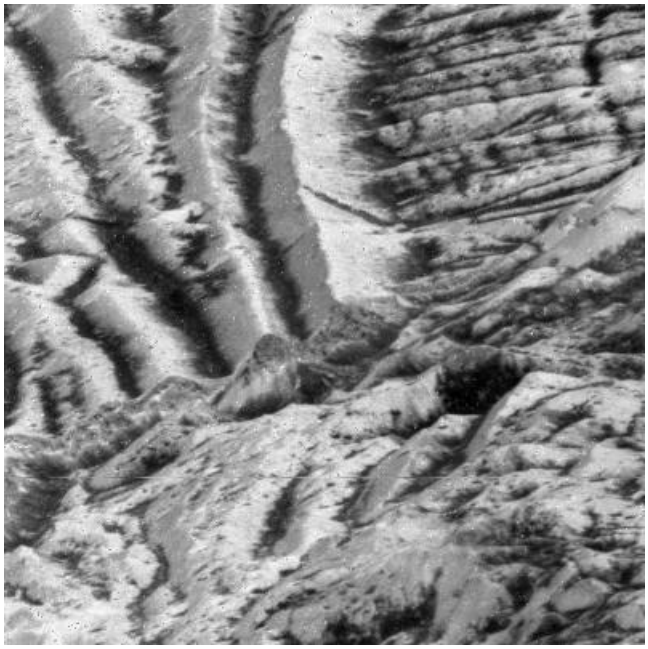


Fig. 15.5. The highest resolution image taken by Galileo (6 m/pixel). The view is oblique (not straight down) as if looking sideways at the ground from an airplane window. In this frame, a couple of ridge pairs come down from the upper left and are cut off by chaos in the foreground, giving a road-cut-style cross-section.

they are associated with tensile cracks in the crust. Because these lineaments seem to comprise complexes of double ridges, it is reasonably assumed that simple double ridges are similarly associated with cracks. Moreover, it has been generally assumed that the double nature is due to ridges running along each side of a crack.

Identification of ridges with cracks has been reinforced by investigations of the locations, orientations, geometries, appearance at higher resolution, displacement of adjacent crust, and sequences of formation over time. The studies have shown that these various characteristics can be explained in terms of tidal stress. For example, distinctive and ubiquitous cycloid-shaped lineament patterns (e.g. Fig. 6) follow from the periodic variation of tidal stress during crack propagation (see Section III.B.1). Cycloidal crack patterns also provide strong evidence for a liquid water ocean under the ice crust, because an ocean is required

Fig. 15.6. An image from the Voyager spacecraft, showing cycloidal ridges: chains of arcs, with each arcuate segment 100 km long. Running from near the lower left corner up to a dark parallelogram is Astypalaea Linea.

in order to give adequate tidal amplitude for these distinctive patterns to form (Hoppa et al. 1999a).

Tectonic displacement Strike-slip displacement (where the surface on one side of a crack has sheared past the surface on the other side, as along California's San Andreas fault) is common on Europa and has important implications (Tufts et al. 1999, Hoppa et al. 1999b, 2000, Sarid et al. 2002). Observed examples of displacement fit a theoretical model driven by tides and requiring that cracks penetrate to a low-viscosity "decoupling layer" such as a liquid ocean (see Section III.B.2.b). Strike-slip generally occurs along ridges (e.g. Fig. 7), rarely along any of the observed simple, ridge-less cracks. Thus, strike-slip displacements suggest that ridges may form along cracks that penetrate entirely through the solid ice crust to the liquid ocean. If so, then under much of Europa's tectonic terrain, the crust must have been thin enough (probably ¡10 km) to allow such penetration (Hoppa et al. 1999a). Strike-slip displacement also has several other crucial implications, including evidence for non-synchronous rotation and for recent polar wander, as discussed in Section III.B.2.b.

Surface displacement also includes opening of cracks, filled by broad bands on the surface (Fig. 8). Recognizable in Voyager and Galileo low-resolution images (e.g. Schenk and McKinnon 1989, Pappalardo and Sullivan 1996), at higher resolution most of these dilation bands exhibit similar characteristics, especially fine parallel ridges within most bands, usually symmetrical about the centerline. Most bands are re-

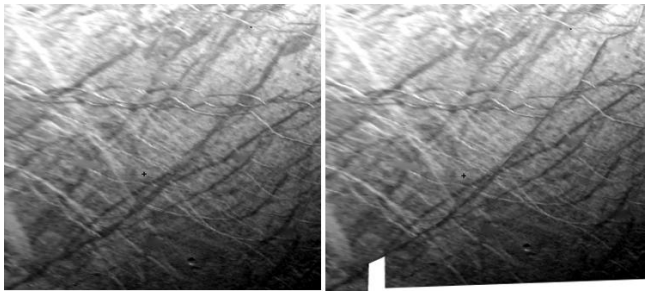


Fig. 15.7. A normal projection of Astypalaea Linea (c.f. Fig. 6), showing the wispy bright lines that end abruptly at Astypalaea. Also note the dark parallelogram at one end. (b) A reconstruction shows that Astypalaea is a strike-slip (shear) fault. Note (going back in time) the realignment of the wispy white lines and the closing of the parallelogram.

constructable (e.g. Fig. 8), demonstrating their dilational nature and they compose a structural continuum with the forms of ridge systems (Tufts et al. 2000). The mobility represented by bands is consistent with the underlying liquid ocean inferred from the mobility of the strike-slip displacement mechanism and from the tidal amplitude of the cycloidal cracks.

The large amounts of new surface created during dilation of the bands, which has been going on as far back in time as we can recognize in the tectonic record, raised the question of how the surface-area budget has been balanced. Evidence for the processes that remove surface on Earth (plate subduction or Himalaya-like mountain building) is not found on Europa. Nevertheless, sites of surface convergence have been found by reconstruction of surface plate motions (Sarid et al. 2002). These sites display bands with internal striations, rounded boundaries, shallow lips at their edges, and only low topography. Unlike on Earth, it appears that there was not much solid material to resist the convergence of surface plates.

15.2.3 Chaotic terrain

The dark splotches in the low-resolution global views (e.g. Fig. 1) represent the other major type of terrain on Europa, *chaotic*, and correspondingly the other major type of resurfacing, thermal. Resolution of 200 m/pixel with appropriate illumination reveals the character of chaotic terrain. The archetypal example is Conamara Chaos, just south

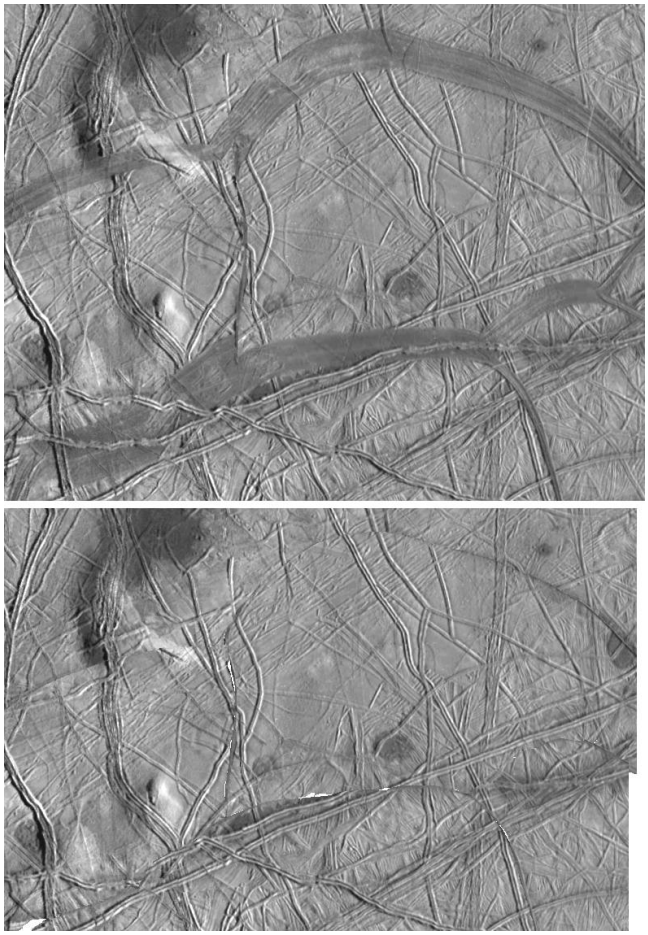


Fig. 15.8. Examples of dark dilational bands (also visible in Fig. 1). (b) Reconstruction of the bands shows them to be dilational, as plates of the surface have separated, allowing in-filling of fresh material.

of the intersection of the crossing global lineaments shown in Figs. 2 and 3. Typical of chaotic terrain, this region has been thermally disrupted, leaving a lumpy matrix with somewhat displaced rafts, on whose surfaces fragments of the previous surface are clearly visible. Conamara, like other chaotic areas, has the appearance of a site at which the crust had melted, allowing blocks of surface ice to float to slightly displaced

locations before refreezing back into place (Carr et al. 1998). Similar features are common in Arctic sea ice and even frozen lakes in terrestrial temperate regions, where the underlying liquid has been exposed. Thus chaotic terrain, like tectonics, appears to represent significant exposure of the liquid ocean at the surface (Greenberg et al. 1999).

While Conamara is a good example, some of the most important results about chaotic terrain were only revealed by identifying and characterizing all chaos regions as completely as possible with the Galileo data set (Greenberg et al. 1999, Riley et al., 2000), while quantitatively accounting for observational selection effects.

Chaotic terrain is very common, covering nearly half of Europa's surface (Greenberg et al. 1999, Riley et al. 2000). The other half is covered by tectonic terrain (the densely packed and overlaid ridges, cracks, and bands). Chaotic terrain is not necessarily young, and it likely formed at various times and places throughout Europa's geological history. Examples of chaotic terrain are found in all degrees of degradation, as cracking, ridge formation, and thermal effects destroy the older terrain. In general, it is more difficult to recognize older or smaller patches of chaotic terrain, which can introduce an observational selection bias favoring younger and larger examples, as demonstrated by Hoppa et al. (2001a). Thus, for example, impressions that chaos is systematically younger than tectonic terrain (e.g. Prockter et al. 1999, Greeley et al. 2000, Figueredo and Greeley 2004) are probably artifacts of this observational bias. No evidence has yet been developed for any systematic change over time in frequency or style of chaos formation. In fact, where high resolution images are available, some chaotic terrain is older than most of the tectonics nearby (e.g. Riley et al. 2006).

15.3 Tides

The activity evidenced on Europa's surface is driven by tides. Tides generate periodic global stresses that crack and displace surface crust, as the thin shell of ice must accommodate to the changing global shape of the ocean below it. Moreover, tidal friction provides the dominant internal heat source, which keeps the most of the thick water layer in the liquid state, as well as allowing occasional local or regional melting of the ice crust, either all the way through or just varying the ice thickness. Tides also generate rotational torque that (as discussed below) tends to maintain non-synchronous rotation, which is especially plausible for Europa given that tidal heating makes it unlikely that there are

frozen-in asymmetries capable of locking onto the direction of Jupiter. Non-synchronous rotation (though very nearly synchronous) also may add important components to the tidal stress as the direction of Jupiter varies all the way around relative to the body of the satellite. Over the long term, tidal deformation of the resonantly coupled Galilean satellites contributes to orbital change, which in turn modifies orbital eccentricities, which changes the amplitudes of periodic (orbit driven) tides.

All of these tidal effects, especially rotation (Section III.A), stress (Section III.B), and heat (Section III.C), not only explain much of what we observe on the surface, but they also provide conditions and rates of change that may be able to support survival and evolution of life.

If Europa were in a circular orbit, rotating synchronously (with one face always toward Jupiter), there would be no variable tide. In general, even if a satellite's orbit were eccentric, tides would quickly circularize it, and then, if it rotated non-synchronously, tides would quickly torque it to synchroneity. If that were the case for Europa, it would be a heavily cratered, geologically inactive body.

However, the orbit of Europa is forced to remain eccentric by an orbital reesonance with other satellites, with enormous consequences for tidal effects and the geology and geophysics. The resonance involves the orbits of Io, Europa and Ganymede (Laplace 1805). Europa's orbital period (and also its day) is about 3.6 days long. The orbital periods of these three satellites are locked into an orbital ratio (or "commensurability") close to 1:2:4. Thus conjunction of Io and Europa (where they line up on one side of Jupiter) is locked to the direction of Europa's orbital apocenter (furthest point from Jupiter in its eccentric orbit). Moreover, conjunction of Europa and Ganymede is always aligned with Europa's pericenter (as reviewed by Greenberg 1982, 1989).

Because of the periodicity of these geometrical configurations, these relationships enhance the mutual gravitational perturbations among the satellites. The effect is to maintain the alignments of conjunction, and to pump the eccentricities of the three satellites. The forced eccentricity has a value that depends on exactly how close the ratio of orbital periods is to the whole number ratio. The closer the system is to the exact commensurability, the greater are the forced eccentricities of the satellites.

The origin of the resonance itself may be a result of tidal effects that modified orbital periods until the system became locked in resonance. More likely, the resonance formed as orbits evolved during the formation of the Jovian system, and then evolved significantly under the influence

of tides. For a complete discussion, see Greenberg (2005). However the resonance formed, it currently drives the significant eccentricity of Europa.

The orbital eccentricity causes tides on Europa to change periodically over the orbital period of 3.6 days. Over the course of each orbit, the position of Jupiter relative to Europa not only gets closer and farther, but it also advances and regresses in Europa-centric longitude. The tidal elongation of the figure of Europa thus changes both in magnitude and in direction. The change in orientation of the elongation of Europa does not represent rotation of the body, but rather a "remolding" of the figure of Europa in response to the changing direction of Jupiter. Those periodic tidal variations are responsible for the character of Europa's surface. They stress the crust causing cracks, they provide frictional heat for melting the ice, they induce rotational torque yielding slightly non-synchronous rotation and modifying stress patterns, and they drive long-term orbital change. These effects are only operative because the resonance forces the orbit of Europa to remain eccentric.

15.3.1 Tidal Torque: Rotation Effects

The lag in the tidal response to Jupiter creates a torque on Europa's rotation which, averaged over each orbit, tends to drive the rotation to a rate slightly different from synchronous. That rate depends on details of the tidal response, but according to theory is expected to be only slightly faster than synchronous (Greenberg and Weidenschilling 1984).

Even given Europa's orbital e, the rotation might be synchronous if a non-spherically symmetric, frozen-in density distribution (like that of the Earth's Moon) were locked to the direction of Jupiter. Given that Europa is substantially heated by tidal friction (see below), it may not be able to support such a frozen-in asymmetry. It is also conceivable that the silicate interior is locked to the direction of Jupiter by a mass asymmetry, while the ice crust, uncoupled from the silicate by an intervening liquid water layer, rotates non-synchronously due to the tidal torque.

Even a completely solid Europa would rotate non-synchronously as long as mass asymmetries are small enough. Non-synchronous rotation would not in itself imply the existence of an ocean. However, both the existence of an ocean and non-synchronous rotation are made possible by the substantial tidal heating.

Observational evidence places some constraints on the actual rotation

rate. A comparison of Europa's orientation during the Voyager 2 encounter with that observed 18 yr later by Galileo showed no detectable deviation from synchroneity. Hoppa et al. (1999c) found that any deviation must be small (as predicted), with a period $> 12,000$ yr, relative to the direction of Jupiter.

In principle, further evidence could come from changes in the orientations of cracking over time, as terrain moved west-to-east through the theoretical tidal stress field. The possibility that the observed tectonic patterns might contain some record of such reorientation was noted by Helfenstein and Parmentier (1985) and McEwen (1986). Studies of cross-cutting sequences of lineaments from early Galileo spacecraft images suggested just such a variation in the azimuths over time (Geissler et al. 1998).

However, more recent studies using higher resolution images and exploring several regions suggest that the systematic result was an artifact of an incomplete data set (Sarid et al. 2004, 2005). The sequence of azimuth formation is not continuous and can only be consistent with non-synchronous rotation if one or two cracks (at most) form during each cycle of rotation. That result is reasonable, because once one crack forms in a given region, it would relieve further tidal stress until it has rotated into a different stress regime. This change over time is important in the evolution of the hypothetical ecosystem discussed in Section V.

Other lines of evidence do suggest non-synchronous rotation. As the tectonic effects of tidal stress began to be better understood, various features were found to have likely formed further west than their current positions, including strike-slip faults (Hoppa et al. 1999b, 2000) and cycloidal crack patterns (see Section III.B.1). From the latter, Hoppa et al. (2001b) inferred a non-synchronous period of ¡250,000 yr, but ¿12,000 yr based on the Voyager-Galileo comparison discussed above. There is also similar evidence from strike-slip faults for substantial "polar wander" within the past few million years, in which the ice shell, uncoupled from the interior, was reoriented relative to the spin axis (Sarid et al. 2002). Supporting evidence for polar wander comes from asymmetries in distributions of small pits, uplifts, and chaotic terrain (Greenberg 2003, 2005).

15.3.2 Tidal Stress: Tectonic Effects

Crack patterns Global scale lineament patterns (e.g. those in Fig. 1) correlate roughly with tidal stress patterns (i.e. major global and regional scale lineaments are generally orthogonal to directions of maximum tension), showing that the lineaments result from cracking (e.g. Greenberg et al. 1998). Because the morphology of these lineaments involves large scale ridge systems, we infer that ridges result from cracks.

A distinct and ubiquitous shape for cracks and ridges on Europa are the chains of arcs known as cycloids, first discovered by Voyager (Fig. 6) (Smith et al., 1979, Lucchitta and Soderblom, 1982). Many are marked by double ridges; cycloidal cracks that have not yet developed ridges are also common; many dilational bands appear to have initiated as arcuate cracks; even the major strike-slip fault Astypalaea appears to have begun as a chain of arcuate cracks. Typically these features have arcuate segments 100 km long and the chains run for 1000 km.

This distinctive style of cracking likely occurs as a result of the periodic changes in tidal stress due to the orbital eccentricity (Hoppa et al. 1999a). Suppose a given crack initiates at a time when the tidal tension exceeds the local strength of the ice. It will then propagate across the surface perpendicular to the local tidal tension. On a timescale of hours, as europa moves in its orbit, the orientation of the stress varies. Hence the crack propagates in an arc until the stress falls below a critical value necessary to continue cracking. Crack propagation comes to a halt. A few hours later, the tension returns to a high enough value, now in a different direction, so crack propagation resumes at an angle to the direction at which it had stopped. Thus a series of arcs is created (each corresponding to one orbit's worth of propagation), with cusps between them.

The characteristics of observed arcuate features can be reproduced theoretically, using strength values plausible for bulk ice (105 Pa) and a speed of propagation of a few km/hr. Because each arc segment represents propagation during one Europan orbit, a typical cycloidal crack must have formed in about one (terrestrial) month. In general, the orientation and curvature of the cycloidal chain as a whole is determined by the location at which the crack formed. Thus one can determine the formation location by fitting the model to those characteristics; Such modeling (Hoppa et al. 2001b, Hurford et al. 2006) constrains Europa's rotation (Section III.A).

In all cases a subsurface ocean was required in order to have adequate

tidal amplitude to form the observed cycloidal geometries. If there were no ocean, and the tidal amplitude were correspondingly small, cracking would require the ice to be even weaker. In that case, cycloids' shapes would be distinctly different from what is seen on Europa. Thus the presence of cycloidal lineaments, and their correlation with theoretical characteristics, provided the first widely convincing evidence for the hypothesized liquid water ocean (Hoppa et al. 1999a). Because many cycloids formed during the past few million years, they suggest that the ocean still exists. More recently, studies of Europa's effect on the Jovian magnetosphere provided corroborating evidence for a current liquid ocean (Khurana et al. 1998, Zimmer et al. 2000).

Tidally active cracks Once a crack forms, it relieves the tidal stress in its vicinity. Further cracking is unlikely for as long as the crack remains "active", that is until it anneals so that it can support transverse tension again. During the time that the crack is active and unannealed, the periodic tidal distortion of the satellite's icy shell results in a working of the crack. That is, depending on time, location, and crack orientation, there may be a regular opening and closing of the crack and/or a periodic shearing along the crack. The repeated opening and closing would pump ice and slush from the ocean to the surface, creating the ubiquitous double ridges and their modified forms. Moreover, such working appears to have also driven strike-slip displacement along many cracks on Europa, which in turn indicates that cracks penetrate all the way down to the liquid ocean, a strong constraint on ice thickness. The next two sections describe how the working of active cracks may build double ridges and drive strike-slip displacement.

Ridge Formation The regular opening and closing of a crack can build a ridge by a process of tidal extrusion (Greenberg et al. 1998), which operates as illustrated in Fig. 9: As the tides open cracks (Fig. 9a), water flows to the float line, where it boils in the vacuum and freezes owing to the cold. As the walls of the crack close a few hours later (Fig. 9b), a slurry of crushed ice, slush, and water is squeezed to the surface and deposited on both sides of the crack (Fig. 9c,d). Given the frequency of the process, ridges of typical size (100 m high and 1 km wide) can grow quickly, in as little as 20,000 years. Identification of a mechanism for ridge formation that is so fast is important, because many generations of Europan ridges have formed over the past 10^8 years

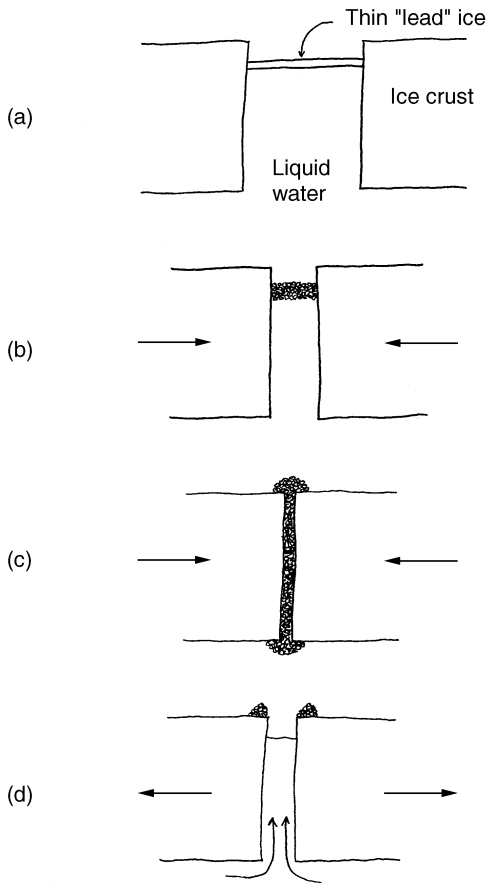


Fig. 15.9. A schematic of the diurnal steps of ridge-building as described in the text.

(the maximum age of the surface on the basis of the paucity of craters); Also, each ridge probably had to form well within one rotational period.

Thus ridges appear to form by a process in which active cracks are bathed in liquid water (including whatever impurities the ocean contains) on a daily basis. Any localized heating due to friction at these cracks would further maintain this process. Fagents et al. (2000) has proposed that the darkening that flanks major ridge systems may be related to heat along the lineament, allowing impurities in the ice to be

concentrated at the surface. Given the exposure of liquid water during the ridge-formation process described above, the dark material flanking major ridge systems is plausibly associated with oceanic substances. This material might have been emplaced at the margins of ridges as oceanic fluid spread through porous ridges (Greenberg et al. 1998), or geyser-like plumes might have sprayed the surface during the daily exposure of liquid water to the vacuum (Kadel et al., 1998).

Another model of ridge formation assumes that long linear diapirs (low-density blobs in the viscous crust) have risen from below the surface, tilting it upward along the sides of cracks (Head et al. 1999). That model was predicated on hypothetical solid-state diapirism in a tens-of-kilometers thick layer of ice, and assumes uniform upwelling over hundreds of km for each major ridge pair, which seems implausible. Moreover, the linear diapir model is inconsistent with the observed properties of ridges in a variety of ways (e.g. Greenberg 2005).

Variations on the theme of double ridge formation are common (Greenberg et al. 1998, Greenberg 2005). The multiple-ridge complexes that compose the center stripes of many of the long triple bands probably formed as parallel, lateral cracks due to the weight of earlier ridges on the thin ice. These lateral cracks then become activated by tidal working and built their own ridges. Other ridges are wider and symmetrically lineated along their length, and are found where the adjacent terrain has separated somewhat. The dilation may have been due to incomplete extrusion of solid ice during daily crack closure (e.g. Fig. 10c), which left jammed ice to gradually pry the cracks open.

Strike-Slip Displacement Strike-slip displacement (e.g. Fig. 7) is common and widely distributed, often with long extent and large offset displacement (Schenk and McKinnon 1989, Tufts 1998, Tufts et al. 1999, Hoppa et al. 1999b, Hoppa et al. 2000, Sarid et al. 2002). Examples include a 170-km-long fault in the far north with > 80 km of shear offset, and a long, bent cycloidal crack whose shear offset indicates that a cohesive plate > 400 km across rotated by about $1°$. The 40 km shear offset fault along 800-km-long Astypalaea Linea near the south pole (Tufts et al. 1999, see Figs. 6 and 7) was originally cycloid-shaped with double ridges. Under shear, ridges on opposite sides of the cracks moved in opposite directions, like trains passing on separate tracks. Those parts of the original crack that were oblique to the shear direction were pulled apart, yielding parallelograms of in-filled material, presumably slush from below. These sites show in detail the structure at a loca-

tion where we know the crust was pulled apart, an important point of reference for interpretation of dilational bands (Section II.B.2).

Tidal stress in the Astypalaea region drives the shear displacement by a mechanism analogous to walking (Tufts, 1998; Hoppa et al., 1999b). Over each orbital period, the tide goes through a sequence starting with tensile stress across the fault, followed 21 hours later by right-lateral shear, followed 21 hours later by compression across the fault, followed 21 hours later by left-lateral shear. Because the left-lateral shear stress occurs after the fault is compressed, friction at the crack may resist displacement, while right-lateral stress occurs immediately after the crack is opened by tension. Thus tides drive shear in the right-lateral sense in a ratcheting process. The mechanism is similar to walking, where an animal repeatedly separates a foot from the ground (analogous to a crack opening), moves it forward (analogous to shear displacement), compresses it to the ground (analogous to compressive closing of a crack), and pushes it backward (analogous to the reverse shear phase), resulting in forward motion. On Europa this process moves plates of crust.

Surveys of strike-slip offsets over much of Europa show that they fit the predictions of the tidal-walking theory quite well (Hoppa et al. 2000, Sarid et al. 2002). Furthermore, the fit was even better if one takes into account some nonsynchronous rotation and at least one polar wander event. Another result of the surveys of strike-slip was identification of large-plate displacements that showed where surface convergence had taken place (Sarid et al. 2002). The latter result is important because it shows how the surface-area budget may be balanced, even given the large amount of new surface created where cracks have dilated (Section II.B.2).

The success of tidal walking in explaining the observed faulting argues strongly that the decoupling layer over which the surface plates slide is a fluid. Penetration to such a layer is required for the daily steps in the "tidal walking" model, because a fluid can deform as necessary on the short timescale of the orbit-driven tides. A thick, ductile, warm ice layer may in principle allow lateral displacement above it (Schenk and McKinnon 1989, Golombek and Banerdt 1990), but no studies have yet shown that tidal walking could be viable in that case.

Thus strike-slip requires that cracks penetrate through most of the crust, which in turn requires that the ice be quite thin, because the tidal tensile stress available for creation of the cracks is not high, only about 40 kPa. Such low tensile stresses are overwhelmed by the compressive hydrostatic overburden pressure at a depth of only 100m. These cracks

could be driven to greater depths by the insertion of liquid water or in-falling solid material. Cracks may also go to greater depth because additional stress is concentrated at the tip of the crack. It is unlikely that cracks could penetrate more than a few km on Europa. Thus our model of strike-slip displacement implies that Europa's ice crust overlies liquid water, that cracks readily penetrate from the surface to the liquid ocean, and therefore the ice must be fairly thin.

Furthermore, Hoppa et al. (1999b) noted that strike-slip offsets are along ridge pairs, not simple cracks, implying that ridges form along cracks that penetrate all the way down to liquid water. This result is consistent with the evidence that ridges form as a result of the working of cracks that link the ocean to the surface (Section III.B.2.a).

The observations and theory of strike-slip displacement have helped constrain the physical character of Europa in ways that are critical to the potential setting for life, including evidence for non-synchronous rotation, for polar wander, for penetration of cracks to the ocean, for significant crust displacement driven by the tidal walking process, and for zones of surface convergence.

15.3.3 Tidal Heating

Liquid water ocean with an ice crust As the figure of a satellite is distorted periodically due to tides, friction may generate substantial heat. Even with assumptions about unknown parameter values (especially the dissipation parameter Q) that would tend to minimize heating, tidal friction could maintain a liquid water ocean (e.g. Cassen et al. 1979, Squyres et al. 1983), but only if one already exists: Without a global ocean, the tidal amplitude would be inadequate to generate the heat to create one. However, the range of plausible parameters could admit much higher heating rates: With plausible material parameters, the tidal dissipation rate could be $> 6 \times 10^{12}$ W, or > 0.2 W/m2 at the surface (O'Brien et al. 2002). In these estimates, the bulk of the heating is in the rocky mantle under the ocean. Geissler et al. (2001) noted that with partial melting of silicate, dissipation could be as much as 9×10^{12} W, or > 0.3 W/m^2 at the surface.

Even though such a total heat flux is very high from the point of view of geophysical implications, it is below the current limit of detectability, swamped out by the reradiated solar energy (Spencer et al. 1999).

A heat flux > 0.2 W/m^2 would imply a steady-state conductive ice layer thinner than 3 km. Convection is probably not possible in such

thin ice. On the other hand, the ice could be much thicker and still transport the same amount of heat, if and only if it were convective. Thus either a thin-ice model (thinner than 10 km) with conductive heat transport or a thick ice model (thicker than 20 km) with convective heat transport could be consistent with the plausible heat flux value estimated above. The thickness required for convection depends on unknown material properties. It must be greater than 20 km (McKinnon 1999, Wang and Stevenson 2000), but convection may require ice so thick that a liquid ocean would be precluded (Spohn and Schubert 2001), a result that favors the thin conducting ice layer.

Thermal transport models may still be too dependent on uncertain parameters for them to definitively discriminate between thin conductive or thick convective ice. Even if the relevant parameters were known, both possibilities might provide a stable steady state. In that case, Europa's actual current physical condition may depend on its thermal history. If the heating rate has increased toward its current value, the ice would likely be thick and convecting. If the heating rate has decreased from significantly higher rates, then the current ice would more likely be thin and conducting. Because tidal heating is driven by orbital eccentricity, the history of Europa's orbit may be crucial.

Formation of Chaotic Terrain Chaotic terrain (Section II.C) has the appearance of sites of melt-through from below, consistent with the thin crust that we infer from the tidal-tectonic model (Greenberg et al. 1998, 1999). The characteristics of chaos, surveyed by Riley et al. (2000), including the rafts, the relations with preexisting ridges, the types of shoreline, the apparent topography, and the association with dark material all are explained by melting from below, so that a lake of liquid water is exposed for some finite time before refreezing (Greenberg et al., 1999). Moreover, morphological similarities between chaotic terrain and craters that appear to have punctured the crust (Section IV.B) reinforce the argument that both represent penetration to liquid.

Dark halos seen around chaotic terrain are also consistent with our interpretations of both ridge formation and chaos formation as processes involving exposure of liquid water. The dark halos are analogous to the diffuse dark margins along major ridge systems (yielding the "triple band" appearance at low resolution). The dark material is spectrally indistinguishable in the two cases (e.g. Fanale et al. 1999). It is consistent in both cases for the dark material to have been brought up from

the ocean and deposited around sites of temporary oceanic exposure, especially given that it includes hydrated salts (McCord 1998a). (The darkening agents are unknown, but may include various organic and sulfur compounds.)

Only modest, temporary concentrations of tidal heat are required for substantial melt-through. For example, if a relatively modest 1

The modest concentrations needed for such melt-through might originate at volcanic vents. Thermal plumes within the ocean could keep the heat localized as it rises to the ice (Thomson and Delaney, 2001, Goodman et al. 2004).

Study of chaotic terrain implies that melt-through occurs at various times and places. It follows that the thickness of ice is quite variable with time and place. Chaos formation has been common, currently covering nearly half the surface (Riley et al. 2000), and continual, displaying a wide range of degrees of degradation by subsequent cracking and ridge formation (Greenberg et al. 1999). The sizes of chaos patches range from over 1000 km across down to as least as small as recognizable on available images, 1 km across. Persistent claims that chaos is relatively young are based on an observational artifact: Old or small examples are harder to see. In fact, over the surface age, formation of areas of chaotic terrain has been frequent and continually interleaved with tectonics (Riley et al. 2006), so at any given location melt-through has occurred at least every few million years.

15.4 Critique of the isolated-ocean model

As described above, the observed characteristics of Europa's surface have been explained as results of tidal processes, and a consistent implication is that connection between the liquid ocean and the surface is common and essential to resurfacing, whether by tectonics or melting. On the other hand, a belief that the ice is so thick that surface processes are independent of a liquid ocean has been actively promoted. If Europa's ocean were actually isolated in that way, the prospects for life would be considerably diminished. Oceanic life would be stifled by the lack of access to oxidants and light, and the frozen crust would lack hospitable openings (Section V) if the ocean were isolated from the surface.

Because impermeable ice would have such major implications for astrobiology, it is important to evaluate the evidence that has been presented for that isolated-ocean hypothesis, specifically interpretations of

(a) a putative class of features called "pits, spots, and domes", (b) the morphologies of craters, and (c) topographic models.

"Pits, spots, and domes" The taxonomy of the features called pits, spots, and domes was never well-defined, and indeed the interpretation was based on a preliminary qualitative interpretation of the appearance of a selected area from an early Galileo image sequence (Pappalardo et al. 1998). Consideration of the examples of these features presented in the literature shows them to be a mix of various types: patches of chaotic terrain, small topographic pits, and isolated uplifts. The features in this ill-defined class were widely described as rounded in shape, with cracks common across the domes, and a typical size of about 10 km across, and a regular spacing. Based on that description, an interpretation was advocated in which the features represent the tops of thermal upwellings in a convective layer of solid ice at least 20 km thick (Pappalardo et al. 1998).

The main problem with that model was that the true characteristics of the types of features cited do not match the qualitative impressions based on the early images. A more complete survey (Greenberg et al. 2003) found no evidence for regular spacing of these features, nor is there any evidence for a typical size, nor for prevalence among uplifts of round domical shapes or summit cracks. The "pits, spots and domes" argument was largely an artifact of over reliance on initial qualitative impressions and over-generalization.

Crater morphologies More recently, isolated-ocean advocates have cited crater characteristics as evidence for their model. On solid bodies, craters that are not simple bowls are usually classified as "complex" (including central peaks, flat floors, and terraced rims), and are generally understood to result from rebound during the later portion of the crater-formation event in solid material. On Europa craters wider than about 5 km tend to have complex interior structure, including various styles of lumpy interior morphologies (Moore et al. 2001). Given the crater-morphology taxonomy developed for solid bodies, Schenk (2002) classified all craters ranging from 4 km to 30 km diameter as "central-peak craters". However, that work was based on problematic assumptions specifically that such a taxonomy is appropriate for an oceanic moon and complex craters on Europa actually form in the same way as those on solid bodies.

In order to form central peak craters by the standard mechanism

for solid bodies, the ice on Europa would need to be thicker than 12 km to prevent interaction with the ocean (Turtle and Pierazzo 2001, Turtle and Ivanov 2002). These results come from simulations that give great insight into the processes of crater formation in ice, and provide quantitative information that is essential to understanding the impact record in the outer solar system, but they would only constrain the thickness of Europa's ice if the morphology of the actual craters was known to have formed that way.

In fact, most, if not all, of the craters in questions have an appearance at least as similar to lumpy chaotic terrain as to solid-body central-peak craters. If that is the case, the character of these craters may result from formation in ice thinner than a few km, rather than thicker than 15 km. Thus craters may represent another mechanism (in addition to tectonics and melt-through) by which the ocean is linked to the surface. Even if a few of the 20 km craters did form without reaching liquid, they would confirm the expected variability of ice thickness, rather than demonstrate that the ocean is isolated.

Hence, the crater record is at least consistent with the larger picture of a surface governed by oceanic linkages. Moreover, penetrating impacts may provide an additional mechanism that occasionally supplements tectonic and thermal processes in linking the ocean to the surface. For a detailed discussion of cratering in this context, see Greenberg (2005).

Topography The more recent argument for the isolated ocean is based on calculations of the variations in surface elevation 1 km over short distances (Prockter and Schenk 2005). The argument is that, if the ice were thin enough to be permeable, it could not support such extreme surface variations. Furthermore, features such as chaotic terrain are identified as upwellings from the solid-state convection, and topographic highs have been reported in that type of terrain (Schenk and Pappalardo 2004, Nimmo and Giese 2005).

Elevations have been quantified by two methods: stereo imagery taken from different angles at different times by Galileo, and photoclinometry, in which the apparent brightness at any point is interpreted as being due to the angle of the surface relative to the illumination and the viewing angle. Both of these methods are susceptible to a variety of systematic and random errors, but the published elevation profiles have not included quantitative discussion of the uncertainties. For stereo, for example, it would be important to understand the errors introduced by using images taken under differing illumination geometries at different

resolutions. For photoclinometry, one concern would be possible correlation of albedo with terrain. Thus for example, the darkening around chaotic terrain could confound the assumption of uniform albedo essential to the method. Thus, any conclusions about the ice thickness based on surface elevation results must await a credible analysis of the potential effects of such sources of errors.

Finally we emphasize that isolated examples of large topographic variations would not necessarily imply that the ice is everywhere too thick to be permeable. Most evidence indicates that permeable ice is the norm for the predominant terrain types on Europa.

15.5 The permeable crust: Conditions for a Europan biosphere

The two major terrain-forming processes on Europa are (1) melt-through, creating chaos, and (2) tectonic processes of cracking and subsequent ridge formation, dilation, and strike slip. These two major processes have continually destroyed preexisting terrain, depending on whether local or regional heat concentration was adequate for large-scale melt-through or small enough for refreezing and continuation of tectonism. The processes that create both chaotic and tectonic terrains (and probably cratering as well) all include transport or exposure of oceanic water through the crust to the surface.

Whether life was able to begin on Europa, or exists there now, remains unknown, but the physical conditions seem propitious (e.g. Greenberg 2005). Change likely occurs over various timescales, which may provide reasonable stability for life, while also driving adaptation and evolution.

Cracks are formed in the crust due to tidal stress and many penetrate from the surface down to the liquid water ocean. The cracks are subsequently opened and closed with the orbit-driven tidal working of the body. Thus, on a timescale of days, water flows up to the float line during the hours of opening, and is squeezed out during the hours of closing. Slush and crushed ice is forced to the surface, while most of the water flows back into the ocean. The regular, periodic tidal flow transports substances and heat between the surface and the ocean.

At the surface, oxidants are continually produced by disequilibrium processes such as photolysis by solar ultraviolet radiation, and especially radiolysis by charged particles. Significant reservoirs of oxygen have been spectrally detected on Europa's surface in the form of H_2O_2, H_2SO_4, and CO_2 (Carlson et al. 1999; McCord et al. 1998b), while

molecular oxygen and ozone are inferred on the basis of Europa's oxygen atmosphere (Hall et al., 1995) and the detection of these compounds on other icy satellites (Calvin et al., 1996; Noll et al., 1996). Moreover, impact of cometary material should also provide a source of organic materials and other fuels at the surface, such as those detected on the other icy satellites (McCord et al., 1998c). In addition, significant quantities of sulfur and other materials may be continually ejected and transported from Io to Europa.

The ocean likely contains endogenic substances such as salt, sulfur compounds, and organics (e.g., Kargel et al. 2000), as well as surface materials that may be transported through the ice. Oceanic substances most likely have been exposed as the orange-brown darkening along major ridge systems and around chaotic terrain. While the coloration displayed in images taken at visible-to-near-IR wavelengths is not diagnostic of composition, near-IR spectra are indicative of frozen brines (McCord et al., 1998a), as well as sulfuric acid and related compounds (Carlson et al. 1999). The orangish brown appearance at visible wavelengths may be consistent with organics, sulfur compounds, or other unknowns. The ocean likely contains a wide range of biologically important substances.

The ice at any location may contain layers of oceanic substances, which are deposited at the top during ridge formation, and then work their way deeper as the ice maintains thickness by melt at the bottom (until a melt-through event resets the entire thickness as a single layer of refrozen ocean). This process can bury surface materials, eventually feeding (or recycling) them into the ocean. For oxidants, burial is especially important to prevent their destruction at the surface, and impact gardening may help with the initial burial (Phillips and Chyba 2001).

Chemical disequilibrium among materials at various levels in a crack is maintained by production at the top and the oceanic reservoir at the bottom, while the ebb and flow of water continually transports and mixes these substances vertically during the tidal cycle. Transport is through an ambient temperature gradient 0.10° C/m, from 0° C at the base of the crust to about 170° C colder at the surface.

The physical conditions in such an opening in the ice might well support life as illustrated schematically in Fig. 10. No organisms could survive near the surface, where bombardment by energetic charged particles in the Jovian magnetosphere would disrupt organic molecules (Varnes and Jakosky, 1999) within 1 cm of the surface. Nevertheless, sunlight adequate for photosynthesis could penetrate a few meters, farther than

necessary to protect organisms from radiation damage (Reynolds et al. 1983, Lunine and Lorenz 1997, Chyba 2000). Thus, as long as some part of the ecosystem of the crack occupies the appropriate depth, it may be able to exploit photosynthesis. Such organisms might benefit from anchoring themselves at an appropriate depth where they might photosynthesize, although they would also need to survive the part of the day when the tide drains away and temperatures drop. Other non-photosynthesizing organisms might anchor themselves at other depths, and exploit the passing daily flow. Their hold would be precarious, as the liquid water could melt their anchorage away. Alternatively, some might be plated over by newly frozen water, and frozen into the wall. The individuals that are not anchored, or that lose their anchorage, would go with the tidal flow. Organisms adapted to holding onto the walls might try to reattach their anchors. Others might be adapted to exploiting movement along with the mixing flow. A substantial fraction of that population would be squeezed into the ocean each day, and then flow up again with the next tide.

A given crack is likely active over thousands of years, because rotation is nearly synchronous and the crack remains in the same tidal-strain regime, allowing for a degree of stability for any ecosystem or organisms within it. Over longer times, with non-synchronous rotation a given site moves to a substantially different tidal-strain regime in 10^3 - 10^5 years. Then the tidal working of any particular crack is likely to cease. The crack would seal closed, freezing any immobile organisms within it, while some portion of its organism population might be locked out of the crack in the ocean below.

For the population of a deactivated crack to survive, (a) it must have adequate mobility to find its way to a still active (generally more recently created) crack, or else or in addition (b) the portion of the population that is frozen into the ice must be able to survive until subsequently released by a thaw. At any given location, a melt event probably has occurred every few million years, liberating frozen organisms to float free and perhaps find their way into a habitable niche. Alternatively, in the timescale for non-synchronous rotation (less than a million years), fresh cracks through the region would cross the paths of the older refrozen cracks, liberating organisms into a niche similar to where they had lived before. Survival in a frozen state for the requisite few million years seems plausible, given evidence for similar survival in Antarctic ice (Priscu et al. 1999). The need to survive change may provide a driver for adaptation and mobility, as well as opportunity for evolution.

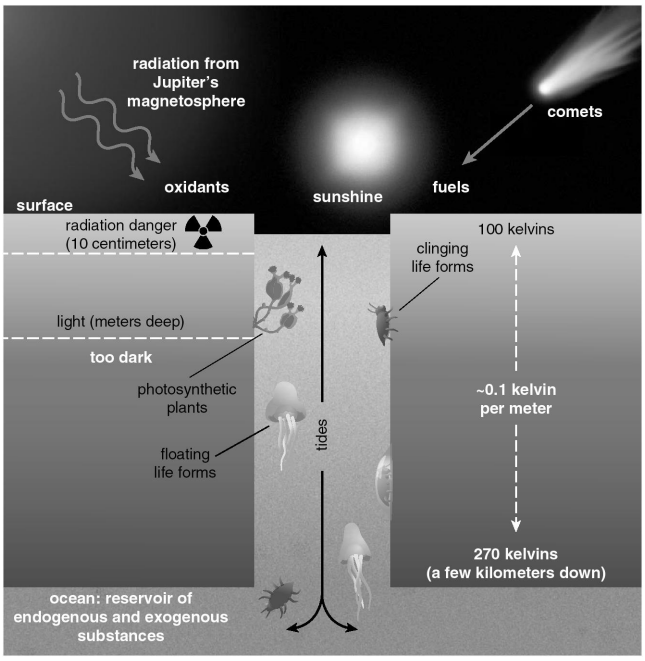


Fig. 15.10. Tidal flow though a working crack provides a potentially habitable setting, linking the surface (with its low temperature, radiation-produced oxidants, cometary organic fuels, and sunlight) with the ocean (with its brew of endo- and exogenic substances and relative warmth). Photosynthetic organisms (represented here by the tulip icon) might anchor themselves to exploit the zone between the surface radiation danger and the deeper darkness. Other organisms (the tick icon) might hold onto the side to exploit the flow of water and the disequilibrium chemistry. The hold would be difficult, with melting releasing some of these creatures into the flow, and with freezing plating others into the wall of the crack. Other organisms (jellyfish icon) might exploit the tides by riding with the flow. This setting would turn hostile after a few thousand years as Europa rotates relative to Jupiter, the local tidal stress changes, and the crack freezes shut. Organisms would need to have evolved strategies for survival by hibernating in the ice or moving elsewhere through the ocean. (Artwork by Barbara Aulicino/*American Scientist*.)

We have shown that this model creates environments in the crust that may be suitable for life. Moreover, it provides a way for life to exist and prosper in the ocean as well, by providing access to necessary oxidants (Gaidos et al. 1999, Chyba and Phillips 2001) and linkage between oceanic and intra-crust ecosystems. Oceanic life would be part of the

same ecosystem as organisms in the crust. Components of the ecosystem might adapt to exploit suboceanic conditions, such as possible sites of volcanism. If there is an inhabited biosphere on Europa, it most likely extends from within the ocean up to the surface. While we can only speculate on conditions within the ocean, we have observational evidence for conditions in the crust, and the evidence points toward a potentially habitable setting.

This model is based on the surface manifestations of tectonic processing and chaotic terrain formation. The entire surface is very young, < 2

Spacecraft exploration of Europa is likely to continue in the future. With the likelihood that the liquid ocean is linked to the surface in multiple ways, Europa's biosphere may be exposed at the surface, facilitating exploration and also contamination. Unless exploration is planned very carefully, we may discover life on Europa that we had inadvertently planted there ourselves. The advantage from the point of view of exploration is that, if landing sites are chosen wisely, it may not be necessary to drill down to the ocean in order to sample the deep. Oceanic materials, possibly including organisms, may be readily accessible at or near the surface. Thus, the search for life in that habitable zone may be less daunting than has been assumed in the past.

Acknowledgments

This chapter reviews a body of evidence produced by many people. Most of the data come from the Galileo mission. For help developing the interpretation presented here, I thank Paul Geissler, Greg Hoppa, Terry Hurford, Dave O'Brien, Jeannie Riley, Alyssa Sarid, and the late Randy Tufts. The images were processed at the Planetary Image Research Laboratory of the University of Arizona. This work was supported by a grant from NASA's Outer Planets Research program.

References

[515] Anderson, J.D., et al.(1998). et al. Europa's differentiated internal structure 281, 2019–2022.

[516] Calvin, W. M., R. E. Johnson, and J. R. Spencer. O2 on Ganymede: Spectral characteristics and plasma formation mechanisms, Geophys. Res. Lett.(1996). R. E. Johnson Geophys. Res. Lett. 23, 673–676.

[517] Carlson, R. W., R. E. Johnson, and M. S. Anderson, Sulfuric acid on Europa and the radiolytic sulfur cycle.(1999). R. E. Johnson Sulfuric acid on Europa and the radiolytic sulfur cycle. Science 286, 97–99.

[518] Carr, M.H., et al., Evidence for a subsurface ocean on Europa.(1998). et al. 391, 363–365.
[519] Cassen, P., R.T. Reynolds, and S.J. Peale, Is there liquid water on Europa? Geophys. Res. Lett.(1979). and S.J. Peale 6, 731–734.
[521] Chyba, C.F.(2000). C.F. Nature 403, 381–382.
[521] Chyba, C.F., and C.B. Phillips.(2001). C.F. Lunar and Planetary Science XXXII (abstract 2140), CD–ROM.
[522] Fagents, S.A., R. Greeley, R.J. Sullivan, R.T. Pappalardo, and L.M. Prockter.(2000). triple band margins Icarus 144, 54–88.
[523] Fanale, F.P., et al.(1999). et al. Icarus 139, 179–188.
[524] Figueredo, P., and Greeley, R.(287). and Greeley 2004, Resurfacing history of Europa from pole–to–pole geological mapping: Icarus 167.
[525] Gaidos, E.J., K.H. Nealson, and J.L. Kirschvink, Life in ice-covered oceans.(1999). and J.L. Kirschvink 284, 1631–1633.
[527] Geissler, P., et al.(1998). Evidence for non–synchronous rotation of Europa 391, 368–370.
[527] Geissler, P., et al., Silicate volcanism on Europa, Lunar and Planetary Sci.(2001). P. Silicate volcanism on Europa, Lunar and Planetary Sci. Conference XXXII.
[528] Goodman, J.C., et al. Hydrothermal plume dynamics on Europa: Implications for chaos formation, J. Geophys. Res.(2004). Goodman et al. Hydrothermal plume dynamics on Europa: Implications for chaos formation, J. Geophys. Res. – Planets 109.
[529] Greeley, R., et al.(2000). 105 559 – 22,578.
[536] Greenberg, R., Orbital evolution of the galilean satellites, in The Satellites of Jupiter, edited by D. Morrison, pp.(1982). edited by D. Morrison University of Arizona Press, Tucson.
[536] Greenberg, R., Time-varying orbits and tidal heating of the Galilean satellites, in Time-Variable Phenomena in the Jovian System, edited by M.J.S. Belton, R.A. West, and J. Rahe, NASA Special Publication No.(1989). edited by M.J.S. Belton and J. Rahe, NASA Special Publication No. 494.
[536] Greenberg, R.(2005). R. the Ocean Moon, Springer– Praxis books.
[536] Greenberg, R., et al.(1998). mechanical response Icarus 135, 64–78.
[536] Greenberg, R., and S.J.(1984). and S.J. Weidenschilling 58, 186–196.
[536] Greenberg, R., et al.(1999). et al. Chaos on Europa 141, 263–286.
[536] Greenberg, R., et al.(2003). R. Icarus 161, 102–126.
[537] Golombek, M.P., and W.(1990). and W.B Banerdt Icarus 83, 441–452.
[538] Head, J. W., R.T. Pappalardo, and R.J. Sullivan, Europa: Morphologic characteristics of ridges and triple bands from Galileo data (E4 and E6) and assessment of a linear diapirism model, J. Geophys. Res.(1999). 104 223–24,236.
[539] Helfenstein, P., and E. M.(1985). Patterns of fracture and tidal stresses due to nonsynchronous rotation: implications for fracturing on Europa 61, 175–184.
[545] Hoppa, G.V., et al., Formation of cycloidal features on Europa.(999a). G.V. Formation of cycloidal features on Europa. Science 285, 1899–1902.
[545] Hoppa, G.V., et al.(999b). Strike–slip faults on Europa: Global shear patterns driven by tidal stress 141, 287–298.
[545] Hoppa, G.V., et al.(999c). Rotation of Europa: Constraints from terminator and limb positions 137, 341–347.

[545] Hoppa, G.V., et al., Distribution of strike-slip faults on Europa, J. Geophys. Res.(2000). J. Geophys. Res.–Planets E9, 22617–22627.
[545] Hoppa, G.V., et al.(001a). et al. Icarus 151, 181–189.
[545] Hoppa, G.V., et al.(001b). et al. Icarus 153, 208–213.
[546] Hubbard, G.S.(2005). Astrobiology: The First Decade Fukuoka, Japan.
[547] Hurford, T.A., et al.(2006). Hurford et al. Cycloidal cracks on Europa: Improved modeling and rotation effects, Icarus submitted.
[548] Kadel, S. D., et al.(1998). S. D. Lunar and Planetary Science Conference XXIX, 1070.
[549] Kargel, J.S., J.Z. Kaye, J.W. Head, G.M. Marion, R. Sassen, J.K. Crowley, O.P. Ballesteros, S.A. Grant, D.L.(2000). Composition Icarus 148, 226–265.
[550] Kasting, J.F., et al.(108). 108. J.F., et al. 1993 Icarus 101.
[551] Khurana, K.K., et al.(1998). Induced magnetic fields as evidence for sub–surface oceans in Europa and Callisto 395, 777–780.
[552] Laplace, P.S., Mecanique Celeste, vol. 4, Courcier, Paris, 1805 (Translation by N. Bowditch, reprinted 1966 by Chelsea, New York.(ork.). Courcier 1805 (Translation by N. Bowditch, reprinted 1966 by Chelsea.
[553] Lipps, J.H., and S.(2005). and S. Rieboldt Icarus 177, 515–527.
[554] Lucchitta, B. K., and L. A. Soderblom, The geology of Europa, in The Satellites of Jupiter, edited by D. Morrison, pp.(1982). edited by D. Morrison University of Arizona Press, Tucson.
[555] Lunine, J. I., and R. D.(1997). and R. D. Lorenz Lunar and Planetary Science Conference XXVIII, 855–856.
[556] Menou, K. and Tabachnik, S.(473). S. 2003 ApJ, 583.
[557] McEwen, A. S.(1986). Tidal reorientation and the fracturing of Jupiter's moon Europa 321, 49–51.
[560] McCord, T.B., et al.(998a). Salts on Europa's surface detected by Galileo's Near–Infrared Mapping Spectrometer 280, 1242–1245.
[560] McCord, T.B., et al. Non-water-ice constituents in the surface material of the icy galilean satellites from the Galileo near-infrared mapping spectrometer investigation, Jour. Geophys. Res.(998b). T.B. Jour. Geophys. Res. – Planets 103, 8603–8623.
[560] McCord, T.B., G.B. Hansen, R.N. Clark, P.D. Martin, C.A. Hibbitts, F.P. Fanale, J. C. Granahan, M. Segura, D.L. Matson, T. V. Johnson, R.W. Carlson, W.D. Smythe G. E. Danielson. Organics and other molecules in the surfaces of Callisto and Ganymede.(998c). R.W. Carlson 278, 271–275.
[561] Moore, J. M., et al.(2001). Impact features on Europa: Results of the Galileo Europa mission 151, 93–111.
[562] Nimmo, F., and Giese, B.(-340). B. Thermal and topographic tests of Europa chaos formation models from Galileo E15 observations, Icarus 177.
[563] Noll, K. S., R. E. Johnson, A. L. Lane, D. L. Domingue, and H. A. Weaver, Detection of ozone on Ganymede.(1996). D. L. Domingue Detection of ozone on Ganymede. Science 273, 341 343.
[564] O'Brien, D.P., P. Geissler, and R.(2002). and R. Greenberg Icarus 156, 152–161.
[566] Pappalardo, R. T., et al.(1998). Geological evidence for solid–state convection in Europa's ice shell 391, 365–368.
[566] Pappalardo, R.T., and R.J.(1996). Evidence for separation across a gray

band on Europa 123, 557–567.
[567] Phillips, C.B., and C.F. Chyba. Impact gardening rates on Europa.(2001). Phillips and C.F. Chyba. Impact gardening rates on Europa. Lunar and Planetary Science XXXII (abstract 2111), CD–ROM.
[568] Pilcher, C. B., Ridgeway, S. T. and McCord, T. B.(1972). Galilean satellites: Identification of water frost 178, 1087–1089.
[569] Priscu et al.(1999). Geomicrobiology of subglacial ice above Lake Vostok Science 286, 2141–2144.
[571] Prockter, L.M., et al.(1999). Europa: Stratigraphy and geological history of the anti–Jovian region 16531 – 16,540.
[571] Prockter, L., and Schenk, P.(-326). P. an anomalous young depression on Europa, Icarus 177.
[573] Riley, J., et al., Distribution of chaos on Europa, J. Geophys. Res.(2000). J. Geophys. Res.–Planets E9, 22599–22615.
[573] Riley, J., et al.(tted). et al. Europa's south pole region: A sequential reconstruction of surface modification processes, GSA Bulletin.
[576] Sarid, A.R., R. Greenberg, G.V. Hoppa, T.A. Hurford, B.R. Tufts, and P.(2002). and P. Geissler Icarus 158, 24–41.
[576] Sarid, A.R., et al.(2004). Crack azimuths on Europa: Time sequence in the southern leading face 168, 144–157.
[576] Sarid, A.R., et al.(2005). A.R. Icarus 173, 469–479.
[579] Schenk, P.(2002). Schenk Nature 417, 419–421.
[579] Schenk, P., and W.B.(1989). Fault offsets and lateral crustal movement on Europa: Evidence for a mobile ice shell 79, 75–100.
[579] Schenk, P.M., and R.T. Pappalardo, 2004, Topographic variations in chaos on Europa: Implications for diapiric formation. Geophys. Res. Lett. 31, 10.(6703). and R.T. Pappalardo Topographic variations in chaos on Europa: Implications for diapiric formation. Geophys. Res. Lett. 31, 10.1029/2004GL019978.
[580] Smith, B. A., et al.(1979). The Galilean satellites and Jupiter: Voyager 2 imaging science results 206, 927–950.
[581] Spencer, J. R., L. K. Tamppari, T. Z. Martin, and L. D.(1999). Temperatures on Europa from Galileo PPR: Nighttime thermal anomalies 284, 1514–1516.
[582] Spohn, T., and G. Schubert. Internal oceans of the galilean satellites of Jupiter, Jupiter Conference, Boulder, Colo.(2001). and G. Schubert. Internal oceans of the galilean satellites of Jupiter Boulder, Colo..
[583] Squyres, S. W., et al.(1983). Liquid water and active resurfacing on Europa 301, 225–226.
[584] Thomson, R.E., and J.R. Delaney, Evidence for a weakly stratified Europan ocean sustained by seafloor heat flux, J. Geophys. Res.(2001). J. Geophys. Res. 106 355–12,365.
[587] Tufts, B. R., Lithospheric displacement features on Europa and their interpretation, PhD thesis, University of Arizona, Tucson, 288 p.(1998). PhD thesis Tucson, 288 p..
[587] Tufts, B.R., et al.(1999). et al. Icarus 141, 53–64.
[587] Tufts, B.R., et al.(2000). et al. Icarus 146, 75–97.
[589] Turtle, E.P., and E.(2001). Thickness of a Europan ice shell from impact crater simulations 294, 1326–1328.
[589] Turtle, E.P., and B.A. Ivanov.(2002). Turtle and B.A. Ivanov. Numerical simulations of impact crater excavation and collapse on Europa: Impli-

cations for ice thickness, Lunar and Planetary Science XXXIII (abstract 1431).
[590] Reynolds, R.T., et al.(1983). On the habitability of Europa 56, 246–254.
[591] Raymond, S.N., and Barnes R.(549). Raymond and Barnes R. 2005 Predicting Planets in Known Planetary Systems II: Testing for Saturn Mass Planets, ApJ 619.
[592] Varnes, E.S., and B.M. Jakosky, Lifetime of organic molecules at the surface of Europa.(1999). E.S. Lifetime of organic molecules at the surface of Europa. Lunar and Planetary Science XXX (abstract 1082), CD–ROM.
[593] Ward, P.D., and Brownlee, D., Rare Earth, Copernicus Books, N.Y.(2000). D. Copernicus Books, N.Y..
[594] Zahnle, K.L., et al.(2003). et al. Icarus 163, 263–289.
[595] Zimmer, C., K.K. Khurana, M.G.(2000). M.G. Kivelson Icarus 147, 329–347.