1.17
Planet Formation

J. E. Chambers

The SETI Institute, Mountain View, CA, USA

1.17.1 THE OBSERVATIONAL EVIDENCE

Modern theories for the origin of the planets are based on observations of the solar system and star-forming regions elsewhere in the galaxy, together with the results of numerical models. Some key observations are:

- The solar system contains eight large planets with roughly circular, coplanar orbits lying 0.4–30 AU from the Sun. There are few locations between the planets where additional large objects could exist on stable orbits.
- The major planets are grouped: small volatile-poor planets lie close to the Sun, with large volatile-rich planets further out. The main asteroid belt (2–4 AU from the Sun) is substantially depleted in mass with respect to other regions.
- The planets and asteroids are depleted in volatile elements compared to the Sun. The degree of fractionation decreases with distance: the terrestrial planets and inner-belt asteroids are highly depleted in volatiles, the outer-belt asteroids are less so, while many satellites in the outer solar...
system are ice rich. Primitive CI meteorites (probably from the outer asteroid belt) have elemental abundances very similar to the Sun except for highly volatile elements.

- Ancient solid surfaces throughout the solar system are covered in impact craters (e.g., the Moon, Mercury, Mars, Callisto). Most of the planets have large axial tilts with respect to their orbits. Earth possesses a large companion with a mass $\sim 1\%$ that of the planet itself.

- The terrestrial planets and many asteroids have undergone differentiation. There is strong evidence that Saturn is highly centrally condensed, with a core of mass $\sim 10M_\oplus$, and weaker evidence that Jupiter has a core of similar mass. These cores have masses comparable to Uranus and Neptune.

- Meteorites from the main asteroid belt show evidence that they once contained short-lived radioactive isotopes with half-lives $<10$ Myr. The main components of primitive meteorites (chondrules and refractory inclusions) have sizes clustered around 1 mm. These components appear to have undergone rapid melting and cooling.

- Young stars generally exist in gas- and dust-rich environments. Many young stars possess massive, optically thick disks with diameters of $10-1,000$ AU. These disks are inferred to have lifetimes of $\sim 1-10$ Myr.

- At least 4% of main sequence (ordinary) stars have planetary-mass companions. The companions have masses of $0.1-10$ Jupiter masses (the lower limit is the current detection threshold), and orbital distances from 0.05 AU to 5 AU (the upper limit is the current detection threshold).

These observations have led to the development and refinement of a theory in which the planets formed from a disk-shaped protoplanetary nebula (Laplace) by pairwise accretion of small solid bodies (Safranov, 1969). A variant of the standard model invokes the gravitational collapse of portions of this disk to form gas giant planets directly. It should be pointed out that the standard model is designed to explain the planets observed in the solar system. Attempts to account for planetary systems recently discovered orbiting other stars suggest that planet formation is likely to differ in several respects from one system to another.

1.17.2 THE PROTOPLANETARY NEBULA AND THE FIRST SOLIDS

1.17.2.1 Circumstellar Disks

The solar system probably formed from the collapse of a fragment of a molecular cloud—a cold, dense portion of the interstellar medium containing gas and dust with a temperature of 10–20 K (Taylor, 2001). Collapse may have occurred spontaneously or been triggered externally, for example, by a supernova (e.g., Cameron, 1962). As the cloud fragment collapsed, the bulk of its mass fell to the center to form a protostar, while the remaining material formed a rotationally supported disk destined to become the protoplanetary nebula (see Chapter 1.04).

Roughly half of young "T Tauri" stars with ages $<10$ Myr are observed to have optically thick disks of gas and dust with masses of $0.001-1M_\odot$ (Beckwith et al., 1990; Strom, 1994). These disks have spectra containing absorption features caused by the presence of water ice and silicates. Ultraviolet and visible emission lines indicate that the central stars are accreting mass from their disks at rates of $10^{-7}-10^{-9}M_\odot$ yr$^{-1}$ (Hartmann et al., 1998). Optically thick circumstellar disks are not observed around stars older than $\sim 10$ Myr (Strom, 1995), which provides an approximate upper limit for the lifetime of the Sun's protoplanetary nebula.

The protoplanetary nebula initially had a mass of at least $0.01M_\odot$. This "minimum mass" is obtained by estimating the total mass of rocky and icy material in all the planets, and adding hydrogen and helium to give a nebula of solar composition (Weidenschilling, 1977a). However, planet formation is probably an inefficient process, suggesting that the protoplanetary nebula was initially more massive than this.

Stars typically form in clusters. If the Sun formed in a large group, such as the Trapezium cluster in Orion, it is likely that one or more massive "OB stars" would have been present. These stars produce large amounts of ultraviolet radiation, which rapidly erodes the outer parts of nearby circumstellar disks by photoevaporation. Smaller clusters, such as the Taurus star-forming region, generally do not contain OB stars. Ultraviolet radiation from the Sun, together with any nearby external sources, would have slowly photoevaporated gas in the protoplanetary nebula beyond $\sim 10$ AU from the Sun, and this gas may have been removed entirely within $\sim 10$ Myr (Hollenbach et al., 2000).

1.17.2.2 Viscous Accretion and Nebula Evolution

The finite lifetime of circumstellar disks, and the fact that young stars are observed to be accreting material, has given rise to a model which views the protoplanetary nebula as a viscous accretion disk in which material was transported radially inwards, ultimately falling onto the Sun (Lynden-Bell and Pringle, 1974). As a consequence of this accretion, the mass of the nebula declined viscosity in Chapter 1.04, ($\sim 0.1M_\odot$), then been accreted by Laughlin to flow inwards transported out disk ($\ll 1$ AU), been partially, interaction in generated ma transported m (Hollenbach et al., 1991). The ab elsewhere in accretion remains for most of th.

The magnetic in disk gave rise material in jet lines out of it, material lost in the mass accretion entrained in the disk, poss et al., 1988).

1.17.2.3 Disk Frac

Currently, the maximum Sun's protop
The approximat
ary material i at some point AU of the $S$ addition, man the Sun preferentially survived to co

stant condensation gradual escay bodies (Hummel, the observed result of rapi
temperature of the outer regions would have been susceptible to gravitational instabilities (e.g., Laughlin and Bodenheimer, 1994), causing mass to flow inwards and angular momentum to be transported outwards. The innermost part of the disk (<1 AU) and the surface layers would have been partially ionized, and in these regions interaction with the Sun's rotating magnetic field generated magnetorotational instabilities, which transported material inwards (Balbus and Hawley, 1991). The absence of viable sources of viscosity elsewhere in the nebula suggests that viscous accretion may have been modest in these regions for most of the disk's lifetime.

Magnetic interactions between the Sun and the disk gave rise to powerful winds which ejected material in jets directed along open magnetic field lines out of the plane of the disk. The amount of material lost in this wind was possibly 30–50% of the mass accreted by the Sun. Some solid particles entrained in the wind would have fallen back into the disk, possibly many AU from the Sun (Shu et al., 1988).

1.17.2.3 Disk Temperatures and Chemical Fractionation

Currently, there is considerable debate about the maximum temperatures reached in the Sun's protoplanetary disk (see Chapter 1.20). The approximate isotopic homogeneity of planetary material in the inner solar system argues that, at some point, most of the material within a few AU of the Sun was vaporized and mixed. In addition, many primitive meteorites are progressively depleted in volatile and moderately volatile elements with respect to the Sun (Palme et al., 1988), independent of cosmochemical behavior (lithophile, siderophile, chalcophile). This fractionation can be understood if elements more volatile than magnesium and iron were initially vaporized in the inner part of the disk. As the nebula cooled and refractory material began to condense, the gaseous phase continued to accrete onto the Sun. Hence, volatile elements were preferentially lost while refractory elements survived to coagulate into planetary bodies (e.g., Wasson and Chou, 1974; Cassen, 2001). The absence of isotopic fractionation of potassium in meteorites is consistent with the idea that chemical fractionation occurred by preferential condensation of refractory elements rather than gradual escape of volatile elements from solid bodies (Humayun and Clayton, 1995). However, the observed fractionation could also be the result of rapid escape of volatile elements due to collisions (Halliday and Porcelli, 2001), or be caused by the addition of a refractory substance to primitive material of solar composition (Hutchison, 2002).

The disks of mature T Tauri stars (ages > 0.3 Myr) are observed to have surface temperatures of 50–300 K at 1 AU from the star (Beckwith et al., 1990). Temperatures in the midplanes of these disks are likely to be higher, in the range 200–750 K at 1 AU, and hot enough to vaporize silicates within 0.2–0.5 AU (Woolfum and Cassen, 1999). Models of viscous accretion disks yield midplane temperatures that are substantially higher than this early in their history, due to the release of gravitational energy as material flows through the disk towards the central star. These models have disk mass accretion rates ~100 times greater than those observed in mature T Tauri stars (Boss, 1996; Bell et al., 1997), and similar mass accretion rates have been inferred for young stars that are still embedded in envelopes of gas from their molecular cloud (Kenyon et al., 1993). Mass accretion rates this high imply that silicates would have been vaporized out to several AU in the disk for a period of ~10^5 yr. However, the ubiquity of presolar grains in meteorites (identified by their unusual isotopic compositions) implies that at least some material in the inner solar system survived vaporization or that such material was added during the accretion of the planets. The variety of oxygen isotope abundances seen in solar system bodies suggests incomplete mixing of nebula material prior to the formation of the planets and meteorite parent bodies (Clayton, 1993), although iron isotopes do seem to have been homogenized (Zhu et al., 2001).

Disk temperatures would have decreased rapidly with distance from the Sun as accretional energy release, optical depth, and solar radiation all declined. For example, some meteorite samples from main-belt asteroids contain hydrated silicates, formed by reactions between anhydrous rock and water ice. This implies that temperatures at 2–3 AU became low enough for ice to condense while the asteroids were forming.

1.17.2.4 Chondrules and Refractory Inclusions

Chondrites—meteorites from parent bodies in the asteroid belt that never melted—represent the most primitive samples available of material that formed in the protoplanetary nebula. Chondrites are mainly composed of chondrules, with smaller amounts of refractory inclusions and a fine-grained matrix of silicate, metal, and sulfide. Chondrules are roughly spherical objects, typically ~1 mm in size, and largely composed of olivine and pyroxene (Taylor, 2001). They appear to have formed from melt droplets that cooled on timescales of
hours according to their texture (Jones et al., 2000). At least 25% of chondrules show signs that they were melted more than once (Rubin and Krot, 1996). The chondrules and matrix grains tend to have complementary chemical compositions (underabundant elements in one correspond to overabundant elements in the other), suggesting that they originated in the same part of the nebula.

As of early 2000s, it is not known how chondrules formed. The great abundance of chondrules in chondrites (up to 80% by mass; Jones et al., 2000) indicates either that chondrule formation was an efficient process, or that they were preferentially retained in objects that grew larger. The presence of moderately volatile elements (e.g., sulfur and sodium) and unmelted relict grains in chondrules implies that chondrule precursors existed in a cool environment with $T < 650$ K, and that chondrules remained molten for only a few minutes (Connolly et al., 1988). These characteristics suggest that chondrules formed in the protoplanetary nebula rather than within a parent body. Plausible formation mechanisms include melting due to lightning (Desch and Cuzzi, 2000) or shocks in the nebula. Models of shock melting have proved the most successful in terms of reproducing the observed properties of chondrules (Desch and Connolly, 2002). Such shocks may have been generated by large-scale gravitational instabilities in the disk if the nebula was still sufficiently massive when chondrules formed (Desch and Connolly, 2002).

Refractory inclusions (also called calcium-aluminum-rich inclusions or CAIs) are an order of magnitude less abundant than chondrules. These objects contain mostly calcium-aluminum silicates and oxides (Jones et al., 2000), and have undergone melting. However, they appear to have remained molten for longer than chondrules—hours rather than minutes—and cooled less rapidly (Jones et al., 2000). In addition, CAIs have uniform oxygen isotope ratios, suggesting that they all formed in the same region (McKeegan et al., 1998). It is possible that chondrules and CAIs both formed in situ, a few AU from the Sun. Models suggest that CAIs are likely to have formed under a narrower range of nebula conditions than chondrules, which would explain their lower abundance in chondrites (Alexander, 2003).

A second possibility is that CAIs formed elsewhere in the nebula and were subsequently transported to the asteroid belt. In the “X-wind” model, solids in the disk drifted inwards, emerged from a partially shielded environment and were melted by solar radiation (Shu et al., 1996). Some of these objects were entrained in the wind of material flowing away from the Sun, and millimeter-sized particles would have subsequently fallen back to the disk at distances of several AU. It seems unlikely that chondrules formed in an X-wind since their volatile components would have been lost due to prolonged heating, while the smallest observed chondrules should have been carried out of the solar system rather than falling back to the disk (Hutchison, 2002).

1.17.2.5 Short-lived Isotopes

Many CAIs, together with some chondrules and samples of differentiated asteroids, contained short-lived radioactive isotopes at the time they formed. This is deduced from the abundances of the daughter isotopes seen in modern meteorites. The short-lived isotopes include $^{44}$Ca, $^{26}$Al, $^{10}$Be, $^{60}$Fe, $^{53}$Mn, and $^{107}$Pd, with half-lives (in units of Myr) 0.13, 0.7, 1.5, 1.5, 3.7, and 6.5, respectively. Many of these isotopes could have been produced from stable ones by absorption of neutrinos in a supernova or the outer layers of a giant star. In particular, $^{60}$Fe can only be produced efficiently by stellar nucleosynthesis and so must have come from an external source (Shukolyukov and Lugmair, 1993). Conversely, some isotopes such as $^{10}$Be are almost certainly formed in the protoplanetary nebula when material was bombarded by solar cosmic rays (McKeegan et al., 2000). Multiple sources are possible for some short-lived isotopes. The abundances of the decay products of $^{26}$Al and $^{44}$Ca in CAIs from carbonaceous chondrites are correlated, suggesting that these isotopes come from a single stellar source (Sahijpal et al., 1995). Models indicate that no more than ~10% of the $^{26}$Al was produced by cosmic ray bombardment, since otherwise $^{44}$Ca and $^{53}$Mn abundances would be higher than observed (Goswami and Vanhala, 2000). The short half-lives of these isotopes favor a scenario in which they were generated by the same supernova or stellar wind that triggered the collapse of the molecular-cloud fragment that went on to form the solar system (Vanhala and Boss, 2000).

The source of short-lived isotopes is important since if these isotopes were homogeneously mixed in the nebula they could be used as chronometers—the relative ages of materials can be obtained by measuring the abundance of the daughter products of the isotopes. CAIs appear to be the oldest surviving material in the solar system (with ages of $4.566 \pm 0.002$ Gyr, measured using the lead–lead chronometer; Allègre et al., 1995). Most CAIs, for which accurate measurements are available, formed with $^{26}$Al/$^{27}$Al ratios of $(4-5) \times 10^{-5}$. The uniformity of these values suggests that $^{26}$Al was thoroughly mixed at the time CAIs formed, and that they formed within a few hours of another, as FUN nuclear of $^{26}$Al anomalies in the nebula and Russ Chondrites in 2001. Measured abundances lower $t$ interpretation of the two may be lifetime $t$ AU, longer $t$ before collision of their up remnants (Huss et al., 2000).

1.17.3

1.17.3.1

The most prominent feature of section corugula nebula the Sun would hot ph primary sulfides were $c$ water $i$ up to 2 $W$ these $v$ in the s "ice line discont trappin" line (St also li $smear ice line the ne"
few hundreds of thousands of years of one another. A few refractory inclusions, referred to as FU Ori inclusions (fractionated and unidentified nuclear anomalies), had different initial amounts of $^{26}\text{Al}$. These CAIs exhibit several isotopic anomalies, suggesting that they formed before the nebula was homogeneously mixed (Wadhwa and Russell, 2000).

Chondrules formed with a variety of $^{26}\text{Al}/^{27}\text{Al}$ ratios in the range $(0.2) \times 10^{-3}$ (Huss et al., 2001). These abundances are lower than those measured in most CAIs, suggesting that chondrules formed $2-5$ Myr after CAIs. The initial abundances of other short-lived isotopes are also lower in chondrules than CAIs, supporting the interpretation that chondrules are the younger of the two (Alexander et al., 2001). This conclusion may be hard to reconcile with the short gas-drift lifetimes of millimeter-sized bodies ($\sim 10^3$ yr at 1 AU, see Section 1.17.3). It is possible that CAIs were incorporated into larger bodies, with longer gas-drift lifetimes, for a few million years, before being returned to the nebula in disruptive collisions. However, some CAIs show signs of thermal alteration and remelting over periods of up to 2 Myr, implying that these objects remained in the nebula until chondrules formed (Huss et al., 2001).

1.17.3 THE ORIGIN OF PLANETESIMALS

1.17.3.1 Dust Grain Compositions and the Ice Line

The earliest stage of planetary accretion is the most poorly understood at present. If planets did not form directly via instabilities in the disk (see Section 1.17.6), they would have formed by coagulation of solid material in the nebula. In a nebula with a similar elemental composition to the Sun, $\sim 0.5\%$ of the mass in the inner region would have formed solids following any initial hot phase of nebula evolution. These solids primarily consisted of silicates, metal, and sulfides. In the outer nebula, where temperatures were colder, icy materials also condensed, with water ice being the most abundant. In this region, up to $2\%$ of material with a solar composition would have formed solids. The boundary between these two regions was marked by a discontinuity in the surface density of solid material called the "ice line" (or snow line). The magnitude of this discontinuity may have been enhanced by cold trapping of ice as gas was recycled across the ice line (Stevenson and Lunine, 1988). However, it is also likely that radial drifting of solid material smeared out the boundary to some degree. The ice line would have moved inwards over time as the nebula cooled.

Lacking gas pressure support, dust grains, and aggregates would have settled towards the midplane of the nebula, increasing the solid-to-gas ratio in this region. The rate of sedimentation depended on the particle size, with larger objects falling faster than small ones. As dust grains fell, they would have coagulated to form larger objects, increasing the rate of sedimentation. Calculations suggest that much of the solid material would have accumulated near the midplane in $10^3-10^4$ yr (Weidenschilling, 1980).

1.17.3.2 Gravitational Instability

What happened next depends sensitively on the vertical thickness of this solid-rich layer and on the relative velocities of the solid particles. Pioneering studies of planetary accretion showed that if the solid-rich layer was very thin, portions of it would become gravitationally unstable, collapsing to form solid bodies $\sim 1$ km in diameter called "planetesimals" (Safranov, 1969; Goldreich and Ward, 1973). Note that this process is different from the large-scale disk instabilities that may have formed Jupiter-mass bodies.

Later studies suggested that things would not be so simple. Gravitational instability is most likely to occur when the volume density of solid particles is high and their relative velocities are low. However, a dense layer containing mostly solid material would revolve about the Sun with a velocity given by Kepler's third law. In the gas-rich layers above and below the midplane, the Sun's gravity was partially offset by gas pressure, so that the gas disk revolved at less than Keplerian velocity. This differential velocity ($\sim 100$ m s$^{-1}$ for objects at 1 AU; Adachi et al., 1976) generated turbulence which would have stirred up the solid-rich layer, rendering it less prone to gravitational instability (Weidenschilling, 1980). In addition, particles of differing sizes would have drifted through the nebula at different rates due to gas drag. This increased the relative velocities of the particles, also frustrating gravitational instability (Weidenschilling, 1988).

Ward (2001) has suggested that the onset of gravitational instability depends sensitively on disk parameters that are poorly constrained at present. There appears to be a critical solid-to-gas surface density ratio necessary for the onset of gravitational instability, which requires enhancement of solids by a factor of 2–10 times above that expected for material of solar composition (Youdin and Shu, 2002). In a turbulent disk, particles would have migrated at different rates in different regions, leading to a pileup of solid material at certain points (Youdin and Shu, 2002). In addition, particles are likely...
to migrate more slowly through regions where the volume density of particles is high, and this can also lead to a local increase in the solid-to-gas ratio (Goodman and Pindor, 1999). These factors may have favored the formation of planetesimals by gravitational instability.

1.17.3.3 Dust Grain Sticking

If planetesimals did not form via gravitational instability, such bodies must have formed as a result of the sticking together of dust grains and aggregates during collisions. Only when bodies reached ~1 km in size could gravity have played a significant role in further accretion. Some grain growth probably occurred in the Sun’s molecular cloud fragment, forming objects up to ~0.1 mm in size (Weidenschilling and Rudolph, 1996). However, many of these grains would have evaporated subsequently when they entered the protoplanetary disk, at least in the inner solar system.

Grain growth in the disk itself must have occurred quickly in order for the objects to survive against gas drag due to the differential velocity of solid objects relative to the nebula gas. Objects <1 m in size were somewhat coupled to the motion of the gas, reducing the effects of gas drag, but larger bodies had no such protection. These objects drifted towards the Sun as they lost angular momentum to the nebula gas, with drift timescales inversely proportional to their size. Meter-sized objects were especially vulnerable, with drift lifetimes ~100 yr at 1 AU (Weidenschilling, 1977b). Differential drift velocities for objects of different size may have aided the accretion of bodies smaller than 1 km, provided that impact speeds were not too high (Cuzzi et al., 1993).

Experiments suggest that small dust grains will stick together if they collide at velocities of less than a few meters per second (Poppe et al., 2000). The probability of sticking decreases with increasing collision speed. Spherical silica grains stick at collision speeds below ~1 m s⁻¹, while irregularly shaped particles can stick at higher velocities, up to 50 m s⁻¹ (Poppe et al., 2000). Sticking is mainly due to van der Waals forces, which are weak. Coagulation of iron grains in the nebula would have been enhanced by the presence of the Sun’s magnetic field (Beckwith et al., 2000), while in cool regions, frost coatings probably aided sticking of centimeter-to-meter size bodies at low collision speeds (Bridges et al., 1996). Collisions between small particles typically lead to charge exchange and formation of dipoles, and this could have enhanced sticking forces between chondrule-sized particles by a factor of ~4,000 (Marshall and Cuzzi, 2001).

1.17.3.4 Larger Bodies and Turbulent Concentration

It is unclear how aggregation continued for boulder-sized and larger objects. The presence of gas may have helped. This is because small fragments, formed by disruptive impacts onto large bodies, would have been blown back onto the body as the fragments became coupled to the gas, itself moving at less than Keplerian velocity (Wurm et al., 2001). Recent attention has also focused on processes that increased the concentration of solid bodies and may have aided further accretion. If the nebula was turbulent, solid objects would have been concentrated in the convergence zones of eddies. Numerical simulations show that particle concentration could be increased by a factor ~100 within 100 yr (Klahr and Henning, 1997), and much higher concentrations may have occurred on longer timescales (Cuzzi et al., 2001). Turbulent concentration is size dependent, working most effectively for particles trapped in the smallest eddies. For plausible disk viscosities, the mean particle size and size distribution would have been similar to those observed for chondrules (Cuzzi et al., 2001), suggesting that turbulence played an important role in the early stages of accretion if gravitational instability was ineffective.

1.17.4 TERRESTRIAL PLANET FORMATION

1.17.4.1 Random Velocities and Gravitational Focusing

Once solid bodies reached sizes ~1 km, mutual gravitational interactions became significant. Objects of this size are traditionally referred to as planetesimals even if they did not form by gravitational instability. Close passages between planetesimals tended to increase their “random velocities” ν (the radial and out-of-plane components of their motion) as a result of their mutual gravitational attraction. Conversely, drag caused by nebular gas and physical collisions between planetesimals damped their random velocities, making their orbits more circular and coplanar. Competition between these excitation and damping mechanisms established an equilibrium distribution of random velocities, which gradually changed over time as the planetesimals gained mass and their gravitational interactions grew stronger.

The frequency of collisions depended sensitively on the random velocities. When ν was small, a pair of planetesimals undergoing a close encounter would have remained close to each other for some time, allowing their mutual gravity to “focus” their trajectories towards each other, increasing the chance of a collision. When ν was large, the focusing would be suppressed, as the planets would have been too far apart to interact.
large, close encounters were brief affairs rendering gravitational focusing ineffective, thus reducing the collision probability and the growth rate of planetesimals.

This intimate coupling between the planetesimals' masses and random velocities can give rise to three different growth modes, each with quite different characteristics and accretion timescales. It seems likely that each of these growth modes operated at different times in the solar nebula.

1.17.4.2 Runaway Growth

The early stages of accretion were marked by "runaway growth." Over the course of many close encounters, the largest planetesimals tended to acquire the smallest random velocities, a process that is often referred to as "dynamical friction." (A similar equipartition of energy is observed in gas molecules.) As a result, the largest bodies experienced the strongest gravitational focusing of their trajectories, and they grew most rapidly. Most of the solid material remained in smaller planetesimals that grew slowly if at all (Stewart and Kaula, 1980; Kokubo and Ida, 1996).

Initially at least, it seems likely that most collisions led to accretion rather than fragmentation (Leinhardt and Richardson, 2002). However, as the planetesimals grew larger, their random velocities increased, and collisions between small bodies were increasingly likely to result in disruption. Small collision fragments were strongly affected by gas drag, and these rapidly acquired circular, coplanar orbits. As a result, collision fragments were quickly accreted by the largest planetesimals due to their high mutual collision probability, and this allowed growth to proceed more rapidly (Wetherill and Stewart, 1993).

1.17.4.3 Oligarchic Growth

Once the largest objects, dubbed "planetary embryos," become more than ~100 times more massive than a typical planetesimal, the random velocities of the planetesimals became largely determined by gravitational perturbations from the embryos rather than by other planetesimals (Ida and Makino, 1993). As a result, gravitational focusing became less effective and runaway growth slowed down. The larger an embryo, the more it stirred up the velocities of nearby planetesimals, and the slower it grew, allowing neighboring embryos to catch up. This state of affairs is known as "oligarchic growth" (Kokubo and Ida, 1998). At the same time, gravitational interactions between embryos tended to keep them apart, so that each carved out its own niche, or "feeding zone" in the protoplanetary disk.

Numerical simulations of runaway growth suggest that bodies of the size of the Moon or Mars could have formed in ~10^5 yr at 1 AU (Wetherill and Stewart, 1993). However, it is likely that oligarchic growth began at much lower masses than this, in the range 10^-5 - 10^-3 M_Earth (Rafikov, 2003; Thommes et al., 2003). As a result, the formation of lunar-to-Mars size bodies took place in the oligarchic growth regime, requiring ~10^6 yr at 1 AU in a minimum mass nebula (Weidenschilling et al., 1997).

Planetary embryos accreted most of their mass from their feeding zones, which had widths of ~10-12 Hill radii, where the Hill radius RH = (mr/3M_Earth)^1/3, where r is the distance from the Sun (Kokubo and Ida, 1998). Hence, embryos probably had different compositions depending on where they formed in the protoplanetary disk, although radial drift of small bodies due to gas drag means that some radial mixing of material would have occurred.

1.17.4.4 Late-stage Accretion and Giant Impacts

The final stage of accretion began when the remaining planetesimals had too little mass to damp the random velocities of the embryos. As the random velocities of the embryos increased, growth slowed dramatically, and embryos' orbits began to cross those of their neighbors. This growth mode, known as "orderly growth," began when roughly half of the total solid mass was contained in embryos (Kokubo and Ida, 2000). Despite its name, the last stage of accretion was the most violent, marked by giant collisions between bodies the size of the Moon or Mars. It is unclear how efficient these collisions were, although numerical simulations suggest that growth continued even if there was significant fragmentation (Alexander and Agnor, 1998). The remaining planetesimals were swept up or lost, either by falling into the Sun or being ejected from the solar system. Numerical models suggest that the formation of a fully formed Earth required ~100 Myr (Chambers and Wetherill, 1998).

The highly noncircular orbits of embryos and the long accretion timescales allowed considerable radial mixing of material over distances of 0.5-1.0 AU (Wetherill, 1994). It is likely that each of the inner planets accreted material from throughout the inner solar system, although the degree of radial mixing depends sensitively on the mass distribution of the embryos at this time (Chambers, 2001). The relative contributions from each part of the disk would have been different for
each planet, however, producing somewhat different chemical compositions. Earth and Mars have similar but distinct oxygen isotopic compositions, which implies that embryos and planetesimals in the inner solar system were not thoroughly mixed before they accreted to form these planets (Drake and Righter, 2001). Alternatively, Mars may represent a surviving planetary embryo, with a unique chemical and isotopic signature, while the more massive Earth is a composite of many embryos.

Giant impacts were common during the final stages of accretion (Agnor et al., 1999). Several impacts would have been energetic enough to completely melt each of the inner planets, forming a magma ocean which homogenized existing material and erased chemical signatures of earlier stages of accretion. The presence of a massive atmosphere containing captured nebular gas would also have melted the surfaces of Earth and Venus (Sasaki, 1990). Impacts would have been particularly energetic in the innermost part of the disk where orbital, and hence collision, speeds were highest. The high density of Mercury compared to the other inner planets can best be explained as the result of a catastrophic collision occurring after the planet had differentiated (Wetherill, 1988). Numerical simulations suggest that a high-velocity impact onto proto-Mercury would have removed much of the planet's mantle, leaving an intact metal core encaised in a thin layer of silicates (Benz et al., 1988).

1.17.4.5 Core Formation and the Late Veneer

Energy released from impacts, together with heat from the decay of radioactive isotopes, led to differentiation in planetary embryos, once these objects became partially molten (Tonks and Melosh, 1992). Iron and siderophile elements (e.g., platinum, palladium, and gold) preferentially sank to the center to form a core, while the lighter silicates and lithophile elements formed a mantle. Differentiation was probably a continuous process rather than a single event, so that large planets like Earth accreted from embryos that were already partially or wholly differentiated.

The time of core formation on the Earth and Mars can be constrained using chronometers based on short-lived isotopes. The W–Hf isotope system is particularly useful in this respect, consisting of a lithophile parent nucleus, $^{183}$Hf, which decays to a siderophile daughter isotope with a half-life of 9 Myr. Recent measurements of an excess of the daughter isotope indicate that the cores of Mars and Earth formed in less than 13 Myr and 30 Myr, respectively (Kleine et al., 2002; Yin et al., 2002; see Chapter 1.20 by Halliday for a more detailed discussion). This estimate for Mars is consistent with the 4.5 Gyr age for ALH 84001, the oldest known martian meteorite.

The affinity for iron of elements such as palladium and platinum, even at the high pressures present in planetary mantles, means that these elements should be essentially absent from the mantles of the inner planets (Holzheud et al., 2000). The fact that siderophiles are present in Earth's mantle and crust implies that some material was accreted as a "late veneer" when Earth's differentiation was largely complete. This material could have originated in the terrestrial–planet region or the asteroid belt. Earth's $^{187}$Os/$^{188}$Os ratio appears to rule out a late veneer consisting primarily of enstatite or carbonaceous chondrites, leaving ordinary chondrites as the most likely source if the material was predominantly from the asteroid belt (Drake and Righter, 2001).

1.17.4.6 Formation of the Moon

The leading model for the origin of Earth's moon is an oblique impact between proto-Earth and a Mars-sized embryo (Cameron and Ward, 1976). Such an impact would have formed an accretion disk in orbit around Earth, consisting mostly of mantle material from the impactor, while the impactor's core coalesced with that of the Earth (Canup and Asphaug, 2001). The small size of the lunar core, and the Moon's extreme depletion in volatile and moderately volatile elements, arose as the result of its accretion from a hot circumplanetary disk containing mostly silicates. Today, the Earth and the Moon have essentially identical oxygen isotopic compositions, which suggests that the lunar impactor originated in the same region of the Sun's protoplanetary disk as proto-Earth (Wiechert et al., 2001). The fact that the Moon's core has remained small since its formation suggests that the moon-forming impact happened in the closing stages of planetary accretion, since subsequent collisions on the Moon would have increased its metal fraction (Canup and Asphaug, 2001).

The oldest known lunar rocks have ages of 4.4–4.5 Gyr (Carlson and Lugmair, 1988; see Chapter 1.21), and hence the Moon must have formed within the first 150 Myr of the solar system. This is consistent with the age of the oldest known terrestrial samples—zircon grains, which formed in crustal rocks ~4.4 Gyr ago (Wilde et al., 2001). However, the age of Earth's core, as determined using the W–Hf isotope system, suggests that the moon formed significantly earlier than this, since the moon-forming impact would probably have strongly affected the W–Hf chronometer.
1.17.4.7 Terrestrial-planet Volatiles

The presence of water and volatile elements (carbon, nitrogen, and the noble gases) on Earth, Mars, and Venus poses a problem for theories of planet formation. It is likely that the inner part of the protoplanetary nebula was too hot for these materials to condense at the time when planetesimals were forming (see Section 1.17.2). Enstatite chondrites, from undifferentiated bodies in the inner asteroid belt, are very dry, and ordinary chondrites also contain little water (Taylor, 2001). This suggests that planetesimals which formed <2.5 AU from the Sun were almost free of volatiles. If true, this implies that Earth acquired its volatiles by accreting material that originally formed beyond 2.5 AU, in regions of the nebula that were cold enough for ices to condense. In fact, Earth probably acquired more water than currently exists in the oceans and mantle, since some water would have been lost by reacting with iron (Righter and Drake, 1999).

Comets are rich in volatile elements, but they probably delivered no more than 10% of Earth’s volatile inventory. There are several reasons for this. Comets have a very low impact probability with Earth over their dynamical lifetime (~10^-6; Levison et al., 2000), limiting the amount of cometary material that Earth could have accreted. In addition, if most of Earth’s water was acquired from comets, it seems likely that Earth’s noble gas abundances would be higher than observed by several orders of magnitude (Zahnle, 1998). Finally, water measured spectroscopically in comets differs isotopically from that of seawater on Earth, with the cometary D/H ratio being greater by a factor of 2 (Lunine et al., 2000).

An asteroidal source of volatiles is more promising. Carbonaceous chondrites contain up to 10% water by mass, in the form of hydrated minerals (Taylor, 2001), while some ordinary chondrites also contain water. The hydrogen in many chondritic hydrated silicates has a similar D/H ratio to that of Earth’s oceans. Numerical simulations show that Earth could have accreted several oceans worth of water from the asteroid belt, especially if lunar-to-Mars size planetary embryos formed in this region (Morbidelli et al., 2000). However, the amount of mass accreted from the asteroid belt depends sensitively on the early orbital evolutions of the giant planets, and these are poorly known at present (Chambers and Cassen, 2002). Oxygen isotope differences between Earth and carbonaceous chondrites imply that the latter contributed no more than a few percent of Earth’s total mass (Drake and Righter, 2001), although this is enough to supply Earth’s water. In addition, the relatively low abundance of siderophile elements in the terrestrial mantle argues that Earth acquired most of its asteroidal material before core formation was complete.

Mars is depleted in highly volatiles relative to Earth and contains water with a higher D/H ratio. Models suggest that it acquired a significant fraction of its volatiles from both comets and asteroids (Lunine et al., 2002). It is possible that Mars has lost much of its initial volatile inventory during large impacts. Giant impacts (including the moon-forming event) may have removed the early atmospheres of the terrestrial planets, leading to depletion of atmosphere-forming elements with respect to geochemical volatiles such as thalium (Zahnle, 1998). Hence, the modern atmospheres of the inner planets were produced by outgassing of material incorporated into the mantle during accretion (see Chapter 1.20 for a detailed discussion).

1.17.5 THE ASTEROID BELT
1.17.5.1 Formation and Mass Depletion

The primary feature of the main asteroid belt is its great depletion in mass relative to other regions of the planetary system. The present mass of the main belt is ~5 x 10^-4 M⊕, which represents 0.1-0.01% of the solid mass that existed at the time planetesimals were forming. There are several ways the main asteroid belt could have lost most of its primordial mass. Substantial loss by collisional erosion appears to be ruled out by the preservation of asteroid Vesta’s basaltic crust, which formed in the first few million years of the solar system (Davis et al., 1994). More plausible models are based on the existence of orbital “resonances” associated with the giant planets.

Half a dozen strong resonances exist at particular heliocentric distances in the region 2-4 AU from the Sun. The orbit of an asteroid in one of these resonances becomes unstable on a timescale ~1 Myr, such that the asteroid ultimately falls into the Sun or is ejected from the solar system (Gladman et al., 1997). At present, the resonance locations are almost devoid of asteroids. However, resonances currently occupy only a small fraction of the orbital phase space in the main belt, so an additional mechanism must have operated to make them more effective at removing mass in the past. Some of the resonances existed in a different location when the protoplanetary nebula was present. As the nebula dispersed, these resonances swept across the asteroid belt and could have removed a substantial amount of mass in the process (Ward et al., 1976; Nagasawa et al., 2000). However, the clearing efficiency depends sensitively on the timescale for nebula removal, such that rapid nebula dispersal corresponds to less efficient clearing. Prior to the dispersal of the nebula, a combination of gas drag and resonances is...
likely to have caused many asteroids with diameters in the range 10–100 km to drift into the region now occupied by the terrestrial planets (Franklin and Lecar, 2000).

The asteroid belt would have also experienced rapid mass loss if planetary embryos accreted in this region via runaway and oligarchic growth. Close encounters between embryos would have caused frequent changes in their orbits until objects entered a resonance and were removed. Numerical simulations show that the most likely outcome is that all planetary embryos would have been lost from the asteroid region by this process (Chambers and Wetherill, 2001). The same mechanism would have removed ~99% of planetesimals and asteroid-sized bodies at the same time (Petit et al., 2001). At present, it is unclear whether embryos did form in the main belt. Accretion simulations suggest that embryos ought to have formed within \( \sim 10^8 \) yr (Wetherill and Stewart, 1993), provided that Jupiter formed later than this. However, if Jupiter formed rapidly, its gravitational perturbations would have increased the random velocities of planetesimals in the asteroid belt and prevented runaway growth from taking place (Kortenkamp and Wetherill, 2000).

The disruptive effect of Jupiter’s gravity suggests that asteroids accreted before the giant planets were fully formed. The fact that Jupiter and Saturn are mostly composed of nebular gas sets an upper limit of \( \sim 10 \) Myr for asteroid formation, based on estimates for the lifetimes of circumstellar disks (Strom, 1995). Rapid formation of some asteroids is confirmed by isotopic chronometers. The initial abundances of \(^{53}\)Mn and \(^{182}\)Hf in HED (howardite, eucrite, diogenite) meteorites, which probably originated on Vesta, suggest that this asteroid formed and differentiated within 4 Myr of the formation of CAIs (Lugmair and Shukolyukov, 1998; Kleine et al., 2002; Yin et al., 2002). This agrees with a formation time of \( \sim 5 \) Myr derived using the \(^{26}\)Al chronometer (Srinivasan et al., 1999). Similarly, the parent bodies of many iron and stony-iron meteorites and ordinary chondrites formed within \( \sim 10 \) Myr, according to the \(^{107}\)Pd and Hf-W systems (Wadhwa and Russell, 2000; Lee and Halliday, 1996).

1.17.6 GIANT-PLANET FORMATION

1.17.6.1 Giant-planet Compositions

The dominant components of Jupiter and Saturn are hydrogen and helium (see Chapter 1.23). These elements would not have condensed or become trapped in solids at temperatures present in the protoplanetary nebula (Lunine et al., 2000), so they must have been gravitationally captured as gases. Uranus and Neptune contain \( \sim 10\% \) hydrogen and helium, and presumably these gases were captured from the nebula too. Protoplanetary disks orbiting young stars are observed to disperse in \( \leq 10 \) Myr (Strom, 1995), which strongly suggests that the giant planets in the solar system took no more than 10 Myr to form. The gravitational fields of Jupiter and Saturn indicate that they possess dense cores, of unknown composition,
with masses of \( \sim 10M_\oplus \) \((\text{Wuchterl et al., 2000})\). Jupiter’s gaseous envelope is enriched in carbon and sulfur by a factor \( \sim 3 \) relative to the Sun, while nitrogen, argon, krypton, and xenon are also enriched \((\text{Owen et al., 1999})\). Elements heavier than helium are also enriched in Saturn relative to a solar composition.

The fact that Jupiter and Saturn have atmospheres enriched in elements heavier than helium suggests that they accreted a large amount of mass in the form of planetesimals. However, the similar enrichments of highly volatile argon and nitrogen compared to less volatile carbon, observed in Jupiter, imply that these planetesimals formed at temperatures \( \leq 30 \text{ K} \) \((\text{Owen et al., 1999})\). In contrast, temperatures at 5 AU are commonly thought to have been \( \sim 160 \text{ K} \) at the time when Jupiter was forming. This suggests that Jupiter formed much further from the Sun than its present location, which is hard to reconcile with the known models for planet formation, or that it was efficient at accreting planetesimals from the region now occupied by the Edgeworth–Kuiper belt. Alternatively, argon and nitrogen may have been trapped at higher temperatures in planetesimals composed of crystalline rather than amorphous ice \((\text{Gautier et al., 2001})\), in which case Jupiter could have formed and accreted planetesimals at its current location.

1.17.6.2 Core Accretion

As of early 2000s, two models for the origin of Jupiter and Saturn are being actively pursued. In the “core accretion” model, planetary accretion in the outer solar system initially proceeded through the same stages as in the inner solar system: the formation of planetesimals followed by runaway and oligarchic growth. The presence of additional solid material in the form of water ice, and the fact that Hill radii (and hence feeding zone widths) increase with distance from the Sun, means that, in principle, larger bodies would have formed here than in the inner solar system. As they grew, the largest planetary embryos would have acquired thick atmospheres of nebula gas. Numerical models indicate that once an object grew to \( \sim 10M_\oplus \), it could no longer support a static atmosphere, and it steadily accreted gas from the nebula \((\text{Pollack et al., 1996})\), eventually forming a gas-giant planet.

Analytic estimates suggest that \( 10M_\oplus \) cores were unlikely to accrete in a minimum mass nebula on a timescale comparable to the lifetime of circumstellar disks. However, such cores could have formed if the surface density of solids at 5 AU was 5–10 times that of a minimum mass nebula \((\text{Lissauer, 1987})\). Numerical models of planetary accretion support this conclusion \((\text{Thommes et al., 2003})\), although it is possible that growth could have stalled when the largest bodies were a few Earth masses, due to loss of solid material via collisional fragmentation \((\text{Inaba and Wetherill, 2001})\). The surface density of solids at 5 AU may have been greater than that in a minimum mass nebula because the nebula itself was more massive than this, implying that planet formation was inefficient. In addition, solid material could have accumulated near 5 AU, either because of cold trapping of ice \((\text{Stevenson and Lunine, 1988})\) or drift of small solids to a local maximum in the gas surface density \((\text{Haghighipour and Boss, 2003})\).

The rate at which the planetary cores accreted gas increased slowly until their masses reached \( \sim 3M_\oplus \). After this, gas accretion was very rapid \((\text{Pollack et al., 1996})\). The growth timescale depends on the opacity of the planet’s gas envelope, since this determines the rate at which the energy of accretion could be radiated away. For interstellar dust opacities, a \( 10M_\oplus \) core would require \( \sim 10 \text{ Myr} \) to grow to Jupiter’s mass \((\text{Pollack et al., 1996})\). However, growth would have been quicker if the opacity was lower due to coagulation of grains in the envelope \((\text{Ikoma et al., 2000})\).

1.17.6.3 Disk Instability

A protoplanetary nebula containing enough mass to rapidly form a \( 10M_\oplus \) giant-planet core would have been quite close to the limit of gravitational stability beyond \( \sim 5 \text{ AU} \) \((\text{Boss, 2001})\). This suggests that the giant planets may have formed in a single step via “disk instability.” This would occur on a timescale that is orders of magnitude shorter than the timescale for core accretion. Numerical simulations show that a marginally unstable disk gives rise to gravitationally bound clumps comparable to the mass of Jupiter on timescales of \( \sim 100 \text{ yr} \) \((\text{Boss, 2001})\). High-resolution simulations show that these objects would remain bound and continue to collapse for at least 1,000 years \((\text{Mayer et al., 2002})\). The question of whether solids in these clumps would settle to the center to form a core similar to those thought to exist in Jupiter and Saturn has not been explored to date.

At present, it is unclear whether the protoplanetary nebula could have evolved to the point where it was marginally unstable, or whether disk instabilities would have redistributed mass in the disk prior to the formation of gravitationally bound clumps. If marginally unstable disks do develop, then giant-planet formation by disk instability seems unavoidable. It is worth noting that current simulations of disk instability tend to generate planets with masses greater than...
Jupiter and Saturn, but comparable to some observed extrasolar planets.

1.17.6.4 Planetary Migration and Disk Gap Formation

The growth of the giant planets took place while the nebula gas was still present. This complicates the picture of planet formation because of gravitational interactions between the gas and the giant planets. In particular, a planet will strongly interact with the disk at “Lindblad resonances,” i.e., at certain locations where the planet’s gravitational perturbations build up constructively over time. A planet launches spiral density waves so that the distribution of gas in the disk becomes nonaxisymmetric. The gravitational attraction of overdense regions exerts a torque on the planet which changes its orbit. Torques due to gas orbiting inside and outside the planet have opposite signs but unequal magnitudes in general, such that a planet is likely to lose angular momentum and drift inwards (Goldreich and Tremaine, 1980). This is referred to as “type-I migration.” The migration speed is proportional to the mass of the planet, and can be very rapid: a 10M_⊕ core at 5 AU would have migrated into the Sun in ~10^9 yr (Tanaka et al., 2002), while the migration timescale for a Jupiter mass planet is 30 times shorter!

Clearly, type-I migration presents a problem for models of planet formation, both in terms of accreting fully formed planets before they migrate into the Sun and in terms of their survival once fully formed. However, it is likely that a sufficiently massive planet would have cleared a gap in the disk gas. Once this gap extended beyond the Lindblad resonances, type-I migration ceased. At present, there is considerable uncertainty about how massive a planet must be to clear a gap in the disk. This depends sensitively on the way in which waves damped in the nebula, and on the disk viscosity, both of which are poorly constrained. A recent estimate is that a body with a mass of 2–3M_⊕ would have cleared a gap at 1 AU, while at 5 AU, a body ~15M_⊕ would do the job (Rafikov, 2002).

Once a planet clears a gap in the disk, its orbital evolution becomes tied to that of the nebula. In a viscous disk, material would have flowed towards the Sun and a giant planet would have migrated with this material, maintaining a gap in the disk as it moved. This is known as “type-II migration” (Lin and Papaloizou, 1986; Ward, 1997). Type-II migration rates depend on the nebula’s viscosity, and so are poorly constrained at present. For plausible viscosities, the migration timescale would have been at least 1–2 orders of magnitude longer than for type-I migration (Ward and Hahn, 2000). Type-II migration provides a plausible explanation for the small sizes of the orbits of many observed extrasolar planets.

The existence of type-I migration implies that the cores of the giant planets must have accreted quickly if these planets formed by core accretion. It is likely that migration would have speeded up accretion as a giant-planet core drifted into regions containing additional planetesimals. However, simulations suggest that a migrating planet would accrete ~10% of these planetesimals, so the need for a massive nebula remains (Tanaka and Ida, 1999).

The giant planets ceased growing when the flow of gas onto their envelopes was cut off. This may have been the result of gap formation or because the nebula dispersed. The latter seems unlikely, since the timescale for gas accretion onto a Jupiter-size planet is small compared to the lifetime of the nebula. However, hydrodynamical simulations suggest that gas would continue to flow onto Jupiter after it cleared a gap in the disk (Lubow et al., 1999), so this explanation is problematic too. In addition, it has been suggested that some gas would remain at the same orbital distance as the planet after it cleared a gap if the disk viscosity was low (Rafikov, 2002), and this would also be accreted by the planet eventually.

1.17.6.5 Formation of Uranus and Neptune

Attempts to model the accretion of Uranus and Neptune from planetesimals orbiting 20–30 AU from the Sun (the current locations of these planets) have met with severe difficulties. Long orbital periods in the outer solar system mean that accretion occurs very slowly. In addition, solar gravity is sufficiently weak here that gravitational interactions between planetary embryos would have ejected a substantial amount of mass from this region of the disk (Levison and Stewart, 2001). Numerical simulations show that it is unlikely that bodies larger than Earth could have accreted in situ at the locations of Uranus and Neptune, even if the nebula was substantially more massive than the minimum-mass nebula (Thommes et al., 2003).

A more likely scenario is that Uranus and Neptune, along with the cores of Jupiter and Saturn, formed in the region 5–10 AU from the Sun. Such a system would have remained dynamically stable until one object (Jupiter) accreted a large H/He-rich atmosphere. At this point at least two of the other bodies would have been perturbed into the region beyond 15 AU. Gravitational interactions with planetesimals in the outer solar system would then have circularized the orbits of Uranus and Neptune by dynamical friction, while at the same time scattering most of these planetesimals onto
unstable orbits that crossed those of Jupiter and Saturn. Numerical simulations have shown that this model is robust provided that cores with masses $\sim 10 M_\oplus$ were able to survive type-I migration (Thommes et al., 1999). The failure of Uranus and Neptune to capture more than a small amount of nebula gas indicates that the nebula dispersed before they could accumulate massive atmospheres similar to Jupiter and Saturn. This may be a consequence of photoevaporation of nebula gas by ultraviolet radiation from the Sun, since this is most effective at distances beyond $\sim 10$ AU (Hollenbach et al., 2000).

ACKNOWLEDGMENTS

I am most grateful to Robbins Bell, Lindsey Bruesch, Jeff Cuzzi, Jack Lissauer, Ignacio Mosqueira, and Kevin Zahnle for helpful comments and discussion during the preparation of this chapter.

REFERENCES


