4. Origin of the Solar System

How do we arrive at a theory of Solar System formation?

It is impossible to "reverse" the equations of motion, chemistry and physical evolution of a complex system like the solar system to arrive back at its origins unambiguously. The current system is too chaotic, indeterminate and possibly lacking in some of its earliest 'ingredients'. Even the things we know very well have error bars on them or small builtin uncertainties which would build up such a large uncertainty when we tried to trace them back as tomake our conclusions worthless. So the only way to try to work out a viable theory of the origin of the solar system (and there is no guarantee this would be a unique solution) is to:

- Postulate a theory of Formation
- Develop that forward in time
- Make predictions
- Compare those predictions to the current system
- Where there are seen to be differences or inconsistencies go back to the beginning with a new or modified theory.

Theories to date

Theories of the origin of the solar system, universe, earth etc are known generally as **cosmogonic theories**. The origin of the solar system is in many ways linked to the origin of the universe; certainly to that of the local star group and the galaxy, but we r group and the galaxy, but we shall restrict ourselves only to solar system origin, bringing only a few ideas from the "wider picture" where that will help us understand the solar system.

The (serious) theories of the origins of the solar system can be divided into a number of different types, generally. There are co-eval and non-co-eval theories - that is those which postulate the planets' birth taking place at the same time as the central star of the system was forming, and those where the planets were "acquired" later. We shall briefly look at a non-co-eval theory later. The co-eval theories we can further subdivide into nebula and non-nebula theories - those where the planets formed out of the gas/dust nebula which was the birthplace of the star, or otherwise. Within the nebula theories there is a further sub-division into low and high mass theories. The non-nebula theories are currently out of favour, so we shall only look at the two types of nebula theories. These both start from the same basic mixture of dust and gas; they just differ in the details of the development of planetary formation.

What evidence do we judge these theories against?

We have to decide first what sort of information we can judge the theories against. As we go through we shall see there is a plethora of facts about the physics and chemistry of the system against which we can measure these ideas. There are ese ideas. There are distributions of elements, the details of the dynamics, overall compositions etc which all need to be fitted in. But we can start by just looking at some of the basic outline information that we have. We know for instance that the planets can be divided into two main groups - the "terrestrial" planets, small rocky bodies with densities 3-6 times that of water in the inner part of the system, and gaseous, low density bodies in the outer part, divided from the terrestrial planets by the asteroid belt. We also know that the

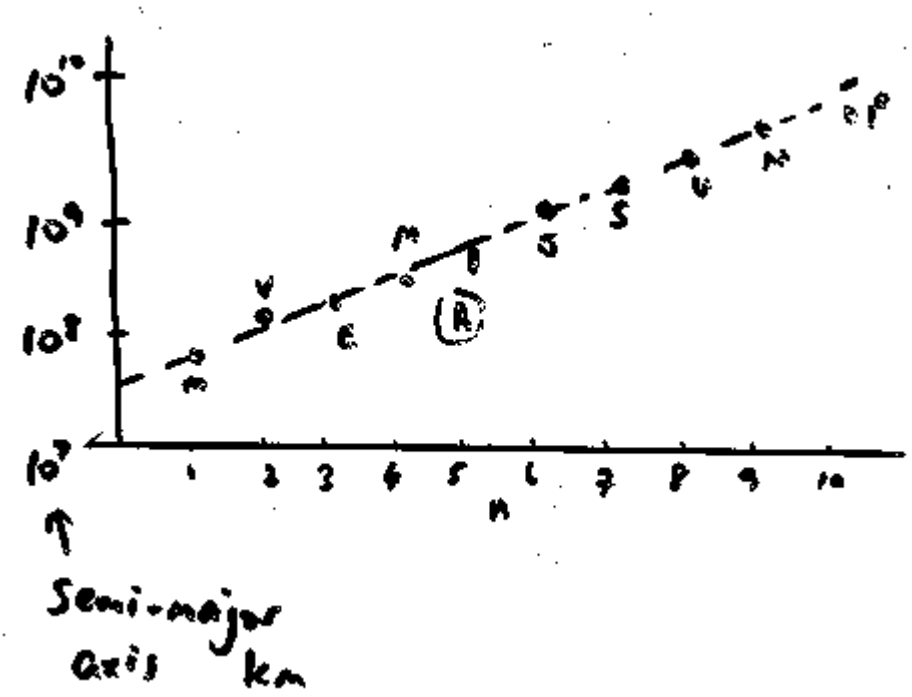
gaseous planets can be further subdivided: there is Jupiter and Saturn, the "Gas Giants" which are roughly the same size (though different in density), and Uranus and Neptune, the "Subgiants"again of similar size. Outside Neptune (and often inside) is the small, icy Pluto with its large companion Charon. This is an interesting exception which may be unimportant in deriving the overall structure of the solar system (as it may be a chance exceptional case) but which must be explained in any comprehensive theory.

We do not have just the major planets to explain, however. Besides the asteroids we must explain the existence and occurrence of comets, and of the "Kuiper belt" bodies which form a large asteroid-like belt outside the Neptune/Pluto region. If the comets come from the "Oort cloud" (see later where we discuss comets) then we also have to explain how that cloud is formed.

There are elements of the morphology of the system which has to be explained: why are the planets all orbiting in or near the same plane - the ecliptic, why all in prograde orbits, and why most spinning in the same direction. (Conversely how important is it to our theory that a planet like Uranus with its spin axis virtually in the ecliptic plane differs from the general rule?) The satellites also generally tend to follow these same trends - orbiting in their primary's equatorial plane, in prograde orbits with prograde spin.

It is important when we examine the theories that we know what are real trends and laws that the planets hold to, and which are spurious or more tenuous. Kepler's Laws are firmly established and "approved" by intimate connection with the basic laws of gravitation. The planetary motions can thus in principle be calculated and extrapolated virtually as far as we like by the accurate application of these principles. However, a good example of what is often called a "Law" but the accuracy of which is really in some doubt is the "Titus-Bode Law" of planetary distance from the sun. This states that the log of the disance from the sun is linear:

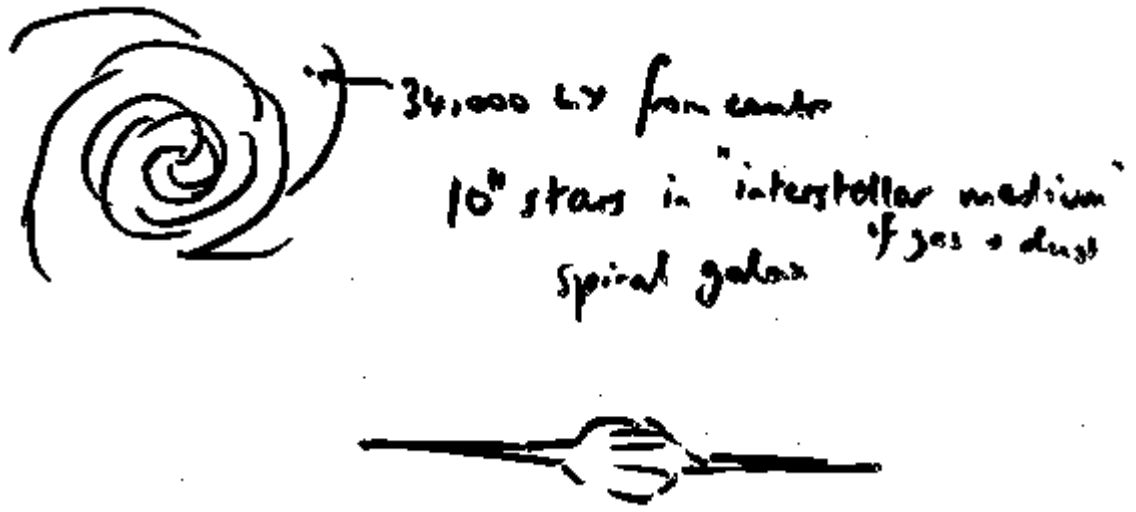
Bode's Law:



When this "law" was first posited we did not know of the existence of planets beyond Saturn. The fact that Uranus and Neptune, when discovered, fitted , when discovered, fitted so well to this curve seemed to give it some credence, though there are some minor discrepancies (in the positions of Mars and Venus) in even the inner part of the system. The discovery of the asteroids - near but generally inside their "Titus-Bode" position - and then of Pluto which is way off the line, cast the theory into doubt, though it could be argued these are explicable exceptions from an otherwise well-maintained rule. It has been pointed out that many satellites of the major planets also seem to follow a similar sort of rule. There have been attempts to explain why planets might tend to form in this sort of spacing, but it is still a matter of disagreement as to whether we should read any significance into this "rule" for the planets' spacing.

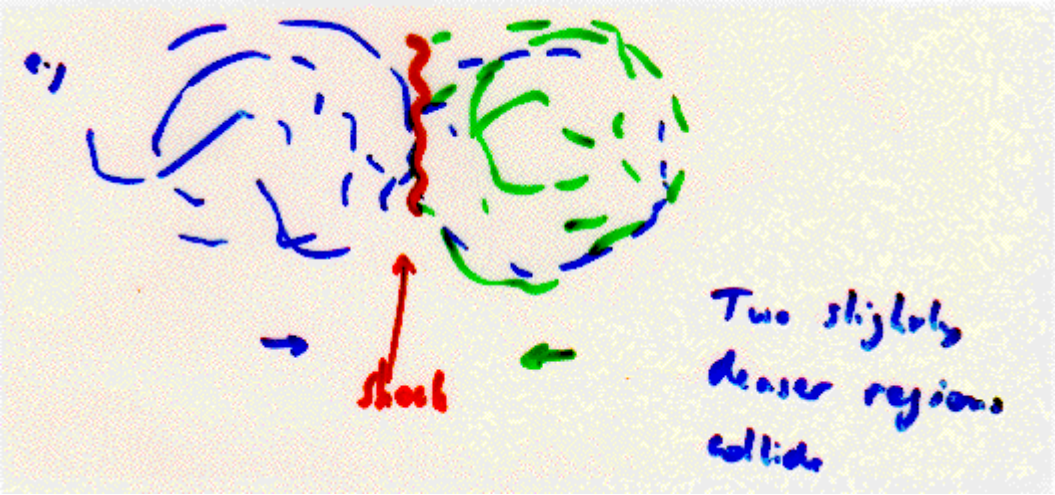
The solar system - location and general information

The solar system is located in the outer reaches of a minor arm of a spiral galaxy which is very similar to many others. "Our" galaxy is 3.2 10^9 AU across (50,000LY - LY = light-year). We are 34,000 LY from the centre. In the galaxy there are 10^{11} stars in an "interstellar medium" of gas and dust.



The gas and dust is thought to be the source of the stars. This gas and dust permeates the galaxy. The mass of this will have a gravitational attraction tending to make the cloud of material coake the cloud of material collapse in on itself, but usually internal forces like spin, electric and magnetic fields and thermal motions prevent this collapse from occurring.

However, there is much chaotic behaviour in the complex gas/dust field and it is possible for regions to get isolated or change their characteristics. Two regions of slightly higher density might collide, for example, and the resultant shock might trigger collapse:



An alternative might be that two sub-critically-sized clouds merge and thence form a body which has too large a density to prevent collapse.

Whatever starts the collapse, as it gathers momentum it will tend to be self-sustaining, as the gravitational force is concentrated by the material moving towards a central point: this collapse gathers pace until a distinct central body is formed - this is the origin of the central star of the system. (Hence planetary system cosmogony is usually linked to stellar system creation.)

Forming stars from molecular Hydrogen clouds

The largest densities and the lowest temperatures are the conditions most likely to produce planets in the clouds of molecular hydrogen which are though to be the birth places of stars and their planetary systems. The theories of formation usually concentrate on the nebulous regions, with the first problem being to explain what triggers the collapse, and the next major one being to explain how you segregate out the individual bodies.

The nebulae are mainly composed of molecular Hydrogen, and are typically from 2 to 300 light years (LY) across, containing from 10's to 10^6 solar masses (M_S) respectively of material. ($1M_S$ = 1.989 x10^{30} kg). The temperature of the hydrogen cloud will be typically a few 10's of K, and the density 10^8 to 10^{12} H_2 particles m^{-2}. This latter may sound a lot but it represents just 0.003 kg in a 10km cube. This nebula will be typically cooler and denser than the rest of the interstellar medium so that it should be in equilibrium position-wise. There is also typically more dust - about 1% of the cloud. It generally also contains heavy elements from previous stellar explosions where complex nucleosynthesis took place. (You need a supernova to get heavier elements than Fe: also to give short-lived isotopes like Al^{26} and Co^{56} which we shall come across later.)

The low temperature, high density is a good condition for leading to a contraction of the cloud. Holding it apart, however, will be thermal motions, and it will be "stiffened" against gravitational collapse also by **B** fields, spin and turbulence. Spin will be the biggest effect. Turbulence can act both ways - both helping to hold a h helping to hold a cloud apart with its chaotic motions, and also aiding compression through shock impact.

Because of its size, the collapsing molecular cloud must fragment, or there would be one immense, very short-lived sun. (Sometimes this happens - large fragments of so-called "prestellar nebulae" with masses around 20 M_{Sun}, collapse to a star, which lasts about 1Ma and then goes nova. In comparison, note that a sol-type sun lasts 10^{10} years.)

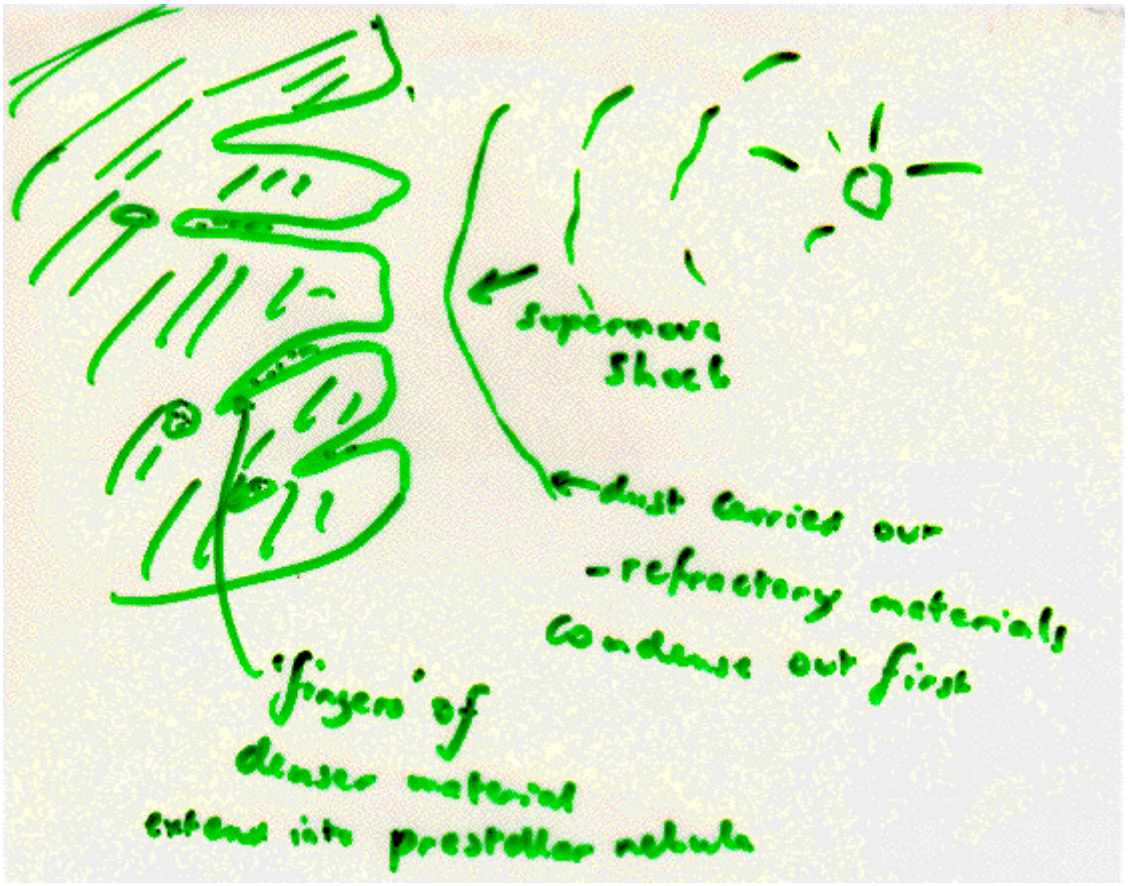
What causes the fragmentation? It could be turbulence, pressure waves or differential cooling. One can calculate that there is a minimum mass (at a given temperature and density) called the Jean's Mass, needed for an individual star to collapse. For a given medium, the larger the mass the more likely the collapse. Hence stars form in preference to planets, and we assume planetary formation follows formation of the sun.

The fact that there is a tendency for fragmentation then stellar formation means that stars often form in "star clusters". There must be a dispersal mechanism though, at least in some cases, as 5-20% of stars are "singles" like the Sun. This is fortunate for planetary formation since it is very difficult to form a stable planetary system with star clusters.

The collapsing cloud becomes denser and denser until it forms a "protostar", a compact infra-red emitting body surrounded by a "cloud". Sometimes the star is visibmes the star is visible at the centre of the cloud, like the T-Tauri stars. These stars are seen to be surrounded by tenuous gas and dust in violent motion (the T-Tauri wind - like a very strong solar wind). A lot of work is currently going in to understanding such systems, but they are still poorly understood. The collapse down into the protostar continues, with a subsequent rise in temperature, until eventually stellar "ignition" takes place and fusion begins. The rapid collapse also starts the system rapidly spinning. The "ignition" of nuclear fusion takes place about 10Ma after maximum luminosity, and by then by most cosmogonic theories, the star probably has a planetary system.

One suggested mechanism for triggering the collapse of some interstellar nebulae, and one that seems may be appropriate for ours, is that a nearby supernove produces a shock wave which "seeds" the collapse. Credence is leant to this theory by the excess of Mg^{26} found in meteorites like the "Allende meterorite". By an "excess" we mean that there is more Mg^{26} than the estimated proportion from the universal abundances - and this is produced by the decay of Al^{26}. This Al isotope is very short lived on a cosmic timescale and so the fact it was there when the meteorite was formed suggests a source of heavy nuclei - presumably a supernova - immediately preceeded the period leading up to condensation out of the first material. Thus supernova and condensation seem linked, suggesting it could easily be that the former triggered the latter. Of course the supernova will also supply heavy elements to the collapsing nebula.

Heavy elements enter the primordial nebula



So far, to this point, there is little to distinguish high and low mass nebula theories. A lot of the difference between them, however, resides in the mechanism for explaining where the angular momentum goes, since all theories of collapse show a very fast spin rate must be built up as the star is formed. To get from that state to our current situation where the sun spins relatively slowly and most of the angular momentum of the solar system resides in the planets (particularly Jupiter) requires a lot of dissipation of the angular momentum to have occured, and it seems only in the early days of formation of the system could this have happened. It is one of the major difficulties of theories of solar system formation to explain where the angular momentum went.

Low Mass Nebulae

Formation from a "low-mass" nebula means that the mass of the cloud which contracts to form the star and planets is around 1.1 times the mass of the star, so in the case of the solar system, 1.1 M_{Sun}. The spin-up as the collapse continues leads to a trade-off between gravitational and spin forces which produces a sheet of material in the spin plane.

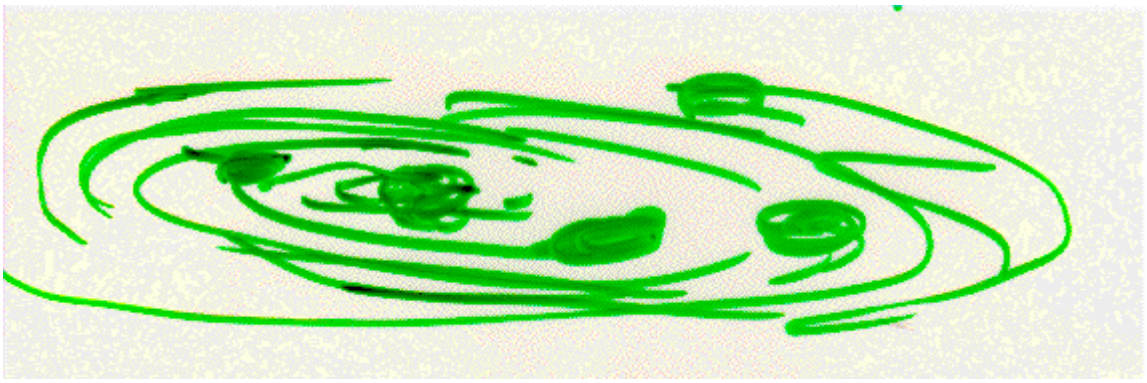
The material settling in to this plane eventually takes up Keplerian orbits, and infra-red cooling leads to further shrinkage. The cooling is partly offset by dust capture. You get some accretion - electrostatic, electromagnetic and collisional mechanisms lead to the formation of planetisimals. The temperature on the fringes of the collapsed disc is 50K but closer in to the centre of mass it can be up to 1000 K. With the higher temperatures you get more mixing of materials.

As the sun heats up it volatilises and "blows away" the volatile material near to it. Within the planetisimals there is a competition between accretion and the disruption due to destructive collisions, with accretion gradually winning and the gas and dust gradually accumulating into large bodies. H_2 and He is lost from the inner system due to the heating from the sun.

Jupiter and Saturn form rocky cores early and so start to capture the Hydrogen and Helium - at least enough to retain similar compositions to the early nebula. Either Jupiter and Saturn swept up so much material that Uranus and Neptune, forming later, had less to accumulate, or Uranus and Neptune formed with smaller cores which were unable to attract the lighter gases as efficiently before they were lost to the system. The sun meanwhile slowed down due to solar wind drag which deposited the angular momentum in the material which was then lost from the system. Once the inner system had been cleared by the solar wind and radiation pressures from the primordial sun, all that was left in a surrounding cloud outside the main planets was icy conglomerates with compositions akin to the early nebula.

High Mass Nebula

A "High-Mass" Nebula is, considered to be one where the mass is greater than about 2 M_{Sun}. Theories considering the solar system were formed in such a region again start with the cloud contracting to the spin plane. The large angular momentum in this massive disc leads to turbulence and increased interaction of the constituents. It is possible also that solar tidal forcing plays a part, but the end result is that the planetary bodies start to separate out:



The cores form from the nebula with roughly primordial compositions, but then gas is lost from the disc due to turbulence and the evolving Sun's action. Angular momentum is transferred away from the Sun by the turbulence and viscosity of the medium. As the Sun warms up it clears the system of its un-bound volatiles, starting with the inner solar system first.

Fitting the evidence to the theory

How do these theories fit the current hese theories fit the current morphology of the system? We know that the density of the planetary bodies tends to be higher in the inner part of the system where we expect the volatiles to have been depleted. (The terrestrial planets have a higher proportion of silicates - this gives way to water ice at larger distances from the sun and then to methane and other more volatile ices.) The nebula theories, both high and low-mass, have support in the Allende meteorite discussed above, but this does not distinguish between them. We can look more generally at meteoritic composition - a topic we shall return to nearer the end of the course, and see if the range of compositions and morphologies favours one theory over another. We can also see what material from elsewhere in the solar system might tell us.

We have material from the Moon. This seems to be differentiated, volatile poor, similar to chondritic meteorites in composition, but sampling was poor (the Apollo landers came down all in a band near the lunar equator and all in fairly similar smooth terrain, for safety reasons).There is anyway the question as to whether the Moon is typical or representative. In the cases of Venus and Mars very little is known of their compositions. We have some surface chemistry information from a limited number of sites on both bodies, but no internal material with the possible exception of the SNC meteorites which are believed to be Martian in origin; even with these though we have no way of knowing how representative the sampling is. Mercury we know little about except that it is dense, probably contains a high proportion of iron, and that FeO has been detected spectroscopically: we conjecture it might have a core. We give more information on this and the other terrestrial planets in a later section, but overall the information is complex and not always reliable or telling as far as one theory or the other is concerned. Similarly with the gas giants - little is know of their interiors and in many ways the theories of their structure are dependent on how we think the solar system formed, rather than vice versa.

On the whole, though, the tendency currently is to favour the low-mass nebula theory and the accretion of many small planetisimals.

Whatever theory we favour, they tend to leave Pluto/Charon as an anomaly along with others like the large relative size of the earth's moon in comparison to its primary, Triton's retrograde orbit about Neptune, Uranus' large inclination etc. These "uncomfortable" facts require additional detail to be added, and each of these has to be explained as exceptions to the general theories. This should

not be unexpected: the theories of solar system formation are all complex, and will all have a lot of scope for chaotic variation.

We get some evidence to support the nebula theories from looking a to support the nebula theories from looking at other star systems. We see stars like Beta Pictoris which have disc-like nebulae, probably made of rock rubble and asteroidal-type bodies. IRAS (the Infra-Red Astronomy Satellite) saw many bright infra-red sources which appeared to be flattened dust discs about young stars. We have found large Jupiter-sized (and larger) planets in nearby star systems (by the "wobble" they impart to the motion of the parent star) and what is believed to be a small terrestrial-sized planet associated with a neutron star, but so far our technology has not allowed us to see small planets in other star systems. However, the evidence is growing that planetary systems are not uncommon, and we believe that our system may not be atypical.

When we start to examine the nebula creation theories in detail we can ask more complex questions. We know from dating meteorites and other evidence that the building blocks of the system were created 4.6×10^9 years ago, but the models of the creation tell us that the nebula only lasted 10^5 years. What are the consequences of this? It certainly seems likely that this short time scale means that many of the chemical processes that we postulate as taking place did not have time to come to an equilibrium state, and we have to take that into account when looking at abundances and distribution of elements.

From meteorites we see evidence of volatiles captured in volatiles captured in large, low-temperature grains, but we also have some material which shows evidence of differentiation, that is processes, probably driven by heat, which have allowed materials to separate out by density or other gradient. Primitive "chondritic" meteorites have a matrix of materials containing hydroxyl silicates, sulphides, halides and reactive iron oxides. Other meteorites with larger crystals (which would have been heated - origin inside asteroids?) have much smaller concentrations of these.

The **Gases** that we find around bodies as we go out from the sun also seems to show trends that would fit with our general picture. Models of the equilibrium concentrations expected in a radial temperature gradient, given the nebula's initial composition, suggest that near the sun one would get a primarily CO and N_2 mixture as the most stable components. We would expect this to dominate in the early nebula inside the radius where the temperatures are above 680 K. Below this (and hence further out) CH_4 and nitrogen would be the most stable combination, and then out beyond the point where the temperature drops to 330K ammonia and methane would be expected to dominate. Unfortunately this picture is drawn from models which assume equilibrium conditions, and as we discussed above, the 10^5 years we have for the material to settle to the final locations from the nebula may not hanebula may not have been enough for this equilibrium to have been properly established. However, it does seem to fit in a crude way with the way volatiles and gases are distributed through the current system.

Non-co-eval Theories

Just for completeness we should note that there are also the non-co-eval theories of planetary formation. An example of these is the theory due to Woolfson, where two molecular clouds, with protostars already forming or formed, meet head on, leading to turbulence and shock waves in both systems. These in turn lead to compression and condensation out of planetary bodies from the nebula:

This model implies that a lone protostar retains its original angular momentum and would have no planets. Any lone star with planets must have captured them from a passing protostar by tidal forces. This "Tidal Capture Theory" can obviously take a number of varying forms, and there are several theories along these lines, though the nebulaic theories currently seem to be holding sway.

Atmospheric origin

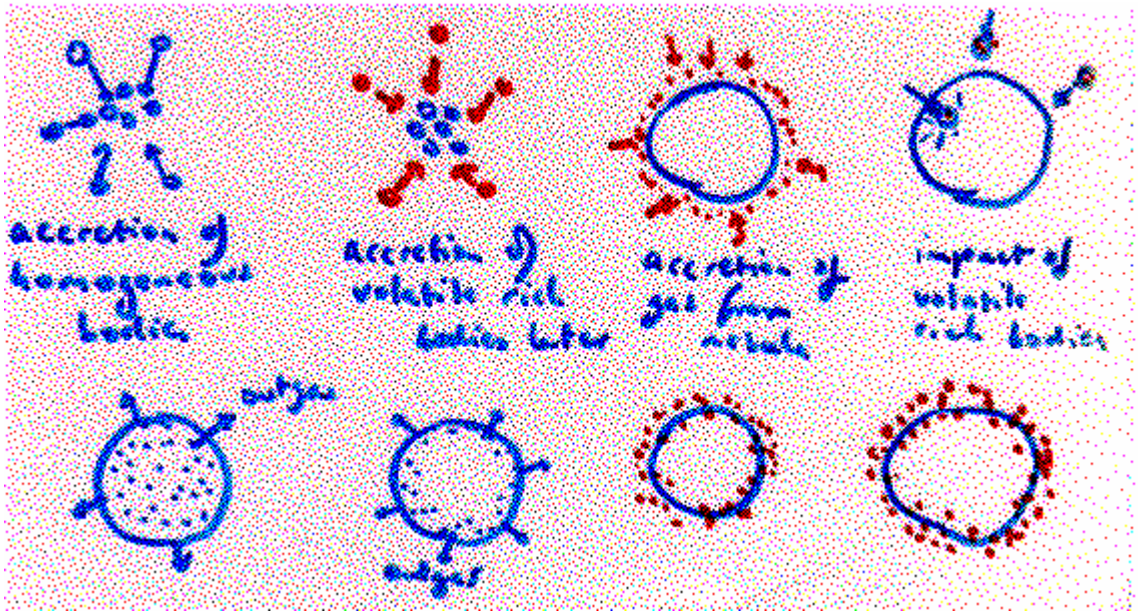
To some extent explaining the origins of the atmospheres of the terrestrial planets represents a more complex problem than the distribution of solid bodies throughout the system. The main question is, where did the volatiles come from?

- Primary capture from the primordial cloud?
- Outgassing?
- Cometary CLI > Outgassing?
- Cometary Capture?

Deciding between these is complicated by the fact the current atmospheres might be the result of a long process of chemical change from the original atmospheres that were created. It is not very east to try to trace back any possible temporal variation to the "source atmosphere". The earth's atmosphere, for example, is largely the result of biological processes, rather than the primordial constituents.

Mg, Si, Fe and O are the most abundant terrestrial elements. The most common elements in the primordial nebula (based on solar abundances) are H, He, C, O (H_2O) and N. On Earth the C, N and H_2O are 10^4 less abundant relative to Mg, Si and Fe than they were in the solar nebula. So the earth, during or since formation, lost a great deal of the volatile material from the primordial nebula. The question is, why did it not lose all in the same process? Is the atmosphere a result of the little that was left, or was this acquired later to a body that was by then volatile-free?

There are a number of possible mechanisms by which we could imagine the earth (and other terrestrial planets, or satellites) acquired their atmospheres:



If water were acquired by local accretion it would mainly be in hydrated form, as seen in meteorites. N, C, O and H would be acquired in complex carbohydrates, and smallex carbohydrates, and small amounts of C and N "dissolved" in the metallic Fe phase. Some water would be derived from reduced hydrogen from organic compounds. As the planetisimals were forming and accretion taking place, the early sun would be "burning off" the volatile material near the sun, but material deflected by Earth and Venus would tend to homogenize the inner solar system.

If accretion of volatiles was from bodies further out than the terrestrial bodies, this would require deflection inwards due to resonances with Jupiter and Saturn and other interactions further out. If the material that arrived was then asteroidal in type, then the water would be in hydrated form (with maybe some water from ice if the material was from far enough out) and C and N from organic molecules. In all cases you would expect the terrestrial planets all to have similar fractional amounts of volatiles. We can check this out by looking at the evidence in terms of the C and N in the different planets. This is difficult though because of the differing distributions and histories. On earth the C is mainly trapped in carbonates, whereas on Venus much of it is still in the atmosphere. Most of the N_2 on Venus is in the atmosphere, but on earth some has been buried in organic material. The evidence for earth is particularly difficult to interpret as we have to take into account "recycling" in the crust/mantle. It does seem, though, that earth and Venus have similar mounts of C and N, but Venus is deficient in H.

In trying to assess the evidence we have to try to guess how an accretion scenario would go. Early on the accretion would be fairly 'gentle' and all the volatiles would be 'held' by the bodies to which they accreted. Later accretion "collisions" would have more power and so would contribute to loss of the volatile material - we would then have a process that both adds and takes away volatile material. Larger bodies would degas more - anything over $0.3R_E$ would heat up enough to outgas the trapped material at its centre - but would have larger gravitational fields to hold onto the volatiles and gases. It is thought that the earth while accreting could have reached a temperature of 1500K from the pressure-generated heat of collisions and redistribution.

In the presence of iron in the bodies, some reduction reactions control the gases formed. Thus Carbon would be seen mainly as CO, N as N_2, Hydrogen in H or H_2 form. In the absence of iron we would get chiefly H_2O, CO_2 and N_2.

In addition to these considerations of the original components and reactions we also have to consider subsequent possible evolution of atmospheres - water can be extracted from the gas phase as ice and "locked in" to the body. Ultra-violet radiation can dissociate water to H_2 and O_2, and the hydrogen flow outwards (hydrodynamic escape) can take other particles with it. (Thus, there is evidence of preferential loss on earth of lighter isotopes of Ne.) A relatively new theory from NASA suggests that regular increases in the amount of oxygen in the earth's atmosphere is caused by the loss of large amounts of organic matter in large-scale tectonic upheavals.

An extra complication is the fact we must also take into account the sun's early history. It was, for example, less luminous early on, but may have been a stronger u.v. radiator. The early earth may have needed more greenhouse gases in its atmosphere in order to keep its oceans liquid (as they have been know to be for at least 3×10^9 years.). Could there have been more CH_4 and NH_3? These could then have been converted to CO_2 and N_2 by photodissociation and loss of H_2 But early volcanic activity producing H_2 and CO would soon have destroyed the methane and ammonia.

Some clues as to the origins of the terrestrial atmospheres can be obtained by looking at the proportions of inert gases in these bodies. One piece of evidence suggesting that they are not of pre-solar nebula origin is that the noble gases are not of solar abundances. The proportion of inert gases is more like that found in chondritic meteorites than in thc meteorites than in the sun. Mars is more deficient in the noble gases than earth, which is in turn more deficient than Venus. It is easiest to understand this overall deficiency if the acquisition was by "veneering" rather than initial accretion at the time of planetary body formation.

Mars is depleted in H and N compared to earth. It might have acquired similar proportions of material but lost more by outgassing, it might have acquired from a less-volatile rich region, or it may have been incapable of attracting volatile rich bodies late in its history because if its size. Venus has about the same amount of C and N as earth but less hydrogen. If Venus has lost its hydrogen why hasn't the earth? In the case of Mercury and the Moon, did they accrete material but lose it almost immediately, or were they incapable of attracting sufficient volatile rich bodies?

Lunar Origin

The origin of the earth's moon has by itself presented an interesting puzzle to man throughout history. There have been a number of theories suggested:

- Fission: the early cloud from which the earth formed pulled apart into two bodies as condensation was taking place. There is then the question as to what might have caused this split. Spin? This seems unlikely. A third body could have been involved.
- Simultaneous accretion: The earth and moon form separately in similar orbits from the original nebula. Instead of accnal nebula. Instead of accreting together as would have happened with all other such near bodies, a resonant orbital configuration allowed the smaller body to be captured by the larger. This probably required the moon to have been originally farther away than currently - possibly on a totally separate planetary orbit.
- Capture: The Moon was a totally separate body in an earth-crossing orbit which was captured by the earth. This might be a derivative of the "simultaneous accretion" theory, or might be capture of a body from much further out. The Moon is rather large to be an errant asteroid, especially an earth-crossing one, but maybe it was dislodged by a close encounter further out in the system. There is a problem here in seeing how capture can occur if the bodies meet at too high a relative velocity.
- Collision: A theory that has been around a long time but which currently seems to be gaining ground again is the idea that the moon was formed following a collision of the earth with another large body. The colliding body would have to be as large as Mars, and the debris from the collision

accrete quickly into the moon in a near-earth orbit, then spiral out. This theory might account for why the moon is more like mantle material with little core. The moon is known to have spiralled gradually out from the earth, at least in relatively recent history.

If you are interested in this subject see Ida et al, Nature vol 389 et al, Natuure vol 389, pp 353-357 and the associated references.