4

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

Paul G. Higgs and Ralph E. Pudritz
Origins Institute, McMaster University, Hamilton, Ontario, Canada.

4.1 Introduction

The robust formation of terrestrial planets, as well as abundant sources of water and organic molecules, are likely to be important prerequisites for the wide-spread appearance of life in the cosmos. The nebular hypothesis of Kant and Laplace was the first to propose that the formation of planets occurs in gaseous disks around stars. The construction of new infrared and sumillimetre observatories over the last decade and a half has resulted in the discovery of protoplanetary disks around most, if not all forming stars regardless of their mass (eg. reviews Meyer et al, 2006; Dutrey et al., 2006). The recent discoveries of extrasolar planets in over a hundred planetary systems provides good evidence that Jovian planets at least, may be relatively abundant around solar-like stars (see chapter by Zucker). These results beg the question of whether protoplanetary disks are also natural settings for the manufacture of the molecular prerequisites for life. Life requires water, organic molecules such as amino acids, sugars, nucleobases, and lipids as building blocks out of which biological macromolecules and cellular structures are made, and many of these can be manufactured in protoplanetary disks.

The first part of this article reviews the properties of protoplanetary disks and how planets are believed to form within them. It then considers the evidence that chemical reactions within the disk may be a major source of the water and biomolecules available for the earliest life on Earth. We focus on amino acids because they are the key components of proteins and because more is known about prebiotic synthesis of amino acids than most other biomolecules. In the second part of this review, we compare the different environments that have been proposed for amino acid synthesis - including the atmosphere, hydrothermal vents in the deep ocean, and protoplanetary disks - and show that there is considerable consensus on which amino acids can be formed prebiotically, even if there is still disagreement on the location.

We will argue that the amino acids that were least thermodynamically costly to form were the ones that were most frequent before life arose. Early organisms could make use of existing amino acids in the first proteins. Later organisms evolved biochemical pathways to synthesize additional amino acids that were not common prebiotically, thus increasing the diversity and functional specificity of the proteins they could make.

Synthesis of specific proteins is only possible after the origin of the genetic code (i.e. the mapping between codons in RNA and amino acids in proteins). We will discuss the origin of the genetic code in the context of our understanding of prebiotic amino acid frequencies. We will also consider the evidence that the genetic code is optimized to reduce translational error and discuss how it came to be this way.

It is likely that early proteins were composed of a smaller set of amino acids than the 20 used currently. We will discuss several experimental studies of proteins composed of deliberately reduced amino acid sets. Another way to estimate amino acid frequencies in early proteins is to use phylogenetic methods to deduce ancestral protein sequences. It might be expected that amino acids that were added early to the code should be more frequent in the past than at present, whereas the late amino acids should be less frequent in the past. We will consider to what extent phylogenetic estimates agree with our expectations from prebiotic sytnthesis.

4.2 Protoplanetary disks and the formation of planet systems

Stars and planets form within dense, cold clouds of gas and dust known as molecular clouds. Surveys of many molecular clouds show that most stars form as members of entire star clusters. These stellar nurseries arise from dense regions within molecular clouds known as clumps which extend over about 1 pc in physical scale (or about 3 light years, or 205,000 Astronmical Units - AU - where 1 AU = 1.5×10^{13} cm is the distance between the Earth and the Sun) and typical initial temperatures around 20 K. The gas in clouds and clumps is vigourously stirred and by supersonic turbulence and has a filamentary structure. A single star or a binary stellar system forms within a small, dense subregion of a clump (known as a core) that for stars like the Sun extends over a scale of 0.04 pc and has a (particle) number density ranging from $10^4 - 10^8$ cm^{-3}.

Numerical simulations of supersonic turbulent gas under the observed conditions show how stars and their protoplanetary disks form. Supersonic turbulence first sweeps up the gas into systems of shocked sheets and filaments. Dense core-like regions are produced as such flows develop. Also, the shock waves that produce the cores are oblique and hence impart spin to the cores (eg. review MacLow and Klessen, 2004). The eventual gravitational collapse of such cores under their own weight preserves most of their angular momentum resulting in the formation of protoplanetary disks (eg. Tilley and Pudritz, 2004).

Surveys of star forming regions have now established that disks are universal around solar mass stars. Stars range over more than 3 decades in mass, and over all of these one has evidence for disks. The least massive of these is the disk around an object that has only 15 M_J (Jovian masses)! (Note, the mass of Jupiter is one thousandth the mass of the sun; $M_J = 0.001 M_{\odot}$). Disks are commonly observed around solar like stars as they form. At the most massive end, rotating disks have been found around B stars that are up to 10 $M_{\odot}$ (eg Schreyer et al 2005).

Most studies generally do not resolve a disk around a star, but infer its presence from the excess of infrared emission that is seen in the spectrum of the star. Disks around forming solar type stars extend typically out to more than 100 AU, with each radius r of the disk being at its own temperature. For reference: in our solar system, the orbital radius of Earth's is 1 AU, while that of its most massive outer planet Neptune, is 30 AU. The total emission from this collection of rings of gas and dust, each of which emits radition like a black body at some temperature, adds up to the observed excess infrared emission (known as the spectral energy distribution or SED).

For solar-type stars, the disk temperature that can be inferred from such observations scales with disk radius as $T \propto r^{-0.6}$, with a temperature at 100 AU of about 30 K. Similarly, one can deduce that the column density of the disk (which is the volume density integrated over the disk thickness at each radius) varies as $\Sigma \propto r^{-1.5}$ and has a value at 100 AU of 0.8 g cm^{-2}. Disks seen at these later stages of their evolution are far less massive than their central stars, having typically $10^{-2} M_{\odot}$ of gas and dust (eg. Dutrey et al., 2006). Hence, their dynamics are governed by the gravitational field of the central star, and the material is expected to be in nearly Keplerian orbit about the star (ie $v_{Kep}(r) = \sqrt{(}GM_*/r))$.

Occasionally, one can also spatially resolve the disk around a young star. The rotation curves of the disks can be measured under such conditions, as has been done in molecular observations (Simon et al., 2000), and these have been found to be nearly Keplerian. At a slightly later stage (a few hundred thousand years), after the surrounding cloud is blown away by the young massive star, stars and their disks still in formation can be seen in optical images. This is seen in Hubble Space Telescope image in Figure 4.1 where we see several images of young stars that are forming in the Orion Nebula Cluster. A massive star has created a strongly illuminated background of glowing nebular gas. The HST image shows a disk around each young star that is seen as a silhouette against this bright background (McCaughrean and Stauffer 1994). Disks are heated primarily by the young star that they surround, as well as by a massive nearby star in the case of clustered star formation. They are also illuminated by ultraviolet radiation as well as X-rays from the central star, and this drives the production of organic molecules, as we discuss in the next section.

Studies of the frequency of disks in Orion also allow us to determine their lifetimes. One can calibrate the lifetimes of protoplanetary disks because one can measure the ages of the young stars that they encompass. The result is that 50% of disks are gone by 3 million years after formation, and 90% are gone by 5 million years. This lifetime for disks translates into a hard upper limit for the time for the formation of massive Jovian planets. While Jupiter has a 8-10 Earth mass core, its bulk consists of gas that was accreted from the protoplanetary disk in which it formed.

The building of planets and molecules depends, in part, on the balance between carbon and oxygen in the molecular gas within these protoplanetary disks (eg. Gaidos and Selsis, 2006). These two elements are the most abundant in the interstellar medium after hydrogen and helium, and are predominantly bound up in the form of CO in molecular clouds. During gravitational collapse however, the presence of high pressure can shift this balance in favour of the formation of water and methane (CH_4) from CO and molecular hydrogen. Thus, if there is an excess of O over C, then the extra O is taken up in the formation of water (through combination with abundant molecular hydrogen). In the converse case, and excess of C over O is used up in the formation of graphite and organic molecules. The balance between C and O depends on the condensation and sedimentation of grains.

While the bulk of the mass of disks is dominated by gas, dust grains play several important roles. Dust grains in more diffuse gas consist of

06/29/06 http://www.ifa.hawaii.edu/faculty/jewitt/images/HSTproplyds.jpg #1

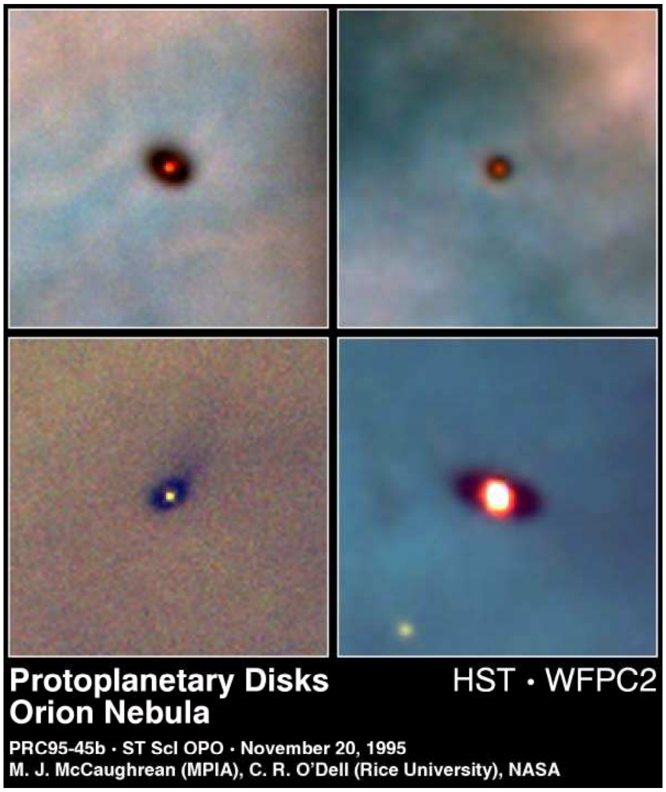
Protoplanetary Disks
Orion Nebula
HST • WFPC2
PRC95-45b • ST ScI OPO • November 20, 1995
M. J. McCaughrean (MPIA), C. R. O'Dell (Rice University), NASA

Fig. 4.1. Hubble images of four, protoplanetary disks in the Orion Nebula. These are seen as silouettes against the bright background of the nebula.

a mixture of silicates and carbons, with sizes ranging typically from 100 Angstroms to 0.2 - 0.3 micometers. Once dust starts to collect in denser environments such as protostellar disks, it grows by agglomeration via collisions. This growth continues up to the scale of kilometer sized objects (called planetesimals) and eventually planets. Dust grains, while still small, are the dominant form of opacity in protostellar disks and effectively absorb the UV radiation that falls on disks from their central and surrounding stars (eg. Natta et al 2006).

Models of terrestrial planet formation agree that the process gets started as dust settles onto the midplane of protoplanetary disks (see chapter by Thommes). In the earliest phases, dust grains condense and grow by collisions with other dust grains, gradually settling to the disk midplane as they become more massive. After the formation of planetesimals, an "oligarchic" growth phase occurs wherein gravitational interactions lead to focusing of orbits and collisions of these planetesimals resulting in a large numbers of Moon to Mars-sized objects. At a radial distance of 1 AU from the central star, this process occurs with a million years (eg. Kokubo and Ida, 2002). The gasous disk is still present at this phase in evolution. The late collisions of these oligarchs over 100 Myr produces the sparse system of massive planets.

If the Earth formed at around 1 AU in such a disk, it must have formed in a very dry zone without much water. This is because the snow-line for water - that radius in the protostellar disk which separates a hot inner region of the disk where water cannot condense from an outer region where the temperature is low enough for water to freeze out and condense in grains and planetesimals - occurs at 2.5 AU. The Earth's ocean therefore probably arose as a consequence of the delivery of water by bodies that formed beyond 2.5 AU and that later collided with the Earth. It is now thought water ice in comets is unlikely to be the source since the deuterium to hydrogen ratio (D/H) of comets is twice that found in Earth's water. Perhaps only 10 % of the Earth's ocean came from a cometary source. It has been suggested (Morbidelli *et al.*, 2000) that water arrived from the asteroid belt through the impact of water-laden planetary embryos discussed above.

Massive planets such as Jupiter play a major role in this process. Recent numerical simulations (Raymond et al., 2004) show that the presence of a Jovian mass planet stirs up planetsimals and perturbs them into ever more eccentric orbits. Such orbits carry the planetary embryos into the hot, dry inner regions of the disk where they can undergo collisions with the Earth. This is a stochastic process and the simulations show that the majority of Earth-like planets that formed in the "habitable zone" (0.8 - 1.5 AU) have water contents that are within an order of magnitude of that of the Earth. There is a broad range however, extending from some dry worlds to a few watery ones that have more than 100 Earth ocean's worth of water.

In fact, massive gas-giant (Jovian) planets play a dominant role in the formation and evolution of entire planetary systems. Their formation is still unclear, but there are two major contemporary models that are under intense investigation. The most popular of these is the core accretion picture wherein a rocky core that grows to about 5-10 Earth masses can then rapidly accrete gas from the disk. This may take a few million years however - uncomforatably close the the observational limit for disk lifetimes (Pollack *et al.*, 1996). The alternate picture is that Jovian planets form as a consequence of rapid gravitational instability in a fairly massive disk (eg. Mayer *et al.*, 2002). Whatever the mechanism of their formation, the final mass of a giant planet is strongly dependent on the structure of the disk (its pressure scale height) as well as the magnitude of its turbulent viscosity. Jovian planets cease to grow when their mass (known as the gap-opening mass) is large enough to exert a torque on their surrounding disks in the face of turbulent viscosity, thereby opening a signficant gap in it. The disappearance of the corresponding ring of emission from the gas at this radius translates into a notch-like feature in the SED, which can be searched in the spectra of forming stars. The advent of the highly sensitive Atacama Large Millimeter Array (ALMA) in the latter part of this decade will allow such gaps to be directly imaged, as will the emission from the forming Jupiter (eg. Wolf and D'Angelo, 2005).

Planets exert significant tidal forces on their natal disks and this results in the inward migration of the planet (Lin and Papaloizou, 1993). This process is quite general and explains why the newly discovered extrasolar planets observed around solar type stars, have orbits well within 5 AU, and sometimes at the equivalent radius of the orbit of Mercury (see chapter by Zucker). Jupiter mass planets are thought to have formed at much larger disk radii, where gas is more abundant, and to have migrated to their present postions. Models show that planets can migrate very quickly - within a million years or so, and are liable to being swallowed by the star. This presents a problem as to why there are any planetary systems in the first place. Also, the migration of massive planets will strongly perturb the smaller terrestrial planets and this is of profound significance for the origin of life in such systems.

Several mechanisms have been proposed for saving planetary systems from such rapid destruction by their central stars. One of possible universal applicability is the role of so-called dead zones that occur in disks (Matsumura *et al.*, 2006). Dead zones are regions that are two poorly ionized to support the kind of hydromagnetic turbulence in disks that makes them "viscous". Such zones may extend up to 15 AU or so and are marked by low viscosity. A massive planet that migrates into such a region will readily open a large gap in the disk. It's radial inward migration is then locked to the slow rate at which gas in the dead zone drifts inwards.

4.3 Protoplanetary disks and the formation of complex organic molecules

Well over one hundred molecules, many of them organic, have been identified in various regions of interstellar gas. The clumps and cores in molecular clouds out of which stars form have densities ranging from $10^4 - 10^8$ cm^{-3} and low temperatures. In this state, cold gas chemistry can account for the formation of the simple molecules (CO, N_2, O_2, C_2H_2, C_2H_4, and HCN. The surfaces of dust grains play host to the formation of more complex organic molecules, which include nitriles, aldehydes, alcohols, acids, ethers, ketones, amines, amides, and long-chain hydrocarbons.

For decades, our knowledge of the chemistry in the solar nebula was based on studies of planetary atmospheres, meteorites, and comets. Current observations allow us to directly study the chemistry of disks around other stars. This work reveals that active chemistry occurs near the surface regions of protoplanetary disks. This chemistry is in disequilibrium and is observationally similar to that found in dense regions directly exposed to UV and X-ray irradiation (Bergin *et al.*, 2006).

Chemistry and reaction rates in protoplanetary disks depend on the local gas density, temperature, and radiation field. Disks are not flat, but are instead observed to be flared. Their surfaces are therefore exposed to radiation from the central star. The vertical structure of disks arises from the fact that the pressure gradient of the gas at any radius and scale height in the disk support the weight of the material above it. The force that must be balanced is the vertical component of the gravitational field of the central star (see eg. Dubruelle *et al.*, 2006).

Self-consistent calculations show that the heating of such flaring disks is due primarily to the UV radiation from the central star that is absorbed by dust grains in such surface layers (Chiang and Goldreich 1997, D'Alessio *et al.*, 1998). The dust in the disk envelope then re-radiates this energy in the form of infrared and submillimeter photons - about half of which escape vertically away from the disk, while the remainder is radiated downwards towards the disk midplane. These infrared photons are then absorbed by the deeper lying dust and ultimately radiated

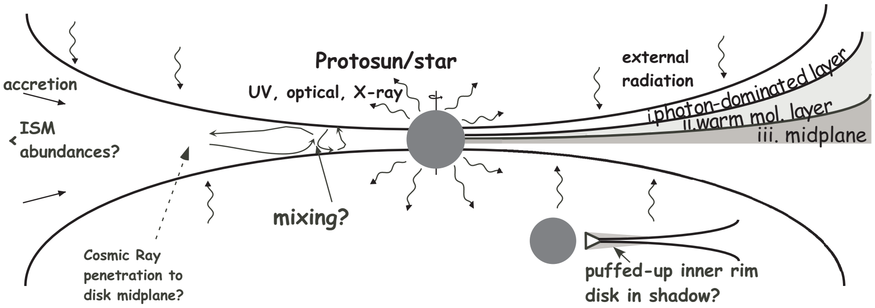

Fig. 4.2. Illustration of the structure of protostellar disks in response to heating and ionization from their central stars and cosmic rays. Note the general appearance of a molecular layer. Adapted from Bergin *et al.* (2006), with permission.

out of the disk. The gas is heated by collisions with the warm dust, and cools by radiating this energy in the form of molecular vibration-rotation emission lines that can be observed. This overall process produces a disk that is hottest near the surface layers, and cooler towards the midplane. By balancing all of these heating and cooling rates, one can determine the local disk temperature distribution and hence the disk's SED. The prediction that the temperature should scale as $T \propto r^{-0.5}$ is close to that observed.

The vertical structure of disks at disk radii beyond 100 AU consists of three layers; the radiation dominated surface layer or "photon dominated region" (PDR) which consists mainly of atomic and ionized species, a molecular layer at greater depth beyond which most of UV has been absorbed by the grains, and the cold midplane layer noted above. This mid-plane region turns out to be so cold (around 20 K) that most heavy species such as CO and other molecules, freeze out on the dust grains. Figure 4.2 shows a schematic of the disk. It is the molecular layer that is of primary interest for the synthesis of complex organic molecules.

Observations of molecules so far are still limited by the sensitivity of telescopes since the emission is so weak. By far the most abundant molecule is molecular hydrogen H_2 but it is generally hard to detect. So far, the most abundant species that have been observed in disks include HCO^+, CN, CS, HCN, H_2CO, and DCO^+ (eg. Dutrey et al 1997; van Zadelhoff et al. 2001). Silicates such as Mg_2SiO_4 (forsterite) have been detected in the hot surface layers of the inner regions of the disk ($r \leq 1-10$ AU). This is important because the shape and strengths of silicate features in disk spectr allows one to track the evolution and growth of dust grains in disks with time. Ices such as H_2O, CO_2, and CO are observed in the outer regions of the disks where tempertures drop to less than 100K. The most abundant organic molecule observed are the polycyclic aromatic hydrocarbons, or PAHs. Their abundance per H atom is of the order of 10^{-7}. Up to 50 % of carbon may be locked up in such carbonaceous solids. They are important for disk chemistry becuse they are absorb UV, head gas, and act as potential sites for H_2 formation.

Dust grains that are coated with simple icy mantles warm up as they are mixed and transported to dense, active protostellar regions. This can occur in regions such as the so-called hot cores, that are hosts to massive star formation, or the innermost (in radius) regions of protoplanetary disks. These regions are particularly rich in organic molecules. Ultraviolet irradiation, or perhaps X-ray bombardment, then breaks bonds, and allows reactions that can produce organic molecules. However, evaporation of these simple dust grain surfaces can drive gas phase production of organics.

It is now well established that organic molecules of extra-terrestrial origin are found in carbonaceous chondrite meteorites. The best studied of these is the Murchison meteorite, in which 8 of the 20 biological amino acids have been reported (Engel and Nagy, 1982). Meteorites also contain many other organic molecules that are relevant in astrobiology (Sephton, 2002), including lipids that have been shown to be capable of membrane formation (see chapter by Deamer). The organic molecules in meteorites may have originally arisen on dust grains, which were then incorportated into comets and meteorites. The possibility that dust grains with icy mantles are a site for amino acid synthesis has been investigated experimentally. Bernstein *et al.* (2002) and Munoz-Caro *et al.* (2002) have shown that amino acids may result from the exposure of icy mantles consisting of HCN, ammonia (NH_3) and formadlehyde (H_2CO) to UV, produces amino acids such as glycine, alanine, and serine.

Alternatively, it has been proposed that amino acids such as glycine and alanine, and the base adenine, can form during the gravitational collapse of a core. Chakrabarti and Chakrabarti (2000) added reactions that make amino acids (eg. by Strecker process) to the existing UMIST database of chemical reactions in molecular clouds creating a network involving 421 molecular species. As an example, they found that glycine was formed which had a mass fraction of order 10^{-12}. It peaks at about 100 AU scales after about a million years after the initiation of the collapse. The astronomical detection of interstellar amino acids, such as glycine, would strengthen the link between the chemistry of disks and observed amino acids in meteorites. A spectroscopic survey for amino acids in clouds and disks will be possible with ALMA.

Organic molecules contained in comets and meteorites are delivered to Earth when impact events occur, and it is known that impacts were frequent in the early history of the solar system. Impacts release huge amounts of energy which might cause thermal degradation of macromolecules. However, Pierazzo and Chyba (1999) predict that molecules such as amino acids can survive such impacts, and hence that the rate of delivery of biomolecules from space could exceed the rate of synthesis on Earth.

4.4 Measurements and experiments on amino acid synthesis

The previous section presents the case for synthesis of biomolecules in protoplanetary disks, but several mechanisms of prebiotic synthesis on Earth have also been debated for many years. In this section, we show that these results combine to give a coherent picture of which amino acids were most frequent on the early Earth. We focus on amino acids, because of their fundamental importance in biology, and because they have been detected in a wide range of non-biological contexts relevant to prebiotic synthesis.

In Table 4.1, each column contains observed amino acid concentrations normalized relative to glycine, unless otherwise stated. In the following discussion, column numbers refer to this table. Column M1 is the H_2O extract from the Murchison meteorite (Engel and Nagy, 1982). Column M2 is the Murray meteorite (Cronin and Moore, 1971), and Column M3 is the interior, hydrolysed sample of the Yamato meteorite (Shimoyama *et al.*, 1979). Several amino acids that are not used biologically are found to be as frequent in meteorites as the biological ones (Kvenvolden *et al.*, 1971). We will not consider non-biological amino acids in this paper. However, the question of why certain amino acids became used in proteins and others did not is an important one, and in many cases, reasons have been proposed as to why the non-biological ones were avoided (Weber and Miller, 1981). Column I1 shows measurements from an experiment simulating the chemistry on icy grains (Munoz Caro *et al.* 2002).

Miller and Urey showed that amino acids could be synthesized by exposure of a mixture of reducing gases to UV or electrical discharge.

	M1	M2	M3	I1	A1	A2	A3	H1	H2	S1	S2	S3	R_{obs}
G glycine	1.00	1.0	1.00	1.000	1.000	1.000	1.000	18	12	1.000	1.000	40	1.1
A alanaine	0.34	0.4	0.38	0.293	0.540	1.795	0.155	15	8	0.473	0.097	20	2.8
D aspartic acid	0.19	0.5	0.035	0.022	0.006	0.077	0.059	10	10	–	0.581	30	4.3
E glutamic acid	0.40	0.5	0.11	–	0.010	0.018	–	6	11	–	–	20	6.8
V valine	0.19	0.3	0.10	0.012	–	0.044	–	1	–	0.006	–	2	8.5
S serine	–	–	0.003	0.072	–	0.011	0.018	8	11	–	0.154	–	8.6
I isoleucine	0.13	–	0.06	–	–	0.011	–	8	9	–	0.002	4	9.1
L leucine	0.04	–	0.035	–	–	0.026	–	3	–	0.001	0.002	7	9.4
P proline	0.29	0.1	–	0.001	–	0.003	–	9	–	–	–	2	10.0
T threonine	–	–	0.003	–	–	0.002	–	2	–	–	0.002	1	11.7
K lysine	–	–	–	–	–	–	–	7	–	–	–	14	12.6
F phenylalanine	–	–	–	–	–	–	–	4	–	–	–	1	13.2
R arginine	–	–	–	–	–	–	–	–	–	–	–	15	13.3
H histidine	–	–	–	–	–	–	–	–	–	–	–	15	13.3
NQCYMW	–	–	–	–	–	–	–	–	–	–	–	–	14.2

Table 4.1. *Frequencies of amino acids observed in non-biological contexts. In the column headings, M denotes meteorites, I denotes icy grains, A denotes atmospheric synthesis, H denotes hydrothermal synthesis, and S denotes other chemical syntheses (details in text).* R_{obs} *is the mean rank derived from these observations. The final 6 amino acids are not observed: N asparagine, Q glutamine, C cysteine, Y tyrosine, M methionine, W tryptophan.*

These experiments were intended to show that synthesis was possible in the atmospere of the early Earth, which was presumed to be reducing. Column A1 shows the results with an atmosphere of CH_4, NH_3, H_2O and H_2 (Miller and Orgel, 1974), and column A2 shows the results with an atmosphere of CH_4, N_2, H_2O and traces of NH_3 (Miller and Orgel, 1974). Although yields are lower in non-reducing atmospheres, column A3 shows an experiment using proton irradiation of an atmosphere of CO, N_2 and H_2O (Miyakawa *et al.*, 2002), which is not a strongly reducing mixture.

Another competing theory is that life originated in high-temperature, high-pressure environments in the deep sea (Amend and Shock, 1998).

Chemical syntheses of amino acids in hydrothermal conditions have been reported. Column H1 is from Marshall (1994), and H2 is from Hennet *et al.* (1992). Both of these columns give the number of times that the amino acid was observed in greater than trace amounts in a large number of experiments. The final three columns are miscellaneous chemical synthesis experiments: S1 is shock synthesis from a gaseous miture (Bar-Nun *et al.*, 1970); S2 is synthesis from ammonium cyanide (Exp. 2) (Lowe *et al.*, 1963); S3 is Synthesis from CO, H_2 and NH_3 at high temperature in the presence of catalysts (Yoshino *et al.*, 1971). This column shows the number of positive identifications of the amino acid from many experiments.

It is possible that the organic molecules that were present at the time of origin of life were synthesized by several different mechanisms. A central issue is to estimate what fraction originated on Earth and what fraction was brought to Earth after synthesis in protoplanetary disks (Whittet, 1997). The yields on Earth are dependent on the mixture of gases in the atmosphere. If the atmosphere is non-reducing, yields on earth are low, and delivery of molecules from space is more relevant (Chyba and Sagan, 1992; Kasting, 1993; Pierazzo and Chyba, 1999). However, there seems to be no consensus on this, and the importance of synthesis in the atmosphere is maintained by some (Lazcano and Miller, 1996; Miyakawa *et al.* 2002). Despite this uncertainty, the results in Table 4.1 are surprisingly consistent as regards *which* amino acids can be synthesized, even if they are in disagreement about *where and how* they were formed.

In his studies of the origin of the genetic code, Trifonov (2000, 2004) collected 60 criteria by which the amino acids can be ranked. A consensus order was obtained by averaging the rankings. This procedure incorporates diverse criteria that are not all consistent with one another, and some of these must be misleading. Nevertheless, the consensus order that emerges appears to be a useful one. We shall call the average ranks derived from genetic code criteria R_{code}. In this section we derive our own ranking, R_{obs}, based only on observed frequencies of amino acids in non-biological syntheses, i.e. we use non-speculative critera.

We ranked the amino acids according to each criterion in Table 4.1 separately. For example, for criterion M2, G is most frequent and is given rank 1. The next two, D and E, are equally frequent, and are both given rank 2.5. The next three are A, V and P, with ranks 4, 5 and 6. All the remaining amino acids are not observed, and are therefore given an equal bottom rank 13.5 (the average of the numbers between 7 and 20). R_{obs} is the mean of the ranks obtained from the 12 columns.

Several points are worth noting from Table 4.1. G is the most frequent amino acid in all but column A2, where it is second. A, D and E are also frequent in most experiments. The next six (VSILPT) are all found in the Miller-Urey experiment (A2), at least one of the meteorites, and several of the other cases. Therefore there is good evidence for prebiotic synthesis of all the amino acids down as far as T on the list. The following four (KFRH) are not found in the Miller experiments or in meteorites and only occur in one or two of the other chemical synthesis experiments. The remaining amino acids (NQCYMW) are not formed in any of the experiments. In our view, the data are not sufficient to make a definite distinction between KFRH and NQCYMW.

4.5 A role for thermodynamics

The first organisms probably had a very different mix of biomolecules and other conditions available than organisms today. Thermodynamic arguments can help us to understand what the earliest mixtures of amino acids may have been like. Table 4.2 lists the amino acids and their properties according to the ranking R_{obs} derived above. We will refer to the top 10 amino acids as 'early' and the bottom 10 as 'late'. Early amino acids were easily synthesized by non-biological means and were therefore available for use by early organisms. Late amino acids were not available in appreciable quantities until organisms learned to synthesize them. The early/late distinction is relevant in the discussion of the origin of the genetic code in the following section. Our ranking, R_{obs} is close to the genetic code ranking, R_{code}, taken from Table V of Trifonov (2004). The top 3 are the same, and 9 amino acids are in the top 10 of both orders. We prefer R_{obs} on the grounds that it is derived directly from experimental observables, however the conclusions drawn from both rankings are very similar.

The ranking can be interpreted on thermodynamic grounds. Amend and Shock (1998) have calculated the free energy of formation of the amino acids from CO_2, NH_4^+, and H_2 in two sets of conditions. ΔG_{surf} in Table 4.2 corresponds to surface seawater conditions (18^oC, 1 atmosphere), and ΔG_{hydro} corresponds to deep-sea hydrothermal conditions (100^oC, 250 atmospheres). Figure 4.3 shows that R_{obs} is closely related to ΔG_{surf}. For the 10 early amino acids there is a strong correlation between the two (r = 0.96). The late amino acids have significantly

	R_{obs}	R_{code}	ΔG_{surf}	ΔG_{hydro}	ATP	MW	dp/dt	Δp
G	1.1	3.5	80.49	14.89	11.7	75	-0.0063	-0.5
A	2.8	4.0	113.66	-12.12	11.7	89	-0.0239	-3.3
D	4.3	6.0	146.74	32.78	12.7	133	-0.0039	-0.4
E	6.8	8.1	172.13	-1.43	15.3	147	-0.0137	0.7
V	8.5	6.3	178.00	-70.12	23.3	117	0.0098	-1.6
S	8.6	7.6	173.73	69.47	11.7	105	0.0167	-0.6
I	9.1	11.4	213.93	-96.40	32.3	131	0.0089	-0.9
L	9.4	9.9	205.03	-105.53	27.3	131	-0.0017	3.2
P	10.0	7.3	192.83	-38.75	20.3	115	-0.0139	0.0
T	11.7	9.4	216.50	53.51	18.7	119	0.0091	-1.4
K	12.6	13.3	258.56	-28.33	30.3	146	-0.0065	1.6
F	13.2	14.4	303.64	-114.54	52.0	165	0.0042	1.5
R	13.3	11.0	409.46	197.52	27.3	174	0.0038	-0.2
H	13.3	13.0	350.52	154.48	38.3	155	0.0073	-1.3
N	14.2	11.3	201.56	83.53	14.7	132	0.0073	-0.6
Q	14.2	11.4	223.36	44.03	16.3	146	0.0020	1.3
C	14.2	13.8	224.67	60.24	24.7	121	0.0067	0.5
Y	14.2	15.2	334.20	-59.53	50.0	181	-0.0005	1.6
M	14.2	15.4	113.22	-174.71	34.3	149	0.0088	-0.3
W	14.2	16.5	431.17	-38.99	74.3	204	0.0002	0.6

Table 4.2. *Thermodynamic and evolutionary properties of amino acids.*

higher ΔG_{surf} than the early ones. The means and standard deviations of the groups are 169.3 ± 42.0 and 285.0 ± 94.1. Methionine (M) apparently has a much lower ΔG_{surf} than the other late amino acids. We have used the figure calculated by Amend and Shock without change, but we suspect that there may be an error in this figure, as it should be larger than for cysteine. Cysteine is the only other sulphur-containing amino acid, and it is substantially smaller than methionine. If the methionine ΔG_{surf} were higher, this would increase the difference between the two groups. Table 4.2 also lists the molecular weight (MW) of each amino acid, and the ATP cost, which is the number of ATP molecules that must be expended in order to synthesize the amino acids using the

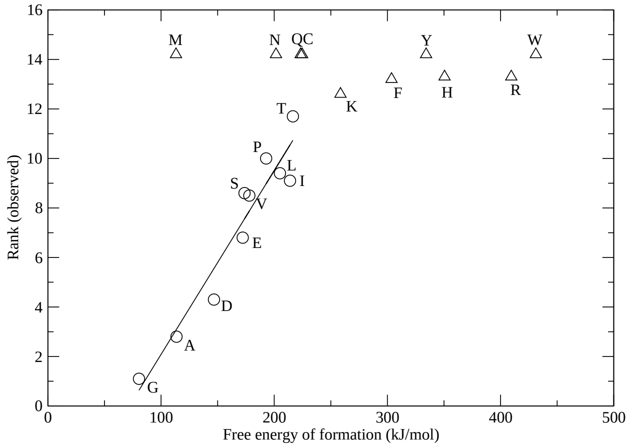

Fig. 4.3. Relationship between the rank of the amino acid and the free energy of formation. Circles - early amino acids. Triangles - late amino acids. The line is the linear regression for the early amino acids.

biochemical pathways in *E. coli* bacteria (Akashi and Gojobori, 2002). The means and standard deviations of MW for early and late groups are 116.2 ± 20.6 and 157.3 ± 23.2, and the figures for ATP cost are 18.5 ± 6.9 and 36.2 ± 17.3. Thus it is clear that the late group are larger and more thermodynamically costly.

All the formation reactions are endergonic ($\Delta G_{surf} > 0$). Those with the smallest ΔG_{surf} should be formed most easily, as they require the least free energy input. Figure 4.3 demonstrates this. The values of ΔG_{surf} are predictive of what we see in a wide range of meteorite and prebiotic synthesis experiments. If the mixture of compounds were in equilibrium, then we would expect the concentrations to depend exponentially on ΔG via the Boltzmann factor $exp(-\Delta G/kT)$. The ranking procedure linearizes this relationship and makes the correlation easier to see. The exponential dependence explains why the amino acids with high ΔG are not seen in experiment: their concentration would be too low to detect. Observed concentrations also depend on the rates of formation and not just on equilibrium thermodynamics. The middle-ranking amino acids show considerable fluctuation between the columns of Table 4.1, being present in some cases and not others. This may reflect differences in rates of synthesis between different experimental conditions. The ranking procedure averages out these fluctuations.

The picture becomes less clear when we consider the free energy of formation under hydrothermal conditions. The central message of Amend and Shock (1998) is that many of the formation reactions are exergonic ($\Delta G_{hydro} < 0$) under hydrothermal conditions, and that even the endergonic ones are less positive than they are at the surface. They use this to explain why hydrothermal vents might be a good place for current life, and to support the idea that the first organisms might have been deep-sea chemosynthesizers. However, there is no correlation between rank and ΔG_{hydro}, and no significant difference in ΔG_{hydro} between early and late amino acids. The two experiments designed to simulate hydrothermal systems (H1 and H2) give results that agree fairly well with the combined ranking from the other data. The ΔG_{hydro} values do not seem to agree with the H1 and H2 experiments any better than they agree with the overall ranking. Although this does not rule out the possibility of a hydrothermal origin of life, what does appear clear is that the 10 early amino acids identified by the ranking procedure can be predicted on the basis of ΔG_{surf} and not ΔG_{hydro}.

4.6 The RNA world and the origin of the genetic code

Having discussed amino acid synthesis, we now want to consider how amino acids became incorporated into proteins, and how protein evolution got underway. To do this we need to make a brief foray into the RNA world. Current life is DNA/protein-based: proteins carry out the majority of catalytic roles in the cell, and DNA stores the genetic information. Proteins are made using the information specified in the DNA sequence (genome) of the cell. Proteins do not have their own hereditary mechanism. Double-stranded DNA does have a hereditary mechanism because each strand can act as a template for synthesis of the complementary strand. In current organisms, DNA synthesis is catalyzed by protein enzymes. Thus, DNA and proteins are mutually dependent, and it is difficult to see how one could have existed without the other. The RNA world hypothesis was proposed as a way out of this chicken and egg dilemma. The RNA world is a period thought to have existed in the early stages of life in which RNA molecules carried out both catalytic and genetic roles. This is supported by the fact that there are many viruses in which the genome is RNA (although no longer any cellular organisms), and by the increasing repertoire of catalytic RNAs that are being synthesized by *in vitro* selection techniques (Joyce, 2002).

It is thought that when organisms evolved the ability to synthesize specific proteins, these proteins took over most of the catalytic roles formerly performed by RNA. Also DNA took over the role of information storage from RNA at some stage. Nevertheless, RNA was retained as an intermediate between DNA and proteins: information in the DNA gene sequence is first transcribed into an RNA sequence and then translated into a protein sequence. There are also a certain number of types of RNAs with key biochemical roles that occur widely in all domains of life, and that are thought to be relics of the RNA world (Jeffares *et al.* 1998). Joyce (2002) likens this situation to a primitive civilization that existed before recorded history, but which left its mark on the modern civilization that followed.

There are still important gaps in our understanding of the RNA world, and it is not known if complex self-replicating RNAs could have originated *de novo*. It may be that there was no single self-replicating RNA, and that life began with a network of reactions of many molecular components (Shapiro, 2006). These components could have been RNAs and/or other types of biomolecules. Nevertheless, we are persuaded that life on Earth passed through an RNA world stage, even if something preceded this. In this article, we will not consider the origin of life itself, but we will focus on a key step that occurred some time later: the origin of translation. Translation is the process by which the sequence information in an RNA molecule is used to make a specific protein. Translation depends on RNA in several ways. Firstly, the messenger RNA contains the sequence information; secondly, ribosomal RNA is the active component of the ribosome - the complex of macromolecules that catalyses formation of proteins (Moore and Steitz, 2002); and thirdly, the transfer RNAs are the ‘adaptors’ that make the link between nucleic acids and amino acids. It is therefore likely that translation originated in organisms in which RNAs played a central role, and in which a mechanism of accurate replication of relatively long RNA sequences had already evolved.

The key to translation is the genetic code. This is the set of assignments between the 64 possible codons (three-letter sequences composed of U, C, A and G nucleotides) in RNA and the 20 possible amino acids in proteins. The genetic code is decoded by complementary base pairing between the codon sequence in the messenger RNA and the anticodon sequence on the transfer RNAs (readers requiring more information should consult any biology textbook). The canonical genetic code shown in Table 4.3 is shared by the three domains of life (archaea, bacteria and eukaryotes). Therefore, it evolved prior to the last universal common

	U	C	A	G	
U	UUU Phe (F)	UCU Ser (S)	UAU Tyr (Y)	UGU Cys (C)	U
	UUC Phe (F)	UCC Ser (S)	UAC Tyr (Y)	UGC Cys (C)	C
	UUA Leu (L)	UCA Ser (S)	UAA Stop	UGA Stop	A
	UUG Leu (L)	UCG Ser (S)	UAG Stop	UGG Trp (W)	G
C	CUU Leu (L)	CCU Pro (P)	CAU His (H)	CGU Arg (R)	U
	CUC Leu (L)	CCC Pro (P)	CAC His (H)	CGC Arg (R)	C
	CUA Leu (L)	CCA Pro (P)	CAA Gln (Q)	CGA Arg (R)	A
	CUG Leu (L)	CCG Pro (P)	CAG Gln (Q)	CGG Arg (R)	G
A	AUU Ile (I)	ACU Thr (T)	AAU Asn (N)	AGU Ser (S)	U
	AUC Ile (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)	C
	AUA Ile (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)	A
	AUG Met (M)	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)	G
G	GUU Val (V)	GCU Ala (A)	GAU Asp (D)	GGU Gly (G)	U
	GUC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)	C
	GUA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)	A
	GUG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)	G

Table 4.3. *The canonical genetic code. Each of the 64 codons is assigned to one of the 20 amino acids (or to a Stop signal).*

ancestor (LUCA). Understanding how it arose is a fundamental question in evolutionary biology.

As prebiotic synthesis of the early group of amino acids seems to be possible in a number of environments, it is likely that these early amino acids were available to RNA-based organisms. Prior to the evolution of the genetic code, amino acids could have played a role in metabolism, and it is possible that polypeptides might also have been synthesized if peptide bond formation was catalyzed by a ribozyme. For more details of possible types of chemistry that could have existed in this period, see Pascal *et al.* (2005). However, before the genetic code evolved, any peptides would have had stochastic amino acid compositions. When the genetic code arose, it became possible to synthesize specific proteins where the amino acid sequence was predetermined by the RNA sequence. Specific proteins can have specific structures and hence specific functions; hence they are much more useful to an organism than stochastic peptides. The invention of the genetic code is thus a stroke of evolutionary brilliance which couples the hereditary mechanism inherent in the template-directed replication of nucleic acids to the catalytic possibilities of proteins. Prior to the genetic code, proteins could not undergo evolution.

Early versions of the code probably used a smaller repertoire of amino acids, each with a larger number of codons. The codon table was gradually divided up into smaller blocks as new amino acids were added. This idea goes as far back as Crick (1968). Each addition would have opened up a whole new world of protein possibilities. Thus there is a selective drive for adding new amino acids in the early stages of code development. This brings us back to question of the order of addition of amino acids to the genetic code. We have shown above that the early group of amino acids can be identified based on their appearance in meteorites, the Miller-Urey experiment and a variety of other chemical syntheses designed to simulate prebiotic conditions. The strong correlation between the ranking, R_{obs} that we derive and the free energy of formation, ΔG_{surf}, suggests that this ranking is a meaningful prediction of the relative frequencies of abiotically synthesized amino acids that would have been available to the first organisms. Essentially the same amino acids are placed early in R_{code} (Trifonov, 2004) as well.

We therefore envisage an early RNA-based organism that learned to use the available amino acids to synthesize proteins in a prescribed way, and found this to be extremely useful. Such an organism would have begun to evolve metabolic pathways to synthesize these amino acids itself from other chemicals so that it was no longer reliant on the prebiotic supply. As this process proceeded, pathways developed for synthesis of the more complex, late group of amino acids that were never present in appreciable quantities prebiotically. At some point along the line, proteins became so useful and so essential to the organisms that the RNA world was replaced by modern DNA/protein based organisms. This scenario has has long been advocated by Wong (1975) as part of what is known as the coevolution theory for the origin of the code. In an updated version of this theory, Wong (2005) proposes the same set of early and late amino acids as us, based on synthesis from atmospheric gases. We have shown above that this set of early amino acids seems likely, even if the place of origin is different.

4.7 How was the genetic code optimized?

There is considerable evidence that the canonical code has evolved to minimize the effect of errors (Freeland *et al.* 2003). Mutations create errors in the gene sequence that are passed to the protein sequences translated from this gene. Translational errors due to codon-anticodon mispairing or mischarging of a tRNA will introduce occasional non-heritable errors into proteins. Both these types of error are likely to involve only one base in the codon most of the time. Errors can be minimized by arranging for codons that differ by only one base to code for the same amino acid or amino acids with similar physical properties. Thus when an error leads to an amino acid replacement, it is usually replaced by a similar amino acid, and the deleterious effect on the protein is minimized. The canonical code has been compared with large numbers of random codes created by reshuffling the amino acid positions in the table of codons. The fraction of random codes for which the average effect of an error is less than in the canonical code is as small as 10^{-6} (Freeland and Hurst, 1998), and can be even smaller than this if the effect of unequal amino acid frequencies is also accounted for (Gilis *et al.* 2001).

It was originally thought that the canonical code was shared by all organisms and that it was frozen and unable to change. However, although the majority of organisms use the canonical code, there are now a large number of cases known where small changes to the code have occurred in specific lineages (Knight *et al.* 2001). There are several different mechanisms by which codon reassignment can occur despite the negative selective effects that are present during the changeover period (Sengupta and Higgs, 2005). The variant codes have all arisen since the establishment of the canonical code. In this article, we are more concerned with the origin of the canonical code itself.

One remarkable pattern in the code is that the 5 highest ranking amino acids, G, A, D, E and V all have a G nucleotide at the first codon position (i.e. they are on the bottom row of Table 4.3). It is possible that only these bottom-row codons were assigned in the first code. Another possibility that we consider likely is that there was a stage in which all codons were assigned to G, A, D (and/or E) and V, and where only the second base was significant for coding (i.e. all first column codons were V, all second column codons were A, etc.). In this four-column pattern, there is still a shift of three bases between successive codons. This means that the code can become more specific later on by assigning significance to first and third positions without destroying the information contained at the second position (Crick, 1968). Later on, prefix and suffix codons may have arisen (Wu *et al.*, 2005), where either the first and second, or the second and third bases were significant. This would have broken up the four columns into smaller divisions.

This proposal is consistent with the observation that the code is error-minimizing. Selection would have acted to ensure each new amino acid was added in a favourable position. This would mean that amino acids would be added into positions that used to be occupied by an earlier amino acid with similar properties. Current work by H. Goodarzi (personal communication) also makes this point. In the first column F, L, I and M are similar to V, and in the second column S, P and T are similar to A. A general property of the code is that amino acids in the same column are much more similar than those in the same row (Urbina *et al.* 2006). This may be a relic of an early four-column stage of the code where all codons in the same column coded for the same amino acid.

Wong (1975, 2005) also supposes that large blocks of codons were divided up into smaller ones as new amino acids were added to the code; however, there are important differences to our argument above. From the pathways of biosynthesis of amino acids in modern organisms it is possible to define precursor-product pairs. Wong argues that later amino acids were assigned to codons that were formerly occupied by their biochemical precursors. The earliest code would have only contained the amino acids that are at the start of the biochemical pathways. Di Giulio and Medugno (1999) show the complex pattern of assignments that would have existed in the early code if every current amino acid were replaced by its earliest precursor. This pattern seems unlikely to us because it would require a very complex molecular recognition process for tRNA charging. The initial four-column pattern that we propose would require straightforward molecular recognition of the second base in the anticodon. The precursor-product argument predicts that precursor-product pairs in the canonical code should occupy neighbouring codons. The statistical significance of this depends on the details of how precursor-product pairs are counted (Ronneberg *et al.* 2000), and is not as high as was previously claimed. Our argument is that later amino acids were added to positions that were formerly occupied by amino acids with similar properties, irrespective of whether the earlier amino acid was a precursor of the later. This predicts that amino acids on neighbouring codons should have similar properties, i.e. the code should be optimized in the sense of Freeland and Hurst (1998). The statistical evidence for this is much more robust, as discussed above.

In contrast to both these theories, it has also been proposed that there were specific interactions between amino acids and their codons or anticodons. Recent evidence for this comes from selection experiments on random RNA pools (Yarus *et al.* 2005). Although this is intriguing, to us it is the error-minimizing properties of the code that most demand an explanation. An error-minimizing code can be reached by a pathway in which natural selection acts each time a new amino acid is added. Precursor-product relationships and RNA-amino acid interactions might influence which codons were most likely to 'tried out' for a new amino acid, but selection would eliminate those trial codes in which an amino acid was added in 'the wrong place', and would favour the addition of a new amino acid in a place that was consistent with its physical properties. Thus our principle for amino acid addition explains how the code became optimized.

4.8 Protein Evolution

The previous section envisages early proteins composed of a small set of amino acids. Several studies have considered whether such proteins could be functional. Using only three amino acids (Q, L, and R), proteins with strong helical structure were found (Davidson *et al.*, 1995). Doi *et al.* (2005) found that random sequences of the five most frequent early amino acids according to our ranking, GADEV, were more soluble than random sequences of 20 amino acids or QLR proteins. Riddle *et al.* (1997) took a 57-residue structural domain of a naturally occurring protein and found that almost all residues could be replaced by a an amino acid from a simple alphabet of IKEAG. The smallest amino acids, G and A, were essential parts of this alphabet. The large aromatic amino acids, W, F and Y, that are late according to our theory, were not able to be replaced by amino acids from the simple alphabet. Babajide *et al* (1997) used inverse-folding programmes to locate sequences that are likely to fold to a specified natural protein structure. Suitable sequences could be found for some very small alphabets (including ADLG) but not others (including QLR). Taken together, these studies suggest that proteins composed of the early group of amino acids could well have formed similar structures to modern day proteins and could have performed useful functions in early organisms.

Another unusual proposal that is worth mentioning is the GADV-protein world hypothesis of Ikehara (2005). The same amino acids are again first in this theory, but proteins composed of GADV evolve replication and metabolism prior to the origin of RNA. It is not clear to us how genetic information could be passed on by proteins alone, however.

We considered above the role that thermodynamics has in prebiotic synthesis of amino acids. It also seems likely that thermodynamics effects the evolution of modern proteins. Akashi and Gojobori (2002) showed that organisms appear to be sensitive to the ATP cost of amino acid synthesis (listed in Table 4.2). Specifically, the most highly expressed proteins had the highest frequecy of the low-cost amino acids. Seligmann (2003) has also found considerable evidence that amino acid usage in proteins is affected by cost minimization. Gutierrez *et al.* (1996) observed an increase in frequency of the base G in the first codon position in highly expressed genes, which can be interpreted as selection for low-cost amino acids GADEV. Alternatively, it might be that GNN codons are preferred because they reduce the degree of frameshifting errors (Trifonov, 1987). This might be another reason to think that the earliest codons were GNN.

It is possible that modern day proteins may still contain a signal of which amino acids were frequent in the early stages of evolution. Phylogenetic methods can be used to reconstruct estimates of ancestral sequences and hence to determine the frequencies of amino acids in those sequences. Brooks *et al.* (2002) estimated frequencies of amino acids in the LUCA, and found them to be noticeably different from current proteins. In Table 4.2, $\Delta p = p_{LUCA} - p_{current}$. Jordan *et al.* (2005) use a method that involves comparing sequences from triplets of related species. The earliest diverging species in each triplet can be used to give a direction to the amino acid substitutions occurring on the branches leading to the other two species. It is found that forward and backward substution rates between pairs of amino acids are not equal, and that amino acids frequencies show an increasing or decreasing trend. dp/dt is the estimate of the rate of change of the amino acid frequencies.

There are some limited points of agreement between dp/dt and Δp. The most rapidly decreasing amino acid is A in both cases, and the three most frequent early amino acids, G, A and D, are all decreasing. Beyond that, there is little agreement: only 7 amino acids have the same sign in both studies. We would expect early amino acids to be decreasing and late amino acids to be increasing. 13 amino acids have the right sign according to Δp, and 14 have the right sign according to dp/dt. This is suggestive, but not fully convincing. Recently, McDonald (2006) has cast doubt on the Jordan *et al.* (2005) study on methodological grounds, and a second study of Brooks *et al.* (2004) produced results that differ from their previous study (Brooks *et al.*, 2002) as to which amino acids are increasing and decreasing. Our position is that trends in amino acid frequencies like this would have occurred over time and it makes sense to look for them. However, it is a long time since the LUCA and it may not be possible to see convincing trends above the noise.

One observation of Brooks *et al.* (2004) is relevant to the extremophile theme of this book. They found that the estimated frequencies of amino acids in the LUCA were closer to those of hyperthermophiles than mesophiles, which might suggest that LUCA was a hyperthermophile. It has also been argued that the genetic code was established at high temperature (Di Giulio, 2000) and high pressure (Di Giulio, 2005) based on comparison of amino acid frequencies in extremophiles and mesophiles. The hyperthermophile nature of LUCA is still controversial, and some cold water is poured on this idea by Brinkmann *et al.* in this book.

4.9 Summary

In this article, we have done our best to synthesize information from astrophysics, organic chemistry, molecular evolution and bioinformatics. It became apparent to us that the strands of knowledge that link these diverse disciplines are becoming ever stronger, and are indeed required in order to understand how life may have arisen in our solar system, and perhaps in others. The generality of planet formation and the universality of the organic chemistry (that is just starting to be explored) in protoplanetary disks is very encouraging. It suggests that the astrophysics of protoplanetary disks could play a major role in understanding how one goes from dust and gas, to watery worlds equipped with the biomolecules that may have served as the first building blocks of life. Beyond this, we have also shown that the abundance and distribution of early amino acids may have played an important role in the origin of the earliest genetic code and how it may have subsequently evolved.

We thank the many excellent speakers at the Origins Institute's 2005 conference on "Astrobiology and the Origins of Life" for the inspiration for this article, and book.

References

Akashi, H. and Gojobori, T. (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillis subtilis. Proc. Nat. Acad. Sci. USA* **99**, 3695-3700.

Amend, J.P. and Shock, E.L. (1998) Energetics of amino acid synthesis in hydrothermal ecosystems. *Science* **281**, 1659-1662.

Babjide, A., Hofacker, I.L., Sippl, M.J. and Stadler, P.F. (1997) Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Folding and Design.* **2**, 261-269.

Bar-Nun, A., Nar-Nun, N., Bauer, S.H. and Sagan, C. (1970) Shock synthesis of amino acids in simulated primitive environments. *Science* **168**, 470-473.

Bergin, E.A., Aikawa, Y., Blake, G.A., and van Dishoeck, E.F. (2006) The chemical evolution of protoplanetary disks. *Protostars and Planets V*, B. Reipurth, D. Jewitt, and K. Keil (eds.) (University of Arizona Press, Tucson), in press.

Bernstein, M.P., Dworkin, J.P., Sandford, S.A., Cooper, G.W. and Allamandola, L.J. (2002) Racemic amino acids from the ultraviolet photolysis of interstellar ice analogues. *Nature*, **416**, 401-403.

Brooks, D.J., Fresco, J.R., Lesk, A.M. and Singh, M. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. (2002) *Mol. Biol. Evol.* **19**, 1645-1655.

Brooks, D.J., Fresco, J.R., and Singh, M. (2004) A novel method for estimating ancestral amino acid composition and its application to proteins of the last universal ancestor. *Bioinformatics* **20**, 2251-2257.

Chiang, E.I. and Goldreich, P. (1997) Spectral energy distributions of T Tauri stars with passive circumstellar disks *Ap. J.* **490**, 368 - 376.

Chakrabarti,S. and Chakrabarti, S.K. (2000) Can DNA bases be produced during molecular cloud collapse? *Astron. Astrophys.* **354**, L6-L8.

Chyba, C.F. and Sagan, C. (1992) Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life. *Nature* **355**, 125-132.

Crick, F.H.C. (1968) The origin of the genetic code. *J. Mol. Biol.* **38**, 367-379.

Cronin, C.R. and Moore, C.B. (1971) Amino acid analyses of the Murchison, Murray and Allende carbonaceous chondrites. *Science* **172**, 1327-1329.

D'Alessio, P., Canto, J., Calvet, N., and Lizano, S. (1998) Accretion disks around young objects I: the detailed vertical structure. *Ap. J.* **500**, 411-427.

Davidson, A.R., Lumb, K.J. and Sauer, R.T. (1995) Cooperatively folded proteins in random sequence libraries. *Nature Struct. Biol.* **2**, 856-864.

Di Giulio, M. (2000) The late stage of genetic code structuring took place at high temperatures. *Gene* **261**, 189-195.

Di Giulio, M. (2005) The ocean abysses witnessed the origin of the genetic code. *Gene* **346**, 7-12.

Di Giulio, M. and Medugno, M. (1999) Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. *J. Mol. Evol.* **49**, 1-10.

Doi, N., Kakukawa, K., Oishi, Y. and Yanagawa, H. (2005) High solubility of random sequence proteins consisting of five kinds of primitive amino acids. *Prot. Eng. Design and Selection*, **18**, 279-284.

Dubruelle, C.P., Hollenbach, D., Kamp, I., and D'Alessio, P. (2006) Models of the structure and evolution of protoplanetary disks. *Protostars and Planets V*, B. Reipurth, D. Jewitt, and K. Keil (eds.) (University of Arizona Press, Tucson), in press.

Dutrey, A., Guilloteau, S., and Guelin, M. (1997) Chemistry of protosolar-like nebulae: the molecular content of the DM Tau and GG Tau disks.

Astron. Astrophys. **317**, L55-L58.

Dutrey, A., Guilloteau, S., and Ho, P. (2006) Interferometric spectro-imaging of molecular gas in proto-planetary disks. *Protostars and Planets V*, B. Reipurth, D. Jewitt, and K. Keil (eds.) (University of Arizona Press, Tucson), in press.

Engel, M.H. and Nagy, B. (1982) Distribution and enantiomeric composition of amino acids in the Murchison meteorite. *Nature* **296**, 837-840.

Freeland, S.J. and Hurst, L.D. (1998) The genetic code is one in a million. *J. Mol. Evol.* **47**, 238-248.

Freeland, S.J., Wu, T., Keulmann, N. (2003) The case for an error minimizing standard genetic code. *Origins Life Evol. Biosph.* **33**, 457-477.

Gaidos, E. and Selsis, F. (2006) From protoplanets to protolife: the emergence and maintenance of life. *Protostars and Planets V*, B. Reipurth, D. Jewitt, and K. Keil (eds.) (University of Arizona Press, Tucson), in press.

Gilis, D., Massar, S., Cerf, N.J. and Rooman, M. (2001) Optimality of the genetic code with respect to protein stability and amino acid frequencies. *Genome Biol.* **2(11)** 49.1-49.12.

Gutierrez, G., Marquez, L. and Marin, A. (1996) Preference for guanosine at first position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. *Nucl. Acids. Res.* **24**, 2525-2527.

Hennet, R.J.C., Holm, N.G. and Engel, M.H. (1992) Abiotic cynthesis of amino acids under hydrothermal conditions and the origin of life: a perpetual phenomenon. *Naturwissenschaften* **79**, 361-365.

Ikehara, K. (2005) Possible steps to the emergence of life: The GADV-protein world hypothesis. *Chemical Record* **5**, 107-118.

Jeffares, D.C., Poole, A.M. and Penny, D. (1998) Relics from the RNA world. *J. Mol. Evol.* **46**, 18-36.

Jordan, I.K., Kondrashov, F.A., Adzhubel, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S. and Sunyaev, S. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**, 633-637.

Joyce, G.F. (2002) The antiquity of RNA-based evoltion. *Nature*, **418**, 214-221.

Kasting, J.F. (1993) Earth's early atmosphere. *Science* **259**, 920-926.

Kokubo, E. and Ida, S. (2002) Formation of protoplanet systems and diversity of planetary systems. *Ap. J* **581** 666-680

Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001) Rewiring the keyboard: evolvability of the genetic code. *Nature Rev. Genet.* **2**, 49-58.

Kvenvolden, K.A., Lawless, J.G. and Ponnamperuma, C. (1971) Nonprotein amino acids in the Murchison meteorite. *Proc. Nat. Acad. Sci. USA* **68**, 486-490.

Lazcano, A. and Miller, S.L. (1996) The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time. *Cell* **85**, 793-798.

Lin, D.N.C. and Papaloizou, J.C.B. (1993) On the tidal interaction between protostellar disks and companions. *Protostars and Planets III*, p. 749-835.

Lowe, C.U., Rees, M.W. and Markham, R. (1963) Synthesis of complex organic compounds from simple precursors: formation of amino acids, amino-acid polymers, fatty acids and purines from ammonium cyanide. *Nature* **199**, 219-222.

MacLow M.-M., and Klessen, R.S. (2004), Control of star formation by supersonic turbulence. *Rev. Mod. Phys.* **76**, 125-195.

Marshall, W.L. (1994) Hydrothermal synthesis of amino acids. *Geochim. Cos-*

mochim. Acta **58**, 2099-2106.

Matsumura, S., Pudritz, R.E., and Thommes, E.W. (2006) Saving planetary systems: dead zones and planetary migration. *ApJ* , submitted.

Mayer, L., Quinn, T., Wadsley, J., and Stadel, J. (2002) Formation of giant planets by fragmentation of protoplanetary disks. *Science* **298**, 1756-1759.

McCaughrean, M.J., and Stauffer, J.R. (1994) High resolution near-infrared imaging of the trapezium: a stellar census. *Astron. J.* **108**, 1382-1397.

McDonald, J.H. (2006) Apparent trend of amino acid gain and loss in evolution due to nearly neutral variation. *Mol. Biol. Evol.* **23**, 240-244.

Meyer, M.R., Backman, D.E., Weinberger, A.J., and Wyatt, M.C. (2006) Evolution of circumstellar disks around normal stars: placing our solar system in context. *Protostars and Planets V*, B. Reipurth, D. Jewitt, and K. Keil (eds.) (University of Arizona Press, Tucson), in press.

Miller, S.L and Orgel, L.E. (1974) *The Origins of Life on the Earth* (Prentice-Hall, Englewood Cliffs, New Jersey).

Miyakawa, S., Yamanashi, H., Kobayashi, K., Cleaves, H.J. and Miller, S.L. (2002) Prebiotic synthesis from CO atmospheres: implications for the origins of life. *Proc. Nat. Acad. Sci. USA* **99**, 14628-14631.

Moore, P.B. and Steitz, T.A. (2002) The involvement of RNA in ribosome function. *Nature*, **418**, 229-235.

Morbidelli, A., Chambers, J., Lunine, J.I., Petit, J.M., Robert, F., Valsecchi, G.B., and Cyr, K.E. (2000) Source regions and time scales for the delivery of water to Earth *Meteorit. Planet. Sci.* **35**, 1309-1320.

Munoz Caro, G.M., Meierhenrich, U.J., Schutte, W.A., Barbier, B., Arcones Segovia, A., Rosenbauer, H., Thiemann, W.H.P., Brack, A. and Greenberg, J.M. (2002) Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature.* **416**, 403-406.

Natta, A., Testi, L., Calvet, N., Henning, T., Waters, R., and Wilner, D. (2006) Dust in proto-planetary disks: properties and evolution. *Protostars and Planets V*, B. Reipurth, D. Jewitt, and K. Keil (eds.) (University of Arizona Press, Tucson), in press.

Pascal, R., Boiteau, L. and Commeyras, A. (2005) From the prebiotic synthesis of α-amino acids towards a primitive translation apparatus for the synthesis of peptides. *Top. Curr. Chem.* **259**, 69-122.

Pierazzo, E. and Chyba, C.F. (1999) Amino acid survival in large cometary impacts. *Meteoritics Planet. Sci.* **34**, 909-918.

Pollack, J.B., Hubickyj, O., Bedenheimer, P., Lissauer, J.J., Podolak, M., and Greenzweig, Y. (1996) Formation of the giant planets by concurrent accretion of solids and gas. *Icarus* **124**, 62-85.

Raymond, S.N., Quinn, T., and Lunine, J.I. (2004) Making other earths: dynamical simulations of terrestrial planet formation and water delivery. *Icarus* **168**, 1-17.

Riddle, D.S., Santiago, J.V., Bray-Hall, S.T., Doshi, N., Grantcharova, V.P., Yi, Q. and Baker, D. (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805-809.

Ronneberg, T.A., Landweber, L.F. and Freeland, S.J. (2000) Testing a biosynthetic theory of the genetic code: fact or artifact? *Proc. Nat. Acad. Sci. USA* **97**, 13690-13695.

Schreyer, K., Semenov, D., Henning, T., Pavlyuchenkov, Y., and Dullemond, C. (2005), The massive disk around the young B-star AFGL 490. *IAU Symposium 231: Astrochemistry - Recent Successes and Current Chal-*

lenges.

Seligmann, H. (2003) *J. Mol. Evol.* Cost-minimization of amino acid usage. **56**, 151-161.

Sengupta, S. and Higgs, P.G. (2005) A unified model of codon reassignment in alternative genetic codes. *Genetics* **170**, 831-840.

Sephton, M.A. (2002) Organic compounds in carbonaceous meteorites. *Nat. Prod. Rep.* **19**, 292-311.

Shapiro, R. (2006) Small molecule interactions were central to the origin of life. *Quart. Rev. Biol.* **81**, 105-125.

Shimoyama, A., Ponnamperuma, C. and Yanai, K. (1979) Amino acids in the Yamato carbonaceous chondrite from Antarctica. *Nature* **282**, 394-396.

Simon, M., Dutrey, A., and Guilloteau, S. (2000) *Ap.J.* **545**, 1034-1043.

Tilley, D.A. and Pudritz, R.E. (2004) The formation of star clusters I: simulations of hydrodynamic turbulence. *Mon. Notices Roy. Astron. Soc.* **353**, 769-788.

Trifonov, E.N. (1987) Translation framing code and frame-monitoring mechanism as suggested by the analysis of messenger RNA and 16S ribosomal RNA nucleotide sequences. *J. Mol. Biol.* **194**, 643-652.

Trifonov, E.N. (2000) Consensus temporal order of amino acids and the evolution of the triplet code. *Gene* **261**, 139-151.

Trifonov, E.N. (2004) The triplet code from first principles. *J. Biomol. Struct. Dynam.* **22**, 1-11.

Urbina, D., Tang, B. and Higgs, P.G. (2006) The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and the structure of the genetic code. *J. Mol. Evol.* **62**, 340-361.

van Zadelhoff, G.-J., Aikawa, Y., Hogerheijde, M.R., and van Dishoeck, E.F. (2003) Axi-symmetric models of UV radiative transfer with applications to circumstellar disk chemistry. *Astron. Astrophys.* **397**, 789-802.

Weber, A.L. and Miller, S.L. (1981) Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* **17**, 273-284.

Whittet, D.C.B. (1997) Is extraterrestrial organic matter relevant to the origin of life on Earth? *Orig. Life Evol. Biosph.* **27**, 249-262.

Wolf, S. and D'Angelo, G. (2005) On the observability of giant protoplanets in circumstellar disks. *Ap. J.* **619** 1114-1122.

Wong, J.T. (1975) A coevolution theory of the genetic code. *Proc. Nat. Acad. Sci. USA* **72**, 1909-1912.

Wong, J.T. (2005) Coevolution theory of the genetic code at age thirty. *BioEssays* **27**, 416-425.

Wu, H.L., Bagby, S. and van den Elsen, J.M.H. (2005) Evolution of the genetic code via two types of doublet codons. *J. Mol. Evol.* **61**, 54-64.

Yarus, M., Caporaso, J.G. and Knight, R. (2005) Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem.* **74**, 179-198.

Yoshino, D., Hayatso, R. and Anders, E. (1971) Origin of organic matter in the early solar system - III. Amino acids: catalytic synthesis. *Geochim. Cosmochim. Acta* **35**, 927-938.