Probabilities of randomly assembling a primitive cell on Earth

Dermott J. Mullan, mullan@bartol.udel.edu

Summary

We evaluate the probability P_r that the RNA of the first cell was assembled randomly in the time available (1.11 billion years [b.y.]). To do this calculation, we first set a strict upper limit on the number of chemical reactions n_r which could have occurred before the first cell appeared.

In order to illustrate the consequences of the finite value of n_r, we make some extremely minimalist assumptions about cells. We consider a cell composed of $N_p = 12$ proteins, each containing $N_a = 14$ amino acids. We refer to the minimum (N_p , N_a) set as a (12-14) cell. Such a cell is smaller than some modern viruses.

The ability to perform any of the basic tasks of the cell is not necessarily limited to a single protein. Many different proteins among all those which were available in the primeval soup may have been able to perform (say) waste disposal. In order to allow for this in estimating P_r, we include a factor Q to describe how many different proteins in the primeval soup could have performed each of the basic tasks of cell operation. The larger Q is, the easier it is to assemble a functional cell by random processes. However, there is a maximum value Q_{max} that is set by phase space arguments.

The hypothesis that life originated by random processes requires that P_r be of order unity. We estimate how large Q must be (Q_{ra} : subscript "ra" denotes "random assembly") in order to ensure $P_r = 1$ in the time that is available (1.11 b.y.). We find that Q_{ra} must be so large as to exceed the maximum permissible value Q_{max} in the phase space of proteins comprised of a set of 14 distinct amino acids. Such a large value of Q_{ra} would have serious consequences for biology: if Q_{ra} were truly as large as Q_{max} in the primeval soup, then essentially all 14-acid proteins must have possessed the ability to perform each of the fundamental tasks in the cell. That is, there was no task specificity among the proteins: a protein which was able (say) to maintain the membrane in a cell would also have been able to control (say) the replication process.

In such a situation, the very concept of a cell, as a well-organized factory in which the task of each department is regulated, and each department must coordinate dependably with all

others, would no longer be valid. A cell would quickly be reduced to an unpredictable entity which lacked robust properties.

In the "real world", where a cell must be able to preserve itself and replicate faithfully from generation to generation, it seems inevitable that the various proteins must be prevented (by nature) from performing multiple tasks. That is, there must be a certain amount of specificity to the task that any given protein can perform: of all available proteins, only a fraction F should be capable of performing the task of (say) membrane repair. In a cell where the number of proteins is N_p, the restraints of specificity require that the value of F can certainly not exceed $1/N_p$. But F might be much smaller than this upper limit. This leads us to introduce a "protein specificity index" m such that the actual value of F in the primeval soup is usefully written as $(1/N_p)^m$. In the modern world, the value of m ranges from 1 to a maximum value between about 10 and 20.

We find that, even assigning the minimum possible specificity (m = 1), the probability P_r of assembling the RNA of a (12-14) cell by random processes in 1.11 billion years using triplet codons is no more than one in 10^{79}. And if the protein tasks are even marginally specific (with m = 2-3, say), the chances of random assembly of RNA for the first cell decreases to less than one in 10^{100}.

In order to improve the chances of random assembly of the first cell, we consider a situation which might have existed in the young Earth. We suppose that proteins could be constructed using a smaller set (numbering N_{aa}) of distinct amino acids: we consider the case of N_{aa} = 5 (instead of the modern 20). If, in these conditions, the number of bases in DNA remained as large as 4, then doublet codons sufficed to encode protein production with the same amount of error protection as occurs in the modern (triplet) genetic code. In such conditions, the probability of randomly assembling the RNA for the first cell in 1.11 b.y. improves. However, it is still small: the optimal probability is no more than one in 10^{63}.

To improve the probability even further, it is tempting to consider the possibility of singlet codons. But we point out that these are not relevant in a realistic biology.

In the context of doublet codons, we can improve the probability P_r of random assembly by considering a larger set of distinct amino acids. The number of distinct amino acids for which doublet-codons can encode ranges from 5 to 14 (allowing for start and stop codons). As N_{aa} increases above 5, there is a marked improvement in

P_r for a (12-14) cell: in fact, P_r may approach a value of order unity when N_{aa} = 11 provided that the specificity index m is smaller than 1.3. (This is far below the average value of m, and represents very marginal specificity.) And P_r formally exceeds unity for N_{aa} in the range from 12 to 14, provided that m does not exceed 2.5. This value of specificity is still well below the average value. It is not clear that a functioning cell could survive for long with such low protein specificities. Nevertheless, the fact that P_r formally reaches a value as large as unity suggests that we may have found a window of opportunity for random assembly of the first (12-14) cell.

However, these cells face a potentially fatal problem: even with 11 amino acids to be encoded by 16 codons in the RNA, there is little redundancy in the genetic code. And for N_{aa} = 14, the redundancy vanishes altogether. As a result, there is a much reduced error protection in the code which translates the information in RNA to proteins. In the limit N_{aa} = 14, there is no error protection at all: transcription from RNA to protein then has no immunity against noise. Moreover, in the limit N_{aa} = 14 (plus a start and stop), proteins would be equally able to encode for RNA, in violation of the Central Dogma of biology. Therefore, although the probability of randomly assembling the RNA for a (12-14) cell in such a world may approach unity in a mathematical sense, it is not clear how useful such a cell would be for biology.

We stress that our assumptions about a (12-14) cell are minimalist in the extreme. In the "real world", it is not obvious that a protein containing only 14 peptides will be able to fold into a stable 3-dimensional shape at the temperatures where water is liquid. And in the "real world", a cell probably requires as many as 250 proteins to function. In such case, even if N_{aa} = 14, P_r approaches unity only if the specificity index m lies in the very restricted range between 1.0 and 1.17. We identify this as a narrow window of opportunity for random assembly of primitive cells. But even this narrow window closes altogether if our estimate of the number of chemical reactions is too large by several orders of magnitude (as it may well be).

Our calculations refer only to the assembling of a cell in which the genetic code is already at work. We do not address the origin of the genetic code itself.

We conclude that, even if we assume that the genetic code was already in existence (by some unspecified mechanism), conditions in the early Earth must have been "finely tuned" in order to

"squeeze through" the narrow window of opportunity and assemble the first cell on Earth in a truly random manner.

1. Introduction

Evolution theory claims that all species of animals and plants that now exist on Earth came into existence as a result of random variations in pre-existing species. It is presumed that life on Earth began as a single cell. An essential aspect of evolution theory is that the first living cell originated in the early Earth also as a result of random processes.

When Darwin proposed his theory of evolution, he did not know the chemical make-up of a cell. Therefore, when he appealed to random processes at work in nature, he could be excused for not knowing what exactly was entailed in such processes. But in our day and age, advances in microbiology and biochemistry have opened up to us the molecular details of the processes that occur in living cells. For example, we now know the make-up of proteins and DNA. In fact, we will need to describe these in some detail in order to proceed with our discussion of the probability of random formation. (We will return to these details below.)

We are now in a position to spell out the chemical processes that must have occurred if the first cell was indeed put together by chance.

2. The challenge of creating the first cell

The question we wish to examine here is the following. If the process of assembling the first cell occurred in a truly random manner in the early Earth, what conditions would be needed?

To address this question, we need to answer two more basic questions: (a) how much time was available before the first cell appeared? And (b) how many chemical reactions of the correct type could have occurred in the time available? The aim here is to answer these questions as quantitatively as possible.

The answer to question (b) will set a limit on the properties of the first cell that would have been created by random processes in the early Earth.

We turn first to the question of how much time was available for the development of the first living cell.

3. The earliest life forms on Earth

The fossil record indicates that the first life forms to appear on Earth existed some 3.45 billion years ago. These are cyanobacteria (formerly called blue-green algae) which are found in rocks from Apex Chert, Australia.). The first life forms on Earth were single-cell organisms. (See http://www.uni-muenster.de/GeoPalaeontologie/Palaeo/Palbot/seite1.html

It is hardly surprising from an evolutionary standpoint that the earliest forms of life on Earth were single-cell organisms. Presumably it is easier for random processes to give rise to a single cell first, before bringing forth a multi-cell organism.

4. How much time elapsed before the first cell appeared on Earth?

The age of the Earth, based on radioactive dating of rocks, is estimated to be 4.56 billion years old. Comparing this with the cyanobacteria ages, we see that the first living cells appeared within a time interval of 1.11 billion years of the formation of the Earth.

Therefore, the time t_{fc} required for the development of the first cell on the Earth is certainly no longer than 1.11 billion years.

Actually, the value of t_{fc} might be much shorter than this. Astronomers who calculate the internal structure of the Sun find that the Sun has not always been as luminous as it is today: the young Sun is calculated to have had a luminosity that is some 20-30 percent fainter than it is today. Therefore, the mean temperature on the early Earth might have been considerably colder than it is today, so cold that the water on Earth's surface was frozen. (This is the "faint young Sun problem": Sagan and Mullen, 1972, Science vol. 177, 52).

It is likely that the development of life requires water to be in liquid form. The solar structure calculations suggest that the energy provided by the Sun to the Earth might not have become sufficient to melt the ice until the Sun was about 700 hundred million years old. This means that the first living cell appeared no more than about 400 million years after liquid water became available.

Moreover, the early Earth would have been subject to a more or less heavy bombardment by the debris of the proto-planetary disk

before the latter was finally cleared out. The impacts of planetesimals (such as that which destroyed the dinosaurs some 60 Myr ago) would have interrupted the processes which were "trying to form" the first cell. Large impacts might have reduced the interval for assembling the first cell to even less than 400 million years.

However, in order to improve the chances of evolution, let us grant a full 1.11 billion years and ask the question: could the first cell have developed by random processes in 1.11 billion years?

The number of seconds of time in 1.11 billion years is $3.5X10^{16}$. We will need this number in what follows.

5. Some essential constituents of cells

Now that we know how much time is available, we move on to the main question that we wish to address: how was the first living cell formed? Evolution theory asserts that it was formed by random processes. We wish to assess the probability of such processes.

To assess realistically the chances of assembling the first cell by chance, we need to know certain fundamental properties of the components that go to make up a cell. Let us first summarize these.

5.1. What do we need to know about proteins?

There are three levels of structure within a protein which are relevant to us here.

(a) Primary Structure

A protein consists of a series of amino acids that are linked (by peptide bonds) into a chain in a specific order. The change of even a single amino acid in a chain of dozens or hundreds of amino acids may in certain cases disrupt the functioning of the protein.

(b) Secondary structure

In order that proteins may function, the primary structure (i.e. a chain of amino acids) is not sufficient. Certain segments of the amino acids in the chain group themselves together into sub-units known as alpha-helixes, beta-sheets, and beta-turns. For example,

an alpha-helix consists of a chain of consecutive amino acids arranged in a twisted three-dimensional structure (including 3.6 acids per turn of the helix) with well-defined angles between neighboring acids in the chain.

These well-defined sub-units form the secondary structure of the protein: they are stable and rigid, like "lego" blocks which can be "fitted together" into a larger structure.

(c) Tertiary structure

Once the "lego" blocks are available, the stage is set for the protein to go beyond the secondary structure: using available thermal energy, the protein twists and folds itself into a certain 3-dimensional structure with specific bumps and hollows. These bumps and hollows, which are referred to as the tertiary structure of the protein, determine where electric charge builds up, and these localized charges control the protein's function, including the reactions that it can catalyze (if it is an enzyme). For example, insulin (one of the shortest proteins in the human body, with 51 amino acids) folds itself naturally into a wedge-like shape which enables groups of six insulin molecules to pack themselves tightly into spherical clusters.

The sequence of amino acids in a particular protein may be highly specific at certain locations. There are certain sites in the protein ("invariant sites") where even a single alteration in the sequence can lead to drastic changes in the shape of the folded protein, thereby disabling the protein. For example, human hemoglobin, the protein that carries oxygen through the blood, contains $N_a = 574$ amino acids arranged in four secondary sub-units, with an overall spherical tertiary structure. Two of the invariant sites in hemoglobin have attracted widespread attention because of the drastic consequences they may have in a certain segment of the population. If one of the amino acids (glutamic acid) in a certain position in two of the sub-units of the hemoglobin molecule is replaced by another amino acid (valine), the result is the painful and deadly disease known as sickle cell anemia. Although it would seem that switching only 2 out of 574 amino acids ought to have an insignificant effect, this is not the case for these two particular sites. Just by changing 2 amino acids and leaving all the remaining 572 as before, the process of folding the molecule is altered so much that the 3-dimensional shape of the hemoglobin changes is no longer spherical. Instead the molecule takes on an elongated structure resembling a sickle.

There are some proteins in which essentially all sites are invariant. For example, histones which have at least 125 amino acids in the peptide chain, have 122 invariant sites. Such proteins are therefore exceedingly specific in the arrangement of amino acids.

However, not all sites in all proteins are invariant. In many proteins, there are sites where the amino acid can be replaced by a number of other amino acids without affecting the functioning of the protein. Yockey (*Information Theory and Molecular Biology*, 1992, Cambridge Univ. press, 408 pp; Table 6.3) discusses the example of a particular protein (iso-1-cytochrome c, with 110 amino acids), with a list of all amino acids which are functionally equivalent at each site. Some sites can have up to 13 different amino acids and still the protein retains functionality, whereas others (the invariant sites) must contain one and only particular amino acid in order to protect against protein dysfunction.

At the primary level, the linear sequence of amino acids in a protein is important to the proper operation of a living cell. But in order to reach the final operating stage (which is fully three-dimensional), the creation of the "lego" blocks (i.e. stable and reproducible secondary structure) is an essential intermediate stage.

(d) How long are the secondary structures?

A central question in the present context is: what is the minimum requirement for the "lego" blocks to be formed? What does it take to be able to create the rigid sub-units which are used in making the final protein? The answer is found in the quantum chemistry of an alpha-helix and a beta-sheet: in principle, a sequence of at least 4 amino acids is required in order to make the smallest alpha helix (this allows for one complete turn of the helix). The minimum size of a beta-sheet may be comparable.

However, the minimum size is not the only factor that is at work in creating the "lego" blocks in proteins: the question of stability also enters, because it is a fundamental requirement for living cells that the secondary structures must be rigid. Otherwise, the shapes of proteins in a cell would be subject to chaotic fluctuations. Studies of reproducible structure of sub-sequences in proteins suggests that chains of at least 7 amino

acids are required in order to create a stable and reproducible "lego" (Sudarsanam and Srinivasan, 1996, abstract E0274, IUCR Seattle meeting). It therefore seems unlikely that stable "lego" blocks can be constructed with a chain that is less than 7 amino acids long.

Now, the tertiary structure of a protein comes into existence only if at least *two* stable "lego" blocks are joined together in a reproducible 3-dimensional structure. (Many proteins require more than 2 secondary structures: e.g. hemoglobin contains 4.) Thus, the bare minimum requirement for a protein is that N_a should be at least twice the bare minimum needed for rigid and stable secondary structure. According to the estimates of Sudarsanam and Sreenivasan, this requires $N_a = N_{min} = 14$.

We emphasize that this assumption of a mere 14 amino acids in a functioning protein is extreme. A protein with only 14 amino acids is very short in terms of the proteins that exist either in the modern world (e.g. insulin, with its 51 amino acids, and hemoglobin, with its 574 amino acids), or even in ancient proteins. For example, bacterial ferrodoxins, with at last 56 amino acids, "are believed to date nearly to the time of the origin of life, and the histones which are also believed to be ancient and have at least 125 amino acids" (Yockey, p. 143). Even in the earliest stages of life on the planet, before the so-called "breakthrough organism" had appeared, the proteins that might have been operational back then have earned the title of "mini-proteins" because the number of amino acids they contained was "perhaps 20 or shorter" (Maniloff, Proc. Natl. Acad. Sci. USA, vol. 93, p. 10004, 1996).

Computational attempts to "construct" proteins which are capable of folding into a certain unique and stable tertiary structure have been made by several groups. Dahiyat and Mayo (Science vol. 278, p. 82, 1997) found that, using only amino acids which occur in modern nature, the shortest protein without sulfides or metals that folds into a stable tertiary structure contains 25 amino acids. An earlier computation (Struthers et al., Science vol. 271, p. 342, 1996) had obtained a stable tertiary structure with a chain of only 23 amino acids: however, one of the 23 was a non-natural amino acid. It seems that polypeptide chains with fewer than 23-25 amino acids can probably not create the tertiary structure which is key to protein function unless they are assisted by sulfides or metals.

How far below the 23-25 limit can a functional protein go when assisted by sulfides and metals? The answer is not clear. However,

it seems unlikely that the limit will be reduced below 14, which is our limit based on the stability properties of at least two "lego" pieces (alpha-helices and beta-sheets). In fact, in terms of the thermal energy which is available, it is not clear that a protein as short as 14 amino acids will be "foldable" or "bendable" at temperatures where water is liquid.

Nevertheless, in the spirit of optimizing probabilities, we assume that polypeptides in the primeval soup could indeed function as proteins while containing no more than 14 amino acids.

5.2. What do we need to know about DNA?

DNA is a molecule that has the shape of a long twisted ladder (the "double helix"). In this ladder, there are "rungs" connecting the long "sidepieces". The "sidepieces" are long linear chains of sugars and phosphates, while each "rung" is composed of two interlocking bases. The four bases consist of two purines and two pyrimidines. The bases in the ladder are arranged in a definite order, just as amino acids are arranged in a definite order in a protein.

When a cell wishes to reproduce a certain protein, the section (or "gene") of DNA that is responsible for that protein must undergo a well-defined process. First, the two bases that are interlocked in each rung of the ladder in that section must be "unzipped" so as to expose a sequence of bases. The exposed sequence then creates a strip of RNA whose task is to assemble amino acids from the cell medium in the correct order.

The order of the bases along the DNA "ladder" (or along the RNA strip) is highly specific, just as the order of acids in the protein is crucial for protein function. The change of even a single base inside a gene may result in the creation of the wrong protein, and the organism may die as a result. This indicates the need for serious error-protection in the process of replication of a cell.

6. Cell structure: high information content

Even a "simple" cell is a complicated system where chemicals of various kinds operate in a synergistic way to provide various functions that are essential to cell viability.

The outer wall (or membrane) provides the cell with its own identity, and separates it from the rest of the world. Apart from the membrane, i.e. inside the body of the cell itself, there are a number of sub-systems that must run cooperatively in order to keep the cell in operation. The most important chemicals are proteins and the DNA that has the capacity to reproduce those proteins.

Some proteins provide the structural characteristics of the different components of the cell. Some proteins serve as catalysts in the various chemical reactions that keep the cell running. There are also regulatory proteins which ensure that each protein performs its function only in its proper location within the cell: it would not do, e.g., to have energy generation occurring in the cell membrane. In a multi-cell organism, these regulatory proteins ensure that (e.g.) kidney cells do not grow in (say) the eye.

It is amazing that there is enough information in a linear object (a DNA strand) to determine a three-dimensional object (a protein). How is it that the sequence of bases in DNA instructs the cell to make proteins, each of which is a "sentence" composed of a specific sequence of various choices from a "vocabulary" of the 20 (or so) amino acids which occur in modern proteins? (There are many more amino acids in nature, but they are non-proteinous, and we do not consider them here.) The beginnings of an answer were first proposed by Gamow (1954: Nature 173, 318): there exists a code which translates the information in the bases in DNA into the amino acids in protein. This was an amazing insight on Gamow's part. As Yockey says (p. 4): "The idea…of a code is so unconventional that had Gamow's paper been submitted by almost anyone else, it would most certainly have been rejected".

The eventual identification of the code at the heart of biology is a triumph of human ingenuity. The bases in DNA are now known to be grouped into 64 "code words", and the sequence of these words contain the information which is eventually translated into the 20-letter vocabulary of proteinous amino acids.

A more difficult question to answer is: how do the amino acids "understand" the "language" of the "words of information" that are contained in the DNA? (For example, a string of letters may mean one thing to a Frenchman, something else to a German, and nothing at all to an Englishman.) It is not obvious that an answer has yet been given to this question. It may in fact be the most difficult question of all to answer. For example, Yockey (2000: Computers and Chemistry 24, 105) argues that the answer may simply be beyond the powers of human reasoning. In the present calculation, we do not address the issue of the origin of the code. We merely assume

that the code is already in existence as a result of unspecified processes in the early Earth.

Returning to a question about the links between DNA and protein that can be answered, the distinction between 64 and 20 is noteworthy and essential for living cells. In terms of coding theory, the fact that 64 greatly exceeds 20 means that DNA code has a lot of built-in redundancy: there are more code words (or symbols) at the source (DNA) than at the destination (protein). Coding theory proves that this redundancy of source relative to destination is an essential feature of a code in order to protect from errors in transmission. One of the theorems of coding theory (Shannon's channel capacity theorem) makes a strong statement which at first sight appears counterintuitive (Yockey, p. 8): even if there is noise in a message, the proper use of redundancy allows one to extract the original message "with as small a probability of error as we please".

Therefore, if we were to attempt to construct a biological system based on a code where redundancy is absent (and we shall mention one such attempt in Section 19 below), the process of cell replication would inevitably be prone to errors in transmission. Since even a single error may prove to have mortal consequences for a protein (and its host organism), it is hard to see how cells that are subject to serious errors in replication could be regarded as "living" in any meaningful sense.

The code words in DNA in the modern world consist of a series of triplets of bases. Each triplet (written as ACG, or UGA, etc, where each of the letters A, C, G, and U is the initial letter of one of the 4 bases) encodes for a particular amino acid. There are 64 such triplets available as a source code. (We will consider below the possibility that triplet codons were not necessary in the primeval soup, but that doublet codons might have sufficed then.)

If a cell contains a particular protein that is a chain of N_a amino acids in a certain sequence, then the DNA of that cell contains a corresponding segment containing $3N_a$ bases also arranged in a sequence that exactly parallels the N_a acids in the protein.

However, this is not all that is required for a gene. Since the DNA consists of a long chain of bases, we need to ask: how does the RNA know where to start "reading" the code for a particular protein? The answer is: in the DNA itself, associated with each gene, there must be a "start code" and a "stop code". In fact, a triplet of bases serves to encode START and another triplet to

encode STOP. (E.g., in modern cells, the triplet AUG encodes for start, while stop has three possible codons: UAA, UAG, UGA.). Therefore, although a strip of RNA needs to have $3N_a$ bases in a particular order, the gene (i.e. the corresponding piece of the DNA) must have $3N_a+6$ bases in a particular order.

As an example, we note that among the shortest proteins that exist in human beings, insulin contains 51 amino acids in a particular order. Such a protein requires a sequence of 153 bases in human DNA in a specific order, plus 6 bases for start and stop.

7. What does a cell need in order to function?

To determine the probability that the first cell was assembled randomly, we first need to answer the following general question: what is required in order to make a functional living cell?

In other words, what is the bare minimum number of proteins for a cell to function at all? If we can answer this, it should help us determine what the very first cell might have looked like.

As a first step in answering this, it is worthwhile to consider the simplest known cell that exists in the world today. This is an organism called "Mycoplasma genitalium" (MG) whose genetic information is many times smaller than the information in the human genome: the number of genes required for the functioning of MG in its natural state is only 517. (Humans have tens of thousands of genes.)

Recently, researchers have raised the interesting issue: are all 517 of these genes really necessary for MG to function properly? The answer is No. By removing genes one at a time, researchers have been able to show that the cell continues to function with fewer than the total complement of 517. By eliminating more and more of the genes, it has emerged that MG continues to function normally as long as there are between 265 and 350 protein-coding genes (see Hutchison et al., Science vol. 286, p. 2165, 1999). An earlier estimate of the minimum cell size in nature had suggested that the minimum number of proteins for cell operation might indeed be about 250 (J. Maniloff, Proc. Natl. Acad. Sci. USA Vol. 93, p, 10004, 1996).

It appears, then, that the simplest cell in the modern world requires at least 250 proteins in order to survive in viable form. Many of the 250 (or so) essential proteins in MG have identifiable

functions. Hutchison et al. list 13 categories of identified functions in the MG genome: (1) cell envelope, (2) cellular processes, (3) central intermediary metabolism, (4) co-factors and carriers, (5) DNA metabolism, (6) energy metabolism, (7) fatty acid metabolism, (8) nucleotides, (9) protein fate, (10) protein synthesis, (11) regulatory functions, (12) transport/binding proteins, and (13) transcription. Each of these 13 categories contains multiple genes, so that (e.g.) protein synthesis does not depend solely on a single protein for its operation: there are backups and multiple redundancies in each category. For example, some 19 proteins are used for membrane maintenance (category (1)). About 150 of the MG proteins can be assigned with some confidence to one of the 13 categories.

However, more than 100 of the MG genes perform functions that are currently unidentified. Nevertheless, the cell certainly requires them: without them, there is empirical proof that the cell fails to function.

8. The first cells to appear on Earth: reducing the requirements to an absolute minimum

It might be argued that the first cells to appear on Earth were smaller than the simplest cells (such as MG) that exist in the world today. Those primitive cells might have been able to operate with many fewer proteins than the 265 needed by MG.

Although we will use this argument below, it is actually difficult to substantiate. The mathematician John Von Neumann estimated the bare necessities which are necessary in order to construct what he referred to as "a self-replicating machine" (Theory of Self-Reproducing Automata: Univ. of Illinois press, 1966). It has been a popular exercise among science fiction writers to use this idea in connection with how a civilization might colonize a galaxy by sending out machines. Von Neumann concluded that the number of parts in one such machine must be in the millions. Other authors have reduced this estimate somewhat, but even according to the most optimistic estimate, the numbers remain very large: the best estimates suggest that there must be between 10^5 and 10^6 parts in a self-replicating machine. This means that the genome needs at least 10^5 bits in order to metabolize and replicate (Yockey, p. 334). Using the information content in a typical modern protein, Yockey concludes that the original genome must have been able to specify at least 267 proteins. The fact that this is close to the minimum number required for a modern cell (such as MG) suggests that one is not necessarily permitted to assume that the original

cell contained significantly fewer proteins than the smallest modern cell.

Nevertheless, other authors have argued that the Von Neumann approach is overly restrictive. E.g., Niesert (1987, origins of Life 17, 155)) estimates that the first cell might have been able to operate with as few as 300-400 amino acids.

Which of these various estimates of minimum requirements for the first cell should we consider? There must be some absolute minimum requirements for making even the simplest cell. For example, one might argue that, among the 12 non-regulatory categories of gene functions listed by Hutchison et al., one representative protein should be present in the first cell. And each of these 12 proteins should have an accompanying protein to serve in a regulatory role. This line of reasoning would suggest that 24 proteins are a minimum for cell operation.

Can we reduce this to an even barer minimum? Examples of minimum cell requirements have been summarized by the paleontologist George Gaylord Simpson. Of the 13 categories listed by Hutchison et al, Simpson narrows down the bare minimum to the following: (i) energy generation, (ii) storing information; (iii) replicating information; (iv) an enclosure to prevent dispersal of the interacting sub-structures; (v) digestion of food; (vi) waste product ejection (Science vol. 143, p. 771, 1964).

In view of these bare-bones requirements, it is hard to imagine how any cell could function without at least the following six types of proteins: (i) those that help to digest food, (ii) those that generate energy for cell operations, (iii) those that carry away waste products, (iv) those that preserve and repair the cell membrane, (v) those that determine when reproduction is to occur, and (vi) those which actually catalyze the tasks of reproduction. Corresponding to each of these six, there must be a regulatory protein which ensures that the corresponding protein does not "express itself" in the wrong location in the cell.

It is hard to imagine how a living cell would exist at all if it failed to contain at least these 12 proteins.

The fact that the simplest cell in the modern world (MG) requires 265 proteins as a bare minimum in order to function makes our estimate of 12 proteins look ridiculously small. But since it is possible that the first living cells may have been much simpler than those we find in the world today, let us make the (perhaps

absurdly reductionist) assumption that the first cells in fact were able to operate on the basis of the bare minimum 12 proteins.

As an illustration of how reductionist our assumption is, we note that in the first cell, we are assuming that a single protein is responsible for ensuring proper functioning of the lipid membrane of that cell. In contrast, the smallest known cell in the modern world (MG) uses 19 genes to encode for lipoproteins (Hutchison et al. Science vol. 286, p. 2166). The use of 19 genes in the modern cell is an example of the large amount of redundancy that nature uses to ensure that the membrane survives. But the first cell may not have had the luxury of redundancy: it may have been forced to survive using only one gene for its membrane. It would have been a precarious existence.

We have argued that each protein must contain at least 14 amino acids: thus our bare minimum cell, with 12 proteins and 14 amino acids in each, contains 168 amino acids. This is even smaller than the bare minimum of 300-400 amino acids described by Niesert (1987, Origins of Life, 17, 155). The DNA of our minimal (12-14) cell would contain only about 500 bases. This is 10 times shorter than the genome of a certain virus (PHI-X 174) which transmits 9 proteins. It is widely believed that a virus cannot be regarded as a "living cell" (it has no self-contained replication system), so this again indicates the extreme nature of our assumption that the first cell could have as few as 12 proteins. But let us proceed in the spirit of optimizing the probability that the first cell appeared by chance.

8.1. The first cell: putting the proteins together by chance

In the early Earth, the commonest concept of conditions back then is that the primeval "soup" consisted of various chemicals that were stirred up and forced into contact with one another as a result of the forces of nature (including rain, ocean currents, lightning). Simple chemical reactions in the soup were easily able to create amino acids: these molecules are so small (containing no more than 10-30 atoms each) that random processes can put them together quickly from the abundant C, O, N, and H atoms in the soup. As a result, we expect to find in the primeval soup, in abundant supply, all of the 22 amino acids that occur in modern life forms. (For the number 22, see Nature vol. 417, 478, 2002). In fact, there are more than 100 amino acids in modern nature, but only 22 are used in proteins. And of those 22, numbers 21 and 22 are rare. Most living material relies on only 20 of these amino acids, and we will use that number here.

To be sure, the “primeval soup” hypothesis is not without its opponents (e.g. Yockey, pp. 235-241). Laboratory experiments which claim to replicate conditions in the primeval Earth generate not only amino acids but also a tarry substance (as the principal product). This substance should have survived as a non-biological kerogen in ancient sedimentary rocks, but no evidence for this has been found. It should not be surprising that, in the primeval soup, other amino acids, not currently used in life forms, could have been formed. (This would include the acids that are used in nylon.) And each of the amino acids which are created randomly in the primeval soup would be created in two forms: the D-variety and the L-variety. (These varieties refer to the ability of the molecule to rotate the polarization of light either right or left: this ability depends on the chirality of the molecule, i.e. on the handedness of its 3-dimensional structure.) For reasons that are not yet obvious, only one of these varieties (the L-variety) is actually used in present-day life forms. However, the basic property of amino acids, that they polymerize, operates only between L alone or D alone: when an L and a D amino acid combine, their opposite chirality has the effect of locking out any possibility of further polymerization.

Another difficulty of a very different nature has to do with reactions in an aqueous solution. The very process of assembling amino acids into a polypeptide chain (so as to make a protein) requires the removal of H from the amino radical and the removal of OH from the acid radical: it is not obvious how these constituents of a water molecule can be removed in an aqueous solution.

Despite these difficulties with the primeval soup hypothesis, the idea of the soup is so widespread in textbooks that it is a natural starting point for an optimized estimate of probabilities. In the spirit of the present approach (where we do whatever we can to optimize the chances of assembling the first cell randomly), we will simply go along with the textbooks. We shall assume that the formation of the first cell in the early Earth began in liquid water where only 20 L-amino acids need to be taken into account.

Other simple chemical reactions in the soup also give rise more or less quickly to the four bases (two purines and two pyrimidines) that form the "rungs" of DNA. Why are these formed relatively readily? Because each base consists on no more than 13-16 atoms, random processes can also assemble these bases rapidly from the abundant C, O, N, and H atoms. It was probably more difficult to form pyrimidines than purines, but the principle is robust:

formation of small molecules is essentially inevitable in the early Earth.

In order for the first cell to come into existence, at least 12 proteins, each with N_a amino acids in a specific order, had to emerge in the same patch of the "primeval soup". To be sure, individual proteins were probably emerging at random at many places around the world. But if our aim is to form a complete living cell, it will not help if the membrane protein emerged (at random) in China, the energy protein in Russia, and the replication protein in South America. That will not do: the only way to have the first cell develop is if all 12 proteins emerge in close enough proximity to one another to be contained within a single membrane.

How might this have happened in random processes? By way of example, let us consider one particular protein, in which the chain of amino acids happens to be denoted by the series of letters ABCDEFGHIJKLMN. In order that this protein be made by chance, amino acid E (say) (one of the 20 commonest in nature) might have started off by entering into a chemical reaction with amino acid F (another of the 20), such that the two found it possible to become connected by a peptide bond. Then amino acid D might have had a chemical reaction so as to join onto the EF pair at the left end, forming DEF by means of a new peptide bond. Note that it is important to form DEF rather than EFD, which would be a very different protein. This process presumably continued until the entire 14-unit protein chain ABCDEFGHIJKLMN was complete.

8.2. The first cell: putting the DNA/RNA together by chance

It is not enough to assemble 12 proteins to have a functional living cell: the cell must be able to reproduce, and for that the cell needs DNA (or at least RNA). In order to ensure reproduction of the cell, there had to be (also in the same patch of the primeval soup) at least 12 genes on an RNA strand, each containing $3N_a+6$ bases in a specific order.

Thus, in the very same patch of "soup" where the protein ABCDEFGHIJ formed by chance, a strand of RNA must have been formed where the three bases that encode for amino acid A were joined in a specific order along the RNA strip by a series of chemical reactions. Then the three bases that encode for amino acid B had to be added in a specific order to the sidepieces, right next to the three bases that encode for A. This process must have continued until the triplets of bases that encode for each of C, D, E, F, G, H, I, J, K, L, M, and N respectively were assembled in

a specific order into a chain of 30 bases. There would also be one triplet at each end of the 30-base sequence to serve as markers for start and stop. This 36-base sequence would then form the gene for the first protein in the first cell.

Now that we know how the first proteins and RNA/DNA were put together, we are in a position to estimate the probability that this will occur by random processes.

9. Probability of protein formation at random

In the example given above, we recall that amino acid (say) E is only one of 20 amino acids that exist in living matter. Amino acid F is also one of 20. Therefore, a process that successfully forms the sequence EF at random out of a soup where all amino acids are present in equal abundances, has a probability p_2 which is roughly equal to (1/20) times (1/20) = 1/400.

Actually, however, pre-living matter contains not only the L-variety of each amino acid, but also the D-variety. Therefore, a better estimate of the probability p_2 that the correct pair of L-amino acids be formed is (1/40) times (1/40), i.e. $p_2 = 1/1600$. However, once an L-acid unites with a D-acid, the opposite nature of their chiralities leads to a "lock-out": no further polymerization is possible. So we will optimize probability by assuming that only the L-variety is present. We therefore take $p_2 = 1/400$.

Another way to state this result is that if we wish to create the combination EF (both L-variety) by chance, the number of chemical reactions that must first occur between amino acids in the primeval soup is about $1/p_2$, or about 400. That is, if we allow so much time to elapse that 400 reactions can occur in the primeval soup, then there is a high probability (close to a certainty) that the combination EF will appear simply at random.

This argument assumes that the only amino acids in the primeval soup are the 20 which occur in modern living organism. However, there were certainly other non-biological amino acids available. As a result, many more than 400 reactions was almost certainly required before the combination EF appeared at random. However, we will optimize the chances for random assembly of the first cell by ignoring the non-biological amino acids.

After creating EF by random processes, the next step is to have the next amino acid to join the chain be the L-variety of (say) G, i.e. only 1 out of the 20 types available. Then the probability

that the three amino acids EFG will be assembled in the correct order is about $(1/20)^3$.

Continuing this all the way through a sequence of N_a amino acids in a protein, the chance f_1 of correctly picking (at random) all the necessary amino acids to create one particular protein is roughly equal to (1/20) raised to the power N_a. This corresponds to $f_1 = (1/10)^x$ where $x = 1.3N_a$. Actually, to the extent that some amino acids may be replaced by others without affecting the functionality of the protein, the above expression for f_1 is a lower limit. (We will allow for this later in this section.) Yockey (p. 73) shows that instead of 20^N for the value of $1/f_1$, a more accurate estimate is 2^{NH} where H is the mean value of a quantity known as the Shannon entropy of the 20-acid set (see below). In the limit where all amino acids have equal probability of being encoded, and are equally probable at all sites in the protein, 2^{NH} turns out (from the definition of H) to be equal to 20^N . In all other cases, 2^{NH} is less than 20^N. This returns us to the previous conclusion: the above expression for f_1 is a lower limit on the true value.

Suppose that the particular protein with probability f_1 has been formed in a particular patch of the primeval soup. Then in order to form a single cell (with at least 12 proteins as a bare minimum to function), eleven more proteins must also be formed in the same patch of soup, in close enough proximity to one another to be contained within a single membrane. Each of these proteins also has a certain number of amino acids: for simplicity let us assume that all have length N_a.

The overall probability f_{12} that all twelve proteins arise as a result of random processes is the product of the probability for the twelve separate proteins. That is, f_{12} is roughly equal to f_1^{12}, i.e. f_{12} is roughly $(1/10)^y$ where $y = 15.6N_a$.

We can now quantify the claim that the first cell was assembled by random processes. If the first cell consisted of only the bare minimum 12 proteins, and if each of these proteins was uniquely suited to its own task, the probability that these particular 12 proteins will be formed by random processes in a given patch of primeval soup is f_{12}.

Now let us turn to the fact that a protein may remain functional even if a certain amino acid is replaced with another one. (Obviously, we are not referring to invariant sites here.) For example, it may be that the protein which we have specified as the one that is responsible for (say) energy generation in the cell is

not unique. There may exist other groupings of amino acids which also have the shape and properties that enable the task of energy production for the cell. Maybe the others are not as efficient as the first one, but let us suppose that they have enough efficiency to be considered as possible candidates for energy production in the first cell. Then we need to ask: how many energy-producing proteins might there be in the primeval soup?

It is difficult to tell: in principle, if N_a has the value 14 (say), then one could examine the molecular structure of all 14-amino acid proteins (of which there are some 20^{14} , i.e. $10^{18.2}$ if all amino acids are equally probable) and identify which ones would be suitable for performing the energy task. Presumably there must be *some* specificity to the task of energy production: otherwise, a protein which is supposed to perform the task of (say) waste removal might suddenly start to perform the task of (say) membrane production in the wrong part of the cell. Therefore, it is essential for stable life-forms that not all available proteins can perform all of the individual tasks.

Suppose the number of alternate energy-producers Q is written as 10^q. In a world where all proteins have $N_a = 14$, the absolute maximum value that q can have is $q_{max} = 18.2$. This is the total number of discrete locations in the "14-amino acid phase space".

In the real world, a more realistic estimate of q_{max} would be smaller than the above estimate. First, not all amino acids have equal probability of being encoded: there are more codons in the modern genetic code for some amino acids than for others. (E.g., Leu, Val, and Ser have 6 codons each, whereas 10 others have only 2 codons each.) When these are allowed for in the probability distribution, it is found that the "effective number" of amino acids in the modern world is not 20 but 17.621 (Yockey, p. 258). Thus, with $N_a = 14$, a more accurate estimate of $q_{max}(eff)$ is 17.4 (rather than 18.2).

As a result, in the real world, $q_{max}(eff)$ may be considerably smaller than 18.2. However, in the spirit of optimizing probabilities, let us continue to use the value 18.2.

The requirement that some specificity of task persists among proteins requires that the value of q must certainly not exceed q_{max}. At the other extreme, in a situation where each protein is uniquely specified, q would have the value $q_{min} = 0$ (so that one and only one protein could perform the task of energy production).

Now we can see that our estimate of f_{12} needs to be altered. We were too pessimistic in estimating f_{12} above. Each factor f_1 needs to be multiplied by 10^q. For simplicity, let us assume that q has the same value for each of the 12 proteins in the cell. Then the revised value of f_{12} is $1/10^z$ where

$$z = 15.6N_a - 12q \ . \qquad \text{(eq. 1)}$$

This result applies to a cell with 12 proteins, each composed of amino acids chosen from a set of 20 distinct entries.

10. Random formation of DNA/RNA

The first cell could NOT have functioned if it consisted only of proteins. In order to merit the description *living*, the cell must also have had the ability to reproduce. That is, it must also have had the correct DNA to allow all 12 proteins to be reproduced by the cell.

In order to estimate the probability of assembling a piece of DNA by random processes, we can follow the same argument as for proteins, except that now we must pick from the available set of 4 bases.

Repeating the arguments given above, we see that for each protein which contains N_a amino acids in a certain sequence (plus one start and one stop), there must exist in the DNA a strip of $B = 3N_a+6$ bases in a corresponding sequence. If we pick bases at random from a set of 4 possibilities, the probability of selecting the correct sequence for a particular protein is $(1/4)^B$. Therefore, the probability of selecting the correct sequences for all twelve proteins, if each protein is unique, is $(1/4)^D$ where $D = 36N_a + 72$. Writing this with the symbol f_{RNA}, we see that f_{RNA} is equal to $(1/10)^E$ where

$$E = 21.7N_a + 43.3. \qquad \text{(eq. 2)}$$

Again, however, if instead of unique proteins for each task, there are 10^q proteins available to perform each task in the cell, then we must increase the above value of f_{RNA} to 10^{-G} where

$$G = 21.7N_a + 43.3 - 12q. \qquad \text{(eq. 3)}$$

11. Probability of random formation of a complete cell

Since both the RNA and all 12 proteins have to be formed in the same patch of primeval soup in order to form a viable cell, the probability f_{cell} that random processes will perform both tasks in the same patch of soup will be the combination of the separate probabilities. That is, f_{cell} is roughly equal to f_p X f_{RNA}, i.e. about 10^{-J} where

$$J = 37.3N_a + 43.3 - 24q. \qquad \text{(eq. 4)}$$

Therefore, once enough time elapsed in the primeval soup to allow the chemicals there to undergo a certain number of reactions, $R_{12p} = 1/f_{cell}$, there would be a high probability (in fact, a near certainty) that the proteins and the requisite DNA for a (12-14) cell could indeed have been assembled by chance in the primeval soup.

In order to optimize the chances of forming the first cell, we ask: is it possible to find ways to make R_{12p} smaller than the above estimate? The answer depends on the theory that one adopts for the development of the first cell.

Suppose one were to theorize that the only thing one would have to provide to get the first cell going was the RNA containing the genetic code for the 12 proteins. (It might be beneficial if the RNA could catalyze its own replication: however, this is not altogether desirable, since it leads to possibilities of ``error-catastrophes" [Niesert et al. 1987, J. Mol. Evol., 17, 348].) According to the "RNA-first theory", one would not have to "wait around" for proteins to be constructed by random reactions in the primeval soup. Instead, once strips of RNA were formed (as a result of random processes), DNA could be assembled from the RNA strips. At that point, proteins should be reproduced more or less automatically, apart from the necessity of certain enzymes (proteins) to catalyze the "unzipping" of the DNA itself, and to catalyze the collection and assemblage of the amino acids.

In order to optimize the chances of cell formation at random, let us assume that the unzipping can be done with the help of a single protein, and that the collection and assemblage of amino acids can also be done with a single protein. (This is a far cry from the modern world, where multiple proteins exist in even the simplest cell to perform each task.) Then the first cell will require the RNA to be assembled by chance (with probability f_{RNA}, as given above) plus just two proteins (with probability f_2) also assembled

by chance. If this theory is correct, then R_{12p}(RNA-first) would be equal to 10^K where

$$K = 24.3N_a + 43.3 - 14q. \qquad \text{(eq. 5)}$$

This may provide a substantial reduction below the original estimate of R_{12p}.

Should we also consider the obvious alternative to the RNA-first theory? That is, should we also consider the "protein-first" theory? The answer is no, provided that the modern genetic code is at work. The structure of the modern genetic code is such that, according to the Central Dogma, proteins do not pass on information to DNA: the flow of information goes only from DNA (or RNA) to protein, and not the reverse. As Yockey (2000) puts it, "The origin of life [as we currently know it] cannot be based on 'protein-first'." However, the "protein-first" theory may need to be considered when we consider a certain "window of opportunity" in the early Earth (see Section 19).

Because we now know how many reactions are required in order to create the first simplest possible cell, we are in a position to test the evolutionary claim that the first cell was assembled randomly. To do this, we proceed to the crucial question that is at the heart of the present argument. This question, and its detailed answer, is the subject of the next section.

12. How many reactions occurred in the primeval soup?

Is random assembly of the first cell possible? To address this, we need to answer the following question: How many chemical reactions (of the sort we are interested in) actually occurred in the primeval soup during the first 1.11 billion years?

We will not be surprised to find that the number of reactions n_r is a "large" number (in some sense). Nevertheless, n_r is a finite number.

Once we obtain n_r, we can then estimate how large the value of q must be in order that the probability of randomly assembling the first cell of order unity. That is, we will equate n_r to 10^J (or to 10^K, if we accept the "RNA-first theory"), and solve for q, assuming that N_a is at least as large as 14. The value of q which we obtain from this estimate will be labelled q_{ra} to denote that this is how large q must be in order that random assembly of the first cell in the primeval soup becomes essentially certain.

We are interested in chemical reactions involving amino acids or bases. To proceed with this discussion, we need to consider in detail what happens during such a reaction. The most basic requirement of a chemical reaction is the following: the two reacting molecules must at the very least come close enough to each other to have a collision. However, the very fact that two molecules collide does not guarantee that a reaction will occur. The reaction is controlled by many factors, e.g. the energy involved, the angle of the encounter, the removal of by-products, etc. As a result of these factors, many collisions may occur before even a single reaction occurs. This explains why it is so difficult to manufacture (e.g.) nylon: the creation of the peptide bonds that hold nylon together (exactly equivalent to those which hold proteins together) requires careful quality control. The quality control which the DuPont engineers are forced to impose in order to create nylon was certainly not available in the primeval soup: therefore, the efficiency of the reactions which led to peptide bonds (i.e. proteins) in the primeval soup was almost certainly very small.

In view of this, we can derive an absolutely firm (and probably very generous) upper limit on the number of two-body reactions n_2 that occurred between two amino acids during any time interval by calculating the number of collisions n_{coll} that occurred between those two amino acids during that interval. In practice, n_2 is probably orders of magnitude smaller than n_{coll}. The purpose of a catalyst is of course to increase n_2 as much as possible: however, even with a "perfect" catalyst, n_2 can never exceed n_{coll} .

So let us turn to estimating n_{coll}. This number, which is "large" but finite, will provide us with a firm piece of quantitative evidence that will allow us to test the assertion that the first cell was assembled randomly.

13. Collisions between amino acids in the primeval soup

We begin the calculation of n_{coll} by estimating the mean time t_c that elapses between successive collisions of molecule A with molecule B. The general formula for t_c is straight-forward. Let us consider molecule A as the projectile, and molecule B as the target. If projectile A moves with mean speed v cm/sec through an ambient medium where there are n_t target molecules per cubic

centimeter, then t_c equals $1/(v\ n_t\ A)$ seconds. Here, A is the area (in square centimeters) presented by the target molecule.

13.1 Mean time interval between collisions

Let us now estimate the three quantities that enter into t_c. First, the area A. Amino acids and bases in nature have linear dimensions of a few Angstroms (where 1 Angstrom = 10^{-8} cm). Therefore, a typical amino acid or base molecule has A equal to about 10^{-15} sq. cm.

Second, as regards v, there is a standard formula for the mean speed of the molecules in a medium at temperature T: $v^2 = R_g T/m$ where R_g is the gas constant (= 8.3 X 10^7 ergs/degree/gram) and m is the molecular weight. Amino acids and bases have m = 100 or so. Moreover, living cells require liquid water in order to survive: this means that T must be in the range 273-373 degrees Kelvin. Taking an average value for T of about 300 K, we find that v for the molecules in which we are interested here is about 10^4 cm/sec. Even if we consider the extremely hot conditions at the ocean bottom, near the hot thermal vents, where temperatures may be as large as 1000 K, this will increase our estimate of v by a factor of no more than 2. This will have no significant effect on our conclusions below.

Third, as regards n_t, we note that at the present time, the total mass of living organisms on Earth is M_{living} = 3.6 X 10^{17} grams (see http://www.ursa.fi/mpi/earth/index.html). In the early Earth, before the first cell appeared, the mass of living material was by definition zero. But there were amino acids and bases present in the primeval soup. So in order to optimize the chances of cell formation, let us make a second gross assumption: let us assume that *all* of the mass that is now in living organisms was already present in the primeval soup in the form of amino acids (if we wish to assemble proteins) or bases (if we wish to assemble RNA).

With a molecular weight of about 100, each amino acid (or base) has a mass m_{aa} of about 1.7 10^{-22} grams. Therefore, the total number n_{total} of amino acids (or bases) in the primeval soup was of order M_{living}/m_{aa}. With this assumption, we find n_{total} = 2 X 10^{39}. Naturally, this estimate is quite uncertain. Other estimates of this number are larger. E.g. Bar-Nun and Shaviv (Icarus 24, 197, 1975) estimate 5.4 X 10^{41}, while Shklovskii and Sagan (1966 Intelligent Life in the Universe) estimated 10^{44}. We shall see that our results are only slightly affected by these uncertainties.

Finally, to derive n_t in the primeval soup, we need to divide n_{total} by the volume of the material where living material existed on the early Earth. In the present Earth, the volume of the biosphere is of order 10^{19-20} cubic cm. However, life probably started in particular locations, and so the relevant volume of the primeval soup was probably much smaller. Let us suppose that the early Earth had a biosphere with a volume that was 10-100 times smaller than it is at present. (This putative decrease in volume will help to speed up reactions.) That is, let us suppose that all of the amino acids which now are present in living matter on Earth were concentrated in the primeval soup into a favored volume of only 10^{18} cubic cm. Combining this with our estimate of n_{total}, we see that the mean density of amino acids in the favored volume of the primeval soup n_t could have been about 2 X 10^{21} per cubic cm.

Is this a reasonable value? To answer this, we note that this value of n_t corresponds to a mean mass density of 0.34 gram/cubic cm for the amino acids in the primeval soup. This density is very high (the molar concentration is about 0.1): it is questionable whether such a high density of amino acids could ever have been dissolved in water. This estimate of mass density is certainly close to the upper limit possible: it could hardly have been any higher. In order to remain dissolved in water (with mean density 1 gram/cubic cm), the mass density of amino acids can certainly not exceed the density of water. Therefore, our estimate of the upper limit on n_t is not unreasonable as we try to optimize the chances of randomly assembling a cell. (If we were to use Bar-Nun and Shaviv's estimate of the total number of amino acids, we would need to dilute them by dissolving them in at least 100 times more volume than we used above in order to keep the mean density less than that of water. With Shklovskii and Sagan's estimate, the volume must be larger still by a further factor of 200.) The actual value of n_t in the primeval soup was probably orders of magnitude less than the estimate given above. Maximum molar concentrations of amino acids in the primeval soup have been estimated to be as low as 10^{-7} or 10^{-8} (Hulett 1969 J. Theor. Biol. 24 56; Dose, 1975, Biosystems 6, 224). Thus, our estimates of n_t are probably too large by 6 or 7 orders of magnitude. However, in the spirit of optimizing the chances of making a cell, let us use the above upper limit as the value of n_t.

Now we have all of the ingredients we need to calculate t_c, the mean time between collisions in the primeval soup. We find t_c = 5 X 10^{-11} seconds.

13.2. Number of collisions by a single amino acid in 1.11 b.y.

Now that we know the mean interval between collisions, we see that in the primeval soup, a given amino acid experienced 2 X 10^{10} collisions every second as an upper limit. Therefore, each amino acid experienced no more than 2 X 10^{10} reactions every second with other amino acids.

How many collisions did an amino acid experience in the primeval soup in the course of a time interval of 1.11 billion years, i.e. in the 3.5 X 10^{16} seconds before the first cell appeared on Earth? The answer is straightforward. Multiplying the above reaction rate by the number of seconds available, we find that each amino acid in the primeval soup experienced at most $n_r(1)$ = 7 X 10^{26} reactions with other amino acids before the first cell appeared on Earth.

13.3. Total number of collisions between amino acids in 1.11 b.y.

Finally, we ask: what was the total number of reactions between amino acids that occurred in the primeval soup before the first cell appeared? The answer is again straightforward: since each amino acid experienced $n_r(1)$ in that time, and since there were n_{total} amino acids in the primeval soup, the total number of reactions n_r between amino acids was about 10^{65} before the first cell appeared.

This is a "large" number. But it is finite.

Moreover, we have artificially forced n_r to be as large as possible by making four extreme assumptions. (i) Every collision produces a peptide-bonding reaction. (ii) The mass of pre-biotic material was as large in the primeval soup as it is in today's biomass. (iii) The entire biomass in the primeval soup was in the form of amino acids (or bases). (iv) All amino acids were concentrated in pools where their mass density could build up to the maximum permissible value. In the real primeval soup, conditions might have been such that any or all of these assumptions could have failed by several orders of magnitude. (In particular, (iv) almost certainly failed by 6-7 orders of magnitude, and (i) almost certainly failed by several orders of magnitude because of reaction kinetics.) Therefore, it is highly likely that the actual total number of collisions which occurred in the primeval soup before the first cell appeared could have been 10 or more orders of magnitude less than 10^{65}.

Of course, our estimates refer to our estimates of the biomass only, and also to binary collisions only. If we were to use the estimates of Bar-Nun and Shaviv or of Shlokskii and Sagan, the number densities per unit volume n_t cannot exceed the value we have

already used above. Therefore, there will be no change in the number of collisions per second. But the total number of collisions would increase by 2-5 orders of magnitude above our estimate.

For the sake of argument, let us assume that these other processes compensated for orders of magnitude deficits associated with the extreme assumptions (i)-(iv) above. That is, we will assume in what follows that n_r was indeed of order 10^{65}. This appears to be a very generous estimate of the total number of reactions in the primeval soup.

14. Random production of the first cell

We are now in a position to estimate probabilities for randomly assembling the first cell.

Let us return to our estimate of the number of reactions that were necessary to create the first cell by random processes. In order to create a cell containing 12 proteins with chains of $N = N_a$ amino acids each, we recall that R_{12p} was required to be 10^J (where J is given in eq. (4) above) if proteins and RNA were both assembled at random.

However, if we accept the "RNA-first theory", we recall that the number of reactions R_{12p}(RNA-first) was "only" 10^K (where K is given in eq. (5) above).

Now that we know how many reactions actually did occur in the primeval soup before the first cell appeared, we can equate n_r to the above values of R_{12p} in order to determine how large q_{ra} must have been in order to have reasonable probability of assembling the first cell at random.

Setting R_{12p} equal to n_r, we find that the value of q_{ra} required for random assembly of the first cell must satisfy the equation

$$37.3N_a + 43.3 - 24q_{ra} = 65 \qquad \text{(eq. 6)}$$

if proteins and RNA were assembled together. On the other hand, if we accept the RNA-first theory, then we find

$$24.3N_a + 43.3 - 14q(\text{RNA})_{ra} = 65. \qquad \text{(eq. 7)}$$

As mentioned above, the value of N_a is no less than 14. Inserting $N_a = 14$ in eq. (6) and (7) leads to $q_{ra} = 20.8$ or $q(RNA)_{ra} = 22.8$. The numerical value of q_{ra} increases linearly with the value of N_a, increasing by 1.7 for each unit increase in N_a. However, q_{ra} is not sensitive to the number of proteins in the cell. Moreover, q_{ra} is not sensitive to errors in our estimates of the number of collisions in the primeval soup: even if our estimated number of collisions is wrong by factors of (say) one million times too large or too small, our estimates of q_{ra} would change by only plus or minus 0.4.

The above estimates of q_{ra} emerge from the two basic points of our argument: (i) a finite time was available for chemical reactions to operate, and (ii) a cell cannot function as a truly living organism with less than the bare minimum of 12 proteins.

However, as we saw in Section 9 above, the total number of all available proteins in the $N_a = 14$ world is such that q has certainly a maximum value $q_{max} = 18.2$. (The actual maximum would be smaller than this for the reasons discussed in Section 9 above, but let us continue to optimize the case for random assembly and retain $q_{max} = 18.2$.) We see that the value of q_{ra} that is required to ensure random assembly of the first cell is larger than q_{max}.

However, it is formally impossible for q to have a value in excess of q_{max}: q_{ra} cannot exceed q_{max} even in optimal conditions. If q_{ra} is equal to, or larger than, q_{max} it implies that every available protein in the primeval soup must have been capable of performing the task of every other protein. This indicates a serious lack of specificity of tasks in the cell.

This conclusion does not depend sensitively on the choice of N_a. If functioning proteins actually require N_a to be as large as (say) 20 (such as the mini-proteins referred to by Maniloff), we would find $q(RNA)_{ra} = 33$. However, the total number of proteins in an $N_a = 20$ world would be of order 20^{20}, i.e. $q_{max} = 26$. The value of $q(RNA)_{ra}$ again exceeds q_{max}, and so the conclusion about non-specificity still applies.

15. Do proteins in the primeval soup have specific tasks?

The result that q_{ra} has a value in excess of q_{max} has significant implications. It implies that there are no distinguishing properties between proteins: each protein would have had the

ability to perform the task of all the other functional proteins in the first cell. If that were to be the case, then there would be no way to regulate the various distinct groups of cell operations: replication could occur in the membrane, or membrane generation could occur in the energy generation sites.

However, the nature of a cell requires that proteins have clearly defined and distinctly specific functions. That is, not all proteins must be capable of (say) membrane production: only a fraction F (<1) of the proteins must have this capability.

What is a likely value for F? At one extreme, the smallest value F can have is $F_{min} = 1/Q_{max}$. Writing $F = 10^{-f}$, this means that the maximum possible value of f is $f_{max} = q_{max}$. In this limit, protein specificity would be maximized: there would then be one and only one protein out of the Q_{max} distinct proteins which could perform any one of the basic tasks of the cell. In such a case, all 14 amino acids in each protein would be an invariant site, forbidding any substitutions.

This extreme specificity is not true of most modern proteins: typically, only a subset of sites are invariant. E.g., Yockey (Table 6.3) discusses a 110-acid protein in which only 14 sites are invariant. At the remaining 96 sites, a number of other amino acids (from 2 to 19) may be substituted without degrading significantly the functioning of the protein. The amino acids which are functionally acceptable at a site are those which do not impede the folding process or the biochemical requirements of the protein. Because of these possibilities for substitution, the probability of randomly "finding" a functional protein in "amino-acid phase space" may be much improved over what one might expect on the basis of the value of Q_{max} alone. Yockey (p. 254) describes in detail how to compute the probability factor 2^{HN} when one knows how many different amino acids can be substituted at each site. For the 110-acid protein discussed by Yockey, the improvement in probability is enormous (from 1 in 10^{137} to 1 in 10^{93}). It is not clear how much improvement will occur in a small protein, where there are only 14 amino acids. For the latter, the phase space is limited to $10^{18.2}$. The 3-dimensional folding of such a small protein might be quite sensitive to amino acid substitutions, more so than for a larger protein. If this is true, then the improvement factor might be quite small.

At the opposite extreme, F can certainly not exceed $F_{max} = 1/12$ if we are to preserve the distinction of 12 separate proteins for each of the cell's tasks. The limit F = 1/12 represents the

minimum possible protein specificity. This means that f cannot have a value less than 1.08 in a cell with N_p = 12 proteins.

In fact, it is probable that F is much smaller than 1/12. If F were as large as 1/12, the prognosis for cell survival would be slim: a single point mutation could convert (say) a membrane-producer in any particular cell into (say) a waste management protein. If this were to happen, the cell and its progeny could hardly expect to survive for long.

This suggests that, in order to ensure long life for the cell, the value of F should be much smaller than 1/12. How small might F be? Let us introduce a "protein specificity index" m such that $F=(1/12)^m$, i.e. f = 1.08m. With this definition, the minimum value that m can have is m_{min} = 1 (the minimum permissible specificity). Values of m in the range (say) m = 3-4 represent conditions where protein functions are only marginally specific. The maximum value that m can have is $m_{max} = q_{max}/1.08$: in the example given above where q_{max} = 18.2, m_{max} would have a value of about 16.9. In the limit m = m_{max}, every protein performs a unique task.

With this well-defined range of the m parameter, we may usefully refer to an "average specificity index" $m_{av} = (m_{min} + m_{max})/2$. With the values just cited, we find m_{av} is about 9. High specificities can be considered as those with m values in excess of m_{av}. Low specificities are those with m values less than m_{av}.

16. What are the chances of creating the first functioning cell randomly?

The fact that the factor F departs from unity has the effect that the Q factor which we used above in estimating the probability of random formation of the first cell must be replaced by the product FQ_{max}. The quantity q in our earlier estimates must be replaced by q_{max}-f where f cannot be less than 1.08.

In view of this, if we adopt the "RNA-first" theory, the necessary number of reactions for random assembly of the first cell is 10^L where

$$L = 24.3N_a + 43.3 - 14(q_{max} - f). \qquad \text{(eq. 8)}$$

Setting N_a = 14, the chance P_r of random assembly of the first cell in the first 1.11 billion years of Earth's existence (during which time there were at most 10^{65} reactions) is one in 10^b where

$$b = 14(f - q_{max} + q_{ra}). \qquad \text{(eq. 9)}$$

With f=1.08m, and $q_{ra} - q_{max} = 4.6$, the chance P_r is about one in $10^{15m+64.4}$. Since m cannot be less than 1, P_r is certainly less than one in 10^{79}. If m takes on its average value $m_{av} = 9$, P_r decreases to 1 in 10^{200}. Even if m takes on values that are much smaller than m_{av} (say 2-3), the probability P_r amounts to only one in 10^{94-109}.

Note that the exponent b increases rapidly as N_a increases: both q_{ra} and q_{max} are proportional to N_a. As a result, if we increase N_a to (say) 21, we would find that $q_{ra} - q_{max}$ would increase from 4.6 to 6.9. Then even with m = 1 (its lowest value), exponent b exceeds 100.

Even if we were to allow for a much older Earth, with an age of (say) 100 billion years, the number 65 in our formula for q_{ra} would increase only to 67. This would lead to a reduction of only 0.14 in q_{ra} in the "RNA-first scenario". This would increase the chance of random cell assembly, but even in the best possible case (m=1), P_r would still be no better than one part in 10^{77}.

The result $P_r < 10^{-79}$ applies to a cell consisting of only the absolute minimum set of $N_p = 12$ proteins. Such a cell is extremely small compared to the smallest known cell in the modern world (where $N_p = 250$). What if the minimum number of proteins in a functional cell is 30 or 50 or 100? In such cases, the requirement of specificity of protein function has the effect that the factor F must be smaller than $1/N_p$, i.e. the exponent f must exceed $\log(N_p)$. In terms of the protein specificity index m,

$$f = m \log(N_p), \qquad \text{(eq. 10)}$$

where m cannot be less than 1. In view of this, the probability of random assembly of the first cell is one in 10^b where

$$b = (N_p+2)[m\log(N_p) - q_{max} + q_{ra}\ ]. \qquad \text{(eq. 11)}$$

Therefore, if the first cell required (say) 30 proteins to become operational, the chance of assembling its RNA at random in the primeval soup after 10^{65} collisions is less than one in $10^{47m+147}$. The exponent in this result rapidly becomes large even if we allow for only marginal specificity. For example, if m has a value of 2, P_r is less than one in 10^{240}. And if m is set equal to its average value $m_{av} = 9$, P_r falls to less than one in 10^{570}.

If the modern genetic code was operative in the first primitive cell (much smaller than the smallest cell in today's world), the above numbers are mathematical statements of the chances that the

RNA for the first cell was assembled by random processes. It is clear that the probabilities are extremely small. We stress that we have optimized a number of parameters in estimating the above probabilities.

17. What about doublet-codons?

We can improve the situation for random assembly of the first cell by considering the following possibility: suppose that, by some means, the proteins in the first cell were assembled from a smaller set of distinct amino acids than the $N_{aa} = 20$ which exist in nature today.

To be specific, let us suppose that the number of distinct amino acids which were used in the first cell was as small as $N_{aa} = 5$. It is not obvious that functional proteins could actually exist with such a small "vocabulary" of amino acids. However, it has been claimed that protein folding is possible with as few as 5 distinct amino acids (Riddle et al. 1997). Therefore, consideration of this case probably does not violate any of the constraints of physical chemistry. It also does not violate any of the limitations of information theory: the quaternary genetic code might have begun as a "first extension" using doublet codons (Yockey, p. 188). (Vestiges of this early code might still exist in modern mitochondria.) Doublet codons might have encoded for as few as 4-5 proteins (see Yockey, Table 7.1).

The major change in our calculation in this case is that the codons in the RNA would no longer need to consist of triplets of bases. Assuming that there are still 4 bases to use for RNA coding, doublets would suffice to provide unique encoding for all 5 amino acids (plus a start and a stop code). Of course, one might suspect that in a world where the number of useful amino acids has been reduced from 20 to 5, there might also be a reduction in the number of useful bases. For example, if there were only 2 useful bases (i.e. if the genetic code were ever binary consisting of one purine and one pyrimidine, a possibility discussed by Yockey (p. 184), then triplet codons would still be needed even to encode for $N_{aa} = 5$. In this case, we would return to the estimates derived above for the triplet codon world. If there were 3 useful bases available, doublet codons would suffice to encode for up to $N_{aa} = 7$ (plus a start and stop code).

However, to optimize chances for random assembly, let us assume that all 4 of the modern bases are available so that we can exploit the possibility of doublet codons for the case $N_{aa} = 5$.

In this case, the probability of assembling the RNA for a cell consisting of 12 proteins, each with N_a amino acids, would be $f_{RNA} = (1/10)^M$ where

$$M = 14.4N_a + 28.9 - 12q. \qquad \text{(eq. 12)}$$

We still need two proteins to allow DNA to do its work: with only 5 different amino acids to choose from, the chances of assembling these two proteins at random are $(1/5)^P \times 10^{-2q}$ where $P = 2N_a$. Therefore f_{RNA} in the 2-codon world would be equal to $(1/10)^R$ where

$$R = 15.8N_a + 28.9 - 14q. \qquad \text{(eq. 13)}$$

In order that RNA for the first doublet-codon cell could have been assembled at random in the first 1.11 billion years of Earth's existence, we must satisfy the equation

$$15.8N_a + 28.9 - 14q_d = 65 \qquad \text{(eq. 14)}$$

where subscript d denotes that we are dealing with doublet codons.

What is the minimum size of a protein in a world with $N_{aa} = 5$? In our previous discussion of our modern world where $N_{aa} = 20$, we have argued that proteins with $N_a = 14$ are the smallest functional units. Does this argument remain valid when N_{aa} is reduced to a value as small as 5? The answer is not obvious. For lack of alternatives, we will assume that N_a cannot be less than 14 in a functional protein in the $N_{aa} = 5$ world.

With this assumption, we find that q_d cannot be less than 13.2. This is many orders of magnitude smaller than the value of q_{ra} which is required in the three-codon world. At first sight, this might appear to represent a large increase in protein specificity. However, results from the three-codon world are not relevant here. Instead, we need to compare the new estimate of q_d with the total number of distinct proteins that are possible in the primeval soup. With 5 distinct amino acids in the soup, and with each protein containing 14 amino acids, we see that there are some 5^{14} distinct possible proteins. Therefore, in this case, $Q_{max} = 10^{9.8}$, i.e. $q_{max} = 9.8$. In view of the requirement that Q be at least as large as $10^{9.8}$, we see that the q_d required for random assembly of the RNA for the first *cell again exceeds its maximum permissible value*, this time by 3.4. That is, once again essentially all proteins are required to perform the task of all other proteins. We are faced once again with the problem of lack of protein specificity.

To satisfy the demands of specificity, we again introduce the fraction $F = 10^{-f}$ of all available proteins which are able to perform the task of (say) energy production. As before, we write $f = m \log(N_p)$ where m lies between 1 and $q_{max}/\log(N_p)$. (With the above numbers, $m_{max} = 9.1$, and the average specificity m_{av} takes on a value of about 5.) In view of this, we see that the probability of assembling RNA for the first cell by chance in the 2-codon world becomes one in 10^c where

$$c = (N_p+2)[m\log(N_p) - q_{max} + q_d\ ]. \qquad \text{(eq. 15)}$$

Since the difference $q_d - q_{max}$ is now "only" 3.4 (as opposed to 4.6 for the 3-codon case), we see that the probability of random assembly of the RNA for a (12-14) cell has increased in the 2-codon case by at least 16-17 in the exponent. This is a great improvement indeed relative to the 3-codon case.

However, even with absolutely marginal specificity of protein tasks, i.e. m = 1, the probability P_r of assembling RNA randomly in the primeval soup for a (12-14) cell which uses only $N_{aa} = 5$ distinct amino acids is no better than one in about 10^{63}. If the specificity has its average value $m_{av} = 5$, then $P_r = 10^{-123}$. Even if the value of m is much smaller than m_{av} (say m = 2-3), and with more realistic numbers of proteins in the cell (say $N_p = 30$), the chances of randomly assembling the RNA for the first cell in the primeval soup using doublet codons is no better than one in 10^{200}.

18. What about singlet codons?

We might (in principle) improve the chances of randomly assembling the first cell if the genetic code were able to operate with singlet codons (instead of doublets or triplets). However, it seems unlikely that such a world can exist. It is known that folding of a protein simply cannot be achieved using an amino acid set that is as small as 3 (Riddle et al. 1997): on the other hand, folding can be achieved if the set of amino acids is as large as 5. For the sake of argument, let us make the extreme assumption that folding CAN occur with an amino acid set consisting of only 4 species in the primeval soup. In this case, a singlet codon (one of the four bases for each amino acid) would in principle suffice for the RNA to encode for the amino acids, although with zero redundancy (and therefore no error protection).

However, in order to assemble an accompanying DNA molecule, we also need to have start and stop codons. That is, we must encode not merely for the 4 amino acids, but also for the start/stop codons. This means that the DNA is required to encode for at least 6 elements. This cannot be done with singlet codons (if only four bases are available.)

We conclude that the doublet-codon world is as simple as we can go and still have access to the flexibility of the genetic code.

19. A window of opportunity

When we considered what was probably the simplest example of a doublet-codon world, with $N_{aa} = 5$, we found that random assembly of the first cell turned out to be more probable than in the triplet codon case with $N_{aa} = 20$. But still, the probability P_r is very small.

However, this is not the only example we might consider. Doublet codons with 4 useful bases can in principle encode for a "vocabulary" of proteins made with N_{aa} in the range from 5 to 14 (allowing for start and stop codes). And if proteins still consist of $N_a = 14$ amino acids, then the maximum available number of proteins Q_{max} increases from 5^{14} to 14^{14} as N_{aa} increases from 5 to 14. That is, q_{max} increases from 9.8 to 16.0. The corresponding values of m_{max} in a 12-protein cell are 9.1-14.8 (with $m_{av} = 5.05$-7.9).

Returning to the expression we obtained for the probability P_r of random assembly of RNA for the first cell in a doublet codon world, 1 in 10^c, we recall from eq. (15) that

$$c = (N_p+2)[m \log(N_p) + q_d - q_{max}]$$

where $q_d = 13.2$ (for proteins with 14 amino acids each) and m has a value of at least 1. Inserting q_{max} values in the range from 9.8 to 16.0, we see that the difference q_d-q_{max} is no longer in all cases positive definite. In fact, when N_{aa} grows to a value as large as 9, the value of q_d-q_{max} becomes for the first time negative (-0.2). This will certainly improve the probability of random assembly.

However, if we insert numerical values, and set the specificity to its average value ($m_{av} = 7.2$), we find that in a (12-14) cell, the value of the exponent c for the case $N_{aa} = 9$ becomes 106. If we allow the protein specificity to fall to a very small value, say m = 2, then c becomes 28. That is, the probability that the RNA of

the first cell with N_{aa} = 9 was assembled by chance in the first billion years of the primeval soup might be as large as 1 in 10^{28}. These represent large improvements over the probabilities we have considered above.

Moving on to even larger values of N_{aa}, the formal probabilities of random RNA assembly become even larger. In fact, with N_{aa} = 11, the probability P_r approaches unity if m has a value less than $1.4/\log(N_p)$. Thus, in a (12-14) cell, a value of m less than 1.3 would ensure that P_r could have a value of order unity if N_{aa} = 11. Such a cell could have had its RNA assembled randomly with high probability in the primeval soup in an interval of 1.11 billion years.

In the limiting case N_{aa} = 14 in the doublet codon world, a (12-14) cell could be assembled randomly with high probability (in fact, with near certainty) in 1.11 billion years as long as $m\log(N_p)$ does not exceed the numerical difference between q_{max} and q_d (i.e. 16.0-13.2 = 2.8), i.e. as long as m does not exceed 2.5. This represents the widest opening of the window of opportunity for the random assembly of the RNA for a (12-14) cell.

We note that a specificity of less than 2.5 is much smaller than the average m_{av}: for the case N_{aa} = 14, m_{av} has the value 7.9. If the protein specificity index in the primeval soup was indeed as large as the value m_{av}, the probability P_r of assembling the first (12-14) cell randomly in a doublet codon world is no more than one in 10^{80}.

The window of opportunity in the doublet-codon world has an interesting property that is relevant to the modern world. For a 14-acid cell where the number of proteins is as large as in the smallest known modern cell (N_p = 250), the probability of random assembly P_r could have approached unity as long as m is in the range 1.0-1.17. This is a very restricted window: but it is a bona fide window. It indicates that, provided all of the various optimized conditions are satisfied, random assembly of a (250-14) cell might have occurred with high probability in the young Earth with N_{aa} = 14.

However, the restricted window for the N_p = 250 cell closes altogether if we have overestimated by too much the number of collisions in the primeval soup. As was mentioned in Section 13.3, our choice of 10^{65} for the value of n_r (the total number of reactions experienced by bases or amino acids in the primeval soup) may be too large by 10 or more orders of magnitude. If n_r is in fact equal to 10^{58} (or less), then q_d increases to 13.7 (or

more). In this case, the probability P_r (= 1 chance in 10^c) falls far below unity even if m has its minimum possible value (m=1): the exponent c takes on the value 24.7 (or more).

Values of m as small as 1.17 or 1.3 (or even 2.5) represent marginal specificities; they are far below the average specificities, and are close to the absolute minimum value of m (=1). Whether living cells could in fact survive (and replicate faithfully) in the present of such marginal specificities is not known. At the very least, it is a cause for concern in the context of cell robustness.

The above calculation suggests formally that random assembly of the first cell could have been achieved in the primeval soup if certain conditions were satisfied. The requirements are: (i) at least 11 distinct amino acids were available for use in the making of proteins; (ii) 4 distinct bases were available for the DNA; (iii) the protein specificity index m did not exceed 2.5 (for a cell with 12 proteins); (iv) the number of amino acids in the polypeptide chain of each protein equals 14; (v) the total number of reactions between bases or amino acids in the primeval soup was 10^{65} ; (vi) we accept the RNA-first theory of cell assembly.

If any of these conditions was violated in the young Earth, the probability of random assembly quickly falls to very low values.

20. Entropy constraints on the window of opportunity

At this point in the argument, we need to ask: is the mathematical scenario described in Section 19 relevant in a robust biological world?

In order to address this, we need to consider a certain aspect of coding theory (Yockey, p. 5). The Central Dogma of biology states that DNA encodes for protein assembly but proteins do not encode for DNA assembly. To ensure this, coding theory states that the "vocabulary" at the source (e.g. DNA) must have significantly more symbols than the "vocabulary" at the receiver (amino acids).

In the modern world, there is no problem with this requirement. With 64 codons in the DNA, and only 20 amino acids in (most) proteins, there is a large excess in the "mutual information entropy" of DNA compared to amino acids. The maximum information content of a DNA sequence is 5.931 bits per codon, whereas the information content of an average protein sequence is 4.139 bits per amino acid (Yockey, p. 175). (These numbers are close to the

definition of Shannon entropy for the source $\log_2(64)$ and receiver $\log_2(20)$ respectively: the slight differences arise because not all modern amino acids are encoded with equal probability.) The difference dH between 5.931 and 4.139 (dH = 1.792 bits per codon) is (in the language of coding theory) a measure of the difference in Shannon entropy between source (DNA) and receiver (proteins). (Shannon entropy has nothing to do with the Maxwell-Boltzmann-Gibbs entropy of thermodynamics). Because of this difference in entropy, DNA can communicate information to amino acids, whereas amino acids cannot communicate information back to the DNA.

The large amount of redundancy (represented by the ratio of 64 to 20) in the modern DNA "vocabulary" relative to the amino acid "vocabulary" allows for error checking in the course of cell replication. With the proper use of redundancy, the channel capacity theorem (Yockey, p. 115) indicates that the error rate in a code can be kept below any specified level. This is essential for cells to ensure reliable and consistent replication in the course of many generations.

As one possible measure of the level of error protection in a code, we may refer to some results obtained by Yockey (p. 73). It turns out that in a protein with N amino acids, the number of high-probability states N(h) in parameter space is 2^{NH} where H is the Shannon entropy per amino acid. In the event that all sites have equal probability of occupation by each and all of the N_{aa} distinct amino acids, the value of N(h) becomes equal to $N_{aa}{}^{N}$, as expected from the probability arguments we have used in this paper. In view of the formula for N(h), it seems reasonable to use, as a measure of error protection in the translation from DNA to proteins, the number $E = 2^{NxdH}$. In the case of a modern protein such as insulin (with N=51), E has a value of 3×10^{27}, and we interpret this to mean that insulin is extremely well protected in the modern world from errors in transcription.

Now let us return to the doublet codon option in the primeval soup. A world containing 14 distinct amino acids in the proteins (plus one start and one stop code) would correspond to a doublet code in which the source has 16 symbols but the receiver also contains 16 symbols. In this situation, where $dH = \log_2(16/16) = 0$, there is zero entropy difference between source and receiver. As a result, E = 1, and the measure of error protection for (say) insulin would be some 27 orders of magnitude smaller than it is in the modern world. Replication of insulin in such a situation would be subject to intolerable uncertainty.

Moreover, the Central Dogma of biology would break down: a protein (such as insulin) would be able to control DNA just as much as DNA controls proteins. This hardly seems like a prescription for hardy life forms: there are too many options for lack of reproducibility.

However, the break-down of the Central Dogma in the $N_{aa} = 14$ world suggests that in such a world, one might consider not only the RNA-first theory, but also a "protein-first" theory. The numerical factors entering into our estimates of the probability of random assembly would then change. Thus, the value we have used above for q_d (=13.2) (obtained from eq. (14)) would have to be changed to a value determined from a modification of the expression for z in eq. (1). We recall that eq. (1) refers to the case where the set of distinct proteinous amino acids contains 20 entries. Here, we have only 14 entries in the set, and as a result, z changes to $13.8N_a - 12q$. Setting z equal to 65 and $N_a = 14$, we find $q_d = 10.7$. The window of opportunity now widens somewhat: for the case $N_{aa} = 14$, the value of P_r approaches unity as long as the specificity index m does not exceed 4.9. This is still well below the average value m_{av} (= 7.9). Thus, we are still forced to confront the requirement that protein specificities are quite small.

A doublet codon world, if it is to be of interest to biology in the context of error-free replication, must certainly contain less than 14 distinct amino acids. How much less than 14 should we consider? We have seen that there is a good probability that RNA can be assembled randomly as long as N_{aa} has a value of 11 or more. Including a start and a stop codon, this means that the genetic code must use 16 symbols at the source to encode for 13 (or more) amino acids. The difference in Shannon entropy between source and receiver for this case is $\log_2(16/13)$, i.e. dH = 0.3. With such a value of dH, the error protection E of insulin would fall to 4×10^4, i.e. some 23 orders of magnitude weaker than the protection which exists in the modern genetic code. And for the cases $N_{aa} = 12$ and 13, the values of dH are 0.19 and 0.09 respectively. The corresponding values of E for insulin would be 826 and 24, i.e. up to 26 orders of magnitude less protection than in the modern world.

Although it is sometimes claimed that error protection "must have been" less in the early genetic codes than in the modern world, this is not necessarily true. On the contrary, to ensure that reliable replication occurs among millions of cells of even a single species, it appears that the earliest genetic codes "must have been nearly as accurate as those of today, otherwise even short proteins could not have been transmitted in sufficient

numbers" (Yockey, p. 338). In other words, if the earliest genetic codes were error prone, biology would not have been possible.

In order to ensure the same error protection between source and receiver which exists in the modern world, there should be similar redundancy to what exists in the modern world. That is, the ratio of the number of codons in the DNA to the number of symbols in the amino acids should be comparable to the modern value (64/20 = 3.2). This suggests that, at an epoch when there were 16 codons in the DNA code (if there was indeed such a "doublet-codon epoch" in the early Earth), the value of N_{aa} should have been 5. This is precisely the case we considered in the Doublet Codon section. The Central Dogma would be just as robustly valid in such a world as it is in today's world. However, the chances of randomly assembling such a cell is (as we have seen) only 1 in 10^{63}.

21. Window of opportunity? or bottleneck?

There is a further constraint on the world of doublet codons in which N_{aa} lies in the range from 11 to 14. This has to do with how well protected the genetic code is from noise-induced mutations. Cullmann and Labouygues (1983, BioSystems 16, 9: hereafter C&L) have discussed this issue in numerical detail.

In order to understand the results of C&L, a brief summary of their terminology is necessary. In a doublet code, with 4 bases, there are 16 possible codons. Of these, only a certain number (the "sense codons") are used to encode for proteinous amino acids. The remainder are "non-sense codons" which serve to terminate the translation. Mutations of various types can occur as a result of noise. There is one class of mutations which causes a sense codon to switch to a non-sense codon. In a second class of mutations, a single mutation causes a sense codon to switch to another sense codon. In the latter case, the protein may still function if there are synonymous code entries. But if we dealing with an invariant site, then the protein function is disabled, and C&L refer to a "mis-sense" codon.

C&L have systematically analyzed all possible doublet codons in a world where the number of amino acids being encoded varies from N_{aa} = 0 to 16 (thus including all numbers of interest to us here). In each case, they count up how many single mutations N lead to non-sense codons, and how many single mutations D(1) belong to synonymous and mis-sense codons. C&L point out that the optimal code (as far as immunization from noise is concerned) is one which

minimizes N and which simultaneously maximizes D(1). Codes which have N not too far from its minimum value also possess significant immunization against noise. C&L find that, starting with N_{aa} = 0 and increasing N_{aa} in steps of unity, there is at first a growing number of doublet codes which satisfy the optimal condition.

In the present context, it is important to note that this growth in available codes continues up to N_{aa} = 8, at which point there are thousands of codes which are not far from optimal. But for N_{aa} = 9 and larger, the number of available codes begins to diminish rapidly. For N_{aa} = 12, the number of codes has decreased to the hundreds, and as N_{aa} approaches 16, the numbers drop off towards a value of 1. Thus, as a doublet-codon system attempts to encode for more and more amino acids, there are less and less options the closer N_{aa} approaches 16.

Yockey (p. 190) refers to this as a "bottleneck" which has evolutionary significance. He suggests that doublet codons might have been successful in operating biology as long as N_{aa} was smaller than 16. But as more and more amino acids became available for inclusion into proteins, and N_{aa} eventually increased above 16, it eventually became necessary to go to triplet codons. However, before this happened, and as N_{aa} increased upward through values of 9, 10,…16, the shrinking size of parameter space in which noise-immunized codes can exist would have exposed the organisms of that time to an increasing lack of immunization against genetic noise.

Now, we recall that, in our discussion above, the probability of randomly assembling the RNA for the initial (12-14) cell first rises to large values when N_{aa} is as large as 11. Using the results of C&L, we now see that this value of N_{aa} has a significant property: *it is already past the peak in available numbers of doublet codes*. Thus, we are already approaching the vicinity of Yockey's "bottleneck". This makes it increasingly difficult for an immunized genetic code to handle the large variety of proteins which one might expect to find in a flourishing biosphere.

22. Overview on the window of opportunity

Let us now take an overall look at the window of opportunity in the light of our discussions of the "bottleneck" (Section 21), the entropy (Section 20), and the requisite marginal specificities of proteins (Section 19). Taken in combination, these discussions suggest that what appears as a window of opportunity for random assembly of the first cell (in a formal mathematical sense) may be

subject to several classes of difficulties in the biological context.

It is true that a scenario in which the doublet-codon window opens up to its widest extent describes a system which is interesting from a mathematical perspective. But from a biological perspective, this system suffers from three serious drawbacks. First, in the encoding process between DNA and proteins, error protection is many orders of magnitude weaker than it is in modern organisms. Second, the phase space of permissible genetic codes shrinks to smaller and smaller volumes. Third, a huge number of the available proteins must be able to perform each and every task in the cell: the number is so large that there would have been almost no specificity in protein tasks within a cell. That is, there is a good chance that a protein which is supposed to be used for (say) membrane repair, may switch to one whose function is (say) enabling reproduction.

Any one of these features could be considered as posing significant difficulties for cell survivability. The combination of all three exacerbates the problem. It is difficult to see how a cell (even of the primitive kind we consider here, no bigger than a modern virus) could have survived. For the first robust cell to have developed randomly in the doublet-codon phase of the primitive Earth, conditions must have been "just right" to allow survival in the presence of the above serious drawbacks.

23. Conclusion

We have numerically evaluated the probability P_r that, in the first 1.11 billion years of Earth's existence, random processes were successful in putting together the RNA for the first cell. In estimating P_r, we initially assumed that the first cell follows the rules which guide modern life-forms. That is, we assume there are $N_{aa} = 20$ distinct amino acids in proteins, and triplet codons in the genetic code.

In calculating P_r, we consider only the random assembly of RNA: we assume that once the RNA is present, it will generate the proteins for the cell. (Thus, we are not requiring that the proteins be assembled randomly: if we were to impose such a requirement, the probabilities of random assembly of the first cell would be even smaller than the results we obtain here.) Furthermore, we consider

a cell which is much smaller than those which exist in the modern world. The latter contain at least 250 proteins. By contrast, we have reduced the requirements of the first living cell to a bare minimum: we assume that that cell was able to function with only 12 proteins. Compared to the smallest known living cell, our choice of 12 proteins seems almost absurdly reductionist. Our "cell" looks more like a modern virus (which cannot reproduce itself) than a bona fide cell. But we proceed anyway.

Moreover we also assume that each protein consists of a chain of no more than 14 amino acids. We refer to this as a (12-14) cell. Again, a chain with only 14 amino acids is considerably shorter than the smallest known protein in the modern world (which contains a few dozen amino acids). It is not clear that a protein with only 14 acids would be subject to the 3-dimensional folding which is essential to protein functioning. Nevertheless, we make these reductionist assumptions about a cell with the aim of optimizing the probability of assembling the first cell.

In this spirit, we start with the assumption that the only amino acids which existed in the primitive Earth were the 20 (or so) distinct types of amino acids which occur in the proteins of modern living cells. Also in the spirit of optimization, we assume that the entire pre-biomass of the Earth was in the form of proteinous amino acids. We specifically exclude the non-biological amino acids (numbering more than one hundred) which may have been produced in the primitive Earth. Moreover, we also assume that all 20 of the proteinous amino acids were present solely in the L-isomer form so that the growth of a protein chain is not ended prematurely by unintentional inclusion of a D-isomer. Furthermore, we assume that the initial cell occurred in the physical conditions which are most commonly cited in textbooks, i.e. in a "primeval soup". This allows us to obtain a firm (and generous) upper limit on the number of chemical reactions which could have occurred before the first cell appeared on Earth.

With all of these assumptions, we find that the probability of assembling the RNA required for even the most primitive (12-14) cell by random processes in the time available is no more than one in 10^{79}.

In order to improve on the probability that random processes assembled the RNA for the first cell, we make the (unproven but likely) assumption that proteins in the earliest cells were constructed from a smaller set of distinct amino acids than those which occur in modern cells. In order to ensure that the primitive life forms had a similar level of error protection in their

genetic code as that which exists in the modern world, we consider a case in which the early proteins consisted of only $N_{aa} = 5$ distinct amino acids. For these, the genetic code can operate with doublet codons. In such a world, the probability of randomly assembling the RNA for the first cell in the time available is certainly larger than in our modern (triplet codon) world. But the probability is still small, no more than one part in about 10^{63}.

We have identified a region in parameter space where, once the genetic code exists, the probability of random assembly of the first cell could have reached formally large values in optimal conditions. These conditions include the following: (i) the first cell contained 12 proteins; (ii) each protein in the cell contained 14 amino acids; (iii) there were 4 bases in DNA; (iv) the protein specificity index was no larger than 2.5 (far below its average value); and (v) conditions in the primitive pre-biosphere were such that chemical reactions occurred at their maximum possible rates. (The last of these conditions almost certainly involves an optimization which is unrealistic by as much as 10 orders of magnitude.)

(Note that we have said nothing about how the genetic code came into existence. We merely assume that it is already in operation. The origin of the code is a more formidable problem than the one we have addressed here.)

If mathematics were the only consideration, our conclusions would suggest that the RNA for the first cell *could* have been assembled randomly in the primeval soup in 1.11 b.y. once there was a code and abundant supplies of between 11 and 14 distinct proteinous amino acids. However, when we take into account considerations of coding theory (especially the necessity to protect the proteins from errors of transcription), it appears that this region of parameter space is hostile to protein production. And the genetic code has to pass through a "bottleneck" in order to enter into the modern world, with its 20 proteinous amino acids. As a result, the first cell might have had serious difficulties surviving as an autonomous biological system.

Finally, the extreme nature of our assumptions regarding the first cell (12 proteins, each containing 14 amino acids) can hardly be overstated. If a cell is to fulfil even the minimum requirements of a Von Neumann self-replicating machine, it probably needs at least 250 proteins. Even with multiple optimizations in our assumptions about the primeval soup, the window of opportunity for creating such a cell in 1.11 b.y. narrows down to a very restricted region in phase space: (I) there must have been exactly

14 distinct amino acids in the cell proteins, (II) the protein specificity index must have been between 1.0 and 1.17, and (III) at least 10^{58} chemical reactions must have occurred between the bases (or amino acids) in 1.11 b.y. The "fine tuning" of such conditions presents a problem. However, there are more serious problems than fine tuning: error protection in the genetic code fails altogether in these conditions. Even the Central Dogma of biology breaks down. A cell formed under these conditions would truly be subject to serious uncertainties not only during day-to-day existence but especially during replication. The cell could hardly be considered robust.

Nevertheless, as Yockey (p. 203) points out, the possibility that an organism from the doublet-codon world might have survived the "bottleneck" may have some empirical support. According to the endosymbiotic theory (L. Margulis 1970, *Origin of Eukaryotic Cells,* Yale Univ. Press, New Haven CT), mitochondria might have been at one time free-living bacteria which now survive in a symbiotic relationship with the cytoplasma of other cells. In mitochondria, the genetic code differs somewhat from the code in other cells. Perhaps mitochondria are representative of organisms which originated in the doublet-codon world, but which could not survive on their own because of the difficulties associated with the hostile zone of parameter space where they originated.

In summary, if the first cell actually originated by random processes, the genetic code must already have existed, and conditions must have been "finely tuned" in order to trace a path through a narrow (and hostile) region of parameter space. The idea that some of the constants of the physical world have been subject to "fine tuning" in order to allow life to emerge, has been widely discussed in recent years (e.g. in the book by J. D. Barrow and F. J. Tipler, *The Anthropic Cosmological Principle*, Oxford University Press, 1994, 706 pp). If we are correct in concluding that "fine tuning" is also required in order to assemble the first cell, we might regard this conclusion as a biological example of the Anthropic Principle.