

# **Probability of randomly assembling a primitive cell on Earth: Part II**

By Dermott J. Mullan, [mullan@bartol.udel.edu](mailto:mullan@bartol.udel.edu)

## **Summary:**

In Part I of this paper (*PCID*, Oct.-Dec. 2002), we estimated the probability of randomly assembling a cell containing 12 proteins, each of which consists of a chain of 14 peptides. Here, we apply some of the discussion in Part I to laboratory experiments, and show that the results of these experiments are not inconsistent with our discussion. Our analysis leads us to consider also the probability of assembling larger proteins, consisting of hundreds of amino acids. We show that, in the “bottleneck” between the doublet-codon and triplet-codons world, there is a high probability that such large proteins can be assembled randomly in the primeval Earth. In the presence of very low protein specificity, even cells consisting of 250 large proteins can be assembled randomly with formally high probability. However, in the absence of error protection in the genetic code in the bottleneck, reliable replication of such cells cannot be guaranteed.

## **1. Probability of cell formation: dependence on number of peptides in a protein**

In Part I, we estimated probabilities of randomly assembling a primitive cell in the first billion years of Earth’s existence. The description “primitive” was used for a cell with requirements that we consider the absolute minimum for a functioning cell: 12 proteins, each containing 14 amino acids (aa). We refer to such an organism as a (12,14) cell. We considered a case in which proteins and RNA both have to be assembled randomly, and a second case in which only RNA has to be assembled. We refer to the latter case as the “RNA-world”.

In the present paper, in order again to optimize the probability of cell formation by random assembly, we restrict attention to the RNA-world. Then referring to eq. (11) in Part I, we see that, for a cell consisting of  $N_p$  proteins, each of which is a polypeptide containing  $N_a$  amino acids (aa), the probability of random assembly is 1 in  $10^b$  where

$$b = (N_p + 2) [ m \log(N_p) - q_{\max} + q_{ra} ] \quad (\text{eq. 1})$$

The notation in eq. (1) is as follows.

The quantity  $m$  is a measure of protein specificity. At one extreme, with  $m$  close to its minimum value ( $m=1$ ), essentially all of the distinct polypeptides in the primeval “soup” can perform the task of any particular protein in a cell (such as energy generation, or membrane maintenance, or waste disposal). We refer to the case  $m=1$  as the minimum possible specificity for protein function. At the opposite extreme, each protein in the cell is highly specific in its task: then only a small number of polypeptides are able to perform the task of a specific protein. In the limiting case, only a single specific polypeptide is

capable of performing each task in the cell. In the latter case, protein specificity is maximized, and  $m$  takes on its maximum value of  $q_{\max}/\log(N_p)$ .

Lower case  $q$  in eq. (1) refers to the logarithm to base 10 of a quantity  $Q$ . The value of  $Q_{\max}$  is the maximum possible number of distinct polypeptides (each containing  $N_a$  aa) which can be assembled by choosing from the available set of distinct proteinous amino acids (of which there are  $N_{aa}$ ). In the world in which we currently live,  $N_{aa}$  is equal to 20. In Part I, we gave reasons for considering the case  $N_a = 14$  as a minimum number of aa in a functional protein. Since in each of the  $N_a$  positions in the protein, we can choose from  $N_{aa}$  distinct aa, the value of  $q_{\max}$  is readily determined to be  $N_a$  times  $\log_{10}(N_{aa})$ . Thus, in our current world, we find that  $q_{\max}$  is equal to  $1.3N_a$ .

The quantity  $Q_{ra}$  (where subscript  $ra$  denotes Random Assembly) is the number of distinct polypeptides which must have functioned as cell proteins in order that the first primitive cell could have been randomly assembled in the primeval Earth in the time available (i.e. in the first billion years). Comparing eqs. (3), (7), (9), and (11) of PRA, we see that  $q_{ra}$  scales as  $1.8N_a$ . The numerical value of the coefficient 1.8 is determined by the properties of the DNA code: with four bases to choose from, a triplet-codon world contains  $N_c = 64$  codons. The coefficient of  $N_a$  in the expression for  $q_{ra}$  is equal to  $\log_{10}(N_c)$ .

In eq. (1), the key feature to which we draw attention in the present paper is the *algebraic difference* between the two quantities  $q_{ra}$  and  $q_{\max}$ . We are especially interested in how the difference ( $q_{ra} - q_{\max}$ ) depends on  $N_a$ .

Using the details given above, we see that the difference  $q_{ra} - q_{\max}$  varies as  $(1.8-1.3)N_a$ , i.e. as  $0.5N_a$ . The numerical value of the coefficient 0.5 is of crucial importance in the following discussion: the numerical value is equal to  $\log_{10}(N_c / N_{aa})$ .

In the language of DNA encoding theory (see, e.g. H. P. Yockey's book "Information Theory and Molecular Biology"),  $N_c$  is the number of "vocabulary words"  $V_s$  in the "source" of the code, while  $N_{aa}$  is the number of "vocabulary words"  $V_r$  in the "receiver". The ratio of  $V_s$  to  $V_r$  is related to the "mutual entropy": in information theory, the mutual entropy is expressed in terms of the logarithm to base 2 of  $(V_s/V_r)$ . For the case discussed above, this mutual entropy  $\log_2(64/20)$  has a numerical value of about  $0.5/\log_{10}(2)$ , i.e. about 1.7. This is (essentially) the numerical constant which appears in a more careful evaluation of the mutual entropy in the genome (Yockey 1992, p. 124). Yockey's estimate for the constant (1.7912) is based on a more precise calculation where he includes the fact that not all amino acids in modern proteins are encoded with equal frequency.

## 2. Application to a single protein

In Part I, we considered the case of assembling an entire (primitive) cell, with a set of 12 different proteins. However, we can also apply the discussion to the case of a single protein. If we are interested in randomly assembling a single protein, rather than a complete cell with  $N_p$  proteins, the exponent which appears in the probability (i.e. the  $b$  in

eq. (1)) is replaced by essentially  $b_1 = q_{ra} - q_{max}$ . (Here, we have also simplified the discussion by assuming that no extra proteins are required for the processes of replication: we assume that the RNA which constitutes the RNA-world act as ribozymes to catalyze replication.) Thus, we have a particularly simple expression to describe the chances of assembling a single protein at random. In our modern world, we find the simple result  $b_1 = 0.5N_a$ .

With this formula in hand, we see that the larger  $N_a$  becomes, the larger the numerical value of the exponent  $b$  in eq. (1) (and also the larger  $b_1$ ). Larger  $b$  corresponds to lower probability for random assembly. That is, as we expect, a protein with a large value of  $N_a$  is less likely to be assembled randomly than is a protein with smaller  $N_a$ . Using the above expressions for  $b$  and  $b_1$ , we now have a quantitative estimate of how much less those chances would be. Specifically, if protein A is (say) 40 peptides longer than protein B, the probability of randomly assembling A is  $10^{20}$  times smaller than the probability of randomly assembling B.

We have already determined out (in Part I: Section 14) that in a triplet codon world in which proteins are composed of 20 distinct amino acids, each with  $N_a = 14$  aa, the numerical value of the difference  $q_{ra} - q_{max}$  is  $22.8 - 18.2 = 4.6$ . This means that the probability of assembling one protein of this kind in a time interval of one billion years in the primeval Earth would be 1 in  $10^{4.6}$ . But the chances of assembling a longer protein, say one with 54 aa, would be 1 in  $10^{24.6}$ , i.e. 20 orders of magnitude smaller. In Part I, we chose  $N_a$  as small as 14 specifically in order to optimize the chances of randomly assembling the first cell.

We stress again that the coefficient of  $N_a$  in the expression for  $b_1$  is determined by the ratio of the vocabulary  $V_s$  in the source ( $V_s = N_c$ ) to the vocabulary in the receiver ( $V_r = N_{aa}$ ).

### **3. RNA re-assembly by Eklund, Szostak, and Bartel (1995)**

Now let us see how we can apply the above discussion to some results which appear in a paper by Eklund, Szostak, and Bartel (Science vol. 269, p. 364, 1995) (hereafter ESB). In that paper, the focus is on a certain class of RNA, namely, those which perform catalysis. Although catalysis in living organisms is usually reserved for proteins (enzymes), there are also certain RNA molecules which exhibit catalytic properties. Molecules in this class of RNA are referred to as ribozymes.

The aim of ESB is to create ribozymes of various lengths by a more or less random process, and determine the frequencies with which long and short ribozymes emerge from their experiment. The ESB experiment provides a meaningful (although indirect) way of testing the  $N_a$ -dependence that we have derived for the frequency of creating proteins of varying lengths.

The essential difference between my calculation (in Part I, and above) and the ESB work is as follows. In Part I, the main focus was on how RNA translates into proteins. But the

aim of the ESB experiment is quite different: it is to start with RNA sequences (containing a certain number of random bases, 72 or 76), and translate this into new RNA sequences.

In view of the nature of their experiment, ESB discuss their results in terms of the number of nucleotides (nt) in the chain, whereas our discussion is cast in terms of the number of aa in the chain. However, the conversion between these two measures of chain length is simple: in view of the triplet codons in DNA, 3 nt correspond to one aa.

ESB describe how they created a number of different ribozymes with various lengths. Their results show that they more frequently create short ribozymes than longer ones. We can be quantitative about this. To be sure, ESB do not find differences in frequency as large as  $10^{20}$ , as we described in the above numerical example. However, ESB do indeed find that sequences of different length appear with different frequency. And the difference is in the sense that our arguments would suggest: namely, short sequences emerge more frequently than long sequences.

Specifically, ESB report that the frequency of their shortest ribozyme (containing 31-56 nucleotides: ESB refer to this ribozyme as a4-10t) is almost  $10^4$  times greater than the frequency of their longest sequence (containing 191-239 nt: ESB refer to it as e3-10t). (See Table 1 on p. 365 of ESB: note especially the numerical values of product/enzyme in 24 hours in a 5 microM substrate.) The factor of  $10^4$  is noteworthy. The work of ESB provides us with an example where it is  $10^4$  times easier to create a short sequence than a long one. Other examples of different frequencies can be found in Table 1 of ESB: the differences in frequency are not as extreme as  $10^4$  for other pairs of ribozymes. In fact, in some cases, the differences in frequency are quite small, as if there was very little (or almost no) difference whether the chain is long or short. ESB find that some long chains are created almost as frequently as some short chains.

For reference, we note that in the extreme case noted above, the a4-10t ribozyme of ESB would translate to a peptide chain containing roughly  $N_a = 10-19$  aa, while their e3-10t ribozyme corresponds to  $N_a = 64-80$  aa. That is, the difference between the  $N_a$  values of the corresponding “proteins” is at least 45, and at most 70.

How can we understand the ESB results in the context of the formulas in Part I and those given above? We note that ESB start with RNA, and “translate it” into new RNA. In the language of coding theory, the source is RNA, and the receiver is also RNA. Both source and receiver in the ESB experiment have “vocabularies” of four elements. As a result,  $V_s = 4$  and  $V_r = 4$ . This is a very different case from translating from  $V_s = 64$  in the (DNA) source to  $V_r = 20$  aa in the (protein) receiver. In the ESB case, the ratio of  $V_s/V_r$  is formally unity.

Now we recall our formula for the difference in probability of randomly creating two proteins with different  $N_a$  values: the probability is 1 in  $10^b$  where b is proportional to the difference in  $N_a$  values. The coefficient of proportionality is equal to  $\log_{10}(V_s/V_r)$ . Formally, this logarithm in the ESB experiment is zero. Therefore, formally, the probability of forming chains of various lengths is *independent* of the difference in chain

length. That is, long chains should be formed with essentially equal probability to short chains. Some of the results of ESB bear out this conclusion.

We can be more quantitative. As noted above, the shortest and longest sequences discussed by ESB correspond to proteins which differ in their  $N_a$  values by an amount that is in the range 45-70. Following the general rule of our formulas, we predict that random processes will create the shorter of these proteins  $X$  times more frequently than the longer, where  $X = 10^x$  and  $x = (45-70) \text{ times } \log(V_s / V_r)$ .

Referring to the results of ESB, where the exponent  $x$  is reported to be roughly equal to 4, we find that the source and receiver in the ESB experiments must differ in their vocabulary sizes by a factor of  $10^{4/(45-70)}$ , i.e. 1.14-1.23. This is certainly much smaller than the vocabulary ratio of  $64/20 = 3.2$  which is pertinent to the encoding of proteins by RNA.

Is the numerical result of 1.14-1.23 for the vocabulary ratio reasonable for experiments involving only RNA? Perhaps, if we use a rough entropy argument as follows. It is well known that although DNA and RNA both consist of sequences of 4 bases, the 4 bases are not exactly the same: in DNA, there are 4 bases which belong to the set TACG, but in RNA the set is UACG. Thus, the vocabulary of the DNA/RNA world in one sense contains 5 “words” in total. Therefore, if some T were to be present (accidentally) in the ESB experiment, the source vocabulary might on occasion be as large as 5, while the receiver vocabulary would be 4. In such a case, the ratio of  $V_s / V_r$  could take on values as large as 1.25. It may be coincidence that this value is close to the upper limit of 1.23 that we have extracted from the ESB experiment. But a coincidence is always worth noting.

#### **4. Minimum constituents of proteins: the work of Wilson et al. (2001)**

Wilson et al. (Proc. Nat. Acad. Sci. vol. 98, p. 3750, 2001) describe an experiment in which, starting with an enormous library of DNA segments, random polypeptides are created by “expressing” the information that is encoded in each DNA sequence. The aim is to determine which of these “expressed” polypeptides have a particular property. The property that is being sought is that the polypeptide should have a significant affinity for a particular compound (streptavidin).

In contrast to the ESB work (where RNA was “translated” into RNA, with mutual entropy of zero), Wilson et al. “translate” DNA into protein. Therefore, the full mutual entropy of 1.79 bits per amino acid plays a role in their results. As a result, if Wilson et al. were attempting to construct a library containing peptides with varying  $N_a$  values, then we might use their results to test the prediction that the frequencies of random generation should scale as  $10^{-b}$  where  $b=0.5N_a$ . However, Wilson et al. create polypeptides which are all of precisely the same length ( $N_a = 88$ ). Therefore, their results are not useful to test the  $N_a$ -dependence of the exponent  $b$ .

However, there is one aspect of the Wilson et al. work which is of interest in the context of a different section of Part I. Wilson et al. found that among the polypeptides which bind to streptavidin, long polypeptides (with  $N_a = 88$ ) are more effective at binding than the shorter polypeptides (with  $N_a = 5-38$ ) are. They cite enhancements in binding by factors of 200-2200.

This raises the question: how easy is it for polypeptides of differing lengths to bind to a particular molecule? The answer to this question has a direct bearing on how well an enzyme will perform its function. As Wilson et al point out, the long polypeptides which bind effectively to streptavidin often contain, at some location in the chain, a certain well-defined sub-sequence (a “motif”) of peptides. This suggests that functionality in proteins may depend on collections of highly degenerate motifs which operate in a modular fashion. The level of protein activity is observed to go up in proportion to the number of copies of the motif that are present (Lu et al., Proc. Nat. Acad. Sci. vol. 97, p. 1988, 2000). It appears as if protein activation depends on a repeated structure composed of small units acting additively. Since longer chains have more opportunities to contain multiple copies of the motif, it is not surprising that long proteins have enhanced affinities, as Wilson et al. report.

This leads us to a point which has a direct bearing on an argument that was central to Part I: how short can an individual motif be? This is related to the question I raised in Part I: what is the shortest possible protein? In Part I, I suggested that at least two units of secondary structure are required: this led to the conclusion that  $N_a = 14$  is a plausible choice. Some of the works that are quoted by Wilson et al. are pertinent to this point. For example, Tanaka and Herr (Mol. Cell Biol. Vol. 14, p. 6056, 1994) report that triplets of  $N_a = 18$  polypeptides are necessary for certain proteins to perform their task. Blair et al. (Mol. Cell Biol. Vol. 14, p. 7226, 1994) suggest that pairs of  $N_a = 11$  polypeptides suffice to activate certain proteins. Giniger and Ptashne (Nature, vol. 330, p. 670, 1987) suggest that  $N_a = 15$  suffices. Abedi et al. (BMC Mol. Biol. Vol. 2, p. 10, 2001) suggest  $N_a = 11$ , while in other cases, a chain as short as  $N_a = 8$  may suffice (Pollack and Gilman, Proc. Nat. Acad. Sci. vol. 94, p. 13388, 1997).

It is not just the numerical value of  $N_a$  that is important for the structural motif: the structural arrangement of the aa in the protein also plays a role. Hope et al. (Nature vol. 333, p. 635, 1988) suggest that the basic unit of activation may be a pair of alpha-helices: if the helices cannot form, the efficiency of protein function disappears (Giniger and Ptashne, loc. cit.). Thus, my idea in Part I of taking at least a pair of secondary structures (such as alpha-helices or beta-sheets) as the minimum construct appears to be a plausible choice.

## **5. Probability of randomly assembling large proteins at the bottleneck**

So far, we have considered the shortest possible proteins, in order to enhance the probability of random assembly of an operational cell. Our choice  $N_a = 14$ , although plausible as a choice for minimum protein length, is actually much shorter than typical

proteins. For example, hemoglobin contains several hundred aa. It is not even clear, from a physical perspective, whether a protein as short as 14 aa could undergo folding at room temperature: such a short structure might have great difficulty in creating the tertiary structure that is basic to protein function.

We should therefore consider the probability of assembling randomly a cell with proteins which are much longer than  $N_a = 14$ . As we have seen, since the exponent in the probability formula  $b_1$  is equal to  $0.5 N_a$ , the chances of assembling larger proteins rapidly become very small compared to the probability of assembling a shorter protein.

However, there is one important exception to this rule. It occurs when we are at the “bottleneck” where the doublet codon world has reached saturation, and there are exactly 16 amino acids to be encoded by 16 codons. In this situation, the vocabulary of the source has  $V_s = 16$  “words”, and the vocabulary of the receiver also has  $V_r = 16$  “words”. Therefore, we are again in a situation analogous to that which we discussed in connection with the ESB experiment above: the quantity  $\log_{10}(V_s / V_r)$  has a value of zero.

Now, it is precisely this quantity which is the coefficient of  $N_a$  in the expression for  $b_1$ . As a result, at the bottleneck, we arrive at the remarkable conclusion that the probability of random assembly is *independent of the numerical value of  $N_a$* . We have already seen (in Part I) that, as we approach the bottleneck, the chances of random assembly of a primitive cell (with  $N_a = 14$ ) become high. Now we see that, as long as we are precisely at the bottleneck, the chances of random assembly remain high even if  $N_a$  is allowed to take on an arbitrarily large value. That is, proteins as large as any of those which occur in modern organisms can be assembled randomly with equal probability to that for small proteins. This highlights the important role that the “bottleneck” may play in the random assembly of the first cell on Earth.

Referring to eq. (1) above, the probability of randomly assembling a cell with  $N_p$  proteins may still be reduced to small values if  $N_p$  is allowed to become too large. But the dependence is only logarithmic. As was pointed out in Part I (section 19), we can increase  $N_p$  to values as large as 250 (the minimum number of proteins that is believed to be required for a functioning cell), and still have high probability of random assembly of the first cell, as long as  $m$  is no larger than 1.17. The window of opportunity for random assembly of the first cell remains open (albeit only slightly) for a cell with as many as 250 proteins, each of which may contain hundreds of aa.

Of course, this does not solve the error-prone nature of the bottleneck. As was discussed in Part I (Sections 20 and 21), replication of a cell in which  $V_s$  equals  $V_r$  takes place without the possibility of error correction in the genetic code. This is an unstable situation for any organism. It is not clear how well cells in the saturated doublet-codon world will survive the process of replication through the bottleneck.