

This is a response to Art (Member No. 179) who posted a criticism of my paper on 23 Nov 2002 at 19.28.

Art's criticism is based on an incomplete reading of my paper. He seems to have read only as far as p. 20, where there is the phrase " $f_{12}$  is roughly  $(1/10)^y$  where  $y = 15.6 N_a$ ". Art uses this formula to arrive at the following conclusion: the probability of random assembly of long peptide chains (large  $N_a$ ) should be much smaller than random assembly of small chains (small  $N_a$ ). Clearly, if the above formula for  $y$  was the only important factor, then two peptide chains which differed by (say) 40 in their  $N_a$  values, would differ in probability by  $10^{624}$ . Art quotes two published papers which show that when small and large chains are assembled randomly in the laboratory, the differences in probability are nowhere near as large as this.

Art concludes that my formula for  $f_{12}$  leads to results which are inconsistent with experiments. If the above formula for  $f_{12}$  were the final derivation of my paper, I would agree with Art's conclusion.

However, a key point in which my article differs from a lot of earlier work (such as that described in the thread Art refers to) is that, in the next paragraph but one after the formula cited by Art, I proceed to a quantitative discussion of protein specificity. This discussion is couched in terms of the index  $q$ , and it leads to a crucial revision of the above expression for  $y$ : the revision appears as eq. 1 on p. 22: the probability of random assembly of 12 proteins is not  $(1/10)^y$  but  $(1/10)^z$ , where  $z = 15.6N_a - 12q$ . And there is analogous equation for random assembly of RNA (eq. 3).

What is essential here is that the probability depends now NOT on a single term (as in the equation for  $y$  quoted by Art), but on the algebraic DIFFERENCE between two terms.

The presence of this algebraic difference makes a great difference. In fact I will show that, contrary to Art's conclusion, my formulas are in quantitative agreement with some results which appear in the paper by Eklund, Szostak, and Bartel Science 269, 364, 1995 (hereafter ESB).

To prove this, I refer to the value which  $q$  must have if random assembly (subscript "ra") of the first cell RNA occurred: the answer is given by the solution of eq. 7 (p. 29). The value I obtain for  $q(\text{RNA})_{\text{ra}}$  is 22.8 (p. 30). And the question which is central to the physics of cell assembly is: how does this value of  $q_{\text{ra}}$  compare with the available volume in phase space? I address this on p. 31, where I revert to the quantity  $q_{\text{max}}$  (introduced on p. 21).

This leads to the key formula in eq. (11): the probability of random assembly is 1 in  $10^b$  where  $b$  depends on the DIFFERENCE  $q_{\text{ra}} - q_{\text{max}}$ . The occurrence of this difference is crucial for my discussion.

Now, I need to note here that the expressions in my paper are all for the case of a 12-protein cell, in which each of 12 proteins must be assembled randomly. However, to

make a proper comparison with the experiments in the laboratory, we need to consider the case of a single protein. In that case, the chance of random assembly  $1$  in  $10^b$  reduces essentially to  $b = q_{ra}(1) - q_{max}(1)$  where the term  $(1)$  denotes the value appropriate to a single protein.

My response to the essential aspect of Art's criticism is the following: how does  $q_{ra}(1)$  depend on  $N_a$ , the number of amino acids in the protein, and how does  $q_{max}(1)$  depend on  $N_a$ ? Answer: both factors depend linearly on  $N_a$ , but the coefficient of the linear dependence is different in  $q_{ra}(1)$  from what it is in  $q_{max}(1)$ .

For  $q_{max}$ , the coefficient is already at hand: on p. 21, we see that  $q_{max} = 1.3 N_a$ . And for the value of  $q(RNA)_{ra}$ , we simply refer to eq. (2) and divide  $E$  by 12. The coefficient of  $N_a$  in  $q(RNA)$  is then readily found: it is 1.8 .

Now we are in a position to estimate quantitatively how likely it is that a peptide chain of length  $N_a$  amino acids will be assembled randomly: it is  $1$  in  $10^b$  where

$$b = (1.8-1.3)N_a = 0.5N_a. \quad (\text{eq. 1a})$$

Let us now revert to the example quoted above in the first paragraph: suppose two proteins differ by 40 in the number of amino acids in their peptide sequences. Then eq. 1a indicates that the probability of randomly assembling the RNA for the shorter sequence should exceed the probability of randomly assembling the longer sequence by  $1$  in  $10^{20}$ .

Art points out that this huge estimate is inconsistent with the lab work of ESB: Art says that in the ESB article, "the frequency of long catalysts was roughly equal to short ones". In view of this, Art concludes that my theory is wrong.

However, I claim that the ESB results are actually consistent with my formulas, provided that appropriate changes are made to reflect the fact that the ESB experiment has a different from the aim of my calculation. My aim is to start with RNA and make proteins: the aim of ESB is to start with RNA and make RNA.

To see the differences which arise from this, we need to understand where the numerical values appearing in the difference in eq. 1a ( $=1.8-1.3$ ) originate. The factor 1.8 is the log (to base 10) of 64 (the number of triplet codons), while 1.3 is the log (to base 10) of 20 (the number of distinct amino acids in proteins). Therefore, my formulas are applicable to the case where information in RNA is being used to construct proteins. The difference  $1.8-1.3 = 0.5$  ( $=\log(64/20)$  in base 10 arithmetic) corresponds in coding theory to the difference in entropy between source vocabulary ( $V_s = 64$  codons in RNA) and receiver vocabulary ( $V_r = 20$  amino acids in protein). In base 2 arithmetic (which is useful for coding theory), the entropy difference is not 0.5 but 0.5 times  $\log_2(10)$  i.e. about 1.7. This is (essentially) the numerical constant which appears in evaluating 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 are encoded equally.

So what does this have to do with the ESB experiment? There, the focus is on a certain class of RNA, namely, those which perform catalysis. Although catalysis is usually reserved for enzymes (proteins), there are also certain ribozymes which are RNA molecules with catalytic properties. It is the ribozymes which are the focus of the ESB study.

In this regard, there is an essential difference between my calculation and ESB. In my paper, I am interested in 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.

My calculation shows how to calculate the probability of this translation. The result is again  $1$  in  $10^b$  where  $b$  is still equal to the difference between two terms, both of which are proportional to the number of items in the sequence. In the discussion above, the number of items in the sequence was  $N_a$ , the number of amino acids. Here, ESB are dealing with sequences of bases, and so we should use  $N_b$ , the number of bases. However, instead of  $b$  being equal to the difference between 1.8 and 1.3 times  $N_b$  (as in eq. 1a), the coefficients must be changed to reflect the difference of entropy between source and receiver in the ESB experiment.

What is the source? RNA, with its vocabulary of 4 bases. What is the receiver? Also RNA, with its vocabulary of 4 bases. The entropy difference between source and receiver is formally zero. Therefore,  $b$  is formally zero. In this limit, the probability of randomly assembling a sequence of  $N_b$  bases is formally independent of the number of bases in the sequence. This is consistent with Art's quotation of ESB results: "the frequency of long catalysts was roughly equal to short ones".

But let us be more precise. Although ESB do not find differences in frequency as large as  $10^{20}$ , they DO find some differences in the frequency of appearance of sequences of different length. 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 catalyst sequence (containing 31-56 nucleotides: ESB refer to it as a4-10t) is almost ten thousand times greater than the frequency of their longest sequence (containing 191-239 nt: ESB call it e3-10t). (See Table 1 on p. 365 of ESB: values of product/enzyme in 24 hours in a 5 microM substrate.) The factor of 10 thousand is noteworthy. ESB provide us with at least one example where it is  $10^4$  times easier to create a short sequence than a long one.

We may even be permitted to draw a quantitative conclusion from the ESB results as follows. Quantitatively, the ESB sequence (a4-10t) with 31-56 nt corresponds in our calculations to a hypothetical protein with  $N_a = 10-19$ . And the ESB sequence (e3-10t) with 191-239 nt corresponds to a hypothetical protein with  $N_a = 64-80$ . The difference between these proteins amounts to at least 45 (and at most 70) in the value of  $N_a$ . 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  $x$  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 creation of proteins from 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, the 4 labels are TACG, but in RNA the labels are UACG. Thus, the vocabulary of the DNA/RNA world actually 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.