**A calculation of the probability of assembling the first protocell**

The probability of assembling some minimal ingredients for a primitive organism into a discrete protocell is found to be surprisingly high.

George H. Shaw 1, Roger Hoerl2

1Geology Department, Union College, Schenectady, NY 12308 ([shawg@union.edu](mailto:shawg@union.edu)), 2Mathematics Department, Union College, Schenectady, NY 12308

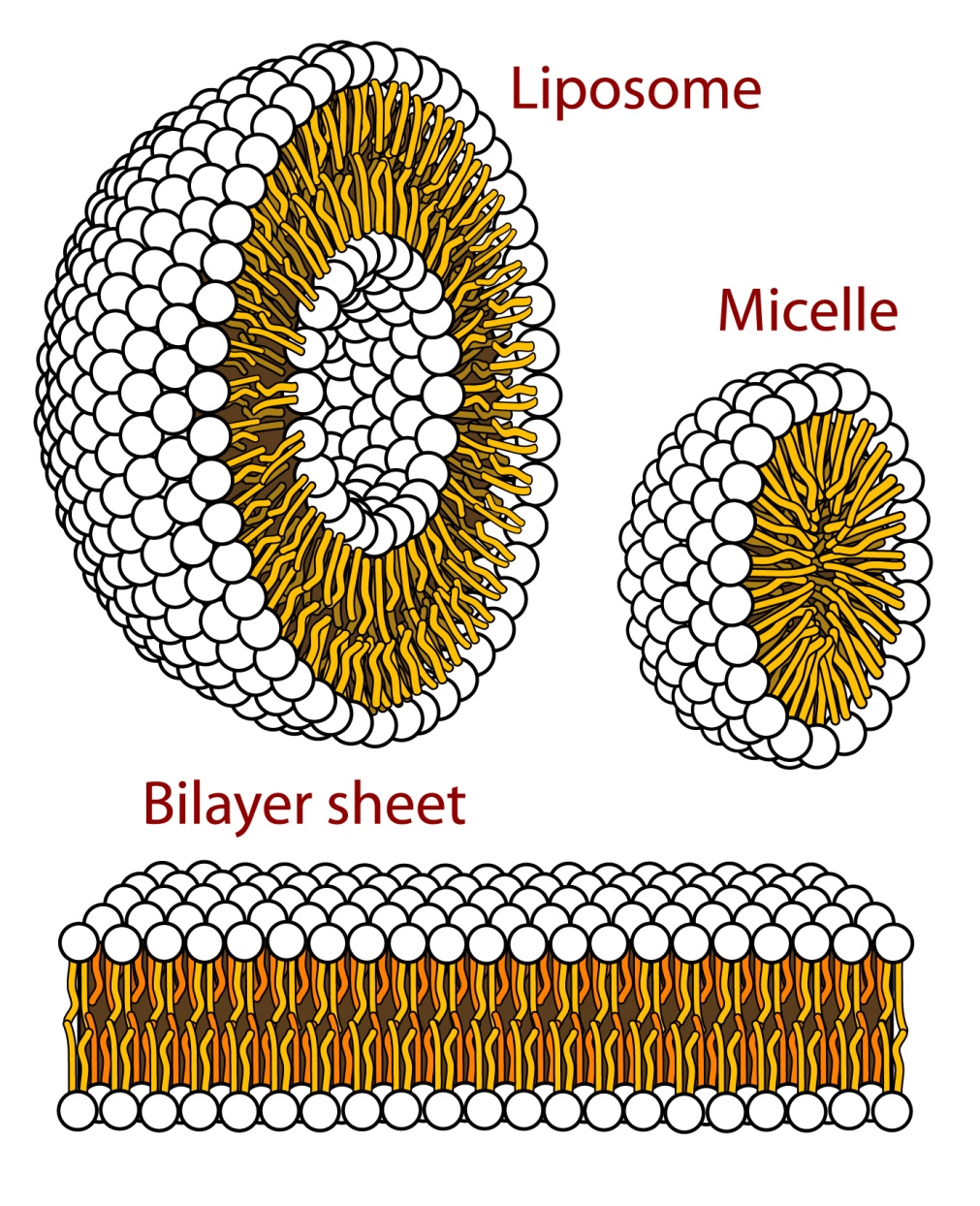
Abstract: Calculating the probability of life emerging from a “primordial soup” appears to be extraordinarily difficult. This is certainly due, in part, to difficulty in formulating just what constitutes the most primitive form of life. If the most primitive form of cellular life (ignoring possible pre-cellular entities) is thought to contain some irreducible collection of molecular ingredients contained within an envelope, it appears to be possible to estimate the likelihood that such a collection would arise. Assuming the existence of possible enveloping particles such as liposomes and a collection of macromolecules resulting from prebiotic chemical processes, and a likely minimum number of molecular components, leads to surprisingly high probabilities that at least one of a large collection of such liposomes will contain the requisite minimal suite. The concentrations of the various components necessary, and the amount of “soup” required, are well within ranges that might be expected on a prebiotic Earth. These results are certainly consistent with the idea that the emergence of life took place very early in Earth’s history, very shortly after the surface environment cooled to the point where liquid water became stable. It also suggests that even if subsequent large impact events sterilized the surface, re-emergence of life would quickly take place.

DRAFT

There have been few attempts to calculate the probability of life emerging from an abiotic environment. The necessary conditions are still rather speculative and quantitative estimates of various factors must necessarily be subject to large errors. Most of the purported calculations revolve around intrinsically improbable occurrences, such as the assembly of a particular enzyme (or collection of enzymes) of significant length, assuming that this is somehow a prerequisite for “life”. Famously, Hoyle (*1*) compared the probability of assembly of the first cell to the probability of an aircraft arising by chance from a pile of parts (or junk). All such calculations are inherently flawed in assuming that something as complex as a modern cell is required for “life”, as has been eloquently shown by Musgrave (*2*). Few scientists interested in the emergence of life would accept this assumption, and there is a growing body of experimental research involving what could be described as chemical “evolution” (*3-6*). It is not our purpose to critique any of the above approaches, few of which are quantified in a way meaningful to the emergence of life as such, although the last mentioned certainly provide quantification of certain specific aspects of molecular (prebiotic) evolution. Our goal is to examine one particular step in the emergence of life, perhaps the key step necessary to arrive at a very primitive proto-organism. Our incentive for formulating these particular probabilities is based on prior ideas concerning the nature of certain minimal requirements for life as we know it (*7*), incorporating ideas first proposed by Eigen (*8*) and expanded on by de Duve (*9*). We are particularly concerned with the emergence of what might be called a “protocell” containing some minimal assemblage of interacting molecules. Previous ideas concerning assembly of parts on a solid surface such as sulfides (*10,11*) or clays (*12,13*) might conceivably resemble life, but would then require the transfer of the molecular community to an organic container. If assembly of life’s components first took place on a catalytic surface which then led to a self-reproducing ensemble, this might be more probable than what we are proposing as a single step from molecules in solution to an organic protocell. As such, our calculations may suggest some kind of minimal probability by comparison.

The point of our calculation may be better understood after a discussion of what we think are the likely prerequisites for a very primitive life form, that is something capable of storing, reproducing and transmitting information while maintaining enough “personal” integrity to do so. Extant organisms all operate using the same macromolecular components: DNA, RNA and proteins, although some rather simple organisms forego DNA, with information storage carried out solely by RNA. DNA information storage appears to be essential for all higher organisms, probably because of the relative stability it provides in maintaining coherence for large genomes. The RNA world (*14-16*) describes an earlier, and much simplified era prior to the emergence of DNA, and encompasses presumably much simpler, and likely more “malleable”, genomes. Most discussions of an RNA-world assume the presence of the necessary translation machinery, particularly a suite of transfer RNAs (tRNA) capable of reading some form of messenger RNA (mRNA) to synthesize polypeptides (proteins). All of these macromolecules are polymers, of nucleotides in the case of RNA and DNA, and amino acids in the case of proteins. There are other compounds probably essential for life including sugars, fatty acids and phospho-lipids, also probably necessary to support metabolism or serve as structural elements of a protocell. It is very likely that such “packaging” was a prerequisite before life could be maintained, probably in the form of something like micelles or liposomes constructed of the fatty acids and phospho-lipids (Fig. 1)

Fig. 1. Liposomes, micelles and lipid bilayers. From Mariana Ruiz Villarreal (*27*).



All of the necessary simple precursor molecules are presumed to have formed by abiotic processes from atmospheric gases and reactions in primordial water bodies. Other possible sources of prebiotic organics have been discussed at length (*17*). It seems likely that whatever the initial delivery mechanism to Earth’s surface, there was considerable reprocessing of organics in hydrothermal, volcanic, atmospheric and aqueous systems. An extensive literature deals with almost all of the necessary prebiotic syntheses of precursor compounds (for a recent rather comprehensive treatment see *6*). It is likely that a “primordial pond” would have included a considerable inventory of these various simple components, but simply having the components is insufficient for the emergence of life. The RNA-world model suggests that some suite of macromolecular polymers must have been present in order to have self-sustaining life, but production of both polypeptides and polynucleotides of 50+ units appears to happen readily in a matter of weeks in silicate-catalyzed conditions (*18-20*); One of the most critical steps in emergence of life is the creation of what Eigen refers to as a hypercycle, a positive feedback between information storage molecules and polypeptide (?) catalysts that influence the production and preservation of same. The requisite step becomes significantly simpler if information storage and translation to polypeptides is carried out by the same suite of molecules (that is, tRNA). Shaw (*7*) has examined the possible role of tRNA as a carrier of information directing polypetptide synthesis. Fig. 2 shows the results from comparing possible short polypeptide sequences that might be coded by primal tRNAs with a large modern protein database.

Fig. 2. Comparison between consensus tRNA coded and randomly generated “tRNA like” coded amino acid strings (*7*).

The low “E-values” for tRNA derived polypeptides, compared to those for random RNA sequences of similar length supports the idea (*8,9*) that tRNA might actually be the first “genes” by recording information for useful polypeptide synthesis. This is a major simplification in the RNA-world, as separate mRNA is not required, the tRNA serving a dual function. This logically leads to the question of whether it is likely that a sufficient suite of tRNA would be present in a protocell in order for translation to proceed. Although somewhat speculative this suggests that assembling a population of tRNA molecules sufficient to carry out translation may be a critical step in producing the biochemical “feedback” (Eigen’s hypercycle) required for life. Our purpose here is to address the question of whether it is at all likely that such an assemblage has a reasonable probability of arising from a “primordial soup” of potential components.

We examine one crucial aspect of this problem, namely whether, given a mixture of tRNA that includes species that are active in translating for specific amino acids, and a collection of possible “packages” such as liposomes, there is a reasonable probability that enough different active tRNA can be found in any one liposome such that translation of a polypeptide might take place. This is obviously still some distance from life, but it appears to be a key step among (perhaps) many. Our basic assumptions are these:

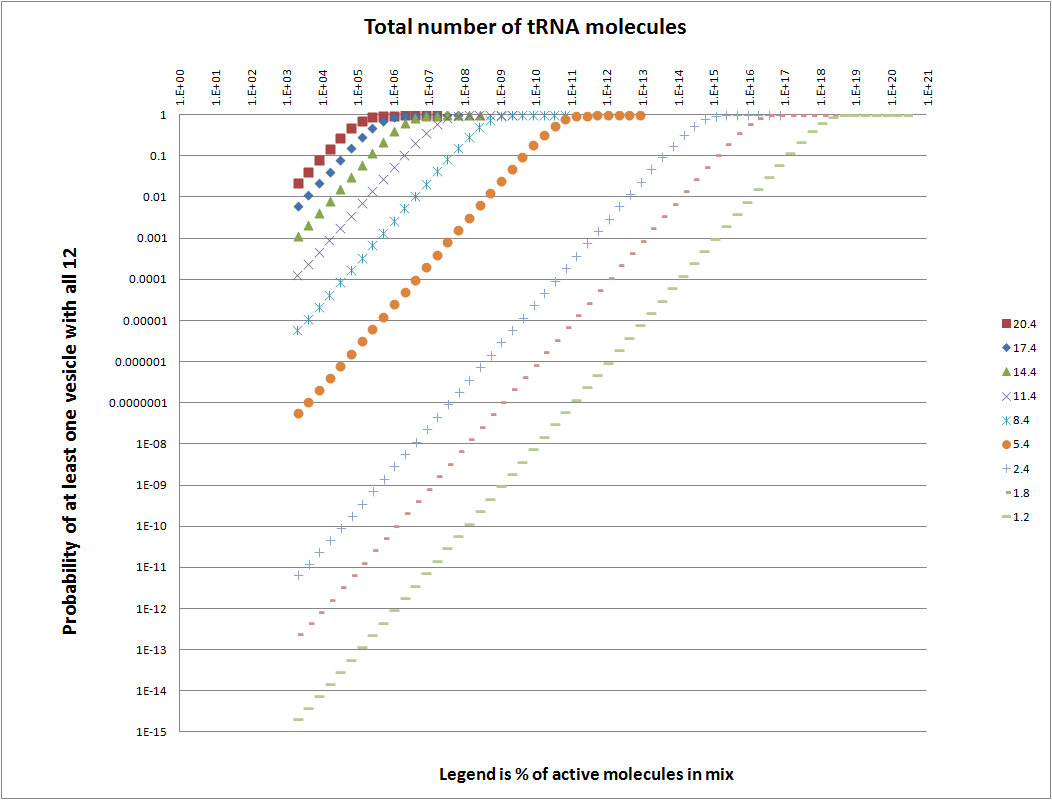
1. Mononucleotides formed abiotically will polymerise under abiotic conditions to the extent of producing tRNA-sized macromolecules (*18-20*)
2. tRNA-like polynucleotides will be selectively preserved in the primordial soup as a consequence of relative chemical stability due to base-pairing and folding. They will thus accumulate over time to provide a significant population of potentially active agents for translation.
3. Some fraction of the accumulated tRNA will be active while (perhaps) the majority will be inert, that is, not useful for translation. It is virtually impossible to know the relative amounts of active and inert tRNA-like molecules, but we err on the side of the majority being inert.
4. Liposomal and/or micellular packages would also be formed in the pool along with amino acids and other prebiotic molecules potentially useful to the earliest life form. Our focus is on the liposomes and the tRNA-like molecules rather than the other components, but it is important to note that the “primordial soup” will likely contain an abundant supply of prebiotic monomers such as amino acids and nucleotides, and various high energy compounds, lipid molecules and small polymers.

Our problem is to calculate the probability that at least one liposome in a primordial pool will contain at least one of each of the essential active tRNA to allow polypeptide synthesis. We assume that the original essential suite of tRNA must number twelve rather than 20. It has been suggested that 10 is the minimal number of amino acids necessary to produce an active polypeptide (*21*). Our assumption that the requisite number is twelve errs on the side of greater difficulty, as it is somewhat less probable that twelve will be present simultaneously in a package than ten.

The calculation scheme is summarized in the supplementary materials. We carry out these calculations for different assumptions about the relative amounts of active and inactive tRNA and different total numbers of tRNA-like molecules. We assume that the number of tRNA-like molecules in a given liposome is 50 and that each active tRNA is present in the same concentration in the “soup”. (Note that a simple bacterium perhaps 1 m in diameter may contain over 100,000 tRNA molecules.) For example, if there are 2000 tRNA-like molecules present and there are 8 of each of the active tRNA, there will be a total of 8 x 12 = 96 active tRNA and 1904 inactive tRNA-like molecules present. In this case there will be 40 vesicles present, each containing 50 molecules (most of which will be inactive). Increasing the total number of molecules by a factor of two while keeping the ratio of active to inactive molecules constant allows us to look at the effect of progressively larger total populations of tRNA (and vesicles) on the probability of finding at least one liposome with the requisite 12 active tRNA. The calculations are also repeated with different ratios of active to inactive molecules to assess the impact of having higher or lower concentrations of the active species.

The results are illustrated in Fig. 3, which shows the probability of finding all twelve active tRNA in at least one liposome for changes in the various parameters.

Fig. 3. Probabilities of finding all 12 translating tRNA in a single liposome, under various assumptions.



Not surprisingly, at higher concentrations of active tRNA, and at larger numbers of total molecules, the probability of finding all twelve active tRNA in a liposome increases. Even at rather low concentrations of active tRNA species the probability of finding all twelve active species in at least one package nears certainty provided the total number of molecules (and also “packages”) increases to sufficiently large numbers. The question remains: How reasonable are these large numbers? And: Would we expect to find them in a primordial pond? The answers to those questions are certainly speculative, but some ideas can be obtained by examining possible amounts of organic precursors that might be available in a given body of water. This, of course, depends on assumptions concerning the amounts of various chemical components that might be available, and certainly assumptions about the processes responsible for generating the various complex units of our calculation scheme. These questions are among the goals of an enormous amount of synthetic chemical research which need not be repeated here (3-6), but which suggest that abiotic processes can produce not only interesting precursors, but can do so in rather large amounts, and can in principle lead to the requisite large polymers (*18-20*).

Even the earliest experiments of Miller (*22,23*) gave yields of 15%+ for interesting products, and this was in experiments without likely natural catalysts such as clays or other silicates, and over rather short periods of time. Prebiotic chemistry experiments are the basis for assumptions made above in the context of the following estimates. As a “test reactor” we posit a small pool of water, perhaps to be thought of as a tide pool (one of an enormous number) on the edge of the primordial sea at the shore of a proto-island arc land mass. If we take the dimensions to be 10m x 10m x 1m deep, it will have a volume (mass) of 108 cm3 (gm). Regardless of the exact composition of the adjacent ocean, in terms of organic molecules, it is certainly conceivable that the organic fraction dissolved in this pool could be as much as 20%. One milliliter of this pond could contain perhaps 20 billion 0.1 m diameter liposomes at a concentration of 3 micromolar. Assuming each liposome contained some 80,000 lipid molecules of approximately 750 atomic mass units each, they would amount to about 2 ppm of the solution (25). If we further assume that there are enough tRNA-like molecules (75 nucleotide units in each) to provide 50 for each liposome, the “tRNA” would amount to somewhat more than a trillion molecules of same in each ml of solution. If the average molecular weight of the tRNA was roughly 30,000 AMU, the weight would be about 6 x 10-8 gm in each ml, for a concentration of about 60 ppb by weight. This figure might be doubled by assuming that an equal number of tRNA was in the pool, but not incorporated in the liposomes. Further, we could also assume a pool of nucleotide monomers (and smaller polymers), maybe ten times as much by mass, bringing the mass of nucleotide molecules to about 1 ppm. In other words, the organic fraction necessary to provide enough “packages” and tRNA components in amounts similar to those required for reasonable likelihood of a least one liposome having a complete suite of active tRNA (see Fig. 3) in one ml of solution requires that only a very small percentage of the organic components in the pool (0.01%) is present as liposomes and tRNA. Considering the size of the entire pool, the number of “experiments” occurring in the pool would be truly enormous, and the number of “attempts” at producing a viable proto-organism similarly large.

To be sure, the emergence of life clearly requires more than the assembly of a suite of active tRNA in a package such as a liposome, but such would seem to be a minimum, and on the face of it one of the more difficult to attain. One likely requirement would a primitive ribosome to catalyze translation and polypeptide synthesis. Recent work on the history of development of the ribosome points to a relatively small active center (or two) consisting of about 100-150 nucleotides (*25*). These are small enough to have developed by the same chemical evolutionary processes, based on comparative stability, as seems to have been the case for tRNA. Because they are somewhat longer polymers it might be thought intrinsically less probably that active units would arise. Against this might be set the idea that particular configurations of the active centers could have been “directed” by association with previously produced tRNA. After all, the active centers clearly must have strong interactions with tRNA in order to serve as catalysts. The implications for our calculations would include adding two additional macromolecules (albeit larger) to the 12 tRNA required for translation. To guess how adding two somewhat different molecular entities to the mix, including an estimate as to their relative concentrations, is beyond what we can contemplate at present. Additional requirements would include amino acids, nucleotides, phosphate, etc., to provide the building blocks for metabolism, as well as energy-rich compounds to drive chemical reactions. Our point here is that one of the most critical barriers to life emerging is the assembly of enough of the complex molecular machinery in a package to provide a positive feedback loop involving information storage and production of biological catalysts. Our calculations suggest that such an occurrence is not of particularly low probability, and that other ingredients could be added to enough “trials” that the emergence of life is highly probable. This is even more true when one considers the much larger volume of test reactors that were likely present on the early Earth, and the considerable amount of time available for trials. The recent extension to earlier times of the presence of photosynthesis, as shown by carbon isotope ratios (*26*) is not as surprising as one might expect, even given that photosynthesis is a very sophisticated biochemical process. Some of the estimates of very short times being required for the emergence of life from abiotic conditions may be very close to the truth (9).

References

1. Hoyle, F., 1983, The Intelligent Universe, Michael Joseph Limited, London, 255 pp.

2. Musgrave, I., 1998, Lies, damned lies, statistics, and probability of abiogenesis calculations, The TalkOriginsArchive, accessed 11/23/2015, last update:12/21/1998, <http://www.talkorigins.org/faqs/abioprob/abioprob.html>.

3. Calvin, M., 1969, Chemical Evolution, Oxforn University Press, New York and Oxford, 277 pp.

4. Ponnamperuma, C., 1977, Chemical Evolution of the Early Precambrian, Academic Press, New York, San Francisco, London, 221pp.

5. Mason, S.F., 1991, Chemical Evolution, Clarendon Press, Oxford, 317 pp.

6. Deamer, D. and J. W. Szostak (eds), 2010, The Origins of Life, Cold Spring Harbor Laboratories Press, Cold Spring Harbor, NY, 318pp.

7. Shaw, G.H., 2015, Earth’s Early Atmosphere and Oceans, and the Origin of Life, Springer, New York, 117 pp.

8. Eigen, M., 1977, The Hypercycle: A Principal of Natural Self-Organization, Naturwissenschaften, 64, 541-565.

9. De Duve, C., 2005, Singularities: Landmarks on the Pathways of Life, Cambridge University Press, Cambridge, 258 pp.

10. Wächtershäuser, G., 1988, Microbiol. Mol. Biol. Rev. **52** (4): 452–84.

11. Wächtershäuser, G., 2007, On the Chemistry and Evolution of the Pioneer Organism, Chemistry & Biodiversity **4** (4): 584–602.

12. Cairns-Smith, A.G., 1982, [Genetic Takeover and the Mineral Origins of Life](http://books.google.com/books?id=mnBqQgAACAAJ), Cambridge University Press, Cambridge, UK, 477 pp.

13. Cairns-Smith, A.G., 2009, An approach to a blueprint for a primitive organism", in Waddington, C. H., The Origin of Life: Towards a Theoretical Biology **1**, Aldine Transaction, Edinburgh University Press, Edinburgh, pp. 57–66*.*

14. Woese C., 1967, The genetic code, Harper and Row, New York, pp. 179–195.

15. Orgel L.L., 1968, Evolution of the genetic apparatus, J Mol Biol 38: 381–393.

16. Crick F.H.C., 1968, The origin of the genetic code, J Mol Biol 38: 367–379.

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

18. Ertem, G. and J.P. Ferris, 1996, Synthesis of RNA oligomers on heterogeneous templates, Nature 379: 123-140.

19. Ferris J.P., A.R. Hill, R. Liu and L.E. Orgel, 1966, Synthesis of long prebiotic oligomers on mineral surfaces, Nature 381:59-61.

20. Ferris, J.P., 2002, Montmorillonite catalysis of 30-50 mer oligonucleotides: laboratory demonstration of potential steps in the origin of the RNA world, Origins Life Evol. Bios. 32: 311-332.

21. Longo, L.M. and M. Blaber, 2012, Protein design at the interface of the pre-biotic and biotic worlds, Archives of Biochemistry and Biophysics, 526, 16–21.

22. Miller, S.L., 1953, Production of amino acids under possible primitive Earth conditions, Science 117, 528-529.

23. Miller, S.L., 1955, Production of some organic compounds under possible primitive Earth conditions, Journal of the American Chemical Society, 77, 2351–2361.

24. Encapsula Nano Sciences, 2015, [The number of lipid molecules per liposome](http://www.liposomes.org/2009/01/number-of-lipid-molecules-per-liposome.html), http://www.liposomes.org/search/label/Number%20of%20lipid%20molecules%20per%20liposome

25. Petrov, A.S., B. Gulen, A.M. Norris, N.A. Kovacs, C.R. Bernier, K.A. Lanier, G.E. Fox, S.C. Harvey, R.M. Wartell, N.V. Hurd and L.D. Williams, 2015, History of the ribosome and the origin of translation, [Proc Natl Acad Sci U S A.](http://www.ncbi.nlm.nih.gov/pubmed/26621738), Nov 30. pii: 201509761.

26. Bella, E.A., P. Boehnke, T.M. Harrison, and W.L. Mao, 2015, Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon, PNAS Early Edition, doi/10.1073/pnas.1517557112, 1-4.

27. Villarreal, M.R., 2007, Phospholipids aqueous solution structures, Licensed under Public Domain via Commons, https://commons.wikimedia.org/wiki/File:Phospholipids\_aqueous\_solution\_structures.svg#/media/File:Phospholipids\_aqueous\_solution\_structures.svg

Figures.

Fig. 1. Liposomes, micelles and lipid bilayers. From Mariana Ruiz Villarreal (17).

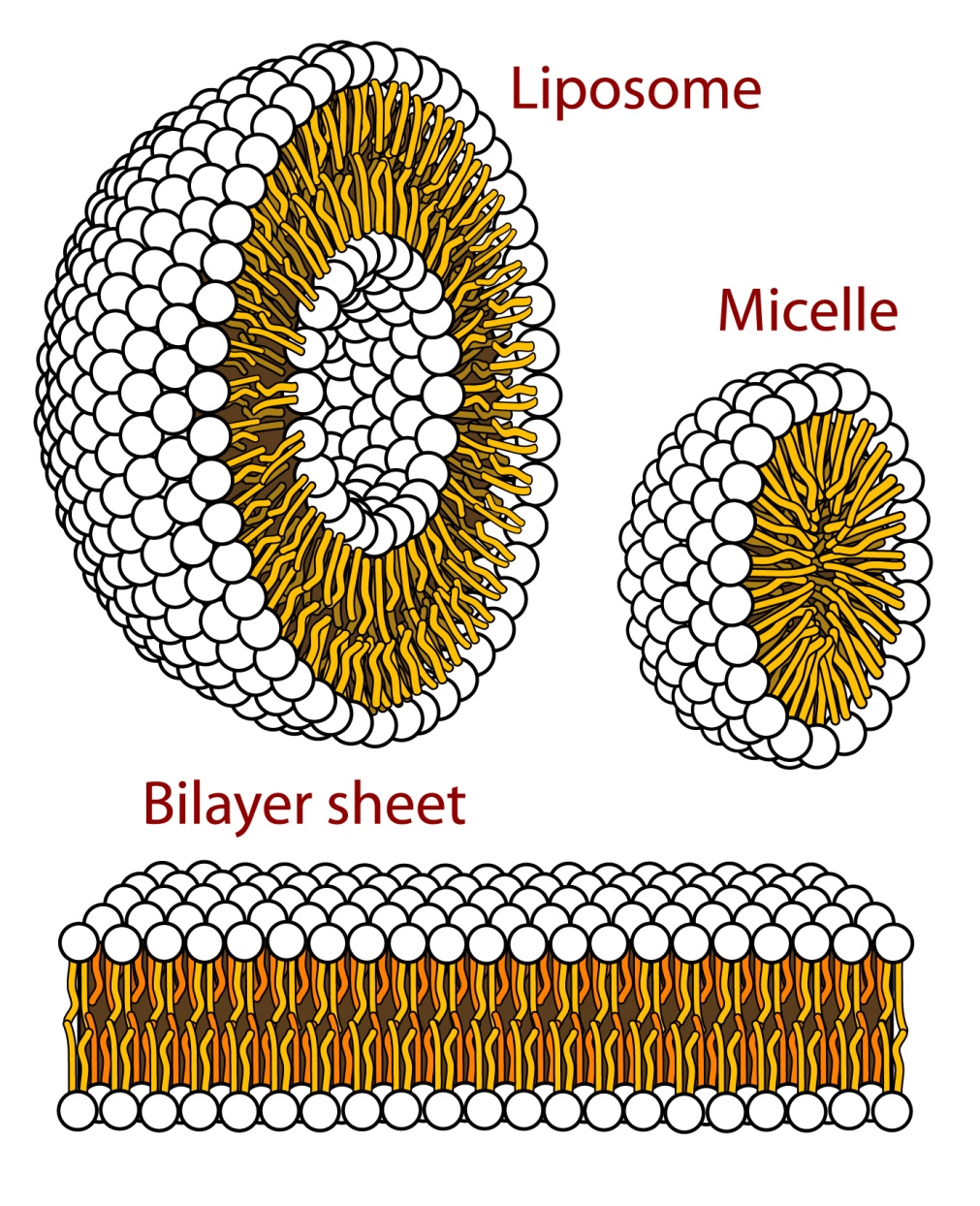
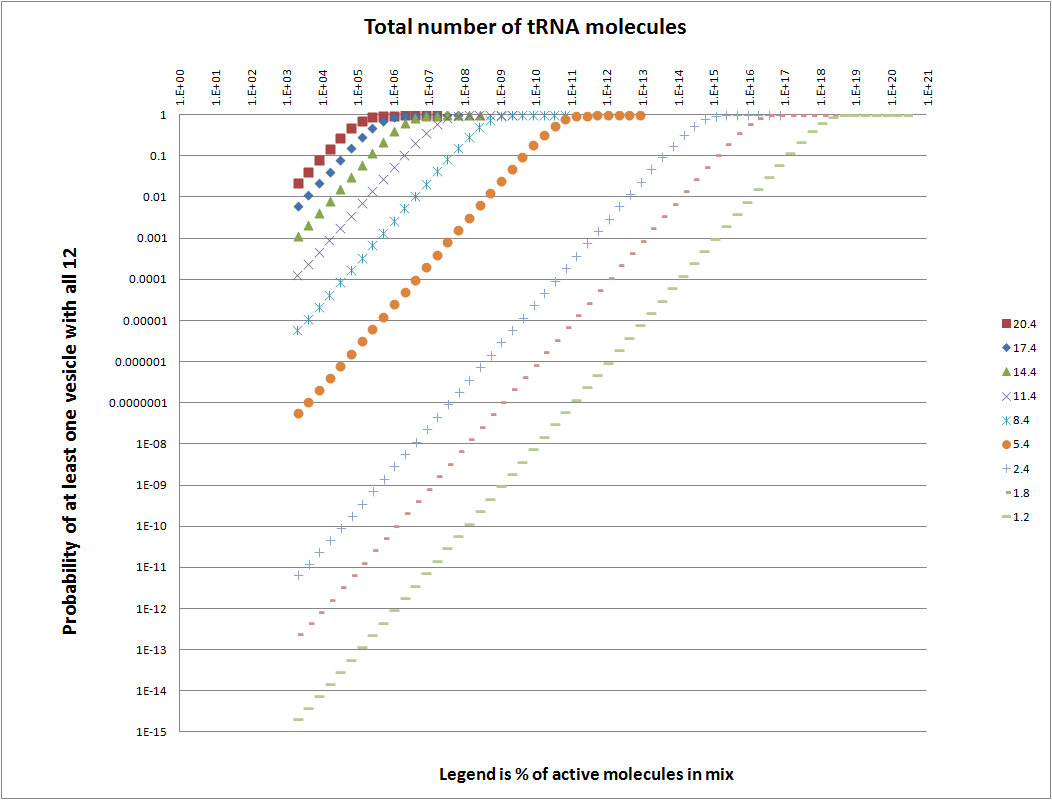


Fig. 2. . Comparison between consensus tRNA coded and randomly generated “tRNA like” coded amino acid strings (7).

Fig. 3. Probabilities of finding all 12 translating tRNA in a single liposome, under various assumptions.



Supplementary materials: method of calculation

Background

* We have a large population of molecules, which we will call N. We look at cases for different values of N up to several trillion.
* Of these, there are 12 types of particular interest that carry out translation. Let ni be the number of molecules of type i, for i = 1-12. We note that ni can take different values for all i depending on the total number of molecules.
* We randomly allocate the molecules to as many vesicles as can be filled by the total number, assuming there are 50 tRNA in each vesicle. That is, for example, vi = 50 for i = 1-20, if we are setting the number of vesicles V = 20 and the total number of molecules to 1000.
* We want to calculate the probably that at least one of the vesicles contains at least one of each of the 12 molecules of interest. In other words, we want the probability that there exists a vesicle for which yi 1 for all i from 1-12, where yi is the number of molecules from type i in the vesicle.

Calculations and Simplifying Assumptions

* In the calculations below we assume that each vesicle contains 50 tRNA-like molecules, assigned at random from a large pool. We vary the size of the tRNA pool and assume a sufficient number of vesicles to accommodate the entire pool of tRNA.
* We will calculate the probability of interest by calculating its opposite, i.e., we first calculate the probably of no vesicles with full matches (i.e., that all vesicles have yi = 0 for at least one i).
* As a first step, we calculate the probability that the first vesicle has no match.
* We will then make the simplifying assumption that the probability of the second vesicle not having a match, given that the first doesn’t have a match, is approximately equal to the probability that the second doesn’t with nothing given, which would be the same as the probability that the first doesn’t.
* We further assume that the probability of no match on the ith vesicle, given no matches on vesicles 1 to (i-1), is approximately equal to the probability of no match on the ith vesicle with nothing given, for i = 1 to 20.
* The resultant probability of none with matches would then be the probability of none on the first to the twentieth power.
* Because of the assumptions on conditional probabilities above, this approach will not give an exact answer, but we feel it will give a reasonable approximation.

Probability of First Vesicle Not Having a Match

* We approach this by considering the probability of a match on the first, and then subtracting this answer from 1. The numbers used in the following discussion/example are for 1000 total tRNA-like molecules, with just two of each active type (that is, a total of 24 active tRNA out of 1000). Obvious adjustments to the numbers can be readily made to explore various options.
* To get a match on the first vesicle, it must have at least one item from all twelve types.
* The probability of getting at least one item from type 1, i.e., of y1 1, is 1 – P(y1 = 0), i.e., one minus the probability of zero molecules of type 1. That is:

P(y1 1) = 1 - P(y1 = 0) = 1 – (998C50)/(1000C50),

because we can select the 50 molecules for vesicle 1 in 1000C50 ways. But if we are to have none from type 1, then we cannot choose either of those items (n1 = 2), so the number of ways to select the 50 items without any from type 1 is 998C50. Therefore, the probability of this happening is the number of ways it can happen, divided by the number of possible outcomes, since each outcome is equally probable.

* We next want to calculate the probability of getting at least one item from type 2, given that we have at least one from type 1. In other words, we want P(y2 1|y1 1).
* We will go through a similar logic to calculate this probability. However, since we know that y1 1, this implies that when we select the molecules for the next vesicle, at most one item from type 1 is available, since one was selected in vesicle 1. We don’t know which was selected, but we know that one molecule from type 1 is unavailable. Therefore:

P(y2 1|y1 1) = 1 - P(y2 0|y1 1) = 1 – (997C49)/(999C49),

Because there are only 999 items available to be selected, and in order to have none from type 2, we cannot select either of these, leaving 997 available to be selected in such as way that y2 = 0. We also only have 49 items to select, since one has been used.

* We then calculate P(y3 0| y2 1 & y1 1) in a similar manner. This would be:

P(y3 1| y2 1 & y1 1) = 1 - P(y3 0|y2 1 & y1 1) = 1 – (996C48)/(998C48),

using similar logic.

* Continuing this process, we end up with the probability of yi 1 for all i from 1-12 being:

P(yi 1 for i = 1-12) = 1-(999-iC51-i)/(1001-iC51-i)] (Equation 1)

Probability of Any Match

* We then use this answer (Equation 1) as the basis of calculating the probability of at least one vesicle having a match. Further, we will first calculate the probability of no vesicles having a match, and then subtract this from 1 to obtain the probability of at least one match.
* The probability of the first vesicle not having a match is just 1-Equation 1, i.e.,

P(no match in V1) = 1 - 1-(999-iC51-i)/(1001-iC51-i)]

* To calculate the probability of at least one vesicle having a match, we make the simplifying assumption noted above, and first calculate the probability of no matches in any of the 20 vesicles. This is:

P(no matches in any vesicle) = P(no match in V1)20

Lastly, P(at least one match) = 1 – P(no matches in any vesicle)