International Journal of Astrobiology 12 (1): 94–98 (2013)
doi:10.1017/S1473550412000468 © Cambridge University Press 2012

The problem of the 'prebiotic and never born proteins'

Gerald E. Marsh

Argonne National Laboratory (Ret) 5433 East View Park, Chicago, IL 60615, USA
e-mail: gemarsh@uchicago.edu

Abstract: It has been argued that the limited set of proteins used by life as we know could not have arisen by the process of Darwinian selection from all possible proteins. This probabilistic argument has a number of implicit assumptions that may not be warranted. A variety of considerations are presented to show that the number of amino acid sequences that need to have been sampled during the evolution of proteins is far smaller than assumed by the argument.

Received 3 August 2012, accepted 24 September 2012, first published online 30 October 2012

Key words: Protein evolution, RNA-world.

The genetic code specifies 20 amino acids, and a typical protein might be made up of a sequence of 200 amino acids. The argument is often made that the probability is negligible than the relatively limited set of about 20 000 proteins coded for human genome (or the far greater number found in the natural proteome) arose by Darwinian selection of random variations out of the 20^{200} possibilities (Chiarabelli & De Lucrezia 2007). As put forth by Chiarabelli and Lucrezia, 'Nature could not have explored all possible amino acid combinations and, therefore, that many proteins with interesting new properties could have never been sampled by Nature. By and large, we are faced with the problem of how the 'few' extant proteins were produced and/or selected during prebiotic molecular evolution … Even knowing a useful method to produce proteins under prebiotic conditions the problem would not be solved. In fact, for example, synthesizing a random 50 mer chain using all 20 different amino acids it is theoretically possible to produce about 10^{65} different sequences and the probability to sample two identical chains is approximately equal to zero.'

Chiarabelli and Lucrezia spoke about proteins being 'produced and/or selected during *prebiotic* molecular evolution'. This means that before the advent of self-reproducing organisms having the capability of both replication and metabolism. When and how the first protein came into being is a critical issue for the origin of life, and will be further discussed later in this paper.

The argument that the number of proteins found in nature could not have arisen by Darwinian selection of random variations brings to mind the Levinthal paradox (Levinthal 1969), which has to do with protein folding times. Levinthal's paradox results from assuming an unbiased random search, and can be resolved by introducing a small energy cost for locally incorrect bond configurations (Zwanzig *et al.* 1992). In this way, the search is transformed into a biased search, which can dramatically reduce the number of configurations that need to be sampled to arrive at a useful one.

From an evolutionary perspective, perhaps the most important proteins are the catalysts known as enzymes. What the probabilistic argument above tells us is that, if one assumes that the synthesis of each of the 20^{200} protein possibilities is equally probable, some of the earliest proteins capable of serving as enzymes could not have arisen from random selection from the set of all possible proteins. That is, they could not result from random variations in the sequence of amino acids followed by the natural selection of primitive organisms utilizing the resulting proteins. The resolution of this conundrum must lie with the origin of early self-reproducing systems. Similar to the Levinthal paradox, the evolution of a set of biologically useful proteins could not have depended on an unbiased selection from all possible proteins. The argument also tells us that during the early development of life, the set of possible proteins had to be sampled in a massively parallel manner.

The way cells currently produce proteins is the result of some three billion years of evolution. Even a cursory examination of the process shows that it is far too complex to have been used by early life: to begin with, amino acids go through the process of 'activation', in which an amino-acid-specific aminoacyl-tRNA synthetase combines a given amino acid with adenosine monophosphate (AMP) – one of the nucleotides that make up RNA – and attaches the activated amino acid to one of some 35 different types of tRNA. These in turn are brought to the ribosome via the guanosine triphosphate (GTP)/guanosine diphosphate (GDP) cycle where mRNA (after the introns are spliced out) instructs the ribosome to produce a specific polypeptide chain. While a majority of proteins produced in this way may fold on their own, many types of cells, from bacteria to higher eukaryotes, use auxiliary folding proteins. This entire process could only have sprung full blown into existence by the intervention of some *deus ex machina*!

The principal difficulty with the evolutionary approach to understanding the problem with protein development is that while the synthesis of amino acids in the reducing atmosphere of the early Earth was demonstrated in the laboratory as early as 1953 (Miller 1953; Miller & Orgel 1974), it is energetically unfavourable for amino acids to combine with polypeptides without the help of a catalyst. In modern cells, peptide bond formation is mediated by the energy in the amino acid–tRNA bond.

There is some recent work that could shed some light on this issue. In studying the origin and evolution of the ribosome (Harish & Caetano-Anollés 2012), it has been found that modern protein synthesis may have evolved from pre-existing functions of primordial molecules. They found that universally conserved, functionally important components at the interface of the ribosomal small subunit and the large subunit are primordial.

It was the discoveries beginning in the early 1980s – that RNA could not only catalyse RNA replication but also direct peptide synthesis – that led to the idea of an 'RNA World' (Orgel 2004) where life forms based on RNA existed before the ability to synthesize proteins from information encoded into DNA evolved. It was the recognition that the information contained in many eukaryotic genes was not contiguous and that introns had to be removed from the mRNA derived from these genes before the mRNA could be used for protein synthesis by the ribosome that led to these discoveries. Some RNAs were even found to be capable of self-splicing. Since the early work in this area, catalytic RNAs (known as ribozymes) have been identified which form amide bonds (known as peptide bonds in a biochemical context) between RNA and an amino acid or between two amino acids (Zhang & Cech 1998).

Both RNA-catalysed aminoacyl-RNA synthesis and RNA-catalysed amino acid activation have now been shown to be possible. Aminoacyl-tRNA synthetases catalyse two essential reactions: the activation of the carbonyl groups of amino acids by forming aminoacyl-adenylates designated as *aa-AMP* and the transfer of the aminoacyl group to a specific RNA. The first can be written as

$$aa + \mathrm{ATP} \rightarrow aa - \mathrm{AMP} + \mathrm{PP}_i,$$

where ATP and AMP stand for adenosine triphosphate and adenosine monophosphate, respectively. The anion $P_2O_7^{4-}$ is abbreviated as PP_i and is formed by the hydrolysis of ATP into AMP. This reaction is followed by the transfer of the aminoacyl group to RNA

$$aa - \mathrm{AMP} + \mathrm{RNA} \rightarrow aa - \mathrm{RNA} + \mathrm{AMP}.$$

RNA-catalysed aminocyl-RNA synthesis was first achieved by Illangasekare *et al.* (1995) and amino acid activation by Kumar & Yarus (2001). The structural basis for specific tRNA aminoacylation by a small ribozyme has been given by Hong *et al.* (2008).

There is also at least one plausible prebiotic polymerization reaction that can produce peptides. It has been shown by Leman *et al.* (2004) that the volcanic gas carbonyl sulphide is capable of polymerizing amino acids to form peptides under reasonable conditions. Whether this reaction could be the basis for a complementary or an alternative scenario to that of the RNA World has yet to be determined.

This brings us back to the issue of whether life first began in the form of creatures capable of rudimentary metabolism coupled with replication or, as put forth by Dyson in his delightful little book *Origins of Life* (Dyson 1985), whether 'life began twice, with two separate kinds of creatures, one capable of metabolism without exact replication, the other kind capable of replication without metabolism'. There is also the co-evolution theory of the genetic code (Wong 2005) based on the postulate that prebiotic synthesis was an inadequate source of all 20 protein amino acids, and consequently some of them had to be derived from the co-evolving pathways of amino acid biosynthesis.

Dyson introduced a 'Toy Model' of molecular evolution along the lines of the Oparin picture of the origin of life (Oparin 1966) where proto-cells came first, followed by enzymes, and subsequently by the use of nucleic acids to store biological information. The model does not allow Darwinian selection. Its purpose was to demonstrate that a population of molecules within a proto-cell could, by random drift, achieve an organized state where active biochemical cycles might exist. Wong's theory, on the other hand, allows for a prebiotic evolution of peptide sequences as well as amino acid biosynthesis, and therefore at least partly makes use of Darwinian selection.

Without Darwinian selection, one cannot preferentially increase the number of proto-cells containing biologically useful molecules formed by random drift. Darwinian evolution can be introduced into Dyson's Toy Model by allowing the proto-cells to grow and reproduce by fission and dissolve in the absence of adequate nutrients. The precise extent and nature of prebiotic molecular evolution in the development of life is unknown, but a resolution of this question is not central to the issue at hand; being able to form proteins via the carbonyl sulphide or a similar route on the prebiotic Earth, or in an RNA World for that matter, does not in and of itself resolve the probabilistic argument given above.

Certain assumptions that may not at first be apparent are built into that argument. The first is that the evolution of life depends on selecting for the specific set of proteins that appears in life forms today. It seems very unlikely that the origin and evolution of life could be so narrowly constrained. The second is that each of the 20^{200} possible proteins is functionally unique. This is also unlikely to be the case. It is far more probable that many of these possible 20^{200} proteins can serve the same biological function. In the case of enzymes, it may have been possible for many structurally unrelated enzymes to catalyse a given reaction, albeit with possibly different reaction rates. This is certainly the case today. Such enzymes are known as 'analogous' enzymes and represent independent paths in enzyme evolution (Galperin *et al.* 1998; Gherardini *et al.* 2007). This contrasts with homologous enzymes that derive from a common ancestor. It is also probable that many of the enzymes of the RNA World were far less efficient than their modern cognates.

Another assumption implicit in the argument that the protein space to be sampled is $\sim 20^{200}$ is that an enzyme can catalyse only a single reaction. The active site of an enzyme is generally only a very small part of the protein and is usually formed by different polypeptide chains. Even the configurations of enzymes often change in the presence of the substrate molecule with which it interacts. It is quite possible that many of the 20^{200} proteins, especially the larger ones, could have several active sites catalysing different reactions. In fact, many enzymes are known to have a number of catalytic sites (Llewellyn & Spencer 2007). Taking into account these built-in assumptions could dramatically reduce the set of possible proteins that need to be sampled to produce an enzyme with a given activity.

There is another even more compelling argument that the set of proteins that were sampled during the early evolutionary process was far smaller than 20^{200}. Nineteen of the 20 amino acid residues in the secondary structures of proteins show a relatively strong tendency to form either α-helices, β-sheets or turns (see Berg *et al.* 2002), i.e., amino acids have different conformational propensities for forming these structures. Thus, from the secondary structural point of view, amino acids fall into three 'fuzzy' equivalence classes. 'Fuzzy' because, strictly speaking, an equivalence relation partitions a set into subsets where each member of the original set belongs to only one member of the partition. Here, amino acid residues only have a 'tendency' to form α-helices, β-sheets or turns.

In addition, it has been shown (Wei *et al.* 2003) that the secondary structure of proteins depends not only on the composition but also on the sequence of amino acids in the protein. Some sequence patterns produce α-helical structures while others produce β-strands.

Most naturally occurring proteins are composed of between 50 and 2000 amino acids. If the size of early proteins was towards the lower end of this range, and if the tertiary structures upon which the protein's properties depend were strongly influenced by this division into three fuzzy amino acid equivalence classes, the set of proteins which would need to be sampled could be reduced. This reduction, in conjunction with the rules that specify the secondary structure of proteins, would dramatically reduce the sequence space that needs to be sampled. While the remaining number of proteins may still be large, it is one that could well be sampled in a reasonable time if the set of proteins is sampled in a parallel manner as discussed below.

Another consideration that could reduce the set of possible proteins is the recognition that more than one-third of all enzymes contain either bound metal ions or require the addition of such ions for catalytic activity. Since 1932, when carbonic anhydrase – a biologically important zinc containing enzyme – was discovered, hundreds of enzymes have been found to contain zinc, but only in the +2 state.

Zinc atoms in such enzymes are essentially always bound to four or more ligands. Carbonic anhydrase, which catalyses carbon dioxide hydration, is particularly important in biological systems. The way this enzyme probably works is that the zinc creates a hydroxide ion from a water molecule by facilitating the release of a proton; the carbon dioxide substrate then binds to the enzyme's active site where it is positioned to react with the hydroxide ion; the hydroxide ion then converts the carbon dioxide molecule into a bicarbonate ion; and finally, the catalytic site is restored by the release of the bicarbonate ion and the binding of another water molecule.

A synthetic analogue of this carbonic anhydrase mechanism, where a simple synthetic ligand binds zinc through four nitrogen atoms – rather than the three histidine nitrogen atoms in the enzyme – accelerates the hydration of carbon dioxide by more than a factor of 100 at a pH of 9.2 (Berg *et al.* 2002).

Zinc is not the only metal ion of interest. It has also been proposed (van der Gulik *et al.* 2009) that specific short peptides 3–8 amino acids long bound to one or more positively charged metal ions such as Mg^{2+} could have served as catalysts during the period of very early life.

What this suggests is that early proto-enzymes may have contained metal complexes that were subsequently incorporated into evolving protein enzymes. If true, this too could have diminished the number of possible proteins that need be sampled – provided large subsets of the possible proteins were able to incorporate a given metal complex.

Thus, what the probability argument is really saying is that the possibility of life arising for a second time in exactly the same way it did is negligible. This may well be true, but many alternative possibilities may have existed for life to begin in the form of self-replicating systems. It is even possible that more than one type of self-replicating organism appeared and all were subject to variation, selection, and in the end, possible convergence.

The idea of an RNA World has its own difficulties. Two of the most important are obtaining, under pre-biotic conditions, nucleotides in sufficient quantity, and the lack of a known mechanism for replication of RNA molecules without the presence of a replicase. These challenges have not yet been fully resolved. However, once self-replicating RNA molecules capable of catalysing polypeptide chains exist, it would be expected that different strands of RNA would produce different polypeptides. Collectively, these effectively sample the set of proteins in a massively parallel manner. Once the minimal necessary set of ribozymes were formed, even if their reaction rates were much lower than their modern cognates, RNA would serve to carry both 'genetic' information and serve as a catalyst for the reactions needed for the first primitive self-replicating systems.

These primitive RNA World proto-cells would be subject to the usual Darwinian variation and selection process. There is, however, a strong incentive for an evolutionary transition from an RNA World to a DNA World (Lazcano *et al.* 1988). This is due to the greater genetic stability of the double-stranded helical structure of DNA compared with single-stranded RNA, and the fact that deoxyribose is chemically less reactive than ribose.

In all modern organisms, the components of DNA are synthesized from those of RNA by ribonucleotide reductases. These convert the base and phosphate groups linked to a ribose sugar that form ribonucleotides to deoxyribonucleotides composed of a base and phosphates linked to deoxyribose sugar. The fact that ribonucleotide reductases take a variety of species-dependent forms increases the probability that a primordial form of these reductases could have formed in the RNA World. Another possibility is that ribozymes were replaced by enzymes formed of amino acids before the evolutionary transition to storing information in DNA.

The transition to the DNA World is complicated because some mechanism for reading the information from the DNA would have to evolve along with the transition to the storage of information in DNA.

One approach to resolve this problem might be the following: the two strands that comprise DNA separate on heating and reform on cooling. If RNA is present during the cooling period, DNA–RNA hybrids will form where sections of the DNA and the RNA have complementary bases. Under natural selection, primitive self-reproducing proto-cells of the RNA World would have a preponderance of the type of RNA needed to support their metabolism and replication. If some of these RNAs were converted to DNAs by early forms of ribonucleotide reductases, a replication of RNAs using these DNAs as a template by temperature cycling might serve in place of the modern enzyme RNA polymerase. Once such DNA-based self-replicating systems come into existence, natural selection would rapidly end the RNA World.

It is interesting, and relevant to this discussion, that the Central Dogma has been questioned. Francis Crick, who coined the phrase in 1957, apparently views RNA editing as the most significant exception (Thieffry & Sarkar 1998). Such editing involves the modification of RNA sequences after transcription and is a commonly found in eukaryotic cells. Although editing implies that the sequence of amino acids in the resulting protein is not entirely encoded in nucleic acids, there is still no known mechanism where the information for an amino acid sequence could flow from protein to nucleic acid.

RNA editing is common in mRNA transcribed from mitochondrial DNA. It was soon discovered, however, that the sequence information needed for RNA editing was supplied by small RNAs transcribed from a second component of the mitochondrial DNA. If this always turns out to be the case, RNA editing would involve the transfer of genetic information from one RNA to another (Seiwert 1996). This would resolve the editing challenge to the Central Dogma. Other challenges to the Central Dogma, such as the discovery of reverse transcriptase, the replication of prions and epigenetic modulation of DNA, have been discussed elsewhere – for example, by Morange (2008).

To summarize, if proteins on an average are composed of 200 amino acids, one possible way of selecting the approximately 20 000 proteins coded in our genes (or the larger number found in the natural proteome) out of the 20^{200} possible is to reduce the set of possible proteins by recognizing that it might have been possible for organisms that first populated the post-RNA World to use many different subsets of the 20^{200} possible proteins; when these proteins are enzymes, more than one might have catalysed any given reaction, although possibly with different reaction rates. In addition, many proteins might have more than one active site and thus served to catalyse more than one reaction. Also, of the 20^{200} possible, many proteins might be able to serve the same biological function. The set of proteins that need to be sampled might also have been reduced if many of the early protein-based enzymes were able to incorporate the metal complexes that might have constituted earlier proto-enzymes. And, finally, if the tertiary structure of early proteins was strongly influenced by the division of amino acids into three fuzzy equivalence classes, the number of proteins that need be sampled could be closer to $\sim 3^{50}$ rather than 20^{200}. The basic point is that the evolution of proteins could not and need not have involved an unbiased random search of all possible proteins.

Once the transition to a DNA World is accomplished, the set of possible proteins would continue to be sampled by variation and selection of early DNA-based cells. These, as many single-celled organisms do today, would presumably share genes through horizontal gene transfer. Owing to this mode of information exchange, the evolutionary process of variation and selection becomes a massively parallel rather than a serial process. This is what allows colonies of micro-organisms to rapidly adapt to changing environmental conditions today.

References

Berg, J.M., Tymoczko, J.L. & Stryer, L. (2002). *Biochemistry*, 5th edn. W.H. Freeman and Co., New York, section 9.2.

Chiarabelli, C. & De Lucrezia, D. (2007). Question 3: the worlds of the prebiotic and never born proteins. *Orig. Life Evol. Bioph*. **37**, 357–361. See also the posts by 'Retread' at http://blogs.nature.com/thescepticalchymist/2008/04/chemiotics_how_many_proteins_c.html

Dyson, F. (1985). *Origins of Life*. Cambridge University Press, Cambridge.

Galperin, M.Y., Walker, D.R. & Koonin, E.V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*. **8**, 779–790.

Gherardini, P.F., Wass, M.N., Helmer-Citterich, M. & Sternberg, M.J.E. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol*. **372**, 817–845.

Harish, A. & Caetano-Anollés, G. (2012). Ribosomal history reveals origins of modern protein synthesis. *PLoS ONE* **7**(3), e32776.

Hong, X., Murakami, H., Suga, H. & Ferré-D'Amaré, A.R. (2008). Structural basis of specific tRNA amonoacylation by a small *in vitro* selected ribozyme. *Nature* **457**, 358–361.

Illangasekare, M., Sanchez, G., Nickles, T. & Yarus, M. (1995). Aminoacyl-RNA synthesis catalyzed by an RNA. *Science* **267**, 643–647.

Kumar, R.K. & Yarus, M. (2001). RNA-catalyzed amino acid activation. *Biochemistry* **40**, 6998–7004.

Lazcano, A., Guerrero, R., Margulis, L. & Oró, J. (1988). The evolutionary transition from RNA to DNA in early cells. *J. Mol. Evol*. **27**, 283–290.

Leman, L., Orgel, L. & Reza Ghadiri, M. (2004). Carbonyl sulfide-mediated prebiotic formation of peptides. *Science* **306**, 283–286.

Levinthal, C. (1969). Mossbauer spectroscopy in biological systems. In *Proceedings of a Meeting held at Allerton House*, ed. Monticello, I.L., Debrunner, P., Tsibris, J.C.M. & Munck, E., pp. 22–24. University of Illinois Press, Urbana, IL.

Llewellyn, N.M. & Spencer, J.B. (2007). Enzymes line up for assembly. *Nature* **448**, 755–756.

Miller, S.L. (1953). A production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529. A recent reanalysis of the residues of one of the original Miller experiments using modern techniques [Johnson AP *et al.* (2008) The Miller Volcanic Spark Discharge Experiment. *Science* **322**:404] showed the presence of 22 amino acids and five amines rather than the five amino acids originally found by Miller.

Miller, S.L. & Orgel, L.E. (1974). *The Origins of Life on the Earth*. Prentice-Hall Inc., Englewood Cliffs, NJ.

Morange, M. (2008). Fifty years of the Central Dogma. *J. Biosci.* **33**(2), 171–175.

Oparin, A.I. (1966). *Life: Its Nature, Origin, and Development*. Academic Press Inc., New York [Translated from the Russian by Ann Synge].

Orgel, L.E. (2004). Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* **39**, 99–123.

Seiwert, S.D. (1996). RNA editing hints of a remarkable diversity in gene expression pathways. *Science* **274**, 1636–1637.

Thieffry, D. & Sarkar, S. (1998). Forty years under the central dogma. *Trends Biochem. Sci.* **23**, 312–316.

van der Gulik, P., Massar, S., Gilis, D., Buhrman, H. & Rooman, M. (2009). The first peptides: the evolutionary transition between prebiotic amino acids and early proteins. *J. Theor. Biol.* **261**, 531–539.

Wei, Y., Kim, S., Fela, D., Baum, J. & Hecht, M.H. (2003). Solution structure of a *de novo* protein from a designed combinatorial library. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13270–13273.

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

Zhang, B., & Cech, T.R. (1998). Peptidyl-transferase ribozymes: *trans* reactions, structural characterization and ribosomal RNA-like features. *Chem. Biol.* **5**, 539–553.

Zwanzig, R., Szabo, A. & Bagchi, B. (1992). Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 20–22.