Orig Life Evol Biosph (2015) 45:163–171
DOI 10.1007/s11084-015-9414-9

ORIGINS 2014

Paradoxes of Early Stages of Evolution of Life and Biological Complexity

Alexey V. Melkikh

Received: 9 September 2014 / Accepted: 15 December 2014 /
Published online: 11 March 2015

Abstract Two of the most fundamental questions concerning the origin of life, how biologically important molecules (RNA, proteins) find their unique spatial configuration, and how coding sequences can evolve beyond a certain critical length, are discussed. It is shown that both of these problems have not been solved. Experiments that could clarify the mechanisms of interaction between biologically important molecules in the simplest cells are discussed.

Keywords Spatial configurations of replicators · Coding · Early stages of evolution · Conformational degrees of freedom

Introduction

The problem of the origin of life remains largely unsolved. Although amino acids have been obtained in such experiments as the classic Miller-Urey experiment (Miller 1953), and observations show the presence of the components necessary for life in different parts of the universe (see, for example, Pizzarello et al. 2012; Callahan et al. 2011), the mechanism of formation of the simplest living system from these components remains unclear. Many questions are unsolved, from the appearance of chiral biological molecules to the origin of the first cells. The generality of these issues for the origin of life is very different. Some of them relate to the origin of life on specific planets, while others are more general and not related to the chemical details of any particular planetary environment.

Most of the outstanding issues are related to the fact that our knowledge of the early stages of evolution is inadequate. However, there are issues related to paradoxes those which are not just a failure of our knowledge, but due to the inconsistency of that knowledge. Without addressing these issues it is difficult to speak about our understanding of life in general and in particular of its origin. This article discusses two of these issues:

Paper presented at ORIGINS 2014, Nara Japan, July 6–11 2014.

A. V. Melkikh (✉)
Ural Federal University, Yekaterinburg, RussiaMira str. 19
e-mail: melkikh2008@rambler.ru

– How do complex macromolecules (replicators) function effectively, bearing in mind that the total number of possible spatial configurations of these molecules is exponentially large?
– How does selection operate on long coding sequences, given the total number of states of such a sequence is exponentially large?

The Problem of Folding and Stability of Replicators

Our confidence that we understand the mechanisms of biochemical reactions is based mainly on the fact that for relatively simple reactions it is possible to carry out theoretical calculations from the first principles of quantum mechanics and to compare these results with experimental data. However, for complex molecules (substrates and catalysts), the situation changes radically. The fact is that for complicated molecules such as proteins, the total number of folded states of the molecule, as well as the total number of possible reactions of such a molecule with all of the other molecules, can become exponentially large (bearing in mind that many macromolecules actually fold into relatively few possible states which are kinetically or thermodynamically stable). How can the potential structural diversity of a molecular machine, in this case a protein, be canalized into a few (one) variants that are capable of performing specific functions? Why then are macromolecules in the cell (or in protocells) not uselessly entangled with each other (because, for example, the same forces act between amino acids of one or different proteins)?

During the earliest stages of evolution, biologically important molecules did not have many degrees of freedom (for example, the most complex molecules detected in space contain only about 10 atoms (see, for example, Belloche et al. 2014), but with increasing polymer length, the number of conformational degrees of freedom grows rapidly. In the process of the emergence of sufficiently long molecules, the number of conformations (as well as of the possible reactions with other molecules) could in principle increase so rapidly that to achieve a certain conformation by enumeration of variants becomes impossible.

Interactions between biological molecules are determined by the interaction potentials between the constituent atoms of the molecules. Most known potentials of interaction are short-range, i.e., these potentials are large only for the nearest atoms. We show that short-range potentials do not allow for most biochemical reactions to be effectively performed (including the folding of macromolecules).

The problem of protein folding (Levinthal's paradox) has been well discussed in the literature (see, e.g., Zwanzig et al. 1992; Onuchic et al. 1997; Berezovsky and Trifonov 2002; Finkelstein and Ptitsyn 2002; Ben-Naim 2012; Grosberg and Khokhlov 2010). It is broadly accepted that resolution of this paradox requires the existence of a funnel-like landscape of free energy.

We estimate, first, at what chain length the problem of enumeration of variants does not arise. The total number of conformationsl states of a protein chain can be estimated (see, e.g., Berezovsky and Trifonov 2002):

$$3^{N_1},$$

where

N_I is the number of protein domains.

Here it is assumed that each protein domain has three different conformations (this is typical value for proteins). If we take the largest possible population of macromolecules (proteins) to

be 10^{50} (mass of this value of molecules is greater than mass of Earth), we obtain:

$$3^{N_1} = 10^{50},$$

whence we get that $N_1 \approx 10^2$. Considering, moreover, that each domain in the protein contains several (25–500) amino acids, we obtain as a rough estimate for total number of amino acids:

$$N \approx 10^3.$$

Longer information chains could not find their native conformation by enumeration of variants over the lifetime of the biosphere. If instead $N << 10^3$, such molecules could in principle find their native conformation through the simple enumeration of variants. The result, however, depends on what restrictions are imposed at the time of folding. For example, if the total time available for state exploration is small (~1 s), as for intracellular processes, then N will also be small.

I now show that short-range potentials do not allow for the realization of funnel-like landscapes, but instead, the energy landscape must be rugged.

First, during the folding process, macromolecular states that are equivalent in energy will always arise and will be implemented with equal probability. For example, for a one- or two-component molecule the existence of such states (a fork during folding) is obvious (see Fig. 1) (see, also Melkikh 2013).

This means that the energy landscape will be fragmented due to the existence of such configurations, and the probability of the folding process converging into a single state will be small. Consider a non-periodic chain with n components, each of which has three possible conformations. For this case, we can write the probability of erroneous folding P_E in one step, connected with the presence of a "fork" in the form:

$$P_E = 1 - \left(1 - \frac{1}{n}\right)^2.$$

For $n >> 1$ (for example, 20 amino acids) the probability takes the form:

$$P_E \approx \frac{2}{n}.$$

Then, the probability of error-free folding for a chain of length N can be estimated by the formula:

$$P = \left(1 - \frac{1}{n}\right)^{2N}.$$

For example, for n=20 and N=100, we obtain $P = 3.5 \times 10^{-5}$. The dependence of $P(N)$ on N at n=20 is represented in Fig. 2.

As seen from Fig. 2 the probability of error-free folding is small even for small N. Note here that the value of n cannot be too large because in general, given a limited energy range of interactions between monomers Δ, the total number of different energy states cannot be greater than Δ/kT (energy states differing by less than kT will behave identically). The value of Δ, as well, cannot be large (because we consider forces that form the quaternary structure of proteins) and is less than 1 eV (40 kT at room temperature).

Thus, due to the presence of energy-degenerate (but spatially different) states, a funnel-like landscape is not formed and the probability of arrival of a molecule in its native conformation is low. With respect to replication, this means that the process itself is unstable. As a result, inaccurate copies will be created, or these sequences will not be able to create copies at all.

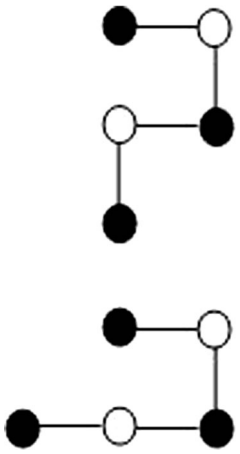

Fig. 1 Equivalent in energy but spatially different states of the molecule (different folding conformations of the same sequence)

There is another reason, which complicated the origin of replicators and protocells. Molecules present at a relatively high concentration (comparable with concentrations in modern cells) within vesicles (protocells) will in many cases interact with each other to form molecular complexes.

The formation of complexes can be both useful and harmful for protocells. On one hand, these complexes can serve to transfer energy (or information) within the protocell by the key-lock principle (for example, active transport of ions and ATP). On the other hand, erroneously arising complexes (which cannot perform useful work) will prevent the normal functioning of the protocell.

Consider the interaction of two sufficiently long (see calculations above) folded molecules. Their interaction does not differ fundamentally from their folding, as the same forces act between the monomers. Although repulsion between the individual monomers is possible, the forces of attraction between the monomers should prevail (otherwise the folding of proteins would be impossible). Then, during the interaction of the two molecules, the complex will be located in one of the metastable states. Because the average energy of the interaction between the monomers is much larger than kT (otherwise it would be impossible for the folded protein to be stable), then the lifetimes of such metastable states should be relatively long in comparison with the characteristic times of intracellular processes (~1 s).

On the other hand, the complex formed can be considered as a single folded molecule. As shown above, in such a complex system, the energy landscape is rugged. All this will lead to the fact that protocells are likely to form molecular complexes that are not able to perform useful work. In other words, the state when the "key" corresponds to its "lock" can be compared with the native configuration of the protein. Because of the large number of forks, the probability of error-free implementation of this process will be small.

Of course, we know that modern cells operate stably, despite the presence of large amounts of protein, RNA and DNA, which may interact with each other. However, the mechanism of such stability remains unclear.

To realize a funnel-like energy landscape in such a system, specific additional requirements on the interaction potential between the molecules (atoms) must be imposed:

– First, this potential should include interaction with far-enough neighbors. Only in this case (at least in principle) does it become possible to significantly reduce the number of forks (and therefore the probability of error).

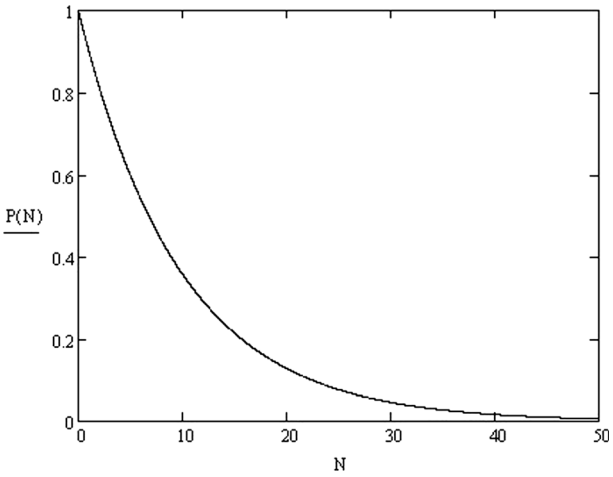

Fig. 2 The dependence of the probability of error-free folding on chain length

– Second, this potential should occur only for certain configurations of the atoms, as the interaction potentials in condensed matter of arbitrary nature have been well studied (for example, Coulomb, Lennard-Jones and Morse potentials), and no additional forces have been found. In other words, the potential must be a collective potential.

However, none of the currently known potentials satisfy these requirements. For example, the Coulomb potential, although it is relatively long-range, is not selective. Lennard-Jones and Morse potentials are short, i.e. at distances equal to about two sizes of atoms their values are already small enough (comparable to kT). Chemical (including hydrogen) bonds are always short-range.

The above observations are confirmed, for example, by the facts that the problem of folding and the problem of protein-protein interaction for long enough proteins have not yet been solved (it is impossible to calculate their structures from first principles) (see, for example, Gruebelle 2010). The current understanding of the interactions of proteins beyond a certain length is primarily only qualitative (see, for example, Lua et al. 2014).

Thus, the existence of a set of spatial structures of biologically important molecules, as well as the many variants of chemical reactions between them, is one of the most important obstacles to the emergence of the simplest living systems. There must be some special mechanism that significantly limits the range of possible variants for such a system. Without this mechanism, we cannot speak about any significant efficiency of molecular machines in the protocell. This means that such protocells must be composed of various types of molecular complexes with different spatial configurations, and these complexes will not be able to perform useful work. As cells become more complex over evolutionary time, and the number and length of biologically important molecules increases, this problem will only get worse.

Origin of the Molecular Code and the Problem of Enumeration of Variants

In the early stages of the evolution of life, biomolecules were involved in various reactions, including self-reproduction. At this stage, the problem of enumeration of variants was not important because molecules arising as a result of reactions could not perform certain types of

work. However, some of the molecules further *encoded* other molecules. This is a crucial step, which creates a problem of enumeration of variants.

It is well known that as the length of informational sequence grows, as the number of variants grows exponentially. We show that the problem of enumeration of variants is not obvious, and requires additional assumptions.

Consider a chain of nucleotides of length N. There are only 4^N variants of such sequences. How large is this number? For example, for $N=1000$ we receive $4^{1000} \approx 10^{602}$.

This number of variants is so great that it could not be enumerated for the lifetime of the universe by all organisms (replicators) that have ever lived in it. However, $N=1000$ corresponds to about only one modern gene. Hence it can be concluded that for the information sequences of length approximately equal to 10^3 and more, evolution occurred in some other way. The main questions are: what is this method and what are the conditions under which it operates?

It has been proposed (see, for example, Bailey and Eichler 2006; Long et al. 2003), that further evolution occurred by molecular exaptation, i.e., by using existing information sequences, blocks, etc. However, this mechanism does not work by itself, it implies the existence of a priori information in the system. Indeed, if the system has no a priori information about what exactly will encode a set of characters, there is no way of knowing about it, but to synthesize the molecular machines using this set and checking whether such an organism will survive or not. This is a simple enumeration of variants. If a priori information is available, it must have some material carrier, i.e., be recorded in some (yet unknown) intracellular structures.

Another argument is related to the general algorithms proposed for solving various classes of search problems. If, for the above situation (i.e., without a priori information), there were an algorithm that allowed for the problem to be solved in a polynomial (relatively small) number of steps, then it, along with all other NP-hard (requiring an exponential number of steps) problems would have been resolved. However, the reduction of NP-algorithms (non-polynomial) to P-algorithms (polynomial) is an unsolved problem now.

Imagine any change in the coding sequence for an efficient molecular machine. What is the probability that this new sequence will still encode an efficient molecular machine? To calculate such probability is of fundamental importance for whether there is a priori information about this machine or not. Consider the information sequence

$$01000101010101000011111,$$

which encodes some molecular machine

$$abcddbcaadcbdcabcdcbadcbadcadb.$$

The problem of obtaining another "good" (able to work effectively) machine in the absence of a priori information about it is fully equivalent to the problem of cracking a password. Indeed, in both cases there is an exhaustive search of some code, and when the correct option is found, some useful action performed. In one case, this is useful work of the synthesized machine, in another—access to any useful information. We emphasize that the information about how close the variant-enumerating system is from the goal, is not available—exhaustive search ends only when a certain (unknown) sequence is received.

However, as is well known (Mao 2003), the task of cracking a password is *NP*-hard, i.e., its solution requires an exponentially large number of steps. For four nucleotides, and a sequence of length N, the number is 4^N.

If a priori information about the target information sequence is (at least partly) available (and stored somewhere), it is possible in accordance with this information to use versions of the code, which has already been found earlier (for example, for other tasks), only slightly changing them. In the limit of complete information, the encoding process of new molecular machines becomes completely deterministic because it is known in advance what to create. In the problem of password cracking, a priori information also plays a key role.

Of course, not all of the amino acids in proteins are equivalent. Some of them form active sites and are crucial for protein function. This disparity, of course, can be typical for arbitrary systems, using information. For example, some of the elements of the code may not carry useful information, but only play a role in correcting errors. However, this property is not essential for the problem of enumeration of variants, for which the main concern is the actual length of the sequence. For example, if neutrality is approximately 20 %, this does not fundamentally change the exponential dependence of the number of variants on the length of the sequence.

A fundamental question is thus raised: is it possible to speed up the search process in the absence of a priori information through any other method (molecular exaptation, horizontal gene transfer, alternative splicing, which emerged later etc.)? The answer to this question is negative: either change in the information sequence occurs randomly, or there must be a system that somehow selects nucleotides and performs certain operations with them. In the second case, the presence of a priori information in the system is a necessary condition for the operation of any such system.

For example, in horizontal gene transfer, which occurs in bacteria, upon receipt of any portion of the genome of a bacterium from the outside, bacteria should a priori "know" where such a piece can be built in and under what circumstances it can be useful. If such knowledge is absent, then it does not matter what the source is of such a piece, be it another species or some random process.

Thus, the paradox of enumeration of variants in coding systems is as follows: of course, the molecular mechanisms of exaptation, block coding, horizontal gene transfer, etc. work (and could exist, apparently, during the early stages of the evolution), but for their operation, a priori information (i.e., existed before the synthesis of molecular machines) is indispensable. The modern theory of evolution does not imply that any such information should exist. If we assume that such information exists, then the question of where it is stored arises, as any information needs to have a physical carrier. On the other hand, information about the changes of genes cannot be stored in the genes themselves. The question of the mechanisms of storage and processing of such information remains open.

Discussion. Possible Directions of Solving Paradoxes

We emphasize that without solutions to these proposed paradoxes, known forms of life could have not arisen and would not function at all. Indeed, in this case, the reactions between biologically important molecules would lead with overwhelming probability to the formation of "parasitic" structures, the effectiveness (the ability to perform useful work) of which would be vanishingly small. On the other hand, the replication of such a system would not lead to the emergence of new types of replicators, i.e., their evolution would be impossible.

Both considered paradoxes give a rough estimate of approximately 10^3 nucleotides, above which the problems of complex systems arise. This number of nucleotides corresponds to approximately one gene. This coincidence does not seem accidental. It is likely that molecular

machines and coding systems cannot be too simple (i.e., they cannot contain only a few monomers) to perform their functions (compare with self-reproducting automata, von Neumann and Burks 1966). Note that the number of nucleotides of the order of 10^3 is much smaller than that of the simplest of modern cells.

These formulated paradoxes are an obstacle to the emergence of life in any form (including alternative forms of life based on other chemical elements, such as silicon, arsenic, etc.). These are general restrictions, relating to all of the known elements of the periodic table.

As for non-living molecular systems, the proposed paradoxes are not typical. The very existence of such paradoxes can be used as the basis for the definition of life. We can give the following definition of life:

> Life is a non-equilibrium self-reproducing system of such great complexity (number of variants) that simple enumeration of variants cannot be realized in it.

We also note that both formulated paradoxes may have a unified solution. Indeed, because any information can exist only if there is a physical carrier for it, then in both cases, the question arises, what forces act between biologically important molecules? This issue has been discussed previously as well (Melkikh 2013, 2014a, b), and it was concluded that quantum mechanics should play an important role in these processes.

What experiments could shed light on the mechanisms of replication and molecular machines operating in the simplest cells? Such experiments should include much more detailed measurement of intracellular molecular processes than is usually done. On the one hand, the experiment should include detailed (at the level of individual nucleotides or atoms) measurements (e.g., using X-ray or neutron diffraction) of evolving systems' status. On the other hand, a detailed registration of reactions between biologically important molecules (in particular—the process of folding) is needed. A number of such experiments were proposed earlier (Melkikh 2013, 2014a, b).

References

Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity, and disease. Nat Rev Genet 7:552–564

Belloche A, Garrod RT, Muller HSP, Menten KM (2014) Detection of a branched alkyl molecule in the interstellar medium: iso-propyl cyanide. Science 345(6204):1584–1587

Ben-Naim A (2012) Levinthal's question revisited, and answered. J Biomol Struct Dyn 30(1):113–124

Berezovsky IN, Trifonov EN (2002) Loop fold structure of proteins: resolution of Levinthal's paradox. J Biomol Struct Dyn 20(1):5–6

Callahan MP, Smith KE, Cleaves HJ II, Ruzicka J, Stern JC, Glavin DP, House CH, Dworkin JP (2011) Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. PNAS 108(34):13995–13998

Finkelstein AV, Ptitsyn OB (2002) Protein physics. Academic, Oxford

Grosberg AY, Khokhlov AR (2010) Giant molecules: here, there, and everywhere, 2nd edn. World Scientific Publishing Company, London

Gruebelle M (2010) Weighing up protein folding. Nature 468:640–641

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4:865–875

Lua RC, Marciano DC, Katsonis P, Adikesavan AK, Wilkins AD, Lichtarge O (2014) Prediction and redesign of protein-protein interactions. Prog Biophys Mol Biol 116(2–3):194–202

Mao W (2003) Modern cryptography: theory and practice. Prentice Hall, Professional Technical Reference, Upper Saddle River

Melkikh AV (2013) Biological complexity, quantum coherent states and the problem of efficient transmission of information inside a cell. BioSystems 111:190–198

Melkikh AV (2014a) Quantum information and the problem of mechanisms of biological evolution. BioSystems 115:33–45
Melkikh AV (2014b) Congenital programs of the behavior and nontrivial quantum effects in the neurons work. BioSystems 119:10–19
Miller SL (1953) Production of amino acids under possible primitive earth conditions. Science 117(3046):528–529
Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48:545–600
Pizzarello S, Schrader DL, Monroe AA, Lauretta DS (2012) Large enantiomeric excesses in primitive meteorites and the diverse effects of water in cosmochemical evolution. PNAS 109(30):11949–11954
von Neumann J, Burks AW (1966) Theory of self-reproducting automata. University of Illinois Press, Champaign
Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. PNAS 89:20–22