Noisy Clues to the Origin of Life

David C. Krakauer
Akira Sasaki

SANTA FE INSTITUTE

Noisy clues to the origin of life

David C. Krakauer (1) & Akira Sasaki (2)

(1) SANTA FE INSTITUTE, HYDE PARK ROAD, SANTA FE, NM 87501 USA.
KRAKAUER@SANTAFE.EDU

(2) INSTITUTE FOR ADVANCED STUDY
PRINCETON, NJ 08540, USA

Abstract

The origin of stable self-replicating molecules represents a fundamental obstacle to the origin of life. The low fidelity of primordial replicators places restrictions on the quantity of information encoded in a primitive nucleic acid alphabet. Further difficulties for the origin of life are the role of drift in small primordial populations, reducing the rate of fixation of superior replicators, and the hostile conditions increasing developmental noise. Thus mutation, noise and drift are three independent stochastic effects that are assumed to make the evolution of life improbable. Here we show to the contrary how noise present in hostile early environments, can increase the probability of faithful replication, by amplifying selection in finite populations. Noise has negative consequences in infinite populations, whereas in finite populations, we observe a synergistic interaction among noise sources. Hence two factors formerly considered inimical to the origin of life - developmental noise and drift in small populations - can in combination give rise to conditions favorable to robust replication.

The origin of life on earth remains a fascinating problem for evolutionary theory as it has required a combination of geophysical, chemical and biological insights. The geophysical insights have helped to explain the emergence of a reducing atmosphere considered to be essential for the production of hydrocarbon molecules and the processes by which these hydrocarbons come into existence [31] [29] [27] [26]. Chemists have provided scenarios for the production of amino acids and nucleic acids from these simple molecules and gases likely to be present in the early atmosphere [14] [23] [20] [5]. Biological theories take these constituents and explore their application in the origins of replication and metabolism [8] [12] [7].

Replication oriented theories for the origin of life suggest that simple RNA-like polymers arose first followed by cells and proteins [8] [25] [19]. Metabolism oriented theories [23] [3] [17] [15] suggest that cells and proteins came first, followed much later by nucleic acid replication. It is important to stress that some form of metabolism is a prerequisite for replication and hence these two approaches reflect only differences in emphasis. The primary data in support of metabolic/protein first theories are the the relative ease of production of amino acids under primordial conditions [13] in contrast to the difficulty of producing nucleic acids [6], and the high levels of noise that can be tolerated by auto-catalytic networks [7]. In support of replication theories are the facts that all contemporary proteins are first encoded as nucleic acids, that RNA molecules possess basic catalytic properties [4] [24] [28] [2], and that modern cells have the appearance of compartmented catalytic hypercycles [11].

In both replication-oriented theories and metabolism-oriented theories, different sources of noise play a very significant role. Noise comes in at least three flavors : (1) high levels of heritable noise or mutation, (2) sampling noise or drift, and (3) developmental noise or thermal flux in the phenotype. In replication-oriented theories, errors that arise through mutation during replication are thought to give rise to an error-threshold [10] [9], limiting the absolute size of a polymer. Whereas in existing organisms the magnitude of these mutations are relatively small, for early replicators, they are likely to have been very high. During the sampling of future generations in finite populations, the fittest genotypes can be lost by chance, compounding the effects of mutation [16]. Drift in effect reduces selective differences among competing replicators and thereby reduces heritability. The effects of

2

non-heritable flux or developmental noise, have been neglected in replication oriented theories. In metabolism-oriented theories, flux limits the specificity of catalysis, whereas random sampling drift plays a constructive role. This is because drift can lead to the selection of ordered catalytic networks, from disordered networks, by surmounting free energy barriers [7].

Here we provide an explicit treatment of all three sources of noise: mutation, drift and developmental noise. We decompose developmental noise into two independent processes or submodels: knockout noise and mutation amplification. We show that non-heritable flux in selection coefficients (developmental noise) when coupled with drift in finite populations, increases the tolerance of replicators to high rates of heritable mutation. Developmental noise in small populations allows more information to be preserved in larger polymers under higher rates of mutation.

For expositional reasons we first describe an infinite population model and subsequently present simulation results and a diffusion approximation for a finite population rendition. Consider a short ribonucleotide polymer of length L. We denote the frequency of a polymer sequence containing i wildtype and $L-i$ mutant nucleotides by p_i. The sequence in which the number of mutant sites $i = 0$ is defined as the wildtype. The fitness of an individual carrying i healthy nucleotides is denoted by W_i, and is assumed to be a monotone non-decreasing function of i ($W_0 \leq W_1 \leq \cdots \leq W_L = 1$). Thus all templates in the ith Hamming class have equal fitness. The mutation rate creating a new defective template is denoted by μ per site. We ignore back mutations restoring the wildtype sequence.

The genotype frequency dynamics for a continuous time change, after the transformation from W_i in the discrete generation model to the Malthusian parameter $w_i = \ln W_i$ in continuous time, is

$$\dot{p}_i = (w_i - \bar{w})p_i - i\mu p_i + (i + 1)\mu p_{i+1}, \qquad (i = 0, 1, \ldots, L) \tag{1}$$

where $\bar{w} \equiv \sum_{i=0}^{L} w_i p_i$ is the population mean fitness (Malthusian parameter), and $p_{L+1} \equiv 0$. By definition, $\sum_{i=0}^{L} p_i = 1$.

There are $L+1$ different equilibria of the system (1), corresponding to solutions differing in the maximum number n of healthy sites found in the population. Let $\bar{w}(n)$ be the mean fitness at equilibrium, defined as

$$\bar{w}(n) = w_n - n\mu, \tag{2}$$

which is derived from $\dot{p}_n = 0$. It can be proven that the equilibrium is locally stable if and only if n, the largest number of healthy sites in the distribution of $\{\hat{p}_i\}$, maximizes $\bar{w}(n) = w_n - n\mu$. In other words, $\bar{w}(n) = w_n - n\mu$ is maximized through evolution acting on the number of wildtype sites in the template. This result can be regarded as a variation of the Haldane-Muller principle: *"the mutation load, the reduction in mean fitness from its maximum, equals the total mutation rate"*. However, the maximum fitness (Malthusian parameter) in this case refers not to the possible maximum, $w_L = \ln W_L = 0$, but the maximum among genotypes maintained in the population, $w_n = \max\{w_i \,|\, \hat{p}_i > 0\}$. The final equilibrium distribution for the number of functional sites depends on the fitness landscape $\{w_i\}$. Here we focus on the linear fitness landscape in continuous time and the corresponding multiplicative landscape in the discrete time, finite population simulations (footnote 1).

Developmental noise can have two possible effects on a replicator. It can either (1) appear to introduce new non-heritable errors into a sequence (knockout noise), or (2) amplify the deleterious effects of existing heritable mutations (mutation amplification). The first model assumes that environmental noise and mutation operate in essentially the same fashion on a sequence. Both lead to the effective replacement of a nucleotide. In the case of noise this replacement is not propagated. This mechanism can be thought of as intervening in the genotype to phenotype map so as to give the impression of a greater number of mutations. For example, in advanced two-step replicators, in which RNA is first transcribed into DNA or DNA into RNA, noise can lead to misincorporation errors without changing the primary replicator. The second mechanism amplifies the deleterious effects of existing mutations either by reducing the ability of a replicator to fold into a desired conformation, or by increasing the structural instability of the mutant phenotype. Thus those replicators with no mutations are left unaffected by noise, whereas those with many mutations, suffer a disproportionately large amount of change to their phenotypes.

Consider the first kind of noise, described as knockout noise. The probability that a non-heritable loss of function takes place at each site is denoted by q. In the genotype with i healthy gene copies, the probability that j genes remain functional after noise is,

$$\binom{i}{j}(1-q)^j q^{i-j}, \qquad (j = 0, 1, \ldots, i). \tag{3}$$

As before, we denote the fitness (Malthusian parameter) of an individual having j functional sites by w_j. The average fitness of the genotype with i healthy sites in the genome is then given by

$$\sigma_i = \sum_{j=0}^{i} w_j \binom{i}{j}(1-q)^j q^{i-j}. \tag{4}$$

With the second kind of noise we envisage an amplification parameter, ω which determines the magnitude of thermal fluctuations. These fluctuations are thought of as small deviations in fitness drawn from a uniform probability distribution ξ, hence

$$\sigma_i = < w_i - \Delta w_i \, \omega \xi > \tag{5}$$

where, $\Delta w_i = w_L - w_i$ is the fitness reduction of genotype i from the wild type, and this effect of defective sites is to be amplified with the magnitude $\omega \xi$. Note that because of the monotonicity in $\{w_i\}$, $\sigma_i < w_i$ holds for all i with both models.

To include the effects of noise, the genotypic frequency dynamics remain the same as (1), except that w_i is now replaced by the appropriate σ_i:

$$\dot{p}_i = (\sigma_i - \bar{\sigma})p_i - i\mu p_i + (i+1)\mu p_{i+1}, \qquad (i = 0, 1, \ldots, L) \tag{6}$$

where $\bar{\sigma} = \sum_{i=0}^{L} \sigma_i p_i$ is the mean fitness, and $p_{L+1} \equiv 0$ as before. Therefore, the mean fitness of the population $\bar{\sigma}$ at the stable equilibrium satisfies

$$\bar{\sigma} = \sigma_n - n\mu = \max_{0 \le i \le L} \{\sigma_i - i\mu\}. \tag{7}$$

4

and the stable frequency distribution is as before with w_i's replaced by σ_i's.

In figure 1 we plot fitness for both submodels of noise against the number of mutations to the wildtype assuming a linear fitness landscape in continuous time. For knockout noise with the maximum number of deleterious mutations (1b), there is no difference between the zero noise model (w_i) and the noisy process (σ_i). This is because there can be no additional sites damaged through noise. As the number of wildtype sites increase, the difference between w_i and σ_i increases. For mutation amplification noise, the opposite is true. When there are no mutations to the wildtype, w_i and σ_i are identical. As the number of heritable mutations increases, so does the difference between w_i and σ_i. With amplification, the greatest difference in fitness between the noisy and noise-free dynamics, is produced when the mutational load is at its maximum. Significantly, knock-out noise reduces the fitness differences between mutants in different Hamming-classes, whereas amplification noise, increases these differences. Thus knockout noise in a linear landscape does not provide any advantage over zero noise. Amplification noise, as we shall see, can provide an advantage.

Early replicators will have emerged in small populations. We are therefore required to consider the effects of drift in finite populations [16]. We have performed extensive Monte Carlo simulations in discrete time, assuming multiplicative fitness landscapes, corresponding to a linear landscape in continuous time for which we have developed a diffusion approximation (footnote 2). In figure 2 we plot the population mean fitness and number of wildtype sites at quasi-stationary state, against the effective selection coefficient to include the influence of noise. We observe that for highly mutable templates in large populations, increasing noise has no effect on the population mean fitness. This is a consequence of the Haldane-Muller principle. At very small levels of noise, corresponding to small selection coefficients ($s << 1$), or large mutational rates (μL), heritability is lost and all sites experience mutation. For very large selection coefficients or small mutations rates, approximately all sites are kept free from mutation and the wildtype template dominates. In large (infinite) populations therefore, the number of wildtype sites increase with an increase in the magnitude of noise, without an increase in fitness. One way of thinking about this, is that increasing noise, selects for increasing levels of genetic redundancy [22] [18].

In small sized populations the Haldane Muller principle is no longer as strong, and changing the magnitude of noise, has an influence on the population mean fitness. For very low levels of noise, once again, all wildtype sites are lost through a combination of mutation and drift. Increasing the magnitude of noise in the region of these low values, reduces the mean population fitness. As the noise level increases still further, variation in mutation rate becomes detectable by selection, thereby leading to an increase in mean fitness. This is because noise enhances the phenotypic expression of mutations in contaminated genomes, leading to their removal from the population. This can be thought of as a negative 'Baldwin effect', in which noise explores deleterious fitness space, rather than the usual beneficial fitness space [1]. At the highest noise levels, there is a significant increase in mean fitness, corresponding to the incremental increase in the number of healthy sites preserved in the quasi-stationary state. Thus fitness is maximized at high and low levels of noise and minimized at intermediate noise levels. Genetic information in contrast, is present only at the highest noise levels. This bistability in fitness and heritability is often ignored in

evolutionary theory, as very large selection coefficients and/or rates of mutation are deemed unrealistic. However, this criticism is less appropriate when considering the several sources of noise in primordial environments. In this scenario, as our noise models illustrate, high rates of mutation and large increases in the effective selection coefficient are quite likely. The most striking difference between the finite population description and the infinite population model, is therefore, the way in which increasing the number of healthy sites through an increase in noise, leads to a graduated increase in mean population fitness (Figure 2). Not only does noise effectively purge mutations from the genome, it increases population mean fitness.

An alternative way of visualizing this process is by plotting the mean population fitness and mean number of wildtype sites, against the mutation rate (μL). In infinite populations, increasing the mutation rate, leads to a gradual decrease in fitness and in the number of wildtype sites. At a sufficiently large rate, all genetic information will be lost. In finite populations, the drop in fitness occurs far more rapidly, with an abrupt loss of wildtype sites and a corresponding rapid drop in fitness. As the magnitude of noise is increased, higher mutation rates can be tolerated, preserving a larger number of wildtype sites. Moreover, the loss of sites and of fitness becomes gradual, more closely approximating the infinite population case. Thus environmental noise in finite populations increases the robustness of replicators to mutation, and has an effect comparable to increasing the effective population size. In other words one source of developmental noise -mutation amplification - has the effect of mitigating a second source of noise - drift. We observe that far from making the evolution of life less probable, stochastic effects by acting in concert can make the evolution of robust replicators more likely.
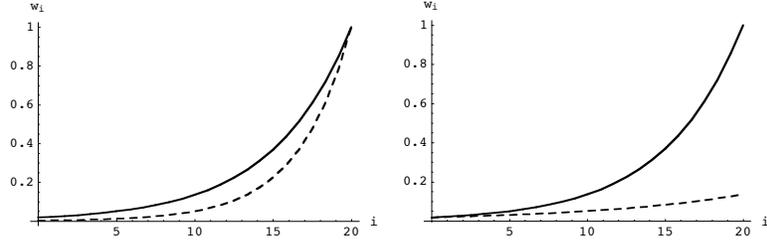
Figure 1: Genotypic fitness modulation assuming the noise amplification model (a), and the knockout noise model (b). Fitness of a replicator is multiplicative, $W_i = (1-s)^{L-i}$, and decreases with the number of healthy sites i (the solid curves in both panels; $s = 0.2$ and the total number of sites $L = 20$). With amplifying noise (a), the deleterious effect of each defective site is increased by $\Delta s = 0.1$, leading to a steeper decline in fitness as the number of defective sites $(L - i)$ increases (dashed curve). With knockout noise, by contrast, each healthy site is lost with a probability q, giving rise to a flatter landscape (dashed curve; $q = 0.5$). The relative difference between the fit nesses with and without noise is greatest at $i = 0$ in the noise-amplification model (a), and at $i = L$ in the knockout-noise model.
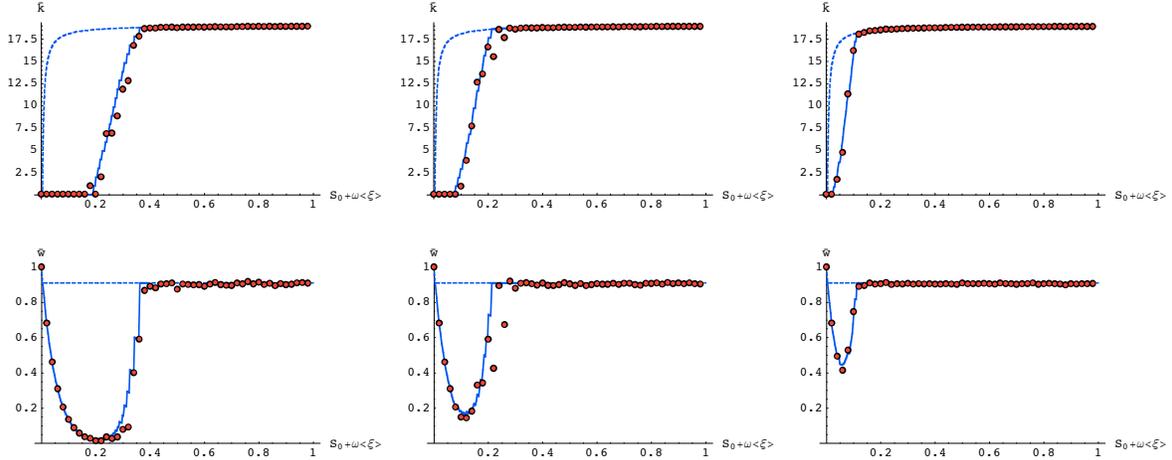
Figure 2: The effect of non-heritable amplifying noise on the mean number of healthy sites maintained in a finite population, and on the population mean fitness. Upper panels illustrate how the mean number of healthy sites, $\bar{k} = \sum_{i=0}^{L} i\hat{p}_i$, averaged over the quasi-stationary distribution $\hat{p}_i$ at generation $T = 30,000$, changes with the effective selection coefficient $\langle S_\xi \rangle = s_0 + \omega \langle \xi \rangle$ with different magnitudes ω of noise ξ. Lower panels illustrate the noise dependency of the corresponding population mean fitness $\bar{W} = \sum_{i=0}^{L} W_i \hat{p}_i$. The initial population is assumed to consist exclusively of wild type replicators (with all sites healthy). The population size is $N = 20$ in a), $N = 50$ in b), and $N = 200$ in c). The dots in each panel show the results of Monte Carlo simulations assuming a multiplicative fitness scheme: $W_i = (1 - S_\xi)^{L-i}$, where i is the number of healthy sites in a template of L sites. The red and blue lines show the expected values obtained from the diffusion approximation (footnote 2), and the infinite population results. Both assume continuous time linear fitness landscape: $w_i = -S'_\xi(L - i)$ corresponding to the multiplicative landscape in the non-overlapping generation Monte Carlo simulations (with $W_i = \exp(w_i)$ and $S'_\xi = -\log(1 - S_\xi)$). In all panels, the the effective selection coefficients (the horizontal axes) are varied by changing the magnitude of noise ω. The total number of sites $L = 19$ and the per-site mutation rate $\mu = 0.005$ are fixed in all panels. Variation in these parameters to include longer templates with suitable scaling do not alter the results.
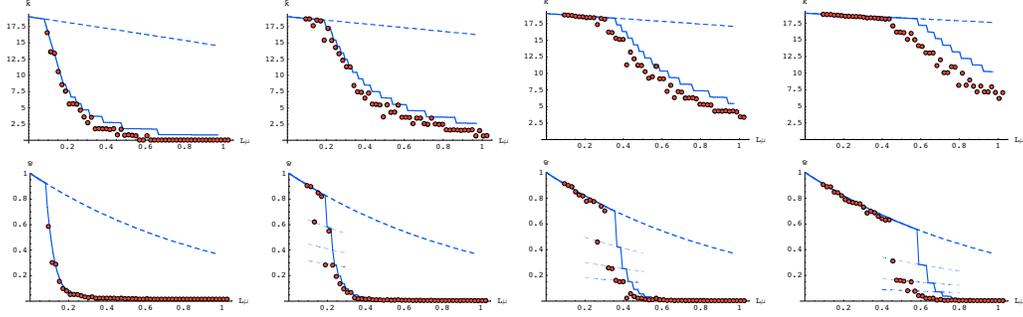
Figure 3: The mean number of healthy sites maintained in a finite population (upper panels), and the population mean fitness (lower panels) plotted against the total mutation rate $L\mu$. The population has a fixed size $N = 50$, and the effective selection coefficients are varied from left to right as: a) $\langle S_\xi \rangle = 0.2$, b) $\langle S_\xi \rangle = 0.3$, c) $\langle S_\xi \rangle = 0.4$, and d) $\langle S_\xi \rangle = 0.5$. As in Fig. 2, dots indicate the results of Monte Carlo simulations; solid curves represent the expectations from a one-dimensional diffusion approximation for the waiting times; dashed curves present infinite population results. Short dashed segments in the lower panels of b)-d) indicate the mean fitness under the infinite population model when the maximum numbers of healthy sites L are 18, 17, and 16 (from top to bottom), rather than the $L = 19$ assumed initially.

Footnote

1. Infinite population model. The expected $\bar{k}$ and $\bar{w}$ in the limit of $N \to \infty$ are easily calculated. If the effective selection coefficient $\langle S_\xi \rangle$ is greater than the threshold ($\langle S_\xi \rangle > \mu$), the stationary distribution for the number of healthy sites under the linear fitness landscape is binomial: $\hat{p}_i = \binom{L}{i}(1 - \mu/s)^i(\mu/s)^{n-i}$, with $s = \langle S_\xi \rangle$. Whereas, if the selection is weaker than the threshold ($\langle S_\xi \rangle < \mu$), the population is monomorphic with all-defective replicators. The mean number of healthy sites ($\bar{k} = L(1 - \mu/s)$)) increases with $s = \langle S_\xi \rangle$ in the region $\langle S_\xi \rangle > \mu$. The population mean fitness, on the other hand, is constant $\bar{w} = -L\mu$ in the region $\langle S_\xi \rangle > \mu$ (the Haldane-Muller principle). In order to compare results of a non-overlapping generation, discrete time, Monte Carlo simulation, with continuous time results, we have to scale our parameters appropriately. If we assume a selection coefficient S_D in a multiplicative fitness scheme ($w_i = (1 - S_D)^{(L-i)}$) where i is the number of healthy sites) in the non-overlapping generation, then we need a selection coefficient S_C in linear fitness scheme ($r_i = -S_C(L - i)$ in continuous time. The selection coefficients are related through the function $S_C = -\log(1 - S_D)$.

2. Diffusion approximation. In a finite population neglecting back mutations, the number of healthy sites irreversibly decreases (Muller's ratchet [21]). Here we calculate the mean time until the loss of the most healthy genotype in the population (i.e., we ask, when starting from a population in which the maximum number of healthy sites is n, what is the expected time, τ_n, until the most healthy class is lost by random drift). This problem represents an n-dimensional diffusion process, but we can reduce the dimension to one, by concentrating on the frequency change in the most healthy class (i.e. p_n). This assumes that the frequencies of the other genotypes are kept close to those values observed in the stable distribution of the deterministic subsystem (with $p_{n+1} = p_{n+2} = \cdots p_L \equiv 0$). The justification for this approximation and the detailed analysis are considered elsewhere (Sasaki & Krakauer 2002). Once we obtain τ_n ($n = L, L - 1, \ldots, 2$), the mean number of healthy sites at generation T is $\bar{k} = n^*_{(T)}(1 - u/s)$ where $n^*_{(T)} = \max(n \,|\, \tau_L + \tau_{L-1} + \cdots + \tau_n > T)$, and the mean fitness $\exp(\bar{w})$ where $\bar{w} = w_{n^*_{(T)}} - n^*_{(T)}\mu$.

References

[1] Ancel, L. W. A quantitative model of the Simpson-Baldwin Effect. J Theor Biol **196**(2), 197-209. (1999)

[2] Benner, S. A., Allemann, R. K., Ellington, A. D., Ge, L., Glasfeld, A., Leanz, G. F., Krauch, T., MacPherson, L. J., Moroney, S., Piccirilli, J. A.,et al. Natural selection, protein engineering, and the last riboorganism: rational model building in biochemistry. Cold Spring Harb Symp Quant Biol **52**, 53-63 (1987)

[3] Cairns-Smith, A. G. The origin of life and the nature of the primitive gene. J Theor Biol **10**(1), 53-88. (1966)

[4] Cech, T. R. RNA as an enzyme. Biochem Int **18**(1), 7-14. (1989)

[5] de Duve, C. Co-chairman's remarks: the RNA world: before and after. Gene **135**(1-2), 29-31. (1993)

[6] de Duve, C. Vital Dust. Life as a Cosmic Imperative, BasicBooks, New York (1995)

[7] Dyson, F. J. A model for the origin of life. J Mol Evol **18**(5), 344-50 (1982)

[8] Eigen, M. Molecular self-organization and the early stages of evolution. Experientia **27**(11), 149-212 (1971)

[9] Eigen, M. Natural selection: a phase transition? Biophys Chem **85**(2-3), 101-23. (2000)

[10] Eigen, M. Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften **58**, 465-523 (1971)

[11] Eigen, M.,Schuster, P. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften **64**(11), 541-65. (1977)

[12] Eigen, M.,Schuster, P. Stages of emerging life–five principles of early organization. J Mol Evol **19**(1), 47-61 (1982)

[13] Fox, S. W. (ed) (1984) Proteinoid experiments and evolutionary theory. Beyond Neo-Darwinism. Edited by Ho, M. W., and Saunders, P. T., Academic Press, New York

[14] Haldane, J. B. S. The origin of life. Rationalist Annual, 148-169 (1929)

[15] Kauffman, S. A. Autocatalytic sets of proteins. J Theor Biol **119**(1), 1-24. (1986)

[16] Kimura, M. The Neutral Theory of Molecular Evolution, Cambridge University Press, New York (1983)

[17] King, G. A. Symbiosis and the origin of life. Orig Life **8**(1), 39-53. (1977)

[18] Krakauer, D. C.,Plotkin, J. B. Redundancy, antiredundancy, and the robustness of genomes. Proc Natl Acad Sci U S A **99**(3), 1405-1409. (2002)

[19] Lazcano, A.,Miller, S. L. The origin and early evolution of life: prebiotic chemistry, the pre- RNA world, and time. Cell **85**(6), 793-8. (1996)

[20] Miller, S. L. A production of amino acids under possible primitive Earth conditions. Science **117**, 528-529 (1953)

[21] Muller, H. J. Our loads of mutations. Am. J. Hum. Genet **2**, 111-176 (1950)

[22] Nowak, M. A., Boerlijst, M. C., Cooke, J.,Smith, J. M. Evolution of genetic redundancy. Nature **388**(6638), 167-71. (1997)

[23] Oparin, A. I. The Origin of Life on Earth, 3rd edn, Oliver and Boyd, Edinburgh (1957)

[24] Orgel, L. E. Evolution of the genetic apparatus: a review. Journal of Molecular Biology **38**, 381-93 (1968)

[25] Orgel, L. E. The origin of life on the earth. Sci Am **271**(4), 76-83. (1994)

[26] Orgel, L. E. The origin of life–a review of facts and speculations. Trends Biochem Sci **23**(12), 491-5. (1998)

[27] Oro, J., Miller, S. L.,Lazcano, A. The origin and early evolution of life on Earth. Annu Rev Earth Planet Sci **18**, 317-56 (1990)

[28] Pace, N. R.,Marsh, T. L. RNA catalysis and the origin of life. Orig Life Evol Biosph **16**(2), 97-116 (1985)

[29] Scherer, S. Could life have arisen in the primitive atmosphere? J Mol Evol **22**(1), 91-4 (1985)

[30] Segre, D., Shenhav, B., Kafri, R.,Lancet, D. The molecular roots of compositional inheritance. J Theor Biol **213**(3), 481-91. (2001)

[31] Woese, C. R. A proposal concerning the origin of life on the planet earth. J Mol Evol **13**(2), 95-101. (1979)