*Journal of Biomolecular Structure &
Dynamics, ISSN 0739-1102
Volume 19, Issue Number 6, (2002)
©Adenine Press (2002)*

Statistical Limitations on Molecular Evolution

Leonid I. Perlovsky

Air Force Research Lab.
80 Scott Road
Hanscom AFB, MA 01731

http://www.jbsdonline.com

Abstract

Complexity of functions evolving in an evolution process are expected to be limited by the time length of an evolution process among other factors. This paper outlines a general method of deriving function-complexity limitations based on mathematical statistics and independent from details of a biological or genetic mechanism of the evolution of the function. Limitations on the emergence of life are derived, these limitations indicate a possibility of a very fast evolution and are consistent with "RNA world" hypothesis. The discussed method is general and can be used to characterize evolution of more specific biological organism functions and relate functions to genetic structures. The derived general limitations indicate that a co-evolution of multiple functions and species could be a slow process, whereas an evolution of a specific function might proceed very fast, so that no trace of intermediate forms (species) is preserved in fossil records of phenotype or DNA structure; this is consistent with a picture of "punctuated equilibrium".

Introduction

Was there enough time in the history of Earth for evolution of life from inorganic matter by mere statistical physical-chemical processes? Simplistic assumptions that life emerged as a single-time fluctuation from elementary atoms or inorganic molecules lead to a vanishing probability for such a fluctuation during the time of Earth existence (as well as anywhere in the Universe during the Universe lifetime). It is therefore often considered more likely that life emerged from pre-life chemical compounds in a series of hierarchical fluctuations, yet such processes leading to non-vanishing probabilities have not been identified. Genetic mechanisms seems to be too complex for such an analysis.

There is an "opposite" issue that is also puzzling: if we succeed with explaining evolution by chemical, physical, and statistical processes, we will have to explain why did it take so long for human to emerge on Earth, and why the Universe is (at least) not clearly over-abundant with life forms (why have not we encountered extraterrestrial intelligence yet?). Current knowledge of environmental limitations is not yet sufficient to answer this question quantitatively.

This paper attempts to approach the problem of evolved function-complexity and its limitations based on Cramer-Rao (CR) theory (1). The CR approach can be used to derive limitations (CR bounds, CRB) which do not depend on the details of mechanisms of the function evolution. It is a general method that can be used to derive complexity limitations for the emergence of life, as well as for evolution of more specific functions, such as anti-agents, or functions of mind, like conceptual thinking. A first suggestion to apply the CR theory to evolution was discussed in (2), and this paper is a first, in many respects a crude step toward development of this method.

Phone: 781-377-1728
Fax: 781-377-8984
Email: leonid.perlovsky@hanscom.af.mil

The next section, *The Cramer-Rao Theory*, briefly summarizes the relevant aspects of the CR theory and illustrates it in the context of molecular evolution. Modifications of the CR theory suitable for evolutionary processes are considered in, *The evolutionary Cramer-Rao Bounds, CRBe*, leading to evolutionary limitations, or "evolutionary" CRBe that indicates a possibility of superfast evolution. In *A Cramer-Rao Limitations for Emergence of Life,* the application of CRBe to emergence of life is discussed using very crude assumptions. The CRBe is related to an RNA structure with the main purpose to illustrate how the method can be used for identifying "fast" and "slow" evolutionary mechanisms. *CRB for the Origin of Species and Genes* considers more complicated CRBg for gene differentiation, origin of species, and coevolution. It is shown that individual functions might evolve very fast, whereas coevolution might be a slow process. This result is consistent with "punctuated equilibrium" (3). The *Conclusion* briefly discusses directions of further extensions of the method toward characterizing more specific functions and relating them to genetic structures.

The Cramer-Rao Theory

The CR theory is a fundamental result of mathematical statistics, serving as a basis for statistical estimation theory, yet relatively little known outside of this area. Therefore, first, I consider a simple example illustrating the relevant ideas, then briefly summarize classical results of the CR theory. Another example considers a "toy" evolutionary process, it illustrates required modifications of classical CRB for evolutionary processes, which are considered in the next section. Whereas the power of the CR theory is that it can be used while no knowledge of the evolutionary mechanism is available, for illustrations here I selected two "toy" examples with known mechanisms. The first example is a most simple one: there is no evolution and a simple well-known mechanism of averaging is used. The second example uses a simple auto-catalytic evolutionary mechanism.

Example 1 (averaging, no evolution). Suppose, an organism O needs to find an object F (say, food), O cannot see F directly, but can perceive its smell. To make things simple, assume that F does not move, its "smell" is due to N molecules emitted all at once and randomly distributed (in two-dimensional space), and O "perceives" at once coordinates, $\mathbf{x}^n$, of all smell-molecules, n = 1,... N. The unknown location of F can be determined by a widely used statistical estimation technique, averaging,

$$\mathbf{X} = \Sigma_n \mathbf{x}^n / N. \qquad [1]$$

In case, when a probability density function (pdf) for each $\mathbf{x}^n$ is Gaussian (e.g., due to Brownian motions of "smell molecules") with a standard deviation σ_o,

$$pdf(\mathbf{x}^n) = (2\pi\sigma_o)^{-1} \exp(-(\mathbf{x}^n - \mathbf{X})^2 / 2\sigma_o^2), \qquad [2]$$

the error of $\mathbf{X}$ estimated according to [1] is given by a well-known (1/√N)-rule:

$$\sigma_o / \sqrt{N}. \qquad [3]$$

For this case, CR theory states that estimation [1] is the best possible in the following sense: no other procedure could yield a smaller error than [3]. In other words, independently of the mechanism O uses to find F, it will not be able to do it with a smaller error. Stated differently, having N observations, O *could* derive the location of food F with the accuracy [3]. Moreover, I will emphasize again, CRB is a way to derive this best possible error without knowing the mechanism (in this case, [1]).

Classical CRB. Consider N "trial" data $\{\mathbf{x}^n, n = 1,\dots N\}$, and their pdf depending on unknown parameters, $\{a_p, p = 1,\dots P\}$, $pdf(\{\mathbf{x}^n\}|\{a_p\})$. Sometimes we will omit index p and denote parameters vector as $\mathbf{a} = (a_1,\dots a_P)$. The unknown parameter values, **a**, are to be learned or estimated from data $\{\mathbf{x}^n\}$ using some unknown algorithm or mechanism (these are the same). The estimation could only be done with some error (in a general case parameter estimates are correlated and their error is characterized by a covariance matrix, $\mathbf{C}_a$); the error of estimated parameters is no smaller than the limit given by CRB; this limit is also a matrix, which we will denote using the same symbol (in bold, when it is a matrix, **CRB**):

$$\mathbf{C}_a \geq \mathbf{CRB},\ \mathbf{CRB} = E\{-\partial^2 \ln \mathrm{pdf}(\{\mathbf{x}^n\}|\mathbf{a}) / \partial\mathbf{a}\ \partial\mathbf{a}\ \}^{-1}. \qquad [4]$$

Here, E{.} is a statistical expectation. In a simple case of uncorrelated estimates, this matrix expression reduces to a set of limitations for each a_p. In particular, for Example 1 case, there is just one unknown parameter, $\mathbf{a} = \mathbf{X}$, $\mathrm{pdf}(\{\mathbf{x}^n\}|\mathbf{X}) = \Pi_n\ \mathrm{pdf}(\mathbf{x}^n)$, and using [4] we derive [3]. In a general case of a correlated Gaussian distribution

$$\mathrm{pdf}(\mathbf{x}^n|\mathbf{a}) = (2\pi)^{-d/2}\ (\det\mathbf{C})^{-1/2} \exp(-0.5\ \mathbf{D}_n^T\ \mathbf{C}^{-1}\ \mathbf{D}_n),\ \ \mathbf{D}_n = (\mathbf{x}^n - \mathbf{a}), \qquad [5]$$

the CRB is computed from [4] and [5],

$$\mathbf{CRB} = \mathbf{C} / N. \qquad [6]$$

Note, that $\mathbf{C}_a$ in [4] is the matrix of error covariances in estimated values of parameters, whereas **C** in [5] and [6] is the matrix of covariances of the data $\mathbf{x}^n$. From an information-theoretical view, the CRB is a limit on how accurately parameters **a** could be learned from information contained in $\{\mathbf{x}^n\}$ (and in the functional form of the pdf, e.g., given by [4]).

Example 2 (autocatalitic evolution). An RNA of the first living (defined here as self-replicating) organism is modeled as a sequence $\mathbf{a} = (a_1,\dots a_P)$, here a_p are from an alphabet of existing "nucleotides", 1 through A. We call this RNA true-RNA. "Trial-RNA" molecules are randomly formed from the nucleotides by a "mutation" mechanism: at every trial one of trial-RNA nucleotides is randomly replaced by any one of the total A nucleotides with a probability given by mutation rate. Trial RNA are also destructed with a certain rate. In addition, trial-RNAs self-replicate with an efficiency, which is [1] not sufficient to overcome the random destruction and sustain growth and [2] increases with an increase in a similarity between the true and trial RNA. The similarity is measured as a number of nucleotides that match the true RNA. The self-replicating efficiency of true RNA is sufficient to sustain growth, but even if one nucleotide mismatches (one unfavorable mutation of a true-RNA), the efficiency is not sufficient for sustained growth. This last property is really nothing more than the definition of the "first or simplest true RNA". The details of this mechanism are described in the Appendix. A rational for such a mechanism is that unfavorable mutations are likely to decrease self-reproduction efficiency, and, say 10 unfavorable mutations are likely to be worse than 1 unfavorable mutation. The main purpose of this "toy" mechanism here is to illustrate the notions related to the CRB in the context of molecular evolution.

With appropriate values of chemical reaction rates, this mechanism results in an average similarity growth with time t (that is, with the number of trials) as illustrated in Figure 1, or equivalently to a reduction of an average error of each nucleotide. The average error can be defined as $\sigma = \mathrm{sqrt}(<(a'_p - a_p)^2>)$, where <.> is averaging over the population.

Considered as a random quantity over the population, the nucleotide a'_p, can be characterized by a probability density, $\mathrm{pdf}(a'_p)$, defined as a ratio of the number of RNA molecules with nucleotide a' at site p, $N(a'_p)$, to the total number of the RNA molecules, N, $\mathrm{pdf}(a'_p) = N(a'_p)/N$. Let us make a few simplifying assumptions, which are

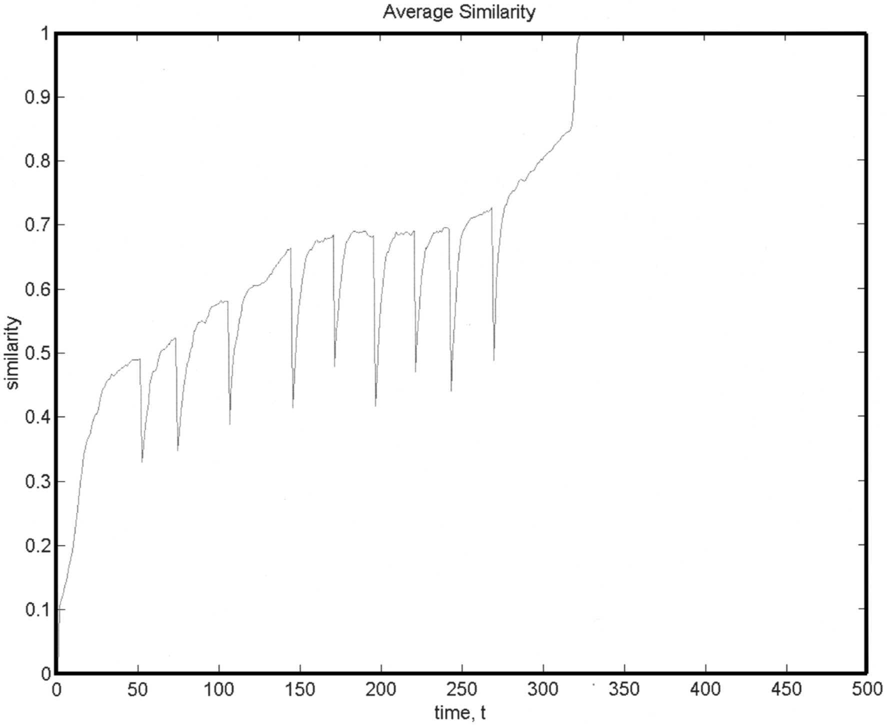

Figure 1: Normalized average similarity <s>/P. It grows with evolution from about 1/P to about 1; because of mutations, it is less than 1, its end value in this simulation run was 0.9998. (Saw-teeth pattern is a greatly exaggerated random variability, it should be ignored).

not essential for the method and which purpose is only to simplify the example. This pdf depends on the difference $(a'_p - a_p)$, this difference does not depend on p; pdf(a'_p) depends on the evolution time, t, which is measured in units of the replication cycles, one cycle is defined as an average time over which true-RNA replicates once. For the purpose of defining the functional shape of this pdf, below we consider $(a'_p - a_p)$ as averaged over a subpopulation, small relative to the entire population yet containing much more than one molecule (say, 10 to 30 molecules). Quantity $(a'_p - a_p)$ satisfies conditions of the Central Limit Theorem (CLT, or "the law of big numbers") (4), according to which it can be characterized by a Gaussian pdf:

$$\mathrm{pdf}(a'_p,t) = (2\pi\sigma_t^2)^{-1/2} \exp(-(a'_p - a_p)^2/ 2\sigma_t^2), \qquad [7]$$

with two parameters, a_p, and σ_t. These properties of pdf(a'_p, t) are illustrated in Figure 2; here pdf(a'_p, t) are shown vs. $(a'_p - a_p)$-axis for various values of t. We see that at the initial moment, t=0, the pdf is flat, as expected, any value of a'_p is equiprobable. In terms of eq. [7] this can be approximately described by $\sigma_t > A$. At later times, even well before life emergence (in this toy model), the average proportion of "correct" nucleotides, $a'_p=a_p$, grows because of the preferential replication, and the average error (standard deviation σ_t) becomes much smaller than one, after relatively small number of replications. The shape of the pdf looks like a Gaussian bell-shaped curve.

The evolutionary Cramer-Rao Bounds, CRBe

The classical CRB [4] was formulated for processes characterized by pdf independent of time. For evolutionary processes pdf depends on time; and the evolutionary CRBe accounting for this dependence through multiple "generations" or evolution cycles is derived here. Mutations, replications, and decays (deaths) at each generation may depend on previous generations due to specific mechanisms of these processes, but random aspects of these processes are statistically independent. In other words, say, a probability of mutation (and mutation rate) for a nucleotide a'_p at site p at time t may depend on the evolution mechanism and its parameters at time (t-1), yet mutation of a specific molecule among the population of exactly same molecules is usually (in genetic theories and models) considered random and independent from the previous generation. Given this, pdf in [4] can be written as

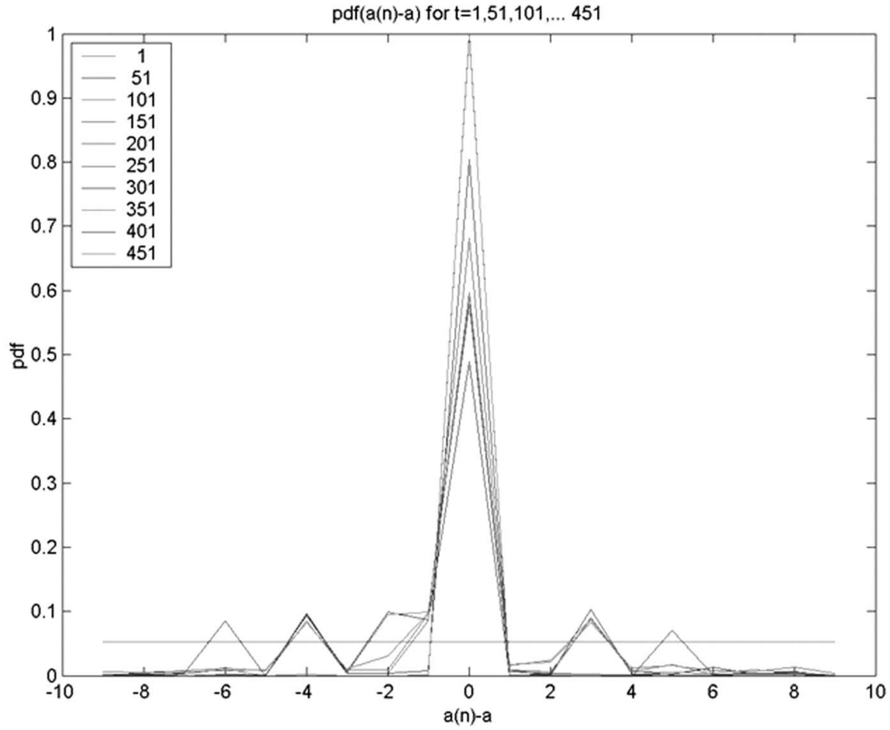

Figure 2: Probability density, pdf as functions of $(a^n_p - a_p)$ for ten values of time t = 1, 51,... 451. As evolution progresses, pdf becomes more narrow, centered around the true-RNA nucleotides, $a^n_p = a_p$. The final "width" seen in the figure is not an adequate representation of the standard deviation, the graph just connects points $(a^n_p - a_p) = 0$ and $(a^n_p - a_p) = \pm 1$. After approximately 300 generations, in this simulation run, the life emerged, and the standard deviation was quickly reduced to a tiny value remaining non-zero due to mutations.

$$\mathrm{pdf}(\{\mathbf{x}^n\}|\mathbf{a}) \;=\; \Pi_t\, \mathrm{pdf}(\{\mathbf{x}^n, t\}|\mathbf{a}, t). \qquad [8]$$

Here, $\{\mathbf{x}^n\}$ are all the "data" that is the trial-RNA molecules and their nucleotide composition, $\{\mathbf{a}'\}$, which are the results of replications, mutations, deaths and survivals during all "generations" from time 0 through the current time, and pdf($\{\mathbf{x}^n$, t$\}|\mathbf{a}$, t) is a pdf at time t for a "single generation" "data" $\{\mathbf{x}^n$, t$\}$ that existed from time (t-1) to time t. According to the previous argument, pdf($\{\mathbf{x}^n$, t$\}|\mathbf{a}$, t) are statistically independent for different times t, and therefore, the expectation E{.} in [4] can be re-written as a product of expectations at each time, E{.}= Π_t E{., t}, resulting in

$$\{\mathbf{CRB}e\}^{-1} \;=\; \Sigma_t\, \{\, \mathbf{CRB}^t \,\}^{-1}. \qquad [9]$$

Here, $\mathbf{CRB}^t$ is the CRB for the errors of parameters (that is unknown nucleotides of true-RNA) estimated at time t using *data* (that is known RNA sequences) that became available during the "generation" t *and information* available at the end of generation (t-1). This information, contained in the pdf at time (t-1), is given by its error-covariance matrix, $\mathbf{C}^{t-1}$, and is limited by the $\mathbf{CRB}^{t-1}$. In case of uncorrelated data (existing RNA sequences), that is pdf($\{\mathbf{x}^n$, t$\}|\mathbf{a}$, t) = $\Pi_{n\in t}$ pdf($\mathbf{x}^n|\mathbf{a}$, t), and uncorrelated errors of parameters (that is true_RNA nucleotides), all having the same standard deviation, σ_t, $\mathbf{CRB}^t$ is reduced to a unit matrix with σ_t^2 on the main diagonal. Similar to [3],

$$\sigma_t^2 = \sigma_{t-1}^2 / N_t = \sigma_o^2 / (\Pi_t\, N_t), \qquad [10]$$

and

$$\mathrm{CRBe} = \{\, \Sigma_t\, [\, \sigma_{t-1}^2 / N_t \,]^{-1} \,\}^{-1} \;=\; \sigma_o^2\, \{\, \Sigma_t\, [\, \Pi_t\, N_t \,]\, \}^{-1}; \qquad [11]$$

Assuming $N_t >> 1$, we obtain,

$$\mathrm{CRBe} \;=\; \sigma_o^2 / (\Pi_t\, N_t). \qquad [12]$$

As already mentioned, CRBe indicates a possibility of an extremely fast, "exponentially fast" evolution. This can be further illustrated as follows. For simplicity, consider a limited population, that is, N_t for all time t are the same, $N_t = N_1$; over T generations, the denominator in [11] is N_1^T; this is a much larger number than would be in the denominator of the classical CRB = T^*N_1, [6]. CRB can be reformulated in terms of the evolution speed as follows. Evolution of a particular biological function, entails evolution of a specific structure of a genetic code, which leads to a faster replication, therefore, to a predominance of this structure in a population, and consequently, to a reduction of the deviations from this structure. The deviations can be measured by an average error or standard deviation, and CRB gives a minimal number of generations necessary to reduce the standard deviation to a certain level. If σ_f indicates a sufficient (in a certain sense) accuracy of the genetic code, the minimal number of generations T_{min} corresponding to the CRB is given by CRB < σ_f^2, or, for the fixed population size N_1,

$$\begin{array}{ll} T_{min} > \ln(\sigma_o^2 / \sigma_f^2) / \ln(N_1) & \text{for CRBe, and} \\ T_{min} > (\sigma_o^2 / \sigma_f^2) / N_1 & \text{for classical CRB.} \end{array} \qquad [13]$$

If the ratio $(\sigma_o^2 / \sigma_f^2)$ is large, T_{min} according to CRBe is smaller than the classical one. This statement here is only given for the illustrative purposes; as discussed previously, an estimate of the time required for evolution is given by the CRBe. This number is not very large at all. To properly interpret this number we have to remember that in the derivation of CRBe we assumed that the entire information existing in the gene pool after each generation is available to the entire population at the next generation. Therefore, in the context of CRBe, a “generation” really refers to many generations, as required for the genetic information to spread around the gene pool. The exact number depends on more detailed models and is beyond the scope of this paper, but we could note, that because of the combinatorially fast nature of genetic information spread, it should not require a very large number of generations, say between 10 and 100.

A statistical meaning of CRB (as a minimal achievable error) can be reformulated in terms of information theory. Learning a function (say, self-replication) in the process of evolution requires certain amount of information, and the minimal amount of the required information is contained in the “trial” mutations and replications that occur during time T_{min}. The conclusion from the CRBe is that based on statistical or information-theoretic measures evolution can occur super-fast. Given the fact that it took more than 4 billion years for our evolution, and that the life is not over-abundant in the Universe (if at all exists anywhere outside Earth), we need to understand if the CR theory gives meaningful estimates, and if so, how could we explain a relatively slow pace of evolution as compared to CRBe?

The mathematical structure of the CRB is such that it is an “asymptotically tight bound”, that is, there are possible mechanisms that come close to the CRB after a large number of trials. This does not mean that such nearly efficient mechanisms can always be realized using existing chemical machinery. Yet, there are no good arguments against versatility of molecular mechanisms. A superfast evolution according to CRBe could be real, and it is important to understand the reasons for the relative “slowness” of the evolution processes. Therefore, below and in the following two sections we emphasize those aspects of the CR theory that could be useful for this purpose.

A general factor that can slow down the evolution of a specific function, relative to super-fast CRBe limits, independent of the details of evolutionary mechanisms is mutations. Let us quantify an effect of mutations on the minimal achievable average genetic spread. For simplicity, assume an equal mutation rate, μ, for each nucleotide. At each time t, there is no less than μ^*N_t errors. Each erroneous nucleotide introduces on average an error $\sigma_1 \sim A$; the average over the N_t RNAs, among which there are μ^*N_t erroneous ones, yields (for $\mu << 1$),

$$\sigma_t / A \geq \mu. \qquad [14]$$

Thus, the minimal attainable error at time t, is larger than either σ_t or μA. This can be approximated by modifying [11] to account for the mutations as follows,

$$\mathrm{CRBm} = \{ \Sigma_t [(\mu A)^2 + \sigma_{t-1}{}^2 / N_t]^{-1} \}^{-1}. \qquad [15]$$

According to CRBm, the minimal achievable error is determined by mutations. As long as the current $\sigma_t = \sigma_{t-1}{}^2 / N_t > (\mu A)^2$, the evolution proceeds at a super-fast rate determined by CRBe; beyond this, CRBm approaches $(\mu A)^2$.

To summarize, the CRBe and CRBm establish limits for an evolution of a particular function. This point will be elaborated further in *CRB for the Origin of Species and Genes*. The CRBe indicates a possibility of a super-fast evolution of a function, and CRBm indicates that this superfast evolution approaches the limit of the specificity of the function. Without the limit, the genetic spread would be eliminated and any adaptivity of this function will be lost. This role of mutations in preventing the loss of adaptivity is well appreciated, as well as existence of various mechanisms developed by evolution to regulate mutation rates for various functions.

A Cramer-Rao Limitations for Emergence of Life

Again, we consider life as an ability or function to reproduce an RNA. An RNA can serve information-retaining, catalytic, and replicating functions and was considered as a possible first living system (5). In this section we explore the CRB accounting for the fact that the number of nucleotides and their chemical compositions are “not known” to the evolutionary mechanism and their “learning” has to be a part of the molecular evolution. In this section we only use the knowledge that nucleotides do not contain elements higher than E=15 (phosphorus); this is not really a “life precursory information”, because occurrences of higher-atomic number elements are much rare (relative to lower elements) higher-atomic number elements do not interfere with formation of nucleotides in a statistically significant way. In other words, evolutionary mechanisms utilizing lower-atomic number elements have to appear first. Each “trial nucleotide site” in a trial-RNA can be occupied by any “appropriate trial-nucleotide”. The number of “appropriate” trial-nucleotides, A, affects the result, yet there is no definitive number in the biochemical literature, therefore we will start in this paper with the worst-case “dumb” estimation.

Let us denote the unknown number of elements in a nucleotide, L. The total number of combinations of L elements out of possible E elements, is E^L. Certainly, this is a great exaggeration, only a minute part of this number of element combinations forms molecules with sufficient lifetime to affect trial-processes of molecular evolution of nucleotides. Yet, let us start with this number, $A = E^L$. For a nucleotide in the RNA site p to be specified unambiguously, the evolution process ought to attain sufficient accuracy or the “narrowness” in terms of the “final” covariance in pdf [5],

$$\mathrm{CRB} < \sigma_f{}^2 << 1, \text{ and initial } \sigma_o \sim E^L. \qquad [16]$$

The previous results [13] let us express this requirement in terms of the evolution time (the number of generations). For illustrative purposes we will continue using the classical CRB alongside the CRBe and CRBm:

$$\begin{array}{lll} \text{Classical-CRB limit} & T_{cl} > (\sigma_o{}^2 / \sigma_f{}^2) / N_1 & >> E^{2L} / N_1, \\ \text{Evolutionary-CRBe limit} & T_e > \ln(\sigma_o{}^2 / \sigma_f{}^2) / \ln(N_1) & > 2L*\ln E / \ln N_1, \\ \text{Mutation-CRBm limit} & T_m > Te \text{ and } \sigma_f > \mu E^L. & [17] \end{array}$$

We see that the classical CRB results in a very large number of replication cycles, possibly exceeding the life of the Earth. The evolutionary CRBe indicates a possibility of a super-fast life emergence within few replication cycles. To properly interpret this result, we have to remember first, that the CRB is "asymptotically tight" (a factor of 10 to 100) and second, that according to a discussion after eq. [13] there is an additional factor of 10 to 100, so the CRBe limit is

Evolutionary-CRBe limit T_e > (100 to 1000) generations. [18]

The mutation CRBm indicates that the required accuracy for "learning nucleotides" cannot be achieved in our "dumb" model for $E = 15$ and $L > 30 : E^L > 10^{35}$, $\mu > 10^{-11}$, so $\sigma_f < 1$ cannot be attained. Of course this argument is not against "learning" chemical structures of nucleotides in the evolution process, but against the "dumb" model. It illustrates that the CRB method can be useful, in particular, it points toward a need for a better model for the "learning of nucleotides".

Surprisingly, the length of the RNA, P, turned out to be not a factor. This is due to an approximation implied in [13] and [16]. For *all* P nucleotides in the RNA to be specified unambiguously, the accuracy of each specification ought to be a little more tight: a probability of an error in *any of* P sites is P times larger than at a single site. Equations [13] and [16] corresponds to a single-site-probability of error ~ exp(-1/CRB); for P sites the probability of error is ~ P*exp(-1/CRB); therefore instead of [16], we have $\sigma_f << 1/\ln P$. This leads to the following modifications for the CRBe limitations [17] and [18]

$$T_e > \ln(\ \ln P * \sigma_o^2 / \sigma_f^2\ ) / \ln(N_1), \text{ and}$$
$$T_e > (100 \text{ to } 1000) + \ln(\ln P) / \ln(N_1).$$

This very weak double-logarithm dependence on the RNA length P suggests that long complex molecules might have been formed early, so the search for life emergence does not have to be restricted to simple molecules; this might be considered as a support for the "RNA world" hypothesis (6).

CRB for the Origin of Species and Genes

Here, we consider more detailed and more complicated models and CR Bounds corresponding to multiple co-evolving functions. We identify situations when the superfast evolution, according to the CRBe is possible, while at the same time, we attempt to identify statistical "states" of a genetic pool that may slow down the superfast evolution.

When a population contains several different types of organisms, conditions of the Central Limit Theorem (CLT) are not satisfied and Gaussian pdf does not describe the pdf of genetic codes in the population. (In this section we will refer to a genetic code as a DNA sequence). The reason for the CLT violation can be summarized as follows (2). In general, CLT conditions are satisfied when there is a single deterministic process with uncorrelated random variations. In Example 1 the deterministic process was given by the position of Food, X (and the standard deviation that evolved through time). Similarly, in Example 2 the deterministic process was specified by the true-RNA (and the standard deviation that evolved through time). When multiple types of organisms are present or evolving in the gene pool, each with its own "true-DNA", the condition of a single deterministic process with random variations is not satisfied. More complicated statistical model than Gaussian is needed. Consider a pdf of a gene pool at a particular time moment with multiple types of organisms; unlike Figure 2, it is likely to exhibit multiple peaks around corresponding true-DNA values. Each peak corresponds to a sub-population of a particular type of organisms. A corresponding statistical model is called a mixture model and is given by

$$\mathrm{pdf}(\mathbf{a}^n, t)\ =\ \Sigma_k\ r_k\ \mathrm{pdf}(\mathbf{a}^n|\mathbf{a}^k, t). \qquad [19]$$

Here, index n numbers any individual member of a genetic pool, k numbers types (true-DNAs), r_k is a proportion of the type k population in the entire pool, and all other notations are as in the previous section, and pdf($\mathbf{a}^n|\mathbf{a}^k$, t) is called a conditional pdf (the condition being that $\mathbf{a}^n$ are of type k). If we consider the population of only one type, k, its deviations from its true-DNA are likely to be random and the one-type sub-population pdf, the conditional pdf($\mathbf{a}^n|\mathbf{a}^k$, t), is likely to be Gaussian, with its mean given by $\mathbf{a}^k$ and a covariance matrix $\mathbf{C}^{k,t}$; in this case a model is called Gaussian mixture. As a first approximation, we are interested in time dependence of covariances $\mathbf{C}^{k,t}$, which determines a speed of evolution of a specific function. In the process of coevolution, the true-DNA, $\mathbf{a}^k$, are evolving with time, but we assume for now that this is a slower process.

This last point brings us to a need to discuss what is the "type" k of organisms, which we consider as characterized by a constant true-DNA^k for a long period of time. It could be a species, a gene, or a chunk of long-living genetic material as considered by Dawkins (7). The definition of "long-living" here is motivated by statistical considerations to make a model [19] more applicable, it is long-living in a statistical sense, as an average over some sub-population. A general statistical applicability of this definition is due to the fact that Gaussian mixture can model any shape of pdf, because Gaussian functions form a complete set of functions (in a space of non-negative functions, like pdf). In case of types k defined as species, n numbers complete-individual DNAs, if types k are defined as genes, n numbers each occurrence of k-th gene in the gene pool (and true-DNA^k refers to this gene sequence).

CRB for model [19] was derived in (2). Let us define

$$P(k|n) = r_k \text{ pdf}(\mathbf{a}^n|\mathbf{a}^k, t) \,/\, \text{pdf}(\mathbf{a}^n|t). \qquad [20]$$

This quantity is called in statistics the a posteriori Bayes probability; it is a probabilistic measure of an "individual" n belonging to a gene (species) k. For example, if for a particular k=k', pdf($\mathbf{a}^n|\mathbf{a}^{k'}$, t) is much larger than all other pdf($\mathbf{a}^n|\mathbf{a}^k$, t) for k≠k', P(k|n) = 1 for k=k', and 0 for k≠k', and the individual gene n is definitely of the type k'. If this holds true for all genes n of type k', k' gene is statistically distinct in the gene pool. Using the a posteriori probabilities, the gene (or species) CRBg can be written as

$$\text{CRBg} = \{ \mathbf{C}_k^{-1} \, [\Sigma_n \, E\{P(k|n)^2 \, \mathbf{D}_{nk}{}^T\mathbf{D}_{nk}\}] \, \mathbf{C}_k^{-1} \}^{-1}, \mathbf{D}_{nk} = (\mathbf{a}^n - \mathbf{a}^k), \qquad [21]$$

Here, the sum extends over the entire gene pool. Note, that for any statistically distinct gene k', only individual genes of the k' type contribute into this sum. For a statistically distinct gene k', P(k'|n) = 1, therefore, the sum in [21] yields $N_{k'}$*$C_{k'}$, where $N_{k'}$ and $C_{k'}$ are the population and covariance of k' type, and CRBg for k' is exactly same as CRBe [9 or 11].

Consider a situation, when a new gene or species k2 just begin to emerge in a population by separating from a previously existing one, k1. Initially, the two genes are not statistically different, there is an "overlapping" population n such that P(k1|n) and P(k2|n) are less than 1. In this case, the value of the sum in [21] is smaller than for a non-overlapping gene with the same population. E.g., if the initial k1 gene was statistically distinct, P(k1|n) + P(k2|n) = 1; for n∈k1, the smallest value of P(k1|n) is 0.5, and the minimum of P(k1|n)^2 is 0.25. Therefore CRBg for overlapping genes is larger than for the statistically distinct ones, and the speed of evolution for overlapping genes is slower than for the distinct ones. For two overlapping genes, this slowing factor, however, is not large, less than 4. A slow down of evolution due to an overlap could only occur if a large number of genes (species) are not statistically different. In biological terms, this is a situation when there are many different ecological niches or many different functions, potentially offering an evolutionary advantage, but the population has not yet differentiated into distinct species. This is a situation of “unstable equilibrium”, which cannot last for long, since every gene that gets differentiated due to some random events evolves very fast, according to CRBe.

The above model with unchanging true-DNAs is not completely general, it is possible to imagine “fast-coevolving” species (or genes) k1 and k2, with means (true-DNAs) changing over time as fast as covariances, $\mathbf{C}^{k1,t}$, $\mathbf{C}^{k2,t}$. In today’s world many (or most) genes and species are co-evolved and inter-dependent with many others. The CRB suitable for this more general case of gene co-evolution has been derived in (2). Let us denote the time-evolving true-DNA sequence of a species (or gene) k as $\mathbf{DNA}(k,\mathbf{a},t)$. Here, $\mathbf{a}$ are parameters that determine the $\mathbf{DNA}(k,\mathbf{a},t)$ model. Whereas actual DNA changes are discrete and non-differentiable, $\mathbf{DNA}(k,\mathbf{a},t)$ models, or evolutionary “pathways”, are statistical expectations, that continuously vary with time and parameters and coincide with actual DNA sequences only at some time points in evolution. Computation of these models require detailed evolutionary models, yet, similarly to previously derived CRB, a detailed knowledge of the evolutionary mechanism is not required. Development of $\mathbf{DNA}(k,\mathbf{a},t)$ models is beyond the scope of this paper, but we touch briefly on this in the next section along with a discussion of biological nature of parameters $\mathbf{a}$. The general CRBg is given by

$$\mathrm{CRBg} = \{ (\partial\mathbf{DNA}(k,\mathbf{a},t)/\partial\mathbf{a})^T \mathbf{C}_k^{-1} [\Sigma_n E\{P(k|n)^2 \mathbf{D}_{nk}^T \mathbf{D}_{nk}\}] \mathbf{C}_k^{-1} (\partial\mathbf{DNA}(k,\mathbf{a},t)/\partial\mathbf{a}) \}^{-1}. \qquad [22]$$

Compared to the previous CRBg expression [21], this more general CRBg contains derivatives $\partial\mathbf{DNA}(k,\mathbf{a},t) / \partial\mathbf{a}$. A distinguishing aspect of [22] is that it can be used to explain “slow” evolutionary processes and analyze “evolutionary delays”. At “turning points” of evolution, when

$$\partial\mathbf{DNA}(k,\mathbf{a},t) / \partial\mathbf{a} \approx 0, \qquad [23]$$

CRBg → ∞, that is the genetic variability in the population becomes large and cannot be quickly reduced. These points are *not* evolutionary “halts”, because the CRB is an averaged property and only gives limitations on statistical expectations; a random event, like mutation, would move the evolution away from the “turning point”. Also, genetic properties and evolutionary processes are not mathematically continuous functions, therefore, the exact zero cannot be attained in [23]. The “turning points” could be minima, maxima, or saddle points of $\mathbf{DNA}(k,\mathbf{a},t)$ and near such points a pace of evolution may significantly slow down.

General properties of the CRBg for co-evolution of genes (or species) derived in this section include possibilities on the one hand, of a superfast evolution of particular functions (“evolutionary pathways”) and on the other, slow evolution “turning points”. This might be considered as a support for the “punctuated equilibrium” hypothesis (3).

Conclusion

Two fundamental questions facing scientific theories of evolution are first, is it possible to explain the origin of life and its complicated forms, like ourselves, by purely physical and chemical processes? And second, if the first question is answered satisfactory, how to explain that life and intelligence are not overabundant in the Universe?

Having in mind these two questions, a new method of studying statistical and information-theoretic limitations on evolution and genetic structures based on the Cramer-Rao theory is described in this paper. It is relatively independent from specific evolutionary mechanisms in that general statistical properties of a gene pool can be used to derive limitations on more specific evolving genetic structures and biological functions. For example, cellular mechanisms of genetic replication do not have to be explicitly considered (the information-theoretic basis for this, of course, is in that DNA contains all the information). The CRB method emphasizes that an evolution is a "learning" process, various biological functions are learned, while the precision of the genetic representation of these functions increases in the evolutionary process.

The CRB method establishes the maximal complexity limitations for the evolved functions, which depend on the information extracted from environment in the process of evolution. The CRB does not specify the actual selection or replication mechanisms of evolution, nor guarantee that such mechanisms actually exist (or existed). Yet, the mathematical nature of the CRB is such that it is an "asymptotically tight" limitation, that is, "in the long run" there exists a mechanism attaining the CRB limitation. Such mechanisms are called efficient, they extract the maximal information from environment (for the purpose of learning a specific function). This property of the CRB (the existence of efficient mechanisms) taken together with the superfast CRBe derived in this paper emphasize that any doubt with respect to superfast evolutionary mechanisms having been realized by available chemical processes would have to be substantiated by scientific understanding of natural laws precluding such mechanisms. In other words, the results obtained in this paper to some extent shift an onus from the emphasis on explaining life as seemingly an "improbable event", to explaining why life and intelligence are not overabundant in the Universe.

Several CR Bounds suitable for evolutionary processes were derived. These bounds indicate on one hand, a possibility of a superfast evolution of specific functions and genetic structures, and on the other hand, that from time to time evolution reaches "turning points" of significant delays in the evolution process, which are related to coevolution of multiple functions and organisms. This might be considered as a support for the "punctuated equilibrium" hypothesis (3).

The derived CR Bounds were used for analyzing limitations on the emergence of life. Being a first such attempt and utilizing by necessity crude models, this is more of an illustration of the method than a detailed analysis. It has illustrated first, a possibility of fast emergence of self-reproducing long nucleotide sequences and second, a need for more detailed analysis of the evolutionary process of "learning the nucleotide alphabet". These results might be considered as a support for the "RNA world" hypothesis (6).

The proposed method is of a general nature: it can be used to relate complexity of evolved functions and gene structures to the previous less complex forms and to the duration of the evolution process. This method may lead to nontrivial limitations on structures and functions without requiring detailed knowledge of evolutionary and genetic mechanisms. Still, a detailed knowledge can be used to construct detailed models, leading to more specific results concerning yet unknown details of the processes. E.g., CRBm accounting for mutation mechanisms is an example of a more detailed, mechanism-specific model than CRBe, albeit the mutation specifics used is quite general. More specific models, for example, might start with a known genome for an ancient species early on an "evolutionary tree" and develop the CR limitations on the complexity of genes and their functions for the evolved species later on the "evolutionary tree". Future research along this direction will show how much of the complexity of the evolutionary pathways can be derived from statistical and informational considerations based on the CR theory.

What is the nature of **DNA**(k,**a**,t) models and their parameters **a**? In a most simple case, **a** is the same as **DNA**, **DNA**(k,**a**,t) = **a**. In a most complex case, imagine we succeeded in the development of a complete mathematical model of an evolutionary process, capable of predicting the evolution from simpler forms to more complex ones, depending on initial composition of the gene pool and environmental parameters. Parameters **a** describe the initial composition and environment. Useful models **DNA**(k,**a**,t) are between these two extremes, possibly capturing few effects for several types k, each isolated from other types (or for a small isolated subset of {k}). A simple approach toward developing relatively complicated models could be illustrated as follows. Say, one is interested in an evolution from an earlier known form of a DNA sequence, **a1** (at time = t1) to a later partially known DNA sequence, **a2** (at time = t2). Define an evolutionary "pathway" model, **DNA**(k,**a**,t) = **a1***(t2-t)/(t2-t1) + **a2***(t-t1)/(t2-t1); the parameters of this model **a** are the unknown nucleotides in **a2**.

Study of complex evolutionary processes will proceed to ever increasing levels of complexity. Moving to the next level will require adequate models representing genetic information at a previous level with all complexity of mutual dependencies. Beyond a certain level of complexity, it might become necessary to model cellular mechanisms: it is quite possible that complex relationships implicitly coded in the DNA are easier to represent via explicit cellular models. Another direction of future research toward relating functions and genetic structure may proceed by studying algorithmic functional models and applying the evolutionary CRB directly to functional models. The CRB for algorithmic functional representations are not different in principle from those for genetic structures. For example, studying the evolutionary CRB for functions of mind, such as conceptual thinking, may help identifying the corresponding molecular genetic structures.

Acknowledgements

The author is grateful to M. Frank-Kamenetskii and V. Petrov for valuable suggestions and comments.

Appendix

This appendix describes the details of the model used in the Example 2 and quantitatively illustrates notions involved in CRB formulation and interpretation, such as pdf for RNA sequences, measures of an accuracy and error of an evolving molecular function, and a gradual growth of the accuracy of the function coding (equivalently, genetic spread reduction in a subpopulation that adapts to this function in the evolution). Again, please keep in mind that the objective is not to introduce a realistic model of a molecular evolution, but to illustrate the notions related to CRB, using a simple model.

The model starts at a time t=0, when there are N_o "mechanisms that attempt to assemble an RNA from nucleotides". This "target RNA" is a sequence of nucleotides $\mathbf{a} = (a_1,\ldots a_P)$, which is not available to the mechanisms (the target RNA does not exist yet). The target RNA performs a self-replicating function, in other words, has a property that it can self-replicate faster than random mutations destroy it. The number of "mechanisms" is the same as the number of trial-RNA-sequences, each mechanism, n=1,...N, assembles a single sequence $\mathbf{a}^n = (a^n{}_1,\ldots a^n{}_P)$ as follows: (1) random mutations occur with rate μ at any of the nucleotide site p = 1,...P; a mutation replaces $a^n{}_p$ with any of a = 1,... A nucleotides; (2) replications occur with rate r, which is proportional to a similarity between the trial RNA and the target RNA; a similarity, s, is measured by the number of "correct" sequence sites (that is, matched to the target nucleotide), $a^n{}_p = a_p$; the similarity, s = 0,...P; and the replication rate is modeled as r(s) = b*s/P, for s = 0,...(P-1); for the target RNA, s = P, and the replication rate jumps to c > b. In this model, only the ratio of replications to mutations, r/μ, is important, so b is a redundant parameter and we choose b = 1.

The relative rate of growth and decay of "trial-RNA" molecules depends on mutation rate, μ, the sequence length P (we assume for simplicity that the length is the same for all sequences), and the number of nucleotides A. At any moment in evolution, t, the number of trial-RNA with similarity s is denoted $N_{t,s}$. The parameters $\{\mu, P, A\}$ are selected in such a way, that the growth of the number of molecules with s < P due to replication is slower than their random mutation rate. Yet, the number of mechanisms is not reduced to zero by mutations, because mutations in this model do not destroy "mechanisms", just make them less efficient replicators. To prevent indefinite growth of inefficient replicators, we introduce a decay mechanism with rate d, which grows with the number of mechanisms, $d = d_0*(1\text{-}\exp(\text{-}N_t/N_d))$. This can be described as assuming that there are "preferred" environmental niches, hosting on the order of N_d mechanisms, and the number of mechanisms beyond this preferred environment are destroyed with higher rates. All parameters are selected so that if c = 1, the total number of all assembled sequences would not grow with time indefinitely, $N_t < N1 < \infty$, for all t; but for c > 1, the target-RNA replicates faster than decayed, and the total number of RNA molecules grows indefinitely with time. This signifies the emergence of life (and the evolution will eventually "turn" to evolving other functions, which is not considered in the current model).

The model numerical simulations and probability density functions. Figure 1 and 2 show results of a particular simulation run with the following values of the model parameters. The number of different "nucleotides" A=10; number of nucleotides in an RNA sequence, P=10; mutation rate, μ=0.002; d_0=1; $N_d=10^8$; true-RNA reproduction rate, c=3; maximal number of generations simulated T=500. The total number of sequences existing at t is N_t. Of these, $N_{t,s}$ have similarity s (that is, s*P nucleotides are same as in the true-RNA), $N_t = \Sigma_s N_{t,s}$. Quantities $N_{t,s}/N_t$ measure the probabilities that a sequence with similarity s occurs, pdf(s) = $N_{t,s}/N_t$. Average similarity, <s> = Σ_s s*pdf(s). Average similarity s grows with evolution, but does not reach P because of mutations, normalized average similarity <s>/P is shown in Figure 1. A saw-teeth pattern is a greatly exaggerated random variability; it is an artifact of the simulation due to the fact that a huge number of RNA sequences were modeled with finite computer capabilities. The final number of RNA sequences after 500 generations was 10^{112}, it was simulated using only the maximum of 210 sequences (the maximal number of randomly generated sequences stored in a computer memory at any generation). This pattern of oscillations should be ignored. Similarly, pdf(a^n_p, t) is a proportion of the sequences with a specific nucleotide a^n at p-th site; asymptotically (after many cycles of mutations and replications) it does not depend on n, but only on a specific nucleotide a'_p at p-th site. Moreover, because in this model a random occurrence of a mutation or replication does not depend on neighbors, asymptotically, pdf(a^n_p, t) only depend on the value of (a^n_p - a_p). For this reason, Figure 2 shows these pdf as functions of (a^n_p - a_p) for ten values of time t = 1, 51,... 451. As evolution progresses, pdf becomes more narrow, centered around the true-RNA nucleotides, $a^n_p = a_p$. The final "width" seen in the figure is not an adequate representation of the standard deviation, the graph just connects points (a^n_p - a_p) = 0 and (a^n_p - a_p) = ±1. After approximately 300 generations, in this simulation run, the life emerged, and the standard deviation was quickly reduced to a tiny value remaining non-zero due to mutations.

References and Footnotes

1. H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton NJ (1946).
2. L. I. Perlovsky, *Neural Networks and Intellect: using model-based concepts*. Oxford University Press, New York, NY (2001).
3. N. Eldredge and S. J. Gould, *Punctuated equilibrium: an alternative to phyletic gradualism. In Models in Paleobiology*, ed. J. M. Schopf, Freeman Cooper, San Francisco, CA. (1972).
4. Here, it is an approximate statement as discussed later; for CLT see(1).
5. A. Brack, *Origin of life*, In Encyclopedia of life sciences, Macmillan Nature (2001).
6. A. Lazcano, *Prebiotic Chemistry*, In Encyclopedia of life sciences, Macmillan Nature (2001).
7. R. Dawkins, *The Selfish Gene*, Oxford Univ. Press, Oxford, GB (1976).

Date Received: February 16, 2002

Communicated by the Editor M. D. Frank-Kamenetskii