Molecular Evolution and Non-extensive Statistics

The origin of life is an open problem that has been deserved much attention during the last decades, due to recent developments on the understanding of the genetic code itself, and on the biochemistry involved in the emergence of life on Earth. The hypothesis on this issue goes from extraterrestrial origin (Oró) to the RNA world. Also this last model can be divided into many other branches, as the Oparin Ocean thesis, or the Woese's model of life emerging inside small droplets formed in the ancient Earth atmosphere.

The main facts constraining all those models are the earliest evidences of structurally complex life, indicated by microfissils, which are as old as $3.5X10^6$ years, the earliest indications of the simpler living structures, considered to be as old as $3.85X10^6$ years, and the so-called late heavy bombardment of the Earth surface by meteors, considered to have happened $3.8X10^6$ years from present, according to lunar records, which is considered to be the older date for beginning of life on our planet. These evidences give us an indication that the time for life evolutes from the simpler forms to the more complex structures are only 300 thousand years, and it is considered by many author to be a too short time for random combination of nucleotides produce such complexity in the life organization.

In vitro studies have shown that the elapsed time for abiotic production of organic molecules necessary for the RNA world are much smaller than the disponibility given by the evidences (Lazcano and Miller), and thus the RNA world hypothesis can not be discarded. But there is still a gap in the model, the explanation of how the nucleotide random combinations give rise to organized living structures in such a short time.

The studies of the emergence of life from the chaotic environment on the Earth surface are usually performed by computer simulation, also called *in silico* studies. However, up to now these simulations were not able to provide the guidelines for the origin of life (Lazcano and Miller). We show here that concepts from non-extensive statistical mechanics, from physics, can give such guidelines for *in silico* studies that must take into account the necessity of fast evolution from the simpler living forms to the first complex organisms.

The non-extensive statistic is better understood in the frame of the Tsallis entropy, which is a well-succeeded generalization of the Boltzmann entropy and has found many applications in different fields of knowledge. The main feature of the Tsallis entropy is its non-extensive character, which means that two systems with entropies S_A and S_B, when put together, give rise to a new system with entropy $S_C \neq S_A+S_B$. In a recent paper, Anteneodo and Tsallis have shown in which conditions some physical systems, known as spin-glasses, may present the non-extensive character.

These kind of system are well described by the Hamiltonian function

where $\mathbf{p}_i$ is the *ith* spin momentum, V_{ij} is the potential energy due to the interaction of spins i and j, r_{ij} is the distance between the relevant spins, and α is a positive exponent which is related to the range of interaction between the spins.

The non-extensivity of the system can be inferred from the parameter α and from the system dimension, d. If $\alpha > d$ the system is extensive, otherwise it is non-extensive. One important consequence of the non-extensivity is that large system present non-chaotic behaviour, i.e., their greater (correctly normalized) Lyapunov exponent tends to zero as the number of spins in the spin-glass system increases.

There are already many models for DNA and RNA evolution based on the spin-glass model. We show here how those models can take advantage of the non-extensive character of these systems to explain not only how the molecular evolution takes place, but also show why the evolution happens and, more important, in which conditions it will be faster. From now on, we will restrict ourselves on the RNA evolution, although the DNA evolution can, in principle, be also included in our study.

According to the RNA-world hypothesis, there were conditions at the Earth for the formation of small catalytic RNA molecules (ccRNA), but these molecules, however, were rather unstable due the environmental conditions, and their mutation rate was probably high. We may define a mutation function, **M**, which describes the mutations on the RNA along the time. This function may be represented, based on the spin-glass model, as

$$M = \sum_i p_i^2$$

where $\mathbf{p}_i$, now, is to be understood as a vector related to the probability that the nucleotide at the *ith* position at the molecule will mutate into the nucleotide A, U, C or G.

Equation 2 describes the mutation due to the environmental conditions only, but the ccRNA may also modify is environment, synthesizing proteins which will interact with other substances that would otherwise induce damages in some of the RNA's codons, or proteins which will act repairing those codons that had been damage, or even proteins which will damage unwanted codons which may be found at the same or in others RNA molecules. In this way, different codon structures located at the same or at different RNA molecules can interact with each other, as schematically shown in figure 1. This kind of interaction is conceptually similar to the interactions between elementary particles, which is mediated by the bosons exchanged by the two interacting particles. In the particle physics, this particle exchange gives rise to the potential energy, while here the exchange of proteins modifies the mutation rate.

We initially suppose that the RNA molecules are found in a close environment, such that all the proteins synthesized by the codons are confined to a small region in space, and therefore all codons are affected by the proteins synthesized by itself or by all the others codons present in that environment, no matter if they are at the same RNA molecule or not. To take into account the effects of the proteins synthesized by the RNA, we introduce another term in the function **M** defined above, which is then given by

$$M = \sum_i p_i^2 + \sum_{\substack{i,j \\ i \neq j}} V_{ij}$$

where the V_{ij} acts as the potential energy in the spin-glass model. Of course we use it here not as a potential energy, but as an inhibitor or stimulator of mutations, depending on its value, which can be positive (inhibits the mutation) or negative (stimulates the mutation). The quantities V_{ij} represent the environmental conditions determined by the presence of different ccRNA molecules and their synthesized proteins.

Comparing eqs. 3 and 1 we observe that they are similar if we put α=0. Therefore, it satisfies the condition for non-extensivity discussed above, and thus this system presents non-chaotic behaviour when the number of codifying agents, or codons, is sufficiently high.

A consequence of the non-extensivity is that, as the number of codons present in the system increases, a few stable configurations emerge from the chaotic mixture. If the codons are forming one RNA molecule, it means that only a few species, which are characterized by different coding sets, are possible to develop in that environment conditions, reducing the chaos and allowing fast evolution toward more complex living structures.

We performed *in silico* tests to verify these conclusions, and they corroborate our findings. Similar models for molecular evolution may be found at the literature, and they also agree with our results.

One important hypothesis in this study was that the potentials V_{ij} do not depend on the distance between the interacting codons. This condition can be clearly satisfied in the Woesel model for early life evolution taking place in small droplets on the atmosphere, but it is not obvious that it still holds in the Oparin Ocean model, since proteins synthesized by one RNA codon can not, in principle, interact with other codon in a molecule far away from the first one.

We can relax the condition for the interaction range by introducing a position dependent potential in the **M** function, as

$$M = \sum_{i} p_i^2 + \sum_{\substack{i,j \\ i \neq j}} \frac{V_{ij}}{r_{ij}^{\alpha}}$$

It is reasonable suppose that the probability for a protein synthesized by a codon *i* to interact with a codon *j* depends on the protein density on the codon *j* site. If the RNA density in the ocean is low, this density decreases as the inverse square of the distance between the two codons, and thus α=3, or faster. We can also suppose that the proteins produced may be inactivated by some reason, as damage or digestion, so that we may expect α≥2. In this case the system may be chaotic, leading to the competition among many different species, and slow evolution rates, which could be in disagreement with the strong constraints to the time available for evolution from the simpler living forms to the more complex ones.

We do not intend to present a complete study on the subject here, since many possibilities on RNA density and distribution and proteins dispersion through the Oparian Ocean must be considered, but the approach delineated here may give useful guidelines in the investigations of the early life evolution on Earth. The physical aspects of the RNA distribution and the diffusion of the substances on the environmental condition existing in the ocean are relevant in the determination of the exponent α, and we hope that the discussion presented in this work may contribute to stimulate researches on this issue.