Journal of Theoretical Biology ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Maximal gene number maintainable by stochastic correction – The second error threshold

András G. Hubai [a], Ádám Kun [a,b,c,*]

[a] Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös University, Pázmány Péter sétány 1/C, H – 1117 Budapest, Hungary
[b] MTA-ELTE-MTM Ecology Research Group, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
[c] Parmenides Centre for the Conceptual Foundation of Science, Kirchplatz 1, D – 82049 Munich/Pullach, Germany

HIGHLIGHTS

- The second error threshold is set by intravesicular competition and assortment load.
- Multilevel selection can support as much as a 100 genes.
- This system can mitigate a limited amount of competition asymmetry.

ARTICLE INFO

Article history:
Received 15 July 2015
Received in revised form
2 February 2016
Accepted 4 February 2016

Keywords:
Origin of life
Stochastic Corrector Model
Coexistence
RNA world

ABSTRACT

There is still no general solution to Eigen's Paradox, the chicken-or-egg problem of the origin of life: neither accurate copying, nor long genomes could have evolved without one another being established beforehand. But an array of small, individually replicating genes might offer a workaround, provided that multilevel selection assists the survival of the ensemble. There are two key difficulties that such a system has to overcome: the non-synchronous replication of genes, and their random assortment into daughter cells (the units of higher-level selection) upon fission. Here we find, using the Stochastic Corrector Model framework, that a large number ($\tau \geq 90$) of genes can coexist. Furthermore, the system can tolerate about 10% replication rate asymmetry (competition) among the genes. On this basis, we put forward a plausible (and testable!) scenario for how novel genes could have been incorporated into early living systems: a route to complex metabolism.

1. Introduction

It has been forty years since Manfred Eigen proposed the theory that mutations in molecular replication, a phenomenon considered conducive to the adaptation and speciation of the extant biota, could have posed a fundamental obstacle to the spontaneous formation of life (Eigen, 1971). The idea can be presented simply: early living systems lacking proof-reading processes had to tolerate a high rate of mutation; such mutation pressure precludes sustaining information in long chromosomes; but shorter genomes are unable to store proof-reading enzymes. For example, in the RNA world scenario "one cannot have accurate replication without a length of RNA, say, 2000 or more base pairs, and one cannot have that much RNA without accurate replication" (Maynard Smith, 1979). This is Eigen's Paradox which still troubles origin of life research: maintenance of information is a central topic of this field (Kun et al., 2015).

The notion of the error threshold was put forward with DNA genomes and peptide enzymes in mind. The 2000 base long RNA in Maynard Smith's example would code for an enzyme of length 600+, a small protein. A quick look at UNIPROT yields DNA-dependent DNA polymerases (E.C. 2.7.7.7) that are smaller than this, albeit mostly DNA polymerase IV, which is quite error prone.

On the other hand, reliable information replication evolved during the RNA world (Joyce, 2002; Kun et al., 2015; Yarus, 2011). The RNA world is the era in the history of Earth during which information was stored in RNA and catalysis was mostly done by RNA enzymes (ribozymes). At the moment, there is no known general RNA-based RNA polymerase ribozyme. There is a ribozyme which can catalyse the template based polymerisation of up to 98 nucleotides (Wochner et al., 2011), and given a very specific

* Corresponding author at: Department of Plant Systematics, Ecology and Theoretical Biology, Eötvös University, Pázmány Péter sétány 1/C, H – 1117 Budapest, Hungary.
E-mail addresses: hubaiandras@gmail.com (A.G. Hubai), kunadam@elte.hu (Á. Kun).

http://dx.doi.org/10.1016/j.jtbi.2016.02.007
0022-5193/

template a ribozyme can copy longer strands as well (Attwater et al., 2013) on par with its size of roughly 200 nucleotides. Still 200 nucleotides is a long sequence to emerge through non-enzymatic processes, when we take prebiotic replication fidelity into account (< 99%, Orgel, 1992).

The error threshold, in the simplified treatment of Maynard Smith (1983), is: $L < \ln s/\mu$, meaning that the maximum sustainable genome size (L) is less than the quotient of the natural logarithm of the selective superiority (s) of the sequence to be copied ('master') and the error rate (μ). Selective superiority is the ratio of the average Malthusian growth rates of selected sequences (here, only the master) versus the rest (here, its mutants). Let us say that the error rate is $\mu = 0.01$ (Orgel, 1992). Based on the above inequality, this only allows the sustainment of sequences shorter than $L < 100$ monomers (with the standard assumption that $\ln s \approx 1$). Thus the 200 bases long putative replicase ribozyme (Wochner et al., 2011) seems to be too long.

Recent advances paint a brighter picture. An order of magnitude longer functional ribozymes can be maintained (with the error rate being equal) if the structure of the ribozymes, and the neutral mutations it allows, are taken into account (Kun et al., 2005; Szilágyi et al., 2014; Takeuchi et al., 2005). Second, it seems that intragenomic recombination may have shifted the threshold by about 30% (Santos et al., 2004). Third, the processivity of replication (i.e. the constraint that during template-based replication, nucleotides have to be inserted one by one into the growing copy) may have somewhat filtered against errors, provided erroneous insertions had slowed down replication (Huang et al., 1992; Mendelman et al., 1990; Perrino and Loeb, 1989): erroneous copies would have thus suffered from an inherent fitness disadvantage (Leu et al., 2012; Rajamani et al., 2010). It may also have alleviated the error threshold by about another 30%.

While such a relaxed error threshold seems less problematic, the replication of whole genomes that could run a primitive metabolism is still out of reach. Ribocells (cells whose metabolism is run by RNA enzymes) require at least one ribozyme of each of the essential enzymatic functionalities to be considered viable: they can produce the biomass component necessary for growth and reproduction. Cells lacking even one of the functions cannot reproduce. Thus all information needs to be replicated, which can be done if all ribozymes replicate individually. Individual known ribozymes are short enough to be faithfully copied (Szilágyi et al., 2014). However, if individual genes are replicated, they have individual growth rates inside the cell. Sequences having the highest growth rates will dominate the ribozyme population, and other genes will be lost (cf. the Spiegelman experiment (Kacian et al., 1972)). Thus while the error catastrophe can be overcome by replicating the whole set of genes required for the cell as individual replicators, it creates another problem, that of non-synchronous replication. How much information can be integrated via the compartmentalisation of individually replicating ribozymes? Is such a system complex enough to overcome the error catastrophe?

The Stochastic Corrector Model (SCM) is a group selection/package model framework; it was developed to investigate the above compartmentalised system, which has the potential to solve the problem of information integration. Szathmáry and Demeter (1987) have shown that given a low number of replicators inside a cell having a far from optimal copy number distribution (the goal distribution can be arbitrary), stochastic separation of the genes into the daughter cells can ameliorate the copy number distribution of the parent cells. Previous works on the SCM have focused on cells with only two (Grey et al., 1995; Zintzaras et al., 2010) or three genes (Zintzaras et al., 2002). A few enzymes can coexist without a problem even without full compartmentalisation, i.e. on surfaces (Boerlijst, 2000; Czárán and Szathmáry, 2000; Hogeweg and Takeuchi, 2003; Könnyű and Czárán, 2013; Takeuchi and Hogeweg, 2009). And in vesicle models the coexistence of a few enzymes was demonstrated (Hogeweg and Takeuchi, 2003; Takeuchi and Hogeweg, 2009). An intellectual forebear to the SCM framework, the package model introduced by Niesert et al. (1981) shows that more than three genes can coexist. They assumed that cell division rate does not depend on its composition as long as at least one copy of each gene is present. A follow up study by Silvestre and Fontanari (Silvestre and Fontanari, 2008) shows the prerequisites for the coexistence of up to 10 genes. But the maximal number of coexisting genes was not investigated except by Fontanari et al. (2006), who have shown that arbitrary number of genes can coexist, if their replication rates are the same and the population size is infinitely large. These assumptions, however, are unrealistic—and as we will show—both of them critically affect gene coexistence.

Here we investigate how many independently replicating genes can coexist in a cell, despite the potential for information loss due to random assortment to daughter cells and non-synchronous replication. Information loss due to mutations in individual ribozymes is not investigated here. We already know that the error threshold limits the amount of information that can be maintained, and including it now would hamper our ability to assess how many genes can coexist despite different replication rates and random assortment into daughter cells. We show that these also limit the sustainable length of information. To distinguish these two sources of limitation, we term Eigen's limitation 'first error threshold' and the limitation investigated here 'second error threshold'.

2. Methods

We follow the dynamics of a population (N) of ribocells. The biomass of the cells is produced by an abstract metabolism requiring τ different enzymatic functions. Ribozymes (catalysts) replicate individually and there could be more than one ribozyme of each type in the cell. The internal composition of the cell, i.e. the number of ribozymes and their distribution among the metabolic functions, determines the metabolic activity (R_i), which in turn affects the growth and replication of the cell. Accordingly, a cell i containing $\nu_i \in [1, \nu_{\max}]$ independently replicating ribozymes distributed among the τ different genes each having $\nu_{i,j}$ copies ($\nu_i = \sum_{j=1}^{\tau} \nu_{i,j}$) has a metabolic activity

$$R_i = \frac{\nu_i}{\nu_{\max}} \cdot \left(c \prod_{j=1}^{\tau} \frac{\nu_{i,j}}{\nu_i} \right)^{\varepsilon}. \quad (1)$$

The main components of Eq. (1) are the effect of the cell size, and the effect of its composition. The greater the size of the cell, i.e. the number of ribozymes it harbours, the faster its metabolic activity will be (cf. the influx of materials through the surface). But the composition, the copy number of the different genes, matters as well. Each reaction (catalysed by its gene) is presumed to be essential in the metabolic pathway (producing intermediers (e.g. monomers) for the replication of the ribozymes). If a cell loses any such genes, it becomes unviable (perishes) as its metabolic activity goes to zero. The optimal composition, maximising the second part of the Eq. (1) is the uniform distribution of copy numbers: where every different gene (ribozyme) is present with an identical number of copies (cf. the inequality of arithmetic and geometric means). A constant $c = \tau^{\tau}$ ensures that a cell of size $\nu_{\max}$ always has a maximal metabolic activity of $R_i = 1$. An arbitrary exponent (ε) weights the two components (size and composition). In preliminary studies employing $0.3 < \varepsilon < 3$, we found that giving higher weight to the composition is beneficial for the sustainability of the genome. We used $\varepsilon = 0.3$ for all results presented below.

Please cite this article as: Hubai, A.G., Kun, Á., Maximal gene number maintainable by stochastic correction – The second error threshold. J. Theor. Biol. (2016), http://dx.doi.org/10.1016/j.jtbi.2016.02.007