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

András Hubai[a] and 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-MTMT 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

* Corresponding author

e-mail address of András Hubai: hubaiandras@gmail.com

e-mail address of Ádám Kun: kunadam@elte.hu

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 (τ >= 90) of genes can coexist. Furthermore, the system can tolerate about 10% of 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 evolving systems: a route to complex metabolism.

Highlights

- We find that the non-synchronous replication of genes and their random assortment into daughter cells result in a threshold-like drop in the maintainable number of individually replicating genes. We term this phenomena the second error threshold in reference to the first error limit caused by mutations (cf. Eigen's Paradox)
- Multilevel selection can supports no less than a 100 genes: the larger the cells are, the more genes they can uphold
- This system can mitigate a limited amount of competition asymmetry, further aiding the coexistence of genes

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 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 polymerization of up to 98 nucleotides (Wochner et al., 2011), and given a very specific 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 when we take prebiotic replication fidelity into account (<99%, (Orgel, 1992)).

The error threshold, in the simplified treatment of John 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 putative replicase ribozyme of 200 bases length (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 ribozymes 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 only 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 compartmentalization 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 compartmentalized 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 compartmentalization, 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). 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. However, neither of these assumptions is realistic—and as we will show–both of them critically affect the outcome.

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 limit 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 $R_i = \prod_{j=1}^{\tau} \left(\frac{\tau \cdot \nu_j}{\nu_i} \right)^{\varepsilon} \cdot \frac{\nu_i}{\nu_{\max}}$. Thus we assume that each gene catalyses an essential reaction in the metabolic pathway, producing intermediers (e.g. monomers) for the replication of the ribozymes. We further assume that there is an optimal distribution in the copy number of the different ribozymes, which corresponds to the highest metabolic activity inside the cell. We arbitrarily assign this optimum to the most even distribution, where every different ribozyme (gene) is present with an identical number of copies. We also presume that the greater the size of the cell, i.e. the number of ribozymes it harbours, the faster its metabolic activity will be. An arbitrary exponent (ε) weights these two components (evenness of distribution and ribozyme number). In pilot studies, we found that a selection focusing on the inner distribution is beneficial for the sustainability of the genome. In the studies to be presented we used $\varepsilon = 0.3$.

The population dynamics is the following: a cell is chosen randomly, in proportion to its metabolic activity (R_i); this cell gains one new ribozyme. Next, a ribozyme (j) is chosen randomly from the ribozymes in the selected cell, proportionately to its replicase affinity (a_j); this ribozyme is copied. The new ribozyme belongs to the same type and has the same replicase affinity as its parent.

If the number of ribozymes inside a cell reaches a maximal number ($\nu_{\max}$), then the cell splits into two. The ribozymes get into one of the daughter cells independently and randomly. The two new cells take the place of the parent cell and another one, which will perish, chosen randomly (with uniform probability): the population dynamics is a Moran process. Thus we assume constant population size, and fitness independent death-rate.

Pilot studies have shown that 50 “generations”, i.e. $50N$ cell divisions, are enough for the system to reach equilibrium.

The initial cells start with cell$\nu_{\max}/2$ ribozymes, with each function (gene) represented evenly.

3. Results

Genes can coexist in the SCM, and the parameter range in which coexistence is possible increases with population size (N) (Fig. 1) and gene redundancy within the cell (Fig. 2), furthermore, the more equal the replication rates of the individually replicating genes, the more genes can coexist (Fig. 3).

A larger population size allows more genes (functionally different ribozymes) to coexist (Fig. 1). This is not surprising, given the fact that an infinite population size guarantees coexistence (Fontanari et al., 2006). However, it is also important to know how infinity is approached. It seems that for most of the parameter space, an increase in population size has a meagre effect on cell viability. Thus, while it is possible to increase population size to achieve the coexistence of any number of genes, the additional population size required for it can be unrealistically large; e.g. nearly two magnitudes of increase in population size do not raise the fraction of viable cell when $\tau > 45$. Because of computational limits, we did most of our simulations with $N = 1000$. Thus our estimates are conservative, as increasing population size would allow for a slightly greater coexistence of replicators.

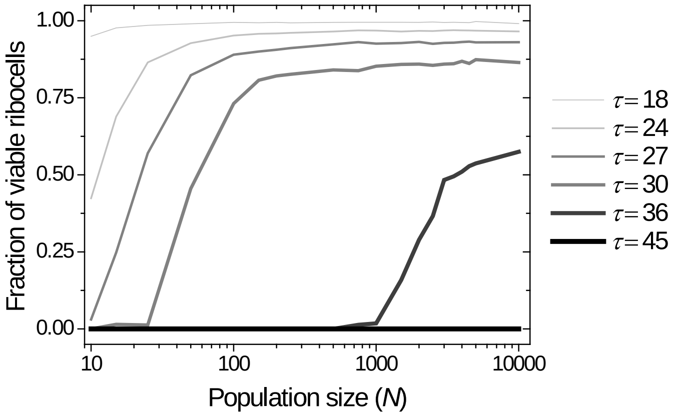

Figure 1. Fraction of viable cells increases with population size. Darker and wider lines represent systems with more required functions (τ). In all cases there are $\nu_{\max} = 2160$ ribozymes in the cells. All cells are viable if $\tau < 18$, and none is viable if $\tau > 45$.

Given a number of genes that need to coexist there exists a redundancy (maximum ribocell size, which translates to more ribozymes of each type) that allows it. The transition between the ribocell size that precludes coexistence and which ensures a viable population is threshold-like (Fig. 2). Increasing the maximum number of ribozymes inside the cells, and with it the achievable redundancy for each gene, increases the fraction of viable ribocells. The increase is sigmoid in shape with a very steep increase at certain point. This point is the second error threshold, i.e. the redundancy below which given number of genes cannot coexist. Thus at any given ribozyme abundance, there is a maximum to how many genes can coexist. Reaching a higher maintainable ribozyme diversity requires an increase in the number of ribozymes a cell can harbour. As a good rule of thumb, the maintainable genetic diversity (number of genes) is equal to the square root of the maximum number of ribozymes. At $N = 1000$ about a 100 different ribozymes can coexist if a cell can house 10,000 ribozymes per cell. Once a population passes the error threshold and becomes viable, further increase in gene redundancy has negligible effect.

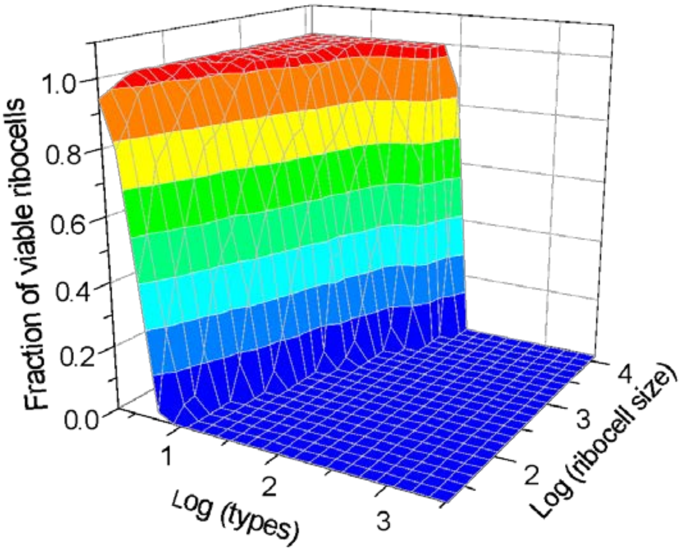

Figure 2. Maximum gene diversity. The fraction of viable ribocells is displayed as function of the number of types (τ) and the number of ribozymes per ribocell (ribocell size, ν). Please note that the logarithm of the number of types and ribocell size is on the X and Y axes, respectively.

The above investigations assume that each functionally different ribozyme (different genes) has the same affinity to the replicase, thus there is no competition asymmetry in the system. Three different scenarios are compared with the above results: (1) all affinities are the same, except for the affinity of one of the ribozymes, which is higher; (2) all affinities are the same, except for one, which is lower; and (3) affinities have two values, low and high, and the ribozymes of half (or about half in the cases of an odd number of genes) the genes have low, the other half have high affinity. Competition asymmetry causes the competitive exclusion of some of the genes, and consequently the loss of viability of the cells. When affinities differ as much as 10%, then in the case of a single competitively superior gene, coexistence is already lost at $\tau = 12$ (Fig.3a). In the case when there is a single competitively inferior gene, a considerable number of genes can coexist. In this case, the results are only different from the equal affinities scenario if $\tau \geq 96$ (Fig.3a). The stochastic corrector can tolerate all of these scenarios when the difference in the affinities is even lower, i.e. 1% (Fig.3b) or 0.1% (Fig.3c).

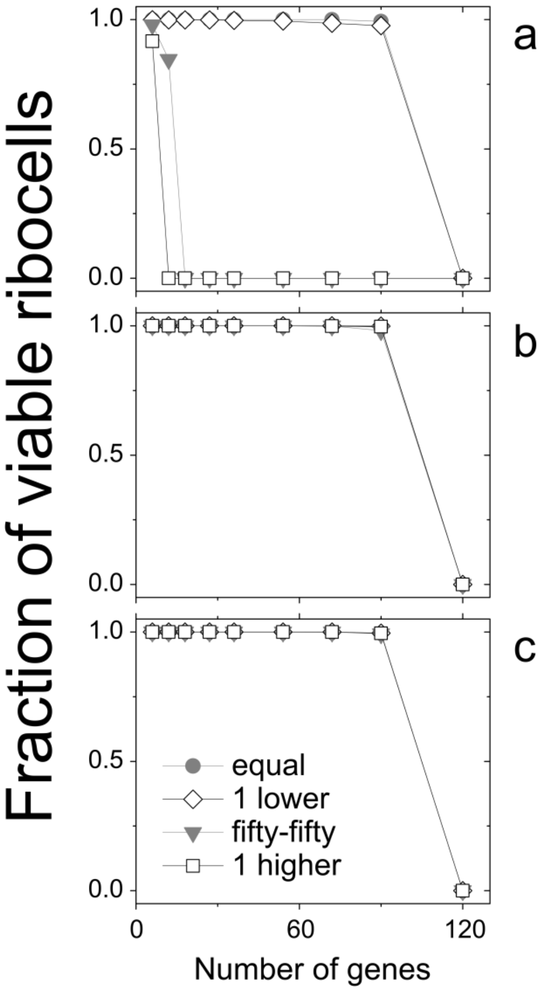

Figure 3. Mean fraction of viable cells with non-equal affinities to the replicase. "Equal" (solid square) $a_{1..\tau} = 1$; "One lower" (open upward triangle) $a_{1..(\tau-1)} = 1$ and $a_{\tau} = \alpha$); "1:1" (solid circle) $a_{1..(\tau/2)} = 1$ and $a_{(\tau/2)+1..\tau} = \alpha$; and "one higher" (open downward triangle) $a_1 = 1$ and $a_{2..\tau} = \alpha$. (a) $\alpha = 0.9$, (b) $\alpha = 0.99$, (c) $\alpha = 0.999$. Symbols represent means from 10 iterations. Other parameters g = 100000; $v_{\max} = 25920$; ε = 0.3.

We found that the relationship between the effect of asynchronous replication and the genetic composition of the ribocell (cf. metabolic activity) depends on the maximal cell size (Fig. 4). While a larger ribocell promotes a more even composition for equal replicase affinities, for an unequal distribution ($\alpha = 0.9$) this leads to a more adverse composition instead. There seems to be an optimal protocell size for asynchronous replication where despite the differences in replicase affinity a mostly even composition is sustained.

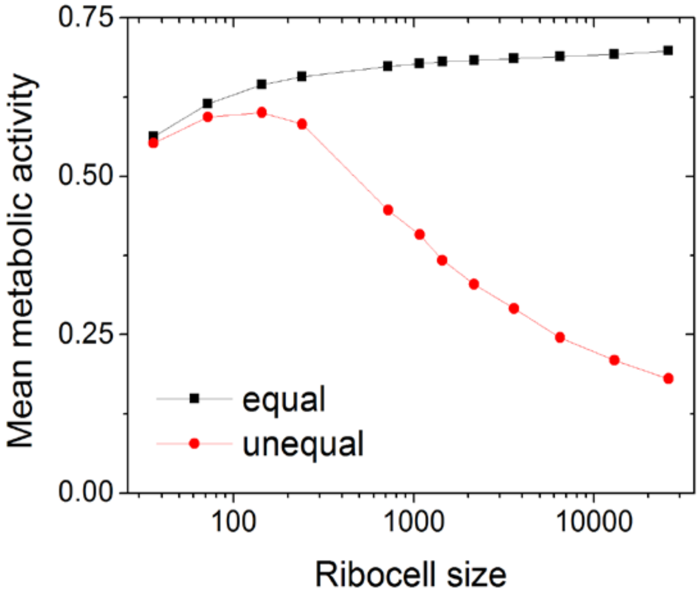

Figure 4. The effect of redundancy on asynchronous replication. In this minimal system of two genes ($\tau = 2$) we compared replicase affinities of equal distribution ($a_1 = a_2 = 1$) with those of an unequal distribution ($a_1 = 0.9$, $a_2 = 1$). The two lines show a divergent trend with the growth of protocell size. Note the logarithmic scale. Parameters g = 100000; N = 1000, t = 2, ε = 0.3.

4. Discussion

Non-synchronous replication and random assortment leads to a threshold-like decrease in the viability of ribocells as the number of type increases. Thus there is a limit to the enzymatic diversity that can be maintained in an ancient cell. We term this limit the second error threshold. Despite the second error threshold a sizable number of genes can coexist. Here we have shown that as many as 100 different genes (types) can coexist if internal copy number is moderately high and affinities do not differ by more than 1%.

4.1 Assumptions of the model

The computational analysis presented in this paper focuses on the phenomena we call the second error threshold. For this reason, we have excluded mutations from our current model. We understand that excluding possible mutations in the enzymatic activities can affect our results. Increasing mutational rate can push the population over the error threshold (Kun et al., 2015; Takeuchi and Hogeweg, 2012). Here we show that assortment load also leads to a threshold-like change in the viability of the population, independently of the first error threshold. Mutations can also produce parasites, sequences that do not contribute to biomass production, but which contribute to cell size. Thus cells might divide harbouring few enzymes and many parasites, leading to deficient daughter cells with a higher probability. On the other hand, such cells have a severe selective disadvantage, and they would divide at a slower rate compared to cell having few parasites. Efficient information integration despite the presence of parasites were demonstrated in the stochastic corrector model framework (e.g. (Zintzaras et al., 2002)), albeit for only 3 genes. The interplay of the two error thresholds will be revisited in a future study.

We assume that the limiting factor in the metabolism of the cells is the number of enzymes present. Food scarcity is irrelevant, as it does not differentiate between the ribocells, thus would not affect selection.

The employed fitness function assumes that an uniform distribution of every different ribozyme has the highest replication rate. It can be understood as either a linear (serial) set of all-essential reactions (for example, a chain of reactions that transform food molecule to monomers), or parallel pathways with equally important end-products (for example, the parallel production of all NTPs). In essence an arbitrary metabolic network can be employed, and the metabolic flux through the network can be used as a proxy for fitness (as used in (Szilágyi et al., 2012)). The added complexity of flux calculation can pose a technical difficulty, as the computational requirement is already quite high, and would necessitate other simplifying assumptions.

4.1. Minimal gene number of a ribo-organism

Minimal genome sizes found in contemporary organism can be as low as 112 kbases: *Nasuia deltocephalinicola* (112 kbasee) (Bennett and Moran, 2013), *Tremblaya princeps* (139 kbase) (McCutcheon and von Dohlen, 2011), *Hodgkinia cicadicola* (144 kbase) (McCutcheon et al., 2009), *Sulcia muelleri* (146 kbase) (Chang et al., 2015; McCutcheon and Moran, 2007; McCutcheon and Moran, 2010; Woyke et al., 2010; Wu et al., 2006), *Carsonella ruddii* (160 kbase) (Tamames et al., 2007), *Zinderia insecticola* (208 kbase) (McCutcheon and Moran, 2010). However, these symbionts of insects are barely alive in the sense that they lack genes for membrane and cell wall synthesis, lack transporters, most of carbon metabolism (McCutcheon and Moran, 2010) and some even lack some genes for DNA replication and translation. Other symbionts and intracellular parasites have genomes of around 600 kbases (*Mysoplasma genitalium*, *Buchnera* sp. (Islas et al., 2004)) and these minimalistic cells contain around 500-600 genes. However, the smallest possible genome size could have been even less (Luisi et al., 2006; Szathmáry, 2005): around 200 (Gil et al., 2004) (Table 1). These estimates pertain to cells having a DNA genome and peptide enzymes. A minimal ribo-organism can do with less. Jeffares *et al.* (1998) suggested that the last ribo-organism had a genome of 10,000-15,000 base pairs. This estimate includes ribozymes involved in translation and RNA replication, but it lacks enzymes for the control of cell division and the estimates for intermediate metabolism is rather arbitrary. The last ribo-organism most probably had translation, but we are more interested in the first cells, and not in the ones just on the verge to switch to DNA genomes.

A ribocell requires enzymes for the replication of its genetic material, chaperons for its ribozymes, maybe some enzymes that alters ribozymes much like post-translational modification alters peptide enzymes. Cellular processes, such as transport, also need some RNA enzymes. Moreover, the NTPs (both as monomers for RNA synthesis and as energy molecules), coenzymes and lipids need to be produced. A good estimate for the minimal intermediate metabolism covering said functionalities is given by Moya and co-workers (Gabaldón et al., 2007), who suggested 50 enzymes to be the minimum. We have to note that this set also included enzymes for dNTP production, which a ribo-organism did not need. A conservative estimate of 88 ribozymes is afforded by this back of the envelope calculation (Table 1). Most probably even fewer ribozymes would be enough, as this set of 88 contains multi-subunit enzymes as well (Gil et al., 2004). We have estimated 60 to be a minimum (Szilágyi et al., 2012), a more detailed analysis of the minimal set of genes required for a ribocell will be proposed later (Kun *et al. in prep*).

It is clear that even with 0.99 replication fidelity, a chromosome packed with 60 genes cannot be maintained due to the first error threshold. Sixty or even a hundred individually replicating genes can be maintained in randomly assorting ribocells. We thus conclude that the information required for a minimal ribocell can be propagated despite the second error threshold.

Table 1. Estimate of a minimal gene set for a ribo-organism

Function	Number of gene in a DNA-peptide organism	Number of gene in a ribo-organism	Notes
Replication of the genetic information	16	16	
translation	106	0	
Enzyme folding, modification and translocation	15	15	
Cellular processes	5	5	
Energetic and intermediary metabolism	56	52	no need for dNTP production
Total	198	88	

4.3 Possible evolutionary route to complex metabolism

Metabolisms having hundreds of enzymes and molecules do not appear at once. Most probably enzymes, and thus functions, were added one at a time (Szathmáry, 2007). A few enzyme can coexist on surfaces (Boerlijst, 2000; Czárán and Szathmáry, 2000; Hogeweg and Takeuchi, 2003; Könnyű and Czárán, 2013; Takeuchi and Hogeweg, 2009) as well as in vesicles (Hogeweg and Takeuchi, 2003; Szathmáry and Demeter, 1987; Takeuchi and Hogeweg, 2009; Zintzaras et al., 2002). How can we get from a few enzymes to nearly a hundred? The enhancement of metabolic capabilities afforded by more enzymes is surely selectively advantageous. On the other hand if the new enzyme cannot establish or coexist with the "old" ones, then this evolutionary step cannot be taken. Based on our results we can propose a possible evolutionary route to increasing metabolic complexity, i.e. more genes.

Equal affinities to the replicator ensure that no replicator outcompete the others. Thus the process could have started by a few (even two) ribozymes with equal replication rates. Now let us assume that any novel enzyme has a lower affinity to the replicase than the already established ones, then this enzyme can establish in the system, even if its affinity to the replicase is lower by as much as 10% compared to the rest of the enzymes (cf. Fig. 3a). Difference in affinities could not be very high: 60% difference is too much for the maintenance of a mere 10 enzymes, which is still too few for a metabolism. However, new enzymes probably evolved from established ones, and thus probably had tag sequences compatible with the replicase. The system then can evolve to equalize all affinities (Kun unpublished results), in this case to increase the affinity of the new enzyme. The simultaneous addition of more enzymes drive the system to extinction, but the addition of a single one seems to be feasible. Thus enzymes can be added one after the other with the requirement of only slight difference in affinities to the replicator.

The proposed evolutionary scenario of gradual increase in metabolic complexity can progress till the coexistence is no longer possible due to internal redundancy (Fig. 2), which can be alleviated by increasing the cell’s size at division. Cell sizes do not need to increase to infinity or even to very high number: at a certain metabolic complexity replication efficiency and fidelity could increase to a level at which a chromosome can be replicated. Then integration of the genetic information a chromosome can evolve (Maynard Smith and Szathmáry, 1993). The chromosome, a major evolutionary transition (Maynard Smith and Szathmáry, 1995; Szathmáry, 2015; Szathmáry and Maynard Smith, 1995), is made possible by overcoming the first error threshold. An intermediate solution to the first error threshold is the individual replication of ribozymes, which introduces the second error threshold. The second error threshold is alleviated by controlling the distribution of chromosome to the daughter cells.

5. Acknowledgement

We are grateful to Eörs Szathmáry, Mauro Santos and Elias Zintzaras for the valuable discussions regarding the topic of this paper. Financial support has been provided by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement no [294332]. ÁK acknowledges support by the European Union and co-financed by the European Social Fund (grant agreement no. TAMOP 4.2.1/B-09/1/KMR-2010-0003); and from the Hungarian Research Grants (OTKA K100299). ÁK gratefully acknowledges a János Bolyai Research Fellowship of the Hungarian Academy of Sciences. This work was carried out as part of EU COST action CM1304 “Emergence and Evolution of Complex Chemical Systems”.

6. References

Attwater, J., Wochner, A., Holliger, P., 2013. In-ice evolution of RNA polymerase ribozyme activity. Nature Chemistry 5, 1011–1018, doi:10.1038/nchem.1781

http://www.nature.com/nchem/journal/vaop/ncurrent/abs/nchem.1781.html#supplementary-information.

Bennett, G. M., Moran, N. A., 2013. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome Biology and Evolution 5, 1675-1688, doi:10.1093/gbe/evt118.

Boerlijst, M. C., 2000. Spirals and spots: Novel evolutionary phenomena through spatial self-structuring. In: Dieckmann, U., et al., Eds.), The Geometry of Ecological Interactions. Cambridge University Press, Cambridge, pp. 171-182.

Chang, H.-H., Cho, S.-T., Canale, M. C., Mugford, S. T., Lopes, J. R. S., Hogenhout, S. A., Kuo, C.-H., 2015. Complete genome sequence of "Candidatus *Sulcia muelleri*" ML, an obligate nutritional symbiont of maize leafhopper (*Dalbulus maidis*). Genome Announcements 3, doi:10.1128/genomeA.01483-14.

Czárán, T., Szathmáry, E., 2000. Coexistence of replicators in prebiotic evolution. In: Dieckmann, U., et al., Eds.), The Geometry of Ecological Interactions. Cambridge University Press, Cambridge, pp. 116-134.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenscaften 10, 465-523.

Fontanari, J. F., Santos, M., Szathmáry, E., 2006. Coexistence and error propagation in pre-biotic vesicle models: A group selection approach. Journal of Theoretical Biology 239, 247-256.

Gabaldón, T., Peretó, J., Montero, F., Gil, R., Latorre, A., Moya, A., 2007. Structural analyses of a hypothetical minimal metabolism. Philosophical Transactions of the Royal Society of London 362, 1761-1762, doi:10.1098/rstb.2007.2067.

Gil, R., Silva, F. J., Peretó, J., Moya, A., 2004. Determination of the core of a minimal bacterial gene set. Microbiology and Molecular Biology Reviews 68, 518-37.

Grey, D., Hutson, V., Szathmáry, E., 1995. A re-examination of the stochastic corrector model. Proceedings of the Royal Society of London B 262, 29-35.

Hogeweg, P., Takeuchi, N., 2003. Multilevel selection in models of prebiotic evolution: Compartments and spatial self-organization. Origins of Life and Evolution of the Biosphere 33, 375-403, doi:10.1023/a:1025754907141.

Huang, M.-M., Arnheim, N., Goodman, M. F., 1992. Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR. Nucleic Acids Research 20, 4567-4573, doi:10.1093/nar/20.17.4567.

Islas, S., Becerra, A., Luisi, P. L., Lazcano, A., 2004. Comparative genomics and the gene complement of a minimal cell. Origins of Life and Evolution of Biospheres 34, 243-256, doi:10.1023/b:orig.0000009844.90540.52.

Jeffares, D. C., Poole, A. M., Penny, D., 1998. Relics from the RNA world. Journal of Molecular Evolution 46, 18-36.

Joyce, G. F., 2002. The antiquity of RNA-based evolution. Nature 418, 214-220.

Kacian, D. L., Mills, D. R., Kramer, F. R., Spiegelman, S., 1972. A replicating RNA molecule suitable for a detailed analysis of extracellular evolution and replication. Proc. Natl. Acad. Sci. U. S. A. 69, 3038-3042

Könnyű, B., Czárán, T., 2013. Spatial aspects of prebiotic replicator coexistence and community stability in a surface-bound RNA world model. BMC Evolutionary Biology 13, 204, doi:10.1186/1471-2148-13-204.

Kun, Á., Mauro, S., Szathmáry, E., 2005. Real ribozymes suggest a relaxed error threshold. Nature Genetics 37, 1008-1011.

Kun, Á., Szilágyi, A., Könnyű, B., Boza, G., Zachár, I., Szathmáry, E., 2015. The dynamics of the RNA world: Insights and challenges. Annals of the New York Academy of Sciences 1341, 75-95, doi:10.1111/nyas.12700.

Leu, K., Kervio, E., Obermayer, B., Turk-MacLeod, R. M., Yuan, C., Luevano, J.-M., Chen, E., Gerland, U., Richert, C., Chen, I. A., 2012. Cascade of reduced speed and accuracy after errors in enzyme-free copying of nucleic acid sequences. Journal of the American Chemical Society 135, 354-366, doi:10.1021/ja3095558.

Luisi, P. L., Ferri, F., Stano, P., 2006. Approaches to semi-synthetic minimal cells: a review. Naturwissenschaften 93, 1-13.

Maynard Smith, J., 1979. Hypercycles and the origin of life. Nature 280, 445-446.

Maynard Smith, J., 1983. Models of evolution. Proceedings of the Royal Society of London B 219, 315-25.

McCutcheon, J. P., Moran, N. A., 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proceedings of the National Academy of Sciences 104, 19392-19397, doi:10.1073/pnas.0708855104.

McCutcheon, J. P., Moran, N. A., 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. Genome Biology and Evolution 2, 708-718, doi:10.1093/gbe/evq055.

McCutcheon, John P., von Dohlen, Carol D., 2011. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. Current Biology 21, 1366-1372, doi:http://dx.doi.org/10.1016/j.cub.2011.06.051.

McCutcheon, J. P., McDonald, B. R., Moran, N. A., 2009. Origin of an alternative genetic code in the extremely small and GC–rich genome of a bacterial symbiont. PLoS Genetics 5, e1000565, doi:10.1371/journal.pgen.1000565.

Mendelman, L. V., Petruska, J., Goodman, M. F., 1990. Base mispair extension kinetics. Comparison of DNA polymerase alpha and reverse transcriptase. Journal of Biological Chemistry 265, 2338-2346.

Orgel, L. E., 1992. Molecular replication. Nature 358, 203-209.

Perrino, F. W., Loeb, L. A., 1989. Differential extension of 3' mispairs is a major contribution to the high fidelity of calf thymus DNA polymerase-alpha. Journal of Biological Chemistry 264, 2898-2905.

Rajamani, S., Ichida, J. K., Antal, T., Treco, D. A., Leu, K., Nowak, M. A., Szostak, J. W., Chen, I. A., 2010. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. Journal of the American Chemical Society 132, 5880-5885, doi:10.1021/ja100780p.

Santos, M., Zintzaras, E., Szathmáry, E., 2004. Recombination in primeval genomes: a step forward but still a long leap from maintaining a sizeable genome. Journal of Molecular Evolution 59, 507-519.

Szathmáry, E., 2005. Life: in search of the simplest cell. Nature 433, 469-470.

Szathmáry, E., Demeter, L., 1987. Group selection of early replicators and the origin of life. Journal of Theoretical Biology 128.

Szilágyi, A., Kun, Á., Szathmáry, E., 2012. Early evolution of efficient enzymes and genome organization. Biology Direct 7, 38, doi:10.1186/1745-6150-7-38.

Szilágyi, A., Kun, Á., Szathmáry, E., 2014. Local neutral networks help maintain inaccurately replicating ribozymes. PLoS ONE 9, e109987, doi:10.1371/journal.pone.0109987.

Takeuchi, N., Hogeweg, P., 2009. Multilevel selection in models of prebiotic evolution II: A direct comparison of compartmentalization and spatial self-organization. PLoS Computational Biology 5, e1000542, doi:10.1371/journal.pcbi.1000542.

Takeuchi, N., Hogeweg, P., 2012. Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life. Physics of Life Reviews 9, 219-263, doi:http://dx.doi.org/10.1016/j.plrev.2012.06.001.

Takeuchi, N., Poorthuis, P. H., Hogeweg, P., 2005. Phenotypic error threshold; additivity and epistasis in RNA evolution. BMC Evolutionary Biology 5, 9.

Tamames, J., Gil, R., Latorre, A., Pereto, J., Silva, F., Moya, A., 2007. The frontier between cell and organelle: genome analysis of *Candidatus* Carsonella ruddii. BMC Evolutionary Biology 7, 181.

Wochner, A., Attwater, J., Coulson, A., Holliger, P., 2011. Ribozyme-catalyzed transcription of an active ribozyme. Science 332, 209-212, doi:10.1126/science.1200752.

Woyke, T., Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., Lapidus, A., Wu, D., McCutcheon, J. P., McDonald, B. R., Moran, N. A., Bristow, J., Cheng, J.-F., 2010. One bacterial cell, one complete genome. PLoS ONE 5, e10314, doi:10.1371/journal.pone.0010314.

Wu, D., Daugherty, S. C., Van Aken, S. E., Pai, G. H., Watkins, K. L., Khouri, H., Tallon, L. J., Zaborsky, J. M., Dunbar, H. E., Tran, P. L., Moran, N. A., Eisen, J. A., 2006. Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biology 4, e188, doi:10.1371/journal.pbio.0040188.

Yarus, M., 2011. Life from an RNA World: The Ancestor Within. Harvard University Press, Harvard, USA.

Zintzaras, E., Mauro, S., Szathmáry, E., 2002. "Living" under the challenge of information decay: the stochastic corrector model *versus* hypercycles. Journal of Theoretical Biology 217, 167-181.

Zintzaras, E., Santos, M., Szathmáry, E., 2010. Selfishness versus functional cooperation in a stochastic protocell model. Journal of Theoretical Biology 267, 605-613.