EÖTVÖS LORÁND UNIVERSITY
FACULTY OF SCIENCE

MASTER’S THESIS

Emergence and evolution of primeval metabolic systems

Author:

András Gábor Hubai
MSc student in biology
Ecology, Evolutionary and Conservation Biology spec.
hubaiandras@gmail.com

Supervisor:

Dr. Ádám Kun
research associate professor
Department of Plant Systematics, Ecology and Theoretical Biology
kunadam@caesar.elte.hu

2013

ABSTRACT

The minimal genome of a protocell at the early stages of the origin of life could not have possibly been stored on a single chromosome: the inaccuracy of enzymatic replication would have caused the quick decay of the majority of information (cf. Eigen’s Paradox). And though short sequences were copied with due precision even by then, the conservation of this minimal genome as separate genes poses its own problems: inevitable differences among replication rates lead to competitive exclusion and thus information loss. Through random segregation of protocell-enclosed gene packages at fission, the stochastic corrector model (SCM) enables the frequent recurrence of protocells with advantageous composition, and so the conservation of the whole set of genes despite intra-package competition.

We used an agent-based modelling framework to infer the maximum number of different genes that can be stably maintained depending on the number of vesicles and the number of molecules inside each vesicle. We show that stochastic correction enables the coexistence of about a 100 genes even with slightly unequal reproduction rates. A minimal living cell requires ca. 60-100 different genes thus it is reasonable to conclude that information integration is successful in our compartmentalized system: it is sufficient for a functioning protocell. We also presented a small set of mechanisms that can explain the observed characteristics of the dynamics. Our results suggest a possible evolutionary route through the serial integration of novel genes into the system while avoiding collapse.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

ACKNOWLEDGEMENTS

Ádám Kun provided me with guidance and encouragement over the last three years.
Eörs Szathmáry, Mauro Santos, and Elias Zintzaras commented on the fitness function, and also on possible research directions.
Tamás Hubai advised me on mathematical and IT topics.
Ádám Maulis facilitated my access to the ELTE IIG *Atlasz* high performance computer.

I wish to express my deepest gratitude for their help.

CHAPTER 1

INTRODUCTION

The spectacular diversity of life is a continuous source of fascination for humanity. Accordingly, there have always been descriptions of, and interpretations to this diversity. Some of the accounts have been ‘more fond of miracles, than apprehensive of truth’ (i.e., consistent with the reality of observations). Some achieved greater recognition than others. The current secular consensus originating from Darwinian theory (Darwin, 1859), the Extended Evolutionary Synthesis (Pigliucci, 2009; Pigliucci és Müller, 2010), offers explanation for phenotypic adjustments to the environment (adaptation and plastic response), the evolution of hereditary (genetic and epigenetic) traits, and speciation. However, it does not explain the origins of life, which necessarily preceded diversification. Thus understandably, scientists of our times investigate open questions of abiogenesis with an ever increasing enthusiasm.

Our scientific understanding of the origin of life is based on some fundamental principles: the same natural laws describe organisms capable of life and the inorganic matter they are composed of (*physicalism*); organisms do not form in a rapid and spontaneous manner (*univocal generation*); but they might form in slow and cumulative processes (*gradualism*); the present environment might not be suitable for the historic process of abiogenesis to recur (because of a different atmosphere, and life already present); but even these unique chain of events had to comply with the—presently observable—laws of nature (*uniformitarianism*); and once life has formed, its selective advantage must have ensured the continuous survival of living systems (*evolution*) (cf. Darwin, 1887, p. 18; Oparin, 1938; and Brack, 1998, pp. 1–2).

We can thus outline abiogenesis as the process between organic compounds capable of abiotic formation and the first, presumably simple, living organism capable of evolution. The criteria for the capability of evolution according to Maynard Smith (1987) is proliferation, trait heritability, and—due to error in the previous—variability. There is no similarly accepted definition of life: even if we believe that self-preservation requires a metabolism (the build-up of internal constituents from external resources), a semipermeable membrane, and a genome (Luisi, 1998), these cannot function independently of the environment, especially not is the earliest organisms (Szathmáry, 2007). Most abiogenesis research focuses accordingly on the early composition of the chemical environment (and as part of this, abiotically formed organic substances), and the metabolism, membrane and genome of the first organisms.

An interesting discovery affecting our concept of the origin of life was that there are chemical substances—most notably, ribonucleic acid (RNA)—with the dual capability of catalyzing reactions and storing information, which are the key functions in a metabolism and a genome, respectively. The resulting hypothesis, that RNA could have an essential role in the origin of life and early organisms, was termed the ‘RNA world’ scenario (Gilbert, 1986). An RNA molecule having a catalytic function is a *ribozyme*; and a hypothetical organism with RNA driven metabolism is called a *riboorganism*. The RNA world scenario has inspired many a scientists (Jeffares *et al.*, 1998; Bartel & Unrau, 1999; Szathmáry, 1999; Yarus, 1999; Murray & Doudna, 2001; Joyce, 2002 b; Orgel, 2004; Poole, 2006; Cech, 2009; Kun, 2011). Our study will refer to this scenario to illuminate certain theoretical considerations, but wishes to retain its generality concerning the origin of life.

1.1. *Preserving information*

Genes are the most likely subjects of early evolution—some suggest that the compositional information of metabolisms (Kauffman, 1986) or membranes (Segré & Lancet, 2000) are also evolvable, but that proposition is fundamentally contested (cf. Vasas *et al*., 2010). The prime feature which makes genes the most probable candidates is their specific structure—their constituent modules (e.g. nucleotides) form a linear, unbranched strand (*straight-chain*)—enabling their template-based, modular replication (Szathmáry & Maynard Smith, 1997). If a copy is to be created of a gene, its module-order (*master sequence*) is preserved through the template effect: the physical proximity of the original strand determined which modules can be incorporated in the appropriate positions of the new strand.

However, replication processes are prone to error: genes with alternate (*mutant*) sequences may form. In the case of modular replication, error can be present (or absent) independently in each module, resulting in a disproportionately large number of possible variants (I): $I = K^L$, where K is the number of different modules (i.e., the possibilities in their chemical repertoire), and L is the length of the sequence (i.e., the number of modules copied). Note that I also measures the information stored in the master sequence. If the number of possible variants is orders of magnitude greater than the obtainable amount of copies in the environment, the genes will qualify as *unlimited hereditary replicators*, capable of 'open-ended' evolution (Szathmáry & Maynard Smith, 1997; Vasas *et al*., 2010).

We shall be reminded that not all genes have an astronomical number of possible variants: that is a property of long sequences. And it is uncertain that long genes were of existence at the origins of life. As Eigen (1971) has recognized, preserving long genes is utmost problematic: first, if the probability of error is independent for each module, then longer sequences will have more mutations in their copies (assuming a constant rate of error); furthermore, there is a critical gene length (for any given error rate) above which the proportion of mutants increases so, that the master sequence practically disappears (Eigen, 1971; Schuster, 2010). The concept that random processes (i.e., error) pose a limit on the length of sustainable information is termed *error threshold*. In light of another phenomenon causing this same effect, we will refer to Eigen's finding as the *first* error threshold.

In the absence of long genomes there is no accurate replication, and without accurate replication there are no long genomes; so goes Eigen's Paradox (Maynard Smith, 1979). For the repair enzymes, which form the basis of accurate replication, most probably have to be long. This is truly a paradoxical situation, since while there seems to have been an obstacle at the origins, we certainly have both long genes and accurate replication. So the questions arise: what accuracy was available at the dawn of life? What genome length could it sustain? And how could an early chemical or biological system circumvent Eigen's Paradox at this accuracy and genome length?

To estimate the primeval error rate, let us start from the minimal accuracy of present living species (a 'top-down' approach; see Table I1). *E. coli* bacteria have a relatively accurate replication with a low error rate of 10^{-8}–10^{-10} per nucleotide (Schaaper, 1993; Kunkel, 2004). Other bacteria may have a higher error rate of up to 10^{-6} per nucleotide; but even that is only possible through complex repair mechanisms, which most certainly have evolved to ensure a specific rate of error (Joyce, 2002 a). Bacteriophage viruses have a higher error rate of up to 10^{-4} per nucleotide; but they are replicated by the same bacterial machinery, that has repair mechanisms—it is their lack of self-sustaining capability (metabolism, self-reproduction) that disqualify viruses from being considered living (cf. Luisi, 1998; Gánti, 2003, pp. 74–80).

Another approach for estimating the primeval error rate is from the 'bottom-up': finding the maximal accuracy provided by a single copying enzyme. Artificially synthesized RNA polymerases can have an error rate as low as $8.8 \cdot 10^{-3}$ per nucleotide (Wochner *et al.*, 2011). And even this might not be the minimum. *In vitro evolution* is a relatively recent method, providing us with increasingly accurate enzymes.

Table I1. The relation between accuracy and error rate. A low accuracy (e.g. 0.9912) corresponds to a high error rate (e.g. $8.8 \cdot 10^{-3}$). Data from Joyce, 2002; Wochner *et al.*, 2011.

	bacterium	virus	ribozyme	
lowest accuracy	(0.999999)	(0.9999)	0.9912	highest accuracy
highest error rate	10^{-6}	10^{-4}	$8.8 \cdot 10^{-3}$	lowest error rate

Calculated values are in parenthesis.

This leads us to our next question concerning the sustainable genome length. We shall first employ a formulation provided by Eigen (1971), and simplified by Maynard Smith (1983). Let us assume that all possible mutant sequences have the same replication rate (α), beneath that of the master sequence (A). Then the selective advantage (s), calculated as the ratio between these replication rates: $s = A/\alpha$; will be constant. And the correlation between the sustainable genome length (L) and the available replication accuracy (q) will thus be:

$$L < \frac{\ln(s)}{(1 - q)}$$

if we neglect back mutations to the master sequence, and consider the mutations independent of module type and position.

To have a more realistic picture of genome sustainability, we should also consider that enzyme activity depends on three-dimensional structure (Anfinsen, 1973), which is unaffected by a significant proportion (λ) of errors. Takeuchi and colleagues (2005) incorporated these neutral mutations into the above formula:

$$L < \frac{-\ln(s)}{\ln(q + \lambda - q\lambda)}$$

The constants in question were determined using empirically measured data and a predictive scoring of naturally occurring ribozymes: $\lambda \approx 0{,}24$ and $s \approx 350$ (Kun *et al.*, 2005). Applying the above presented accuracy of ribozymes (q = 0.9912), the sustainable genome length was found to be 872 modules (nucleotides). Such length might be capable of storing a sufficiently accurate copying ribozyme, but it is unable to maintain the whole genome of a supposed riboorganisms (Kun *et al.*, 2005).

1.2. *Information integrating systems*

Eigen and Schuster (1977) assumed, that if a single molecule cannot sustain the genome, it might be stored separately on multiple genes. Then, each molecule could be shorter, with its content seamlessly preserved, while their composition would hold the sum of information required for the genome. However, it is not evident that such a composition can be sustained. While once the genome has been partitioned, its sustainability would depend on the coexistence of all genes. They proposed a system, the Hypercycle (Figure I1a), to ensure this function of 'information integration'. It consists of moderately autocatalytic genes (i.e., independent replicators storing different sequences), each of which assist the copying of another gene (through heterocatalysis) according to a circular topology.

The basis of this topology is that each gene assists exactly one other; and it is also assisted by exactly one other; so that their relative stoichiometry remains proportionate, while on the other hand, all the genes benefit indirectly from the assistance of every other. The coexistence of genes, and thus the survival of the Hypercycle, is assured by the fact that the overgrowth of either gene entails the accelerated copying of the others—providing a mechanism for sustained equilibrium. It is also worth noting, that the loss of either gene breaks the cycle of heterocatalytic aid, thus slowing down the replication of all the genes significantly.

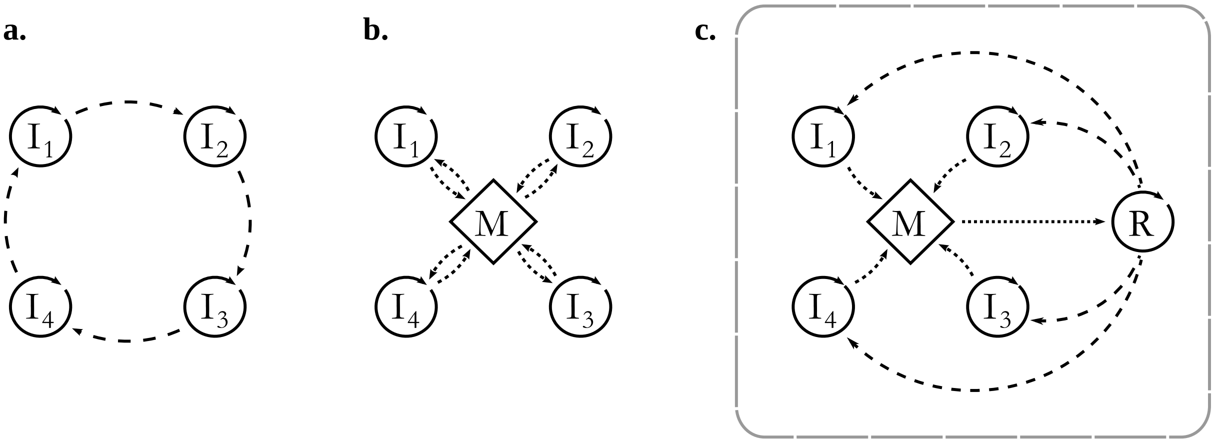


Figure I1. Information integrating model systems. (a) The Hypercycle, **(b)** the Metabolic Replicator, and **(c)** the Stochastic Corrector. Solid arrow: autocatalysis (e.g. through template effect); dashed arrow: heterocatalysis; dotted arrow: metabolic contribution; grey box: membrane. I: information storing molecule, M: metabolism, R: aspecific replicase. Modified from Czárán & Szathmáry, 2000; Scheuring *et al.*, 2003; and Könnyű *et al.*, 2008.

However, there are fundamental problems with the Hypercycle (Maynard Smith, 1979). The development of a new heterocatalytic connection (*shortcut*) leads to the overgrowth of the resulting shorter cycle, and so the loss of the shortcut genes. Thus it is almost impossible to accumulate information in this system: any peripherally joining genes are to be immediately shortcut; the heterocatalytic cycle would have to undergo extensive reorganization to incorporate any new gene. Furthermore, if new genes do join by accepting heterocatalytic aid from a member of the Hypercycle, while not providing any assistance themselves (i.e., being *parasites*), they are able to extract all the 'nutrients' (monomer modules) from the system, thus destroying the Hypercycle—also ending their temporary sudden growth in the process.

Since providing heterocatalytic aid is an altruistic behaviour, whose failure does not lead to immediate negative feedback, there is no punishment for parasitism. Nevertheless, it is possible to introduce such punishment into the system, e.g. by restricting the dispersion of genes (Boerlijst & Hogeweg, 1991): if parasites are unable to leave the Hypercycles they exploited, the damage they cause will eventually have their repercussions ('backlash').

Information loss by shortcut is avoidable, and also the incorporation of new genes is feasible, if genes connect to a central system (e.g. a metabolism) independently of each other, providing their mutual assistance through this system: this is the Metabolic Replicator Model (Figure I1b; Czárán & Szathmáry, 2000). The harm of parasitism is yet again reducible by restricting dispersion: either by attaching the genes to a surface (Szabó *et al.*, 2002); or by packaging them into separate membrane compartments: which is the Stochastic Corrector Model (SCM; Figure I1c; Szathmáry & Demeter, 1987). This latter model considers metabolic genes nourishing a central system, and a replicase gene assisted by this central system, while copying—without specificity—the molecules of the compartment (*protocell*).

Comparing these three information integrating systems, clearly the SCM is the most complex; still, we shall not deem its presumptions excessive. Membrane envelopment is a simple way of impeding dispersion; even early membranes could have been impermeable to polymers (Schrum *et al.*, 2010), protecting against the dilution of genes, and the spread of parasites. Furthermore, it is likely that even genetic monomers (e.g. nucleotides) could not escape from protocells (Yarus, 1999), so the effect of their neighbourship should be negligible. The survival of protocells (and their fitness) within the population must have depended on their harboured genetic composition: the number of metabolic genes contributing to the synthesis of monomers (Unrau & Bartel, 1998; also see Box M1)—we will consider these genes highly essential. And though we have yet to find an aspecific replicase ribozyme that can create its own copies (Bartel és Unrau, 1999; Johnston *et al.*, 2001; Zaher és Unrau, 2007; Wochner *et al.*, 2011), we see no theoretical obstacles to their existence—we consider it a mere matter of (finding the appropriate) chemistry. The fission of the earliest protocells was most likely an unregulated process (cf. Koch, 1985), and so the assortment of the molecular content to the daughter protocells was random.

A central component of the SCM is hierarchical selection. With each gene contributing differently to the central metabolism, and thus the replication of the genes inside the protocell, it is reasonable to assume that the different genetic compositions affect the proliferation of the protocell. The highest rate of genomic growth should correspond to optimal composition(s). We cannot be certain about the nature of such composition(s), since we lack insight even into the functions of the primeval genes. But assuming the existence of such optimum (optima), we can be sure about the detrimental effect that stochastic and misdirected processes can have on protocell growth and proliferation; influencing the sustainability of the genome.

The main difficulties of sustaining the optimal composition in the SCM are "assortment load, and mutation load" (Zintzaras *et al.*, 2010): the drop in fitness due to the random loss of any gene (because of fission), and due to the different replication rates of genes (because of mutation), respectively. Suboptimal compositions undoubtedly lead to information loss beyond a certain point, so we shall presume that assortment and mutation loads set a limit on the sustainable number of genes—and thus the length of maximal genetic information. We term this correlation the *second error threshold* for its similarity with the correlation between the mutation rate and the maximal gene length (Eigen, 1971), which we have referred to as the *first error threshold*.

1.3. *Objectives*

The coexistence of independently replicating genes is known to be supported by the SCM (Szathmáry & Demeter, 1987), thus is it possible that the SCM can sustain a larger genome size, than what would be possible on a single gene (e.g. 872 nt, Kun *et al.*, 2005). Most investigations, however, did not consider more than two (Szathmáry & Demeter, 1987; Grey *et al.*, 1995), or three genes (Zintzaras *et al.*, 2002; 2010; Santos *et al.*, 2004). Fontanari and colleagues (2006) have shown analytically that there is no theoretical maximum to the sustainable gene number: an infinite population size, and an equal replication rate for every gene, can sustain an arbitrary number of genes.

We wished to give a quantitative estimate to the maximum amount of genes that the SCM can sustain within realistic conditions (i.e., finite, and sometimes unequal parameters). We also wanted to investigate the dynamics of the SCM on different hierarchical levels: that of the genes, the protocells, and the population. By “carry[ing] out a thorough analysis of the space of parameters that determine the evolution [of genes and protocells]”, to remedy this deficiency in our knowledge, pointed out by Fontanari and colleagues (2006). We could not accept the conclusion of Silvestre & Fontanari (2007), that the whole class of package models “should be discarded as possible solutions to the prebiotic information crisis” as “the information gain derived from the coexistence of the distinct templates is not significant”. We set out to determine the second error threshold.

CHAPTER 2

MODEL

We consider the dynamics of *primeval metabolic systems* (PMS-s) inside a population of *protocells*. Each protocell harbours a PMS composed of an array of *ribozymes* that perform various enzymatic functions. Ribozymes performing the same function are *copies* of the same *gene*—different genes serve complementary roles in the PMS. Presumably, one of the genes function as a replicase (e.g. a general polimerase, Wochner *et al.*, 2011), catalyzing the template-based copying of not only its own but the other genes as well, using up *metabolites* (e.g. nucleotides) in the process. There are several ways for the other genes to contribute to the replication, thus preserving the PMS analogously (see Box M1). Consequently, we will not distinguish the replicase from the other ribozymes. The structure of this population is summarized in Figure M1.

The dynamics of the population is governed by three central processes of different timescales: the *replication* of ribozymes (which is also the growth of protocells), the *fission* of protocells, and the *recruitment* of genes with novel enzymatic function (Figure M2); *mutation* is an integral part of both replication and recruitment. Notation is presented in Table M1.

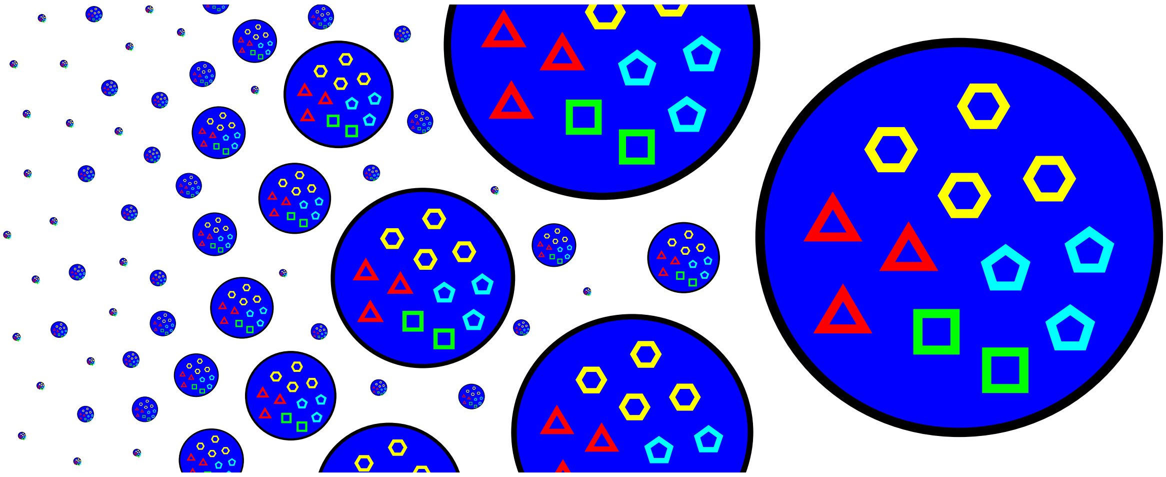


Figure M1. A schematic illustration of the protocell population. Circles represent protocells, with their boundaries (black) dividing space into separate compartments (deep blue). Polygons represent ribozymes, those of identical shape and colour are copies of the same gene.

Box M1. Realistic functions for ribozymes in assisting a PMS.

(*1*) Ensuring the appropriate structure of metabolites: aiding their stability (Mansy & Szostak, 2008); preventing clutter (and enzyme promiscuity, cf. Szilágyi *et al.*, 2012) from disrupting synthesis (Joyce, 2002 b); and maybe even initiating chiral homogeneity (Garay, 2011).

(*2*) Speeding up synthesis: catalysing the substrate-to-metabolite reaction by their own (Jeffares *et al.*, 1998), by binding coenzymes (Szathmáry, 1993), or by forming autocatalytic biochemical cycles (which, by the way, are analogous to enzymes, Gánti, 2003, pp. 22, 26); providing reducing power through redox reactions (Wächtershäuser, 1990) or chemical energy through catabolism (Joyce, 2002 b, pp. 218–219).

(*3*) Increasing the yield of metabolites: protecting the integrity of the protocell membrane (Bartel & Unrau, 1999, p. M12); retarding the efflux of metabolites through this membrane (Khvorova *et al.*, 1999; Szathmáry, 2007); stimulating the influx of substrates (e.g. by generating a transmembrane pH gradient, Chen & Szostak, 2004) or increasing the specificity of the influx (Sacerdote & Szostak, 2005).

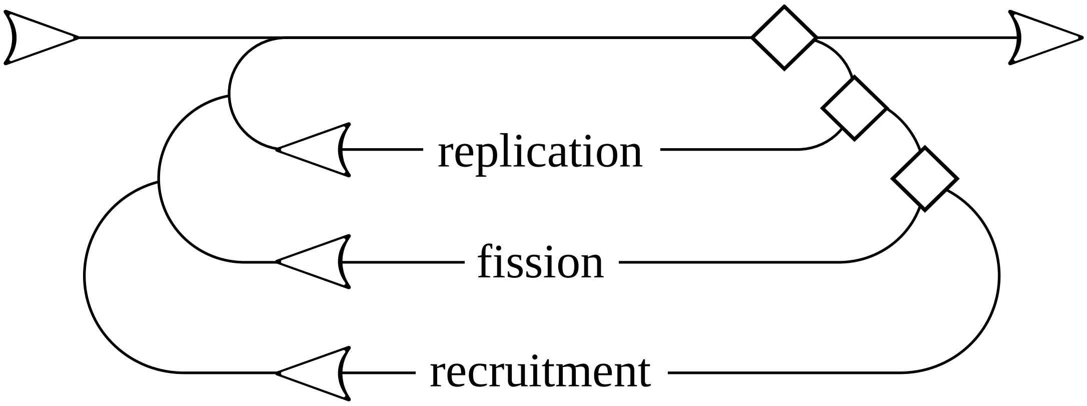


Figure M2. An overview of the hierarchical nature of processes, taking place inside the protocells and within the population of protocells. The most repeated cycle is the smallest (replication), while the largest (recruitment) has the least iterations. A general measure of time passed is calculated from the repeats of the mid cycle (fission). Arrowheads indicate the direction of processes, diamonds are conditional branching points in the algorithm.

Table M1. Parameters of the model.

agent	attribute	notation	variability*
ribozyme	function (gene)		+ (4)
	replicase affinity	a	+ (2)
	enzymatic affinity	b	−
	(hosting PMS)		+ (3)
PMS	size	v	+ (1, 3)
	redundancy	c_i, $i \in [1, \tau]$	+ (1, 3, 4)
	activity	R	+ (1, 3, 4)
	(held ribozymes)		+ (1, 3)
population, environment	protocells	N	−
	critical protocell size	$v_{\max}$	−
	genes available	τ	−
	weighting exponent	ε	−
	background activity	b_0	−
	cost of losing a gene	C	−
	mutation rate	f	−
	variance of severity	$p = \sigma^2$	−
	generations (runtime)	g	−
	interval of new genes	h	

* May change at (1) replication, (2) mutation, (3) fission, and (4) recruitment.

2.1. *Replication*

The replication of ribozymes depend on (1) the *activity* (R; see Box M2 for details) of their hosting PMS, which is their protocellular milieu, and (2) their own *replicase affinity* (a).

An active PMS requires (1a) an abundance of ribozymes in (1b) a proper composition of enzymatic functions. We presumed the maximal abundance of ribozymes ($v_{\max}$) to be an environmental constant. We conferred the optimal composition ($Q_{\max}$) on the PMS having an even enzymatic contribution for all the available functions—this corresponds in our study to having the same amount of copies of every gene. To underscore the essentiality of the functions, we assigned a steep cost to losing any gene (a relative measure of b and b_0, see Box M3); if there is no background activity ($b_0 = 0$) every gene is truly essential (cf. Szathmáry & Demeter, 1987, p. 473). In compliance with these specifications, we have devised a 'fitness' function to quantify the activity of PMS-s (Box M2). We are confident that our model retains generality despite the arbitrariness of some of these distinctions (cf. Szathmáry & Demeter, 1987, p. 479).

The initial conditions of the population is an optimal composition of all present genes (τ_0; or if unchanged, τ), and a medium protocell size ($v_0 = v_{\max} / 2$), for every protocell. We ignore the dynamics of the population until this initial orderliness completely disappears.

While the activity of the PMS focuses on accumulating metabolites inside the protocell—a cooperative venture of the various ribozymes—the replicase affinity of a ribozyme is the ability to exploit these metabolites for its own replication. Replication occurs in a stochastic fashion: the PMS and then the contained ribozyme is chosen randomly, but proportionately to the activity and replicase affinity, respectively. During replication the ribozyme may be subject to mutation when its replicase affinity can change. First, we determine whether mutation occurs or not according to the frequency of mutations (f), then we set its severity—we pick a random number from a discretized normal distribution of zero mean and a given ($p = \sigma^2$) variance. This will lead to ribozymes having different replicase affinities which undermines the deterministic coexistence of genes.

Box M2. An algebraic formulation of PMS activity (or 'protocell fitness').

The activity of the PMS (colloquially, the fitness) of protocell *i* is

$$R_i = P_i \cdot Q_i^{\varepsilon}$$

where P_i is the quantitative component ('quantity'), Q_i is the qualitative component ('quality')—both values between 0 and 1—and ε is a weighting exponent. The quantity is

$$P_i = \frac{v_i}{v_{\max}}$$

where v_i is the size of protocell *i* (i.e., the abundance of ribozymes inside), and $v_{\max}$ is the maximal protocell size, an environmental constant. The quality is

$$Q_i = \left(\frac{G_i}{A_i}\right)^{\tau} \text{ with } A_i = \frac{1}{\tau}\sum_{j=1}^{\tau} B_{ij} \text{ and } G_i = \sqrt[\tau]{\prod_{j=1}^{\tau} B_{ij}}$$

where B_{ij} is the enzymatic contribution of gene *j* in protocell *i*, τ is the number of genes, A_i is the arithmetic mean, and G_i is the geometric mean of B_{ij}. According to the AM–GM inequality $A_i \geq G_i$. The enzymatic contribution of gene *j* is

$$B_{ij} = \sum_{k=1}^{c_{ij}} b_{ijk} + b_0$$

where c_{ij} is the redundancy (i.e., the number of copies) of gene *j* in protocell *i*, b_{ijk} is the enzymatic affinity of ribozyme *k* of gene *j*, and b_0 is the background activity, an environmental constant independent of gene (i.e., considered equal for all genes).

Thus the activity of the PMS of protocell *i* in its extended form is

$$R_i = \frac{v_i}{v_{\max}} \left(\frac{\sqrt[\tau]{\prod_{j=1}^{\tau}\left(\sum_{k=1}^{c_{ij}} b_{ijk} + b_0\right)}}{\frac{1}{\tau}\sum_{j=1}^{\tau}\left(\sum_{k=1}^{c_{ij}} b_{ijk} + b_0\right)} \right)^{\tau \cdot \varepsilon}$$

Box M3. The 'fitness cost' of losing a gene.

When a gene is lost from a PMS, the enzymatic contribution of that function (*B*) will decrease to the level of the background activity (b_0). What makes a difference is the enzymatic affinity of the last copy (*b*). Thus the cost (*C*) of losing the gene in terms of enzymatic contribution is:

$$C = \frac{B+}{B-} = \frac{b_0 + b}{b_0}$$

Losing a gene also has a detrimental effect on the activity of the PMS (*R*, see Box M2.). The change of the geometric mean follows that of the enzymatic contribution ($G+^{\tau} = G-^{\tau} \cdot C$), while the arithmetic mean and the quantitative component does not change significantly ($P+ \approx P-$ and $A+ \approx A-$). So the overall effect on the activity of the PMS (the 'fitness cost') would depend on the weighting exponent:

$$\frac{R+}{R-} = \frac{P+}{P-}\left(\frac{G+}{G-} \cdot \frac{A-}{A+}\right)^{\tau \cdot \varepsilon} \approx \left(\frac{G+^{\tau}}{G-^{\tau}}\right)^{\varepsilon} = \left(\frac{B+}{B-}\right)^{\varepsilon} = \left(\frac{b_0 + b}{b_0}\right)^{\varepsilon}$$

Now, we would prefer the fitness cost of losing a gene to be independent of the exponent ($R+ \approx R- \cdot C$) so that we could compare the result of simulations with different weighting exponents. For the fitness cost of losing a gene to remain equal, we decided to adjust the background activity to the exponent: b_0 will pertain to $\varepsilon = 1$, the adjusted value b_0' to $\varepsilon \neq 1$. Also, the background activity will be relative to the unit enzymatic affinity ($b = 1$).

$$\frac{b_0 + 1}{b_0} = \left(\frac{b_0' + 1}{b_0'}\right)^{\varepsilon}$$

$$\log\left(1 + \frac{1}{b_0}\right) = \varepsilon \cdot \log\left(1 + \frac{1}{b_0'}\right)$$

For $\varepsilon \neq 1$ the adjusted value of the background activity should be:

$$b_0' = \frac{1}{\sqrt[\varepsilon]{1 + \frac{1}{b_0}} - 1}$$

Note that if $b_0 = 0$ the above derivation holds no meaning. Each gene is then essential, and the fitness cost of losing a gene is infinite.

2.2. *Fission and recruitment*

The protocell grows (i.e., acquires ribozymes) in the above manner. When these ribozymes reach a certain number, and thus the protocell a certain size, the protocell undergoes binary fission—e.g. because the increasing surface-volume ratio causes ever larger invaginations (Koch, 1985); but the exact mechanism may also be different (cf. Zhu & Szostak, 2009). The content of the *parent* protocell is assorted into two *offspring* protocells stochastically: each ribozyme has an equal chance of getting into either offspring. The size of the offspring protocells—and also the abundance of copies for each gene—will have a binomial distribution.

We consider the carrying capacity of the environment by maintaining a constant population size (N): whenever a protocell undergoes fission, a random protocell will have to die. Each protocell has an equal probability of dying—if the parent protocell is chosen, one of its offspring will perish. The dynamics of the protocells thus follows a Moran process.

To be able to compare the relative speeds of different features of the dynamics, we have introduced a general measure of time: a *generation* is N number of fissions—i.e., the time during which on average each protocell undergoes fission once.

Once in a while, the mutation of ribozymes can lead to gain of a novel function. When a new gene thus appears the PMS might recruit this function—i.e., incorporating the new ribozyme into the existing metabolic network—, which might confer a selective advantage upon the hosting PMS (Horowitz, 1945; Jensen, 1976; for a recent review see Emiliani *et al.*, 2010, p. 41–49). We consider the incidence of these macro-evolutionarily important—i.e., successfully incorporated gain-of-function—mutations to be proportionate to the overall amount of mutations, which in turns is proportionate to the frequency of replications of the ribozymes. So every now and then, after a specified interval of time (h) passes, we decide which ribozyme to endow with a new function. We do this by choosing a PMS and a contained ribozyme the same way we do it for replication—proportionately to the activity and replicase affinity, respectively. The introduced function is always one that is previously absent from the PMS.

2.3. *Scenarios*

According to our interest in the dynamics, we have outlined three main avenues of inquiry: (1) the *equilibrium* scenario explores the necessary conditions for a stochastic coexistence of genes; (2) the *synchrony* scenario investigates the effects of mutation on asynchronous replication; and (3) the *assembly* scenario holds the key to understanding the growing complexity of early metabolisms.

These scenarios differ only in the range of their parameters. However, some parameter values result in complete processes being omitted. If $f = 0$, mutation never occurs, so the replicase affinitiy of every gene (a) is constant, which is the case in the equilibrium scenario. Also, we did not examine how PMS-s lose genes in the equilibrium and the synchrony scenarios, so we had $h \geq g$, $b_0 = 0$ and $C = \infty$. And since we had both mutation and recruitment in the assembly scenario, it was unavoidable to have $f \geq 0$ and $b_0 > 0$ (see Table M2).

We believe that the best approximation of the primeval dynamics is offered by the assembly scenario. On the other hand, the equilibrium and synchrony scenarios contribute to our understanding by allowing us to separate mechanisms in the coexistence of early replicators.

Table M2. The pursued scenarios and their distinctive features.

Name	Processes involved	Fixed attributes	Explored phenomena
Equilibrium	replication, fission	$f = 0$ $h \geq g$, $b_0 = 0$, $C = \infty$	hierarchical selection, stochastic correction of assortment load
Synchrony	replication, mutation & fission	$f \geq 0$ $h < g$, $b_0 > 0$, $C < \infty$	synchronization of replicase affinities, competitive exclusion
Assembly	replication, mutation, fission & recruitment	$f \geq 0$ $h < g$, $b_0 > 0$, $C < \infty$	sustainable diversity, decreasing redundancy

CHAPTER 3

RESULTS

This review will follow the order prescribed by the three scenarios. We will profess our motivation for each inquiry; then refer the reader to the respective graphs; a brief description of the findings will ensue. For a summary of the investigated parameter combinations consult Table R1.

Table R1. An overview of the examined set of parameters.

fig.	g	N	$v_{\max}$	τ	ε	d_0	D	f	p	b_0	I
1.	10000	↔	1000	5	0.3	0	E	0	-	0	-
2.	100	1000	6480	↔	↔	0	E	0	-	0	-
3.	100	↔	2160	↔	0.3	0	E	0	-	0	-
4.	100	1000	↔	↔	0.3	0	E	0	-	0	-
5.	100	1000	25920	↔	0.3	↔	↔	0	-	0	-
6.	100	1000	↔	2	0.3	0.1	↔	0	-	0	-
7.	100	100	2000	5	0.3	0	E	0.01	↔	0	-
8.	N/A	5000	1000	5	N/A	0.68	L	N/A	N/A	0	-
9.	100	100	↔	↔	0.3	0	E	0.01	1	0.1	1
10. a–b	↔	100	1000	50	0.3	0	E	0.01	1	0.1	↔
c–f	200	100	5000	50	0.3	0	E	↔	↔	0.1	10
11.	1000	100	5000	50	0.3	0	E	0	-	0.1	50

The symbol ↔ indicates that several values of the respective parameter were investigated. *D* denotes the distribution of replicase affinities (E: equal, L: '1 lower'). *I* denotes the number of genes introduced, of which recruitment is possible. N/A: incommensurable value.

The autocatalytic growth of the ribozymes is exponential: in a deterministic scenario any difference in the initial concentration of the genes, or their replication rates (replicase affinities), will lead towards competitive exclusion inside the protocell. However, stochasticity of replication, coupled with group selection on the level of protocells, is known to be capable of impeding this course. We show that our model has a dynamic equilibrium state in which several gene can stably coexist (Figure R1). The manifest fluctuation of the average fitness (of the protocell population) in equilibrium indicates the even larger fluctuation of its components—the size and the quality of the protocells; their quality is the evenness of their genetic composition (Figure R1a–b; and see Box M2). The effect of the population size is also observable: the fewer the protocells are the larger the variance becomes.

To ensure that our results describe equilibrium states of the system—and knowing that the initial conditions of our simulations are ‘inordinately ordered’—our curiosity took aim at the approach of this equilibrium. We found that the equilibrium is invariably reached during the first 50 generations (Figure R1c–d). We thus decided to run subsequent simulations for a 100 generations; and to calculate equilibrium properties (e.g. the fitness of protocells) from the average of the final 50 generations.

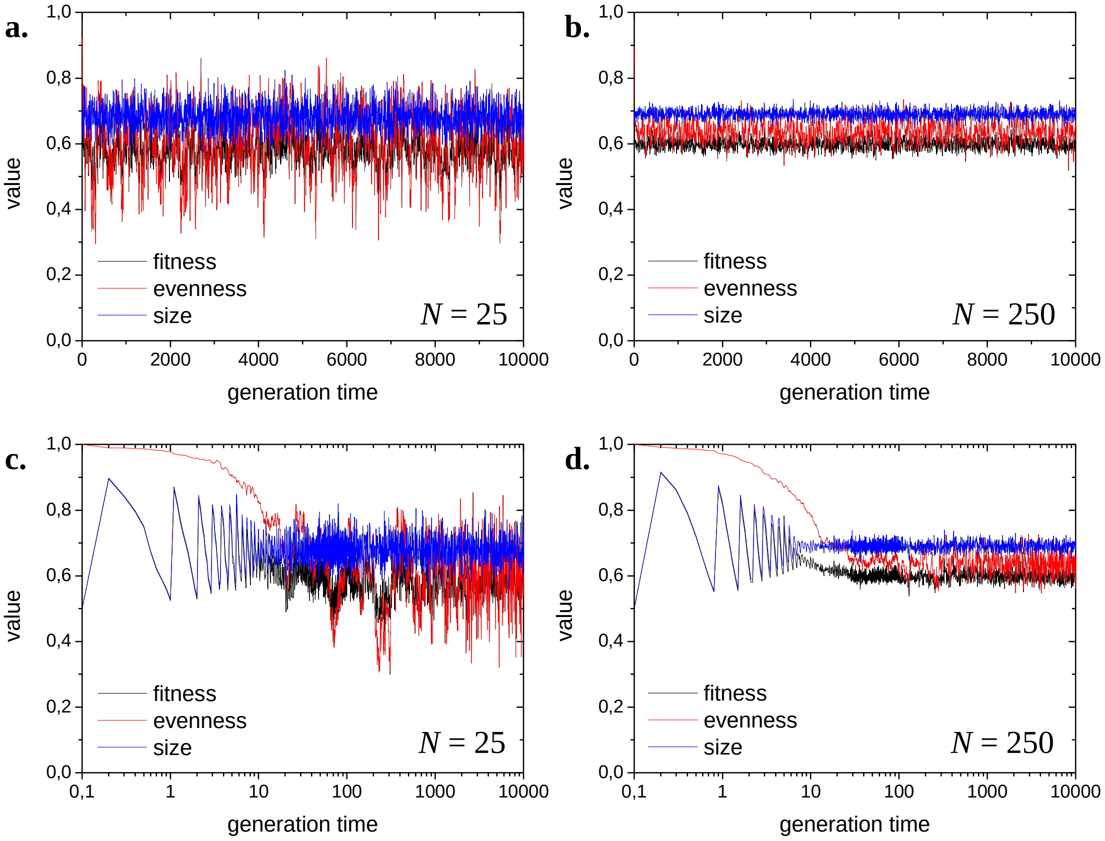


Figure R1. Dynamic equilibrium. The upper panel **(a–b)** shows the dependence of the amplitude of fluctuation on the population size (*N*). The lower panel **(c–d)** reveals the rapid onset of the equilibrium. Note the different scales. Parameters: g = 10000, c_0 = 100, τ = 5, ε = 0.3.

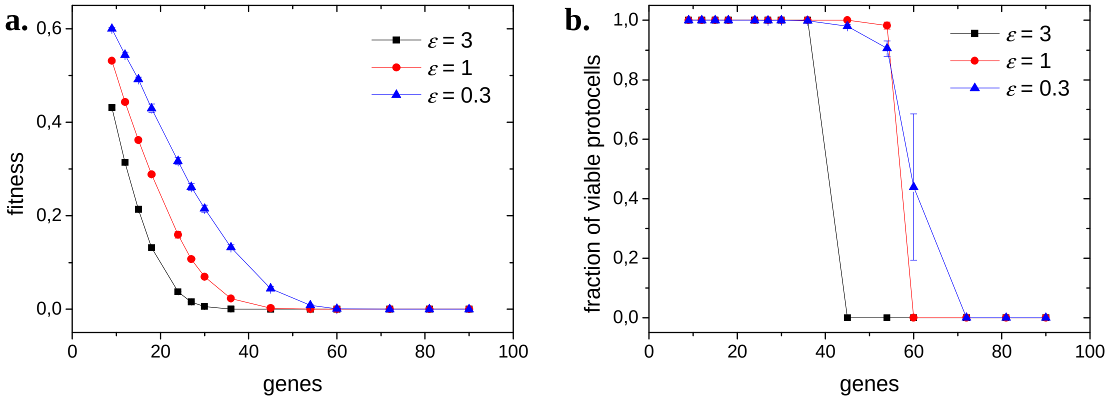


Figure R2. The selection regime's effect on genetic integration. A focus on the quality component ($\varepsilon < 1$) results in a less steeply decreasing fitness **(a)** and a larger sustainable genome size **(a–b)**. Each symbol marks the average of 6 repeats, totalling 252 simulations. Horizontal bars show standard deviation. Parameters: $g = 100$, $N = 1000$, $v_0 = 3240$.

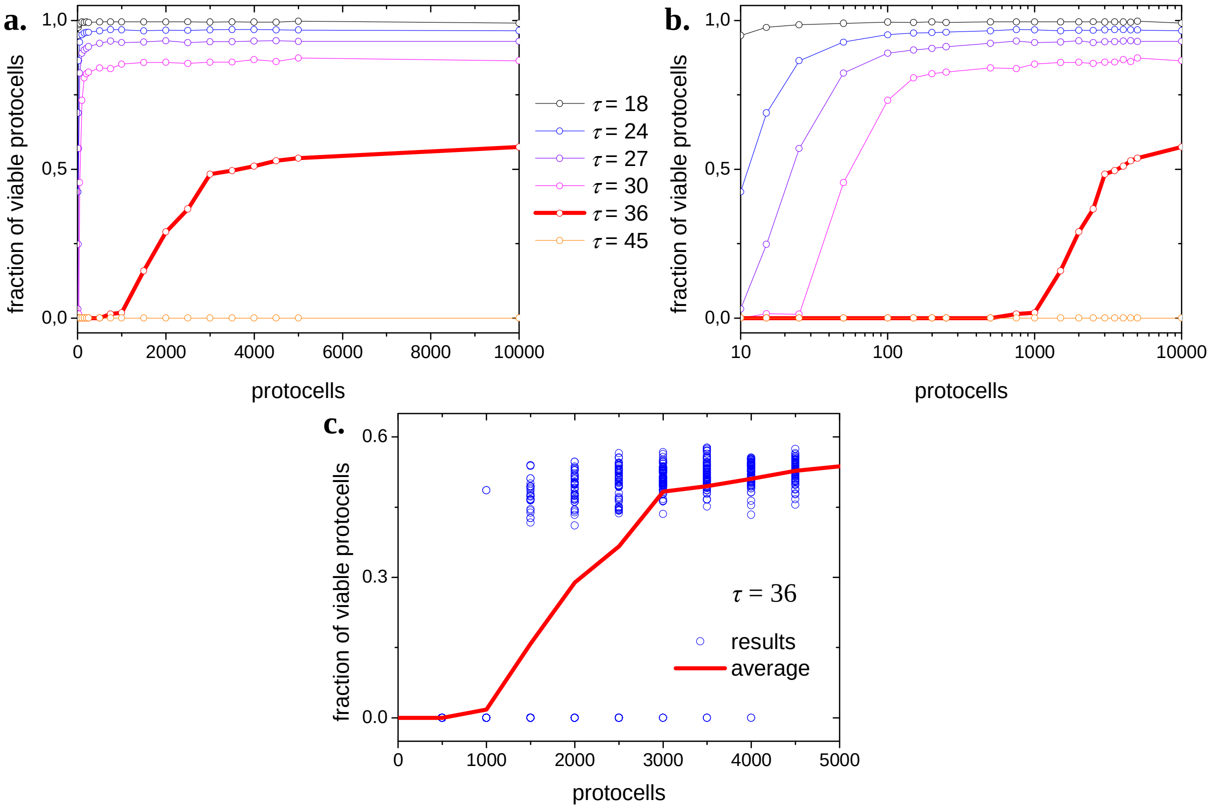


Figure R3. The minor effect of population size. (a–b) A larger population can indeed sustain a larger genome but the correlation is disproportionate. Symbols show the mean of at least 7 repeats ($\tau = 36$ and $200 < N < 5000$ have 77 repeats), totalling 1729 simulations. **(c)** When depicting each result separately, they form two separate clusters corresponding to survival and extinction. Note the different scales. Parameters: $g = 100$, $v_0 = 1080$, $\varepsilon = 0.3$.

3.1. *Equilibrium*

In exploring the optimal conditions for our system to uphold genetic diversity, we considered the effect of the selection regime (Figure R2), the carrying capacity of the environment (Figure R3), and the critical protocell size, where fission occurs (Figure R4). These are the main results of the equilibrium scenario.

The fitness function contains a weighting exponent that describes how the selection regime favours the quality of the protocells. If selection is 'fair' (i.e. $\varepsilon = 1$) then the growth rate of equal-sized protocells are proportionate to their quality. If selection is 'helping' (i.e. $\varepsilon > 1$) then medium-quality protocells will have a higher growth rate: protocells of suboptimal composition will have a realistic chance to proliferate and result in higher quality offspring (cf. stochastic correction). On the other hand, if selection is 'vigilant' (i.e. $\varepsilon < 1$) then medium-quality protocells will have a subproportional growth rate: practically, only the best compositions will proliferate. We found that a 'vigilant' regime—when selection on the quality of protocells is the strongest—results in the highest sustainable genome size (Figure R2). Our following investigations will therefore pertain to a 'vigilant' selection regime
($\varepsilon = 0.3$).

Population size has only a minor effect (Figure R3). Protocell viability curves show saturation with the increase of the population size. Also it is known that an infinite population size supports an arbitrary number of genes (Fontanari *et al.*, 2006). But while the population size increases by three orders of magnitude (a factor of 1000) the sustainable gene number grows only from $27 \leq \tau < 30$ to $36 \leq \tau < 45$ (a factor of approximately 1.5). We have mostly used $N = 100$ or $N = 1000$ in our simulations. Higher population sizes could result in more genes coexisting, thus our results are conservative estimates of the maximal sustainable genome size. It is interesting to note that although the average viability of the protocells increases gradually, equilibrium viabilities show a discontinuity between high and low values: either many a protocell sustains the genome, or none of them does so (Figure R3c).

The effect of the critical protocell size, however, is spectacular (Figure R4). At their respective positions on the depicted parameter space (genes × copies) we have marked simulation results: either the integration or the loss of genetic information. We found that the subspaces corresponding to these two kind of results divide along a perfectly fitting linear threshold. Its slope is 5/3 (Figure R4a–b); we have yet to find an express meaning to this value. Linear correlation between the genome size and the (maximal) amount of copies inside the minimal sustaining protocell ($\tau_{\text{sust}} \cdot 5/3 \leq c_{\max}$) corresponds to a root-like correlation between the sustainable genome and the minimal protocell size ($\tau_{\text{sust}} \leq \sqrt{v_{\max} \cdot 3/5}$, since $v_{\max} / \tau = c_{\max}$) (Figure R4c).

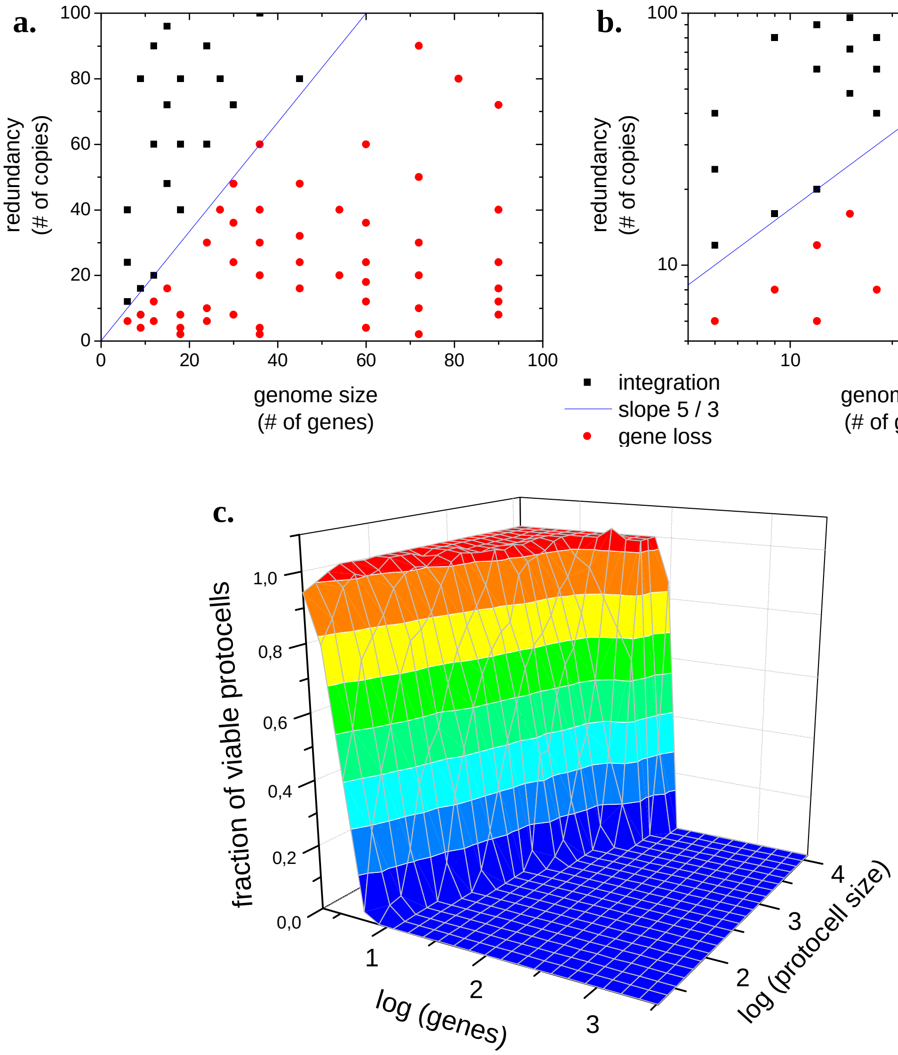


Figure R4. The direct proportionality between redundancy and sustainable genome size. The upper panel **(a–b)** shows the parameter space of the number of copies versus genes. Each dot marks an examined combination where the 6–8 repeats showed unanimous integration or gene loss. The thus divided parameter space has a linear threshold with a slope of 5/3. Note the different scales. The lower panel **(c)** is an interpolation of results, from 1031 simulations, reinforcing this observation on a much larger extent of the parameter space. Parameters: $g = 100$, $N = 1000$, $\varepsilon = 0.3$.

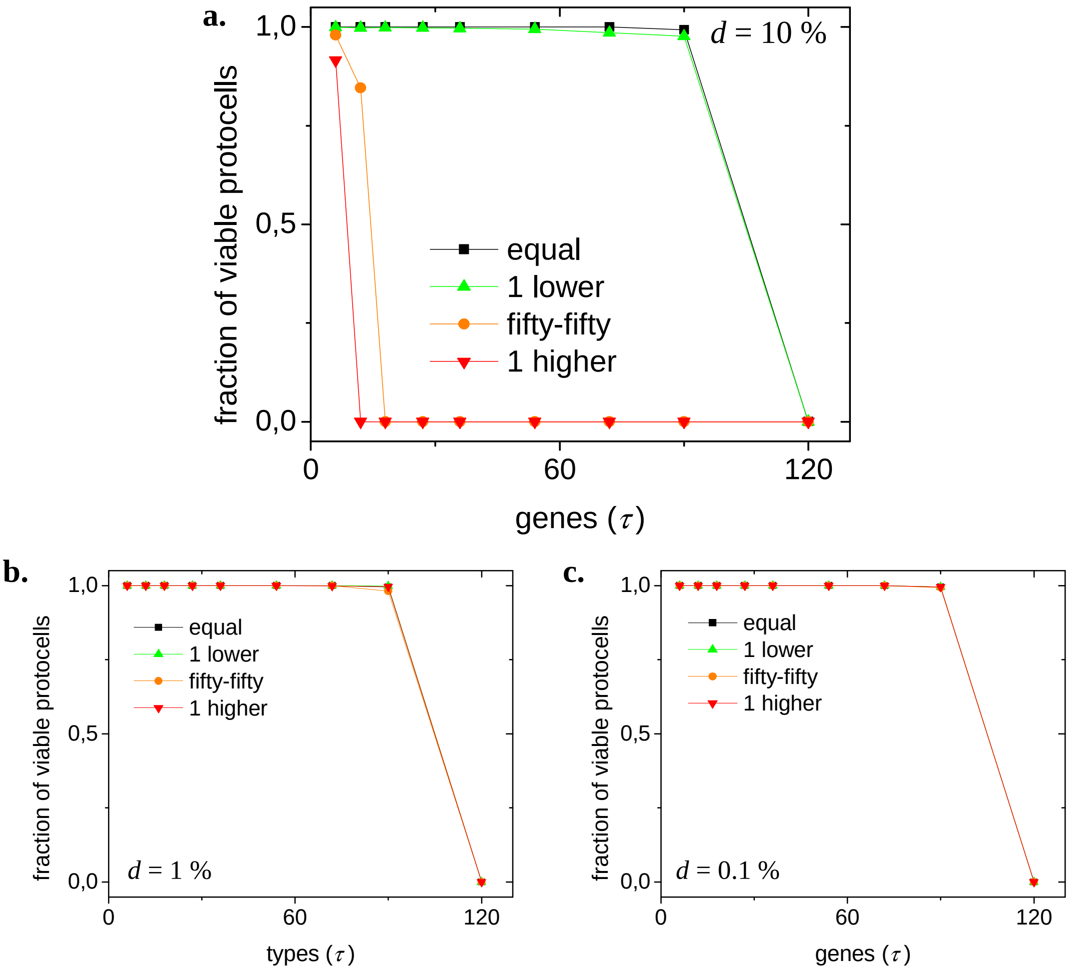


Figure R5. Enduring asynchronous replication. The effect of different replicase affinity distributions on the sustainable genome size. The higher affinity is H = 1, while the lower differs among the panels: **(a)** L_a = 0.9, **(b)** L_b = 0.99, **(c)** L_c = 0.999. Compared distributions are: ‘equal’ ($a_i = H, i \in [1; \tau]$), ‘1 lower’ ($a_1 = L$ and $a_i = H, i \in [2; \tau]$), ‘fifty-fifty’ ($a_i = L$ and $a_j = H, i \in [1; [[\tau/2]]], j \in [[[\tau/2]] + 1; \tau]$), and ‘1 higher’ ($a_i = H$ and $a_\tau = L, i \in [1; \tau - 1]$). The result of 566 simulations are shown. Parameters: g = 100, N = 1000, v_0 = 12960, ε = 0.3.

3.2. *Synchrony*

As a transition towards the synchrony scenario we have examined the sustainability of asynchronous replication (Figures R5). Genes with a higher replicase affinity (H) tend to outreplicate other genes of lower affinities (L), while the composition becomes increasingly uneven. We compared different distributions of unequal affinities: we found that if only a single gene had a unique replicase affinity it was not indifferent whether that was higher than average, or lower ('1 higher' and '1 lower' distributions, respectively). When half of the genes had a higher, and the other half a lower affinity ('fifty-fifty' distribution), it behaved similarly to the '1 higher' case. An affinity difference of 10% (i.e., $d = (H - L) / H = 0.1$) resulted in the equal and '1 lower' distributions sustaining a sizeable genome ($90 \leq \tau < 120$); while the 'fifty-fifty' and '1 higher' distributions even failed to uphold one fifth of that genome ($\tau < 18$) (Figure R5a). We also found, however, that this difference in genome sustainability disappears if the affinity difference is lower ($d \leq 0.01$) (Figure R5b–c).

Furthermore, we found that the connection between asynchronous replication (i.e., different replicase affinities, thus rates) and uneven genetic composition is dependent on the critical protocell size (Figure R6). While a larger protocell promotes a more even composition for equal replicase affinities, for an unequal distribution ($d = 0.1$) this guarantees instead a more adverse composition. There seems to be an optimal protocell size for asynchronous replication where despite the differences in replicase affinity a mostly even composition is sustained.

So far, we have disregarded mutation. In our ensuing studies we have presumed a mutation frequency of one per a hundred replications ($f = 0.01$). As part of the synchrony scenario we have investigated how mutations of different severity impact the sustainability of the genome (Figure R7). We have found that if the variance of severity corresponds to that of the standard normal distribution (i.e., $p = \sigma^2 = 1$) then the composition of the protocells (characterized by their average quality) remains comparatively the same as in mutation-less simulations. However, a higher severity ($p \geq 4$) results in gene loss.

We could not refrain from presenting here an earlier result from a similar model (Figure R8). It demonstrates that initially different ($d_0 = 0.68$) replicase affinities of a '1 lower' distribution are capable of reaching an almost equal distribution through random mutations—an event we term synchronization.

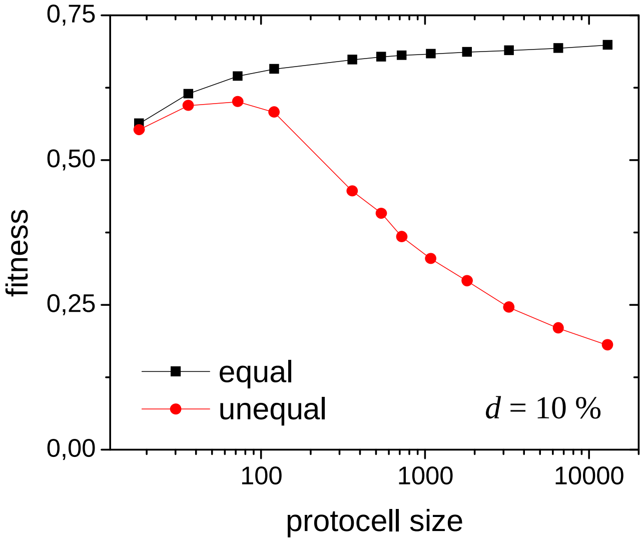


Figure R6. The effect of redundancy on asynchronous replication. In this minimal system of two genes ($\tau = 2$) we compared replicase affinities of equal distribution ($a_1 = a_2 = 1$) with those of an unequal distribution ($a_1 = 0.9$, $a_2 = 1$). The two lines show a divergent trend with the growth of protocell size. Note the logarithmic scale. Each symbol marks the result of a single run, totalling 24 simulations. Parameters: $g = 100$, $N = 1000$, $\tau = 2$, $\varepsilon = 0.3$.

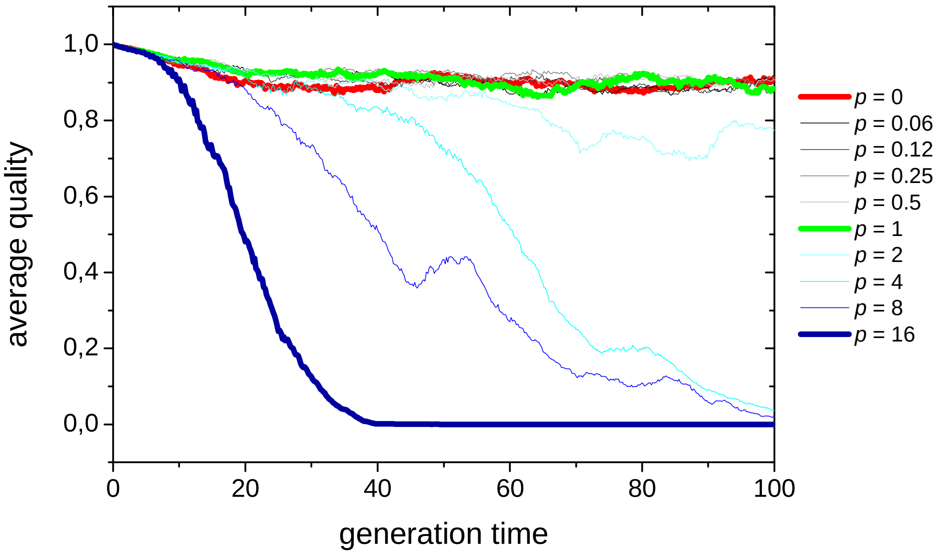


Figure R7. The effect of error severity on the endurance of the population. Quality corresponds to the evenness of copies. The other component of fitness, quantity, remains in equilibrium—plotting the fitness would only mean more noise. Parameters: $g = 100$, $N = 100$, $c_0 = 200$, $\tau = 5$, $\varepsilon = 0.3$, $f = 0.01$, $a_i = 1$, $i \in [1; \tau]$.

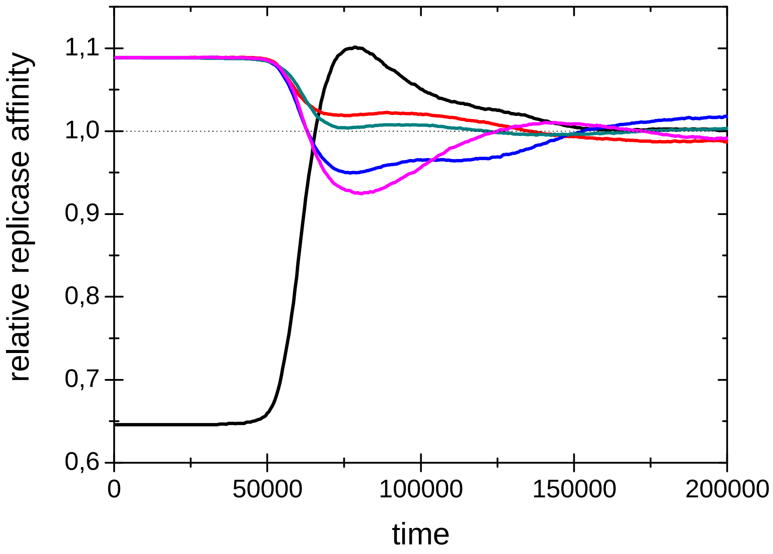


Figure R8. Synchronization of affinities through chance mutation. Each line corresponds to a gene’s mean replicase affinity, normalized by their overall mean. Note that the data shown here was created with a similar but different model than the one described in the text. Parameters: N = 5000, v_0 = 500, τ = 5, a_1 = 1, and a_i = 1.68, where $i \in [2; \tau]$.

3.3. *Assembly*

Finally, we examined the dynamics of genes being recruited to, and lost from, PMS-s. While until now, the background activity of metabolic functions was considered nonexistent ($b_0 = 0$), and thus the 'fitness cost' of losing a gene infinite ($C = \infty$, see Box M3), in the assembly scenario we decided along more realistic values: the background activity will be an order of magnitude smaller than the enzymatic affinity of a single ribozyme ($b_0 = 0.1$), and the 'fitness cost' will be finite ($C = (0.1 + 1) / 0.1 = 11$). In this scenario we explored the rate at which a metabolism can recruit or lose a single gene (Figure R9); the role of the evolution of new functions in sustaining the genome size (Figure R10); and the possible extent of genome assembly (Figure R11).

We found that both losing and recruiting a gene can unfold in a matter of generations (Figure R9). If a larger genome size is sustainable, and a new gene appears somewhere in the population, this new gene will invigorate its hosting PMS—e.g. through the extra metabolites it helps to produce—to proliferate and overgrow other PMS-s of lower quality. It is more exciting to reflect on how losing a gene can be similarly rapid. And why is it that just before its extinction, a few protocell will hold unto it for a little while? As an answer is not evident from these results, we will mention our strongest hypothesis in the discussion. The rate of losing further genes eventually decelerates, and the genome size becomes sustainable.

One might imagine that a PMS conducive to the evolution of new functions, thus facing the opportunity of recruiting genes more often, will have a higher equilibrium genome size than other, less conducive PMS-s. We found that there is no such connection (Figure R10); rather that there is an extrinsic limit on the sustainable genome size, which every PMS is forced to respect. The only difference it makes to receive a persistent stream of new functions is a smoother approach towards the limit (Figure R10b). The importance of gain-of-function mutations is more pronounced when converging to this limit from below (i.e., starting from smaller genome size) (Figure R10c–d): they are an obvious prerequisite to recruitment. Interestingly, the average evenness decreases as the newly recruited gene invades the population (Figure R10e–f). A compensatory increase after integration signals the system's capacity for further recruitment. The assembly of a multigene metabolism through sustained recruitment and integration is thus attainable (Figure R11).

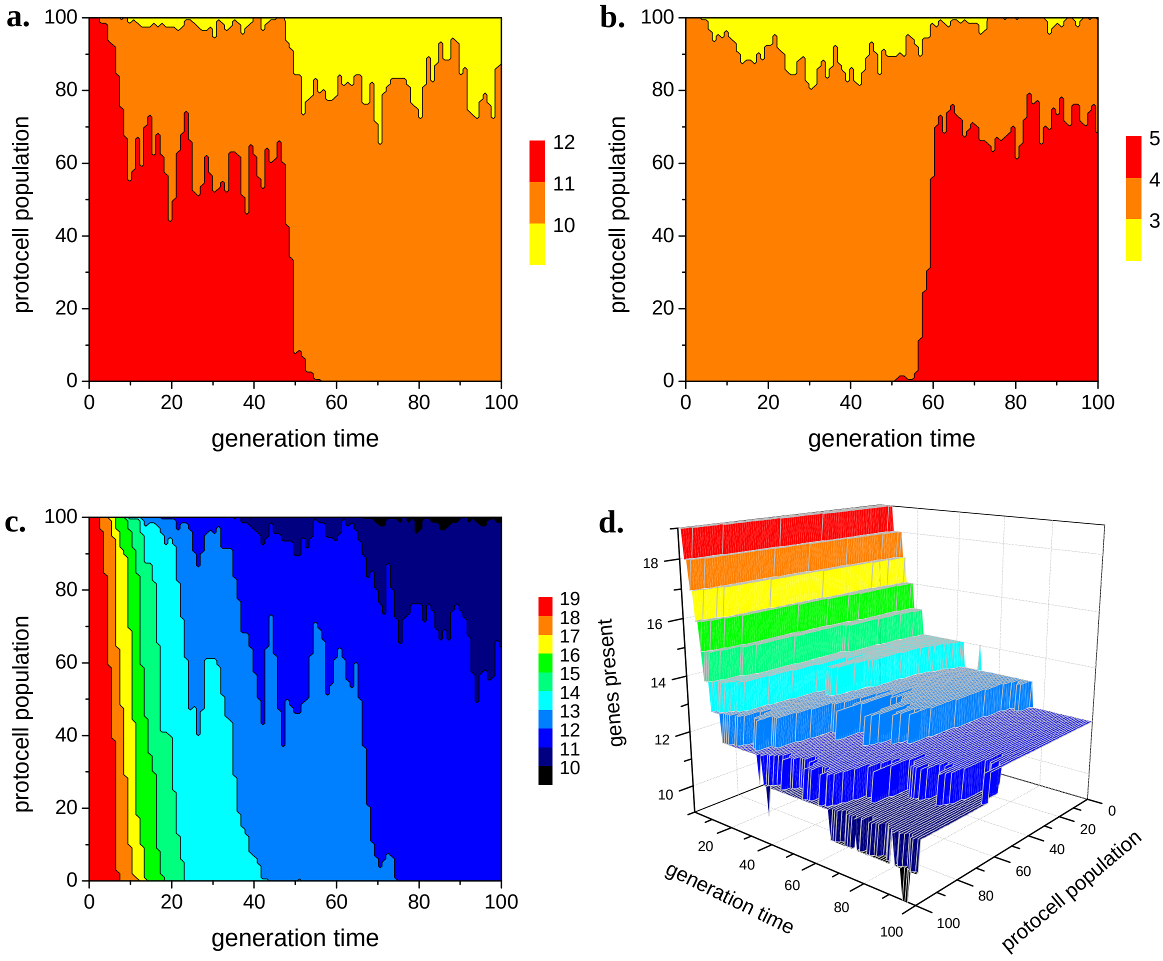


Figure R9. The rapidity of gene loss and recruitment. It is surprisingly quick to **(a)** lose or **(b)** recruit a gene. However, **(c–d)** as the genome turns more sustainable, the rate of losing more genes slows down. Note that the protocells' genome sizes are sequenced in monotonic order. Parameters: g = 100, N = 100, c_0 = 10, $\tau_0 = \tau - 1$, ε = 0.3, d_0 = 0, f = 0.01, p = 1, b_0 = 0.1, I = 1. **(a)** $v_{\max}$ = 260, τ = 13. **(b)** $v_{\max}$ = 100, τ = 5. **(c–d)** $v_{\max}$ = 400, τ = 20.

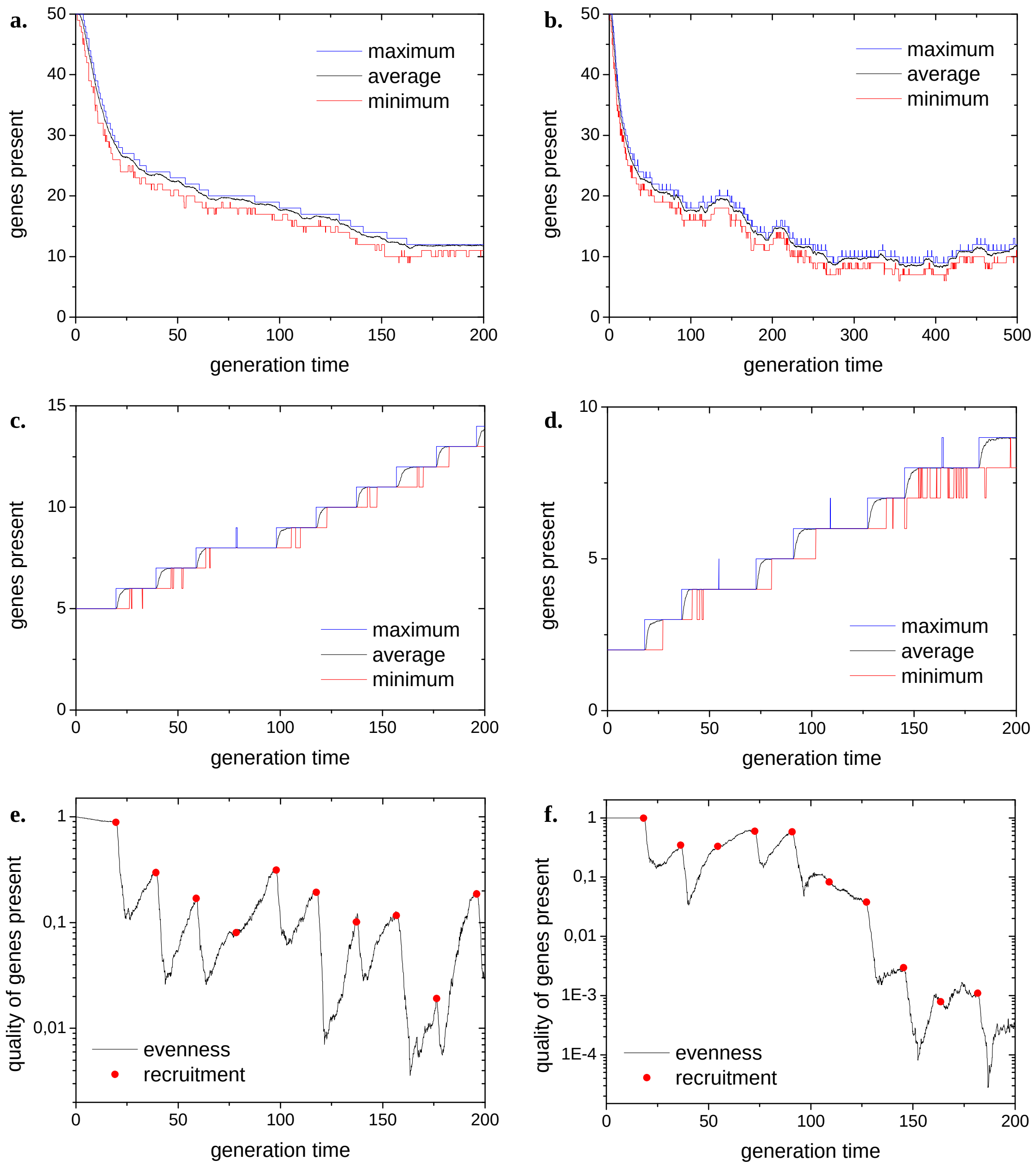


Figure R10. The ultimate behaviour of primeval metabolic systems. (a–b) Approaching the limit of sustainability: long term gene loss as determined chiefly by the critical protocell size. An unsustainable genome loses most of is genes, independently of whether **(a)** there is no occasion for recruitment (I = 0) or **(b)** there are many (I = 100); but note the different timing. **(c–f)** On the other hand, genomes will quickly recruit many genes, if their size is well below the sustainable limit and the occasion presents itself (I = 10). **(c–d)** While this tendency is independent of having or not having mutations, **(e–f)** the evenness of compositions change differently. Parameters: N = 100, τ = 50, ε = 0.3, d_0 = 0, b_0 = 0.1, p = 1. **(a–b)** $v_{\max}$ = 1000, c_0 = 10, τ_0 = 50, f = 0.01. **(a)** g = 200, I = 0. **(b)** g = 500, I = 100. **(c–f)** g = 200, $v_{\max}$ = 5000, I = 10. **(c, e)** c_0 = 500, τ_0 = 5, f = 0. **(d, f)** c_0 = 1250, τ_0 = 2, f = 0.01.

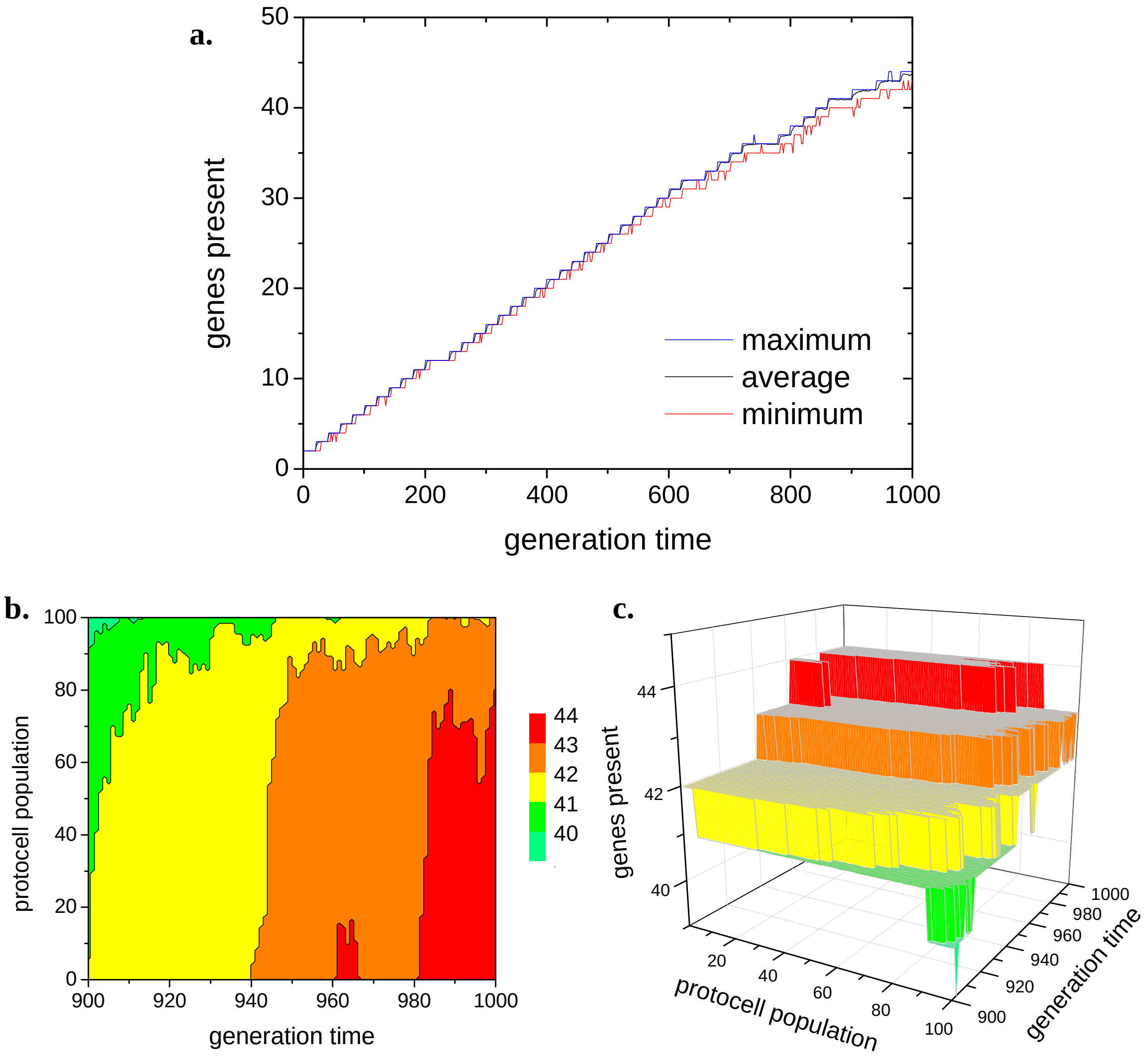


Figure R11. Towards a complex metabolism: a success story. (a) The protocell population acquires a genetically diverse composition. Starting from a mere two genes, they expand their metabolism of ribozymes to 44 genes over the course of a thousand generations. **(b–c)** For the last 100 generation, the genome size distribution of the protocell population is shown. Note that the genome sizes are sequenced in a monotonic order. Parameters: g = 1000, N = 100, $v_{\max}$ = 5000, c_0 = 1250, τ = 50, τ_0 = 2, ε = 0.3, d_0 = 0, f = 0, b_0 = 0.1, I = 50.

CHAPTER 4

DISCUSSION

We have shown that despite asynchronous replication and random assortment, the coexistence of a large number of individually replicating genes is possible in a compartmentalized system (e.g. ≥ 90 genes, Figure R5a). Thus we were able to ascertain that compartmentalization is an efficient way to integrate information. Moreover, we have shown that gene diversity can increase in this system, allowing the gradual buildup of a metabolism (Figures R10c–f, R11).

In order to elucidate the fundamental conditions required for genome sustainability (i.e., coexistence of the genes), we need to identify the mechanisms by which a compartmentalized system can overcome the obstacles posed by random assortment and asynchronous replication (Figure D1). Both of the aforementioned processes can contribute to the loss of information (i.e. loss of all copies of a gene): (*1*) if the chance allocation of ribozymes upon fission results in all the copies of a certain gene getting into the same offsprint protocell, the other offspring will certainly lack that gene; and (*2*) if, as a consequence of having a smaller replicase affinity, a ribozyme gets overgrown and thus diluted inside the protocell, its few copies will be unable to facilitate the spread of that gene (competitive exclusion), while having a fair chance of eventually getting lost by witnessing the death of their protocellular host(s).

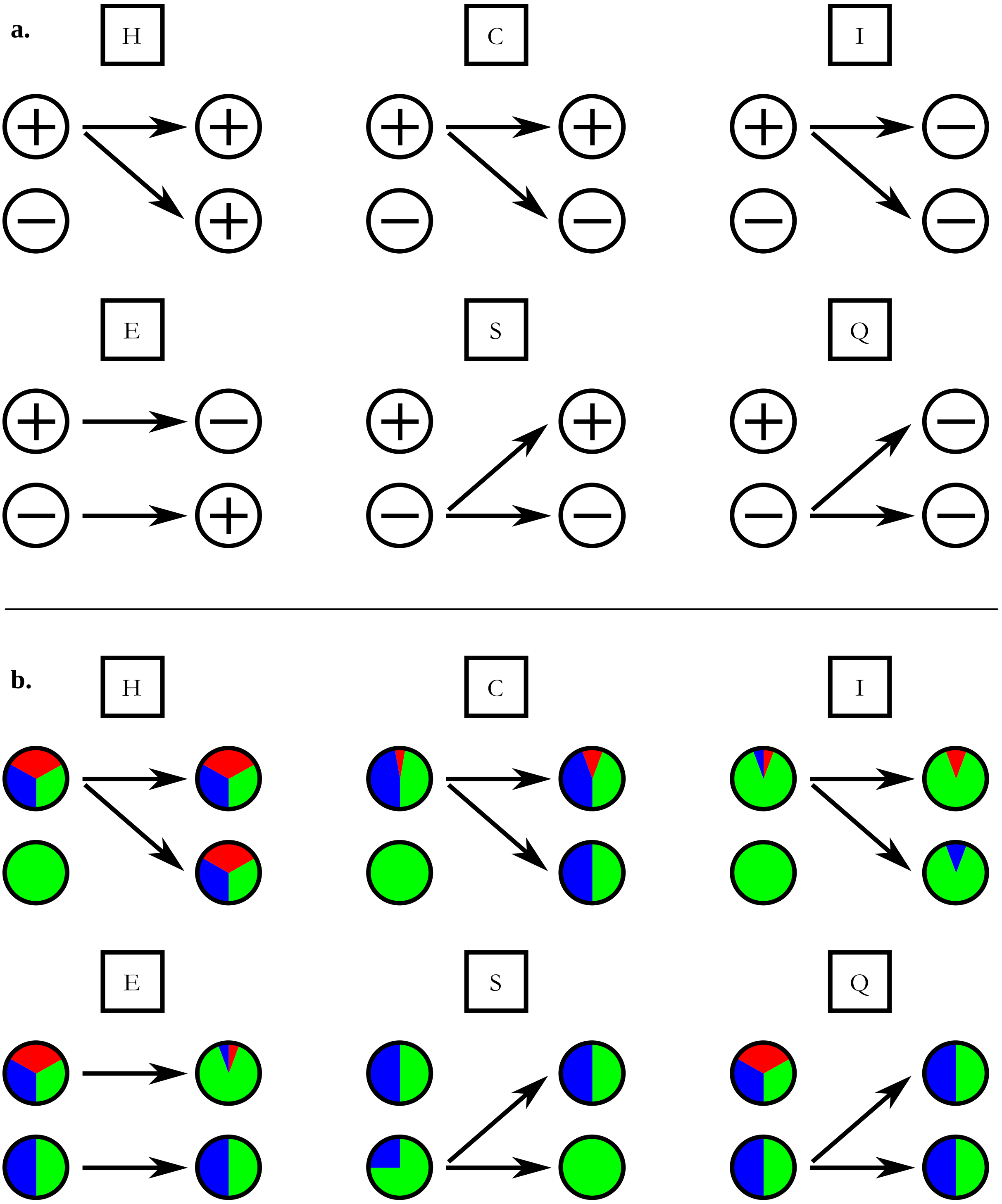


Figure D1. The presumed basic mechanisms of the observed dynamics, (a) in terms of fitness and **(b)** composition. H: hierarchical selection, C: complementation, I: isolation, E: exorbitance, S: stochastic correction, Q: quasispecies effect.

Table D1. Observed phenomena categorized by their presumed underlying mechanism.

#	fig.	description
Hierarchical selection		
1.	R1	The variance of fitness decreases with an increasing population size.
2.	R3ab	The chance of sustaining the genome increases with the population size.
3.	R5bc	Small differences in replicase affinity do not hinder genome sustainability.
4.	R7	Mutations of low severity do not endanger sustainability.
5.	R6	Asynchronous replication guarantees uneven composition in large populations.
6.	R2	A ‘helping’ selection regime reduces the sustainable genome size.
7.	R9a–d	Momentary genome sizes vary in a fairly small range at most times.
8.	R10b	New genes can temporarily spread even above the sustainability limit.
9.	R8	Mutations can lead towards evenness if the initial distribution is unequal.
10.	R9b	New genes start their spread among the protocells at a supralinear rate.
11.	R10ef	The integration of new functions decreases the evenness of compositions.
Complementation		
12.	R10ef	After gene spread is complete, the average quality of the population increases.
13.	R5a	One low affinity gene can be sustained, even with big differences in affinity.
Isolation		
14.	R5a	Several low affinity genes are hard to sustain with big differences in affinity.
15.	R7	Severe mutations undermine sustainability even at initially equal affinities.
16.	R4a–c	Redundancy is essential for genome sustainability.
17.	R10b	Even frequently emerging new genes do not raise the sustainability limit.
18.	R3c	The genome is either sustained in several protocells, or in none.
19.	R9cd	The rate of losing further genes slows down as genome size decreases.
Exorbitance		
20.	R9acd	When losing a gene, the number of hosting protocells follow a sigmoid curve.
Stochastic correction		
21.	R6	Replication asynchrony does not reduce evenness at small population sizes.
Quasispecies effect		
22.	R9a–d	A steady percentage of the protocells have a less than maximal genome.

4.1. *Hierarchical selection*

The general mechanism for maintaining favourable compositions is *hierarchical selection* (Figure D1H): among a population of protocells, those having a higher fitness (i.e., metabolic activity of the PMS inside them) will proliferate faster; upon fission, their ribozyme content will be divided among the offspring almost evenly (cf. binomial distribution); resulting in more protocells of favourable composition.

If one of the genes inside the protocell have a significantly different ($d \geq 10\%$) replicase affinity than the others, it will produce a suboptimal (i.e., lopsided) composition: higher affinity genes will outcompete and dilute lower affinity genes. This helps the spread of genes with exceptionally high affinities. However, as a large difference among replicase affinities inside the protocell will cause suboptimal composition, it will hinder the growth and proliferation of such protocells. Meanwhile, other protocells containing genes of more equal replicase affinities will proliferate seamlessly. Thus on the population level, protocells harbouring genes of exceptionally high affinities will be outcompeted by others. The selection pression on the affinity difference will therefore restrain the spread of ‘selfish’ genes (i.e., of exceptionally high affinities), and favour ‘cooperation’ among the genes (i.e., preserving, or evolving, an equal distribution of affinities).

Much of our results (#1–11, Table D1) can be understood as an outcome of hierarchical selection. When stochastic effects cause moderate variation (#1–4), hierarchical selection can ensure a good quality in sustained compositions. On the other hand, a decrease in stochasticity means fewer opportunities for selection to occur; this can be detrimental if asynchrony is significant (#5). If selection can focus on the quality of compositions—which is hereditary, instead of the quantity, which is not—then larger genomes have better chances of survival (#6–8). Selection can also support the self-reproduction of newly formed, better quality compositions (#9–11) which can lead to accelerated proliferation.

4.2. *Random assortment*

We distinguish three mechanisms of random assortment (at fission). The possible recurrence of favourable compositions, called *stochastic correction*, depends on the variance of the distribution. The sheltering of rare genes, termed *complementation,* is a mostly deterministic property of random assortments. The separation of different genes into different offspring, named *isolation,* depends on stochasticity according to the evenness of the composition prior to fission.

Complementation occurs when a single gene is scarce in the genome of fit protocells (cf. the fitness cost of losing a gene, Box M3): though there is a good chance that all copies of the rare gene will be assorted into only one offspring at fission, the offspring receiving the rare gene will most likely retain all the genes of its parent (Figure D1C). Through benefiting the rare gene, complementation can aid the 'catch up' replication of a new gene which has already spread among the protocells (#12), or balance the asynchrony caused by a single low affinity gene (#13, '1 lower' distribution).

However, if several genes of the protocell are scarce, there is a risk of isolation (Figure D1I), that upon fission, neither offspring will receive all the parental genes. An exceptionally low affinity in several genes will results in their becoming rare at the same time; isolation makes such compositions unsustainable (#14). Severe mutations undermine sustainability for the same reason: by endowing a few genes with very high affinities, most other genes will soon become rare (#15). Only a certain level of redundancy can ensure that most genes are frequent (#16–17); or a sufficiently large population, so that inaccuracies accumulated by isolation can be corrected for by hierarchical selection. In the absence of such conditions abrupt gene loss will occur (#18). A decrease in the genome—and a constant critical protocell size—will result in an increased average redundancy, causing the rate of losing genes to slow down (#19).

Stochastic correction, the process whereby a protocell of uneven composition produces an offspring with a better composition (Figure D1S; Szathmáry & Demeter, 1987), can be recognized by its dependence on the variance of the protocell size distribution at fission. It is capable of equalizing any kind of compositional disparity, even the asynchrony of genes with an exceptional difference in their affinities (#21). Still, in most circumstances the impact of this mechanism is apparently low. Note, however, that a prerequisite for stochastic correction, the fission of protocells with uneven composition is infrequent under a ‘vigilant’ selection regime. It is thus conceivable that stochastic correction would be more significant if the regime was ‘helping’.

4.3. *Shifting complexity*

We propose a mechanism, we term *exorbitance*, whereby growing protocells with a harder to sustain, larger genome deteriorate in their composition, while protocells with less genes can maintain a steady composition (Figure D1E). Eventually, the fitness of the former will fall below that of the latter, initiating a logistic invasion in the population (#20), incidentally causing gene loss (through hierarchical selection).

Until many protocells of the maximum genome size have an even composition, protocells with less genes can hardly grow. But as the composition of the former deteriorates, and their numbers dwindle, less fit protocells will have the opportunity to even proliferate. The occasion will present itself when these protocells of submaximal genome size will lose further genes, recreating a cloud of ‘mutant’ compositions, a quasispecies, around their ‘master compositions’. But the growth and proliferation of mutant protocells (of lower fitness) will yet again be impeded. We distinguish the fission of less fit protocells as the *quasispecies effect* (Figure D1Q), since it is their rareness (or nonexistence) which contributes to the constant ‘population size’ of the quasispecies during both gene loss and recruitment (#22).

On a more important note, we found a feasible evolutionary path that a protocell can take towards a larger genome. It is composed of an elementary cycle of gene integration, which can be repeated serially. The three phases of the cycle is as follows: (*1*) the recruitment of a novel ribozyme, of average or below average replicase affinity, into the metabolic system, (*2*) the synchronization of its replicase affinity with the other ribozymes, through mutation, and finally (*3*) the sustainment of this genome until yet another ribozyme comes along, leading to the reiteration of this cycle. This evolutionary path shows us how a primeval metabolic system could increase in complexity.

4.4. *Sustainable information*

We find the available redundancy the key component in determining the maximal sustainable genome size of a protocell. Such a limited sustainablitiy also restricts the achievable complexity through serial integration of ribozymes. We have established an estimate for this limit on sustainability (L), based on our findings in Figure R4:

$$L_{\mathrm{est}} = \sqrt{v_{\max} \cdot 3/5}$$

We find that this estimate successfully pinpoints the order of magnitude of the sustainability limit, as illustrated by our results (see Table D2).

Table D2. Testing our estimate. Comparing our estimate (L_{est}) with the observed results: the maximum sustainable (τ_{sust}) and the minimum unsustainable (τ_{ext}) genome sizes.

fig.	$v_{\max}$	L_{est}	τ_{sust}	τ_{ext}	OK?
R1	1000	24,5	5	-	☑
R2	6480	62,4	60	72	☑
R3	2160	36,0	36	45	☑
R4	25920	124,7	90	120	☒
R5	25920	124,7	90	120	☒
R6	25920	124,7	2	-	☑
R7	2000	34,6	5	-	☑
R8	1000	24,5	5	-	☑
R9a	100	7,7	5	-	☑
R9b	260	12,5	-	12	☒
R9c	400	15,5	-	13	☒
R10a–b	1000	24,5	-	~15	☒
R10c–f	5000	54,7	14	-	☑
R11	5000	54,7	44	-	☑

There is a final task remaining: elaborating on the possible genome size of a riboorganism. Persuant a top-down approach, we can find contemporary organism having a minimal genome size of around 600 kilobases (*Mysoplasma genitalium*, *Buchnera* sp. (Islas et al., 2004)), composed of around 500-600 genes. The minimal gene number is, of course, estimated to be less than this figure (Szathmáry, 2005; Luisi et al., 2006): around 200 (Gil et al., 2004). Lets remind ourselves that these estimates are for cells having DNA genome and peptide enzymes, thus a full machinery for translation and DNA replication is included. So a riboorganism can have an even smaller genome. Jeffares and colleagues (1998), for example, suggested that the last riboorganism had a genome of 10,000–15,000 base pairs. While this includes ribozymes involved in translation and RNA replication, it still lacks enzymes for the control of cell division, and the estimates for an intermediate metabolism are rather arbitrary. For the minimal intermediate metabolism, a good estimate is given by Gabaldón and colleagues (2007), who suggested 50 enzymes to be the minimum. Apart from the intermediate metabolism, RNA replication, RNA degradation, transport and cell-division requires enzymes. We could argue that around 60 enzymes would be the bare minimum (Szilágyi et al., 2012).

It is clear that with 0.99 replication fidelity (an error rate of 10^{-2}) a chromosome packed with 60 genes cannot be maintained due to the error threshold. On the other hand, 60 individually replicating genes can be sustained. Furthermore such complexity can be reached by serial integration of genes, a gradual increase in gene number. Our findings thus indicate that individually replicating genes could already store enough information for a minimal organism, allowing life to emerge and evolve toward the next major evolutionary transition, the chromosome.

4.5. *Perspectives*

The effect of compositional recombination of PMS-s on information integration should be examined. Our implicit assumption, that compartmentalized early genomes were completely isolated from one another, is fairly implausible. Protocellular fusion, horizontal gene transfer or both may have occurred (Emiliani *et al.*, 2010, p. 49). And while it raises the general question whether selfish replication will hinder information integration, it can even be advantageous to have an influx of genes (cf. Vogan & Higgs, 2011).

The phylogenetic dynamics of group selection should also be investigated. It is of major importance to understand how quickly the descendants of a single protocell could over-reproduce competing protocells in their effort to populate the environment. Is it beneficial to have high selective advantages, or does it lead to deleterious endogamy? Examining the impact of different selection regimes could lead to interesting insight.

REFERENCES

Anfinsen CB (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–230

Bartel DP & Unrau PJ (1999) Constructing an RNA world. *Trends Cell Biol* **9**, M9–M13

Boerlijst MC & Hogeweg P (1991) Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. *Physica D* **48**, 17–28

Brack A (1998) Introduction. In: Brack A (ed) *The Molecular Origins of Life: Assembling Pieces of the Puzzle*. Cambridge University Press, Cambridge, 1–10

Cech TR (2009) Crawling Out of the RNA World. *Cell* **136**, 599–602

Chen IA & Szostak JW (2004) Membrane growth can generate a transmembrane pH gradient in fatty acid vesicles. *Proc Natl Acad Sci USA* **101**, 7965–7970

Czárán T & Szathmáry E (2000) Coexistence of competitive-mutualist replicators in prebiotic evolution. In: Dickmann U, Law R & Metz JAJ (eds) *The Geometry of Ecological Interactions: Simplifying Spatial Complexity*. Cambridge University Press, Cambridge, 116–134

Darwin C (1959) *The Origin of Species by Means of Natural Selection*. John Murray, London

Darwin F (1887) *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter*. John Murray, London, **3**, 18

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

Eigen M & Schuster P (1977) The Hypercycle, part A: Emergence of the Hypercycle. *Naturwissenschaften* **64**, 541–565

Emiliani G, Fondi M, Liò P & Fani R (2010) Evolution of metabolic pathways and evolution of genomes. In: Barton L L, Mandl M & Loy A (eds) *Geomicrobiology: Molecular and Environmental Perspective*. Springer, Netherlands, 37–68

Fontanari JF, Santos M & Szathmáry E (2006) Coexistence and error propagation in pre-biotic vesicle models: A group selection approach. *J Theor Biol* **239**, 247–256

Gabaldón T, Peretó J, Montero F, Gil R, Latorre A & Moya A (2007) Structural analyses of a hypothetical minimal metabolism. *Phil Trans R Soc Lond B* **362**, 1751–1762

Gánti T (2003) *The Principles of Life*. Oxford University Press, USA

Garay J (2011) Active centrum hypothesis: the origin of chiral homogeneity and the RNA-world. *BioSystems* **103,** 1–12

Gil R, Silva FJ, Peretó J & Moya A (2004) Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* **68**, 518–537

Gilbert W (1986) The RNA world. *Nature* **319**, 618

Grey D, Hutson V & Szathmáry E (1995) A re-examination of the stochastic corrector model. *Proc R Soc Lond B* **262**, 29–35

Hirao I & Ellington AD (1995) Re-creating the RNA world. *Curr Biol* **5**, 1017–1022

Horowitz NH (1945) On the evolution of biochemical synthesis. *Proc Natl Acad Sci USA* **31**, 153–157

Islas S, Becerra A, Luisi PL & Lazcano A (2004) Comparative genomics and the gene complement of a minimal cell. *Orig Life Evol Biosph* **34**, 243–256

Jeffares DC, Poole AM & Penny D (1998) Relics from the RNA World. *J Mol Evol* **46**, 18–36

Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* **30**, 409–425

Johnston WK, Unrau PJ, Lawrence MS, Glasner ME & Bartel DP (2001) RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science* **292**, 1319–1325

Joyce GF (2002 a) Booting up life. *Nature* **410**, 278–279

Joyce GF (2002 b) The antiquity of RNA-based evolution. *Nature* **418**, 214–221

Kauffman SA (1986) Autocatalytic sets of proteins. *J Theor Biol* **119**, 1–24

Khvorova A, Kwak YG, Tamkun M, Majerfeld I & Yarus M (1999) RNAs that bind and change the permeability of phospholipid membranes. *Proc Natl Acad Sci USA* **96**, 10649–10654

Koch AL (1985) Primeval cells: Possible energy-generating and cell-division mechanisms. *J Mol Evol* **21,** 270–277

Könnyű B, Czárán T & Szathmáry E (2008) Prebiotic replicase evolution in a surface-bound metabolic system: Parasites as a source of adaptive evolution. *BMC Evol Biol* **8**, 267

Kun Á, Santos M & Szathmáry E (2005) Real ribozymes suggest a relaxed error threshold. *Nat Genet* **37**, 1008–1011

Kun Á (2011) Az RNS-világ. *Természet Világa* **142**, 455–456

Kunkel TA (2004) DNA replication fidelity. *J Biol Chem* **279**, 16895–16898

Luisi PL (1998) About various definitions of life. *Orig Life Evol Biosph* **28**, 613–622

Luisi PL, Ferri F & Stano P (2006) Approaches to semi-synthetic minimal cells: a review. *Naturwissenschaften* **93**, 1–13

Mansy SS & Szostak WS (2008) Thermostability of model protocell membranes. *Proc Natl Acad Sci USA* **105**, 13351–13355

Maynard Smith J (1979) Hypercycles and the origin of life. *Nature* **280**, 445–446

Maynard Smith J (1983) Models of evolution. *Proc R Soc Lond B* **219**, 315–325

Maynard Smith J (1987) How to model evolution. In: Dupré J (ed) *The Latest on the Best. Essays on Evolution and Optimality*. MIT Press, Cambridge, 119–131

Murray JM & Doudna JA (2001) Creative catalysis: pieces of the RNA world jigsaw. *Trends Biochem Sci* **26**, 699–701

Oparin AI (1936) *The origin of life*. Moscow Worker Publisher, Moscow

Orgel LE (2004) Prebiotic chemistry and the origin of the RNA World. *Critical Reviews in Biochemistry and Molecular Biology* **39**, 99–123

Pigliucci M (2009) An extended synthesis for evolutionary biology. *The Year in Evolutionary Biology 2009: Ann NY Acad Sci* **1168**, 218–228

Pigliucci M & Müller GB (eds) (2010) *Evolution - the Extended Synthesis*. MIT Press

Poole AM (2006) Getting from an RNA world to modern cells just got a little easier. *BioEssays* **28**, 105–108

Sacerdote MG & Szostak JW (2005) Semipermeable lipid bilayers exhibit diastereoselectivity favoring ribose. *Proc Natl Acad Sci USA* **102**, 6004–6008

Santelices B (1999) How many kinds of individual are there? *TREE* **14**, 152–155

Santos M, Zintzaras E & Szathmáry E (2004) Recombination in primeval genomes: A step forward but still a long leap from maintaining a sizable genome. *J Mol Evol* **59**, 507–519

Schaaper RM (1993) Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. *J Biol Chem* **268**, 23762–23765

Scheuring I, Czárán T, Szabó P, Károly G & Toroczkai Z (2003) Spatial models of prebiotic evolution: Soup before pizza? *Orig Life Evol Biosph* **33**, 319–355

Schrum JP, Zhu TF & Szostak JW (2010) The origins of cellular life. *Cold Spring Harbor Perspectives in Biology* **2**

Schuster P (2010) Mathematical modeling of evolution. Solved and open problems. *Theory Biosci* **130**, 71–89

Segré D & Lancet D (2000) Composing life. *EMBO Reports* **1**, 217–222

Silvestre DAMM & Fontanari JF (2007) Package models and the information crisis of prebiotic evolution. arXiv:0710.3278v1 [q-bio.PE]

Szabó P, Scheuring I, Czárán T & Szathmáry E (2002) *In silico* simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. *Nature* **420**, 340–343

Szathmáry E (1993) Coding coenzyme handles: A hypothesis for the origin of the genetic code. *Proc Natl Acad Sci USA* **90**, 9916–9920

Szathmáry E & Demeter L (1987) Group selection of early replicators and the origin of life. *J Theor Biol* **128**, 463–486

Szathmáry E & Maynard Smith J (1997) From replicators to reproducers: The first major transitions leading to life. *J Theor Biol* **187**, 555–571

Szathmáry E (1999) The origin of the genetic code: amino acids as cofactors in an RNA world. *Trends Genet* **15**, 223–229

Szathmáry E (2005) Life: in search of the simplest cell. *Nature* **433**, 469–470

Szathmáry E (2007) Coevolution of metabolic networks and membranes: The scenario of progressive sequestration. *Phil Trans R Soc B* **362**, 1781–1787

Szilágyi A, Kun Á & Szathmáry E (2012) Early evolution of efficient enzymes and genome organization. *Biology Direct* **7**, 38

Takeuchi N, Poorthuis P & Hogeweg P (2005) Phenotypic error threshold; additivity and epistasis in RNA evolution. *BMC Evol Biol* **5**, 9

Unrau PJ & Bartel DP (1998) RNA-catalysed nucleotide synthesis. *Nature* **395**, 260–263

Vasas V, Szathmáry E & Santos M (2010) Lack of evolvability in self-sustaining autocatalytic networks constraints metabolism-first scenarios for the origin of life. *Proc Natl Acad Sci USA* **107**, 1470–1475

Vogan AA & Higgs PG (2011) The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biology Direct* **6**, 1

Wächtershäuser G (1990) Evolution of the first metabolic cycles. *Proc Natl Acad Sci USA* **87**, 200–204

Wochner A, Attwater J, Coulson A & Holliger P (2011) Ribozyme-catalyzed transcription of an active ribozyme. *Science* **332**, 209–212

Yarus M (1999) Boundaries for an RNA world. *Curr Opin Chem Biol* **3**, 260–267

Zaher HS & Unrau PJ (2007) Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *RNA* **13**, 1017–1026

Zhu TF & Szostak JW (2009) Coupled growth and division of model protocell membranes. *JACS* **131**, 5705–5713

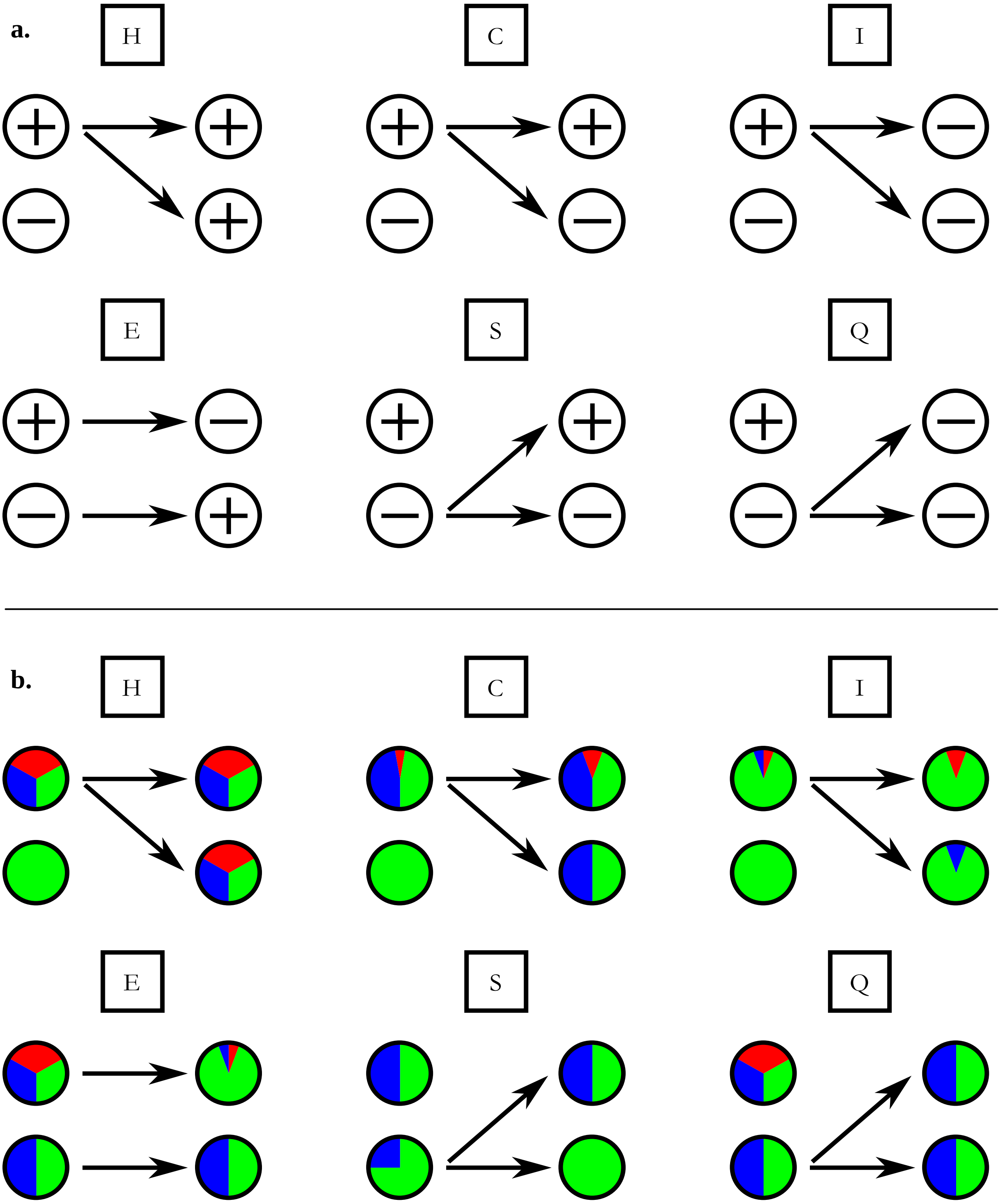
Zintzaras E, Santos M & Szathmáry E (2002) „Living” under the challenge of information decay: The Stochastic Corrector Model vs. Hypercycles. *J Theor Biol* **217**, 167–181

Zintzaras E, Santos M & Szathmáry E (2010) Selfishness versus functional cooperation in a stochastic protocell model. *J Theor Biol* **267**, 605–613

MAGYAR NYELVŰ ÖSSZEFOGLALÓ

Az élet kialakulásának korai szakaszára jellemző (ún. minimális) genomot nem volt lehetséges egyetlen kromoszómán eltárolni: az enzimatikus másolás pontatlansága miatt az információ jelentős része hamar elveszett volna (vö. Eigen Paradoxona). És bár rövid szakaszok már ekkor is kellő pontossággal másolódtak, különálló gének formájában sem problémamentes e minimális genom fenntartása: a szaporodási ráták közötti elkerülhetetlen különbségek kompetitív kizáráshoz, ezáltal pedig információvesztéshez vezetnek. A Sztochasztikus Korrektor Modell (SCM) a protosejtekbe csomagolt gének osztódáskori véletlenszerű szétválása révén teszi lehetővé, hogy az előnyös összetételű protosejtek folyamatosan újra felbukkanjanak, s így a belső versengés ellenére is fennmaradjon a teljes génkészlet.

Egyedalapú modell segítségével kerestük a vezikulaszám, illetve a vezikulán belüli molekulaszám függvényében fenntartható különböző gének maximális számát. Megmutattuk, hogy a sztochasztikus korrekció lehetővé teszi közel 100 gén együttélését, még bizonyos egyenlőtlen szaporodási ráták mellett is. Egy minimális élő sejthez körülbelül 60-100 különböző gén szükséges, így elmondható, hogy kompartmentalizált rendszerünkben az információ integrációja sikeres: elegendő a protosejt működéséhez. Bemutattuk az elemi mechanizmusok egy szűk körét (D1. ábra), amely segítségével értelmezhetővé válnak a dinamika megfigyelt jellemzői. Eredményeink felvetnek egy lehetséges evolúciós utat, amely során újabb és újabb gének épülnek be a rendszerbe, annak összeomlása nélkül.



D1 ábra. A megfigyelt dinamika feltételezett elemi mechanizmusai, (a) a rátermettség illetve **(b)** az összetétel vonatkozásában ábrázolva. H: többszintű szelekció, C: kiegészítődés, I: izoláció, E: mértéktelenség, S: sztochasztikus korrekció, D: kvázispeciesz hatás.

NYILATKOZAT

Név:	Hubai András Gábor
Neptun azonosító:	E9AC16

ELTE Természettudományi Kar, **biológus mesterszak**

Diplomamunka címe:	Emergence and evolution of primeval metabolic systems

A diplomamunka szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló munkám eredménye, saját szellemi termékem, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam.

Tudomásul veszem, hogy plágiumnak számít:

– szó szerinti idézet közlése idézőjel és hivatkozás megjelölése nélkül;
– tartalmi hivatkozás a forrás megjelölése nélkül;
– más személy publikált gondolatainak saját gondolatként való feltüntetése.

Kijelentem továbbá, hogy a diplomamunka leadott nyomtatott példányai és elektronikus változata szövegükben, tartalmukban megegyeznek.

Budapest, 2013. május 17.

____
a hallgató aláírása