Journal Name

RSCPublishing

COMMUNICATION

The dynamics of the RNA world: Insights and challenges

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2014,
Accepted 00th January 2014

DOI: 10.1039/x0xx00000x

www.rsc.org/

Ádám Kun[a,b], Gergely Boza[c], Balázs Könnyű[c], András Szilágyi[a,d], István Zachar[a], Eörs Szathmáry[a,c,d*]

[a] *Parmenides Center for the Conceptual Foundations of Science, Munich/Pullach, Germany*

[b] *MTA-ELTE-MTMT Ecology Research Group, Department of Plant Systematics, Ecology and Theoretical Biology, Budapest, Hungary*

[c] *Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös University, Budapest, Hungary*

[d] *MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group, Department of Plant Systematics, Ecology and Theoretical Biology, Budapest, Hungary*

[*] *corresponding author*

The problem of the origin of life is not only one of structure but also that of dynamics. Ever since the seminal result of Manfred Eigen in 1971 showing that early template replication suffers from an error threshold, research has tackled the issue of how early genomes could have been dynamically stable without highly evolved mechanisms such as accurate replication and chromosomes. We review the theory of the origin, maintenance and enhancement of the RNA world as an evolving population of dynamical systems. Investigation of sequence space has revealed how structures are allocated in sequence space and how this affects the nature of the error threshold that sets the selectively maintainable genome length. New applications of old dynamical theory are still possible: the application of Gause's principle of competitive exclusion, based on resource utilisation, to RNA replication predicts that at most four pairs (plus and minus strands) can stably be maintained on four nucleotides. Other mechanisms of early template coexistence should be regarded as additional means to raise the number of coexisting species above the number set by the competitive exclusion principle. One such example is the hypercycle in which templates were postulated to help replication of the next member in a cycle superimposed on individual replication cycles. Although the hypercycle is ecologically unstable it is evolutionarily unstable because it cannot efficiently compete against emerging parasites. Population structure can modify this conclusion but not without further qualification. The simplest form of population structure is limited diffusion on a surface. This simple mechanism can ensure the coexistence of competing ribozymes contributing to surface metabolism as well as the spread of efficient replicases despite the parasite problem. Hypercycles can only be saved by active compartmentalization when replicators are enclosed in reproducing protocells. Once there are protocells there is no need for internal hypercyclic organization, however. Finally we review two crucial adaptations that enhanced the RNA world: chromosomes and enzymatic metabolism. Interestingly, it was shown that these two have been presumably coevolutionarily linked because protocells harbouring unlinked, competing ribozymes are better off if the ribozymes remain inefficient but generalists. The appearance of chromosomes alleviates intragenomic conflict and is enabling constraint for the emergence of specific and efficient enzymes.

The possibility of an RNA world, a period in the origin of life on Earth, when RNA molecules acted both as enzymes and as genetic material, was suggested well before the name was coined by Gilbert in 1986[1]. The history of the research on the origin of life[2] tells us that the potential prebiotic importance of RNA was suggested as early as in the late 50's. When it became established that living cells harbour much more RNA than DNA some biologist have proposed that RNA preceded DNA during evolution[3,4]. The discovery of the details of protein synthesis[5] revealed a plethora of RNA molecules involved in a diversity of processes within the contemporary cells, which prompted speculation on the possible prebiotic/ancestral role of RNA. Carl Woese[6], Leslie Orgel[7], and Francis Crick[8] independently proposed that RNA acted both as catalyst and as information carrying molecule. Tibor Gánti[9] presented a detailed account of the origin and embedding of catalytic RNA molecules in a metabolising and dividing chemical supersystem: the chemoton[10]. The idea of catalytic RNA received prime experimental proof by the discovery of natural

RNA enzymes (ribozymes), found independently by the groups of Sidney Altman[11] and Thomas Cech[12].

Jeffares[13] proposed that if we encounter a catalytic RNA in a modern organism, it could be a relic from a bygone era—especially if it is found in all domains of life. Unfortunately, not many naturally occurring ribozymes are known. Besides the firstly discovered RNase P[11] and the group I introns[12], there are group II introns[14], the hammerhead ribozyme[15], the hairpin ribozyme[16], the Hepatitis Delta Virus and like ribozymes[17], the *Neurospora* Varkud Satellite Ribozyme[18], the *glmS* ribozyme[19], and the twister ribozymes[20]. These molecules can, however, only cleave RNA molecules[21,22], not a repertoire upon which a metabolism could have been built. A convincing argument says that these ribozymes have been retained in evolution because the large size of the products limits the attainable catalytic enhancement, hence a replacement by protein enzymes could not been selected for[13]. At the same time the idea of an RNA world was built upon the diverse roles of RNA in contemporary metabolism and not upon the limited catalytic role of these natural ribozymes. We are only beginning to unravel the world of functional RNA molecules, but it is already clear that they are much more than simple information storages (RNA viruses) or information carriers between DNA and peptides (mRNA).

After revealing the catalytic role of RNAs, a smoking gun was found: in translation RNA serves as a direct link between DNA and peptides, the two essential actors of contemporary life. Before this discovery, the central component of translation, the ribosome, was thought to be a normal protein enzyme, with an inordinate amount of RNA bundled within. It took decades to unveil the surprising fact that the ribosome is actually a huge ribozyme[23,24] (or at least a ribozyme with peptide structural elements thrown in). The fact that RNA molecules involved in all aspects of the translation process lead many to propose the RNA world hypothesis[6-8], even without knowing its central role yet. Hence this finding provides extremely strong evidence in favour of the theory. It is significant that evidence is accumulating in favour of accepting the spliceosome as a ribozyme[25,26] of which the RNA core has been conserved for over one billion years.

The "fossil record" of the RNA world does not stop with translation, many of the important coenzymes contain a nucleotide part[27]. NAD(P), FAD, Coenzyme A, S-adenozyl-metionin, 3'-phosphoadenosine-5'-phosphosulfate (PAPS), ATP contain an adenine part, while thiamine pyrophosphate, THF, pyridoxal phosphate have cyclic nitrogenous bases that could have been derived from a nucleobase. Interestingly, their biological activities do not depend on the adenine part. Then why is the nucleotide part present at all? It could have been a handle through which ribozymes got hold of the coenzyme before the protein world[28,29]. Aptamers evolved to bind CoA are always binding the coenzyme through the adenine part, and never through the sulfonated pantothenic acid part[30]. These coenzymes are the ones found to be autocatalytic in metabolism[31], which also suggests their ancient origin. Although much better coenzymes might had evolved in a purely protein world, but once many of the reactions already relied on a particular (and crucial) coenzyme evolved in an RNA world, replacing them was nearly impossible, thus many ancient coenzymes evolved in the RNA world are still with us.

Szathmáry[32,33] has proposed a way to evolve novel aptamers *in vitro*, which was realized many times over the next decade. The success of the SELEX technique[34] to obtain ribozymes for many important reactions convincingly demonstrates that RNA can have a rich catalytic repertoire[35-39]. All types of reactions necessary for nucleotide and peptide syntheses can be catalyzed by such ribozymes[36]. We also need to mention redox ribozymes[40,41] which demonstrate that energy production is within the capabilities of ribozyme-run metabolisms. A fully functioning ribo-organism also has to have a membrane besides metabolism, and thus needs membrane transporters. RNA can change the permeability of the membrane[42] and ribozymes can even act as membrane transporters[43] allowing control over the exchange of material with the environment.

The facts listed above support the existence of an ancient RNA world. In themselves they strongly suggest, if not outright prove, that the RNA world held sway at the invention of the genetic code and translation. The chemical nature of coenzymes and the enzymatic repertoire of *in vitro* evolved ribozymes indicate that the RNA world could have had a rich metabolism. RNA involvement in the translation suggests that peptide synthesis was evolved in the RNA world, before DNA. Modern metabolism is arguably a palimpsest of the ancient RNA world[44].

Any proof of its existence however does not mean that we understand all aspects of the RNA world. The truth is, we are quite far from it, as now there might be more questions about the origins of the RNA world than answers. This review focuses on how the RNA world could have emerged after the appearance of self-replicating molecules and how it could have provided the first scaffolding to the living cell ultimately orchestrating the transition to peptide enzymes and to the DNA-encoded genetic material. We survey the current status of the RNA world from the point of view of theoretical evolutionary biology. The story is considerably more coherent than even a decade ago, but burning open questions still remain: these missing details provide the second target of this review. While the followings might be seen as an embarrassing list of ignorance, we see them as successive steps of a research plan: a list of well-defined questions that need to be tackled. One has to appreciate that the up surging of open questions in a field does not only indicate the increasing attention, but also that it gets momentum and progresses, as without progress, no new questions would surface. Let us quote Orgel's optimistic words about the RNA world: "*We are very far from knowing whodunit. The only certainty is that there will be a rational solution.*" (Orgel, 1998). This review complements and updates another one that was written about a decade ago[45].

I. Establishment of the RNA world: The nature of RNA template replicators

"… the Struggle for Existence amongst all organic beings throughout the world […] inevitably follows from their high geometrical powers of increase…"
(Darwin: Origin of Species, 1859)

The RNA world, as any complex adaptive system, has its own problem-of-origins: while a fully RNA-based genetic system seems feasible in light of findings about the catalytic repertoire of ribozymes (see Section III. on metabolism), assuming the spontaneous appearance of a general and effective RNA polymerase ribozyme is highly unrealistic. On the other hand, how could evolutionary search gradually select for a replicase when there is no replication and inheritance, thus no evolution yet? This was termed by Robertson and Joyce[46] as a chicken-and-egg paradox, pointing out that the problem of the origin of genetic systems having multiplication, variability and heredity before an effective RNA world, was relegated, but not vanquished. Here we discuss the possibilities and consequences of some hypothesis that try to remedy the paradox of the very origin of the RNA world.

I/1. The combinatorial approach

RNA sequences can form both in solution[47,48] and on mineral surfaces such as montmorillonite clay[49,50]. A generally accepted scenario[51-53] postulates that the first RNA sequences emerged by random, non-enzymatic synthesis of oligomers and their ligation and recombination produced longer sequences on a surface[54]. The ensuing RNA pool was diverse in structure and thus had some catalytic activity. This scenario, however, is problematic on more than one account. While small ligases most probably would form in a prebiotic environment and the sequence diversity would be undoubtedly huge, one has to keep in mind that the sequence space is vast. Even if we consider only sequences up to length $L = 50$ nt [49,50], the total number of variants is $\sum_{L=1}^{50} 4^L$, which is around 10^{30}. The important enzymes need to emerge from this sequence space, from which at any particular time only an infinitesimally small fraction can be realized. Of course, early evolution did not wait for a single, specific sequence to appear, as many different nucleotide orders could have provided useful enzymatic activity.

Ribozymes activity can be maintained if the structure is kept intact[55,56], or can even withstand minor structural mutations, as the sequence → structure map is highly injective[57]. Thus it makes more sense to look for an appropriate structure than for a sequence. The number of structures for a sequence space of sequence length L is 2.35^L instead of 4^L [57], since there are much fewer structures than sequences for a given sequence length[57,58]. Moreover, some RNA structures are more common than others[59]; for shorter sequences (around $L = 30$) it was shown that more than 90% of sequences fold to common structures. A structure is considered common if it is formed by more sequences than the average structure[59], *i.e.* $N_c > 4^L / S_L$, where N_c is the number of sequences folding into a common structure, S_L is the number of distinct structures of length L. The rest of the structures, while being quite numerous, are represented only by a few sequences. These rare structures are hard to find, as they exist only in some corners of the sequence spaces, as opposed to common structures which exist everywhere. Furthermore, even if a rare structure is found, it can be easily lost as mutations always result in a different structure, while for common structures mutations often result in the same structure (see Section I/3.).

Thus, due to combinatorial and physical necessities, ribozymes fold more probably into these common structures[39,60]. Common structures are easily reached from any starting point in sequence space via evolution; most common structures are within a distance of maximum 15-20 mutations from any arbitrary sequence of length $L = 50$ [58,61-63]. As the composition of RNA sequences is not random[50], the reachable sequence space is constrained. If we assume that such sequence constraints do not restrict the reachable structural space, then a smaller sequence space needs to be searched to find a useful structure. Unfortunately this fraction of the sequence space still contains ~10^{23} sequences (considering sequences only up to length 50). Let us say that short ligases and nucleases emerge. If the reaction network of short oligomers results in the uncontrollable duplex formation, dissociation, ligation and breakage of RNA sequences, the hypothesis of *de novo* emergence of ribozymes has to face another serious blow: this leads to an unavoidable elongation of sequences (called the *elongation catastrophe*)[64]. As the template length increases, the number of possible elongation-events suffers a combinatorial explosion. Consequently the diversity increases in the population, instead of producing a restricted but useful set of sequences.

The hypotheses about the early development of the RNA world usually conclude that if a restricted set of RNA sequences can exhibit a large enough structural variation, then the useful molecules can be enriched as such an enrichment (selection) was frequently demonstrated by *in vitro* selection experiments[54,65]. The problem with such a line of thought is that these techniques, such as SELEX[66], are evolutionary techniques employing the template directed replication of the genetic material. Evolution requires variation, multiplication and heredity. Random generation of RNAs offers variability and multiplication, but no heredity.

Another theory possibility rests on von Kiedrowski-type replicators[67]: two trimers can form a hexamer guided by another hexamer then two hexamers can form a dodecamer guided by an existing one, and so on. Potentially, quite long replicators can be synthesized in a dynamically stable and exponential way. This scenario of "convergent synthesis" has been analysed in the model of Fernando *et al*. who concluded that spontaneous elongation and parallel replication of short oligomers do not allow this mechanism to raise itself above noise level[64]. While this system could show multiplication and heredity, a further problem would be that tolerable variation is

limited as the oligomers and the template have to be very specific and many mutations would ruin the templating effect. Thus the system would lack the potential for open-ended evolution, though even fully fledged ribozymes can replicate in such manner[68].

As a more feasible alternative to the *de novo* emergence of a replicase, short functional replicators (that can emerge spontaneously without enzymes) may form a diverse cross-catalytic set that in turn might be responsible for the replicase functionality as a whole (note that no autocatalysis is assumed for members at this point, *cf.* [69,70]) or they might self-assemble to be a functional ribozyme[71]. Vasas *et al.*[72] analysed the kinetic stability of a simple two-membered autocatalytic loop, where each member catalyses the inclusion of one non-catalytic molecule. If there are large differences in catalytic efficiencies (as it is probable in prebiotic context) the system shows kinetic instability. In this case the deterministic equilibrium concentration of one of the members is very low, so loss by chance in a stochastic system is likely. Thus, even if a replicase appears in a diverse, prebiotic RNA pool, it would still be subject to stochastic loss because initially its concentration is too low.

Along similar lines, an interesting system has been presented as a possible solution of the problem of early RNA replication by Meyer *et al.*[73]. In the proposed network a polymerase helps the replication of RNA oligomers (but not that of complete polymers) and a ligase helps the formation of itself as well as of the replicase out of these oligomers. The system is collectively autocatalytic but there is no direct mutual catalysis of replication: the polymerase helps the replication of the oligomers but the latter contribute stoichiometrically rather than catalytically to the formation of the polymerase and the autocatalysis of the ligase.

We want to understand the transition from activated monomers and short (or not so short) oligomers to an evolving ensemble of RNA replicators. Starting from synthesized (as opposed to replicated, based on a template) RNA sequences, finding a replicase ribozyme that could kick-start evolution is problematic because of the vast sequence space that needs to be searched. Moreover, maintaining a fledging replicase in the realm of population stochasticity is not easy. In summary, the emergence of the first template replicator is far from solved, we are only beginning to understand the problem itself.

I/2. Resource competition: Gause's principle

A self-replicating ribozyme—even if appeared somehow—would still had to compete with other sequences and side reactions for a limited set of resources (activated nucleotides) and fight information loss due to erroneous replication. Simple ecological considerations could help to establish the baseline for coexistence of prebiotic replicators.

The competitive exclusion principle is one of the major organizing aspects of ecology, formulated first by Gause in the Golden Age of theoretical ecology[74]. It states—in a rough-and-ready way—that the number of coexisting species must be less or equal to the number of resources that the species compete for. This obviously puts a limit on the diversity of coexisting species and remains valid beyond the scope of classical ecology. Two major refinements are in order: firstly, the above statement is only valid for steady state situations. Secondly, "resource" does not mean nutrients only but includes many other factors affecting coexistence as well (the so-called regulating factors, *cf.* [75,76]).

Exponential growth is generally used as a "reference case" for modelling in population dynamics. The underlying assumption is simple: the change in the amount of a given species is proportional to its actual amount; the (asexual) mitotic division of a protist is a fitting example. The corresponding differential equation is the following:

$$\mathrm{d}x(t)/\mathrm{d}t = k\,(x(t))^p,$$

where $x(t)$ denotes the concentration of the species at time t, k is the Malthusian parameter of growth (per capita growth rate) and $p = 1$. In this case, the population growth is exponential ($x(t) = x(0)\cdot\exp(k\cdot t)$), until it reaches ecological (extrinsic) constraints. If competing species have different Malthusian parameters, the type with the higher k ultimately excludes all other variants in the absence of mutations. In Eigen's quasispecies model (see *e.g.*[77]), sequences are competitors living on a shared pool of limiting resources (one type of monomer) thus the fastest replicator with its mutational neighbourhood (the quasispecies, see Section I/3.) always excludes others.

When taking the above results into consideration, there is an obvious question: what is the limit of diversity of coexisting replicator molecules competing for the same resources (nucleotides in the RNA world) that can still be maintained? Mutation-free pure resource competition can provide a lower bound on the diversity of coexistence. Both numerical and analytical results of such resource competitions of polynucleotides agree with Gause's principle[78]: asymptotically stable coexistence is only possible when the number of replicators does not exceed the number of resources (nucleotides) and the nucleotide composition of replicators is sufficiently different ("niche-segregation" in the RNA world). Interestingly, the two complementary strands (the plus and minus strands) can be counted as one replicator from an ecological point of view as they are strictly stoichiometrically coupled, thus for example on four nucleotides at most four pairs (eight sequences) are able to coexist. The coexistence is affected not only by the nucleotide composition but by the nucleotide order too: parts of sequences that are copied earlier have a larger influence on the dynamics due to the higher concentration of the corresponding replication intermediates than parts copied later. This sequence effect affects coexistence (*e.g.* it can allow the coexistence of two replicators with identical nucleotide compositions but adequately different sequences), but does not permit more species to coexist. For a simple replication system, the number of nucleotides applies a strict and rather low bound on the number of coexisting sequence-pairs, hence the sustainable diversity.

Taking into account the phenomenon that double-stranded RNA molecules are replicationally inert, the dynamics of coexisting replicators changes dramatically, yielding a more permissive criterion for coexistence. Both von Kiedrowski[67] and Zielinski and Orgel[79] have constructed systems of hexa- and tetranucleotides (respectively) with the ability to non-enzymatic self-replication. Instead of exponential growth, they have found that the growth rate is proportional to the square root of the actual concentration, that is $p = ½$ in Eq. (1). This limited growth is due to three different factors: (*i*) only the single-stranded nucleotide can act as template for replication; (*ii*) the concentration of single-stranded templates is proportional to the square-root of the total concentration; and (*iii*) the immediate product of replication is a replicationally inert double-stranded form. Due to such dynamics, there is always an *advantage of rarity*: any species can invade the population when rare[80-82]. This is true not just for $p = ½$, but for any value in the range of (0, 1). While the exponential case ($p = 1$) means "survival of the fittest", the $p = (0, 1)$ interval corresponds to "survival of everybody": in this so-called parabolic regime, an arbitrary number of competing populations can coexist in a globally stable way[83]. Note that there is enhanced selectivity in the system though relative to the linear growth case: if $p = ½$ then the ratio of equilibrium concentrations of the competing species is the square of the kinetic rate constants.

While the regime of parabolic replication can sustain an arbitrary large diversity and could overcome the restriction posed by Gause's principle, such replicators cannot be *real* information-integrators as evolution cannot act on them. If we assume that any new mutant is also subject to duplex formation, they can gain no selective advantage and thus no evolution is expected to happen in such a regime, because for Darwinian selection, exponential growth of competing replicators is necessary[80,82].

If parabolic growth is coupled with additional physically and chemically feasible assumptions (like degradation, binding of replicators to the surface in an adsorption-desorption process, see [81,84-86]), the outcome of the dynamics (whether it will be survival of everybody or the fittest) becomes a quantitative issue depending on external parameters. Accordingly, such a system of replicators would be able to switch between coexistence (parabolic) regime and a selective (Darwinian) regime, which could provide the necessary selective edge for the system to become a real unit of evolution.

While Gause's principle limits the number of coexisting species by the number of independent resources, there could have been many ecological and dynamical factors that extends the number of "resources" and thus relaxes this limit. When it comes to coexistence, molecular replicators are not that much different from the multi-cellular organisms of supraindividual biology. And thus the results of ecology might apply: it was demonstrated that extrinsic variation in space and time[87,88], intrinsically generated fluctuations[89,90] and chaotic mixing[91] introduce other regulating factors and can increase the number of coexisting species. The possible role of these factors in the prebiotic context is the scope of further research. All this offers a glimpse of hope that a variety of replicators could have coexisted in plausible prebiotic environments and they could have evolved to more complex systems. But from a theoretical-ecological point of view, we do not yet have a conclusive answer.

I/3. Error threshold

Too high a degree of variability undermines heredity. This sounds obvious but it was discovered rather late that the mutation rate (inversely proportional to replication accuracy) sets a limit on the amount of genetic information that can be maintained by selection. Eigen[77] was the first who analysed the amount of maintainable information in the context of a reaction kinetic model of molecular evolution: this landmark study presents the flip side of the coin of the mutational load, known to population geneticists since the investigations of Haldane[92]. Eigen's theoretical model described the dynamics of a large population of replicating sequences (genotypes) in a well-mixed flow reactor. In case of error-free replication, the equilibrium population consists only of replicators with the highest fitness (assuming only one fittest type, the master). If there is even the smallest chance of mutation during replication, the mutation-selection balance results in a new equilibrium: a cloud of mutants appears in the mutational neighbourhood of the master sequence which nevertheless remains the most abundant. This well-defined distribution of mutants (together with the master phenotype) is the quasispecies, introduced by Eigen and Schuster[93] and it becomes the target of selection. By decreasing replication accuracy, the quasispecies collapses at a critical value with a sharp transition; beyond this point the master sequence is lost, the system diffuses randomly in genotype space and further evolution is impossible as no information can be selectively maintained. This critical value of replication accuracy is the error threshold.

The loss of information is inevitable in any such mutation-selection system, but the exact position of the error threshold depends on the fitness landscape (i.e. the phenotype-fitness mapping) and parameters of the population dynamics of replicators (*e.g.* degradation rate, population size, interaction between molecules, etc.). In case of the single-peak fitness landscape of the original model (the master sequence has fitness > 1, all others have 1), the critical per-base replication accuracy (q^*) that defines the error threshold can be approximated analytically as $q^* = s^{-1/L}$ (where s is the selective superiority of the master sequence). Assuming that the logarithm of $s \approx 1$, the maximum chain length roughly equals the inverse of mutation rate per site per replication. Without peptide enzymes, the per-nucleotide copying fidelity is approximately 96-99% [94-96]. This approximation (as a rule of thumb) suggests a very strict limit on the sustainable sequence length that is far from what is thought to be necessary for "minimal life". This sets the so-called Eigen paradox, or with the words of John Maynard Smith "*the 'Catch-22' of the origin of life: no large genome without enzymes, and no enzymes without a large genome*"[97].

There are, however, many subtleties that must be discussed to evaluate the severity of an early error threshold. The single peak fitness landscape is an abstraction with limited biological relevance. While a huge body of literature deals with calculating the error threshold for further fitness landscapes, the selection criteria for each landscape was unfortunately almost always analytical tractability[98] and not biological relevance. For example, the perturbation theory of quantum mechanics can be used to estimate the equilibrium distribution of concentrations in the quasispecies, but this method is applicable only when all fitnesses are different (for details, see [99]). From a biological point of view, this is rather implausible as it excludes individuals sharing the same baseline fitness. Another example rests on the formal analogy between the two (purine and pyrimidine) bases of a binary template and a 2D Ising system with nearest neighbour interaction. There is an exact correspondence between the equilibrium properties of the 2D Ising lattice and Eigen's model, see [100] (for a more general statistical physics approach, see [101-104]). In this context, the error catastrophe corresponds to the magnetic order-disorder transition. Some analytically partially tractable solutions for very simple fitness landscapes can be derived using this analogy. However, the required simplifications on the fitness landscape to achieve tractable solutions makes the model biologically implausible (*e.g.* fitness decreasing with square root of the Hamming-distance from the master genotype[101]; or decreasing in a stepwise manner[102]). Consequently, in a general case, a numerical solution is possible only either by numerically integrating the set of differential equations or via computing the leading eigenvalue of the value matrix of the system (see *e.g.* [99]).

As already discussed, it is not a particular sequence but a structure that needs to be replicated. Thus instead of a genotypic error threshold, we should look for the phenotypic error threshold: the critical mutation rate above which the functional phenotype cannot be maintained selectively. As the number of structures is considerably fewer than the number of sequences, genotypes sharing the same phenotype form a neutral network (or neutral set) in the genotype space. The percolated topology of neutral sets allows for easier evolutionary adaptation: finding a given secondary structure (function) by a mutation-selection process is easier than expected[105-107] and losing an already acquired function is also less probable. Therefore, the so-called phenotypic error threshold is more permissive than the original (genotypic) error threshold.

The connectivity of neutral paths characterized by the fraction of mutants having the same phenotype can account for the more permissive phenotypic error threshold[108,109]. Below a critical replication accuracy (the phenotypic error threshold) the population diffuses randomly over the whole genotype space and the master phenotype is lost. At a relatively high replication accuracy, the population randomly drifts on the neutral network of the master phenotype preserving the secondary structure[110]. Traversing the neutral network is not entirely random, instead the population tends to move to a highly connected part of the neutral set[107] (see Figure 1). A reliable estimate of the structure of neutral networks can only come from fitness landscapes based on real world data. Available data on the activity of mutated hairpin ribozyme[55] and *Neurospora* VS ribozyme[56] allows the construction of a fitness landscape[111]. The phenotypic error threshold allows sequences nearly a magnitude longer to be maintained[112] than presumed from the Eigen's model (i.e. 700 *vs.* 100 nt with 10^{-2} error rate). The whole genetic material required for a minimal ribo-organism, however, cannot be replicated unless the error rate falls below 10^{-3}. On the other hand, individual ribozymes, even relatively longer ones like replicases, can be stably replicated at this accuracy[113].

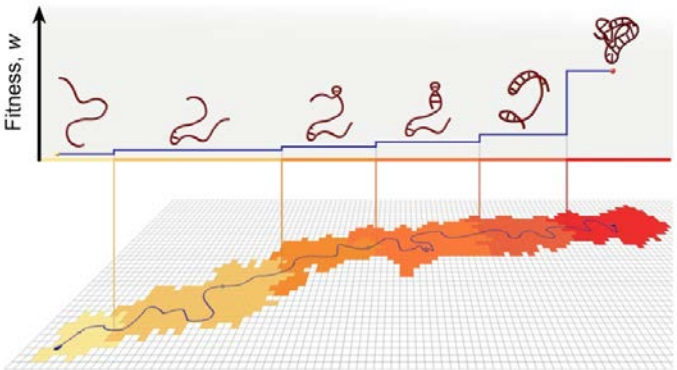

Figure 1. The evolution of RNA phenotypes. While the evolutionary search in the sequence space (represented by the rectangular grid; the evolutionary trajectory is represented by a blue thick line, lower panel) is continuous, multiple sequences may form similar phenotypes (represented by the different shading colours on the grid), hence the fitness increase is non-linear (blue line in the upper panel).

The selectively maintainable information can be further increased by taking into account the stalling of replication after mismatch. Stalling after a Watson-Crick base pair mismatch has been observed for many DNA polymerases[114-116] (the factor of slowdown is between 10 to 10^6). Rajamani and colleagues[117] have demonstrated that in case of non-enzymatic polymerization of DNA, the speed of polymerization slows down by two orders of magnitude after base-pair mismatches. Thus accurate copying without mismatch has an advantage of faster replication. Remember that the original Eigen model assumed that the speed of polymerization as such is not affected by accuracy. Consequently, if stalling is not omitted, more information can be maintained and thus the error threshold is mitigated. These experimental results concern only DNA replication; extension to the RNA world is at least speculative as the authors stated: "*Our data may not be representative of mutations in the RNA world itself, but our results do demonstrate that a nonenzymatic system exhibits stalling after mutations and that such a system could be capable of propagating sequences long enough to be functional because of this effect.*"[117]

Recombination, a mechanism usually ignored in the study of early molecular evolution, could have had its role in the

alleviation of the error threshold[118]. Santos and colleagues[119] have found a beneficial effect of recombination on the sustainable genome size. The authors assumed compartmentalized populations (see Section II/3.) of genes with internal competition among unlinked templates. Recombination during the replication of a gene was allowed. An increase of roughly 30% in length could be achieved by recombination.
Mutation rate sets a limit to the length of an RNA molecule that can be faithfully copied. Replication accuracies at the dawn of life were not sufficiently high to stably replicate all the necessary genes stringed into a chromosome. Maintenance of structure coupled with stalling at mutations and recombination between different copies of the same game can relax the error threshold to the level where individual genes can be faithfully replicated.

II. Maintenance of the RNA world: Coexistence and evolvability of early replicators

[…] differentiation is the necessary condition for coexistence.
(G. Hardin, Science, 1960)

The error threshold, as we have discussed in the previous section, prevents the stable maintenance of information above a certain size[77,112,120,121]. While the whole genetic information cannot be accurately copied as one molecule, a coexisting set of shorter replicators can still provide the same information content. Although such a collection could overcome the error threshold, it also introduces a new problem: during replication, all replicator types have to be replicated *together* to maintain the complete information content of the system. This requirement poses a serious problem as replicators competing for common resources are subject to *competitive exclusion* which ultimately means the survival of only as many replicators as the number of resources (discussed in Section I/2.).
A feasible solution for coexistence is when the full information content of such a system is shared among *functionally interacting* shorter replicators. Phenotypically different replicators assemble to create a molecular community in which each member is a replicator and is essential for the maintenance of the whole system as well, thus it is a collectively autocatalytic system. A functionally coupled replicator system is vulnerable to any member that does not contribute to the maintenance of the whole community jeopardizing the integrity of the system. Therefore, the resistance against such *parasites* has to be in the focus when the coexistence of early replicator communities is investigated. Please note that coexistence in prebiotic molecular communities touch on the same or similar mechanism as coexistence in ecological contexts (for a review of coexistence in ecological settings see [76]). Here we discuss three hypotheses for the coexistence of early replicators, showing if it is possible to achieve and selectively maintain a molecular diversity required to advance to the next stages of the RNA world.

II/1. The hypercycle versus cross-catalytic networks

The hypercycle was the first theoretical model in which functionally coupled replicators could form a molecular community[77,93,120]. In the original model, an arbitrary number of replicators are directly linked together to form a cyclic catalytic loop, thus each member of this loop catalyses both its own replication and the replication of the next member. Accordingly, members of the hypercycle are autocatalytic both individually and collectively, thus forming a cooperative system, see Figure 2A. This hypercyclic connection is responsible for the stable coexistence of replicators[120]. The hypercycle is indeed ecologically stable.

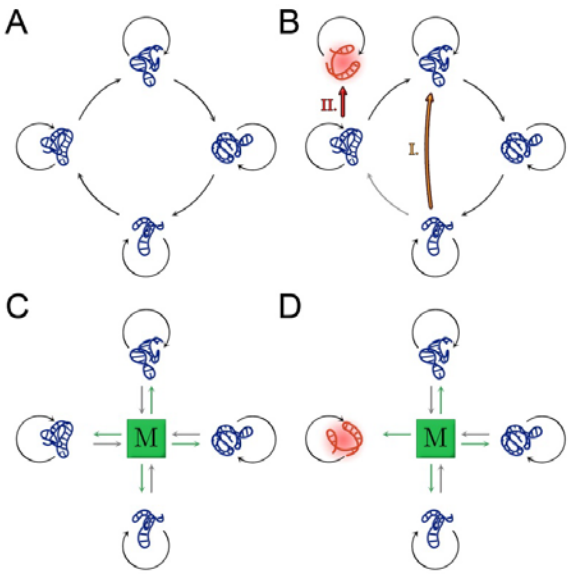

Figure 2. The interaction between replicators can be direct or indirect. A*: In the hypercycle model, replicators catalyse the replication of the next molecule, thus they have a direct effect on the replicative success of another member of the replicator community.* B*: Parasites of the hypercycle interrupt the cooperation of replicators either by creating shortcuts (I.) or by accepting catalytic help without reciprocating it (II.).* C*: The Metabolically Coupled Replicator System is built on the assumption that the interaction among members of the replicator community is indirect, all members contribute to a common metabolism that replicators feed on.* D*: A parasite in the Metabolically Coupled Replicator System consumes products of the common metabolism without contributing to production*

There are, however, two major issues with the hypercycle. Firstly, if any of the members is diluted due to stochastic effects it can ruin the whole hypercycle when running at low concentrations (*cf.* demographic stochasticity). Secondly, the original model has a serious oversight not including mutations[122,123], leaving thus an enormous theoretical gap of explaining the evolutionary origin and survival of the hypercycle in a biologically relevant way. Allowing mutations in a hypercycle can give rise to various mutants: any “selfish” mutant that is a better target for replication will destroy the hypercycle by channelling resources out of the cooperative cycle (towards the parasites), see Figure 2B. Furthermore,

short-circuit mutants introduce shortcuts in the reaction loop, severing cut-off members of the original cycle, reducing thus the diversity maintained. Moreover, any mutant with better catalytic activity would not increase the efficiency of the system as they are not evolutionary units and cannot be selected for[122-124]. As a conclusion, hypercycles are not able to overcome the danger of information decay: they cannot compete against harmful parasites. Moreover, while members of the hypercycle can be units of evolution, the cycle as a whole is not as it is not subject to selection with heritable variability[125,126].

The classical ecological solution to temper the invasion of parasites is to assume local effects (spatially explicit models) which however could only provide a partial solution for the hypercycle[127]. In the model of Boerlijst and Hogeweg, local replicator interactions produce moving spiral waves in which selfish parasites move out to the edge of spiral-arms to finally die out. In a specific case, if parasites are dropped exactly into the centre of spiral waves, they can survive in an inert "cyst". Unfortunately, even if the selfish parasite is contained, the hypercycle cannot be maintained stably when short-cut parasites appear, neither in spatially implicit nor in explicit models[127]. A further problem with this mechanism to save the hypercycle is that it is extremely fragile. Random perturbation in the adhesion of the replicators in the different patches of the surface ruins the spirals and with them goes away the resistance against parasites[128].

On the experimental side it should be noted that contrary to erroneous claims no instantiation of the molecular hypercycle has been realized (Szathmáry[70] presents a survey of the propagation of this conceptual error). In the hypercycle replication is a second-order process: template *replication* is catalytically aided by the previous member in the cycle. In cross-catalytic systems members aid the *formation* rather than the replication of other members. The first such system was realized in the von Kiedrowski lab[129] a recent, more complex example was presented by[69]. An important question is the evolvability of such systems exactly because template replication is *not* a component process.

II/2. Surface-bound replicators

Surfaces, besides their favourable kinetic and thermodynamic effects on an unfolding chemical network[130,131] have an important role in providing population structures in which evolution is known to proceed differently from its course in a well-mixed flow reactor (*cf.* Figure 3). A potential interaction network was explored by the Metabolically Coupled Replicator System (MCRS[132], see Figure 2C). Replicators in the MCRS interact with each other *indirectly*, namely every replicator catalyses only one reaction in a hypothetical metabolic reaction network carrying out monomer production, but all of the replicators are essential, otherwise monomer production breaks down. Moreover, replicators compete for monomers, and replicators with higher replication rate can utilize monomers faster and can become dominant in the system. In the spatially implicit version of the MCRS, there is no compensatory mechanism against superior replicators, therefore they competitively exclude all other replicators and the metabolic, and hence the replicator, system collapses[132]. In the spatially explicit model, however (called the Metabolic Replicator Model, MRM), replicators stably coexist in most parts of the parameter space[132,133]. Local interactions and limited mixing of replicators in the spatially explicit model ensures that the metabolic network is more likely to be complete in the neighbourhood of rare replicators than in the vicinity of dominant replicators (see Figure 3B), providing thus a control over the dominant species (*advantage of rarity*)[132,133].

The MCRS has a double advantage against parasites over the hypercycle. Since the main coupling is indirect via metabolism, the short-circuit parasite (in contrast to the hypercycle case) has no meaning, see Figure 2D. Moreover, harmful effects of parasites occur only locally in the MRM: parasites overwhelm—due to their higher replication rate—their own *local* metabolic community and break the metabolic process down terminating their own replication as well (*cost of commonness*). As long as parasites are able to "infect" new local metabolic communities, they coexist permanently with metabolic replicators[132,133].

Consequently, the MCRS has the ability to incorporate a new replicator (i.e. a new functionality) as long as it does not impair the established metabolic process, therefore a new replicator, even being parasitic, can permanently coexist with the metabolic replicators. Moreover, evolution is able to "transform" parasites into beneficial members of the system without inhibiting the metabolic process[134].

A related model concerns the spread of efficient replicase ribozymes on surfaces. In a well-mixed case shorter, dysfunctional molecules would displace longer-competent replicases by the virtue of their faster replication rate. This is not so in the surface model of Szabó *et al.*[135] with limited diffusion. Local accumulation of parasites is self-limiting, since in such a patch an average parasite finds only other parasites around itself and thus cannot replicate. In the model, elongation activity and accuracy as enzyme and replication rate as template are in a three-way trade-off. Despite this severe constraint a stable, bimodal distribution of short parasites and long replicases emerges as a result of simulated evolution.

To summarize, local indirect interactions and limited mixing of replicators are required for the coexistence of genes. The presence of local interactions is one of many properties linking theoretical and experimental prebiotic approaches. Mineral surfaces could have played an influential role in the evolution of prebiotic information-carrying molecules at multiple levels. They may have been responsible for the homochirality of nucleotides[136,137], may have catalysed the polymerisation of monomers[50], and may have protected polymers from degradation[138]. The properties of mineral surfaces coupled with the theoretical demonstration of potential replicator coexistence hints that life may have originated on surfaces, most probably without a soup phase (albeit chemical intermediates could have formed in the prebiotic ocean, or even in the atmosphere [139-141]).

II/3. Active compartmentalization

Surface-bound replicators could have kick-started life, but the number of coexisting replicators, hence metabolic complexity, was limited[133]. Compartments provide a more articulated population structure and it has further advantages by effectively increasing local concentrations within the small volume of cells compared to free solutions, which significantly improves the efficiency of (bio)chemical reactions[142]; and it can provide an efficient way to spatially segregate different genomes composed of several unlinked replicators[143]. Surface bound models often assume that small molecules produced locally do not diffuse, or do not diffuse faster than the macromolecules catalysing their formation. This is unrealistic since the small molecular products probably leak from the system. Properly compartmentalized catalysts can benefit from the products of their own reactions or that of their cooperative partners.

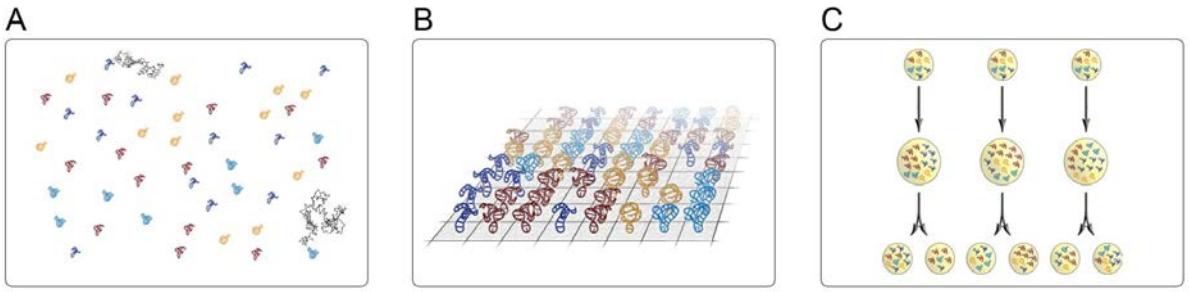

Figure 3. Schematic representation of population structures in models of prebiotic replicators. A: In a well-mixed model, no population structure is assumed, hence free movement of molecules is allowed. B: In the case of surface-bound molecules, the replicators have a limited number of immediate neighbours and their translocation (dispersal, diffusion) is limited. C: The stochastic corrector model assumes that replicators multiply inside vesicles with a membrane boundary, and successful replication of the community will accelerate vesicle growth and division, which defines the fitness of these protocells.

The transition from surface bound to compartmentalized replicators is the first major evolutionary transition[144]. So far we have only considered RNA replicators, and while RNA can fulfil both the roles of template and catalyst, it cannot form membranes. The first membranes were most probably single-chain fatty acids[143,145-148]. Once simple amphiphilic molecules were present, spontaneous formation of vesicles became possible under specific circumstances[145,147,149]. Interestingly, spontaneous membrane assembly is also catalysed by certain mineral surfaces[150], offering a conceptual bridge between the surface bound and encapsulated replicators in the evolution of the RNA world. A feasible mechanism to encapsulate macromolecules into compartments was proposed by Deamer's lab[151,152]. A drying-wetting cycle, in which empty compartments and macromolecules (*e.g.* RNA) are mixed in a solution which is dried, in which phase the compartments dehydrate and produce multilamellar structures with macromolecules among the layers. Then in the wetting phase the compartment re-hydrated which means new compartment are formed encapsulating macromolecules.

Compartmentalisation of the individually replicating genes provided the basis of the first living cells. A minimal living system fulfilling all criteria of life was proposed by Gánti[10,153,154]. The *chemoton* (since its reconceptualization in 1975) has a membrane, an information subsystem and a metabolism. In 2001, Szostak has proposed very similar construct, the *ribocell*[155,156], in which one ribozyme synthesizes the membrane components and another is responsible for genome replication. Remarkable experimental advances have been made in recent years toward the *in vitro* realization of such minimal system[145,157]. *In silico* investigations of (proto)cells could also provide valuable insight to the problems faced by these early systems: how the lipid bilayer could self-assemble from the metabolic products of the vesicle[158]; how membrane permeability affects metabolism[159], or how vesicles transform and divide[160]? Furthermore, compartmentalized models can address the problem of maintaining the genetic information and the effect of parasites—these will be discussed in turn.

In order to understand the mechanisms behind the coexistence of genes in a compartmentalized system a simple yet effective model, the stochastic corrector model was proposed[161,162]. In the model, it is assumed that encapsulated replicators catalyse their own replication and the growth of the membrane (and thus the cell as a whole) as well. As a further natural assumption, competition is allowed among replicators. There is an optimal replicator composition that yields the fastest cell growth. The cells are in selective disequilibrium, maintaining thus a variety of different compositions. At a critical size, cells undergo fission and form two daughter cells with random segregation of replicators. Because the replicators replicate individually, faster growing ones can be overrepresented in the offspring. Several mechanisms act against coexistence, such as the internal competition of the replicators for monomers, their competition for the replicase, and the potential for gene loss due to random assortment of genes to daughter cells. Nevertheless, both the stochasticity of replication dynamics and the stochasticity of cell division increases variability among the cells, and thus selection can act on this variation. It was shown that the stochastic nature of the daughter cell composition in fact facilitates coexistence, as by chance, daughter cells could inherit a balanced gene set, even if the parental cell had a

suboptimal distribution of genes[161]. Hence the name of the model: stochastic corrector model (SCM, see Figure 3C). This process protects the population from extinction and results in evolutionary dynamics yielding a stable quasispecies at the level of compartments[162]. The SCM is inherently stable against parasites: it has the ability to select *against* inferior and select *for* superior mutants[163-166]. The cells of the SCM are individuals subject to selection and are thus evolutionary units. Selection at the level of compartments can be considered as group selection[167,168] since (*i*) the number of templates in protocells is much smaller than the number of compartments, (*ii*) each protocell has only one parent, and (*iii*) there is no migration among groups. Compartmentalization can save also the hypercycle[169-171]: if a favourable mutant appears in a compartment, after random segregation of templates into daughter cells there is a chance of appearance of a superior template composition that can outcompete compartments with inferior compositions. Consequently, the compartmentalized HPC can be the subject of selection and can integrate information successfully. Note, however, that the reaction topology assumed in the Metabolically Coupled Replicator System tolerates higher mutational loads than a compartmentalized hypercycle of similar gene diversity[172], making the hypercycle still a less favoured model of early information integration.

The above models mostly focused on the coexistence of only a few (1-3) genes in the face of stochasticity and parasites. An important question arises next: how many genes can actually coexist within compartments? In an infinite population with replicators having exactly the same replication rate (*i.e.* there is no internal competition), arbitrary number of genes can coexist[173]. Finite population and internal competition however leads to a finite maximum maintainable gene number. Recently Hubai and Kun[174] have shown that as much as 100 genes can coexist within the SCM. We will discuss in turn if 100 genes are enough for a minimal ribo organism. Nevertheless, we can conclude that the compartmentalized systems are not just stable solutions against parasites but are also capable information integrators and are units of evolution.

III. Enhancing the RNA world: Chromosomes and metabolism

> *"Biologists must first of all be concerned with this chemical motor, since the system of chemical cycles is the basis of the functioning of life."*
> *(T. Gánti: The principle of life, 1987).*

Limited diffusion or compartmentalization allows for some genes to coexist. While we have argued that a single RNA dependent RNA replicase is sufficient for the start of evolution, a functional ribo organism requires considerably more enzymes. Comparative analyses of bacteria and theoretical considerations place the minimal set of genes for a present day organism to around 200. Many of these enzymes are involved in translation and DNA metabolism[175], and thus are not required for a ribo organism. Even among the 50 genes suggested as the minimum for an intermediate metabolism, we find genes for the conversion of ribonucleotides to deoxyribonucleotides. The minimal gene set for a functional ribo-organism might lie in the 60-100 range. Compartmentalized systems can harbour such a diverse set of genes. The evolution of the chromosome is the next step toward more complex life forms, as internal competition and the threat of stochastic loss of genes limits the number of individually replicable genes (see above). The invention of the chromosome allows the further expansion of metabolism to the point where the evolution of the genetic code and then translation become feasible. This section discusses the challenges the RNA world faced in its evolution from the first cells through the evolution of complex metabolism to the RNA-protein world.

III/1. Chromosomes

The quest for the chromosome, a single RNA molecule containing all necessary genes of the organism, has been behind the scenes in the previous sections. Once the replication apparatus can copy RNA with all genes simultaneously and with sufficient fidelity, the problem of individually replicating genes and the Eigen's paradox is gone. As an addition to a highly accurate replicase, a chromosome requires an endonuclease that cleaves the embedded genes (ribozymes) from it. Evolving RNA cleavage capability is not very complicated, in fact all naturally occurring ribozymes can do it[21]. Simple structural motifs exhibited by the hairpin or the hammerhead ribozymes, the smallest natural ribozymes, are common even in random pools of short RNA molecules[54]. Thus, chromosomes will be able to evolve once replication is accurate enough (see [176] for plausible molecular steps leading to the establishment of chromosomes from unlinked, replicating ribozymes).

This evolutionary transition probably did not result in (bacterial) chromosomes as we know them now: a single copy of genes per cell which precisely double for cell division. An intermediate evolutionary step could have been individually replicating chromosomes, when all genes are linked together. In such an ensemble, no one can cheat by replicating faster than the others. Furthermore the inheritance of a full set of genes is ensured if at least one copy of the chromosome gets into the daughter cell. The reliable allocation of two chromosomes to the two daughter cells requires a separator mechanism such as the one provided by the cytoskeleton, which is probably a late prokaryotic invention. Without the evolved facilities of a cytoskeleton-like system, multiple copies of chromosomes might still be able to ensure that no daughter cells end up missing any gene. We can argue that if there are higher number of copies of the chromosome, and chromosomes are assigned to daughter cells randomly, both daughter cells will have at least one set of genes with high probability. For example, at seven copies per chromosome, the chance of an empty daughter cell is less than 1% (binomial distribution, with $p = 0.5$). A system with randomly assorting chromosomes can actually outcompete cells with individually replicating genes[177]. But the aforementioned seven copies per cell are still more than the two copies required for cell division. One of them goes to one of the

daughter cells, and the other to the other daughter cell. As simple as it sounds, it either requires a cytoskeleton (like in contemporary organisms[178,179]), or attachment to the cell membrane (as in the replicon model[180]). As RNA polymerases are powerful motors, they could have exerted force when travelling along the strand being copied. As the membrane was growing simultaneously, it could have aided the segregation of chromosomes by letting them move to opposite poles of the early cell (*cf.* bacterial chromosome segregation[181,182]). Thus an RNA polymerase ribozyme could have, besides replicating the genetic material, pushed the two copies to opposite ends of the cell, ensuring cell division.

The main obstacle in the path to a chromosome is the error threshold. When gradual increase of the fidelity of replicase overcame the critical threshold above which the whole genetic material could be replicated as one molecule, chromosome evolved. The evolution of accurate chromosome segregation and bacterial-type cell division remains to be elucidated.

III/2. Enzymatization

Metabolism is the fundamental core function of the living cell[31,153,183]. As we have argued previously (see Section II/2.) there must have been a small but essential set of molecules that catalysed a minimal metabolism, even at the surface bound stage. At least a minimal metabolism was required for RNA replication. Metabolism, however, might sound like a bewilderingly complex network of reactions, as it usually is for contemporary species. How could have evolution proceeded then from a few coexisting genes catalysing their own replication to a complex and intertwined metabolism, with multitude of specialized enzymes? There are three main angles that aim to explain the origins of a complex metabolism: (*i*) discovering the catalytic repertoire of ribozymes, (*ii*) assembling reaction-networks, and (*iii*) understanding the increasing specificity of enzymes.

Firstly, the catalytic repertoire of ribozymes shows that almost all reactions necessary for a ribo-organism can be catalysed by ribozymes. The real challenge is to develop an efficient and accurate replicase ribozyme. Unfortunately, at the moment there is no known ribozyme that can stably replicate itself. On the other hand, template directed polymerization was proved to be possible[95,184] albeit only up to 14-20 nucleotides could be copied. The copying fidelity of these ribozymes is around 96% per nucleotide per copying. Efficiency was further enhanced to be able to copy 98 nucleotides[96] with accuracy increased to 99%. These experiments lend credence to theoretical models that the gradual refinement of copying fidelity is possible and the error threshold can be overcome[135,185]. The latest ribozyme artificially generated is able to copy a longer albeit very specific template[186]. Recent advances seemed to indicate that a self-replicating ribozyme is just around the corner, although the research took almost two decades, which also indicates that a general replicase ribozyme is not something easily evolved.

Surface-bound metabolism can enhance the formation of RNA strands. Apart from the replicase, the availability of nucleotides is critical. Let us assume that nucleobases and ribose are available from the environment. In order to form activated monomers the sugar needs to be phosphorylated twice and then the constituents put together to form the nucleotide. Kinase ribozymes[187] could produce the *D*-ribose-5-phosphate and then 5-phospho-*D*-ribose-1-diphosphate (PRPP). Ribozymes can catalyse the formation of the glycosidic bond between PRPP and a pyrimidine[188-190] or a purine[190,191] nucleobase. Moreover, almost all biologically important reactions could be catalysed by ribozymes to some extent[36].

Upon leaving the mineral surface, replicators were probably encapsulated into vesicles. Compartmentalization raised new problems, for example that of permeability: how could small molecules (raw and waste) cross the membrane? Although RNA molecules cannot be proper transmembrane molecules as they lack a hydrophobic part, it is possible to select for oligonucleotide sequences that efficiently bind to membranes[42,43,192-195], presumably in the form of collaborative hetero-oligomeric complexes[43,192,193]. These complexes can significantly change the permeability of membranes for larger ionic compounds[42,192,196], and serve as specific transporters for more complex compounds such as amino acids[43]. Nucleotides can spontaneously diffuse across fatty-acid membranes[145,197]. Interestingly, ribose has the best permeability coefficient among aldopentoses and hexoses, both for fatty acid and phospholipid membranes, which promotes its accumulation within the protocells[198]. If one considers the formose reaction[199] as a possibly prebiotic pathway for autocatalytic carbohydrate synthesis, such passive sorting and accumulation of ribose, one of many products of the formose reaction, in membrane bound vesicles could have supplied ribose for nucleotide synthesis[198]. Consequently, evidence suggests that when the evolution of the RNA world arrived at the stage of compartmentalized replicators, scenarios considering RNA molecules as mediators of transmembrane transport proved to be possible.

Secondly, four reaction pathway-evolution scenarios are known. According to the *backward (or retrograde)* evolutionary scenario[200] the last step in a pathways leading to important molecules were enzymatized first. Only pathways that operate without enzymes can be populated by enzymes this way. The last product will be depleted first, and then the last but one, and so on. Cells evolving an enzyme for the last non enzymatic step have an advantage as they can secure resources faster. The *forward pathway evolution* postulates that enzymes appear first for the early steps of a pathway, and later steps become catalysed later in succession[201]. Such an evolutionary scenario could work for catabolic pathways, in which more and more energy can be extracted by successive processing of a molecule. The *patchwork evolution* postulates that enzymes are recruited from other pathways [202]. And finally, the *shell hypothesis* proposes that there was a core metabolic process (*e.g.* the reductive citric acid cycle) and new pathways may have been recruited and attached to this core [203]. Obviously, these scenarios cannot be entirely separated from each other, they may all have played essential roles in the evolution of metabolic-reaction networks [204].

The third problem is the evolution of enzymatic efficiency, which raises further problems both from the biochemical and from theoretical point of view. First and foremost, modelling the evolution of enzymes is a challenging task. If the crystal structure of a given (protein-) enzyme is given, the interaction between the enzyme and a small molecule as a ligand can be analysed either on quasi-classical or on quantum-mechanical level. Since the structure of early enzymes is unknown, and the structural-functional evolution of the enzymes on the molecular level cannot be modelled, these approaches fail.

There are two possibilities to overcome the hurdles of modelling the evolution of specific enzymes: either by using a fully artificial chemistry or applying a major simplification of real chemical structures that preserves the major properties of the receptor-ligand interactions. In artificial chemistry approaches, atomic-types, chemical bonds, reaction routes and the interaction between molecules are defined in arbitrary but consistent ways. Dittrich *et al.* [205] have argued that "*artificial chemistries are 'the right stuff' for the study of prebiotic and biochemical evolution*". Such chemistries are applicable for a wide variety of models (from biochemical to ecological systems) with a continuously growing literature. Setting up an artificial chemistry model for studying the evolution of enzymes is a straightforward task (see *e.g.* [206]). In this context, the increasing chemical-functional complexity, the interactions between molecules and the analysis of the system can be handled in a relatively easy way, although the relevance of any results so obtained is at least doubtful. We suggest that such models are nevertheless useful to understand larger-scale phenomena, like metabolic network expansion or self-assembly, in which the abstraction of individual reactions does not affect the behaviour of the system.

The second method of modelling is to reduce the complexity of the structure of real enzymes and to simplify the treatment of the receptor-ligand interaction. This approach could capture the essential features of both the evolution of enzyme-functions and the thermodynamics of the receptor-ligand interaction. An early study using the above method is done by Kacser and Beeby[207]. They approximated enzymes and ligands with 3D cavities, and blocks fitting in cavities. The enzymatic activity is assumed to be proportional to the Lennard-Jones interaction energy between enzyme and substrate (if the substrate can enter into the cavity, zero otherwise). With this choice there is one optimal enzyme size for a given substrate. This approach respects both the effect of the geometry of the participants and the basics of the thermodynamics of interactions. During the evolution, the enzyme sizes can change altering the catalytic activity on a given substrate. A possible extension was made by Szathmáry and colleagues[208]. In this model substrates and enzymes are $D > 3$ dimensional hyper-blocks and cavities with "active sites" on their faces. Instead of increasing the complexity of 3D structures, introducing higher dimensionality provides a way to model the geometrical complexity of enzymes much easily. For proper catalysis, the active sites must meet their complementary partners (otherwise the catalytic product is waste) and for high catalytic activity, enzyme cavities must optimally fit the size of their substrates.

Based on such a model, Szathmáry and colleague[208] have concluded that the formation of a chromosome is a prerequisite for complex metabolism run by specific enzymes. The reason for this is that while replicating ribozymes are unlinked, there is a considerable assortment load due to chance in protocell division (genes are assorted to offspring compartments randomly) which selects for generalist enzymes at the expense of specificity and efficiency. Small metabolic repertoire and promiscuous ribozymes (*e.g.*[190]) were the norm. The invention of the chromosome seems to be the pinnacle of the RNA world, as from its very beginning it was striving for this elusive target (see Section I/3.), but its invention paves the road out of the RNA world. In contrast to peptide synthesis, ribozyme production required only RNA copying: the genetic material is copied to produce ribozymes, and ribozymes or a chromosome harbouring them is copied to replicate the genetic material. Peptide enzymes are more efficient catalysts but their production requires many enzymes (the ribosome, tRNAs and aminoacyl tRNA synthetases). The evolution of life could have arrived to the proliferation of such enzymatic activity at this stage of complexity, which we can refer to as the peptide-RNA world.

What remains to be done

The dynamical theory of the RNA world has advanced considerably over the last two decades. Of course the existence of the RNA world is taken for granted that there once upon a time an RNA world did in fact exist. As Orgel[47] and Joyce[35] note it is quite likely that RNAs were not the first replicating templates. It is also certain that they were not the last either: today we are living in a DNA-RNA-protein world. What are the main goals for dynamical theory in the further clarification of the evolution of the RNA world? Maybe template replication was preceded by collective autocatalysis of molecules lacking template replication at all. This view, forcefully advocated by Kauffman [209,210] received surprising support by the demonstration of limited evolvability of such networks in compartmentalized form[211]. Computationally demanding further examination of such systems may turn out to be very important, but a survey of the relevant details has been beyond the scope of the present review.

More detailed and integrated models of protocells harbouring ribozymes is needed, extending our view towards the evolutionary build-up of a complex, connected metabolism and the establishment of resilient membranes with regulated permeability. The name of the game is undoubtedly detailed modelling of coevolution of metabolism, membrane and templates, much in the spirit of Gánti's chemoton concept.

The RNA world has been left behind by evolution. One could argue that the origin of the genetic code and translated protein enzymes was the greatest, yet in a sense self-defeating invention of the RNA world. How this could have happened and what role theory can have in the elucidation of this

important evolutionary transition will be the subject of a different review.

Acknowledgements

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]. ASz and 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). BK acknowledges financial support from the Hungarian Research Foundation (OTKA Grant No. K100806). GB and ÁK acknowledge support 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".

Notes and references

1. W. Gilbert, *Nature*, 1986, **319**, 618.
2. A. Lazcano, *Cold Spring Harbour Perspective in Biology*, 2010, **2**, a002089.
3. A. N. Belozerskii, in *The Origin of Life on Earth*, eds. A. I. Oparin, A. G. Pasynskii, A. E. Braunshtein and T. E. Pavloskaya, Pergamon Press, New York, 1959, pp. 322–321.
4. J. Brachet, in *The Origin of Life on Earth*, eds. A. I. Oparin, A. G. Pasynskii, A. E. Braunshtein and T. E. Pavloskaya, Pergamon Press, New York, 1959, pp. 361–367.
5. F. H. C. Crick, *The Symposia of the Society for Experimental Biology*, 1958, **12**, 138-163.
6. C. R. Woese, *The Genetic Code*, Harper & Row, New York, 1967.
7. L. E. Orgel, *J. Mol. Biol.*, 1968, **38**, 381-393.
8. F. H. C. Crick, *J. Mol. Biol.*, 1968, **38**, 367-379.
9. T. Gánti, *Biológia*, 1979, **27**, 161-175.
10. T. Gánti, *Chemoton Theory*, Kluwer Academic/Plenum Publishers, New York, 2003.
11. C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace and S. Altman, *Cell*, 1983, **35**, 849-857.
12. K. Kruger, P. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling and T. R. Cech, *Cell*, 1982, **31**, 147-157.
13. D. C. Jeffares, A. M. Poole and D. Penny, *J. Mol. Evol.*, 1998, **46**, 18-36.
14. C. L. Peebles, P. S. Perlman, K. L. Mecklenburg, M. L. Pertillo, J. H. Tabor, K. A. Jarrell and H.-L. Cheng, *Cell*, 1986, **44**, 213-223.
15. A. C. Forster and R. H. Symons, *Cell*, 1987, **49**, 211-220.
16. A. Hampel and R. R. Tritz, *Biochemistry*, 1989, **28**, 4929-4933.
17. L. Sharmeen, M. Y. P. Kuo, G. Dinner-Gottlieb and J. Taylor, *Journal of Virology*, 1988, **62**, 2674-2679.
18. B. J. Saville and R. A. Collins, *Cell*, 1990, **61**, 685-696.
19. W. C. Winkler, A. Nahvi, A. Roth, J. A. Collins and R. R. Breaker, *Nature*, 2004, **428**, 281-286.
20. A. Roth, Z. Weinberg, A. G. Y. Chen, P. B. Kim, T. D. Ames and R. R. Breaker, *Nature Chemical Biology*, 2014, **10**, 56-60.
21. J. A. Doudna and T. R. Cech, *Nature*, 2002, **418**, 222-228.
22. E. Westhof, *Science*, 1999, **286**, 61-62.
23. P. Nissen, J. Hansen, N. Ban, P. B. Moore and T. A. Steitz, *Science*, 2000, **289**, 920-930.
24. P. B. Moore and T. A. Steitz, *Nature*, 2002, **418**, 229-235.
25. S. E. Butcher, *PNAS*, 2009, **106**, 12211-12212.
26. S. Valadkhan, A. Mohammadi, Y. Jaladat and S. Geisler, *PNAS*, 2009, **106**, 11901-11906.
27. H. B. White, *J. Mol. Evol.*, 1976, **7**, 101-104.
28. L. Orgel, *J. Mol. Evol.*, 1989, **29**, 465-474.
29. V. R. Jadhav and M. Yarus, *Biochimie*, 2002, **84**, 877-888.
30. D. Saran, J. Frank and D. H. Burke, *BMC Evolutionary Biology*, 2003, **3**, 26.
31. Á. Kun, B. Papp and E. Szathmáry, *Genome Biology*, 2008, **9**, R51.
32. E. Szathmáry, *Nature*, 1990, **344**, 115.
33. E. Szathmáry, *Oxford Surveys in Evolutionary Biology*, 1989, **6**, 169-205.
34. C. Tuerk and L. Gold, *Science*, 1990, **249**, 505-510.
35. G. F. Joyce, *Nature*, 2002, **418**, 214-220.
36. X. Chen, N. Li and A. D. Ellington, *Chemistry & Biodiversity*, 2007, **4**, 633-655.
37. A. D. Ellington, X. Chen, M. Robertson and A. Syrett, *The International Journal of Biochemistry & Cell Biology*, 2009, **41**, 254-265.
38. L. F. Landweber, P. J. Simon and T. A. Wagner, *BioScience*, 1998, **48**, 94-103.
39. G. F. Joyce, *Annual Review of Biochemistry*, 2004, **73**, 791-836.
40. S. Tsukiji, S. B. Pattnaik and H. Suga, *Nat Struct Mol Biol*, 2003, **10**, 713-717.
41. S. Tsukiji, S. B. Pattnaik and H. Suga, *Journal of American Chemical Society*, 2004, **126**, 5044-5045.
42. A. Khvorova, Y.-G. Kwak, M. Tamkun, I. Majerfeld and M. Yarus, *PNAS*, 1999, **96**, 10649-10654.
43. T. Janas, T. Janas and M. Yarus, *RNA*, 2004, **10**, 1541-1549.
44. S. A. Benner, A. D. Ellington and A. Tauer, *PNAS*, 1989, **86**, 7054-7058.
45. E. Szathmáry, S. Mauro and C. Fernando, *Topics in Current Chemistry*, 2005, **259**, 167-211.
46. M. P. Robertson and G. F. Joyce, *Cold Spring Harbor Perspectives in Biology*, 2012, **4**, a003608.
47. L. E. Orgel, *Critical Reviews in Biochemistry and Molecular Biology*, 2004, **39**, 99-123.
48. M. W. Powner, B. Gerland and J. D. Sutherland, *Nature*, 2009, **459**.
49. W. Huang and J. P. Ferris, *Chem Commun (Camb)*, 2003, 1458-1459.
50. J. P. Ferris, *Philos. T. R. Soc. B.*, 2006, **361**, 1777-1786.
51. W. Ma, C. Yu, W. Zhang and J. Hu, *RNA*, 2007, **13**, 2012-2019.
52. S. C. Manrubia and C. Briones, *RNA*, 2006, **13**, 97–107.
53. S. D. Copley, E. Smith and H. J. Morowitz, *Bioorganic Chemistry*, 2007, **35**, 430-443.
54. C. Briones, M. Stich and S. C. Manrubia, *RNA*, 2009, **15**, 743-749.
55. M. Fedor, *J. Mol. Biol.*, 2000, **297**, 269-291.
56. D. A. Lafontaine, D. G. Norman and D. M. J. Lilley, *Biochem. Soc. Trans.*, 2002, **30**, 1170-1175.
57. C. Haslinger and P. F. Stadler, *Bull. Math. Biol.*, 1999, **61**, 437-467.
58. P. Schuster, W. Fontana, P. F. Stadler and I. L. Hofacker, *Proc. Roy. Soc. Lond. B*, 1994, **255**, 279-284.
59. P. Schuster, *Biophysical Chemistry*, 1997, **66**, 75-110.
60. J. Geveretz, H. H. Gan and T. Schlick, *RNA*, 2005, **11**, 853-863.

61. W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler and P. Schuster, *Monatsh Chem*, 1996, **127**, 375-389.
62. W. Fontana, D. A. M. Könings, P. F. Stadler and P. Schuster, *Biopolymers*, 1993, **33**, 1389-1404.
63. T. Jörg, O. Martin and A. Wagner, *BMC Bioinformatics*, 2008, **9**, 464.
64. C. Fernando, G. Von Kiedrowski and E. Szathmáry, *J. Mol. Evol.*, 2007, **64**, 572–585.
65. M. Yarus, *Life from an RNA World: The Ancestor Within*, Harvard University Press, Harvard, USA, 2011.
66. A. D. Ellington and J. W. Szostak, *Nature*, 1992, **355**, 850-852.
67. G. von Kiedrowski, *Angew. Chem. Int. Ed. Engl.*, 1986, **25**, 932-935.
68. N. Paul and G. F. Joyce, *PNAS*, 2002, **99**, 12733-12740.
69. N. Vaidya, M. L. Manapat, I. A. Chen, R. Xulvi-Brunet, E. J. Hayden and N. Lehman, *Nature*, 2012, **491**, 72-77.
70. E. Szathmáry, *Journal of Systems Chemistry*, 2013, **4**, 1.
71. E. J. Hayden, G. von Kiedrowski and N. Lehman, *Angew. Chem. Int. Ed. Engl.*, 2008, **47**, 8424-8428.
72. V. Vasas, C. Fernando, A. Szilágyi, I. Zachár, M. Santos and E. Szathmáry, *System Chemistry*, 2014, **accepted**.
73. A. J. Meyer, J. W. Ellefson and A. D. Ellington, *Acc Chem Res*, 2012, **45**, 2097-2105.
74. G. F. Gause, *The Struggle for Existence*, William and Wilkins, Baltimore, 1935.
75. R. MacArthur and R. Levins, *PNAS*, 1964, **51**, 1207-1210.
76. G. Meszéna, M. Gyllenberg, L. Pásztor and J. A. J. Metz, *Theor. Pop. Biol.*, 2006, **69**, 68-87.
77. M. Eigen, *Naturwissenscaften*, 1971, **10**, 465-523.
78. A. Szilágyi, I. Zachar and E. Szathmáry, *PLoS Computational Biology*, 2013, **9**, e1003193.
79. W. S. Zielinski and L. E. Orgel, *Nature*, 1987, **327**, 346-347.
80. E. Szathmáry and I. Gladkih, *J. Theor. Biol.*, 1989, **138**, 55-58.
81. S. Lifson and H. Lifson, *J. Theor. Biol.*, 1999, **199**, 425-433.
82. E. Szathmáry, *TREE*, 1991, **6**, 366-370.
83. Z. Varga and E. Szathmáry, *Bull. Math. Biol.*, 1997, **59**, 1145-1154.
84. G. von Kiedrowski and E. Szathmáry, *Selection*, 2000, **1**, 173–179.
85. P. R. Wills, S. A. Kauffman, B. M. R. Stadler and P. F. Stadler, *Bull. Math. Biol.*, 1998, **60**, 1073-1098.
86. I. Scheuring and E. Szathmáry, *J. Theor. Biol.*, 2001, **212**, 99-105.
87. P. Chesson, *Theor. Pop. Biol.*, 1994, **45**, 227-276.
88. P. Chesson, *Theor. Pop. Biol.*, 2000, **58**, 211-237.
89. M. Scheffer, S. Rinaldi, J. Huisman and F. Weissing, *Hydrobiologia*, 2003, **491**, 9-18.
90. J. Huisman and F. J. Weissing, *Ecological Research*, 2002, **17**, 175-181.
91. G. Károlyi, Á. Péntek, I. Scheuring, T. Tél and Z. Toroczkai, *PNAS*, 2000, **97**, 13661-13665.
92. J. B. S. Haldane, *Am. Nat.*, 1937, **71**, 337-349.
93. M. Eigen and P. Schuster, *Naturwissenscaften*, 1977, **64**, 541-565.
94. E. C. Friedberg, G. C. Walker and W. Siede, *DNA repair and mutagenesis*, ASM Press, Washington D.C., 1995.
95. W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasen and D. P. Bartel, *Science*, 2001, **292**, 1319-1325.
96. A. Wochner, J. Attwater, A. Coulson and P. Holliger, *Science*, 2011, **332**, 209-212.
97. J. Maynard Smith, *Proc. Roy. Soc. Lond. B*, 1983, **219**, 315-325.
98. E. Baake and W. Gabriel, *Annual Reviews of Computational Physics VII*, 2000, 203-264
99. M. Eigen, J. S. McCaskill and P. Schuster, *Adv. Chem. Phys.*, 1989, **75**, 149-263.
100. I. Leuthäusser, *The Journal of Chemical Physics*, 1986, **84**, 1884.
101. D. B. Saakian and C.-K. Hu, *PNAS*, 2006, **103**, 4935-4939.
102. D. B. Saakian, C. K. Biebricher and C.-K. Hu, *PLoS ONE*, 2011, **6**, e21904.
103. P. Tarazona, *Physical Reviews A*, 1992, **45**, 6038-6050.
104. S. Bonhoeffer and P. F. Stadler, *J. Theor. Biol.*, 1993, **164**, 359-372.
105. M. A. Huynen, P. F. Stadler and W. Fontana, *PNAS*, 1996, **93**, 397-401.
106. W. Fontana and P. Schuster, *Science*, 1998, **280**, 1451-1455.
107. E. van Nimwegen, J. P. Crutchfield and M. A. Huynen, *PNAS*, 1999, **96**, 9716-9720.
108. P. Schuster and P. F. Stadler, in *Origin and Evolution of Viruses*, eds. E. Domingo, R. G. Webster and J. Holland, Academic Press, New York, 1999, pp. 1-24.
109. N. Takeuchi, P. H. Poorthuis and P. Hogeweg, *BMC Evolutionary Biology*, 2005, **5**, 9.
110. C. Reidys, C. V. Forst and P. Schuster, *Bull. Math. Biol.*, 2001, **63**, 57-94.
111. Á. Kun, M.-C. Maurel, M. Santos and E. Szathmáry, in *The aptamer handbook*, ed. S. Klussmann, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2005, pp. 54-92.
112. Á. Kun, S. Mauro and E. Szathmáry, *Nature Genetics*, 2005, **37**, 1008-1011.
113. A. Szilágyi, Á. Kun and E. Szathmáry, *submitted*, 2014.
114. M.-M. Huang, N. Arnheim and M. F. Goodman, *Nucleic Acids Research*, 1992, **20**, 4567-4573.
115. F. W. Perrino and L. A. Loeb, *J. Biol. Chem.*, 1989, **264**, 2898-2905.
116. L. V. Mendelman, J. Petruska and M. F. Goodman, *J. Biol. Chem.*, 1990, **265**, 2338-2346.
117. S. Rajamani, J. K. Ichida, T. Antal, D. A. Treco, K. Leu, M. A. Nowak, J. W. Szostak and I. A. Chen, *J. Am. Chem. Soc.*, 2010, **132**, 5880-5885.
118. N. Lehman, *J. Mol. Evol.*, 2003, **56**, 770-777.
119. M. Santos, E. Zintzaras and E. Szathmáry, *J. Mol. Evol.*, 2004, **59**, 507-519.
120. M. Eigen and P. Schuster, *The Hypercycle: A Principle of Natural Self-organization*, Springer-Verlag, Berlin, 1979.
121. N. Takeuchi and P. Hogeweg, *BMC Evolutionary Biology*, 2007, **7**, 15.
122. J. Maynard Smith, *Nature*, 1979, **280**, 445-446.
123. C. Bresch, U. Niesert and D. Harnasch, *J. Theor. Biol.*, 1980, **85**, 399-405.
124. V. Niesert, D. Harnasch and C. Bresch, *Parasitology*, 1981, **100**, S5-S18.
125. E. Szathmáry, *TREE*, 1989, **4**, 200-204.
126. I. Zachar and E. Szathmáry, *BMC Biology*, 2010, **8**, 21.
127. M. C. Boerlijst and P. Hogeweg, *Physica D*, 1991, **48**, 17-28.
128. I. Scheuring, T. Czárán, P. Szabó, G. Károlyi and Z. Toroczkai, *Origins of Life and Evolution of the Biosphere*, 2003, **33**, 319-355.

129. D. Sievers and G. von Kiedrowski, *Nature*, 1994, **369**, 221-224.
130. G. Wächtershäuser, *Microbiological Reviews*, 1988, **52**, 452-484.
131. G. Wächtershäuser, *Prog. Biophys. Mol. Biol.*, 1992, **58**, 85-201.
132. T. Czárán and E. Szathmáry, in *The Geometry of Ecological Interactions*, eds. U. Dieckmann, R. Law and J. A. J. Metz, Cambridge University Press, Cambridge, 2000, pp. 116-134.
133. B. Könnyű and T. Czárán, *BMC Evolutionary Biology*, 2013, **13**, 204.
134. B. Könnyű, T. Czárán and E. Szathmáry, *BMC Evolutionary Biology*, 2008, **8**, 267.
135. P. Szabó, I. Scheuring, T. Czárán and E. Szathmáry, *Nature*, 2002, **420**, 340-343.
136. P. C. Joshi, M. F. Aldersley and J. P. Ferris, *Origins of Life and Evolution of Biospheres*, 2011, **41**, 213-236.
137. R. M. Hazen, T. R. Filley and G. A. Goodfriend, *PNAS*, 2001, **98**, 5487-5490.
138. E. Biondi, S. Branciamore, M.-C. Maurel and E. Gallori, *BMC Evolutionary Biology*, 2007, **7**, S2.
139. S. L. Miller, *Science*, 1953, **117**, 528-529.
140. S. Miyakawa, H. J. Cleaves and S. L. Miller, *Origins of Life and Evolution of the Biosphere*, 2002, **32**, 195-208.
141. J. P. Ferris, in *Chem. Eng. News*, 1984, pp. 23-35.
142. S. S. Mansy and J. W. Szostak, *Cold Spring Harbor Symp. Quant. Biol.*, 2009.
143. J. P. Schrum, T. F. Zhu and J. W. Szostak, *Cold Spring Harbor Perspectives in Biology*, 2010.
144. J. Maynard Smith and E. Szathmáry, *The Major Transition in Evolution*, W.H. Freeman, Oxford, UK, 1995.
145. S. S. Mansy and J. W. Szostak, *PNAS*, 2008, **105**, 13351-13355.
146. I. Budin and J. W. Szostak, *PNAS*, 2011, **108**, 5249-5254.
147. T. Laiterä and K. Lehto, *Origins of Life and Evolution of Biospheres*, 2009, **39**, 545-558.
148. D. W. Deamer, J. P. Dworkin, S. A. Sandford, M. P. Bernstein and L. J. Allamandola, *Astrobiology*, 2002, **2**, 371-381.
149. I. A. Chen and P. Walde, *Cold Spring Harbor Perspectives in Biology*, 2010, **2**.
150. M. M. Hanczyc, S. M. Fujikawa and J. W. Szostak, *Science*, 2003, **302**, 618-622.
151. D. Deamer, S. Singaram, S. Rajamani, V. Kompanichenko and S. Guggenheim, *Philos. T. R. Soc. B.*, 2006, **361**, 1809-1818.
152. D. W. Deamer, *Microbiology and Molecular Biology Reviews*, 1997, **61**, 239-261.
153. T. Gánti, *The Principle of Life (in Hungarian)*, Gondolat, Budapest, 1971.
154. T. Gánti, *BioSystems*, 1975, **7**, 15-21.
155. J. W. Szostak, D. P. Bartel and P. L. Luisi, *Nature*, 2001, **409**, 387-390.
156. F. Mavelli, *BMC Bioinformatics*, 2012, **13**, S10.
157. K. Adamala and J. W. Szostak, *Science*, 2013, **342**, 1098-1100.
158. F. Mavelli and K. Ruiz-Mirazo, *Philos. T. R. Soc. B.*, 2007, **362**, 1789-1802.
159. G. Piedrafita, K. Ruiz-Mirazo, P.-A. Monnard, A. Cornish-Bowden and F. Montero, *PLoS ONE*, 2012, **7**, e39480.
160. S. Piotto and F. Mavelli, *Origins of Life and Evolution of the Biosphere*, 2004, **34**, 225-235.
161. E. Szathmáry and L. Demeter, *J. Theor. Biol.*, 1987, **128**.
162. D. Grey, V. Hutson and E. Szathmáry, *Proc. Roy. Soc. Lond. B*, 1995, **262**, 29-35.
163. P. Hogeweg and N. Takeuchi, *Origins of Life and Evolution of the Biosphere*, 2003, **33**, 375-403.
164. M. Santos, E. Zintzaras and E. Szathmáry, *Origins of Life and Evolution of the Bioshpere*, 2003, **33**, 405-432.
165. E. Zintzaras, M. Santos and E. Szathmáry, *J. Theor. Biol.*, 2010, **267**, 605-613.
166. N. Takeuchi and P. Hogeweg, *PLoS Computational Biology*, 2009, **5**, e1000542.
167. E. G. Leigh, *PNAS*, 1983, **80**, 2985-2989.
168. D. S. Wilson, *PNAS*, 1975, **72**, 143-146.
169. R. Michod, *American Zoologist*, 1983, **23**, 5-14.
170. M. Eigen, W. C. Gardiner Jr and P. Schuster, *J. Theor. Biol.*, 1980, **85**, 407-411.
171. M. Eigen, P. Schuster, R. Winkler-Oswatitsch and W. Gardiner, *Scientific American*, 1981, **244**, 78-94.
172. E. Zintzaras, S. Mauro and E. Szathmáry, *J. Theor. Biol.*, 2002, **217**, 167-181.
173. J. F. Fontanari, M. Santos and E. Szathmáry, *J. Theor. Biol.*, 2006, **239**, 247-256.
174. A. G. Hubai, in *Plant Systematics, Ecology and Theoretical Biology*, Eötvös Loránd University, Budapest, 2013.
175. R. Gil, F. J. Silva, J. Peretó and A. Moya, *Microbiol Mol Biol Rev.*, 2004, **68**, 518-537.
176. E. Szathmáry and J. Maynard Smith, *J. Theor. Biol.*, 1993, **164**, 447-454.
177. J. Maynard Smith and E. Szathmáry, *J. Theor. Biol.*, 1993, **164**, 437-446.
178. F. Hayes and D. Barilla, *Nature Reviews Microbiology*, 2006, **4**, 133-143.
179. E. Toro and L. Shapiro, *Cold Spring Harbor Perspectives in Biology*, 2010, **2**.
180. F. Jacob, S. Brenner and F. Cuzin, *Cold Spring Harbor Symp. Quant. Biol.*, 1963, **28**, 329-348.
181. J. Dworkin and R. Losick, *PNAS*, 2002, **99**, 14089-14094.
182. K. P. Lemon and A. D. Grossman, *Genes Dev.*, 2001, **15**, 2031-2041.
183. T. Gánti, *The principles of life*, Oxford University Press, Oxford, 2003.
184. H. S. Zaher and P. J. Unrau, *RNA*, 2007, **13**, 1017-1026.
185. I. Scheuring, *Selection*, 2000, **1**, 13-23.
186. J. Attwater, A. Wochner and P. Holliger, *Nature Chemistry*, 2013, **5**, 1011–1018.
187. J. R. Lorsch and J. W. Szostak, *Nature*, 1994, **371**, 31-36.
188. P. J. Unrau and D. P. Bartel, *Nature*, 1998, **395**, 260-263.
189. K. E. Chapelle, D. P. Bartel and P. J. Unrau, *RNA*, 2003, **9**, 1208-1220.
190. M. W. L. Lau and P. J. Unrau, *Chemistry & Biology*, 2009, **16**, 815-825.
191. M. W. L. Lau, K. E. C. Cadieux and P. J. Unrau, *J. Am. Chem. Soc.*, 2004, **126**, 15686-15693.
192. A. Vlassov, A. Khvorova and M. Yarus, *PNAS*, 2001, **98**, 7706-7711.
193. T. Janas and M. Yarus, *RNA*, 2003, **9**, 1353-1361.
194. T. Janas and T. Janas, *Cell Mol Biol Lett*, 2011, **16**, 25-39.

195. T. Janas, T. Janas and M. Yarus, *Nucleic Acids Research*, 2006, **34**, 2128-2136.
196. M. S. Kaucher, W. A. Harrell and J. T. Davis, *J. Am. Chem. Soc.*, 2005, **128**, 38-39.
197. S. S. Mansy, *Cold Spring Harbor Perspectives in Biology*, 2010, **2**.
198. M. G. Sacerdote and J. W. Szostak, *PNAS*, 2005, **102**, 6004-6008.
199. R. Breslow, *Tetrahedron Letters*, 1959, **1**, 22-26.
200. N. H. Horowitz, *PNAS*, 1945, **31**, 153-157.
201. S. Granick, *Annals of the New York Academy of Sciences*, 1957, **69**, 292-308.
202. R. A. Jensen, *Annu. Rev. Microbiol.*, 1976, **30**, 409-425.
203. H. Morowitz, *Complexity*, 1999, **4**, 39-53.
204. E. Szathmáry, *Philos. T. R. Soc. B.*, 2007, **362**, 1781-1787.
205. P. Dittrich, J. Ziegler and W. Banzhaf, *Artificial Life*, 2001, **7**, 225-275.
206. A. Hintze and C. Adami, *Plos Computational Biology*, 2008, **4**, e23.
207. H. Kacser and R. Beeby, *J. Mol. Evol.*, 1984, **20**, 38-51.
208. A. Szilágyi, Á. Kun and E. Szathmáry, *Biology Direct*, 2012, **7**, 38.
209. S. A. Kauffman, *The origins of order*, Oxford University Press, Oxford, 1993.
210. J. D. Farmer, S. A. Kauffman and N. H. Packard, *Physica D*, 1986, **22D**, 50-67.
211. V. Vasas, C. Fernando, M. Santos, S. Kauffman and E. Szathmary, *Biology Direct*, 2012, **7**, 1.

