CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Repository of the Academy's Library

Evolution of the division of labor between genes and enzymes in the RNA world

Gergely Boza[1,5], András Szilágyi[1,2,4], Ádám Kun[1,2,5], Mauro Santos[3] and Eörs Szathmáry[1,2,4,§]

1 Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös Loránd University, Budapest, Hungary. 2 Parmenides Center for the Conceptual Foundations of Science, Pullach, Germany. 3 Departament de Genètica i de Microbiologia, Grup de Biologia Evolutiva, Universitat Autònoma de Barcelona, Barcelona, Spain. 4 MTA-ELTE Research Group in Theoretical Biology and Evolutionary Ecology, Budapest, Hungary 5 MTA-ELTE-MTMT Ecology Research Group, Budapest, Hungary. § Corresponding author

Abstract

The RNA world is a very likely interim stage of the evolution after the first replicators and before the advent of the genetic code and translated proteins. Ribozymes are known to be able to catalyze many reaction types, including cofactor-aided metabolic transformations. In a metabolically complex RNA world early division of labor between genes and enzymes could have evolved, where the ribozymes would have been transcribed from the genes more often than the other way round, benefiting the encapsulating cells through this dosage effect. Here we show, by computer simulations of protocells harboring unlinked RNA replicators that the origin of replicational asymmetry producing more ribozymes from a gene template than gene strands from a ribozyme template is feasible and robust. Enzymatic activities of the two modeled ribozymes are in trade-off with their replication rates, and the relative replication rates compared to those of complementary strands are evolvable traits of the ribozymes. The degree of trade-off is shown to have the strongest effect in favor of the division of labor. Although some asymmetry between gene and enzymatic strands could have evolved even in earlier, surface-bound systems, the shown mechanism in protocells seems inevitable and under strong positive selection. This could have preadapted the genetic system for transcription after the subsequent origin of chromosomes and DNA.

Funding: 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] and the Hungarian National Office for Research and Technology (NAP 2005/KCKHA005). MS was funded by grant CGL2010 –15395 from the Ministerio de Ciencia e Innovación and the ICREA Acadèmia Programme. 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). ÁK gratefully acknowledges a János Bolyai Research Fellowship of the Hungarian Academy of Sciences. GB and ÁK acknowledge support from the Hungarian Research Grants (OTKA K100299). This work was carried out as part of EU COST action CM1304 "Emergence and Evolution of Complex Chemical Systems".

E-mail: szathmary.eors@gmail.com

Introduction

The RNA world is "almost a logical necessity", for example by the fact that aminoacyl-tRNA synthetases are not among the most ancient proteins [1]. Despite eminent attempts [2,3] we still lack a generalized RNA replicase that would be able to unzip and copy general, long RNA templates, similar to the contemporary activity of, say, the Qβ replicase [4], made of protein. A way out could be the assembly, out of replicable shorter pieces, of a replicase and an associated ligase [5], encouraged by the recent finding of a collectively autocatalytic ligase-based RNA network [6]. Twenty years ago the possibility of an early evolution of a division of labor between gene (+) and enzymatic (−) RNA strands was raised: "The fate of both the plus (+) and minus (−) strands is important for the following discussion. If both strands are to be replicated, both of them must be recognized by the replicase: the 3' and 5' ends of the same strand must therefore be complementary (it is assumed that replication goes in the 5' → 3' direction as today). Interestingly, violation of such a complete symmetry opens up the possibility for a very early origin of "transcription" in the form of replication bias. If the plus strand is the gene, and the minus strand is the ribozyme, naturally it pays to make more enzymes than genes. If the tag of the minus ribozyme acts as a weaker target (owing to some point mutations,

Author Summary

The RNA world refers to the stage of early evolution when RNA macromolecules were responsible both for storing hereditary information and performing enzymatic activities. Conflict arises between these two functions, however, as enzymatic activities of the ribozymes are in tradeoff with their replication rates. Here we address this problem by investigating the evolutionary emergence of a primordial transcription-like system in model protocells inhabited by unlinked replicators. Our numerical analysis demonstrates that division of labor between genes and enzymes could have emerged, given that there was a moderate to strong tradeoff between the enzymatic and template efficiency of one strand of the ribozymes. This division of labor results in a strong asymmetry in the numbers of the enzymatic and genetic strands of the macromolecules, in favor of the former. We offer insight into the emergence of the first transcription-like system, which is today characteristic of all known life forms.

for example) for the replicase, this shift in "emphasis" is guaranteed" ([7], p. 448). The authors noted that there is such asymmetry in contemporary RNA viruses [8].

Besides their target affinity, the complementary strands of RNA molecules also have to be different regarding enzymatic activities. It is not inconceivable that complementary strands of RNAs can act as enzymes: Sergei Rodin has convincingly argued that this could have been the case for at least some tRNA [9] and aminoacyl-tRNA synthetase [10] species. It is thus biologically plausible to assume a system where both RNA strands would be weakly enzymatic, but in general this would imply different functions (unless the two strands are palindromic or the contexts in which the strands must act are highly comparable, as in the Rodin case). To conclude, a truly symmetric initial condition in enzymatic activities cannot be very common. Having said that, it is probable that one strand would lose the weak enzymatic function, whereas the complementary strand would be optimized for its enzymatic activity.

As it is likely that some surface-bound metabolic complexity preceded the advent of protocells (e.g. Ref. [11]), earliest ribozymes may also have acted on surfaces [12,13], including evolving replicases [14]. It is in the context of such a surface-bound replicase population that the evolution of strand asymmetry has been dynamically investigated by the technique of cellular automata [15]: the authors have shown that strand asymmetry evolves (assuming a strand-displacement replication mechanism), but depending on diffusion and decay rates in a complex manner; sometimes genes rather than enzymes dominated the population. No model in the context of metabolically active ribozymes [13,16] is known.

Results and discussion

Here we address the problem of RNA strand asymmetry in the context of metabolically active ribozymes encapsulated in reproducing protocells, relying on the stochastic corrector model [12,17] for the basic dynamics. There are two different ribozymes ($T=2$) that are assumed to be essential for protocell growth and reproduction (Figure 1). In contrast to previous treatments plus ($+$) and minus ($-$) strands are explicitly considered. For simplicity we assume that only minus strands are enzymatically active. All templates grow stochastically within each protocell, and protocells also grow and divide stochastically. There is selection at two levels: faster replicating templates within protocells have an advantage, but protocells with a balanced and adequately abundant ribozyme

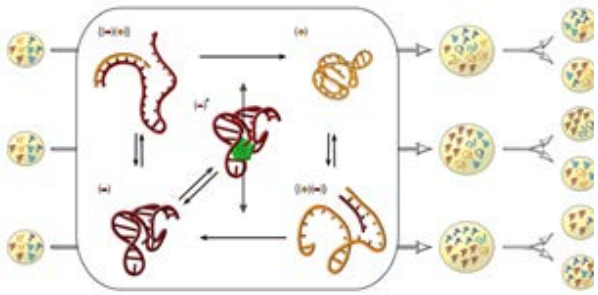

Figure 1. Schematic representation of the main reactions and components of vesicles with complementary replicating strands. Vesicles are composed of two types of macromolecules (type 1 as red, and type 2 as blue), and with two strand types (plus ($+$) strands with light, and minus ($-$) strands with dark shading). The minus ($-$) strands (molecules colored dark red) serve both as enzymes (enzymatic activity indicated with asterisk) for producing monomers (molecule colored green) from source material, and as templates for producing plus ($+$) strands (molecules colored orange). The monomers are used as the building blocks (grey arrow) for the productions of replicators (replication complexes are indicated in curly brackets). The plus strand only serves as template for producing minus strands. For molecule type 2, the metabolic and replication processes are similar to those of molecule type 1 described above, except that the minus ($-$) strand catalyzes a different chemical reaction.

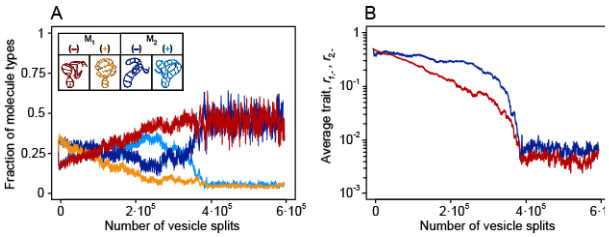

Figure 2. The evolution of division of labor between minus (−) and plus (+) strands. (A) A representative example of simulations resulting in asymmetric strand separation averaged over the population of V vesicles ($M_{1,-}$: red; $M_{1,+}$: orange; $M_{2,-}$: dark blue; $M_{2,+}$: light blue). Starting from an initially symmetric state, i.e. all strand types are represented in equal numbers ($M_{T,+/-,J}=N/2T$), and of equal replication rates ($r_{T,+/-,J}=0.5$) (J denotes the mutation class with trait $r_{T,+/-,J}=0.5$). The trade-off in this case is assumed to be strong between the replication affinity and the catalytic activity. Hence the trait $r_{T,-,i}$ of the minus strand (B) gradually evolves towards lower replication rates ($r_{T,-,i}\to 0$) in order to achieve higher metabolic activity ($m_{T,-,i}\to 1$). During trait evolution the ratio of minus (dark shadings) and plus (light shadings) strands changes, and the minuses significantly increase in numbers. At stable equilibrium, for the very extreme cases, only 4-8% of the macromolecules, on average 2 or 3 per vesicle, are plus strands. Other parameters: $V=1000$, $N=100$, $s=10$, $Z=10$, $n=1000$, $r_{T,+,i}=0.5$, $\mu=0.03$, $\lambda=3$, $\delta=10^{-5}$, $k=0.2$, $\hat{R}=10$, $K_1=10$ and $K_2=1$.

composition are favored [17]. Although we assume their existence, we do not explicitly model replicase molecules, except that a limited number of templates can be replicated at the same time. Their effect is assumed to allow for copying of plus strand from minus strands and *vice versa*, including neat strand separation (which is still an unsolved problem in the origin of life studies [18]). It is assumed that minus strands being copied cannot perform enzymatic function at the same time, due to the opening of the catalytic sites. The two ribozymes are assumed to contribute to the production of the nucleotide monomers of the RNAs. One of the ribozymes (type 1) transforms a source material R available in the environment to intermediate L_1 , which in turn is transformed by the other ribozyme (type 2) to the monomer L_2 . The monomer L_2 is then consumed to build up the four different kinds of strands present in the vesicle. Concrete examples of similar ribozymes that could have helped sustain the RNA world have been successfully selected *in vitro* [19], including nucleoside synthesis, phosphorylation of nucleosides, activation of nucleotides, and processive RNA primer extension. The rates of these reactions are determined by the catalytic activities of the ribozymes. The enzymatic activities of the ribozymes are in trade-off with their replication rates (e.g., active ribozymes are more difficult to unfold due to a denser structure and substrate binding), and the *relative* replication rates compared to those of complementary strands are evolvable traits of the ribozymes. Both higher and lower relative replication rates of the minus strands are allowed to evolve. The traits can change at each replication due to mutations. When the within-vesicle concentration of RNAs reaches a critical level the vesicle splits into two and its content is divided randomly, without replacement, between the two resultant daughter vesicles. See Models and methods for details and Table 1 for parameters and their values used throughout this study.

Average copy number of plus strands can be reduced through evolution even to 1 or 2 gene strands per protocell in cases when trade-off is strong between replication and enzymatic rates. The survival of plus strands in such cases is ensured by the fact that they can be copied from the ribozymes. Figure 2 shows an example of such a successful division of labor between enzymes and genes. In Figure 3 we demonstrate that evolutionary trajectories converge to the same equilibrium ratio of division of labor from different initial states, even when the evolution of replication rates and enzymatic rates is not bound, but only limited by the trade-off function assumed, and when the replication affinity of the plus strand is also allowed to evolve (Figure 3). Less pronounced division of labor is observed for weaker trade-off between replication rate and metabolic efficiency (Figure 4A), for higher numbers of molecules per protocell (Figure 4B), and for higher food concentration and kinetic rate constants (Figure 4C). By far the strongest effect is that of the trade-off, which is understandable, since it is a trait that affects every ribozyme individually. The mild decrease with protocell size is due to the fact that if there are many RNA molecules in total, there are likely to be many enzymes present anyhow, thus the force of selection should decline with protocell size. Similarly, higher food concentrations and higher kinetic rate constants reduce the force of selection for very high enzymatic efficiency. We note that some division of labor evolves even with negligible trade-off: this we attribute to the

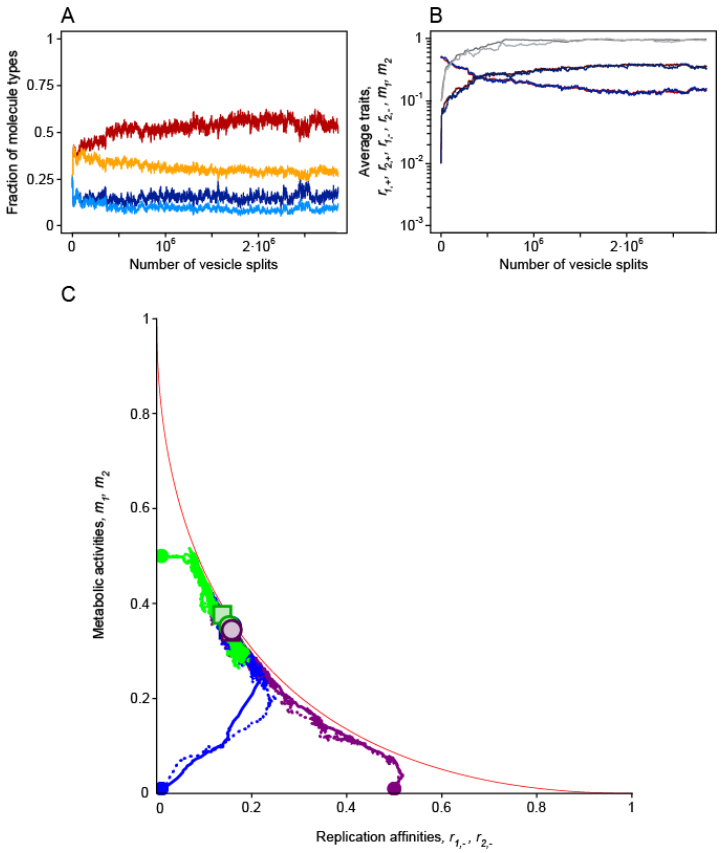

Figure 3. The evolution of division of labor when both replication affinity and metabolic activity of replicators are allowed to evolve separately. (A) A representative example of simulations resulting in asymmetric strand template reaction averaged over the population of V vesicles ($M_{1,-}$: red; $M_{1,+}$: orange; $M_{2,-}$: dark blue; $M_{2,+}$: light blue). Simulations begin from an initially symmetric state, i.e. all strand types are represented in equal numbers ($M_{T,+/-,J}=N/2T$) and equal template replication rates ($r_{T,+/-}=0.5$). We assume low initial metabolic activity of the minus strands ($m_{T,-}=0.01$) and a trade-off between the maximum values of the replication affinity and the catalytic activity of the replicators (see red line in **C**), i.e. no replicator can evolve traits above this boundary, but any rate combination below the curve is accessible (i.e. $r_{T,-,j}^{k}+m_{T,-,j}^{k}\le 1$, see Models Eq. 1b). **(B)** As metabolic activity gradually evolves towards high values (brown and dark blue lines, $m_{T,-,i}\to 1$) the minus strands trade in replication affinity (red and blue lines, $r_{T,-,i}\to 0$) in order to reach the optimum. When the replication affinity of the plus strand can also evolve, evolution further optimizes the protocell composition in favor of strand asymmetry by evolving the highest possible affinity for the plus strand (grey and dark grey lines, $r_{T,+,j}\to 1$). Here $r_{T,+,:}$ is allowed to evolve without any trade-off ($r_{T,+,:}\in[0,1]$, and the initial condition is $r_{T,+,:}=0.1$). **(C)** Trajectories from different initial conditions (green: $r_{T,-}=0.01$ and $m_{T,-}=0.5$; purple: $r_{T,-}=0.5$ and $m_{T,-}=0.01$; and blue: $r_{T,-}=0.01$ and $m_{T,-}=0.01$) converge to the same equilibrium. Solid and dotted lines depict molecule types 1 and 2, respectively. Filled circles represent the initial data points, while light shaded circles and rectangles represent the evolutionary endpoints for traits of molecules 1 and 2, respectively. For the above results we employed a continuous-trait model, in which traits were allowed to change continuously between 0 and 1, and mutant traits were drawn from a normal distribution with the resident trait as a mean and with variance σ. Other parameters: $V=1000$, $N=100$, $s=10$, $Z=10$, $r_{T,+,i}=0.5$, $\mu=0.025$, $\sigma=0.025$, $\delta=10^{-5}$, $k=0.5$, $\hat{R}=10$, $K_1=10$ and $K_2=1$.

metabolic cost of the templates. In short, for the same total template copy number, protocells harboring more enzymes than genes are better off than those with reversed proportions, since the former carry a smaller load of "useless" templates (redundant genes). This effect becomes more pronounced with low food concentrations and kinetic rate constants, as in these cases the selective advantage of protocells with more enzymes increases (Figure 4C). Of course, assortment load (i.e., the drop in average fitness due to the random loss of any essential gene after stochastic assortment of templates in the two daughter protocells), and the fact that high enzymatic efficiency can already be reached without evolving high rate of strand asymmetry, prevents the system from evolving stronger asymmetry without strong trade-off.

High degradation rates can narrow the potential for the evolution of pronounced division of labor.

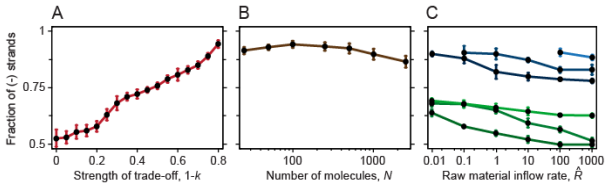

Figure 4. Factors affecting the rate of asymmetry between the minus and the plus strands. (A) In cases when the strength of trade-off is high ($1-k \approx 0.8$), the asymmetry between the minus and plus strands is strong, however as the strength of trade-off decreases ($1-k \to 0$), since in these cases molecules can achieve high metabolic activity without trading off their replication affinities, the asymmetry becomes less pronounced. **(B)** As the number of the initial number of molecules (N) per vesicle is increased ($N = 25, 50, 100, 250, 500, 1000, 2500$) the rate of asymmetry gradually decreases ($1-k = 0.8$). **(C)** The effect of kinetic parameters for strong trade-off (blue lines: $1-k = 0.8$) and for weak trade-off (green lines: $1-k = 0$). Here we increased the inflow rate of source material from the environment into the vesicle ($\hat{R}$) (light blue and green lines: $K_1 = 1$ and $K_2 = 1$; middle dark blue and green lines: $K_1 = 10^2$ and $K_2 = 10^2$; dark blue and green lines: $K_1 = 10^4$ and $K_2 = 10^4$). For low inflow rate and kinetic constants, high metabolic activities of minus strands evolve, which results in high rate of asymmetries between the two strands. However lowering the inflow rate or the kinetic rate of reactions beyond a threshold results in the extinction of replicators (notice the absence of equilibrium ratio of asymmetry, for example $R = 1$, $K_1 = 1$ and $K_2 = 1$, i.e. left hand side of the light blue curve). The results are averaged over 5 replicate model runs, and over 1.000.000 molecular update steps after reaching equilibrium. Whiskered bars represent the standard errors of the replicate runs. Other parameters (if not stated otherwise): $V = 1000$, $N = 100$, $s = 10$, $Z = 10$, $n = 1000$, $r_{T,+,i} = 0.5$, $\mu = 0.03$, $\lambda = 3$, $\delta = 10^{-5}$, $k = 0.2$, $\hat{R} = 10$, $K_1 = 10$ and $K_2 = 1$.

Extreme trade-off between replication and metabolic activity selects for only few gene strands per protocell, hence a higher degradation rate easily eliminates them, and the few new genes synthesized from the ribozymes as templates may well suffer a similar fate: in the end the ribozymes cannot increase in number either, so all in all higher degradation rates lead to weaker admissible trade-off and result in weaker strand asymmetry (Figure 5). Larger protocells could, however, survive at higher degradation rates, potentially allowing for strand differentiation at strong trade-offs (the lower right part of Figure 5, where populations do not survive at the parameter values employed).

Division of labor between genes and enzymes can only be partial in our model, as expected in an RNA-based system, since a complete replication cycle requires that both strands act as templates to some extent. Division of labor implies that entities

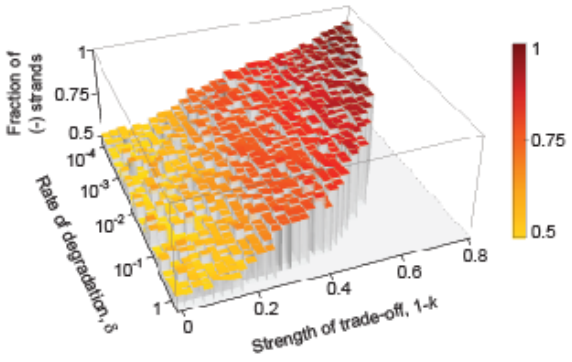

Figure 5. The effect of degradation rate of macromolecules on strand asymmetry. The equilibrium ratio of the minus and plus strands (indicated by the heights as well as the colors of the bars; red: 0.9 → yellow: 0.5) is not affected significantly by the rate of degradation, however increasing the degradation rate above a threshold results in the extinction of the replicators (notice the flat grey area on the right hand side of the graph). For strong trade-off ($1-k \approx 0.8$), this threshold is at a lower rate of degradation, whereas higher degradation rates are tolerated as the strength of trade-off decreases ($1-k \to 0$).The results are averaged over 3 replicate model runs. Other parameters: $V = 250$, $N = 100$, $s = 10$, $Z = 10$, $n = 1000$, $r_{T,+,i} = 0.5$, $\mu = 0.03$, $\lambda = 3$, $\hat{R} = 10$, $K_1 = 10$ and $K_2 = 1$.

required to perform two different tasks end up with one doing (mostly) one of the tasks while the other do the other task. In our case this means that one strand acts mostly as an enzyme, while the other acts mostly as an information carrier. As only one of the strands in our model has enzymatic activity, that strand can be called an enzyme. Both of the strands need to act as templates, otherwise information is lost, but as one of the strands mainly acts as template and does not have enzymatic activity, we can call this a gene.

We would like to note here, that division of labor does not require sharp, and full specialization in different tasks: intermediate degree of specialization with the interchangeability of task-performing entities suffices too [20]. The important point is that the gain in performance due to specialization in one task exceeds the cost of loss of performance due to specialization in another task [21]. Let us give two unrelated examples from biology to illustrate our point. In eusocial insects, such as bees, division of labor often implies high degree of specialization in different tasks, as for example the queen and the workers are quite different in morphology and in behavior. But among

the workers there are behavioral, but no morphological, castes, groups that tend to perform different task (cleaning, foraging, tending the young, etc.) [22]. Moreover, in primitively eusocial wasps, like the *Ropalidia marginata*, even the queen cannot be distinguished from others except for the presence of well-developed ovaries [4]. The other example comes from clonal plants (such as strawberry), where the members are connected physiologically. In these plants, one "plant member" (called the ramet) may specialize in the uptake of belowground nutrients, and thus develop an extensive root system, while the other specializes in the capture of light and develops bigger leaves [5]. Only the relative investments change into shoot or root, but ramets still have both functioning root systems and leaves. In biology, it is thus common to observe intermediate levels of division of labor and functional specialization within the boundaries set by physiological and developmental constraints of an organism.

Chemical difference between the enzymes and the templates is not a requirement for division of labor between genetic and enzymatic functions. The present neat chemical distinction between genes (DNA) and enzymes (proteins) is a rather late invention. Comparative analysis of the genes involved in DNA replication [1,2] and the age of protein domain fold required for dNTP synthesis [3] suggest that the emergence of DNA genome was a late phenomenon which could have happened after the LUCA, thus was most likely a successor to the RNA world. As the authors of a somewhat related theoretical work note: "DNA releases RNA from the trade-off between template and catalyst that is inevitable in the RNA world and thereby enhances the system's resistance against parasitic templates" [23] (p. 2). It is exactly this trade-off that drives the evolution of the division of labor in our protocellular system. (We note in passing that the analysis in Ref. [23] is not enough by itself to explain the advantage of DNA, since DNA molecules can also be selected to act as enzymes [24]).

We investigated the evolution of division labor between enzymatic and genetic strands based on the implicit assumption that minus and plus strands can have very different secondary structures. This indeed proves to be the case: on a sample of 10 million sequences, the distances between the secondary structures of minus and plus strands are slightly higher than those between pairs of randomly generated sequences (Figure 6A). Furthermore, there is asymmetry in the complexity of secondary structures (Figure 6A, C, D); and the difference between the free energies of folding can reach levels up to 20 kcal/mol (Figure 6B). Thus, there is a fraction of complementary, folded strand pairs for which one member is more readily opened by a replicase than the other, due to the looser structure of the former (Figure 6C). Here we have only considered the minimum free energy (MFE) structures of the RNAs. It is known that there are suboptimal structures that could be quite close energetically to the MFE structure [25], and thus provide additional ways in which the two strands can be different (albeit evolution can lead to well-defined structures with little ambiguity in their energetically close sub-optimal structures [26]). Co-folding of the RNA with smaller RNAs can further increase the structural diversity of RNAs [27], again possibly promoting functional diversification of the strands. Our conservative estimate of structural difference is sufficient for strand separation, and incorporation of further mechanisms can further foster the effect demonstrated above.

The origin of basic genetic operations, including replication and transcription, belongs to the key questions of the origin of life. While there has been considerable progress with template copying [2,3], unzipping remains an open problem [18] (but see [28]). In this paper we have shown that once evolution had reached the stage of reproducing compartments with unlinked ribozymes inside, division of labor between enzymatic and gene strands readily followed provided there was moderate to strong tradeoff between the enzymatic and template efficiency of ribozymes. This is to be expected due to the tightly folded structure of ribozymes (for example the Qβ replicase replicates the X-motif ribozyme [29] very slowly compared to other, less complex secondary structures; A. Griffiths, personal communication). Furthermore, analysis of the minimum free energy structures of real ribozymes and aptamers indicates that there is a tendency of them being more thermodynamically stable than random sequences (81.9% are more stable than half of the random sequences; 59.6% are more stable than 75% of the random sequences; and 27,5% are more stable than 95% of the random sequences). This transcription-like process could have been augmented by the evolution of tags recognized by the replicase as envisaged by Szathmáry and Maynard Smith [7], although we have not included this component in

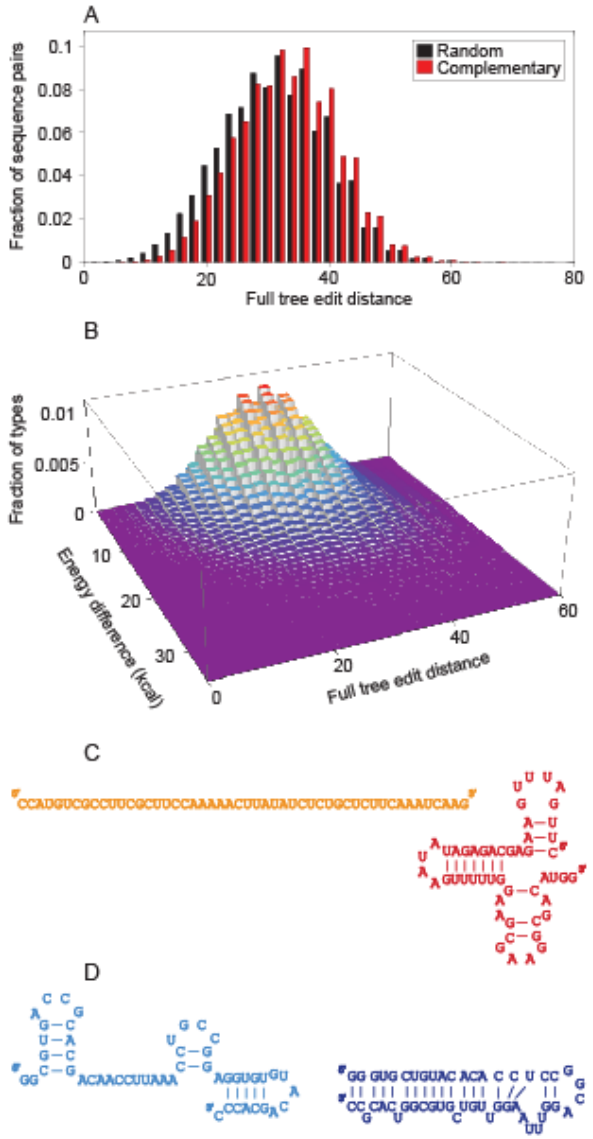

Figure 6. Characteristics of secondary structures of complementary strands. The characteristics of minimum free energy secondary structures are measured on a sample of 10^7 randomly generated sequences of length 50. In case of complementary strands, the complementary sequences of the randomly generated strands are also analyzed. **(A)** Complementary strands have higher full tree edit distance between them (red bars) than random sequence pairs (black bars). **(B)** Energy difference between members of pairs of complementary, folded strands. Around tree edit distance 30 most complementary, folded structures have negligible energy difference, but a decreasing proportion of pairs show a difference of up to 40 kcal. **(C)** Example of a complementary pair of strands in which one of the strands does not have a structure, while the other has a rich structure. The difference of their minimum free energies is (6.6 kcal). **(D)** Example of a complementary pair of strands in which the two strands have very different (tree edit distance 68) but still rich structures. The difference of their minimum free energies is (7.0 kcal).

the present model. We conclude that division of labor between genes and enzymes was under strong positive selection in the RNA world.

Models and methods

Characteristics of the RNA molecules

An RNA molecule M_i of length s ($s=10$) is characterized by its type T ($T=2$), its role of being a ribozyme ($-$) or an informational strand ($+$), and a combined trait r representing the replication affinity and the polymerization rate of the given molecule. We assume $r_{T,+,\cdot}=0.5$ for the ($+$) strands (except for Figure 3, in which case $r_{T,+,\cdot}\in[0,1]$). The traits $r_{1,-,\cdot}$ and $r_{2,-,\cdot}$ ($r_{T,-,\cdot}\in[0,1]$) are the evolvable traits of our model.

The ribozymes catalyze reactions with metabolic activity $m_{T,-,i}$. We assume that the ribozymes cannot perform any metabolic function during the replication process, as the molecule is in an unfolded state and cannot form the pocket responsible for enzymatic activity. Thus there is a trade-off between the processes of replication and catalytic activity which is characterized by the following one-parameter function

$$r_{T,-,i}^k + m_{T,-,i}^k = 1, \quad (1a)$$

and for additional investigations (see Fig. 3) we also allow

$$r_{T,-,i}^k + m_{T,-,i}^k \le 1, \quad (1b)$$

where k characterizes the strength of this trade-off ($k \le 1$).

Mutation can occur with probability μ at each replication of the molecules. We allow $r_{T,-,\cdot}$ to change in a discrete manner:

$$r_{T,-,\text{mutant}} = r_{T,-,\text{original}} \pm \gamma / n \qquad (2)$$

where n is the number of mutation classes and γ is randomly drawn from Poisson-distribution with parameter λ. There is an equal probability of having mutants with higher or with lower traits compared to the original trait. We opted for discrete traits as it facilitates faster convergence to evolutionary equilibrium, as our additional studies indicated similar result can be attained employing continuous traits (see Fig. 3). In the latter case traits are allowed to change on a continuous scale, and mutant traits are drawn from a normal distribution with the resident trait as a mean and with variance σ.

Chemical reactions in the vesicles

Reactions involving the macromolecules M fall into four classes: (1) catalyzed conversion C_1 of the raw material (R) into the intermediate (L_1); (2) catalyzed conversion C_2 of the intermediate (L_1) into the monomer (L_2); (3) polymerization P of a new strand; and (4) degradation D of a macromolecule. As there are n mutational classes, the trait $r_{T,-,\cdot}$ (and thus $m_{T,-,\cdot}$) can have n different values. Accordingly, the total number of possible reactions are: n conversions C_1, n conversions C_2, Tns polymerization P_+ reactions involving the plus strands as templates, Tns polymerization P_- reactions involving the minus strands as templates, and $2Tn$ reaction of degradation D.

$$R + M_{1,-,i} \xrightarrow{K_1 \cdot m_{1,-,i}} L_1 + M_{1,-,i} \quad (C_{1,i})$$

$$L_1 + M_{2,-,i} \xrightarrow{K_2 \cdot m_{2,-,i}} L_2 + M_{2,-,i} \quad (C_{2,i})$$

$$L_2 + M_{T,+,i} \xrightarrow{r_{T,+,i}} \left\{ M_{T,+,i} M^{(1)}_{T,-,\cdot} \right\} \quad (P_{T,+,1,i})$$

$$L_2 + \left\{ M_{T,+,i} M^{(l)}_{T,-,\cdot} \right\} \xrightarrow{r_{T,+,i}} \left\{ M_{T,+,i} M^{(l+1)}_{T,-,\cdot} \right\},$$

$(l = 1,\ldots,s-2) \quad (P_{T,+,2,i},\ldots,P_{T,+,s-1,i})$

$$L_2 + \left\{ M_{T,+,i} M^{(s-1)}_{T,-,\cdot} \right\} \xrightarrow{r_{T,+,i}} M_{T,+,i} + M_{T,-,j} \quad (P_{T,+,s,i})$$

$$L_2 + M_{T,-,i} \xrightarrow{r_{T,-,i}} \left\{ M_{T,-,i} M^{(1)}_{T,+,\cdot} \right\} \quad (P_{T,-,1,i})$$

$$L_2 + \left\{ M_{T,-,i} M^{(l)}_{T,+,\cdot} \right\} \xrightarrow{r_{T,-,i}} \left\{ M_{T,-,i} M^{(l+1)}_{T,+,\cdot} \right\},$$

$(l = 1,\ldots,s-2) \quad (P_{T,-,2,i},\ldots,P_{T,-,s-1,i})$

$$L_2 + \left\{ M_{T,-,i} M^{(s-1)}_{T,+,\cdot} \right\} \xrightarrow{r_{T,-,i}} M_{T,-,i} + M_{T,+,j} \quad (P_{T,-,s,i})$$

$$M_{T,-,i} \xrightarrow{\delta} \varnothing \quad (D_{T,-,i})$$

$$M_{T,+,i} \xrightarrow{\delta} \varnothing \quad (D_{T,+,i})$$

where K_1, K_2 and δ are kinetic constants for the corresponding reactions, $\left\{ M_{T,+,i} M^{(l)}_{T,-,\cdot} \right\}$ or $\left\{ M_{T,-,i} M^{(l)}_{T,+,\cdot} \right\}$ denotes the complex involving a template strand and the an intermediate forms of the complementary strand consisting of l monomers ($i = 1,\ldots,n$, $1 \le l < s$).

The full replication cycle is completed after two steps of copying: $M_{T,+,i} \xrightarrow{sL_2} M_{T,-,j} \xrightarrow{sL_2} M_{T,+,h}$. We assume a limited number of replicase enzymes in a vesicle, hence we limit the number of simultaneous replication processes to Z.

We apply the Gillespie algorithm [30] to follow the reactions within the vesicle. We introduce the quantity $a_v(t)\mathrm{d}t$ that characterizes the probability of reaction $v \in \left\{ C_{1,i}, C_{2,i}, P_{T,+,1,i} \ldots P_{T,+,s,i}, P_{T,-,1,i} \ldots P_{T,-,s,i}, D_{T,+,i}, D_{T,-,i} \right\}$ ($i = 1,\ldots,n$, $T = 1,2$) in the time interval $(t, t+\mathrm{d}t)$. $a_v(t)$ is the product of two factors: the chemical constant for the given reaction type v and the number of possible reactions within a given vesicle. For the reaction $C_{1,i}$

$$a_{C_{1,i}}(t) = \hat{R} \cdot M_{1,-,i}(t) \cdot K_1 \cdot m_{1,-,i}. \quad (3)$$

We note that the input material R has a fixed concentration $\hat{R}$. Similarly, for reaction types e.g. $P_{T,+,2,i}$ and $P_{T,-,2,i}$

$$a_{P_{+,2,i}}(t) = L_2(t) \cdot \left\{ M_{T,+,i} M^{(2)}_{T,-,\cdot} \right\}(t) \cdot r_{T,+,i} \quad (4)$$

$$a_{P_{+,2,i}}(t) = L_2(t) \cdot \left\{ M_{T,-,i} M^{(2)}_{T,+,\cdot} \right\}(t) \cdot r_{T,-,i} \quad (5)$$

Degradation is a monomolecular reaction; its probability is proportional to the present amount of the given molecules. The chemical constants δ for degradation is common for all types of macromolecules M. The degradation and dissociation of replication complexes is neglected in our model.

We define the sum of all a_v's as

$$a_0(t) = \sum_v a_v(t) \quad (6)$$

The time τ after t at which the next reaction will take place is drawn from an exponential probability density function of rate a_0:

$$p(\tau) = a_0 e^{-a_0 \tau} \quad (7)$$

At time $t + \tau$, we choose reaction v as the next reaction with probability a_v / a_0 in the vesicle. We then update the number of different molecules according to reaction scheme v and the process is reiterated.

Population dynamics of protocells

The population is composed of V number of protocells, with the initial number of N replicating molecules, and $\hat{L}_1$ number of intermediate and $\hat{L}_2$ number of building block molecules. The number of RNAs can increase up to $2N$, at which point the vesicle splits randomly assorting all the replicator molecules into two daughter vesicles. During splitting, small molecules L_1 and L_2, as well as the initiated replication complexes are also randomly allocated to the daughter vesicles. One daughter vesicle is replacing the parent, while the other

replaces another random vesicle in the population (i.e. it is a Moran process [31]).

Structural similarity of complementary strands

We have assumed that complementary strands can be quite dissimilar in structure, so that one of them can fold to be a ribozyme while the other has a structure that can be more readily processed by the replicase enzyme. We check if complementary strands can be dissimilar enough to potentially achieve such a state. We have determined the minimum free energy structure of 10^7 random RNA sequences of length 50. We have also determined the minimum free energy (MFE) structure of 10^7 random complementary pairs of RNAs of length 50. Each individual sequence's structure is compared to the structure of the next sequence to obtain the full tree edit distance [32] between the two structures. Similarly, the distance between each complementary pair of sequences is also determined. All computations are done with the Vienna RNA Package 2.0.7 [33].

Thermodynamic stability of ribozymes and aptamers

We analyzed the set of 305 ribozyme and aptamer sequences mainly from the Aptamer Database [34] and from the review of Chen and co-workers [35] (the full list is reported in Supplementary Table S1 of [36]). For each sequence the MFE was determined. Then we generated 100,000 random sequences of the same length and recorded their MFE. Then we counted the number of random sequences having lower MFE (i.e. being more stable) than the ribozyme/aptamer.

Acknowledgements

The authors would like to thank B. Könnyű and the anonymous reviewers for the valuable comments and suggestions in improving the manuscript.

References

1. Aravind L, Mazumder R, Vasudevan S, Koonin EV (2002) Trends in protein evolution inferred from sequence and structure analysis. Current Opinion in Structural Biology 12: 392-399.
2. Wochner A, Attwater J, Coulson A, Holliger P (2011) Ribozyme-catalyzed transcription of an active ribozyme. Science 332: 209-212.
3. Attwater J, Wochner A, Holliger P (2013) In-ice evolution of RNA polymerase ribozyme activity. Nature Chemistry 5: 1011–1018.
4. Blumenthal T, Carmichael GG (1979) RNA replication: Function and structure of Q*b*-replicase. Annual Review of Biochemistry 48: 525-548.
5. Ellington AD, Chen X, Robertson M, Syrett A (2009) Evolutionary origins and directed evolution of RNA. The International Journal of Biochemistry & Cell Biology 41: 254-265.
6. Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, et al. (2012) Spontaneous network formation among cooperative RNA replicators. Nature 491: 72-77.
7. Szathmáry E, Maynard Smith J (1993) The evolution of the chromosome II. Molecular mechanisms. Journal of Theoretical Biology 164: 447-454.
8. Wintersberger U, Wintersberger E (1987) RNA makes DNA: a speculative view of the evolution of DNA replication mechanisms. Trends in Genetics 3: 198-202.
9. Rodin S, Ohno S, Rodin A (1993) Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? Proceedings of the National Academy of Sciences of the USA 90: 4723-4727.
10. Rodin A, Szathmary E, Rodin S (2009) One ancestor for two codes viewed from the perspective of two complementary modes of tRNA aminoacylation. Biology Direct 4: 4.
11. Wächtershäuser G (1988) Before enzymes and templates: theory of surface metabolism. Microbiological Reviews 52: 452-484.
12. Maynard Smith J, Szathmáry E (1995) The Major Transition in Evolution. Oxford, UK: W.H. Freeman.
13. Könnyű B, Czárán T (2011) The evolution of enzyme specificity in the metabolic replicator model of prebiotic evolution. PLoS ONE 6: e20931.
14. Szabó P, Scheuring I, Czárán T, Szathmáry E (2002) *In silico* simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. Nature 420: 340-343.
15. Takeuchi N, Salazar L, Poole A, Hogeweg P (2008) The evolution of strand preference in simulated RNA replicators with strand displacement: Implications for the origin of transcription. Biology Direct 3: 1-22.
16. Czárán T, Szathmáry E (2000) Coexistence of replicators in prebiotic evolution. In: Dieckmann U, Law R, Metz JAJ, editors. The Geometry of Ecological Interactions. Cambridge: Cambridge University Press. pp. 116-134.
17. Maynard Smith J, Szathmáry E (1993) The origin of the chromosome I. Selection for linkage. Journal of Theoretical Biology 164: 437-446.
18. Kováč L, Nosek J, Tomáška Lu (2003) An overlooked riddle of life's origins: Energy-dependent nucleic acid unzipping. Journal of Molecular Evolution 57: S182-S189.
19. Joyce GF (2002) The antiquity of RNA-based evolution. Nature 418: 214-220.
20. Anderson C, Franks NR (2001) Teams in animal societies. Behavioral Ecology 12: 534-540.
21. Rueffler C, Hermisson J, Wagner GP (2012) Evolution of functional specialization and division of labor. Proceedings of the National Academy of Sciences 109: E326–E335.
22. Wilson EO (1971) The insect societies. Cambridge, USA: Harvard University Press.
23. Takeuchi N, Hogeweg P, Koonin EV (2011) On the origin of DNA genomes: Evolution of the division of labor between template and catalyst in model replicator systems. PLoS Computational Biology 7: e1002024.

24. Joyce GF (2004) Directed evolution of nucleic acid enzymes. Annual Review of Biochemistry 73: 791-836.
25. Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244: 48-52.
26. Wuchty S, Fontana W, Hofacker IL, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49: 145-165.
27. de Boer FK, Hogeweg P (2012) Less can be more: RNA-adapters may enhance coding capacity of replicators. PLoS ONE 7: e29952.
28. Acevedo A, Brodsky L, Andino R (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505: 686-690.
29. Tang J, Breaker RR (2000) Structural diversity of self-cleaving ribozymes. Proceedings of the National Academy of Sciences of the USA 97: 5784–5789.
30. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22: 403-434.
31. Moran PAP (1958) Random processes in genetics. Proc Camb Phil Soc 54: 60-71.
32. Tai K-C (1979) The tree-to-tree correction problem. Journal of the Association for Computing Machinery 26: 422-433.
33. Lorenz R, Bernhart S, Honer zu Siederdissen C, Tafer H, Flamm C, et al. (2011) ViennaRNA Package 2.0. Algorithms for Molecular Biology 6: 26.
34. Lee JF, Hesselberth JR, Meyers LA, Ellington AD (2004) Aptamer Database. Nucleic Acid Research 32: D95-100.
35. Chen X, Li N, Ellington AD (2007) Ribozyme catalysis of metabolism in the RNA World. Chemistry & Biodiversity 4: 633-655.
36. Szilágyi A, Kun Á, Szathmáry E (2014) Local neutral networks help maintain inaccurately replicating ribozymes. submitted.

Table 1. Parameters of the model.

Parameter	Definition and value(s)
$r_{T,-,i}$	replication rate and affinity of the $(-)$ strand, $r_{T,-,i}=0.5$
$r_{T,+,i}$	the initial replication rate and affinity of the $(+)$ strand, $r_{T,+,J}=0.5$
V	number of vesicles in the population, $V=250,1000$
N	initial number of molecules per vesicle, $N=25,50,100,250,500,1000,2500$
T	number of replicator types, $T=2$
n	number of mutant classes, $n=1000$
s	number of monomers per macromolecule, $s=10$
Z	maximal number of replication complexes, $Z=10$
K_1	kinetic parameter of conversions C_1, $K_1=10^0,10^1,10^2,10^4$
K_2	kinetic parameter of conversions C_2, $K_2=10^0,10^1,10^2,10^4$
$\bar{R}$	fixed concentration of the input material R, $\hat{R}=10^{-2},10^{-1},10^0,10^1,10^2,10^3$
δ	degradation rate of macromolecules, $\delta=\left[10^{-5},2\right]$
μ	mutation rate, $\mu=0.03$ and $\mu=0.025$
λ	mutational variability, parameter of the Poisson distribution, $\lambda=3$
σ	mutational variability, parameter of the normal distribution, $\sigma=0.025$
k	strength of trade-off, $k=\left[0.2,1\right]$