RNA REPLICATION AND THE ORIGINS OF LIFE

GERALD F. JOYCE
The Salk Institute, La Jolla, CA 92037 USA

RNA-based evolving systems

All known organisms rely on DNA as the genetic material and proteins as the chief agent of function, but the machinery needed to copy DNA and express proteins is far too complex to have arisen spontaneously. In the late 1960s, as the principles of molecular biology came into focus, it was first suggested that the earliest form of life instead relied on RNA as both the genetic material and the agent of function [1–3]. Special attention has been directed to what Francis Crick called the "first enzyme" of life: an RNA molecule that catalyzes the replication of RNA and thus is both gene and enzyme [2]. Such a molecule could provide the basis for a living, evolving system.

There are several known examples of RNA enzymes in biology, but none that have the ability to copy RNA. A larger number of RNA enzymes have been developed in the laboratory using directed molecular evolution, including those that can copy an RNA template by joining together the nucleotide building blocks of RNA (A, U, G, and C) [4–6]. Like natural evolution, directed evolution relies on processes of amplification, mutation, and selection to enrich a population with individuals that are most fit, but in directed evolution the experimenter defines the fitness criteria.

RNA viruses also undergo processes of Darwinian evolution, resulting in the emergence of novel variants with increased fitness. Copying of the viral RNA is dependent on protein enzymes that are encoded within the viral genome, but those proteins must be synthesized by the machinery of the host cell. Thus an RNA virus cannot be regarded as a living system in its own right. The fitness of a virus is ultimately determined by the functional copy number of its genome over time, but that fitness takes into account the material properties of the viral genome and the virally-encoded proteins.

The directed evolution of RNA serves as a model of both RNA-based life and viral evolving systems. In fact, the first directed evolution experiment, carried out by Sol Spiegelman and co-workers in 1967 [7], involved the viral genomic RNA of Qß bacteriophage, which was replicated in the test tube using Qß replicase protein. A portion of this multi-subunit protein is encoded within the viral genome, with the remainder supplied by the host cell. The original Spiegelman experiment, and others that followed, demonstrated the evolution of variants of the Qß genome with increased fitness. Fitness was no longer coupled to viral infectivity, but simply a reflection of the increased copy number of the virus-derived RNA under the chosen set of experimental conditions.

Modern directed RNA evolution experiments seek to drive RNA to perform novel functions, sometimes with a practical application in mind, but also to explore the catalytic potential of RNA. RNA viruses have the benefit of encoding proteins with broad functionality, including the critical function of replicating the RNA genome. The functional superiority of proteins over RNA likely explains why there is no known example of a viral RNA that catalyzes its own replication. Yet if we are to address the question of how the first living systems arose, before the advent of instructed protein synthesis, then it is important to seek RNA molecules that can function as an RNA-dependent RNA polymerase, with the ability to catalyze the replication of RNA.

Toward an RNA enzyme with RNA replicase activity

Early attempts to develop an RNA-dependent RNA polymerase focused on stringing together a few letters of RNA by adding activated nucleotides (NTPs) to the end of a template-bound RNA primer [4,8,9]. Further improvements enabled several dozen nucleotides to be copied, but only for unstructured, repetitive templates [5,10]. Our laboratory entered the fray by evolving polymerases that can copy "difficult" templates to yield a functional RNA product. Selection of the polymerase was made dependent on the function of the synthesized product, requiring the synthesis of progressively more complex products. Those efforts resulted in the evolution of RNA polymerases that are faster, more accurate, and more general in copying RNA [6,11,12].

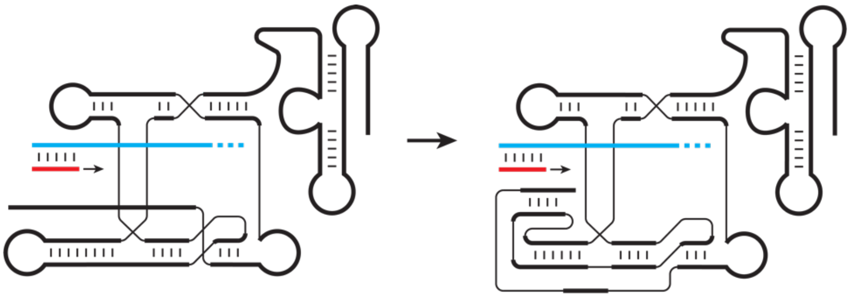

Fig. 1. Secondary structure of the RNA polymerase (black), which extends an RNA primer (red) on an RNA template (cyan). Over the course of directed evolution, the RNA enzyme underwent a tertiary structural rearrangement whereby an existing stem element became shortened while a new stem element was formed, together creating a pseudoknot structure.

During the directed evolution process, the RNA enzyme underwent a dramatic structural rearrangement of its catalytic core [12]. Through the accumulation of 15 mutations within the core, an existing stem element became shortened while a new stem element was formed, together creating a pseudoknot structure that lies in close proximity to the enzyme's active site (Fig. 1). Three important attributes emerged together with this structural rearrangement. First, the catalytic rate improved by ~4,000-fold compared to the starting enzyme. Second, the polymerase gained the ability to copy templates of almost any sequence, including those with structure. Third, the polymerase gained the ability to bind the template-primer complex through high-affinity tertiary interactions, comparable to those seen with modern polymerase proteins.

This advanced form of the RNA polymerase can copy more than 100 nucleotides in 10 minutes and can operate with an accuracy of 92–94% per nucleotide [12]. However, the polymerase itself contains 184 nucleotides and is especially difficult to copy. Furthermore, RNA replication requires copying both the template and its complement, which doubles the challenge. Thus, further improvement of the catalytic rate, copying accuracy, and sequence generality of the polymerase will be needed to recreate the first enzyme of life.

The accuracy of polymerization is critical for an RNA replicase to be able to support the self-sustained evolution of RNA. If the error rate is too high, then the copies will be riddled with mutations, exceeding the ability of selection to cull deleterious mutations [13]. Deep sequencing was used to assess the position-specific frequency of mutations for both partial- and full-length extension products on templates ranging from the most favorable to the most difficult. This analysis revealed that when the polymerase is pushed to the limits of its activity, the accuracy of synthesis declines [6,12]. For a short, unstructured template of 11 nucleotides, the average fidelity is 97% per nucleotide, with the majority of mutations due to G•U wobble pairing. Excluding wobble mutations, the fidelity is >99%. For a longer, more structured template of 33 nucleotides, the average fidelity drops to 92% overall and 96% excluding wobble mutations. For an even longer and highly structured template of 77 nucleotides, pushing the limit of polymerase activity, the average fidelity is 84% overall and 88% excluding wobbles.

Examination of the partial-length products revealed that fidelity is lowest for the last added nucleotide and increases monotonically for positions further upstream from the last nucleotide. The longer the polymerization reaction is allowed to continue, the greater the overall yield, but also the lower the fidelity of the full-length products [6]. Taken together, these facts indicate that the polymerase stalls after adding a mismatched nucleotide, but over time can extend past the mismatch to incorporate the mutation within full-length products. It will not be sufficient to evolve a faster polymerase unless the polymerase also evolves either a lower frequency of mismatched NTP addition or a reduced propensity to extend mismatched termini.

We are continuing the directed evolution process to develop ever more capable forms of the polymerase, focusing especially on improving the fidelity of template copying. The polymerase now has sufficiently high activity that it can synthesize its own evolutionary ancestor, an RNA-joining enzyme that contains 97 nucleotides. By challenging the evolving population of polymerases to synthesize a functional copy of its ancestor, we are placing unprecedented selection pressure on improving polymerase fidelity because about half of the nucleotides within the synthesized product cannot be mutated without loss of activity [14]. Furthermore, the RNA being synthesized has the same catalytic domain as the polymerase itself, thus training the polymerase to synthesize an RNA of similar composition.

As a result of the most recent rounds of directed evolution, the fidelity of polymerization has improved from 84% to 89% for synthesis of the 97-nucleotide functional product. This is the first time that the ability to synthesize longer products has been accompanied by improved fidelity. We appear to have entered the long-anticipated virtuous cycle, where the ability to synthesize longer products enables us to impose selective pressure to drive further improvement of fidelity due to the greater number of immutable nucleotides within those longer products. In turn, every improvement in fidelity enables the synthesis of ever longer functional products.

The threshold of heritable information

The propagation of heritable information requires both efficient and accurate copying of that information. It has long been recognized that there is an "error threshold" based on the relative advantage of a selectively advantageous individual compared to the population as a whole, taking into account the probability of producing error-free copies [13]. For the copying of RNA genomes, there is an inverse relationship between the per-nucleotide fidelity of polymerization and the maximum length of RNA that can be maintained through successive rounds of replication. For the RNA polymerase we have been studying, an average fidelity of >98% will be needed to achieve self-sustained Darwinian evolution. The greater the efficiency and fidelity of the polymerase, the more readily it can be evolved toward further improvements in efficiency and fidelity because one can then impose greater selection pressure to drive those improvements. This bootstrapping process is analogous to what is thought to have driven the evolution of more complex genomes during the early history of life on Earth [6,15].

The genomes of RNA viruses typically contain 10^3–10^4 nucleotides, and the error rate of the corresponding viral RNA polymerase proteins that copy those genomes are in the range of 10^{-3}–10^{-4} [16]. Some RNA viruses, such as HIV-1 and poliovirus, operate very close to the error threshold, which facilitates their rapid evolutionary adaptation, but also places them close to overstepping the error threshold and no longer able to maintain heritable information.

Considerable effort has been devoted toward pushing RNA viruses over the error threshold by exposing them to mutagens, an approach that has been termed "lethal mutagenesis" [17]. This effect must be distinguished from the way in which a mutagen can reduce copying efficiency. Lethal mutagenesis is the result of a cascade of copying errors that cannot be balanced by selection [18]. For example, the purine analogue ribavirin, in addition to inhibiting viral replication, exerts an antiviral effect through enhanced mutagenesis [19]. A more contemporary example is the cytidine analogue molnupiravir, which has been approved for the treatment of patients with symptomatic SARS-CoV-2 infection [20]. This compound promotes G-to-A mutations and is resistant to the proofreading exonuclease encoded by the virus [21].

Both self-replicating RNA enzymes and RNA viruses lie close to the edge of life. Both are able to maintain heritable genetic information and undergo Darwinian evolution. However, both lie precariously close to the error threshold, beyond which it is no longer possible to maintain that genomic information. In addition to being interesting in their own right, these systems serve as simplified models to study the fundamental processes of Darwinian evolution. These processes provide the basis for all known life, from the time of its origins and throughout its natural history.

Acknowledgments

This work was supported by grant NSSC22K0973 from NASA and grant 287624 from the Simons Foundation.

References

1. C. Woese, *The Genetic Code*, Harper & Row, New York, pp. 179–195 (1967).
2. F. H. C. Crick, *J. Mol. Biol.* **38**, 367 (1968).
3. L. E. Orgel, *J. Mol. Biol.* **38**, 381 (1968).
4. W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner, D. P. Bartel, *Science* **292**, 1319 (2001).
5. A. Wochner, J. Attwater, A. Coulson, P. Holliger, *Science* **332**, 209 (2011).
6. K. F. Tjhung, M. N. Shokhirev, D. P. Horning, G. F. Joyce, *Proc. Natl. Acad. Sci. USA* **117**, 2906 (2020).
7. D. R. Mills, R. L. Peterson, S. Spiegelman, *Proc. Natl. Acad. Sci. USA* **58**, 217 (1967).
8. K. E. McGinness, G. F. Joyce, *Chem. Biol.* **9**, 585 (2002).
9. H. S. Zaher, P. J. Unrau, *RNA* **13**, 1017 (2007).
10. J. Attwater, A. Wochner, P. Holliger, *Nat. Chem.* **5**, 1011 (2013).
11. D. P. Horning, G. F. Joyce, *Proc. Natl. Acad. Sci. USA* **113**, 9786 (2016)
12. X. Portillo, Y.-T. Huang, R. R. Breaker, D. P. Horning, G. F. Joyce, *eLife* **10**, e71557 (2021).
13. M. Eigen, *Naturwiss.* **58**, 465 (1971).
14. E. H. Ekland, D. P. Bartel, *Nucleic Acids Res.* **23**, 3231 (1995)
15. G. F. Joyce, J. W. Szostak, *Cold Spring Harb. Perspect. Biol.* **10**, a034801 (2018).
16. E. C. Holmes, *The Evolution of RNA Viruses*, Oxford University Press, New York (2009).
17. L. A. Loeb, J. M. Essigmann, F. Kazazi, J. Zhang, K. D. Rose, J. I. Mullins, *Proc. Natl. Acad. Sci. USA* **96**, 1492 (1999).
18. J. Summers, S. Litwin, *J. Virol.* **80**, 20 (2006).
19. S. Crotty, C. E. Cameron, R. Andino, *Proc. Natl. Acad. Sci. USA* **98**, 6895 (2001).
20. A. J. Bernal, M. M. Gomes da Silva, D. B. Musungaie, E. Kovalchuk, A. Gonzalez, *et al.*, *N. Engl. J. Med.* **386**, 509 (2022).
21. C. J. Gordon, E. P. Tchesnokov, R. F. Schinazi, M. Götte, *J. Biol. Chem.* **297**, 100770 (2021).