nature genetics

© 2005 Nature Publishing Group http://www.nature.com/naturegenetics

Real ribozymes suggest a relaxed error threshold

Ádám Kun[1,2], Mauro Santos[3] & Eörs Szathmáry[1,2,4]

The error threshold for replication, the critical copying fidelity below which the fittest genotype deterministically disappears, limits the length of the genome that can be maintained by selection. Primordial replication must have been error-prone, and so early replicators are thought to have been necessarily short[1]. The error threshold also depends on the fitness landscape. In an RNA world[2], many neutral and compensatory mutations can raise the threshold, below which the functional phenotype[3], rather than a particular sequence, is still present[4,5]. Here we show, on the basis of comparative analysis of two extensively mutagenized ribozymes, that with a copying fidelity of 0.999 per digit per replication the phenotypic error threshold rises well above 7,000 nucleotides, which permits the selective maintenance of a functionally rich riboorganism[6] with a genome of more than 100 different genes, the size of a tRNA. This requires an order of magnitude of improvement in the accuracy of *in vitro*–generated polymerase ribozymes[7,8]. Incidentally, this genome size coincides with that estimated for a minimal cell achieved by top-down analysis[9], omitting the genes dealing with translation.

The origin of life has been plagued by the fundamental obstacle to increasing in complexity summarized by Eigen's[1] paradox: no enzymes without a large genome and no large genome without enzymes[10]. This question applies in an RNA world also. How many different genes (ribozymes) can be selectively maintained in a primordial genome? Eigen's insight of an error threshold quantifies the problem. Following a simplified treatment[11,12], we have

$$L < \ln(s)/(1-q), \qquad (1)$$

where L is the maximum allowed genome size in nucleotides, q is the critical (threshold) copying fidelity, $s = A/a$ is the 'selective superiority' of the fittest (master) sequence and A and a are the Malthusian growth rates of the master and the inferior mutants, respectively. In this simplified treatment, all mutants share the same replication rate, and neutral mutations of and back mutations to the master are ignored.

The error threshold was first defined in relation to a particular genotype. In an RNA world, however, there will be many neutral and compensatory mutations, which allow the preservation or the restoration of the fittest phenotype[3] rather than of a single genotype. Other things being equal, this will increase the error threshold (thus, longer genomes will become maintainable). It is crucial to quantify this effect. Because in an RNA world the functional ribozymes will have the strongest effect on fitness[12], one should gather the pertinent data from known ribozymes. There is just enough empirical evidence to formulate an encouraging statement.

Nobody has yet seen or constructed a functional riboorganism, but we know how mutations affect the functionality of certain ribozymes. Although ribozymes in a metabolically rich riboorganism[13] will interact through a network in various ways, in this study we will use ribozyme activity as a proxy for organism fitness. The aim of the study is therefore to infer the fitness landscape of ribozymes from existing data on ribozyme mutagenesis and then to estimate a revised phenotypic error threshold as a function of copying fidelity.

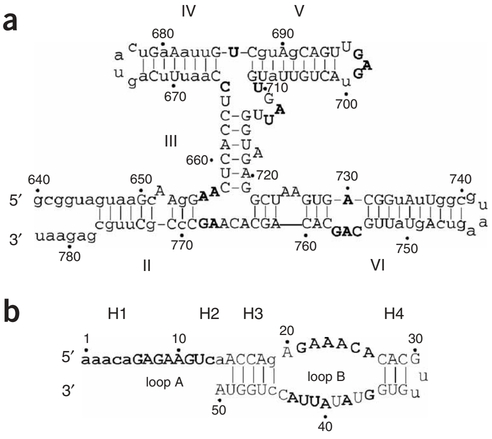

a
IV
V
680
690
a CuGaAauuG-U-CguAgCAGUUG
u gaCuUuaaC GUaUUGUCAu GA
670
C
U•710
700
U
G
C-GUUA
C-G
III
A-UA
660• C-GA
U-A•720
640
650
730
740
5′ gcgguaguaaGcA AgGAA C-G GCUAA GUG-A-CGGuAuUggcgu
cguuCg-CcCGAACACGA—CACGAC GUUaUgAcuga a
3′ uaagag
780
770
760
750
II
VI
b
H1
H2
H3
H4
1
10
20
30
5′ aaacaGAGAAGUcaACCAg AGAAACA CACG
loop A
loop B
u
3′
AUGGUCCAUUAUAUG GUGu
50
40

Figure 1 Mutagenized ribozymes. Secondary structures of (**a**) the *Neurospora* VS ribozyme and (**b**) the hairpin ribozyme indicating the different regions. Position numbering follows standard conventions[25–27]. Capitalized nucleotides specify those sites that have been subjected to mutagenesis experiments and for which enzymatic activities of mutants are available. For the *Neurospora* VS ribozyme, 183 mutants affecting 83 of 144 positions, excluding insertions and deletions, were considered. For the hairpin ribozyme, 142 mutants affecting 39 of 50 positions of the ribozyme and some part of the substrate region were considered. Nucleotides marked in bold are the critical sites.

[1] Collegium Budapest, Institute for Advanced Study, Szentháromság u. 2. Budapest H-1014, Hungary. [2] Department of Plant Taxonomy and Ecology, Eötvös University, Pázmány Péter sétány 1/C, Budapest H-1117, Hungary. [3] Departament de Genètica i de Microbiologia, Grup de Biologia Evolutiva, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain. [4] Research Group of Ecology and Theoretical Biology, Eötvös University, Hungarian Academy of Science, Pázmány Péter sétány 1/c, H-1117 Budapest, Hungary. Correspondence should be addressed to E.S. (szathmary@colbud.hu).

Published online 28 August 2005; doi:10.1038/ng1621

© 2005 Nature Publishing Group http://www.nature.com/naturegenetics

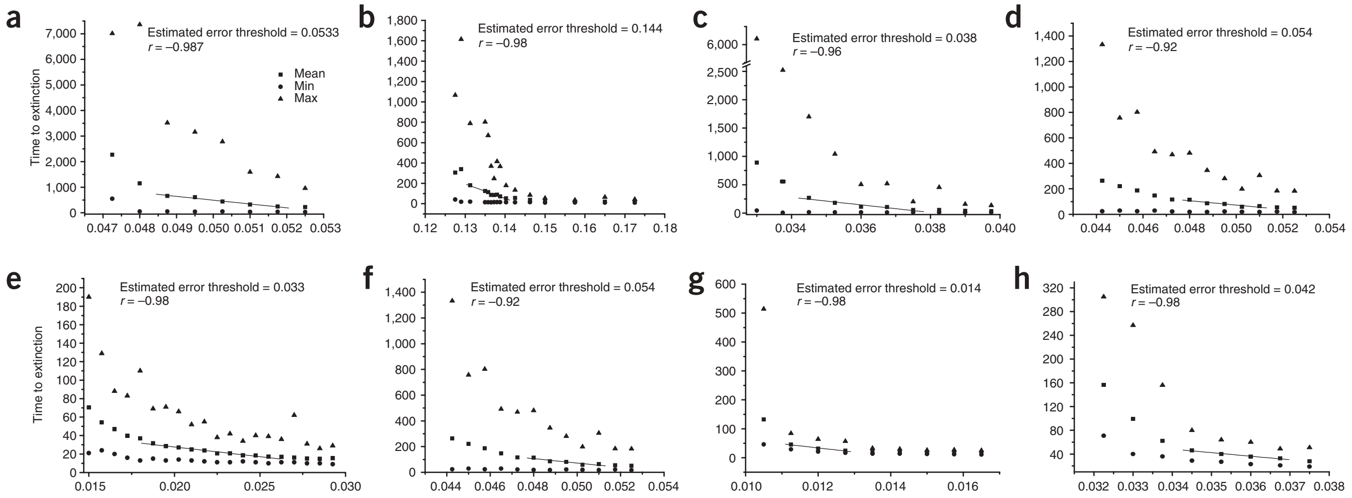


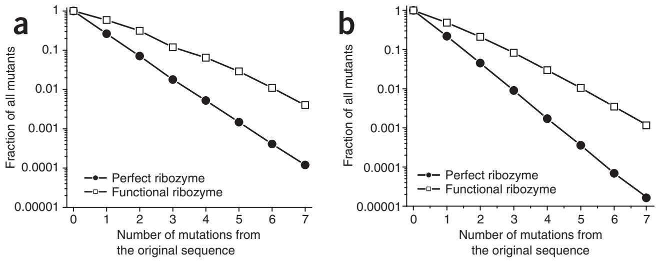
Figure 2 Time to extinction in generations as a function of the per digit effective mutation rate (μ*) in a population of constant size with $N = 10{,}000$ molecules for the *Neurospora* VS ribozyme (**a**,**c**,**e**,**g**) and the hairpin ribozyme (**b**,**d**,**f**,**h**). For each mutation rate, 100 independent runs were obtained, and the mean (solid squares), minimum (solid circles) and maximum (solid triangles) times to extinction were calculated. Solid lines represent the linear fit after extrapolating to an infinite population size to estimate the phenotypic error threshold and *r* is the correlation coefficient. (**a**,**b**) Both structural and functional information are incorporated to infer the fitness landscapes of the ribozymes. (**c**–**f**) From Mount Fuji fitness landscapes that do not take into account ribozyme secondary structure but otherwise use empirically estimated enzymatic activities at those positions where experimental information is available and uniformly predefined values with 80% (**c**,**d**) or 20% (**e**,**f**) wild-type activity at positions where empirical information could not be derived from a different nucleotide. (**g**,**h**) From Fisher's single-peak fitness-landscape assuming Eigen's[1] model with mutant sequences having enzymatic activity 0.217 at each position for the *Neurospora* VS ribozyme (*i.e.*, the average activity of all experimentally tested one-point mutants) and 0.188 for the hairpin ribozyme.

To construct a fitness or functionality landscape for a ribozyme, several requirements must be met: (i) its secondary structure must be experimentally determined; (ii) this secondary structure cannot contain a pseudoknot, a special structural element that conventional RNA folding algorithms cannot cope with satisfactorily; (iii) mutagenesis experiments must have identified all important sites and nucleotides; and (iv) the size of the ribozyme must not be too long, or calculations would be practically unfeasible. The first requirement excludes most known ribozymes, because, apart from the function, only the sequence has been determined. The naturally occurring ribozymes generally fulfill the third requirement, but hepatitis delta virus does not meet the second requirement, and group I and II introns, as well as RNAase P, do not meet the fourth requirement. This leaves the hammerhead, the hairpin and the *Neurospora* VS ribozymes as possible candidates. We used the hairpin and the *Neurospora* VS ribozymes for our study (**Fig. 1**). Both are relatively short, naturally occurring, self-cleaving ribozymes, which can be divided into a *trans*-acting enzyme-substrate system in which the *trans*-acting enzyme part does not contain a pseudoknot.

The construction of the fitness or functionality landscape is based on four general observations: (i) the maintenance of the secondary structure is a key factor in retaining enzymatic activity, but the nature of most individual base pairs is not important and many can be reversed or replaced by a different pair without loss of activity, so long as a base pair is retained at a given position[14,15]; (ii) the structure can have slight variations, which, in most cases, manifest in some mismatched base pairs or some deletions or elongation in a helical region; (iii) there are crucial regions in the molecule where the nature of the base is also important (**Fig. 1**); (iv) the effect of multiple mutations is multiplicative (*i.e.*, the product of the activities of single mutants provides the activity of the multiple mutants)[16].

Accordingly, we devised an algorithm to assemble the fitness or functionality landscapes for any ribozyme with enough directed mutagenesis data available and applied it to the *Neurospora* VS and the hairpin ribozymes. The proposed algorithm assigns a relative activity to each of the 4^L possible RNA sequences of length L. For simplicity, we restricted the sequence space to sequences of a given length, but the algorithm could also be applied if insertions or deletions (indels) were considered (**Supplementary Note** online). There are four basic steps: (i) compatible structure (a mutant molecule should fold into a compatible secondary structure for the ribozyme to retain any activity); (ii) mispairs (those allowed mispairs decrease activity to some extent); (iii) critical sites (an empirically measured activity to every possible nucleotide at well studied critical

Figure 3 Fraction of mutants with full activity (filled circles) or any activity (open squares), as a function of the number of point mutations. For both ribozymes, all possible one-, two- and three-mutant neighbors from the original sequence were generated, and their fitness recorded. Furthermore, from each sequence containing four, five, six, seven, eight, nine and ten point mutations, a total of 10^6 randomly generated sequences were evaluated. According to the fitness landscapes (**a**), for the *Neurospora* VS ribozyme, 114 (26.4%) of 432 possible single mutants are selectively neutral. (**b**) This fraction in the 150 single mutants of the hairpin ribozyme is 22% (33 sequences).



© 2005 Nature Publishing Group http://www.nature.com/naturegenetics

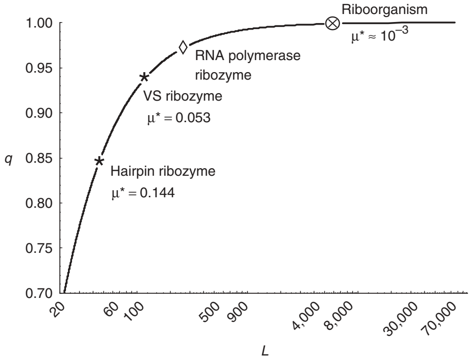


Figure 4 Relationship between the per digit replication accuracy (q) and the permissible genome size (L) estimated from equation 2 with $\lambda = 0.22$ and $s = 351$. The stars indicate the estimated error threshold ($\mu^* = 1 - q$) of the two studied ribozymes. The open rhombus indicates the average fidelity of the artificially selected template-dependent RNA polymerase ribozyme[7]; the filled circle, the estimated replication accuracy required to run a functionally rich riboorganism[6] (*Riborgis eigensis*?).

sites is assigned); and (iv) predicted structure (the predicted secondary structure the mutated sequence will fold into is matched against the formerly resolved compatible structure). An activity value pertaining to each step is calculated, and the resulting relative activity (fitness) of the sequence (A_{sequence}) is the product of all combined activities: $A_{\text{sequence}} = A_{\text{structure}} \bullet A_{\text{mispair}} \bullet A_{\text{critical}} \bullet A_{\text{energy}}$.

For the *Neurospora* VS ribozyme, we relied on 183 mutants that affect 83 of 144 nucleotide positions of the ribozyme; for the hairpin ribozyme, we used 142 mutants that affect 39 of 50 nucleotides of the ribozyme and some parts of the substrate region. Existing mutagenesis information and enzymatic activities reported at critical sites for both ribozymes is given in **Supplementary Note** and **Supplementary Tables 1** and **2** online. Very limited information on orthologous sequences is available for the *Neurospora* VS ribozyme, and the predicted enzymatic activity for those slightly different sequences is quite high (**Supplementary Note** online). For the hairpin ribozyme, orthologous sequences uncover a new structure that can be easily incorporated in the algorithm, which would predict a slight increase in enzymatic activity (**Supplementary Note** online). Because we have restricted the sequence space to sequences of fixed length, however, no mutant ribozyme would fold into that structure. Therefore, all experiments used the wild-type sequences (**Fig. 1**).

From the fitness or functionality landscapes, the estimated phenotypic error thresholds for the *Neurospora* VS and hairpin ribozymes are $\mu^* = 0.0533$ and $\mu^* = 0.144$, respectively, where μ^* is the effective mutation rate per nucleotide per replication (**Fig. 2a,b**). As expected, these values are substantially higher than those inferred from fitness landscapes that do not take into account the secondary structure of the ribozymes but do include information on single mutational effects. Thus, for the Mount Fuji–type fitness landscape, the error thresholds are $\mu^* = 0.033$–0.038 for the *Neurospora* VS ribozyme and $\mu^* = 0.054$–0.134 for the hairpin ribozyme, depending on the assumed functional importance of the different nucleotide positions (**Fig. 2c–f**). For the Fisher's single-peak fitness landscape, similar to the one originally used by Eigen[1], using the average enzymatic activity of the single point mutations as surrogate of fitness, the error threshold estimates are substantially lower: $\mu^* = 0.014$ for the *Neurospora* VS ribozyme and $\mu^* = 0.042$ for the hairpin ribozyme (**Fig. 2g,h**).

Fitness and functionality predictions are quite good for critical sites but less accurate for other positions, simply because fewer data are available and some assumptions are required. For example, base pair changes do not generally alter enzymatic activities very much, and these range between 0.7 and 1.3, which is simply taken as 1 in the landscape. In addition, we disregarded epistatic interactions that might exist for some sites in the form of diminishing return epistasis. Thus, it seems that, for some positions in the molecules, there is diminishing epistasis (*i.e.*, less than multiplicative decline) only if the fitness (activity) is lower than a certain threshold, which would imply that our estimates of the error threshold are slightly conservative.

This is the first time to our knowledge that the fitness landscape in terms of functionality has been inferred from real ribozymes. Our first conclusion is that deleterious mutations tend to affect function approximately independently (**Fig. 3**), as was found for some protein enzymes[17–19]. Our second conclusion is that the phenotypic error threshold thus inferred alleviates Eigen's paradox. This relates to the finding that the fitness landscapes are sufficiently similar. Equation 1 cannot be used to assess the effect of the landscape on the error threshold, owing to its restrictive preconditions. A recently derived expression[5] offers a much more pertinent approximation:

$$L < -\ln(s)/\ln(q+\lambda-q\lambda), \tag{2}$$

where λ is the fraction of neutral single substitutions. For the *Neurospora* VS ribozyme, $L = 144$, $q = 0.947$ and $\lambda = 0.26$, and for the hairpin ribozyme, $L = 50$, $q = 0.856$ and $\lambda = 0.22$. Thus, for $\ln(s)$ of the *Neurospora* VS and the hairpin ribozymes, we obtain 5.761 and 5.957, respectively.

One can raise the objection that equation 2 was derived under the assumption of only two replicator phenotype classes (master and mutants). This is no problem if one considers the following explanation. If, as is legitimate for long enough sequences[20], back mutations to the master are ignored, then in mutation-selection balance, the subpopulation of the mutant classes (without the master) can be substituted by an average mutant sequence or phenotype with the appropriate fitness value. Incidentally, from the values of $\ln(s)$ we obtain the s values 318 and 386 for the *Neurospora* VS and hairpin ribozymes, respectively. Thus, for example, the master phenotype class has a fitness advantage 318 times greater than that of this average mutant competitor.

The fitness values obtained allow us to reconsider Eigen's paradox. Although within-gene recombination can raise the error threshold to some extent[21], the required accuracy of a sufficient replicase ribozyme in a riboorganism was not known. Substituting an accuracy of $q = 0.999$, in the lower bound of viral RNA replicases[22], into equation 2 and using the two obtained values for λ, we find that $L \approx 7{,}000$–$8{,}000$ (**Fig. 4**). Such a ribozyme could replicate a genome consisting of more than 100 different genes of sequence length 70 each or more than 70 different genes of sequence length 100 each. This would be sufficient to run a functionally rich riboorganism, estimated to carry about this number of genes[6]. A recent analysis of a core minimal bacterial gene set places the value at ~200 genes[9]. If we take away the genes coding for the whole contemporary translation system, we are in the same range.

The artificial template-dependent RNA polymerase ribozyme previously selected[7] has an average fidelity $q = 0.97$. Using equation 2 and the fitness or functionality landscape (**Fig. 3**) obtained for the *Neurospora* VS and the hairpin ribozymes (an admitted leap), we conclude that the accuracy of this ribozyme would allow the maintenance of replicators with length $L \approx 250$, which means that this ribozyme could replicate itself if other conditions (such as

© 2005 Nature Publishing Group http://www.nature.com/naturegenetics

processivity) were favorable. To eliminate the burden of Eigen's paradox, a replicase with an error rate of 10^{-3} per nucleotide per replication might have been sufficient to provide minimal life requirements in the RNA world (**Fig. 4**).

METHODS

Compatible structure. A sequence is said to be compatible with the secondary structure if for every base pair *i-j* the nucleotides at the *i*th and *j*th positions in the sequence can form one of the allowed base pairs (A-U, U-A, G-C, C-G, U-G, G-U). Enforcing strict compatibility might result in an overestimation of the negative effects, as some mispair mutants can retain a relatively high level of enzymatic activity (**Supplementary Tables 1** and **2** online). As a result, even sequences with partial compatibility should be considered fully compatible in this step; the negative effects of the tolerated mispairs will be taken into account in the next step of the algorithm. On the other hand, two contiguous mispairs are usually not tolerated for either the *Neurospora* VS (**Supplementary Table 1** online) or the hairpin (**Supplementary Table 2** online) ribozymes. If a sequence is not compatible, even considering the possibility of mispairs, with any possible structures, then it was assumed to have no activity and its fitness was set to 0. The factor in this step ($A_{\text{structure}}$) is the activity associated with the structure to which the sequence can fold. The activities of the various possible structures can be different.

Mispairs. When a sequence is perfectly compatible with a structure (*i.e.*, there are no mispairs in it) then $A_{\text{mispair}} = 1$; otherwise every single allowed mispair decreases activity to some extent. In other words, every mispair ($i \bullet j$) has an associated relative enzymatic activity $A_{\text{mispair},\ (i \bullet j)}$, and the activity factor for this step is the cumulative product of the individual activities: $A_{\text{mispair}} = \Pi A_{\text{mispair},\ (i \bullet j)}$.

Critical sites. The nature of nucleotides at critical sites of the molecule is taken into account in the third step of the algorithm. Those sites are well studied, and so we can assign an empirically measured activity to nearly every possible nucleotide at these positions. All possible single mutants of the single-stranded regions of the hairpin ribozyme have been analyzed (**Supplementary Table 2** online). As before, the product of the individual activities ($A_{\text{critical},\ i}$) gives the activity factor for this step: $A_{\text{critical}} = \Pi A_{\text{critical},\ i}$.

Predicted structure. The last step of the algorithm consists of predicting the secondary structure the mutant sequence will fold into and comparing it with the structure resolved in the first step. We used the Vienna RNA package[23,24] for secondary structure prediction. The predicted minimum free energy structure of the wild-type ribozyme sequence does not always correspond with the actual secondary structure. In this case, that structure can also be accepted as a good structure. Furthermore, if mispairs are allowed then they must be taken into account during structure comparisons. When the predicted and the target structure are the same, then $A_{\text{energy}} = 1$; otherwise, $A_{\text{energy}} = 0$. This step is undoubtedly the most costly in terms of CPU time.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank I. Hofacker for his help with the Vienna RNA package, P. Mezey for allowing us to use the computer facilities at MUN (Canada) and F. Kondrashov for comments on an earlier version of the manuscript. Computer facilities were provided by Microdigit. This work was supported by grant and postdoctoral fellowship from the Hungarian National Research Fund (OTKA) to Á.K. M.S. is partially supported by Fundación Ramón Areces (Spain). This work was also supported by the COST D27 action (Prebiotic chemistry and early evolution).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 3 May; accepted 7 July 2005
Published online at http://www.nature.com/naturegenetics/

1. Eigen, M. Self organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **10**, 465–523 (1971).
2. Gilbert, W. The RNA world. *Nature* **319**, 618 (1986).
3. Huynen, M.A., Stadler, P.F. & Fontana, W. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA* **93**, 397–401 (1996).
4. Reidys, C., Forst, C.V. & Schuster, P. Replication and mutation on neutral networks. *Bull. Math. Biol.* **63**, 57–94 (2001).
5. Takeuchi, N., Poorthuis, P.H. & Hogeweg, P. Phenotypic error threshold; additivity and epistasis in RNA evolution. *BMC Evol. Biol.* **5**, 9 (2005).
6. Jeffares, D.C., Poole, A.M. & Penny, D. Relics from the RNA world. *J. Mol. Evol.* **46**, 18–36 (1998).
7. Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasen, M.E. & Bartel, D.P. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292**, 1319–1325 (2001).
8. Müller, U.F. & Bartel, D.P. Substrate 2′-hydroxyl groups required for ribozyme-catalyzed polymerization. *Chem. Biol.* **10**, 799–806 (2003).
9. Gil, R., Silva, F.J., Peretó, J. & Moya, A. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68**, 518–537 (2004).
10. Maynard Smith, J. Hypercycles and the origin of life. *Nature* **20**, 445–446 (1979).
11. Maynard Smith, J. Models of evolution. *Proc. R. Soc. Lond. B* **219**, 315–325 (1983).
12. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Oxford Univ. Press, Oxford, 1995).
13. Benner, B., Ellington, A.D., Ge, L., Gasfeld, A. & Leanz, G.F. Natural selection, protein enzgineering, and the last riboorganism: rational model building in biochemistry. *Cold Spring Harb. Symp. Quant. Biol.* **52**, 56–63 (1987).
14. Fedor, M. Structure and function of the hairpin ribozyme. *J. Mol. Biol.* **297**, 269–291 (2000).
15. Lafontaine, D.A., Norman, D.G. & Lilley, D.M.J. The structure and active site of the Varkund satellite ribozyme. *Biochem. Soc. Trans.* **30**, 1170–1175 (2002).
16. Lehman, N. & Joyce, G.F. Evolution *in vitro*: analysis of a lineage of ribozymes. *Curr. Biol.* **3**, 723–734 (1993).
17. Takeda, Y., Sarai, A. & Rivera, V.M. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl. Acad. Sci. USA* **86**, 439–443 (1989).
18. Sandberg, W.S. & Terwilliger, T.C. Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc. Natl. Acad. Sci. USA* **90**, 8367–8371 (1993).
19. Skinner, M.M. & Terwilliger, T.C. Potential use of additivity of mutational effects in simplifying protein engineering. *Proc. Natl. Acad. Sci. USA* **93**, 10753–10757 (1996).
20. Eigen, M., McCaskill, J.S. & Schuster, P. The molecular quasispecies. *Adv. Chem. Phys.* **75**, 149–263 (1989).
21. Santos, M., Zintzaras, E. & Szathmáry, E. Recombination in primeval genomes: a step forward but still a long leap from maintaining a sizeable genome. *J. Mol. Evol.* **59**, 507–519 (2004).
22. Domingo, E. & Holland, J.J. in *The Evolutionary Biology of Viruses* (ed. Morse, S.S.) 161–183 (Raven, New York, 1994).
23. Hofacker, I.L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188 (1994).
24. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
25. Beattie, T.L., Olive, J.E. & Collins, R.A. A secondary-structure model for the self-cleaving region of the *Neurospora* VS RNA. *Proc. Natl. Acad. Sci. USA* **92**, 4686–4690 (1995).
26. Butcher, S.E. & Burke, J.M. Structure-mapping of the hairpin ribozyme. Magnesium-dependent folding and evidence for tertiary interactions within the ribozyme-substrate complex. *J. Mol. Biol.* **244**, 52–63 (1994).
27. Butcher, S.E. & Burke, J.M. A photo-cross-linkable tertiary structure motif found in functionally distinct RNA molecules is essential for catalytic function of the hairpin ribozyme. *Biochemistry* **33**, 992–999 (1994).