Supplementary Note for:

Real ribozymes suggest a relaxed error threshold

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

[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*

A word about fitness landscapes

The idea to study evolution by visualizing the distribution of fitness values as a kind of landscape was introduced by Wright [*1*]. Maynard Smith [*2*] was the first to coin the term "protein space"; namely, a high dimensional space where each sequence of length $\mathcal{L}$ amino acids (out of $20^{\mathcal{L}}$ possible sequences) represents one point and is next to $19\mathcal{L}$ points representing all the one-mutant neighbours of each other. This breakthrough can obviously be extended to RNA sequences. Indeed, the relationship between the primary and secondary structures of RNA molecules has been widely studied as a realistic example of multidimensional landscapes [*3*]. However, the actual relationship between RNA sequence, secondary structure, and fitness/functionality landscape of a ribozyme can only be appraised from empirically driven information that will be necessarily sparse for obvious reasons linked to the extremely large dimensionality of sequence space.

Construction of the empirically-supported fitness landscape for the *Neurospora* VS ribozyme

Compatible structures. – Some mispair mutants can retain a relatively high level of enzymatic activity: the C662G mispair mutation decreases activity to 23% that of the wild-type (taken to be equal to one; [*4*]), and the relative activity of the G33U mispair mutant of the hairpin ribozyme is 0.7 [*5*]. Therefore, even sequences with partial compatibility should be considered fully compatible in this step, but the negative effects of the tolerated mispairs are taken into account in the next step of the algorithm. On the other hand, two contiguous mispairs are usually not tolerated: four out of six adjacent mispairs documented for the VS ribozyme did not show any measurable activity ([G727C, U728A]; [G722C, C723G]; [G762C, C763G]; [A759U, C760G]), and the remaining two had very low activities ([G716C, U717A] with activity 0.02; [A661U, C662G] with activity 0.06).

In addition to the wild-type structure (Fig. 1a) we have also considered structures with the A652 bulge deleted ($A_{structure} = 0.013$; [*6*]), the A718 bulge deleted ($A_{structure} = 0.136$) or paired ($A_{structure} = 0.82$) [*6*], stem III lengthened to 8 base-pairs ($A_{structure} = 0.087$; [*6*]), stem V shortened to 7 base-pairs ($A_{structure} = 0.045$;[*7*]), stem IV shortened to 4 ($A_{structure} = 0.59$) or 6

($A_{structure} = 0.47$) base-pairs [*6*], and stem VI shortened to 6 ($A_{structure} = 1.00$) or 8 base-pairs ($A_{structure} = 0.97$)[*7*]. In order to maintain the total length constant (i.e., $\mathcal{L} = 144$), nucleotides were added or deleted from the ends of the sequence in all possible combinations.

Mispairs.– Mispairs are generally tolerated but decrease enzymatic activity to some extent. The only exceptions are: two mispairs cannot be adjacent; a mispair cannot be next to a junction; a mispair cannot occur at positions 653-655; and a stem-loop cannot contain more than two mispairs. In addition, a mispair next to the active site (i.e., positions [731:754] and [729:758]) dramatically decreases activity to $A_{mispair,(i \bullet j)} = 0.025$, and a mispair next to stem V (at positions [695:701]) decreases activity to $A_{mispair,(i \bullet j)} = 0.05$. Finally, a mispair next to the loops of stem IV and VI decrease activity to $A_{mispair,(i \bullet j)} = 0.80$.

Critical sites.– The 16 identified critical sites are located in the junctions (positions 656, 657, 665, 686, 710, 712, 713, 767 and 768; [*7,8*]), in the substrate binding region (positions 697-699; [*9*]), and in the A730 internal loop (positions 730, 755, 756 and 757) that is the active site of the ribozyme [*10-12*] (Fig. 1a). Supplementary Note Table A lists the enzymatic activities associated with the critical sites.

Predicted structure. – The MFE structure of the original sequence of the VS ribozyme is not the experimentally determined secondary structure. The two structures differ in only 3 base pairs in stem-loop VI, and the energy difference between the two is a mere 0.3 kcal. It must be noted, however, that the nuclear magnetic resonance structure of the isolated stem-loop VI is the same as in the MFE structure [*13*]. Thus, both structures might play an important role in the catalysis. Accordingly, we accept the sequence if it folds either to the MFE or the experimentally predicted structure of the wild-type VS ribozyme.
The enzymatic activity of the molecule is simply the cumulative product of the activity factors estimated in the four steps of the algorithm. If the resulting activity was less than the lowest activity that can be reliably measured (that is, $A_{sequence} \le 10^{-3}$), then enzymatic activity (fitness) was set to 0.

Additional sequences. – As indicated by Saville and Collins [*14*] (see also Guo *et al.* [*15*]) there are 6 known isolates of the *Neurospora* VS RNA. Five of them have been isolated from *N. intermedia* (#1805, #1809, #1821, #1823, #2635) and one from *N. sitophila* (#2492). In the region of the ribozyme (from positions 640 to 783) the isolates #1809, #1821, #1823 and #2653 are identical. The isolate #1805 differs at position 677, where it has U instead of C. This causes the loop at the end of stem IV to be enlarged by 6 nucleotides and there is no 673A:678U base pair, which according to the fitness landscape would result in a 20% decreased in enzymatic activity with respect to the sequence shown in Fig. 1a. The isolate #2492 differs at position 769, where it has U instead of C. This would cause no change in the secondary structure and the predicted activity is 1.0.

Construction of the empirically-supported fitness landscape for the hairpin ribozyme

Compatible structures. – Aside from the wild-type structure (Fig. 1b) we considered the following structures having various lengths of H1, H2 and H4 but still retaining full activity ($A_{structure} = 1.0$) [*16-21*]: the first two nucleotides deleted and a base-pair inserted into H4; the first four nucleotides deleted and two base-pairs inserted into H4; deletion of positions 13 and 14 and insertion of a base-pair to H4; deletion of positions 1, 2, 13 and 14 and insertion of two base-pairs to H4. Moreover, the deletions of U49 and A50 are also tolerated with minimal loss of activity [*22*], as well as structures with deletion of A50 and adding a nucleotide to the beginning of the sequence ($A_{structure} = 0.49$), removal of both U49 and A50 and adding two nucleotide to the beginning of the sequence ($A_{structure} = 0.38$), or insertion of a base-pair into H4 ($A_{structure} = 0.38$).

Mispairs.– Similarly to the VS ribozyme mispairs are also generally tolerated in the hairpin ribozyme but decrease enzymatic activity to some extent. Exceptions are: two mispairs cannot be adjacent, and no stem-loop can contain more than two mispairs. A mispair in H4 decreases activity to $A_{mispair,(i \bullet j)} = 0.2$, and in H3 to $A_{mispair,(i \bullet j)} = 0.05$.

Critical sites.– All possible single mutants of the single stranded regions of the hairpin ribozyme have been analyzed [*23*,*24*]. The substrate binding region (positions 1-14) is considered a critical region, as well as most nucleotides in loop A (Fig. 1b). Supplementary Note Table B lists the enzymatic activities associated with the critical sites.

Predicted structure.– The MFE structure predicted for the hairpin ribozyme is not the experimentally determined secondary structure. The two structures differ considerably: H3 is predicted correctly, but the hairpin loop at the end of H4 is predicted to be 2 nucleotides longer (i.e., H4 starts at position 25) and extraneous base-pairs are predicted inside the single stranded regions. The substrate binding region (H1 loop A and H2) was not included during the folding process. We accepted any structure with H3 and H4 in place – either according to the real secondary structure or according to the MFE structure –, and we disregarded any extraneous base-pairs predicted inside loop B. As before, activities lower than $A_{sequence} \le 10^{-3}$ were set to 0.

Additional sequences. – The usually used hairpin sequence is from tobacco ringspot virus (sTRSV). Hampel and co-workers [*25*] report two homologous sequences derived from the satellite RNA of sTRSV: one of them is found in the satellite RNA of the chicory yellow mottle virus (sCYMV1) and the other in the arabis mosaic virus (sArMV). The ribozymes derived from these RNAs have very similar activity when compared to the sTRSV hairpin ribozyme (sCYMV has relative activity 0.89, and that for sArMV is 0.72) [*25*]. 34 out of 50 positions are conserved in the three homologous sequences including all critical sites in loop B, and the whole H4. As they differ in their substrate there are some changes in the substrate binding parts.
sCYMV1 sequence differs by 13 nt when compared to sTRSV. Five differences are found in the substrate binding region, but as the substrate sequence is also different we can safely disregard this variation. The differences in the single-stranded regions (A15G, A20G, C44U, U49C) cause no change in enzymatic activity whatsoever. The additional difference is the

extension of the hairpin at the end of H4, from 3 to 5 nucleotides. Such structure had not been included in the algorithm as a possible structure for the hairpin, but upon inclusion the sCYMV1 sequence would be accepted as a fully active ribozyme.
The sequence of sArMV differs from sTRSV by 16 nt. Five of these differences are found in the substrate binding region, and they were also disregarded. The remaining differences either affect a single stranded region (A20C, U49C, A50G), change two base pairs in H3 (C16A:G48U, A18G:U46C), or change the end of H4 as above. The first two changes have no consequence in enzymatic activity. When the changed H4 structure is incorporated in the algorithm, the sArMV sequence would be accepted as a fully active ribozyme.

Estimation of the error threshold

In order to resolve the "realistic" error threshold for (e.g.) the VS ribozyme we have explored the dynamics of a population of RNA molecules with $\mathcal{L} = 144$ at various mutation rates per nucleotide per replication (μ) in a so-called Moran process [*26*]. Thus, consider an initial homogeneous population of wild-type sequences of size *N*. At each time step a sequence is chosen for replication with a probability proportional to its fitness:

$$p_i = \frac{\mathrm{A}_{\mathrm{E}i}}{\sum_j \mathrm{A}_{\mathrm{E}j}} \ ,$$

where $\sum_j \mathrm{A}_{\mathrm{E}j}$ is the sum of activities for all sequences in the population. The sequence is then replicated with error rate μ (only point mutations were considered). Because one quarter of the times (assuming equal probability for each nucleotide) no effective change will occur in the position even though there is a mutational event, the effective mutation rate is $\mu^* = 0.75\mu$. Finally, the new sequence replaces a randomly chosen one, which allows keeping a constant population of molecules and is also equivalent to assuming that the rate of degradation is the same for all molecules and independent of enzymatic activity [*27*].
We should emphasize here that the occurrence of thresholds for error propagation was originally derived as a deterministic kinetic theory that is only valid in the limit case of an infinite number of molecules. Alves and Fontanari [*28*] have extended it to finite populations and found that the critical error rate per site per replication decreases linearly with $1/N$. In our present case we extrapolated to an infinite population size by recording the time to extinction (that is, the number of generations when no functional ribozymes would remain in the population) at various error rates and fitting a straight line to those last few points which still showed a downward trend. The error threshold μ^* is then the intersection of the line with the error rate axis (Figure 2).

Supplementary Note Table A: Activities of the critical sites in the *Neurospora* VS ribozyme.

Critical site	$A_{critical,i}$			
	A	**U**	**C**	**G**
656	1.000	0.002	0.003	0.061
657	1.000	0.063	0.063 [a]	0.063 [a]
665	0.014	0.01 [b]	1.000	0.01
686	0.006	1.000	0.006 [a]	0.006 [a]
697	0.000	0.001	0.000	1.000
698	1.000	0.000	0.041	0.000
699	0.006	0.018	1.000	0.000
710	0.002	1.000	0.029	0.002
712	1.000	0.005	0.005 [b]	0.006
713	0.019	1.000	0.136	0.019
730	1.000	0.036	0.055	0.011
755	0.840	0.170	1.000	0.019
756	1.000	0.001	0.002	0.002
757	0.044	0.018	0.014	1.000
767	1.000	0.047	0.047 [a]	0.047 [a]
768	0.573	0.015	0.573 [c]	1.000

[a] No data available. We assume that enzymatic activity is the same as for the known mutant.

[b] No data available. We assume that enzymatic activity is the same as for the known mutant with the lower activity.

[c] No data available. We assume that enzymatic activity is the same as for the known mutant with the higher activity.

Supplementary Note Table B Activities of the critical sites in the hairpin ribozyme.

Critical site	$A_{critical,i}$			
	A	**U**	**C**	**G**
1	1.000	0.700 [a]	0.700 [a]	0.700 [a]
2	1.000	0.700 [a]	0.700 [a]	0.700 [a]
3	1.000	0.300 [a]	0.300 [a]	0.300 [a]
4	0.000 [a]	0.000 [a]	1.000	0.000 [a]
5	1.000	0.050 [a]	0.050 [a]	0.050 [a]
6	0.000	0.000	0.000	1.000
7	1.000	0.260	0.164	0.520
8	0.000	0.000	1.000	0.000
9	1.000	0.001	0.005	0.002
10	1.000	0.000	0.000	0.049
11	0.000	0.000	0.000	1.000
12	0.000 [a]	1.000	0.000 [a]	0.000 [a]
13	0.000 [a]	0.000 [a]	1.000	0.000 [a]
21	0.068	0.075	0.130	1.000
22	1.000	0.000	0.000	0.000
23	1.000	0.000	0.000	0.000
24	1.000	0.000	0.005	0.007
25	0.000	0.000	1.000	0.000
26	1.000	0.008	0.178	0.005
36	0.027	0.123	0.021	1.000
38	1.000	0.000	0.000	0.000
40	1.000	0.045	0.046	0.003
41	0.018	1.000	0.092	0.000
42	0.002	1.000	0.002	0.000
43	1.000	0.002	0.000	0.085

[a] Derived from the activity of a mispair, containing mutation in the substrate part.

References

1. Wright, S. in *Proceedings of the Sixth International Congress on Genetics* 355-366 (1932).
2. Maynard Smith, J. Natural selection and the concept of protein space. *Science* 225, 563-564 (1970).
3. Schuster, P. Evolution *in silico* and *in vitro*: The RNA model. *Biol. Chem.* 382, 1301-1314 (2001).
4. 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).
5. Berzal-Herranz, A., Joseph, S., Chowrira, B. M., Butcher, S. E. & Burke, J. M. Essential nucleotide sequences and secondary structure elements of the hairpin ribozyme. *The EMBO Journal* 12, 2567-2574 (1993).
6. Hofacker, I. L. et al. Fast folding and comparison of RNA secondary structures. *Monatchefte für Chemie* 125, 167-188 (1994).
7. Lafontaine, D. A., Norman, D. G. & Lilley, D. M. J. The global structure of the VS ribozyme. *The EMBO Journal* 21, 2461-2471 (2002).
8. Lafontaine, D. A., Norman, D. G. & Lilley, D. M. J. Structure, folding and activity of the VS ribozyme: importance of the 2-3-6 helical junction. *The EMBO Journal* 20, 1415-1424 (2001).
9. Rastogi, T., Beattie, T. L., Olive, J. E. & Collins, R. A. A long-range pseudoknot is required for activity of the *Neurospora* VS ribozyme. *The EMBO Journal* 15, 2820-2825 (1996).
10. Lafontaine, D. A., Wilson, T. J., Norman, D. G. & Lilley, D. M. J. The A730 loop is an important component of the active site of the VS ribozyme. *Journal of Molecular Biology* 312, 663-674 (2001).
11. Sood, V. D. & Collins, R. A. Identification of the catalytic subdomain of the VS ribozyme and evidence for remarkable sequence tolerance in the active site loop. *Journal of Molecular Biology* 320, 443-454 (2002).
12. Lafontaine, D. A., Wilson, T. J., Zhao, Z.-Y. & Lilley, D. M. J. Functional group requirements in the probable active site of the VS ribozyme. *Journal of Molecular Biology* 323, 23-34 (2002).
13. Flinders, J. & Dieckmann, T. The solution structure of the VS ribozyme active site loop reveals a dynamic "hot-spot". *Journal of Molecular Biology* 341, 935-949 (2004).
14. Saville, B. J. & Collins, R. A. A site-specific self-cleavage reaction performed by a novel RNA in *Neurospora* mitochondria. *Cell* 61, 685-696 (1990).
15. Guo, H. C. T. et al. Nucleotide sequence requirements for self-cleavage of *Neurospora* VS RNA. *Journal of Molecular Biology* 232, 351-361 (1993).
16. Chowrira, B. M. & Burke, J. M. Extensive phosphorothioate substitution yields highly active and nuclease-resistant hairpin ribozymes. *Nucleic Acid Research* 20, 2835-2840 (1992).
17. Sargueil, B., Pecchia, D. B. & Burke, J. M. An improved version of the hairpin ribozyme complex functions as a ribonucleoprotein complex. *Biochemistry* 34, 7739-7748 (1995).
18. Nesbitt, S., Hegg, L. A. & Fedor, M. J. An unusual pH-independent and metal-ion-independent mechanism for hairpin ribozyme catalysis. *Chem. Biol.* 4, 619-630 (1997).
19. Donahue, C. P., Yadava, R. S., Nesbitt, S. M. & Fedor, M. J. The kinetic mechanism of the hairpin ribozyme *in vivo*: influence of RNA helix stability on intracellular cleavage kinetics. *Journal of Molecular Biology* 295, 693-707 (2000).
20. Porschke, D., Burke, J. M. & Walter, N. G. Global structure and flexibility of hairpin ribozymes with extended terminal helices. *Journal of Molecular Biology* 289, 799-813 (1999).
21. Pinard, R., Lambert, D., Pothiawala, G., Major, F. & Burke, J. M. Modifications and deletions of helices within the hairpin ribozyme-substrate complex: An anctive ribozyme lacking helix 1. *RNA* 10, 395-402 (2004).
22. Anderson, P. et al. Mutagenesis of the hairpin ribozyme. *Nucleic Acid Research* 22, 1096-1100 (1994).
23. Shippy, R., Siwkowski, A. & Hampel, A. Mutational analysis of loop 1 and 5 of the hairpin ribozyme. *Biochemistry* 37, 564-570 (1998).
24. Siwkowski, A., Shippy, R. & Hampel, A. Analysis of hairpin ribozyme base mutations in loop 2 and 4 and their effects on *cis*-cleavage *in vitro*. *Biochemistry* 63, 3930-3940 (1997).
25. DeYoung, M. B., Siwkowski, A., Lian, Y. & Hampel, A. Catalytic properties of hairpin ribozymes derived from chicory yellow mottle virus and arabis mosaic virus satellite RNAs. *Biochemistry* 34, 15785-15791 (1995).

26. Moran, P. A. P. Random processes in genetics. *Proc. Camb. Phil. Soc.* 54, 60-71 (1958).
27. Bonhoeffer, S. & Stadler, P. F. Error thresholds on Correlated Fitness Landscapes. *Journal of Theoretical Biology* 164, 359-372 (1993).
28. Alves, D. & Fontanari, J. F. Error threshold in finite populations. *Physical review E* 57, 7008-7013 (1998).