Advances in Complex Systems
Vol. 16 (2013) 1250052 (17 pages)
© World Scientific Publishing Company
DOI: 10.1142/S021952591250052X

World Scientific
www.worldscientific.com

SHAPE MATTERS: EFFECT OF POINT MUTATIONS ON RNA SECONDARY STRUCTURE

SUSANNA C. MANRUBIA
Centro de Astrobiología, INTA-CSIC, Carretera de Ajalvir km. 4,
Torrejón de Ardoz, 28850 Madrid, Spain
scmanrubia@cab.inta-csic.es

RAFAEL SANJUÁN
Institut Cavanilles de Biodiversitat i Biologia Evolutiva,
Departament de Genètica,
and
Unidad Mixta de Investigación en Genómica y Salud (CSISP-UV),
Universitat de València, c/Catedrático Agustín Escardino 9,
Paterna 46980, València, Spain
rafael.sanjuan@uv.es

Received 31 August 2011
Revised 9 April 2012
Published 22 June 2012

A suitable model to dive into the properties of genotype-phenotype landscapes is the relationship between RNA sequences and their corresponding minimum free energy secondary structures. Relevant issues related to molecular evolvability and robustness to mutations have been studied in this framework. Here, we analyze the one-mutant neighborhood of the predicted secondary structure of 46 different RNAs, including tRNAs, viroids, larger molecules such as Hepatitis-δ virus, and several random sequences. The probability distribution of the effect of point mutations in linear structural motifs of the secondary structure is well fit by Pareto or Lognormal probability distributions functions, independent of the origin of the RNA molecule. This extends previous results to the case of natural sequences of diverse origins. We introduce a new indicator of robustness, the average weighted length of linear motifs (AwL) and demonstrate that it correlates with the average effect of point mutations in RNA secondary structures. Structures with a high AwL value display the highest structural robustness and cluster in particular regions of sequence space.

Keywords: RNA secondary structure; sequence-structure map; mutational effects; linear motifs.

1. Introduction

Knowledge of the functional relationship between mutation and phenotypic change is essential to develop phenomenological theories of evolution and adaptation. Still, fitness depends on a number of different features whose relative importance is often determined by environmental conditions. The efforts to derive a unique relationship between mutations and their effects thus meet a profound difficulty.

Early attempts to solve this conundrum began with simple assumptions on the dependence between fitness and the number of mutations in a genome [18], thus accepting the view that evolution was occurring on a smooth, Fujiyama-like landscape. Since the distribution of fitness effects corresponding to beneficial or deleterious mutations were thought to take different functional forms, some studies focused only on mutations having beneficial effects [12]. The assumption of a random landscape for the relation between genotype and phenotype — by ignoring correlations between the phenotype yielded by a genome and its neighbors — predicts an exponential shape for the distribution of beneficial effects [22]. Other studies have identified the dependence of mutational effects on the environment where they occur [21]. As the knowledge of the field has advanced through an accumulation of empirical observations [30, 33, 19, 23], it has become increasingly clear that current theories are of no general applicability [27, 26, 34].

A step forward may result from considering partial aspects of fitness that, though unable of fully explaining the viability of an organism in a given environment, are amenable to quantification. Previous works [16, 9, 38, 3] have carried out systematic computational studies of *in silico* populations of RNA sequences at different stages of adaptation using the secondary structure of RNA sequences as a proxy for a fitness trait. The effect of a mutation in a given molecule can be directly evaluated as the structural distance between the two folded conformations, the native and the mutant. Computational studies have revealed that the distribution of effects of point mutations on RNA secondary structure is characterized by a large number of mutations having small effect and, however, a significant number of mutations able to cause major disruptions of the secondary structure [32]. The probability distribution that better explained the numerical data is a function with a sizeable amount of large effects, either a Lognormal or a Pareto distribution [34]. These functions are related to the actual landscape that maps RNA sequence to the corresponding minimum free energy secondary structure, and support the existence of nontrivial correlations between neighboring sequences, as identified in other studies [8, 41]. The complex structure of the RNA-folding map has important implications for evolution and adaptation at the molecular [7] and likely at higher levels.

In this work, we focus on the functional form of the distribution of effects of mutations on the RNA secondary structure of a number of natural, evolved sequences. To this end, we have analyzed a set of natural RNA sequences of various origins, ranging from length 71 (tRNA of glycine of *Danio rerio*) to 1682 (genome of the Hepatitis-δ virus). For comparison, we have also analyzed five randomly generated sequences with lengths 100, 200, 500, 1500 and 3000. We have performed a systematic study of the one-mutant neighborhoods of all sequences and of the functional form of the distribution of effects of mutations on their predicted minimum free energy secondary structure. In all cases, the distributions are well fit either by Pareto or Lognormal distributions, both characterized by having a large weight in small values of the relevant variable and by fat tails (i.e., their decay is sub-exponential at large values of the variable). Among the sequences studied, we include 29 viroids analyzed in a previous work [29] in which it was suggested that viroids appear to have increased their structural robustness along their predicted phylogeny by evolving increasingly rod-like structures. We here further explore the relationship between RNA secondary structure and structural robustness, establishing a clear correlation between the average length of linear motifs in the structure and the effect of mutations, independently of the origin of the sequence.

2. Methods

2.1. *RNA sequences*

Table 1 lists the set of RNA sequences used in this work. The sequences of each viroid species, the *Escherichia coli* DH1 16S RNA, and Hepatitis-δ virus were

Table 1. RNA sequences used and their main properties. Analyzed sequences are listed in the first column. They have been ordered according to their length. Numbers 1 to 10 correspond to tRNA sequences, where the indicated triplet corresponds to the anticodon sequence; numbers 13 to 41 to viroids of the Pospiviroidae family (except for 14, 26, 27 and 41, which belong to the Avsunviroidae family); 11, 12 and 42 to 46 are as described in the Table. Second column: Average effect of mutations (σ_i) as defined in Eq. (1). Third column: Fraction of mutations with less than 10% effect on secondary structure. Fourth column: Function yielding the best fit to the distribution of effects of mutations on secondary structure and the corresponding correlation coefficient r^2. Fifth column: Sequence length. Sixth column: Average weighted length of linear motifs (AwL) in each secondary structure, as defined in Eq. (2).

RNA sequence	σ_i	Frac. with <10% effect	Best fit[a]	Length	AwL
1. *Danio rerio* tRNA Gly ACC	0.174	0.435	P, 0.902	71	27.9
2. *Caenorhabditis elegans* tRNA Asp GUC	0.318	0.311	B, 0.931	72	72
3. *Saccharomyces cerevisiae* tRNA Asp GUC	0.183	0.331	E, 0.993	72	11.2
4. *Homo sapiens* tRNA Val UAC	0.273	0.312	P, 0.994	73	20.9
5. *Bacillus subtilis* tRNA Arg CCG	0.160	0.461	P, 0.975	76	19.9
6. *Escherichia coli* tRNA Ala UGC	0.450	0.326	B, 0.983	76	76
7. *Thermus thermophilus* tRNA Lys UUU	0.223	0.459	L, 0.995	76	14.3
8. *Salmonella enterica* tRNA Pro CGG	0.326	0.258	G, 0.993	77	14.6
9. *Arabidopsis thaliana* tRNA Tyr GUA	0.194	0.359	L, 0.985	85	18.3
10. *Yersinia pestis* tRNA Tyr CGG	0.277	0.382	E, 0.992	85	16.5
11. Random RNA $n = 100$	0.212	0.152	P, 0.951	100	30.1
12. Random RNA $n = 200$	0.192	0.312	P, 0.946	200	17.1
13. Coconut cadang-cadang viroid	0.0192	1.000	L, 0.977	246	246
14. Avocado sunblotch viroid	0.0496	0.766	L, 0.972	247	247
15. Citrus IV viroid	0.0986	0.718	P, 0.909	248	248
16. Coconut tinangaja viroid	0.0219	0.912	L, 0.990	254	254
17. Hop latent viroid	0.0680	0.780	P, 0.965	256	256
18. Coleus blumei 1 viroid	0.0152	0.977	P, 0.972	284	284
19. Citrus III viroid	0.0724	0.619	P, 0.964	294	221.4
20. Hop stunt viroid	0.0351	0.839	L, 0.994	297	297
21. Coleus blumei 2 viroid	0.0616	0.721	P, 0.928	301	301
22. Apple dimple fruit viroid	0.0620	0.746	P, 0.967	306	306

Table 1. (*Continued*)

RNA sequence	σ_i	Frac. with <10% effect	Best fit[a]	Length	AwL
23. Pear blister canker viroid	0.2430	0.397	L, 0.978	315	58.7
24. Citrus bent leaf viroid	0.0275	0.885	L, 0.986	318	318
25. Apple scar skin viroid	0.0263	0.819	L, 0.979	329	329
26. Eggplant latent viroid	0.0331	0.794	P, 0.968	335	64.8
27. Peach latent mosaic viroid	0.1241	0.568	P, 0.960	337	73.9
28. Chrysanthemum stunt viroid	0.0538	0.672	P, 0.985	356	107.6
29. Potato spindle tuber viroid	0.0163	0.878	L, 0.990	359	359
30. Mexican papita viroid	0.0212	0.740	L, 0.994	360	360
31. Tomato apical stunt viroid	0.0638	0.544	P, 0.927	360	360
32. Tomato planta macho viroid	0.0184	0.809	L, 0.986	360	360
33. Tomato chlorotic dwarf viroid	0.0177	0.945	L, 0.990	360	360
34. Coleus blumei 3 viroid	0.0127	0.974	L, 0.984	361	361
35. Grapevine yellow speckle 2 viroid	0.0607	0.682	P, 0.939	363	363
36. Grapevine yellow speckle 1 viroid	0.1281	0.673	P, 0.885	367	367
37. Australian grapevine viroid	0.0298	0.731	P, 0.980	369	369
38. Columnea latent viroid	0.0318	0.836	P, 0.981	370	370
39. Iresine 1 viroid	0.0167	0.947	L, 0.986	370	370
40. Citrus exocortis viroid	0.0567	0.579	P, 0.927	371	371
41. Chrysantemum chlorotic mottle viroid	0.1820	0.547	P, 0.985	399	47.3
42. Random RNA $n = 500$	0.1235	0.557	L, 0.994	500	42.1
43. Random RNA $n = 1500$	0.0404	0.685	P, 0.995	1500	32.8
44. *E. coli* 16S	0.0988	0.513	P, 0.985	1542	33.5
45. Hepatitis-δ virus	0.0041	1.000	L, 0.998	1682	1525.7
46. Random RNA $n = 3000$	0.0660	0.647	P, 0.995	3000	33.9

[a] B = Beta, G = Gamma, E = Exponential, L = Lognormal, P = Pareto.

downloaded from GenBank. Accession codes for viroids are reported in Ref. 29. The accession code for *E. coli* DH1 is CP001637 and the 16S RNA corresponds to genome positions 1152612 to 1154141. The Hepatitis-δ virus genome accession code is NC001653. tRNA sequences were downloaded from the Genomic tRNA database web site (http://gtrnadb.ucsc.edu). Random sequences were generated by drawing one of the four nucleotides with equal probability at each position.

2.2. *RNA folding*

For secondary structure folding and calculation of Hamming distances, we used the Vienna RNA package [14], version 1.5. Folding temperature was set to 25°C for viroids and 37°C for the other sequences. Given each original sequence of length n, we compared its corresponding minimum free energy secondary structure to the $3n$ structures yielded by all its one-substitution mutants. The Hamming distance was evaluated through a position-wise comparison of each pair of secondary structures in dot-bracket notation. This yields the distance $\Delta_{i,j}$ between sequence i and its one-substitution mutant j, which we always took with positive sign. Since the studied sequence and its mutants are of the same length, the Hamming distance performs (statistically) as well as any other more complex definition of distance.

Table 2. Functions used to fit data. We have used five different probability distribution functions to fit the effects of point mutations on RNA secondary structure. We show the functional form of the probability distribution function $P(x)$ and of the corresponding cumulative probability distribution (CDF), $Q(x \leq \Lambda) = \int_0^{\Lambda} P(x)dx$ for each case. In our fits, we have used the latter, with parameters that yield the minimum least squared deviation from data.

	Exponential $[\lambda]$	$\Gamma[a, b]$	$\beta[a, b]$	Pareto $[k, a]$	Lognormal $[m, \sigma]$
$P(x)$	$\lambda e^{-\lambda x}$	$\frac{e^{-x/b}}{b^a \Gamma[a]} x^{a-1}$	$\frac{\Gamma[a+b]x^{a-1}(1-x)^{b-1}}{\Gamma[a]\Gamma[b]}$	$\frac{ak^a}{x^{a+1}}$	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-m)^2}{2\sigma^2}}$
$Q(x \leq \Lambda)$	$1 - e^{-\lambda\Lambda}$	$\frac{\gamma[a,\Lambda/b]}{\Gamma[a]}$	$\frac{\beta_{\Lambda}[a,b]}{\beta[a,b]}$	$1 - \left(\frac{k}{\Lambda}\right)^a$	$\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left[\frac{\ln(\Lambda)-m}{\sigma\sqrt{2}}\right]$

The normalized probability distribution of changes in secondary structure (i.e., distance) for each sequence i is called $\Pi_i(\Delta)$.

2.3. *Distributions of effects of mutations*

As in a previous study with computationally evolved RNA sequences [34], we used five different probability density functions to fit the distributions of mutational effects. Table 2 shows the functional form of the probability distributions assayed $P(x)$ and of the corresponding cumulative probability distribution, $Q(x \leq \Lambda) = \int_0^{\Lambda} P(x)dx$. The average effect of mutations σ, irrespectively of whether they preserve or modify the native secondary structure is obtained from $\Pi_i(\Delta)$ as

$$\sigma_i = \int_0^{\infty} \Delta \Pi_i(\Delta) d\Delta. \tag{1}$$

2.4. *Linear motifs*

The complexity of a secondary structure [11] can be quantified by measuring the abundance of its structural motifs [15]. Here we introduce linear motifs, defined as any structural element that contains at least a hairpin loop (a loop of degree 1) closed by a stem and finishes either in a multi-loop (loop of degree 3 or higher) or in the $3'$ and $5'$ ends (dangling ends are included and counted as part of the linear motif). A linear motif might contain bulges and interior loops (both are loops of degree 2), but not branches (which, by definition, attach through a multi-loop and hence form a new linear motif). The size of a linear motif is defined as the total number of nucleotides it contains, including paired and nonpaired nucleotides. A more detailed definition of structural elements can be found in Ref. 40 or 15.

The structure corresponding to a sequence of length n will contain, in a general case, k_n linear motifs of sizes g_j, $j = 1, \ldots, k_n$, calculated as just defined. The average weighted length (AwL) of linear motifs in that structure is evaluated as

$$AwL = \frac{1}{n} \sum_{j=1}^{k} (g_j)^2. \tag{2}$$

The contribution of each motif is weighted through the fraction g_j/n: the more motifs, the smaller (typically) the latter quantity.

In order to get further insight on the structural meaning of linear motifs and AwL, let us estimate the latter quantity in some representative case-examples. Previous works have analytically calculated the asymptotic number of different structural motifs using a combinatorial approach where preferential pairing between nucleotides or energetic constraints are not considered [15]. Still, that approach is useful in that it retrieves the functional behavior of the expected number of structural elements (as a function of sequence length, for example) that are qualitatively reproduced in more realistic (numerical) studies [37].

The expected number k_n of linear motifs in a secondary structure of length n for a two-letter alphabet can be put in correspondence with the number $l_n(d)$ of interior loops of degree d:

$$k_n \sim \sum_{d>2} (d-1) l_n(d), \tag{3}$$

for n large enough. If we assume that all linear motifs are approximately of the same length,[a] the average size of linear motifs fulfills $AwL \sim n/k_n$. For the particular case of structures with stems of length one or larger and hairpin loops formed by at least $m = 3$ three unpaired nucleotides, the use of Eq. (3) predicts that

$$\lim_{n\to\infty} AwL \sim \frac{n}{k_n} = 4.159\ldots, \tag{4}$$

where we have used

$$l_n(d) = \frac{\alpha^2(1-\alpha)}{(1-2\alpha)^2(2+m-2m\alpha)} \left(\frac{1-2\alpha}{1-\alpha}\right)^d n, \tag{5}$$

as obtained in Ref. 15 to calculate first k_n, and finally substituted the value $\alpha = 0.4369$ that corresponds to the case $m = 3$ when stacks have length larger than or equal to one.

An improvement to the result above is obtained by considering the four-letter alphabet A, C, G, U, which yields a different value of $l_n(d)$, but still finite and independent from n, for $n \to \infty$ [15]. Numerical studies of random RNA sequences folded in their minimum free energy structure reveal that the distribution of loop degrees is essentially dominated by combinatorial principles, though stacks tend to be longer [37]. Hence, we should expect that the asymptotic value of AwL for the random sequences analysed in this work saturates at a finite value of AwL larger than the one above.

The previous result for random sequences differs from the functional behavior expected under different structural assumptions. If the number k of linear motifs would be a constant independent of the sequence length, for example, then $g_j = n/k$, $\forall\, j$, and $AwL = n/k$. In this case, AwL increases with n since the relative size of linear motifs also grows. The limit case is represented by rod-like structures without branches (as many viroids), for which $AwL = n$.

[a] This is an acceptable assumption in the light of numerical studies of random sequences which show that the distribution of stack sizes has a well defined, narrow maximum [8].

2.5. *Computational resources*

All simulations were performed in a Linux cluster using a Perl script available upon request. For statistical analyses we have used the packages in Mathematica 5.2.

3. Results

3.1. *Effect of mutations on RNA secondary structure*

Figure 1 shows four representative examples of the cumulative distribution of effects of mutations on the predicted RNA secondary structure for four sequences folding into a rod-like structure. The robustness of such conformations has been previously studied [29] and is confirmed here by the fact that most mutations affect the secondary structure in less than 20%. Figure 2 shows the location of the remaining mutations. In the case of Coconut cadang-cadang viroid, there is no mutation able to change the structure above 10%, the mutation with the highest effect being 224A → C (9.8% effect). The second example, Fig. 2(b), corresponds to Mexican papita viroid. Mutation 199G → A, U has a large effect of 85.6% in the secondary structure. This is an exceptional effect considering that 99% of the mutations modifying the secondary structure of this viroid change it in less than 20%. The case of Columnea latent viroid, shown in Fig. 2(c), yields additional information. It shows how mutations with a significant effect on fitness (above 20%) cluster in a well defined region of the folded molecule. A similar situation is observed with Avocado sunblotch viroid, shown in Fig. 2(d), where, in addition to the clustering of mutations, we observe a large number of positions changing the secondary

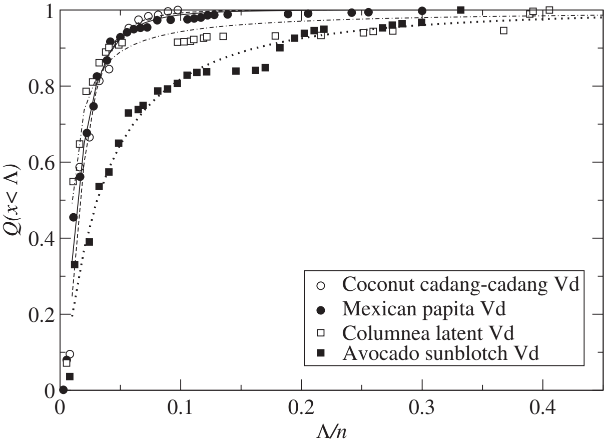
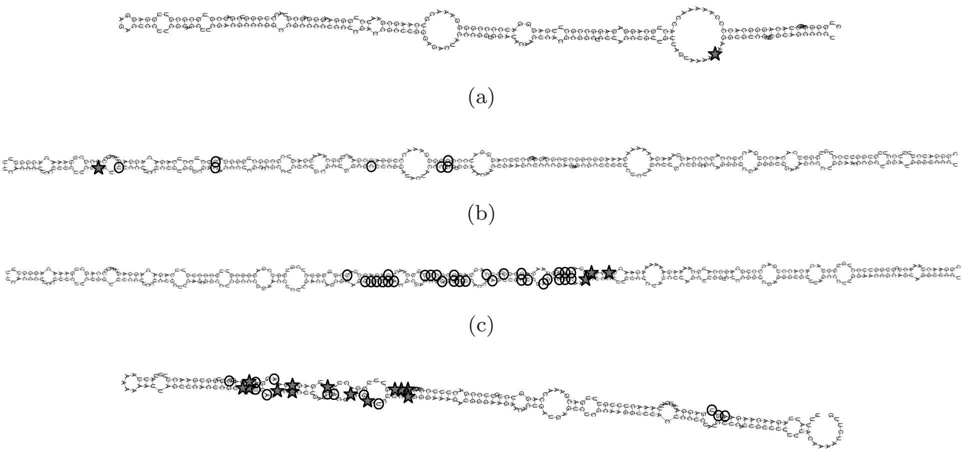


Fig. 1. Fits to the cumulative distribution of effects of mutations for rod-like structures. Four viroids folding into a rod-like secondary structure are shown. In each case, the function yielding the best fit is represented (see Table 1). Since $AwL = n$ in these cases, the functions are fit in the whole range of variation of Δ. Most mutations have mild effects on secondary structure, only two mutations in the Mexican papita viroid (not shown) affect the structure in more than 45%. See Fig. 2.



(a)

(b)

(c)

(d)

Fig. 2. Rod-like viroids and mutations with large effect on structure. Circles indicate mutations causing changes above 20% in the secondary structure. Stars show the position of the mutation(s) with the largest effect. (a) Coconut cadang-cadang viroid, the most disrupting mutation is 224A → C, which modifies the structure in a 9.8%. (b) Mexican papita viroid. Mutation 199G → A, U has a large effect of 85.6% in the secondary structure. (c) Columnea latent viroid. There are three mutations (69G → A, 71G → A, and 300C → G) causing a change of 40.5% in structure. (d) Avocado sunblotch viroid. Up to 17 different mutations in 12 different positions cause a reestructuring of 33.2%.

structure in exactly the same amount (33.2%). These observations indicate that secondary structures may respond in a modular fashion to mutations: There are specific regions where mutations have a fixed effect arising from disruption of a well-defined substructure within the molecule. The cumulative distributions of effects of mutations shown in Fig. 1 allow us to distinguish values of Δ/n (rescaled distance to the predicted minimum free energy structure) where the distribution jumps (as around 0.17 for Avocado sunblotch viroid), indicating the presence of a number of mutations causing similar structural rearrangements.

Figure 3 further clarifies the interplay between structure and the effect of mutations. We depict three branched structures and indicate those positions able to induce structural changes above 40% when mutated. In Fig. 3(a) we give as first example *Arabidopsis thaliana* Tyr tRNA GUA. Most mutations in the two outer short branches cause major disruptions of the secondary structure. In contrast, the longest stem holds only a few positions able to induce major changes. This implies as well that mutations along that element do not affect so easily other structural motifs. A more obvious case is that of Peach latent mosaic viroid [Fig. 3(b)]. None but four large-effect mutations are located along the longest stem, and they occur at the boundary of this motif (see also Fig. 2 in Ref. 29). This observation can be generalized to any linear motif, as the third example shown in Fig. 3(c) further supports. The fraction of point mutations causing large disturbances in the

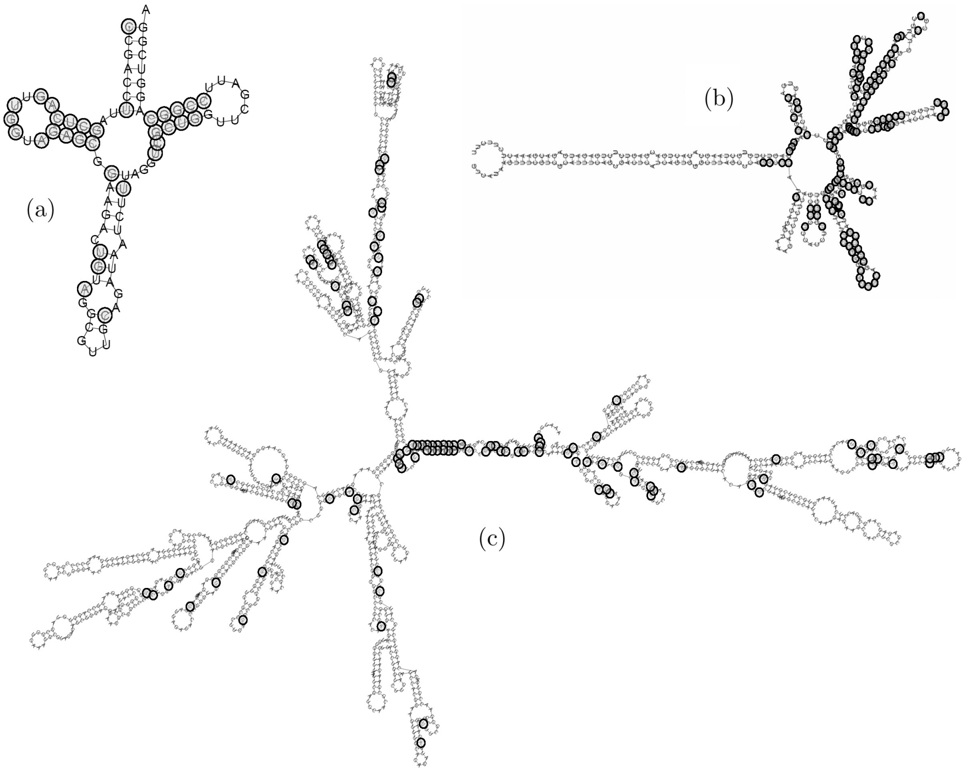


Fig. 3. Examples of branched structures and mutations with large effect. Positions highlighted through a circle correspond to mutations affecting the secondary structure in more than 40%. Mutations occurring in long peripheral motifs are less able of affecting the structure. This effect becomes clearer the larger the RNA sequence. (a) tRNA of Tyrosine of *Arabidopsis thaliana*; (b) Peach latent mosaic viroid (Avsunviroidae family); (c) 16S unit of *E. coli*.

secondary structure decreases as the length n increases. They also tend to occur close to potentially weak positions, near internal loops, branching points or regions with a large fraction of unpaired nucleotides, which might more easily participate in large structural rearrangements.

3.2. *Functional form of the distribution of effects of mutations on linear motifs*

In a previous computational work with RNA sequences [34], it was observed that none of the probability distribution functions commonly used to fit the effect of mutations in fitness (compiled in Table 2) could account for the analyzed data in the whole range of Δ values. Actually, jumps in $\Pi(\Delta)$ are frequent and do not disappear by using larger sequences. Instead, they seem to characterize the structure under study by identifying those positions which, when mutated, disrupt the structure in a precise fashion, as shown in the previous section for several case-examples. For mutations causing large rearrangements, there seems to be a cascade effect where more than one simple (linear) motif is affected, thus opening pairs in the native structure and forming new pairs at previously distant positions. The situation is thus complex and, for this reason, we have decided to study the quantitative properties of the distribution of effects in the secondary structure only when the effect of mutations is below the typical size of linear motifs in the structure. With this prescription we mostly exclude mutations causing large structural rearrangements, which are of a different nature and largely depend on the structure considered. We will use the average weighted length AwL to set the maximal value of the change in the secondary structure to fit the data. In the case of rod-like viroids, functions are fit in the whole range while, as the number of branches in the structure increases, the range shrinks. In the particular case of random sequences, $\Lambda/n \rightarrow 0$ as n grows: Since linear motifs have a typical size that does not increase with sequence length (see Sec. 2.4), the range where the distribution of effects of mutations is fitted shrinks with n.

Figure 4 gives three examples of cumulative distributions $Q(x < \Lambda)$ for sequences of different lengths (shown in Fig. 3). As an illustration of how different functions fit the distribution up to AwL/n, we show for each case the five functions assayed, as specified in the legend and in Table 2. For the sake of comparison, the x-axes have been rescaled by the length of the sequence. Amongst the functions assayed, note that the exponential distribution performs worst.

The analysis above has been carried out with all 46 sequences in Table 1. In all cases, we observe a large number of mutations with small effect. However, the decay of the distribution is not fast, as revealed by the fact that exponentially decaying functions fit data typically worse than probability distributions with fat tails. In most cases data are best fit by a Pareto or by a Lognormal distribution. The common feature of these two distributions is their large weight at small values of the variable and the slow, sub-exponential decay at large values of the variable (up to the typical length of linear motifs). For some of the smallest sequences (tRNAs), however, the Exponential and other distributions with an exponential tail, such as the Gamma or Beta, can also provide the best fit. This nonetheless, it is worth to observe that there are fewer, and statistically less reliable, points for the fit in tRNAs, such that the distributions are more subjected to intrinsic noise due to the short length of those sequences. The analysis has been completed by performing a Kolmogorov–Smirnov two-sample test. The accumulated probabilities predicted by the best-fitting model were not significantly different from the observed values, except for the Coleus blumei 1 viroid ($D = 0.642$, and $P < 0.01$, for 12 data points).

3.3. *Structural motifs and the effect of mutations*

The study of a number of case examples has shown some facts about the relationship between structural motifs in the secondary structure and the effect of point mutations. As one may intuitively expect, longer sequences tend to show smaller

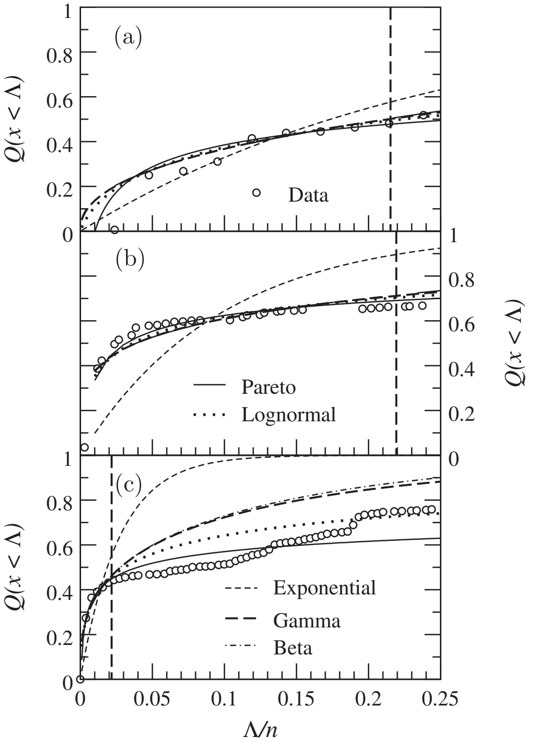


Fig. 4. Examples of probability distributions of effects of mutations. The corresponding RNA structures are shown in Fig. 3. Five different functions have been fitted up to the value $\Lambda = AwL$, which characterizes the weighted length spanned by a typical linear motif. That value is signaled by a vertical, long-dashed line in each plot. (a) tRNA of Tyrosine of *A. thaliana* has size $n = 84$ and $AwL = 18.3$; (b) Peach latent mosaic viroid has length $n = 337$ and $AwL = 73.9$; (c) 16S unit of *E. coli*, with $n = 1542$ and $AwL = 33.5$.

mutational effects (Spearman's $\rho = -0.538$, $P = 0.003$). However, this association is mainly dependent on tRNAs, which are both the smallest sequences considered and those with the greatest mutational effects. If we remove this group, the above correlation is essentially lost ($\rho = 0.118$, $P = 0.494$). The native structure free energy ΔG correlates strongly with the sequence length n: Spearman's $\rho = 0.948$ for all sequences and $\rho = 0.910$ without tRNAs, $P < 0.001$ in both cases. Hence, ΔG behaves similarly to n in relation to the average effect of mutations on secondary structure σ. We obtain $\rho = 0.612$, $P < 0.001$ for all sequences and $\rho = -0.260$, $P = 0.126$ if tRNAs are removed. Considering sequence length (or ΔG) alone also fails to explain some clear patterns, such as the greater mutational effects among avsunviroids (0.097 ± 0.035) compared with pospiviroids (0.051 ± 0.010), despite their similar average sequence length (330 and 327 nucleotides, respectively) or the over twenty-fold difference between the average mutational effect in HDV and a random sequence of similar length (Table 1).

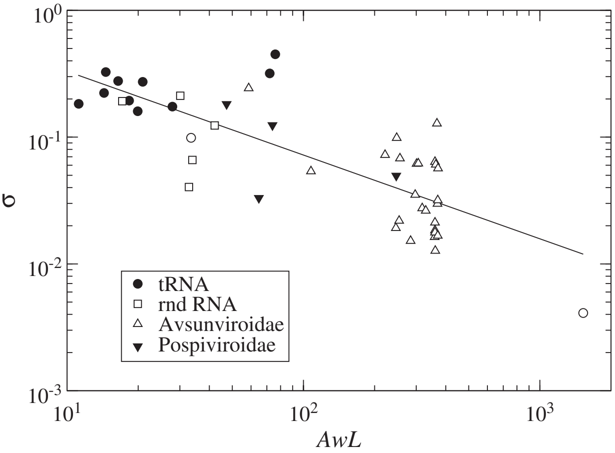


Fig. 5. Dependence of the response to mutations on structural properties. We represent the relationship between the average effect of mutations σ in the secondary structure of each of the 46 RNA sequences studied as a function of the average length of weighted linear (AwL) motifs. Different groups are as described in the legend; the two open circles correspond to *E. coli* 16S (left) and to Hepatitis-δ virus (right). The dependence of all data is well fit by a function of the form $y = ax^{-b}$, with $a = 1.5 \pm 1.5$ and $b = 0.66 \pm 0.08$.

We then used the average weighted length of linear motifs AwL to evaluate the relationship between the size of linear substructures within a secondary structure and its mutational robustness. Notice how random sequences cluster around $AwL \simeq 30$, in qualitative and quantitative agreement with the analytical expectations in Sec. 2.4. On the other hand, the AwL coincides with sequence length for rod-like structures and is close to that length if a rod-like structure is interrupted by another, short stem (e.g., in the Citrus III viroid, Table 1). The lower the AwL values, the more evenly branched the structure is (random sequences and tRNAs being good examples). As shown in Fig. 5, AwL presents a negative correlation with the average effect of mutations ($\rho = -0.738$, $P < 0.001$) which is not lost after removal of the tRNA group ($\rho = -0.589$, $P < 0.001$). The partial correlation between AwL and σ taking the sequence length n into account yields $\rho = -0.493$, $P = 0.001$, and when the tRNA group is removed $\rho = -0.479$, $P = 0.004$. Furthermore, AwL allows us to account for the greater mutational robustness of the pospiviroid rod-like structures compared to those of avsunviroids or the remarkably small effects shown by HDV, whose structure is also rod-like. Therefore, the robustness of a secondary structure depends on its degree of linearity in addition to its size.

4. Discussion and Conclusion

Statistical analysis of the effects of all single nucleotide substitutions on the predicted secondary structure of a wide variety of RNAs has allowed us to reach several conclusions which appear to be widely general. First, mutations of small effect are more common than those of large effect, a property that is shared by essentially all biological systems examined to date [6, 28]. This feature represents another quantification of the local correlations observed in RNA sequence space when mapped to the secondary structure [42, 36]. Second, exponential-tail distributions such as the Exponential, the Gamma, or the Beta provide a less satisfactory fit than fat-tail distributions such as the Lognormal or the Pareto, again similar to what has been empirically observed in a variety of natural systems [6, 28]. This second observation is in agreement with the remarkable degree of accessibility of significantly different RNA secondary structures within a few point mutations from almost any randomly chosen sequence [25, 13]. Third, mutations occurring in neighboring positions usually have similar effects on the secondary structure, because they affect the same set of motifs and probably cause similar rearrangements. In the most extreme situation, there is a set of mutations with exactly the same effect, as was the case of Avocado sunblotch viroid. Such modular behavior of RNA structures could at least partially account for the fact that a certain fraction of mutations has unusually large effects and thereby explain the good fit obtained by fat-tail distributions. We can hypothesize that, analogously, a modular organization of metabolic or regulatory pathways [24, 1] might explain why mutational fitness effects are well described by this family of distributions in many organisms. Fourth, the overall robustness of a given RNA secondary structure depends on its shape rather than just on its size: Rod-like structures with long linear motifs are more resilient to structural changes than branched structures. Possibly, mutations causing small effects in an RNA secondary structure limit their action to the linear motif where they are located, whereas propagation of the effect of a mutation beyond the substructure where it is found might cause a cascade of rearrangements. This further speaks for an uneven distribution of secondary structures in sequence space. Despite an extremely high interwoveness of RNA secondary structure neutral networks [13], our results support that structures with a higher-than-average amount of linear motifs tend to cluster in the space of sequences. This reveals an aspect of robustness additional to neutrality — usually defined as the maintenance of the secondary structure under the action of mutations, a quantity which is maximized in highly connected regions of the neutral network for every given structure [16]. In cases where not only the sequence, but also the secondary structure can be modified to a certain extent, a selection pressure to increase the number and length of linear motifs, which in turn increase structural robustness, might appear. Eventually, the RNA secondary structure of functional molecules should result from a combination of different mechanisms, including how easily it is found (how large the corresponding neutral network is [4] and how long it takes to find and fix a given structure [35]), and what selection pressures act to keep functionality while minimizing the effect of mutations on structure, among others.

An open question is whether the distribution of effects of mutations has a well-defined functional form beyond AwL. Our study points out that this is not the case for individual sequences, though a universal distribution might exist in a statistical sense. Future studies should address this point by averaging over the effects caused by point mutations in independent sequences with the same value of AwL first, and subsequently over sets of sequences folding into secondary structures of arbitrary shape and length. Preliminary results indicate that, with independence of the universality of the distribution of effects of mutations in RNA sequences, that distribution should have a fat tail as well, since changes of size $\Delta \simeq n$ are common for random sequences.

Here, we have focused on the predicted minimum free energy secondary structure of each sequence for simplicity and computational tractability, but RNA can also fold into usually transient thermodynamically suboptimal structures. It is easily conceivable that a given biological function associated with a sequence remains active even if the structure is partially altered. Furthermore, a given sequence might fold into alternative structures, each with a different function, as has been shown for ribozymes [31]. As a case in point, the active structure of hammerhead ribozymes does not coincide with the predicted minimum free energy structure [20]. If the folding is rod-like, alternative structures are highly unlikely, whereas these can be much more frequent in highly branched structures. Whereas our interpretation is that the former are structurally more robust, we cannot discard that certain biological functions can be performed by alternative RNA structures of the same sequence, or even that structural plasticity is required for function. This type of ambiguity appears when one tries to extrapolate the predicted RNA structures to a given function or even to overall fitness. As such, our results should be interpreted solely in the context of *in silico* predicted structures, regardless of whether these structures match the ones found *in vivo* or whether they can be associated to specific biological functions.

Despite the above limitation, the role of structural robustness in RNA sequence evolution is supported by previous studies suggesting that viroids with branched secondary structures (Avsunviroids) occupy a basal phylogenetic position compared to those folding into more rod-like structures (Pospiviroids) [5, 29]. The observation that some viroids show extremely high mutation rates and that most nucleotide substitutions are significantly deleterious [10] increases the efficiency with which selection can favor the evolution of neutrality and structural robustness. However, it is obvious that selection can also operate independently of RNA secondary structure even in noncoding RNA. Indeed, a type of nonstructural selection is needed to explain why the fitness effects of mutagenesis appear to be more severe in Pospiviroids than in Avsunviroids [2]. Again, factors other than thermodynamic stability alone may shape a given RNA structure. For instance, double-stranded RNA triggers post-transcriptional gene silencing and other responses leading to the degradation or underexpression of the target RNA [39, 17]. Therefore, some sequences might have evolved nonrod-like secondary structures to minimize RNA pairing and thus avoid these responses. The extent to which selection for neutrality and selection for structural robustness is effective in face of other molecular functions and the analysis of their interdependence is left for future studies.

Acknowledgments

Support of Spanish Ministerio de Ciencia e Innovación through research projects FIS2008-05273 and BFU2008-03978/BMC, through the Ramón y Cajal program, and of Comunidad de Madrid (project MODELICO, S2009/ESP-1691) is gratefully acknowledged.

References

[1] Barabási, A. L. and Albert, R., Emergence of scaling in random networks, *Science* **286** (1999) 509–512.

[2] Codoñer, F. M., Darós, J. A., Solé, R. V. and Elena, S. F., The fittest versus the flattest: Experimental confirmation of the quasispecies effect with subviral pathogens, *PLoS Path.* **2** (2006) e136.

[3] Cowperthwaite, M. C., Bull, J. J. and Ancel Meyers, L., From bad to good: Fitness reversals and the ascent of deleterious mutations, *PLoS Comp. Biol.* **2** (2006) e141.

[4] Cowperthwaite, M. C., Economo, E. P., Harcombe, W. R., Miller, E. L. and Ancel Meyers, L., The ascent of the abundant: How mutational networks constrain evolution, *PLoS Comp. Biol.* **4** (2008) e1000110.

[5] Elena, S. F., Dopazo, J., Flores, R., Diener, T. O. and Moya, A., Phylogeny of viroids, viroidlike satellite RNAs, and the viroidlike domain of hepatitis δ virus RNA, *Proc. Natl. Acad. Sci. USA* **88** (1991) 5631–5634.

[6] Eyre-Walker, A. and Keightley, P. D., The distribution of fitness effects of new mutations, *Nat. Rev. Genet.* **8** (2007) 610–618.

[7] Fontana, W., Modelling 'evo-devo' with RNA, *BioEssays* **24** (2002) 1164–1177.

[8] Fontana, W., Konings, D. A. M., Stadler, P. F. and Schuster, P., Statistics of RNA secondary structures, *Biopolymers* **33** (1993) 1389–1404.

[9] Fontana, W. and Schuster, P., Continuity in evolution: On the nature of transitions, *Science* **280** (1998) 1451–1455.

[10] Gago, S., Elena, S. F., Flores, R. and Sanjuán, R., Extremely high mutation rate of a hammerhead viroid, *Science* **323** (2009) 1308.

[11] Giegerich, R., Voß, B. and Rehmsmeier, M., Abstract shapes of RNA, *Nucl. Acids Res.* **32** (2004) 4843–4851.

[12] Gillespie, J., A simple stochastic gene substitution model, *Theor. Pop. Biol.* **23** (1983) 202–215.

[13] Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F. and Schuster, P., Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering, *Monatsh. Chem.* **127** (1996) 375–389.

[14] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. and Schuster, P., Fast folding and comparison of RNA secondary structures, *Monatsh. Chem.* **125** (1994) 167–188.

[15] Hofacker, I. L., Schuster, P. and Stadler, P. F., Combinatorics of RNA secondary structures, *Discrete Appl. Math.* **88** (1998) 207–237.

[16] Huynen, M. A., Stadler, P. F. and Fontana, W., Smoothness within ruggedness: The role of neutrality in adaptation, *Proc. Natl. Acad. Sci. USA* **93** (1996) 397–401.

[17] Kaempfer, R., RNA sensors: Novel regulators of gene expression, *EMBO Rep.* **4** (2003) 1043–1047.

[18] Kimura, M. and Maruyama, T., The mutational load with epistatic interactions in fitness, *Genetics* **54** (1966) 1337–1351.

[19] MacLean, R. C. and Buckling, A., The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*, *PLoS Genetics* **5** (2009) e1000406.

[20] Martick, M. and Scott, W. G., Tertiary contacts distant from the active site prime a ribozyme for catalysis, *Cell* **126** (2006) 309–320.

[21] Martin, G. and Lenormand, T., The fitness effect of mutations across environments: A survey in the light of fitness landscape models, *Evolution* **60** (2006) 2413–2427.

[22] Orr, H. A., The distribution of fitness effects among beneficial mutations, *Genetics* **163** (2003) 1519–1526.

[23] Peris, J. B., Davis, P., Cuevas, J. M., Nebot, M. R. and Sanjuán, R., Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage f1, *Genetics* **185** (2010) 603.

[24] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A. L., Hierarchical organization of modularity in metabolic networks, *Science* **297** (2002) 1551–1555.

[25] Reidys, C., Stadler, P. F. and Schuster, P., Generic properties of combinatory maps: Neutral networks of RNA secondary structures, *Bull. Math. Biol.* **59** (1997) 339–397.

[26] Rokyta, D. R., Beisel, C. J., Joyce, P., Ferris, M. T., Burch, C. L. and Wichman, H. A., Beneficial fitness effects are not exponential for two viruses, *J. Mol. Evol.* **67** (2008) 368–376.

[27] Rokyta, D. R., Joyce, P., Caudle, S. B. and Wichman, H. A., An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus, *Nat. Gen.* **37** (2005) 441–444.

[28] Sanjuán, R., Mutational fitness effects in RNA and single-stranded DNA viruses: Common patterns revealed by site-directed mutagenesis studies, *Phil. Trans. R. Soc. B* **365** (2010) 1975–1982.

[29] Sanjuán, R., Forment, J. and Elena, S. F., In silico predicted robustness of viroids RNA secondary structures. I. The effect of single mutations, *Mol. Biol. Evol.* **23** (2006) 1427–1436.

[30] Sanjuán, R., Moya, A. and Elena, S. F., The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus, *Proc. Natl. Acad. Sci. USA* **101** (2004) 8396–8401.

[31] Schultes, E. A. and Bartel, D. P., One sequence, two ribozymes: Implications for the emergence of new ribozyme folds, *Science* **289** (2000) 448–452.

[32] Schuster, P., Fontana, W., Stadler, P. F. and Hofacker, I. L., From sequences to shapes and back: A case study in RNA secondary structures, *Proc. R. Soc. London B* **255** (1994) 279–284.

[33] Silander, O. K., Tenaillon, O. and Chao, L., Understanding the evolutionary fate of finite populations: The dynamics of mutational effects, *PLoS Biol.* **5** (2007) 922–931.

[34] Stich, M., Lázaro, E. and Manrubia, S. C., Phenotypic effect of mutations in evolving populations of RNA molecules, *BMC Evol. Biol.* **10** (2010) 46.

[35] Stich, M. and Manrubia, S. C., Motif frequency and evolutionary search times in RNA populations, *J. Theor. Biol.* **280** (2011) 117–126.

[36] Sumedha, Martin, O. C. and Wagner, A., New structural variation in evolutionary searches of RNA neutral networks, *BioSystems* **90** (2007) 475–485.

[37] Tacker, M., Stadler, P. F., Bornberg-Bauer, E. G., Hofacker, I. L. and Schuster, P., Algorithm independent properties of RNA structure predictions, *Eur. Biophys. J.* **25** (1996) 115–130.

[38] van Nimwegen, E., Crutchfield, J. P. and Huynen, M., Neutral evolution of mutational robustness, *Proc. Natl. Acad. Sci. USA* **96** (1999) 9716–9720.

[39] Voinnet, O., Origin, biogenesis, and activity of plant microRNAs, *Cell* **136** (2009) 669–687.

[40] Waterman, M. S., Secondary structure of single-stranded nucleic acids, *Studies in Foundations and Combinatorics. Advances in Mathematics Supplementary Studies* **1** (1978) 167–212.

[41] Wilke, C. O. and Adami, C., Interaction between directional epistasis and average mutational effects, *Proc. Roy. Soc. B* **268** (2001) 1469–1474.

[42] Wilke, C. O., Lenski, R. E. and C., A., Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding, *BMC Evol. Biol.* **3** (2003) 3.