Template switching in DNA replication can create and maintain RNA hairpins

Heli A. M. Mönttinen and Ari Löytynoja

Institute of Biotechnology, HiLIFE, P.O.Box 56, 00014 University of Helsinki, Finland

The evolutionary origin of ribonucleic acid (RNA) stem structures (1, 2) and the preservation of their base-pairing under a spontaneous and random mutation process have puzzled theoretical evolutionary biologists (3, 4). DNA replication-related template switching (5, 6) is a mutation mechanism that creates reverse-complement copies of sequence regions within a genome by replicating briefly either along the complementary or nascent DNA strand. Depending on the relative positions and context of the four switch points, this process may produce a reverse-complement repeat capable of forming the stem of a perfect DNA hairpin, or fix the base-pairing of an existing stem (7). Template switching is typically thought to trigger large structural changes (8–10) and its possible role in the origin and evolution of RNA genes has not been studied. Here we show that the reconstructed ancestral history of ribosomal RNA sequences contains compensatory base substitutions that are linked with parallel sequence changes consistent with the DNA replication-related template switching. In addition to compensatory mutations, the mechanism can explain complex changes involving non-Watson-Crick pairing and appearances of novel stem structures, though mutations breaking the structure rarely get fixed in evolution. Our results suggest a solution for the longstanding dilemma of RNA gene evolution (1, 3, 4) and demonstrate how template switching can both create perfect stem structures with a single mutation event and maintain their base pairing over time with matching changes. The mechanism can also generate parallel sequence changes, many inexplicable under the point mutation model (11), and provides an explanation for the asymmetric base-pair frequencies in stem structures (12).

Introduction

The characteristic feature of RNAs are the helical structures formed by intra-molecular pairings of two to ten consecutive complementary bases (12). The basis of helices are the hydrogen bonds between classical Watson-Crick pairs and less stable G-U pairs, and their structures are further stabilized by the stacking interactions between successive base pairs. While the underlying sequences may evolve, the locations of RNA helices are highly conserved among related sequences. Computational methods for inference and validation of RNA secondary structures exploit this (13, 14), and correlated base changes turning one Watson-Crick pair to another, e.g., A-U→G-C, are considered the ultimate evidence of a functionally important stem. Such compensatory mutations (CMs) are commonly believed to happen via an intermediate state involving G-U base pairing (14). However, the rate of simultaneous substitution of both members of a base pair has been shown to be greater than zero (3, 12), suggesting that either the intermediate state with non-Watson-Crick pairing is very short or RNA stem sequences can also evolve through double substitutions. Apparent simultaneous double substitutions have been explained with population models where the initial mutations are kept at low frequency by selection and only increase in frequency after another change has restored the base pairing (12). This still requires a very high rate of matching mutations within the population.

We showed earlier that DNA replication-related template switch mutations (TSMs) can produce reverse-complement repeats needed for perfect DNA hairpins and fix the base-pairing of existing stems (7). The mechanism has been most thoroughly studied in microbes (5, 6, 15, 16) and been aware of in eukaryotic research (17), but few studies have looked at the role of the resulting mutations in genes (though see (18)). Here we set out to investigate the role of TSMs in the evolution of RNA sequences by studying CMs in the RNA hairpin stem structures. We used inferred ancestral sequences to reconstruct the mutation history for sets of closely related RNA sequences and located individual base changes into specific tree branches (Fig. 1). We focused on CMs consistent with the "two-step process" (12) where the intermediate state is observed, i.e. cases in which the initial mutation breaking the base pairing is placed in one tree branch and the compensatory change in another branch (Fig. 1b). In these, template switching could either trigger a matching change in the unmutated stem, or restore the original base pairing by a copying of the unmutated sequence. We hypothesized that if TSMs are involved in the CM process, they should occasionally leave patterns where the CM is associated with parallel changes in its immediate proximity (Fig. 2). Within RNA sequences, such associated changes would appear as multiple compensatory changes in the stem, matching mutations within the loop (Fig. 2d), or as fully inverted loops (Fig. 2f).

We tested these hypotheses with ribosomal RNA (rRNA) sequences that evolve at a higher rate than many smaller structural RNA types (12) and for which large quantities of data are available due to their use as phylogenetic markers (19). Ribosomal DNA (rDNA) genes usually appear as multiple copies in the genome and evolve in a concerted fashion (20). While this complicates some analyses, the high copy number also elevates the overall mutation rate and variants temporarily segregating within gene clusters may be detectable with DNA sequencing. We found this to be case and identified CMs consistent with the TSM mechanism both among recent mutations, likely destined for removal by drift and selection, and among historical changes separating established evolutionary lineages. While our analyses demonstrate a dual role for the TSM mechanism in the evolution of novel RNA structures and in the maintenance of existing ones, unexpectedly, they also suggest a solution for two other dilemmas in the RNA evolution, jumps in the sequence space (11) and the asymmetry of the base-pair substitution process (12).

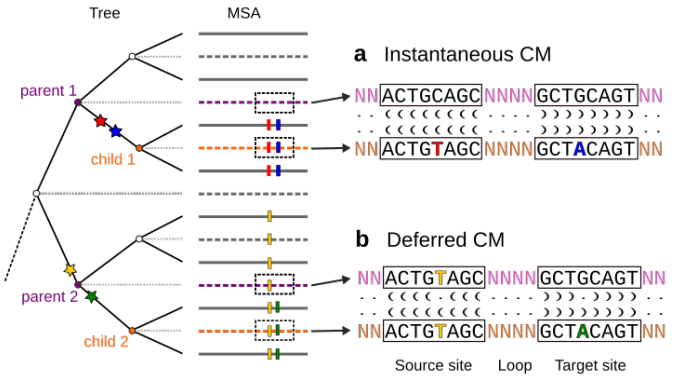

Fig. 1. Identification of TSM-explainable compensatory substitutions. By inferring ancestral sequences (dashed lines) for the internal nodes of a phylogenetic tree, one can locate mutations (stars) to specific tree branches. **a** If the initial and restoring mutation happen in the same branch (red and blue), they appear as an instantaneous CM and the perfect stem base pairing is retained. **b** When the evolutionary lineage splits before the CM takes place, the the initial and restoring mutation appear in different branches (yellow and green) and the mechanism triggering the CM can be studied. We examined whether the TSM mechanism could explain the CM with (partial) reverse-complement copying of the source stem to the target stem. Dot-bracket notations show predicted secondary structures. Opening and closing brackets indicate pairing bases in the stem, unpaired bases are marked with dots.

Results

Reconstruction of ancestral sequence histories. We downloaded 614,502 large and small subunit rRNA sequences from Rfam database (v. 14.1) (21) and Silva rRNA database (v. 138) (19), and clustered them by minimum sequence identity of 90% and 75% pairwise coverage (Fig. S1; see Methods). To remove highly similar sequences while retaining local dissimilarities, we used a sliding window approach and kept sequences showing at least two differences within a 20-bp window. This resulted in 5,525 clusters with at least 10 sequences, containing in total 289,024 sequences. After aligning the sequences, we inferred a maximum likelihood tree for each cluster and reconstructed ancestral sequences for the internal nodes (Fig. 1).

Using the reconstructed sequence history, we analyzed each *parent node* - *child node* pair in the tree and identified mutation clusters between the corresponding sequences that were explainable by a TSM. We considered secondary structures for the sequences involved as well as for their immediate parent and descendant sequences (see Methods), and extracted the cases in which a CM restores the mismatch in the parent sequence to perfect base-pairing in the child sequence (Fig. 1b). Due to the short length of rRNA stems, a statistic proposed for genomic TSM events (22) could not be applied for the evaluation of inferred events. Instead, we required a minimum length and number of sequence changes explainable by the TSM (Table 1) and applied strict criteria on sequence and structure conservation (see Methods). We confirmed the quality of structure predictions by comparing them against a set of solved structures in Protein Data Bank. We found the locations of hairpins to be correctly inferred while the loop boundaries and the locations of bulges and internal loops contained minor error (Fig. S2), largely explained by the structure prediction method only considering the noncanonical pair G-U.

Footprints of compensatory substitutions. In total 13,387 cases of two-step CM passed our quality criteria. Of these, 96.0% were isolated substitutions and uninformative about the underlying mutation process (Table 1). However, in 531 cases the CM was associated with one or more sequence differences within proximity of 10 bp (Fig. 3a-b, Fig. S3a-b). In 458 cases, a single TSM could explain multiple CMs in the stem, and in 73 cases, it could explain the CM and one or more changes in the loop sequence. Most frequently CMs restored G-U pairs in stems, and 58 % of the multi-nucleotide changes explainable by a single TSM did not lead to structural changes.

In our phylogenetic approach, novel mutations appear as terminal tree branches that split out as sister taxa for the consensus sequence of the main form. Slightly older mutations

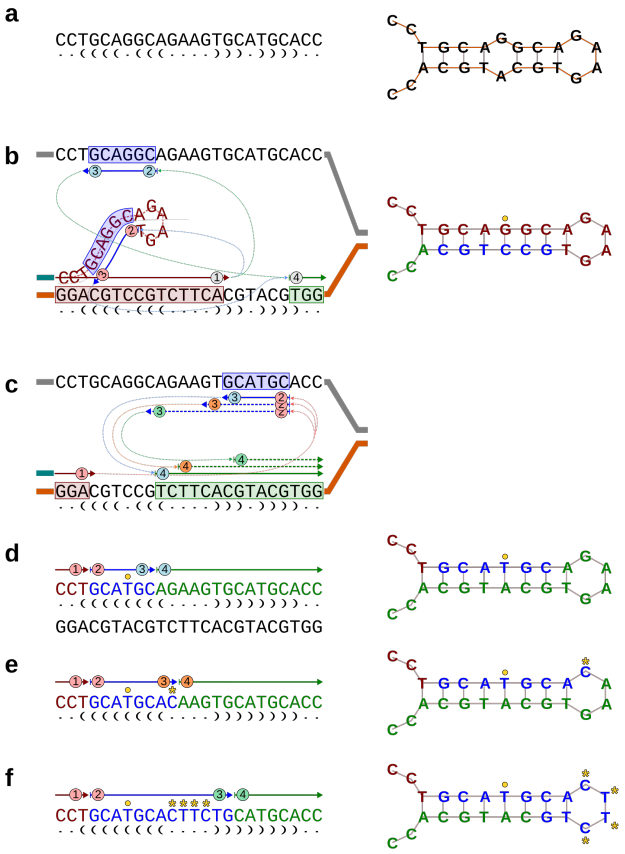

Fig. 2. CMs via template switching can leave footprints in sequence. **a** A hypothetical sequence forms a stem structure with a G-A mismatch. **b** The stem can be corrected by DNA replication (solid arrows; first red) briefly switching to another template (blue arrow) and then returning to the original strand (green arrow). The outcome of the process depends on the relative positions of the four switch points, ① - ④. In intra-strand switch (blue dotted arrows), the newly synthesized sequence (in red) forms a hairpin and templates the replication; in inter-strand switch (green dotted arrows), the complementary strand templates the replication. A backwards jump (② < ①) corrects the mismatch with a G-C pair (right) but cannot cause associated mutations within the loop. **c** Inter-strand switch forwards (② > ①) corrects the mismatch with a T-A pair and, depending on positions of the switch points, may cause associated mutations. **d** When the source sequence (in blue) is within the stem region, only the mismatch is fixed (yellow circle). **e** Source sequence overlapping the loop region causes a parallel change in the loop sequence (yellow asterisk). **f** Source sequence fully covering the loop region inverts the loop sequence and causes multiple parallel changes.

Table 1. Different mutation patterns explainable by a single TSM.

Pattern type	Tree position		Filtering criteria	
	Terminal	Internal	length	diff.
One CM only	8,112	4,744	6+	1
CM + parallel mut. in stem	314	144	6+	2+
CM + parallel mut. in loop	38	24	6+	2+
CM + loop inversion	6	5	6+	2+
Loop inversion only	89	50	6+	1-2
New hairpin loop	70	20	8+	2+
Asymmetric between stems	985	484	6+	2+
Within a stem half	642	291	6+	2+
From loop to stem	173	66	6+	2+

make internal tree branches with a few descendants, while fixed variants are deep in the tree and define subtrees with descendants inheriting the derived form (Fig. 3). Due to uncertainties in the ancestral sequence reconstruction, deep events could be studied only in the most conserved regions. The reconstruction was more robust for sub-terminal nodes, and we recorded the inferred TSMs on the first three levels of phylogenetic trees and normalized the counts with the total number of branches at the same tree level. When the associated changes were in the loop sequence (cf. Fig. 2e), the TSM-explainable CMs were relatively evenly distributed between the terminal and internal nodes (Fig. 4). In contrast, events associated with parallel mutations in the stem were twice as frequent in the terminal than in sub-terminal (-1, -2) branches, consistent with influence of negative selection. One should note though that an event being placed in a sub-terminal tree branch only indicates that variants of the derived form have been independently found in multiple rDNA sequencing experiments. While this suggests that the associated mutation is not lethal, different rDNA gene variants may segregate in a gene cluster and an internal position does not yet prove a mutation passing evolutionary selection.

For comparison, we searched for mutation patterns suggesting an instantaneous CM (Fig. 1a). For that, we identified loop stems where the parent and child nodes had nonidentical stem sequence but both showed perfect Watson-Crick base-pairing (see Methods). We found 126,929 such stems, 57,026 in internal and 69,903 in terminal branches, but the total number of CMs was higher as 16.9% of the stems contained more than one CM (Fig. S4). The numbers of instantaneous CMs and patterns consistent with the two-step CM process obtained using our ancestral sequence reconstruction approach are in line with earlier studies, although the ratio of more than ten to one is higher than expected from the evolutionary models (3, 12). We did not find the affected stems to differ from typical rRNA stems and 81.2% of the CMs were in 3-7 bp-long perfectly base-pairing regions, the average length being 4.5 bp (Fig. S4). While 81.8% of the identified two-step CMs affected G-U base pairs and restored Watson-Crick pairing, 63.4% of instantaneous CMs were of type of A-T↔G-C or C-G↔T-A and could thus occur via a G-U intermediate (Fig. S5). The fact that a greater proportion of two-step CMs involve G-U pairs may reflect their more benign impact on the stem structure, enabling higher frequency in the populations and increased chances of being observed (23, 24).

Selection maintains RNA secondary structures. When searching for CMs consistent with the TSM mechanism, we identified TSM-like mutation patterns that did not originate symmetrically from the opposite stem but were nevertheless predicted to retain the secondary structure. These cases were much more common than the pure CMs causing multi-base changes, totalling 2,641, and could be classified into three classes (Table 1): asymmetric copying of sequence from one stem side to the other; copying of sequence within one side of a stem (Fig. 3d); copying of loop sequence to a stem (Fig. 3e). Half of the inferred asymmetric (48.4%) TSMs maintained the stem structure, indicating that multi-base changes through TSM events can both amend the stem sequence without a structural change and introduce new internal loops and bulges. Interestingly, we found asymmetric TSM patterns compatible with two independent TSM events and could infer the possible mutation history and the intermediate state of the CM process (Fig. 5). Patterns, where one TSM first breaks the stem and another fixes it, can appear as multiple instantaneous CMs between the parent and the child node, possibly partially explaining the high number (21,506) of such cases.

In addition to TSMs maintaining existing RNA secondary structures, template switching could create novel RNA hairpins (7). To identify those, we searched for TSM patterns

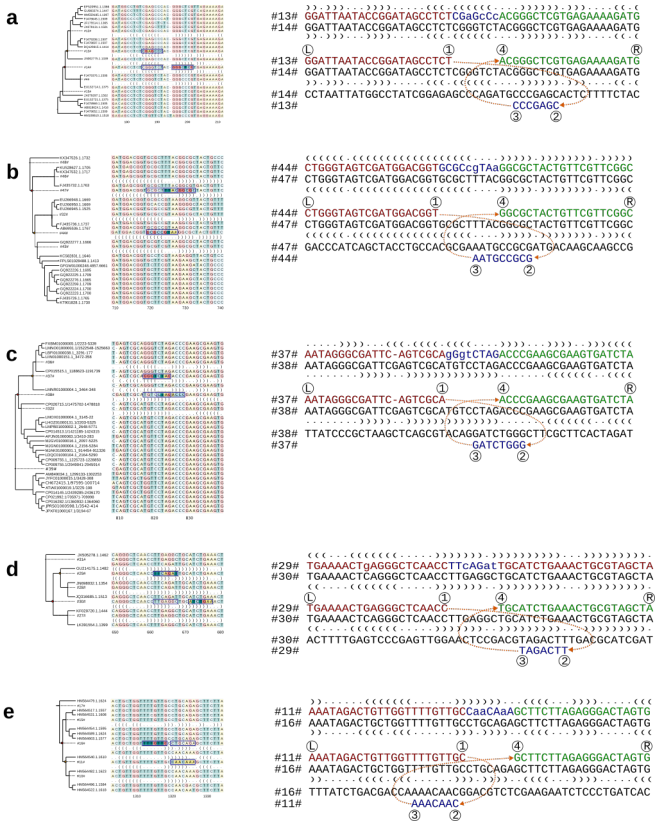

Fig. 3. TSM-like patterns with different effects on sequence and secondary structure. Alignment shows inferred sequences for selected ancestral nodes and predicted secondary structures for the affected nodes and their immediate child nodes. The source and target regions for the inferred TSM event are highlighted and the homologous region in the parent is framed. Sequences on the right show the reference (parent; below) and the query (child; on top) sequences and the inferred TSM processes explaining the differences. The sequence differences are shown in lowercase and the secondary structure is given with the dot-bracket notation. **a** TSM causing a CM with a parallel sequence change in the loop. **b** TSM causing a CM with a parallel inversion of the loop sequence. **c** TSM triggering a formation of a new loop. **d** TSM within one side of a stem. **e** TSM copying loop sequence to a stem.

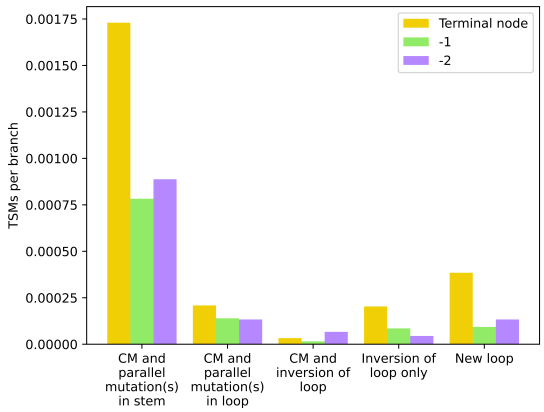

Fig. 4. Frequencies of CM-associated TSM patterns per branch. The frequency of events in terminal and sub-terminal (-1, -2) branches was calculated by dividing the number of inferred TSM events by the total number of branches on that tree level. Enrichment of TSM-like patterns causing structural changes in terminal branches suggests that many mutations are detrimental and eliminated by selection.

overlapping inconsistent stem predictions for the parent and child sequences. We identified 90 strong TSM candidates creating novel RNA hairpins while retaining high sequence similarity (Table 1, Fig. 3c; see Methods). TSM-like mutation patterns creating novel loops were typically found in regions showing pre-existing similarity between the target and source sequences and, in 96% of the detected cases the inferred TSM was supported at most by four base changes. A phylogenetic analysis showed TSM-patterns associated with structural changes to be 3.7 times more frequent in the terminal nodes than in the first two levels of internal nodes (Fig. 4). This differs from TSM-patterns associated with classical CMs and is compatible with studies indicating rRNA secondary structure conservation over that of the corresponding sequence (25). A number of positive cases nevertheless suggests that the TSM mechanism has the potential to shape the rRNA structure and form completely new hairpins.

CMs with motifs forming stable hairpins in DNA. We hypothesized that CMs associated with a full inversion of the loop sequence would provide strong evidence for the involvement of the TSM process and the mechanism occurring inter-strand. The cases matching the expected pattern were very rare (Fig. 3b), however, totalling 11 events. On the other hand, inversions of the loop sequence could also occur at a perfectly pairing hairpin sequence and those would not leave any footprint in the adjacent stems. We observed 50 mutation clusters in the loop sequence consistent with such inversions. The affected loop sequences were typically four nucleotides long and partially reverse complement (76% of the cases), minimizing the impact of the inversion, but we also detected inverted loops of 7-8 nuc in length and containing multiple sequence differences (Fig. S3c-d). In further 89 cases, the apparent loop inversion was supported by a single change in the middle of an odd number of loop bases, the loop sequence being otherwise reverse complement. Majority of these (86.5%) were between sequences AUU-AAU and UAA-UUA.

The rarity of the loop inversions and the high reverse complementarity of the observed cases is consistent with the conservation of the hairpin loop sequences (25), and enrichment of four nucleotide patterns giving stable RNA structures (26). To understand the effect of this, we computed the frequencies of different rRNA hairpin loops and their reverse complements in our data. We predicted secondary structures to all sequences, and extracted in total 7,184,663 hairpin loop sequences from non-root internal nodes (see Methods). As expected, we observed enrichment of hairpin sequences belonging to GNRA, UNCG and CUUG families (Table S1) known to form exceptionally stable hairpins (27). The forward sequences for the three classes were, respectively, 57, 60 and 7 fold enriched in comparison to their reverse complements, the most common hairpin sequence GAAA being 91 times more frequent than its opposite UUUC. While this suggests strong selection against inversion of selected loop sequences, many less common loop sequences had nearly equal numbers of forward and reverse complement versions (Table S1).

A sequence giving an exceptionally stable loop structure is unlikely to be inverted (Fig. 2f), but it could promote intra-strand template switching (Fig. 2b) by bending the single-strand DNA into a loop conformation. To test this, we computed the frequencies of motifs associated with stable hairpin loop structures, loop closing pairs and loop lengths within hairpins containing instantaneous CMs, and compared those to background frequencies computed from all hairpins of non-root ancestral sequences. Interestingly, we observed that four-nucleotides-long loops, which represent the most stable rRNA and rDNA loop lengths (28), were strongly enriched among the instantaneous CMs (p-value<0.00001, one-proportions z-test). Of the most frequent RNA loop motifs, GNRA and CUUG were enriched in our data but UNCG was not. Interestingly, GNRA and CUUG are exceptionally stable both in DNA and RNA, whereas UNCG is stable only in RNA (29).The most common closing pair among the loops linked with instantaneous CMs is C-G. However, it was less frequent (p<0.0001, one-proportion z-test; Table 2) and its reverse complement G-C was more frequent than expected (p<0.001, one-proportion z-test; Table 2). Although both closing pairs are common, the enrichment towards G-C is surprising as the C-G closing pair is associated with exceptionally stable hairpin loops in RNA and DNA (29, 30). The effect of the closing pair may depend on the loop sequence (29, 31) and some previ-

Table 2. Frequencies of loop sequences, lengths and closing pairs.

	Frequency in hairpins		
Motif	with instant. CM	of full data	p-value
GNRA	35.13%	16.58%	0.0
TNCG	4.76%	5.48%	3.41e-33
CTTG	2.01%	1.06%	1.04e-127
Loop length			
3	3.43%	6.63%	0.0
4	52.98%	44.50%	0.0
5	12.56%	14.90%	6.16e-140
6	4.61%	9.42%	0.0
7	10.79%	7.37%	0.0
8	9.91%	9.02%	1.77e-26
Closing pair			
CG	39.11%	43.44%	2.87e-219
GC	38.51%	40.00%	1.34e-4
AT	15.24%	10.37%	0.0
TA	7.14%	8.20%	8.23e-49

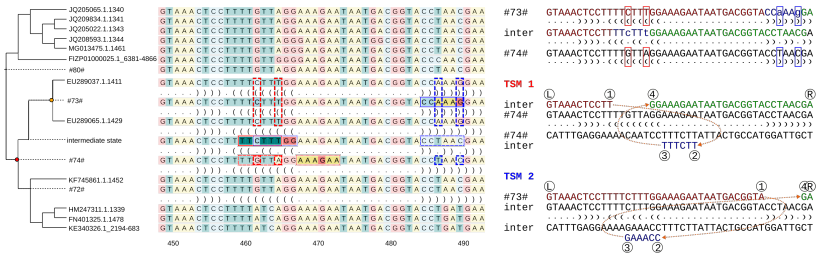

Fig. 5. Inferred two-step mutation history for a pair of instantaneous CMs. The sequence history indicates the stem sequences at nodes #74# and #73# to differ by two CMs, G-C and A-T changing to C-G and T-A (red and blue dashed boxes). The differences can be explained by two TSM events and an inferred intermediate state: the first asymmetric TSM breaks the base pairing by copying loop sequence into the stem (in red); the second TSM fixes the base pairing (in blue). In the alignment, TSM source and target regions are framed with dotted and solid line, homologous regions with dashed line. In the inferred TSM process, affected sites are indicated with red and blue boxes.

ously unidentified *motif - closing pair* sequence combinations may be enriched in the data set. Alternatively, the bias in closing pairs could be explained by a CM-causing mutation mechanism that is asymmetric relative to the direction of the rDNA genes.

Discussion

We found footprints of template switch mutations (TSMs) (5, 6) in rRNA sequences and propose the mechanism as the explanation for the origin of perfect stem structures and for their evolution through compensatory substitutions. Consistent with our hypothesis, we identified compensatory mutations (CMs) associated with nearby parallel sequence changes explainable by a single TSM event. Parallel changes were most common within the stem sequences while those affecting the loop sequence were rare, consistent with the high sequence conservation of the latter (25). Although rare, the presence of complete inversions of the loop sequence indicates that at least some of the TSMs have occurred via the inter-strand mechanism capable of turning the sequence in place (Fig. 2f, Fig. 3b). Interestingly, we also observed TSM-patterns that retain the RNA secondary structure with sequence coming from outside the opposite stem. These demonstrate the flexibility of RNA sequences and suggest that the mechanism could provide an elegant explanation for the complex mutation patterns and jumps in the sequence space observed in evolving RNA sequences (11).

Our study suggests that the TSM mechanism has both a constructive and a destructive role in the evolution of RNA sequences and structures. The constructive role is supported by the enrichment of structurally exceptionally stable motifs in hairpins associated with instantaneous CMs. Formation of a loop structure is defined by the underlying sequence and 1-2 nucleotides adjacent to the loop region (32, 33). Existing rRNA hairpin loop structures may trigger intra-strand TSMs with the nascent strand forming a loop to prime DNA replication, the sequence for the opposite stem functioning as the template. Crucially, such a mechanism fixes the stem sequence with CMs while retaining the secondary structure and the loop sequence. Similar hairpin loop-priming has been described for inverted repeat-containing oligonucleotides replicated by DNA polymerase I of *Escherichia coli* (34), and replication models of parvovirus (single-strand DNA virus) (35) and poxvirus (double-strand DNA virus) (36). The destructive potential of TSMs is represented by the novel loops and radical changes of the secondary structure. Although they are the ultimate source of novelties and enable the evolution, in an established molecule such as rRNA they tend to be deleterious and were mostly observed in terminal branches.

Unexpectedly, we observed an enrichment of G-C closing pairs over the more stable C-G pairs among the CM containing hairpins (Table 2). The asymmetry suggests that mutations do not happen uniformly in both replication directions, and a plausible explanation for the signal is that the mutation mechanism is somehow associated with the unidirectional gene expression. Interestingly, this signal may be connected to another unexplained feature of RNA sequences, the asymmetry of frequencies of G-C and C-G pairs, and of U-A and A-U pairs in the stem structures, the former of each couple appearing at the higher frequency (12). This could be explained by the TSM mechanism if the direction of the switch jump and the strand of the switch were conserved. One scenario conserving the direction of switch jump would be the intra-strand TSM events: in those, the point ② is always left of point ① (Fig. 2b) and the base pairing is thus corrected according to the 5' base. Given this, the possible two-step CMs involving G-U pairs would always lead to either G-C or U-A base pairing:

A-U	→	G-U	→	G-C,
U-A	→	U-G	→	U-A,
C-G	→	U-G	→	U-A,
G-C	→	G-U	→	G-C.

With the right mixture of intra-strand and inter-strand switching, one would obtain base pair frequencies similar to those observed in empirical data (12). Plausible mechanisms conserving the strand of TSM events could be the coordination of transcription and replication to the same directions, or mutations resulting from clashes of the two systems (37, 38).

We found rather modest numbers of positive cases. The reasons for that are many and reflect the technical challenges of working with data from genes that are both multi-copy and RNA-coding. First, the sequencing of rDNA genes is not straightforward and many sequences probably represent a consensus formed of multiple, slightly different rDNA copies. Stochasticity of the sequence assembly and true recombinations within rDNA clusters create hybrid sequences with mosaicism of the sequence origin and homoplasies incompatible with tree-like presentation (39). Second, ribosomal DNA is a challenging target for phylogenetic analyses, because (i) covariance in stems can distort sequence alignment (40), (ii) stem structures violate the independence of homologous sites (40), and (iii) rDNA evolves in concerted fashion, which homogenizes the gene copy variants within genomes (41). All the phenomena can lead to inaccuracies in phylogenetic reconstruction that in turn cause errors in ancestral sequence inference. The high number of instantaneous CMs observed despite these challenges suggests that our data and methodology were adequate for the identification of mutation patterns in ancestral sequence histories. The low numbers of two-step CMs may be explained by the selection keeping the initial mutations at low frequency (12): as such they are unlikely to show up in

consensus-based sequences, potentially rendering our numbers of two-step CMs gross underestimates.

Another factor lowering the number of a observed two-step CMs is our focus on TSM-patterns of at least six nucleotides in length while the perfect Watson-Crick base-pair regions in rRNA genes are usually only a few bases long (42). Due to the low complexity of a four-nucleotide alphabet, a perfect reverse-complement match could be found for nearly any short sequence and the hits of the length of the shortest RNA stems would have little statistical power (22). Rather counterintuitively, the largest TSM-like pattern that we observed, 38-59 bp in size, were of little interest as they typically caused major structural changes, altering the loop location or even breaking the hairpin structure, and did not pass our quality criteria. While gene regions are known to be more conserved than intergenic regions and large TSMs are not tolerated (18, 22), our results indicate that multiple short TSMs may delicately shape and maintain the rDNA sequence and secondary structure.

We did not make assumptions about the biological mechanism(s) behind the observed mutations but focused on local events within stem loop structures. Whatever the mechanism, it could allow a longer distance between the source and target regions and thus have a greater role in the RNA evolution than suggested by our study. While spatially more distant copying events would not create hairpins, they would still generate multiple parallel changes – inexplicable by classical models - and may be crucial for RNA tertiary structures. Interestingly, multi-copy gene clusters may be affected by non-replication-related mutation mechanisms (43) with the potential of creating similar mutation patterns. This has no impact on our findings, however, and the explanation for compensatory substitutions and perfect stem structures would still be copying DNA from another template.

Materials and Methods

Data collection and preprocessing. The ribosomal RNA sequences were downloaded from Silva metagenomic database v. 138 (http://www.arb-silva.de (19)) and Rfam database v. 14.1 (ftp://ftp.ebi.ac.uk/pub/databases/Rfam (21)) (RNA families RF00177, RF01959, RF01960, RF02540, RF02541 and RF02543). Taxonomies for the downloaded RNA sequences were received from Silva database (v. 138) and NCBI databases (ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/ for taxonomy identifiers and ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/ for scientific names; downloaded on 22 March 2020). Information on the RNA families were parsed from the Rfam seed file for v. 14.1 (ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.1/Rfam.seed.gz). A custom python script (downloadable at xxx) was used to remove sequences with degenerate IUPAC symbols and to replace U's with T's. Identical sequences were removed using the cd-hit-est tool of the CD-HIT package v. 4.7 (44, 45) with wordsize 11 and retaining the longest representative. Sequences were then clustered using the cd-hit-est applying wordsize 8, 90% sequence identity and 75% minimum alignment coverage for the longest and shortest sequences (parameters -n, -c, -aL and -aS respectively), and divided into the clusters of minimum ten sequences using the make_multi_seq.pl script. Within clusters, the sequences were compared pairwise and if two sequences did not have at least two mismatches within a 20 bp window, latter of the pair was discarded.

Ancestral sequence history and inference of TSMs. Sequence clusters were aligned with MAFFT (v. 7.310; FFT-NS-i, 1,000 iterations) (46), and then trimmed with TrimAl (47) and 'automated' mode. Alignments shorter than 200 columns were discarded. Maximum likelihood (ML) trees were computed with IQ-TREE (v. 1.6.1) (48) using automated model selection (49) and tree finder, and resulting trees were midpoint-rooted using ete3 python library (50). Sequences within each cluster were realigned and ancestral sequences inferred according to ML trees with PAGAN2 (51). Phylogenetic trees were traversed with a custom python script applying ete3 library (50): each non-root parent node was compared to its child node (query node) using the FPA tool (downloadable at xxx; see Table S2 for details) and mutation patterns consistent with TSM events of at least six nucleotides in length were recorded.

Instantaneous CMs differ from two-step CMs potentially explainble by TSMs and have the initial mutation and the restoring mutation present in opposite stems of the same sequence. They were identified by screening parent-descendant pairs in each tree and locating loops that a) were in identical positions in both sequences, b) formed perfect Watson-Crick base pairing, and c) differed in sequence. Uncertain IUPAC characters were not allowed in reference hairpin sequences.

Quality control for inferred TSMs. In inferred TSM events, the target region of a query differs from the homologous sequence at its parent node, and the changes are attributed to the TSM mechanism copying sequence from the source region in the parent. Using a custom python script, the quality of the ancestral sequence inference was confirmed for all TSM target and source regions. For the TSM target regions, the parent sequence was considered reliable if the target region was at least 90% identical to i) that of its own parent (grandparent of the query), and ii) that of its second child (sister node of the query). The query sequence was considered reliable if the target region was at least 90% identical to i) that of one non-leaf child node, or ii) those of two leaf child nodes. For the TSM source regions, the parental node sequence reconstruction was considered reliable if source sequence was at least 90% identical to i) that of its parent sequence, and ii) that of either query or its sister.

The secondary structures were predicted using RNAfold (52) with default parameters, and uniformity of predictions was assessed for the TSM target region of the (non-leaf) query node and its parental node. The predicted secondary structure of the parental node was required to be at least 70% identical to the predicted structure of the query's sister. The predicted structure of the (non-leaf) query was required to be at least 70% identical to i) that of one non-leaf child node, or ii) those of two leaf child nodes.

Sequences of all solved structures in PDB were downloaded (ftp://ftp.wwpdb.org/pub/pdb/derived_data/pdb_seqres.txt.gz; downloaded on 28 October 2020). The sequences for rRNA were extracted based on the title (982 sequences in total), a BLAST database was created, and all terminal sequences were screened against it using blastn (v. 2.6.0+). Hits with e-value 0.00 and sequence identity of at least 95% for a sequence containing a TSM hit were selected using awk and python scripts. The structures for these were downloaded from PDB (ftp://ftp.ebi.ac.uk/pub/databases/pdb/compatible/pdb_bundle), and the affected hairpin chain was extracted. Predicted and pdb structures were visualized using RNApdbee 2.0 (53) (http://rnapdbee.cs.put.poznan.pl) with default settings with the exception of the visualization setting 'PseudoViewer-based procedure'.

TSMs' impact on secondary structures. A *compensating mutation* was attributed to a TSM event if i) either the TSM source or target sequence overlapped with a sequence difference between the query and its parent, ii) the TSM source and target sequences were in the separate stems of the same RNA hairpin loop, and iii) the source and target sequences were equally distant from the loop region. The distance was counted as nucleotides from the first and last nucleotide of the predicted loop. Internal loops or bulges were not taken into account. A *loop inversion* was attributed to a TSM event if i) both TSM-source and the TSM-target sequences overlapped with both stems of the same RNA loop and ii) the length of the loop sequence was maintained. The loop sequence inversion was considered symmetrical if the overlap of the TSM-source sequence in the first stem was equally long to the overlap of the TSM target sequence in the second stem. Formation of a *new loop* was attributed to a TSM event if i) the stem was not present in the parent sequence, and ii) the TSM source and target sequences of the query were symmetrically in the separate stems of the same hairpin RNA structure. The TSM associated with a new loop had to be at least 8 bp in length and to explain two or more base changes.

Calculation of TSM and base frequencies. TSM-patterns associated with CMs, loop inversions or new loops were counted at each tree

level and the counts were divided by the total number of branches of that tree level in the affected trees. The identification of the branch level was done with post-order tree traversal, assigning each branch to the lowest possible level.

Hairpin loop sequences and closing pair bases in secondary structures of internal nodes were extracted and counted. The sequence and structural quality of the hairpins were confirmed as described above; in addition, the closing pair had to form a perfect Watson-Crick base-pair and uncertain IUPAC characters were not allowed in the hairpin sequence. The frequencies of different closing pairs, loop lengths and loop sequence motifs were compared between hairpins with instantaneous CMs and all hairpins. The significance of the differences was studied using two-sided one proportion z-test. Mutation types in the stems (base pair in parent vs. child) were counted separately for instantaneous and two-step CMs, and the counts were visualized as a heatmap using matplotlib Python package. Only base-pairs with at least one count exceeding 150 were retained.

Script availability. The scripts used in the study will be available at Github.

1. PB Moore, Structural Motifs in RNA. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
2. N Leontis, E Westhof, *Modeling RNA Molecules*, eds. N Leontis, E Westhof. (Springer Berlin Heidelberg, Berlin, Heidelberg) Vol. 27, pp. 5–17 (2012).
3. ER Tillier, RA Collins, High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**, 1993–2002 (1998).
4. MT Dixon, DM Hillis, Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.* **10**, 256–267 (1993).
5. LS Ripley, Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc. Natl. Acad. Sci.* **79**, 4128–4132 (1982).
6. BE Dutra, ST Lovett, Cis and Trans-acting Effects on a Mutational Hotspot Involving a Replication Template Switch. *J. Mol. Biol.* **356**, 300–311 (2006).
7. A Löytynoja, N Goldman, Short template switch events explain mutation clusters in the human genome. *Genome Res.* **27**, 1039–1049 (2017).
8. DK Butler, LE Yasuda, MC Yao, Induction of Large DNA Palindrome Formation in Yeast: Implications for Gene Amplification and Genome Stability in Eukaryotes. *Cell* **87**, 1115–1122 (1996).
9. L Costantino, et al., Break-Induced Replication Repair of Damaged Forks Induces Genomic Duplications in Human Cells. *Science* **343**, 88–91 (2014).
10. CMB Carvalho, JR Lupski, Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
11. MV Meer, AS Kondrashov, Y Artzy-Randrup, FA Kondrashov, Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **464**, 279–282 (2010).
12. PG Higgs, RNA secondary structure: physical and computational aspects. *Q. Rev. Biophys.* **33**, 199–253 (2000).
13. Woese, C R and Pace, N R, *Probing RNA Structure, Function, and History by Comparative Analysis*, ed. R. F. Gesteland and J F Atkins. (Cold Spring Harbor Laboratory Press), pp. 91–117 (1993).
14. RR Gutell, N Larsen, CR Woese, Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Mol. Biol. Rev.* **58**, 10–26 (1994).
15. ED Ladoukakis, A Eyre-Walker, The excess of small inverted repeats in prokaryotes. *J. Mol. Evol.* **67**, 291–300 (2008).
16. B Lavi, E Levy Karin, T Pupko, E Hazkani-Covo, The Prevalence and Evolutionary Conservation of Inverted Repeats in Proteobacteria. *Genome Biol. Evol.* **10**, 918–927 (2018).
17. RR Sinden, RD Wells, DNA structure, mutations, and human genetic disease. *Curr. Opin. Biotechnol.* **3**, 612–622 (1992).
18. ED Ladoukakis, A Eyre-Walker, Searching for sequence directed mutagenesis in eukaryotes. *J. Mol. Evol.* **64**, 1–3 (2007).
19. C Quast, et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–596 (2013).
20. JF Elder, BJ Turner, Concerted Evolution of Repetitive DNA Sequences in Eukaryotes. *The Q. Rev. Biol.* **70**, 297–320 (1995).
21. I Kalvari, et al., Non-Coding RNA Analysis Using the Rfam Database. *Curr. Protoc. Bioinforma.* **62**, e51 (2018).
22. CR Walker, A Scally, N De Maio, N Goldman, Short-range template switching in great ape genomes explored using pair hidden markov models. *PLOS Genet.* **17**, 1–30 (2021).
23. W Stephan, The rate of compensatory evolution. *Genetics* **144**, 419—426 (1996).
24. PG Higgs, Compensatory neutral mutations and the evolution of RNA. *Genetica* **102-103**, 91–101 (1998).
25. S Smit, J Widmann, R Knight, Evolutionary rates vary among rRNA structural elements. *Nucleic Acids Res.* **35**, 3339–3354 (2007).
26. G Varani, Exceptionally Stable Nucleic Acid Hairpins. *Annu. Rev. Biophys. Biomol. Struct.* **24**, 379–404 (1995).
27. HA Heus, A Pardi, Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253**, 191–194 (1991).
28. DR Groebe, OC Uhlenbeck, Characterization of RNA hairpin loop stability. *Nucleic Acids Res.* **16**, 11725–11735 (1988).
29. VP Antao, SY Lai, I Tinoco, A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res.* **19**, 5901–5905 (1991).
30. M Nakano, EM Moody, J Liang, PC Bevilacqua, Selection for thermodynamically stable dna tetraloops using temperature gradient gel electrophoresis reveals four motifs: d(cGNNAg), d(cGNABg), d(cCNNGg), and d(gCNNGc). *Biochemistry* **41**, 14281–14292 (2002).
31. CR Woese, S Winker, RR Gutell, Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci.* **87**, 8467–8471 (1990).
32. MT Woodside, et al., Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins. *Proc. Natl. Acad. Sci.* **103**, 6190–6195 (2006).
33. G Bonnet, O Krichevsky, A Libchaber, Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc. Natl. Acad. Sci.* **95**, 8602–8606 (1998).
34. E Uhlmann, An alternative approach in gene synthesis: use of long selfpriming oligodeoxynucleotides for the construction of double-stranded DNA. *Gene* **71**, 29–40 (1988).
35. SF Cotmore, P Tattersall, Parvovirus diversity and DNA damage responses. *Cold Spring Harb. Perspectives Biol.* **5** (2013).
36. B Moss, Poxvirus DNA replication. *Cold Spring Harb. Perspectives Biol.* **5** (2013).
37. S Hamperl, MJ Bocek, JC Saldivar, T Swigut, KA Cimprich, Transcription-Replication conflict orientation modulates R-Loop levels and activates distinct DNA damage responses. *Cell* **170**, 774–786.e19 (2017).
38. YH Chen, et al., Transcription shapes DNA replication initiation and termination in human cells. *Nat. Struct. Mol. Biol.* **26**, 67–77 (2019).
39. MJ Sanderson, JJ Doyle, Reconstruction of Organismal and Gene Phylogenies from Data on Multigene Families: Concerted Evolution, Homoplasy, and Confidence. *Syst. Biol.* **41**, 4–17 (1992).
40. H Letsch, K Kjer, Potential pitfalls of modelling ribosomal rna data in phylogenetic tree reconstruction: Evidence from case studies in the metazoa. *BMC evolutionary biology* **11**, 146 (2011).
41. TH Eickbush, DG Eickbush, Finely Orchestrated Movements: Evolution of the Ribosomal RNA Genes. *Genetics* **175**, 477–485 (2007).
42. J Wolters, The nature of preferred hairpin structures in 16S-like rRNA variable regions. *Nucleic Acids Res.* **20**, 1843–1850 (1992).
43. D Liao, Concerted Evolution: Molecular Mechanism and Biological Implications. *The Am. J. Hum. Genet.* **64**, 24–30 (1999).
44. L Fu, B Niu, Z Zhu, S Wu, W Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
45. W Li, A Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
46. K Katoh, DM Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
47. S Capella-Gutiérrez, JM Silla-Martínez, T Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
48. LT Nguyen, HA Schmidt, A von Haeseler, BQ Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
49. S Kalyaanamoorthy, BQ Minh, TKF Wong, A von Haeseler, LS Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
50. J Huerta-Cepas, F Serra, P Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
51. A Löytynoja, AJ Vilella, N Goldman, Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* **28**, 1684–1691 (2012).
52. R Lorenz, et al., ViennaRNA Package 2.0. *Algorithms for Mol. Biol.* **6**, 26 (2011).
53. M Antczak, et al., RNApdbee—a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. *Nucleic Acids Res.* **42**, W368–W372 (2014).