614–622 *Nucleic Acids Research, 2007, Vol. 35, No. 2* *Published online 18 December 2006*
doi:10.1093/nar/gkl1036

Encoding folding paths of RNA switches

A. Xayaphoummine[1], V. Viasnoff[2], S. Harlepp[1] and H. Isambert[1,2,*]

[1]Laboratoire de Dynamique des Fluides Complexes, CNRS-ULP, Institut de Physique, 3 rue de l'Université, 67000 Strasbourg, France and [2]RNA Dynamics and Biomolecular Systems, Physico-chimie Curie, CNRS UMR168, Institut Curie, Section de Recherche, 11 rue P. & M. Curie, 75005 Paris, France

Received July 17, 2006; Revised November 2, 2006; Accepted November 7, 2006

ABSTRACT

RNA co-transcriptional folding has long been suspected to play an active role in helping proper native folding of ribozymes and structured regulatory motifs in mRNA untranslated regions (UTRs). Yet, the underlying mechanisms and coding requirements for efficient co-transcriptional folding remain unclear. Traditional approaches have intrinsic limitations to dissect RNA folding paths, as they rely on sequence mutations or circular permutations that typically perturb both RNA folding paths and equilibrium structures. Here, we show that exploiting sequence symmetries instead of mutations can circumvent this problem by essentially decoupling folding paths from equilibrium structures of designed RNA sequences. Using bistable RNA switches with symmetrical helices conserved under sequence reversal, we demonstrate experimentally that native and transiently formed helices can guide efficient co-transcriptional folding into either long-lived structure of these RNA switches. Their folding path is controlled by the order of helix nucleations and subsequent exchanges during transcription, and may also be redirected by transient antisense interactions. Hence, transient intra- and inter-molecular base pair interactions can effectively regulate the folding of nascent RNA molecules into different native structures, provided limited coding requirements, as discussed from an information theory perspective. This constitutive coupling between RNA synthesis and RNA folding regulation may have enabled the early emergence of autonomous RNA-based regulation networks.

INTRODUCTION

RNA molecules exhibit a wide range of functions from essential components of the transcription/translation machinery (1) to natural or *in vitro* selected ribozymes (2,3) or aptamers (4–6) and different classes of gene expression regulators (e.g. miRNA, siRNA and riboswitches) (7–13). The functional control of many of these RNA molecules hinges on the formation of specific base pairs in *cis* or *trans* and secondary structure rearrangements between long-lived alternative folds. Yet, because of their limited four-letter alphabet and strong base pair stacking energies, RNAs are also prone to adopt long-lived misfolded structures (14), as observed for instance upon heat renaturation. Hence, efficient RNA folding paths leading to properly folded structures bear an important role in the regulatory function of non-coding RNAs and mRNA untranslated regions (UTRs) (14).

It has long been proposed (15–17) that, during transcription, the progressive folding of nascent RNAs limits the number of folding pathways, presumably facilitating their rapid folding into proper native structures. However, it is not clear whether native domains fold sequentially and independently from one another or whether co-transcriptional folding paths result from more intricate interactions between domains and individual helices. Transcriptional RNA switches provide interesting examples of co-transcriptional folding paths with local competition between newly formed and alternative helices. Such natural RNA switches are typically found in virus or plasmid genomes (18–22) and in bacterial mRNA UTRs where they regulate gene expression at the level of transcription elongation (e.g. through termination/anti-termination mechanism) (23–25) or at the level of translation initiation (e.g. through sequestration of Shine–Dalgarno motifs) (26–31). The structural changes controlling their regulatory function may correspond to a switch in equilibrium structure or in co-transcriptional folding path caused by binding an effector (e.g. a protein, a small metabolite or an antisense sequence) (32). Alternatively, RNA switches may operate through spontaneous or assisted relaxation of an initially metastable co-transcriptional fold (33). Hence, RNA switches can have stringent needs to control their folding between alternative structural folds, which makes them ideal candidates to dissect RNA co-transcriptional folding mechanisms and estimate the minimal sequence constraints to encode them.

*To whom correspondence should be addressed. Tel: +33 1 42 34 64 74; Email: herve.isambert@curie.fr
Present addresses:
A. Xayaphoummine, Université Paris 7, UMR 7056, 2, place Jussieu, 75251 Paris cedex 05, France
V. Viasnoff, ESPCI, 10 rue Vauquelin, 75005 Paris, France
S. Harlepp, IPCMS–GONLO, 23 rue du Loess BP 43, 67034 Strasbourg cedex 2, France

Several inspiring reports have demonstrated the importance of co-transcriptional folding (21,34–40), and folding pathways of *Escherichia coli* RNaseP RNA (41,42) and *Tetrahymena* group I intron (43,44) have been probed using circularly permutated variants of their wild-type sequences (see Discussion). Yet, dissecting folding paths of natural RNAs remains generally difficult due to two fundamental issues: (i) sequence mutations or circular permutations generally affect both RNA folding paths and equilibrium structures (hence preventing independent probing of folding paths on their own) and (ii) many natural non-coding RNAs have likely evolved to perform multiple interdependent functions, which are all encoded on their primary sequence and thereby all potentially affected, directly or indirectly, by sequence mutations.

To circumvent these limitations, we propose to use artificial RNA switches, presumably void of biological functions, and investigate how to efficiently encode their folding paths by exploiting simple sequence symmetries, instead of extensive (and possibly non-conclusive) mutation studies. Beyond specific examples of natural or designed RNA sequences, we aim at delineating general mechanisms and coding requirements for efficient co-transcriptional folding paths.

In a nutshell, we have designed a pair of synthetic RNA switches sharing strong sequence symmetries, so that both molecules partition, at equilibrium, into equivalent branched and rod-like nested structures with nearly the same free energy. Yet, in spite of this structural equivalence between the two RNA switches at equilibrium, we demonstrate that their folding path can be encoded to guide the first RNA switch exclusively into the branched structure, while the other switch adopts instead the rod-like nested structure by the end of transcription. This shows that folding paths do not simply result from the sequential formation of native helices in their order of appearance during transcription (i.e. sequential folding, see Discussion). Instead, efficient folding paths rely on the relative stability between native and non-native helices together with their precise positional order along the 5′–3′ oriented sequence (i.e. encoded co-transcriptional folding). Furthermore, we show that efficient folding path can be redirected through transient antisense interaction during transcription, suggesting an intrinsic and possibly ancestral coupling between RNA synthesis and folding regulation.

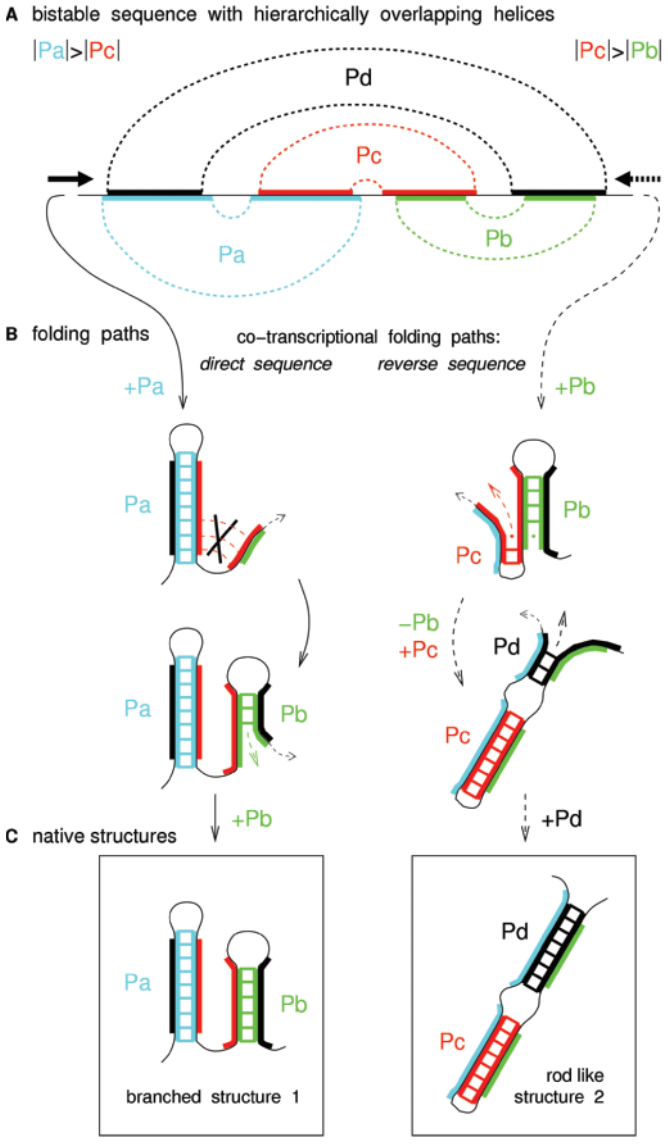

Figure 1. Encoded co-transcriptional folding path of a bistable RNA switch. (**A**). Bistable generic sequence with hierarchically overlapping helices. (**B**). Opposite co-transcriptional folding paths for the direct and reverse sequences rely on small asymmetries in helix length in the direction of transcription (i.e. |Pa| > |Pc| direct sequence and |Pc| > |Pb| reverse sequence). (**C**). *Either* branched or rod-like native structures are obtained depending on the direction of transcription, although both structures can be designed to co-exist at equilibrium.

MATERIALS AND METHODS

RNA switch design

RNA switches with encoded folding paths depicted in Figures 1 and 2 were designed using Kinefold online server (45,46) (http://kinefold.curie.fr). A starting GG doublet was chosen to ensure efficient T7 transcription. The sequence of the 'direct' RNA switch (i.e. 5′-ABCD-3′, 73 nt, Figure 2) is: 5′-GGAACCGUCUCCCUCUGCCAAAAGGUAGAGGG-AGAUGGAGCAUCUCUCUCUACGAAGCAGAGAGAGAC-GAAGG-3′. The 'reverse' RNA switch has exactly the opposite sequence (or reversed orientation), i.e. 5′-DCBA-3′. A 'reverse' sequence with a single mutation U_{38}/C_{38} was also studied to unambiguously establish the correspondence between branched versus rod-like structures and the two migrating bands on non-denaturing polyacrylamide gels (Figure 3); i.e. 5′-GGAAGCAGAGAGAGACGAAGCAUCUCUCUCUACGA-GGC$_{38}$AGAGGGAGAUGGAAAACCGUCUCCCUCUGCC-AAGG-3′. Complementary DNA oligonucleotides including T7 promoter and KpnI/StuI/BamHI restriction sites at sequence extremities were bought from IBA-Naps, Germany.

Sequence cloning and in vitro transcription

Sequences were inserted into pUC19 plasmid (between KpnI and BamH1 restriction sites) using enzyme removal kits (Qiagen) and cloned into calcium competent *E.coli* (DH5α strain) following standard cloning protocoles. Following

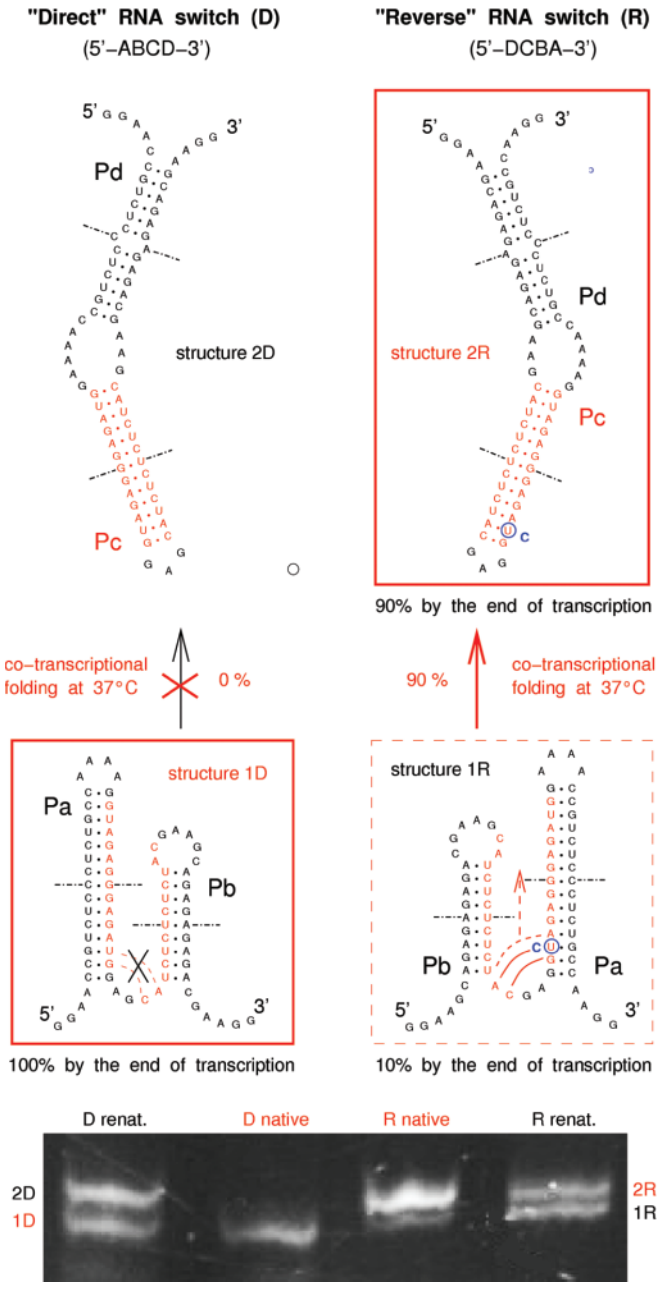

Figure 2. Opposite co-transcriptional folding paths of a pair of RNA switches with 'direct' and 'reverse' sequences (i.e. 5′-ABCD-3′ versus 5′-DCBA-3′). Structures 1D and 1R (respectively, 2D and 2R) of the direct and reverse switches are energetically equivalent because of helix symmetries; dashed lines indicate mirror symmetry of Pa, Pb, Pc and Pd which are therefore conserved under sequence reversal relating direct and reverse switches. Despite these strong similarities between D and R structures at equilibrium, direct and reverse switches display 'opposite' co-transcriptional folding paths (direct switch into structure 1D and reverse switch into structure 2R) guided through a helix encoded persistence (left) or exchange (right) during *in vitro* transcription using T7 RNA polymerase (see Materials and Methods).

plasmid extraction (Genomed kit), inserts were sequenced and cut at the StuI blunt restriction site located at the end of the desired DNA template. Run off transcription was performed *in vitro* using T7 RNA polymerase (New England Biolabs) for up to 4–5 h at 37 or 25°C. Heat renaturation was performed from 85°C to room temperature in ∼10 min

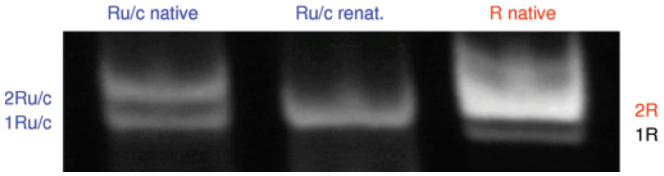

Figure 3. Correspondence between branched versus rod-like structure and migrating bands. A single mutation U_{38}/C_{38} on the reverse sequence, Ru/c (see blue u/c mutation in Figure 2) unambiguously demonstrates the correspondence between the stabilized branched structure and the lower band on the gel (see text).

(starting from 95°C gave the same results). Renatured and co-transcriptional native structures were then separated on 12% 19:1 acryl-bisacrylamide non-denaturing gels (1× TAE, temperature <10°C) and observed using ethidium bromide staining (0.1 μg/μl): Ethidium bromide slightly re-equilibrates the molecule equilibrium partition between branched and rod-like structures during heat renaturation but has no measurable effect at room temperature on the strongly biased partition between co-transcriptional native structures. Controls using denaturing gels (6%, 19:1 acryl-bisacrylamide, RNA in formamide and 8 M urea, 50°C) showed that >90% of transcripts had the expected run off transcription length. Virtually no other bands (i.e. <5% of total transcript) were observed on non-denaturing gels either (i.e. apart from the single or double bands shown in Figures 2–4).

RESULTS

This section is organized into two complementary subsections. The first one is primarily experimental and demonstrates, using sequence symmetries, the basis for encoding efficient folding paths with a pair of 'symmetrically equivalent' RNA switches adopting either their branched or rod-like structure by the end of transcription. The second subsection is theoretical and discuss, from an information content perspective (47,48) and beyond sequence-specific examples, the coding requirement for such efficient co-transcriptional folding.

Co-transcriptional folding of 'direct' and 'reverse' RNA switches

We decided to investigate the basic mechanisms and coding requirements for efficient RNA folding paths with a stringent test case. Following the RNA switch design depicted on Figure 1, we set out to encode two (oppositely oriented) folding paths on the same RNA sequence. The proposed bistable RNA switch should form a branched structure (1) and a rod-like structure (2) with approximately the same free energy at equilibrium, and yet be guided into either one of these structures only, depending on the direction of synthesis (49), (Figure 1).

In practice, however, 5′–3′ versus 3′–5′ folding paths cannot be probed on the same RNA sequence, as there is no RNA polymerase known to perform transcription in 'opposite' (3′–5′) direction. Hence, instead of studying a single RNA sequence, we have actually used a pair of RNA switches with exactly opposite sequences, i.e. 5′-ABCD-3′ and 5′-DCBA-3′ (see Materials and Methods). It is important to

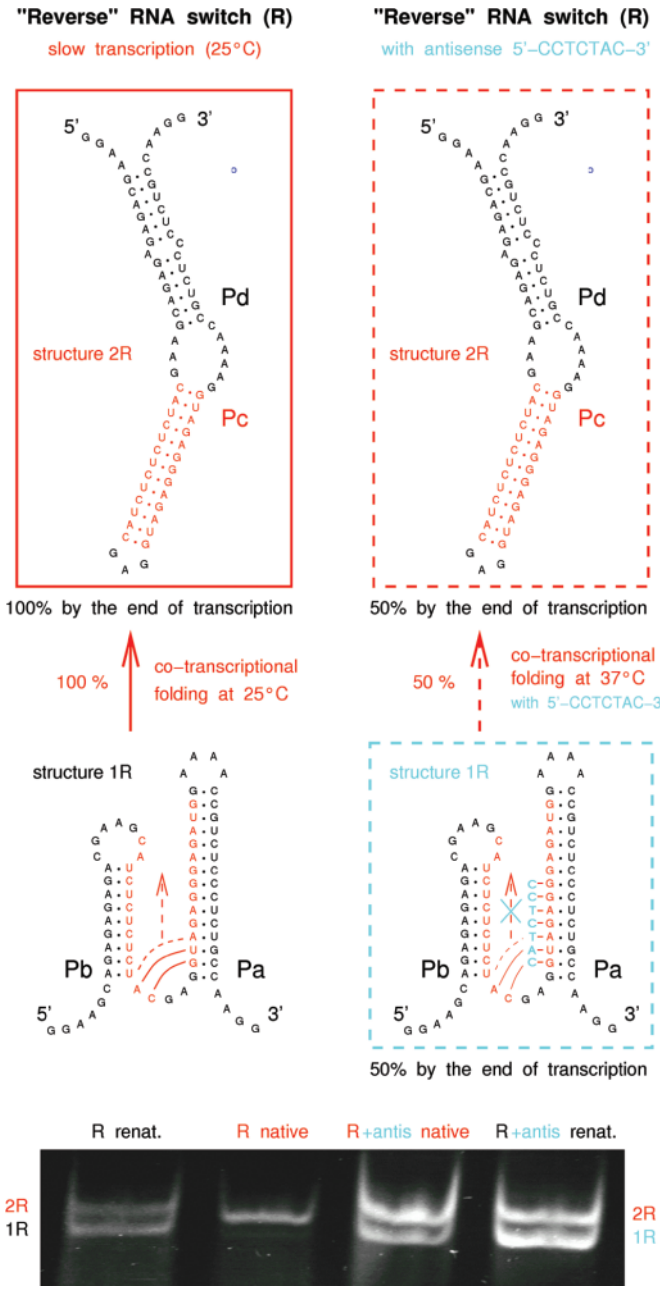

Figure 4. Influence of temperature and transient antisense interactions on co-transcriptional folding. Equilibrium and native structures of reverse switch (R) with *in vitro* T7 transcription at 25°C (left, see text) and under *in vitro* T7 transcription in presence of 0.3 nmol/µl of the 7 nt antisense DNA oligonucleotide CCTCTAC (right, see text). Structures are separated on a 12% 19:1 acryl-bisacrylamide non-denaturing gel (temperature <10°C) and observed using ethidium bromide staining as on Figure 2 (see Materials and Methods).

note that, in general, such pairs of RNA molecules do not adopt related structures at equilibrium, due to the large asymmetry between free energies of stacking base pairs with reversed orientation (e.g. 5′-GC/GC-3′ $\simeq -3.4$ kcal/mol and 3′-GC/GC-5′ $\equiv$ 5′-CG/CG-3′ $\simeq -2.4$ kcal/mol). For this reason, the pair of direct (D) and reverse (R) RNA switches, we have designed (Figure 2), forms at equilibrium a branched structure (1) and a rod-like structure (2) constructed around symmetric helices, that are exactly conserved under sequence reversal (dashed lines on Figure 2 indicate mirror symmetry of Pa, Pb, Pc and Pd helices). Thus, comparing transcription products of the direct and reverse sequences probes the directionality of their folding paths while keeping the equilibrium structures of both switches essentially equivalent by symmetry; since structures 1D and 1R (respecively, 2D and 2R) of the direct and reverse switches are built on the same helices Pa and Pb (respectively, Pc and Pd), their sole free energy difference concerns the small sequence dependent contribution of single-stranded regions in the branched (respectively, rod-like) structure (GNRA tetra-loops and all other sequence-dependent tabulated loops have been avoided).

Figure 2 demonstrates that, in spite of these strong similarities between equilibrium structures, the two RNA switches are indeed guided towards two distinct native structures upon *in vitro* transcription (see Materials and Methods). The correspondence between branched versus rod-like structures and migrating bands on polyacrylamide gels was unambiguously established using a single mutation U_{38}/C_{38} on the reverse sequence (Figure 2). This mutation stabilizes the branched structure (UG > CG) relative to the rod-like structure (AU < AC) at equilibrium, hence demonstrating the correspondence between branched structure and lower band, (Figure 3). Note, however, that this mutation also perturbs the co-transcriptional folding path by redirecting about half of the molecules into the branched structure, hence illustrating the difficulty to dissect independently folding paths from equilibrium structures with sequence mutations only (see Introduction).

These results strongly support the co-transcriptional folding principles depicted on Figure 1 which primarily rely on the difference in helix length (i.e. $|Pa| > |Pc|$ for the direct switch and $|Pc| > |Pb|$ for the reverse switch) to code for the co-transcriptional formation of structures 1D and 2R, respectively. This small asymmetry between successive overlapping helices in the direction of transcription induces a divergence in structural cascades between the two co-transcriptional folding paths; namely, the red helix Pc cannot displace and replace the longer (and stronger) helix Pa previously formed by the direct switch during transcription ($|Pa-Pc|/|Pc| \simeq 15\%$), while Pc does displace and replace the shorter (and weaker) helix Pb initially formed by the nascent reverse sequence ($|Pc-Pb|/|Pc| \simeq 30\%$). The efficacy of these folding paths using T7 polymerase appears maximum ($\simeq 100\%$) for the direct sequence while it is $\sim 90\%$ for the reverse switch at 37°C, suggesting that the branch migration exchange between Pb and Pc is not always successful in these conditions (T7 transcription rate is ~200–400 nt/s at 37°C); however, we found that the folding bifurcation is almost always achieved ($\simeq 100\%$, Figure 4) at lower temperature 25°C which decreases 3-to 4-folds T7 transcription rate (50, 51). This small improvement in bifurcation efficiencies suggests that the decreasing elongation rate indeed prevails over other opposite kinetic factors at lower temperature. In particular, nucleation of Pc (which probably involves the opening of one or two base pairs of Pb, as sketched in Figure 2) and the subsequent branch migration between Pb and Pc are probably both slowed down at lower temperature.

Overall, this demonstrates that the competition between native and non-native helices can lead to efficient

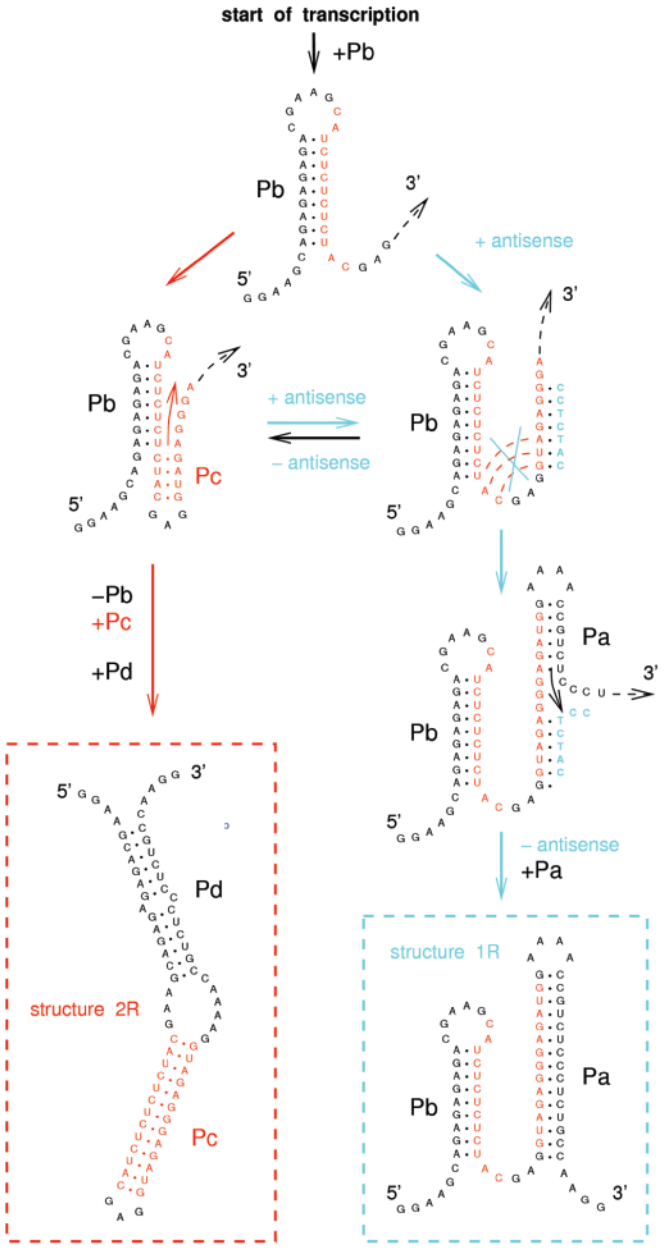

Figure 5. Antisense regulation of co-transcriptional folding paths. Interpretation of the encoded (left) and redirected (right) co-transcriptional folding paths of the reverse switch (Figure 4). This is based on simulations performed using the kinefold server (46) (http://kinefold.curie.fr), To simulate the effect of antisense interaction, the 7mer and RNA switch sequences are actually attached together via an inert linker (made of 'X' bases that do not pair).

co-transcriptional folding paths of RNA switches independently from their actual equilibrium structures.

Moreover, we found that the folding path of the reverse switch could be significantly redirected toward structure 1R (≃50%) through transient antisense interactions, (Figure 4 right panel). This is simply achieved using a 7 nt long antisense oligonucleotide designed to interfere through competing interactions, with the encoded exchange between helices Pb and Pc, (Figure 5). Note, in particular, that the hybridized antisense probe is eventually displaced by the longer (and stronger) downstream helix Pa of each nascent reverse sequence (no global shift of equilibrium bands is observed between transcriptional folds formed in the presence or absence of antisense probe; Figure 4). This shows that transient antisense interactions can, in principle, control multiple turnovers of redirected folding pathways.

Hence, transient intra- and inter-molecular base pair interactions can efficiently regulate the folding of nascent RNA molecules between alternative long-lived native structures, irrespective of their actual thermodynamic stability. Indeed, once formed, the co-transcriptional structures 1D and 2R remain trapped out-of-equilibrium for more than a day at room temperature (data not shown) demonstrating that these RNA switches can reliably store information on physiological time scales with their co-transcriptionally folded structures. In another context, the ability to control folding between distinct long-lived structures of nucleic acids using electrical (49) or thermal (52) stimuli, instead of transcription, could also lead to nanotechnology applications.

Although our conclusions are based on particular examples of synthetic RNA switches related by sequence reversal and helix symmetries, we want to stress that these strong symmetry constraints are solely instrumental in demonstrating the possible independence between encoded folding paths and low-free energy RNA structures. These symmetries are not directly used nor necessary to achieve efficient co-transcriptional folding. On the contrary, imposing such strong sequence symmetries greatly limits the additional 'information content' that can possibly be encoded on the sequence. In the next subsection, we discuss how this use of sequence symmetries can actually be formalized to provide quantitative estimates on the minimum coding requirement for selective folding paths of generic RNA swiches.

Bounding coding requirement through sequence symmetries

In this subsection, we discuss how sequence symmetries can be utilized to estimate necessary base pairing conditions to encode efficient co-transcriptional folding paths. This requires, however, to reformulate base pairing conditions from an information content perspective, following the approach developped for biomolecular sequences (47,48).

In the following, we first establish a simple conservation law for information content. We then argue that upperbounds for the coding requirement of selective folding paths (or other molecular features) can be estimated by restricting the available coding space with strong sequence symmetries. Ultimately, upperbounds on coding requirements are related to the likelihood that a particular feature might arise from natural or *in vitro* selection.

Let us first recall what the information content of a biomolecule is, before showing how it can actually be estimated for designed RNA switches using sequence symmetries.

The information content I of a functional biomolecule corresponds to the number of sequence constraints that have to be conserved to maintain its function under random mutations (47,48). Expressed in nucleotide unit, the maximum information content that can be encoded on an N-nucleotide-long RNA sequence is precisely $I_{\max} = N$ nucleotides, which define a unique sequence amongst all D^N different RNA sequences with N nucleotides, where D is the size of the

coding alphabet ($D = 4$ for nucleic acids; $D = 20$ for proteins). The fact that neutral mutations can accumulate on an RNA sequence without altering its function implies $I < I_{\max} = N$ and can be simply translated into a conservation law, $I + J = N$, where J, the sequence entropy, corresponds to the number of unconstrained nucleotides which generate $\Omega = D^J$ RNA sequences with the same function. Hence $J = \log_D(\Omega)$ and $I = N - J = \log_D (D^N) - \log_D (\Omega)$ (47,48). Although J and I can be inferred by sampling sequence space as demonstrated in (48), their contributions to N do not usually correspond to a simple partition between J 'meaningless' and I 'meaningful' bases since many sequence constraints actually arise from non-local base–base correlations, as shown, in particular, with base pair covariations between homologous RNA sequences. For this reason, it is usually instructive to quantify information content separately within paired and unpaired regions, when considering RNA structures. Since ~70% of all bases are usually paired in low-energy RNA structures, these base pairs typically contribute the most to total information content and thereby to the minimum coding requirement for a given RNA function. Hence, in the following, we will focus, for simplicity, on short RNA sequences (e.g. $N < 100$ nt) and consider, at first, only base paired regions, ignoring both wobble base pairs and unpaired regions.

With these crude initial assumptions, the information content of a short RNA sequence adopting a unique stable secondary structure can be estimated as $I \simeq J \simeq N/2$, since the first base of each Watson–Crick base pair can be chosen arbitrarily. Hence, overall, short RNA sequences adopting a unique stable secondary structure present a large sequence entropy $J = J_u \simeq N/2$ that can, in principle, be used to encode additional features such as alternative, low-energy structures (53,54) or possibly other molecular properties like co-transcriptional folding pathways, as shown in the previous subsection. For example, encoding the simple bistable RNA of Figure 1A but with exactly overlapping helices ($|\mathrm{Pa}| = |\mathrm{Pb}| = |\mathrm{Pc}| = |\mathrm{Pd}|$) requires that $I \simeq 3N/4$ nt be fixed once the initial $J = J_b \simeq N/4$ bases are chosen arbitrarily (e.g. in the first pairing region). Including also as stable structure the pseudoknot constructed around the same four complementary regions (so as to obtain a tristable RNA molecule) then implies that each pairing region is self-complementary and that only $J \simeq N/8$ bases can be chosen arbitrarily (e.g. in the first half of the first pairing region). Similarly, designing the same bistable RNA of Figure 1A but with symmetrical helices (conserved under 5′–3′ sequence inversion) implies that only around $J = J_{bs} \simeq N/8$ bases can be chosen arbitrarily (e.g. in the first half of the first pairing region).

Similar estimates can be made including wobble base pairs (GU and UG) in addition to Watson–Crick base pairs (GC, CG, AU and UA). In that case, the available sequence entropy becomes $J_u \simeq \log_4 (6) \cdot N/2 \simeq 0.65N$ for a molecule with a unique structure (i.e. with 6 possible base pairs), while we get for the previous bistable RNA, $J_b \simeq \log_4 (14) \cdot N/4 \simeq 0.48N$ (i.e. with 14 possible quadruplet circuits including circular permutations: $2\times^{\mathrm{GC}}_{\mathrm{CG}}$, $2\times^{\mathrm{AU}}_{\mathrm{UA}}$, $2\times^{\mathrm{GU}}_{\mathrm{UG}}$, $4\times^{\mathrm{GU}}_{\mathrm{CG}}$ and $4\times^{\mathrm{AU}}_{\mathrm{UG}}$) or $J_{bs} \simeq \log_4 (14) \cdot N/8 \simeq 0.24N$ with additional symmetric helix restriction, as above. Hence, the possibility of wobble pairs tends to increase sequence entropy J, and to concomitantly decrease the number of sequence constraints I. On the other hand, including a significant fraction of wobble pairs (e.g. one-third) in designed structures tends also to facilitate the formation of unwanted, alternative low-energy structures (e.g. with fewer wobble pairs). Thus, including wobble and WC pairs on an equal footing effectively underestimates coding requirements, since preventing the formation of unwanted alternative structures then requires additional information constraints, especially for longer sequences (e.g. >150 nt). In practice, limiting the fraction of wobble pairs in designed structures can efficiently prevent the formation of alternative structures with limited additional sequence constraint. (e.g. we have used 3 wobble pairs out of a total of 50 bp in structures 1 and 2). This also justifies *a posteriori* the initial crude estimate we have made based on WC pairs only. Ignoring opposite effects of wobble base pairs and unwanted alternative structures is a reasonable first approximation of global information constraint requirement. More precised estimates are difficult to obtain, in general, and sequence candidates should always be tested for possible alternative structures with an appropriate RNA folding algorithm including base pair stacking free energies (we have used kinefold (46) which also includes pseudoknots and knots in RNA structures). Alternatively, it is usually possible, owing to the available sequence entropy, to implement highly constraining heuristics that prevent the formation of alternative structures. Typical heuristics are based on the limitation of short complementary substring occurrences in the sequence (52,55,56).

The previous coding requirement estimates demonstrate that structural information I and symmetry constraints I_s encoded on an RNA sequence can equivalently limit its available entropy J. In particular, combinations of structure and symmetry constraints can provide tighter upperbounds to possible coding increments I' of any new feature encoded on the sequence, via the conservation law $I + I_s + J = N$, i.e. $I' < J = N - I - I_s$.

This can be applied to estimate the minimum information that might be required to obtain two efficient opposite folding paths from the generic bistable RNA switch sequence of Figure 1A. From the previous estimate, we conclude that the two efficient folding paths for the direct and reverse sequences do not require to constrain more than ~1/8th of the 42 bases that are paired in both low-energy structures (i.e. overlapping base pairs). It corresponds to assigning, for each folding path, a maximum of two or three overlapping base pairs not already constrained by the combination of branched and rod-like low-energy structures.

This limited coding requirement concerning overlapping base pairs reinforces, *a posteriori*, our intuitive design principles focussing, instead, on a few bases paired in only one of the two low-energy structures (i.e. non-overlapping base pairs), Figures 1 and 2. Thus, efficient co-transcriptional folding of the direct switch (Figure 2) primarily relies on the sole terminal GC base pair at the base of Pa to prevent the nucleation of Pc, while the exchange between Pb and Pc for the reverse switch hinges on the AC terminal mismatch at the base of Pb to facilitate the nucleation of Pc.

Hence, if all non-functional sequence symmetries are lifted, we expect that selective folding paths can indeed be readily achieved for a wide class of RNA sequences, as they require little encoded information beyond small

asymmetries between alternative helices to guide or prevent their successive exchanges during transcription. Interestingly, this pivotal role of a few unpaired or transiently paired bases for efficient folding paths is also observed for other encoded molecular functions of RNAs. For instance, a few unpaired conserved bases usually prove essential for ribozyme functions or *in vitro* selected aptamers showing remarkable binding efficiency to specific target molecules (48).

DISCUSSION

Sequential versus encoded co-transcriptional folding

Although many convincing reports have shown the importance of co-transcriptional folding (21,34–40), the underlying mechanisms behind efficient folding paths has remained elusive. As recalled in Introduction, this is mainly due to intrinsic difficulties to probe natural RNA folding paths independently from their equilibrium structures and possibly multiple functions.

In particular, studying the equilibrium folds of increasingly longer 3′-truncated transcripts has been argued to miss important out-of-equilibrium intermediates on the folding path of full-length molecules (57). This problem was, however, circumvented by using circularly permutated variants of wild-type sequences to study the folding pathways of *E.coli* RNaseP RNA (41,42) and *Tetrahymena* group I intron (43,44). In this approach depicted with a simple RNA switch on Figure 6, the natural 5′ and 3′ ends of the molecule are genetically connected, while two new ends are engineered from an alternative break point in the circularized sequence. This results in a circular permutation of the various domains on the linearized RNA. The resulting permutation in their transcription order indirectly probes folding paths of the full-length wild-type sequence. In particular, altering the connectivity of the primary sequence may lead to alternative co-transcriptional folds as illustrated in Figure 6. The rationale underlying such circular permutation scenario assumes that co-transcriptional folding primarily relies on sequential folding of the nascent chain into independent native domains with no specific coding for transient base pair interactions during transcription. This suggests, in particular, that co-transcriptional folding favors branched secondary structures (43) (as illustrated in Figure 6) and that the 5′–3′ directionality of RNA transcription plays little role as long as the different native domains can all successively fold during transcription.

In contrast, our results demonstrate that folding paths can efficiently guide RNA transcripts into distinct alternative structures even when competing branched-like conformations exist and could, in principle, form during transcription. This competition between local overlapping helices and even global alternative structures is, in fact, ubiquitous to the folding dynamics and thermodynamics of RNA molecules. For instance, co-transcriptional folding has long been known to induce structural rearrangements as the nascent RNA chain is being transcribed (16,17), demonstrating that transient helices almost inevitably participate in co-transcriptional folding paths. More recently, transient helices were also shown to affect force-induced unfolding paths of single

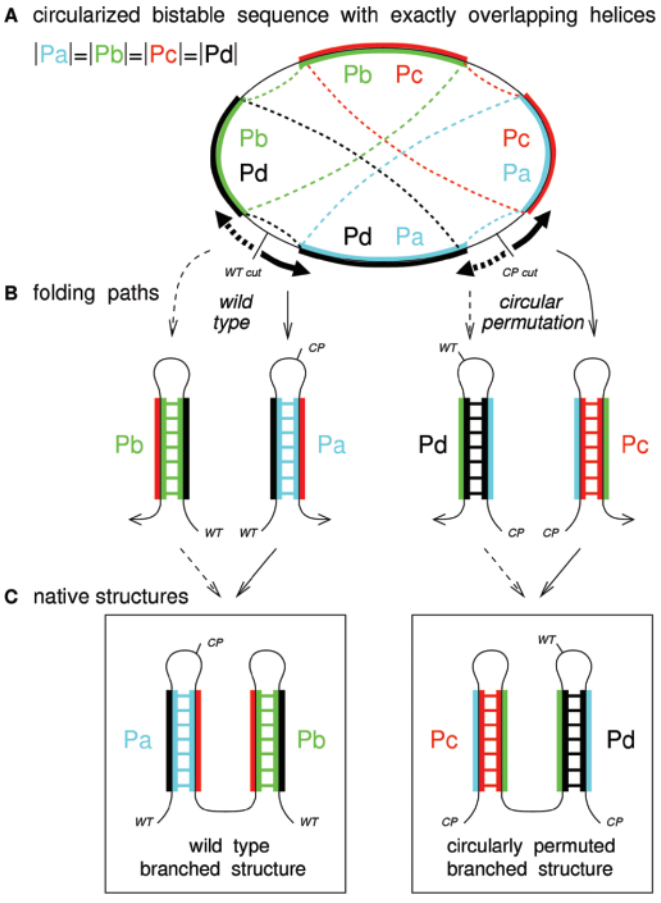

Figure 6. Simple sequential folding of a bistable RNA switch under sequence reversal and circular permutation. (**A**). Bistable generic sequence with exactly overlapping helices Pa, Pb, Pc and Pd. A permutation between the starting and ending regions of the wild-type sequence can be obtained by genetically connecting its 5′ and 3′ ends and engineering two new ends from an alternative break point in the circularized sequence. (**B**). Sequential folding path of the wild-type versus circularly permuted sequences. (**C**). Different branched structures obtained for wild-type and circularly permuted sequences independently from the direction of transcription or sequence reversal (solid versus dashed arrows). The alternative rod-like structures (wild-type Pd–Pc and circularly permuted Pa–Pb) are not formed through sequential folding, although they are expected to coexist with branched structures at equilibrium (not drawn). Figures 1 and 2 show, however, that small asymmetries (2–3 bases) between overlapping helices are sufficient to efficiently guide RNA switches into either branched or rod-like structures (see text).

RNA molecules in micromechanical experiments (58) (e.g. for *E.coli*, 1540 nt long 16S rRNA).

The present study, following earlier stochastic folding simulations reported in (45,46,57,58), demonstrates that this unavoidable competition between alternative base pairs can be exploited to precisely encoded co-transcriptional folding paths of RNA sequences. It is primarily achieved through branch migration exchanges between transient and native helices forming successively as the transcription proceeds from 5′–3′ ends of the sequence. This experimental finding is, in fact, corroborated by a recent statistical analysis of non-coding RNA sequences by Meyer and Miklós (59) who demonstrated the existence of a 5′-3′ versus 3′-5′ asymmetry in the relative positional correlation between native and non-native helices along primary sequences.

Information content and RNA evolution

Non-coding RNAs typically tolerate a significant number of neutral mutations and co-variations in their sequence, which presumably facilitates their continuous adaptation to environmental changes. From an information content perspective, this tolerance to (concerted) mutations also suggests that (partly) unconstrainted nucleotides may be used to encode other alternative structures and functions on the same RNA sequence, a feature which might have favored the emergence of new functional RNAs and RNA switches in the course of evolution (60,61). In fact, much more functional, as well as non-functional information can be encoded on an RNA sequence. For instance, the pair of RNA switches we have designed (Figure 2) demonstrates that not only alternative structures but also selective folding pathways and strong sequence symmetries can all be encoded simultaneously on the same RNA sequence. Although such strong sequence symmetries are both non-functional and probably too stringent constraints to possibly emerge and adapt through natural or *in vitro* selection, they can be used to provide quantitative information on other encoded features of interest. For instance, equalling stacking contributions between alternative folds using helix symmetries also provides a powerful differential approach to uncover other free energy contributions from non-canonical tertiary structure motifs. In the present study, small but reproducible differences in band separation between direct and reverse switches at equilibrium (Figure 2) may possibly reflect a difference in tilt angle between helices Pc and Pd, due to differently structured interior loops in their respective rod-like structures.

In this study, we showed that co-transcriptional folding can efficiently guide RNA folding either toward branched structures (as for the 'direct' switch, Figure 2) or toward elongated rod-like structures (as for the 'reverse' switch, Figure 2) even though their helices are mutually identical. We also argued that only limited information is necessary to encode such selective folding paths for generic RNA switches: it essentially amounts to encoding the relative lengths of helices forming successively during transcription. Moreover, this strict hierarchy between successively exchanging helices can be somewhat alleviated by resorting to topological barriers based on 'entangled' helices (i.e. simple co-transcriptional knots) (46). Hence, we expect that the present findings concerning short RNA sequences (i.e. <100 bases) may be applied to design efficient folding paths for a wide class of larger RNA target structures. This could be achieved by encoding different series of local folding events leading to a succession of either rod-like or branched motifs at the 3′ end of the nascent RNA molecule during transcription. Such folding scheme also provides a theoretical frame to analyze selective folding paths of natural non-coding RNAs (57).

Finally, these results suggest that efficient folding pathways might have easily emerged and continuously adapted in the course of evolution the same way functional native structures have done so through mutation drift in sequence space; non-deleterious mutations are mostly neutral and conserve sequence folds and activity, while new functions may occasionally arise by rare hopping between intersecting networks of neutral mutations (neutral networks) (53,60,61). Furthermore, the fact that encoded folding paths may be redirected through transient antisense interactions (Figures 4 and 5) provides simple 'all-RNA' mechanisms to regulate the functional folding of RNAs in the absence of any elaborate control at the level of transcription initiation. From an ancestral 'RNA World' perspective, this constitutive coupling between RNA synthesis and RNA folding regulation may have also enabled the early emergence of autonomous RNA-based networks relying solely on intra- and intermolecular base pair interactions. Indeed, RNA molecules cross-regulating their respective encoded folding paths could, in principle, be combined to perform essential regulation tasks, characteristic to all natural and engineered control networks (e.g. negative and positive feedback loops, feed-forward loops, toggle switches, oscillators, etc.).

ACKNOWLEDGEMENTS

We thank H. Putzer and L. Hirschbein for critical reading of the manuscript and R. Breaker, D. Chatenay, B. Masquida, K. Pleij, P. Schuster, J. Robert and S. Woodson for discussions. We acknowledge support from CNRS, Institut Curie, Ministère de la Recherche (ACI grant no. DRAB04/117) and HFSP grant (no. RGP36/2005). Funding to pay the Open Access publication charges for this article was provided by Institut Curie.

Conflict of interest statement. None declared.

REFERENCES

1. Dahlberg,A.E. (2001) The ribosome in action. *Science*, **292**, 868–869.
2. Kruger,K., Grabowski,P., Zaug,A.J., Sands,J., Gottschling,D.E. and Cech,T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, **31**, 147–157.
3. Bartel,D.P. and Szostak,J.W. (1993) Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411–1418.
4. Joyce,G.F. (1989) Amplification, mutation and selection of catalytic RNA. *Gene*, **82**, 83–87.
5. Ellington,A.E. and Szostak,J.W. (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
6. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
7. Mironov,A.S., Gusarov,I., Rafikov,R., Lopez,L.E., Shatalin,K., Kreneva,R.A., Perumov,D.A. and Nudler,E. (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, **111**, 747–756.
8. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by metabolite binding mRNA. *Chem. Biol.*, **9**, 1043–1049.
9. Winkler,W.C. and Breaker,R.R. (2003) Genetic control by metabolite-binding riboswitches. *ChemBioChem.*, **4**, 1024–1032.
10. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism and function. *Cell*, **116**, 281–297.
11. Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11–17.
12. He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nature Rev. Genet.*, **5**, 522–531.
13. Winkler,W.C., Nahvi,A., Roth,A., Collins,J.A. and Breaker,R.R. (2004) *Nature*, **428**, 281–286.
14. Uhlenbeck,O.C. (1995) Keeping RNA happy. *RNA*, **1**, 4–6.
15. Boyle,J., Robillard,G. and Kim,S. (1980) Sequential folding of transfer RNA. A nuclear magnetic resonance study of successively longer tRNA fragments with a common 5′end. *J. Mol. Biol.*, **139**, 601–625.

16. Kramer,F. and Mills,D. (1981) Secondary structure formation during RNA synthesis. *Nucleic Acids Res*., **9**, 5109–5124.
17. Nussinov,R. and Tinoco,I.,Jr (1981) Sequential folding of a messenger RNA molecule. *J. Mol. Biol*., **151**, 519–533.
18. Biebricher,C.K. and Luce,R. (1992) *In vitro* recombination and terminal elongation of RNA by Qβ replicase. *EMBO J*., **11**, 5129–5135.
19. LeCuyer,K.A. and Crothers,D.M. (1994) Kinetics of an RNA conformational switch. *Proc. Natl Acad. Sci. USA*, **91**, 333–3377.
20. Gultyaev,A.P., van Batenburg,F.H. and Pleij,C.W. (1998) Dynamic competition between alternative structures in viroid RNAs simulated by an RNA folding algorithm. *J. Mol. Biol*., **276**, 43–55.
21. Repsilber,D., Wiese,S., Rachen,M., Schroder,A., Riesner,D. and Steger,G. (1999) Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel electrophoresis. *RNA*, **5**, 574–584.
22. Olsthoorn,R.C., Mertens,S., Brederode,F.T. and Bol,J.F. (1999) A conformational switch at the 3′ end of a plant virus RNA regulates viral replication. *EMBO J*., **18**, 4856–4864.
23. Putzer,H., Gendron,N. and Grunberg-Manago,M. (1992) Coordinate ex-pression of the two threonyl-tRNA synthetase genes in *Bacillus subtilis*: control by transcriptional antitermination involving a con-served regulatory sequence. *EMBO J*., **11**, 3117–3127.
24. Putzer,H., Condon,C., Brechemier-Baey,D., Brito,R. and Grunberg-Manago,M. (2002) Transfer RNA-mediated antitermination *in vitro*. *Nucleic Acids Res*., **30**, 3026–3033.
25. Wickiser,J.K., Winkler,W.C., Breaker,R.R. and Crothers,D.M. (2005) The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol. Cell*., **18**, 49–60.
26. Romby,P. and Springer,M. (2003) Bacterial translational control at atomic resolution. *Trends Genet*., **19**, 155–161.
27. Caillet,J., Nogueira,T., Masquida,B., Winter,F., Graffe,M., Dock-Bregeon,A.C., Torres-Larios,A., Sankaranarayanan,R., Westhof,E., Ehresmann,B. *et al.* (2003) The modular structure of *Escherichia coli* threonyl-tRNA synthetase as both an enzyme and a regulator of gene expression. *Mol. Microbiol*., **47**, 961–974.
28. Brunel,C., Romby,P., Sacerdot,C., de Smit,M., Graffe,M., Dondon,J., van Duin,J., Ehresmann,B., Ehresmann,C. and Springer,M. (1995) Stabilised secondary structure at a ribosomal binding site enhances translational repression in *E.coli*. *J. Mol. Biol*., **253**, 277–290.
29. Moller-Jensen,J., Franch,T. and Gerdes,K. (2001) Temporal translational control by a metastable RNA structure. *J. Biol. Chem*., **276**, 35707–35713.
30. van Meerten,D., Girard,G. and van Duin,J. (2001) Translational control by delayed RNAfolding: identification of the kinetic trap. *RNA*, **7**, 483–494.
31. de Smit,M.H. and van Duin,J. (2003) Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J. Mol. Biol*., **331**, 737–743.
32. Soukup,G.A. and Breaker,R.R. (1999) Engineering precision RNA molecular switches. *Proc. Natl Acad. Sci. USA*, **96**, 3584–3589.
33. Nagel,J.H. and Pleij,C.W. (2002) Self-induced structural switches in RNA. *Biochimie*, **84**, 913–923.
34. Groeneveld,H., Thimon,K. and van Duin,J. (1995) Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *RNA*, **1**, 79–88.
35. Gultyaev,A.P., van Batenburg,F.H. and Pleij,C.W. (1995) The influence of a metastable structure in plasmid primer RNA on antisense RNA binding kinetics. *Nucleic Acids Res*., **23**, 3718–3725.
36. Poot,R.A., Tsareva,N.V., Boni,I.V. and van Duin,J. (1997) RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proc. Natl Acad. Sci. USA*, **94**, 10110–10115.
37. Gerdes,K., Gultyaev,A.P., Franch,T., Pedersen,K. and Mikkelsen,N.D. (1997) Antisense RNA-regulated programmed cell death. *Annu. Rev. Genet*., **31**, 1–31.
38. Gultyaev,A.P., Franch,T. and Gerdes,K. (1997) Programmed cell death by hok/sok of plasmid R1: coupled nucleotide covariations reveal a phylogenetically conserved folding pathway in the hok family of mRNAs. *J. Mol. Biol*., **273**, 26–37.
39. Franch,T., Gultyaev,A.P. and Gerdes,K. (1997) Programmed cell death by hok/sok of plasmid R1: processing at the hok mRNA 3′-end triggers structural rearrangements that allow translation and antisense RNA binding. *J. Mol. Biol*., **273**, 38–51.
40. Nagel,J.H., Gultyaev,A.P., Gerdes,K. and Pleij,C.W. (1999) Metastable structures and refolding kinetics in hok mRNA of plasmid R1. *RNA*, **5**, 1408–1418.
41. Pan,T., Fang,X. and Sosnick,T. (1999) Pathway modulation, circular permutation and rapid RNA folding under kinetic control. *J. Mol. Biol*., **286**, 721–731.
42. Pan,T., Artsimovitch,I., Fang,X., Landick,R. and Sosnick,T.R. (1999) Biochemistry folding of a large ribozyme during transcription and the effect of the elongation factor NusA. *Proc. Natl Acad. Sci. USA*, **96**, 9545–9550.
43. Heilman-Miller,S.L. and Woodson,S.A. (2003) Perturbed folding kinetics of circularly permuted RNAs with altered topology. *J. Mol. Biol*., **328**, 385–394.
44. Heilman-Miller,S.L. and Woodson,S.A. (2003) Effect of transcription on folding of the Tetrahymena ribozyme. *RNA*, **9**, 722–733.
45. Xayaphoummine,A., Bucher,T., Thalmann,F. and Isambert,H. (2003) Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Nat Acad. Sci. USA*, **100**, 15310–15315.
46. Xayaphoummine,A., Bucher,T. and Isambert,H. (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res*., **33**, 605–610.
47. Adami,C. and Cerf,N.J. (2000) Physical complexity of symbolic sequences. *Phys. D*, **137**, 62–69.
48. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc*., **126**, 5130–5137.
49. Isambert,H. (2002) Voltage addressable nanomemories in DNA? *C. R. Physique*, **3**, 391–396.
50. Chamberlin,M. and Ring,J. (1973) Characterization of T7-specific ribonucleic acid polymerase I. General properties of the enzymatic reaction and the template specificity of the enzyme. *J. Biol. Chem*., **248**, 2235–2244.
51. Chamberlin,M. and Ring,J. (1973) Characterization of T7-specific ribonucleic acid polymerase II. Inhibitors of the enzyme and their application to the study of the enzymatic reaction. *J. Biol. Chem*., **248**, 2245–2250.
52. Viasnoff,V., Meller,A. and Isambert,H. (2006) DNA nanomechanical switches under folding kinetics control. *Nano Lett*., **6**, 101–104.
53. Reidys,C.M., Stadler,P.F. and Schuster,P. (1997) Generic properties of combinatory maps: neural networks of RNA secondary structures. *Bull. Math. Biol*., **59**, 339–397.
54. Flamm,C., Hofacker,I.L., Maurer-Stroh,S., Stadler,P.F. and Zehl,M. (2001) Design of multistable RNA molecules. *RNA*, **7**, 254–265.
55. Seeman,N.C. (1982) Nucleic acid junctions and lattices. *J. Theor. Biol*., **99**, 237–247.
56. Dirks,R.M., Lin,M., Winfree,E. and Pierce,N.A. (2004) Paradigms for computational nucleic acid design. *Nucleic Acids Res*., **32**, 1392–1403.
57. Isambert,H. and Siggia,E.D. (2000) Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl Acad. Sci. USA*, **97**, 6515.
58. Harlepp,S., Marchal,T., Robert,J., Léger,J.-F., Xayaphoummine,A., Isambert,H. and Chatenay,D. (2003) Probing complex RNA structures by mechanical force. *Eur. Phys. J. E*, **12**, 605–615.
59. Meyer,I.M. and Miklós,I. (2004) Co-transcriptional folding is encoded within RNA genes. *BMC Mol. Biol*., **5**, 10.
60. Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
61. Schuster,P., Fontana,W., Stadler,P.F. and Hofacker,I.L. (1994) From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. London B*, **255**, 279.