FEBS Letters

FEBS PRESS science publishing by scientists

Regulation of the ThiM riboswitch is facilitated by the trapped structure formed during transcription of the wild-type sequence

Chengyi Du[1], Yujie Wang[2] and Sha Gong[1]

1 Hubei Key Laboratory of Economic Forest Germplasm Improvement and Resources Comprehensive Utilization, Hubei Collaborative Innovation Center for the Characteristic Resources Exploitation of Dabie Mountains, Huanggang Normal University, China
2 Department of Physics and Telecommunication Engineering, Zhoukou Normal University, China

Correspondence
S. Gong, Hubei Key Laboratory of Economic Forest Germplasm Improvement and Resources Comprehensive Utilization, Hubei Collaborative Innovation Center for the Characteristic Resources Exploitation of Dabie Mountains, Huanggang Normal University, Huanggang 438000, China
Tel: +86 0713-8835186
E-mail: shagong@whu.edu.cn

(Received 18 June 2021, revised 17 September 2021, accepted 20 September 2021, available online 30 October 2021)

doi:10.1002/1873-3468.14202

Edited by Michael Ibba

The *ThiM* riboswitch from *Escherichia coli* is a typical mRNA device that modulates downstream gene expression by sensing TPP. The helix-based RNA folding theory is used to investigate its detailed regulatory behaviors in cells. This RNA molecule is transcriptionally trapped in a state with the unstructured SD sequence in the absence of TPP, which induces downstream gene expression. As a key step to turn on gene expression, formation of this trapped state (the genetic ON state) highly depends on the co-transcriptional folding of its wild-type sequence. Instead of stabilities of the genetic ON and OFF states, the transcription rate, pause, and ligand levels are combined to affect the *ThiM* riboswitch-mediated gene regulation, which is consistent with a kinetic control model.

Keywords: co-transcriptional folding; gene regulation; TPP riboswitch

RNAs play various cellular roles in organisms, including bacterial immunity [1,2], catalysis [3–5], genome maintenance, and regulation by forming different structures [2,6–10]. Gene regulation is an important function that implemented by a kind of structured noncoding mRNA domains, called riboswitches [11–15]. These natural genetic switches can control their downstream gene expression via sensing intracellular signals, such as temperature and metabolite concentrations [9,11,16–18]. Ligand-specific riboswitches are widely distributed among bacteria and other organisms, which often contain two domains: an aptamer and an expression platform. The metabolite-sensitive aptamer is responsible for ligand recognizing and binding, while the expression platform has a translation-related Shine-Dalgarno (SD) sequence, a transcription-related termination region or splice sites. Since the two functional units often partly overlap with each other, the structure folded within the aptamer directly influences the folding pattern of the second domain, which decides the fate of downstream gene expression.

To date, there are a dozen of riboswitch classes that have been identified [18], and riboswitches from the same class usually share many common features [9,16,18–21]. The Thiamine pyrophosphate (TPP) riboswitch is one of the earliest discovered riboswitch representatives, which are intensively studied. For this class, the aptamer domain often forms a similar three-way junction structure to specifically recognize TPP. After ligand docking, the bound aptamer forms a

Abbreviations
E. coli, *Escherichia coli*; MC, Monte Carlo; *N. crassa*, *Neurospora crassa*; NMR, nuclear magnetic resonance, nucleotide; RNAP, RNA polymerase; SD, Shine-Dalgarno; TPP, thiamine pyrophosphate.

compactly folded structure with two parallel stacking helices. In contrast to the highly conserved aptamer domain, the expression platform varies in sequence and structure, so TPP riboswitches can function at different levels. For example, the *NMT1* TPP riboswitch from *Neurospora crassa* regulates the expression of a protein involved in TPP metabolism by controlling alternative RNA splicing [22], while the *ThiM* TPP riboswitch from *Escherichia coli* (Fig. 1) modulates the expression of the enzyme hydroxyethylthiazole kinase through structural changes in the SD sequence [23].

Besides regulation of translation initiation, a recent study indicates that the *ThiM* riboswitch also controls premature transcription termination by using the Rho transcription factor [24]. In the presence of sufficient TPP, the *ThiM* riboswitch folds into the bound OFF state consisting of a long repression stem R and a bound aptamer arranged by a three-way junction (stems P1, P2, and P3). As stem R covering its SD sequence, formation of this structure will inhibit downstream gene expression by occluding the ribosome binding and then lead to Rho-dependent transcription termination [24]. Due to its stability and compact global shape, this ligand bound structure has been solved to high resolution by experimental approaches [25]. When TPP is at a low level, the *ThiM* riboswitch adopts an alternative structure (ON state) with the unstructured SD sequence, which is the only structural information directly hinted from reporter-gene assays [26]. Unlike the already solved bound functional state, the highly flexible ligand-free ON state has different proposed structural models. Winkler *et al.* [27] reported an ON structure model with the pseudoknot formed by P8* (Fig. 1). The high similarity between this model and OFF state can explain the comparable binding affinities of the isolated aptamer and the full-length riboswitch. Another structural model, organized by helices P2, P3a P4a, P4b, and P4c, is inferred via different binding affinities of the RNAs with different lengthens measured from the fluorescence spectroscopic experiment [28]. According to chemical and enzymatic probing data, Rentmeister *et al.* [23] also proposed a structural form, which mainly had helices P2 and P4. In addition, their team found that the GG/CC mutated riboswitch variant (Fig. 1) lost genetic control and acted as a constitutive-OFF switch [23,26].

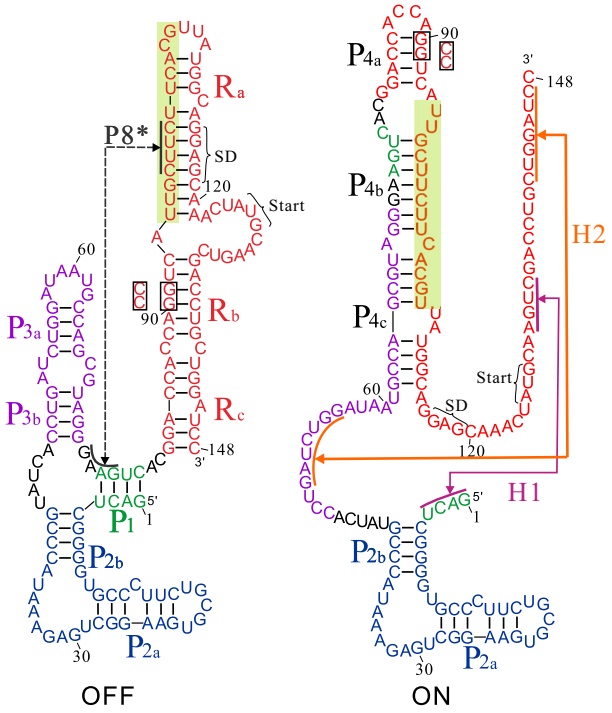

Fig. 1. Structural models of the *Escherichia coli ThiM* riboswitch. The helix region of P1, P2, P3, and the repression stem (R) are colored differently. The three-branch aptamer structure in OFF state can specifically bind TPP to form the bound OFF state. Long stems are divided by interloops and shown by subscript a, b, and c. Nucleotides in the green and black boxes denote the pausing region found in the experiment and the mutation that two Gs are substituted by two Cs. The nucleotides at the two ends of the arrows are formed helices P8*, H1, and H2. The SD sequence and translation start regions are shown.

The different structural characterization of ON state comes from an unclear understanding of the transcriptional folding, since formation of the functional states is linked to the transcription process *in vivo* [29]. Compared to *in vitro* RNA folding, it is more challenge to explore transcriptional folding, as it follows a sequential progression in intracellular environments [29–33]. Despite the extremely complicated cellular condition, the optical-trapping assay has been developed to successfully describe individual transcripts of *PbuE* riboswitch *in vitro* [34]. By incorporating native secondary structural information, the kinetic Monte Carlo (MC) simulation gives access to investigate co-transcriptional folding on the secondary structure level via comparison with experimental results to get the relationship between real time and MC steps [29]. Based on the similar parameters as kinetic MC, the helix-based RNA folding theory has been proven to be a reliable approach to study mechanisms of riboswitches and functions of other mRNAs [35–40]. Compared to the above methods, it is quite suitable to feature the incorporation of competing conformations and ligand-binding kinetics for riboswitches [36,38,41]. Here, we used this approach to investigate the regulation details and structural model of ON state via predicting transcriptional folding behaviors of the *ThiM* riboswitch. Our results suggested that the *ThiM* aptamer structure is formed but quickly broken by a long helix P4

during the transcription without TPP. At the cost of breaking P1 and P3 to fully form P4, a state consisting of helices P2, P4, and a small helix (H1 or H2), traps almost all the *ThiM* riboswitches. Although it is not the most stable ligand-free state, as the dwell time is long enough for translation, this long-trapped state with the flexible SD sequence becomes the genetic ON state of the riboswitch. In the presence of high ligand concentrations, docking of TPP to the aptamer will prevent the invasion of helix P4 and direct folding of the repression stem R within the express platform. By preventing this invasion, the GG/CC mutated riboswitch variant loses the function of gene regulation, since the ligand binding cannot induce structural changes in the SD sequence anymore [23,26]. During the refolding process, most of the full-length *ThiM* RNAs fold into OFF state without being trapped into ON state. These results imply that the riboswitch-mediated gene regulation depends on sequential folding of the wild-type *ThiM* sequence *in vivo*.

Materials and methods

To investigate RNA folding details, most studies focus on *in vitro* RNA folding (refolding), with a random coil of a fixed-length RNA under a given condition. Compared to the complex cellular environment, this folding scenario, where details are easy to describe, has been intensively used to study RNA folding behaviors and related mechanisms [32,42–49]. However, due to the sequential nature of RNA synthesis [30,31,33,50–54], functional mRNAs usually fold co-transcriptionally with growing chains *in vivo*. Besides the varied length, another obvious difference from refolding is the spatial restriction arising from the interaction with RNA polymerase (RNAP). Considering RNA residues covered in RNAP (forming RNA–DNA hybrid and locating in the exit channel) [31,55–56], only the outside nucleotides which locate further than about 15nt from the 3′-end, are free to form base pairs. At the same time, spatial constraint of the narrow exit channel allows the synthesized nucleotides to leave RNAP one by one. According to these characters, in our model every left nucleotide is treated as one transcriptional step, during which the length of RNA is fixed [41] To predict co-transcriptional RNA folding, we used similar procedures as that employed in refolding at every transcriptional step. In this section, we would briefly introduce the theory of RNA refolding kinetics, co-transcriptional folding kinetics, and the way to incorporate ligand-binding kinetics into the current model.

RNA refolding kinetics

To calculate RNA refolding kinetics, there are three main steps: generating conformational space, calculating transition rates between different states, and solving the master equation to get the populational kinetics [57–59]. In the helix-based RNA folding theory [40,60], states in the conformational space are constructed by compatible helices. For a given RNA, we first enumerated all possible helices based on its sequence and the principle of complementary base pairing. By using these helices as building blocks, all possible states which sample the conformational space are constructed, and their free energies are calculated according to the nearest neighbor model [61,62].

In the conformational space, whether one state can directly transit to another via one kinetic move (forming/disrupting a helix, or exchanging two helices) depends on their structural relationship. The transition rates are calculated based on the corresponding transition pathway.

If a state only has one more helix than another state except all the same helices, they will transit to each other by formation or disruption of the helix. For instance, state A with an open chain in Fig. 2A could fold to state B by forming a hairpin with several continuous stacks. Since the entropic cost of forming a stack near an existed stack is generally smaller than that of forming a loop-closing stack, the zipping pathway in Fig. 2B becomes the most probable way to form a helix. In most cases, three continuous stacks are enough to stabilize a loop. The landscape of the folding process along a zipping pathway often shows a downhill profile after the first three stacks are nucleated [59]. The helix folding rate k_f through the zipping pathway in Fig. 2 B therefore can be estimated to equal the rate of forming the first three stacks:

$$k_f = k_{a\to 1} K_1 \left(1 - K_2' K_1' \frac{1}{1 - K_2' K_1}\right)$$

K_i and K_i' in above equation are the forward and reverse probabilities for state i: $K_1 = k_{1\to 2}/\ (k_{1\to 2} + k_{1\to A})$, $K_1' = k_{1\to A}/(k_{1\to A} + k_{1\to 2})$, $K_2 = k_{2\to 3}/(k_{2\to 3} + k_{2\to 1})$, $K_2' = k_{2\to 1}/(k_{2\to 1} + k_{2\to 3})$, where $k_{i\to j}$ is the rate of the transition from state i to j, which involves formation or disruption of a stack [63]. This kind of transition is able to be described by a two-state transition process [64–67], and their kinetic rates are calculated as below: [57,64] $k = k_0 e^{\Delta G/k_B T}$. Where k_0 equals 6.6×10^{12} s^{-1} and 6.6×10^{13} s^{-1} for an AU and GC base pair, k_B is the Boltzmann constant and T is the temperature. The free energy barrier ΔG is estimated to be entropic and enthalpic changes (ΔS_{stack}, ΔH_{stack}) upon stack formation and disruption, respectively. If a stack closes a loop, the free energy barrier will include the entropic change of the loop (ΔS_{loop}). For example, the transition rate form state A to the intermediate state 1 in Fig. 2B is calculated as below:

$$k_{A\to 1} = k_0 e^{(\Delta S_{stack} + \Delta S_{loop})/k_B T}.$$

As the first stack can be formed anywhere inside of the paired region, the rate of the helix formation $k_{A\to B}$ equals

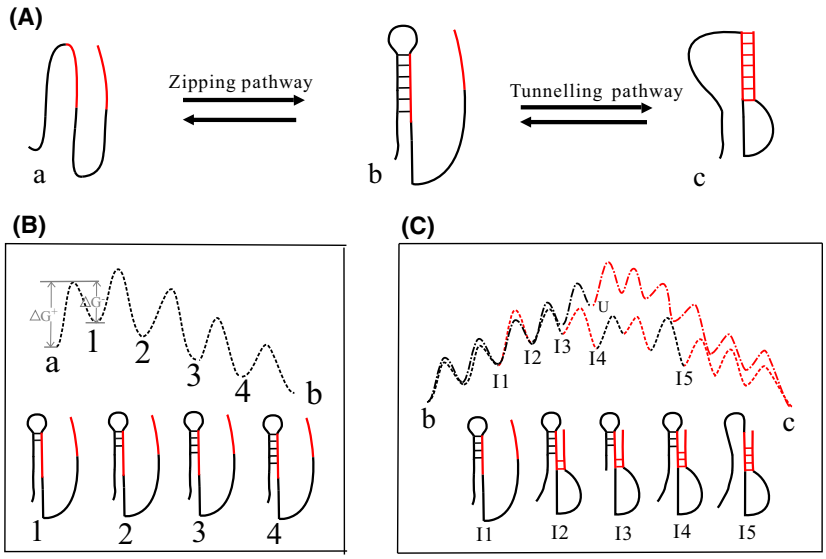

Fig. 2. Transitions between two states through formation (disruption) of a helix and exchange between two helices. The nucleotides in the paired region in (c) are colored red. The relevant pathways in (A) labeled along the arrow are shown in (B and C). The dotted lines denote the schematic energy landscape of zipping and tunneling pathways. The unfolding–refolding pathway is shown with dash-dot lines and formation of a stack in (C) on the landscape is colored red in (c). U is the unfolded, open chain. ΔG^+ and ΔG^- denoted the energy barriers for closing and opening a stack.

the sum of the rates along all zipping pathways with different loop-closing stacks. For the reversed transition, the rate $k_{B\to A}$ is calculated according to the detailed balance condition. If the free energy difference between state A and state B is ΔG_{AB}, the transition rates for formation and disruption of the helix will be summarized as below:

$$k_{A\to B} = \sum_{\text{pathway}} k_f$$

$$k_{B\to A} = k_{A\to B} e^{-\Delta G_{AB}/k_B T}. \quad (1)$$

The rate for helix exchange is calculated based on stack formation and disruption as well. For two overlapped helices, such as state B and state C in Fig. 2A, there are two kinds of transition pathways between them: the unfolding–refolding pathway and tunneling pathway (Fig. 2C). The unfolding–refolding pathway always yields a high energy barrier to completely unfold one helix and then form another helix. In fact, after breaking two stacks in state B to release two nucleotides, the first stack of the helix in state C can be closed (I2). At the following steps (I2–I5), disrupting one stack in B is followed by concurrently forming another stack in C. Formation of the rest helix region in C (from I5 to C) is a zipping process with a downhill profile on the landscape. This transition pathway has a lower energy barrier than that of the unfolding–refolding pathway, called the tunneling pathway. The rate through the tunneling pathway is calculated by Eqn (2) [59], where k_i and k_i' are the rate constants to disrupt (form) a stack in B (C).

$$k_{B\to C} = \frac{\prod_i^n k_i}{\sum_{j=0}^{n-1}\left(\prod_{i=1}^{j} k_i' \prod_{m=j+2}^{n} k_m\right)}$$

$$k_{C\to B} = k_{B\to C} e^{-\Delta G_{BC}/k_B T}. \quad (2)$$

To get the populational kinetics of each state, the master equation $dp_i(t)/dt = \sum_j \left[k_{j\to i} p_j(t) - k_{i\to j} p_i(t)\right]$ is first written in its matrix form of $d\mathbf{p}(t)/dt = \mathbf{M}\cdot\mathbf{p}(t)$, where $p_i(t)$ is the population of state i at time t and $\mathbf{p}(t)$ denotes the fractional populational vector. If n_m and $-\lambda_m$ are the m-th eigenvector and eigenvalue of the rate matrix $\mathbf{M}\left(M_{ij} = k_{i\to j}(i \neq j),\ M_{ii} = -\sum_{i\neq j} k_{i\to j}\right)$, the solution of the equation will be written as:

$$\mathbf{p}(t) = \sum_{m=1} C_m n_m e^{-\lambda_m t}. \quad (3)$$

The initial condition of RNA folding determines the coefficient C_m in the above equation. For *in vitro* folding, the initial state is the open, unfolded state U.

RNA co-transcriptional folding kinetics

In our model, RNA sequential folding during the transcriptional process is divided into many transcriptional steps, each of which corresponds to releasing one nucleotide by RNAP. The emerged RNA chain is one nucleotide shorter at the beginning of one step than that of the next step. That is to say, the nascent RNA chain is fixed-length at each step and its folding window is the time required for RNAP to release the new nucleotide. In a simple case, if the transcription is at a constant rate of v nucleotides per second ($\text{nt}\cdot\text{s}^{-1}$), the folding time of each step will be $(1/v)$ s.

Similar procedures are used to calculate the populational kinetics of co-transcriptional folding as that employed in refolding. Given that the emerged RNA chain is L-nt at step

N, the conformational space of the current step is first sampled by the L-nt RNA chain. Rates of transitions between states in the conformational space are calculated by using Eqns (1) and (2). The population relaxation within step N is obtained by solving the master equation. Here, the folding time is limited to $(1/v)$ s and the initial state may not be the open, unfolded chain. Unlike RNA refolding, the initial condition of step N is determined by the population distribution at the end of the previous step $(N-1)$. However, as a new nucleotide is released, some new structures which do not exist at step $N-1$ may be formed at step N. These structures have an initial population of zero at step N. If the new nucleotide adds to the 3′ tail or pairs with an upstream nucleotide to extend an existed helix, the initial populations of these structures are equal to their ending populations at step $N-1$. The population relaxation within step N is described by Eqn (3) and the populational kinetics of the whole transcription process is calculated by repeating these procedures from the first step to the last one.

To improve the computational efficiency, the transition node approximation is employed to decrease the calculated conformational space when the nascent RNA chain is longer than 120 nt [35,37,38]. Assuming that min$E$ is the highest free energy of the states with big initial populations (a cutoff population of 10%) at one step, the newly formed states with energies lower than (min$E$ + 3) kcal·mol^{-1} possibly have accumulated populations. From the 120-th step, if the states with energies lower than (min$E$ + 3) kcal·mol^{-1} are newly formed at the current step, we will search transitions between these new stable states and the existed states with big initial populations. Considering the time to transcribe the full-length *ThiM* riboswitch (around several seconds), the pathways of the above transitions with relevant rates of each kinetic move greater than 0.001 s^{-1} are remained as they may contribute the overall folding. Among all newly formed states, only the stable states and those located at the remained transition pathways are kept in the reduced conformation space. To improve the calculation efficiency, we first search the transition within 10 kinetic steps because most structures formed by the *ThiM* riboswitch have less than 10 helices. If no such transition pathway is remained, we will continue to search more kinetic moves to avoid isolate states.

Ligand-binding kinetics

Ligand binding to some ligand-competent structures can introduce tertiary interactions to stabilize the structures. When the ligand concentration is much higher than that of RNAs, the binding process could be described as a liner relation. In the presence of the ligand, the ligand bound states are first added to the conformational space. Supposing the association and dissociation rates are k_{on} and k_{off}, the extraenergy introduced by ligand binding is $\Delta G_{binding} = k_B T \ln(k_{on}[L]/k_{off})$, where [L] is the concentration of the ligand. The free energy of a bound state is equal to the sum of $\Delta G_{binding}$ and the energy of its secondary interactions. In the conformational space, rates of the transitions between bound states, which only involve secondary interaction rearrangements, are calculated from Eqns (1) and (2). Under assumption of the liner relation, the ligand-competent structure binds the ligand and forms the bound state with the effective binding rate $k_{eff} = k_{on}$ [L], and the reverse transition rate is k_{off}. For the *ThiM* riboswitch, these rate constants are taken from a recent study of Guedich *et al.* ($k_{on} = 8.7 \times 10^4$ M^{-1}·s^{-1}, $k_{off} = 2.5 \times 10^{-2}$ s^{-1}) [68], which are similar to those measured in the fluorescence spectroscopic experiment [28].

Results

The *ThiM* riboswitch is trapped in the genetic ON state during the transcription without TPP

The transcriptional folding of the *ThiM* riboswitch is first modeled at a typical transcription rate of 50 nt·s^{-1} [69]. From the calculated populational kinetics in Fig. 3A, the main folding pathway can be identified (Fig. 3B). As soon as the first 17 nucleotides emerge from RNAP, the nascent chain U folds into a hairpin structure C1 with a free energy of −2.54 kcal·mol^{-1} (Table 1). This folding is done within one transcriptional step, since forming a hairpin is much faster (around 10^4 s^{-1}) than transcribing a nucleotide [41,69,70]. With elongation of the RNA chain, this structure will be quickly replaced by the more stable state C2. By forming an interloop between the upstream nucleotides and the newly released nucleotides, state C2 transits to state C3 where P2 helix is fully folded. Form about step 65 when the upper part of P3 is nucleated to step 77, almost all the RNAs stay at the two-branch structure C4 with helices P2 and P3. At step 78, state C4 folds into the aptamer structure C5 by formation of the nonlocal helix P1 to close a three-way multiloop.

The ligand-competent structure C5 is the most stable state before step 92. When the upper part of P4 ($P4_a$) can be nucleated at step 92, state C5 folds to state C6 with a lower energy because of $P4_a$ formation. However, with more nucleotides are synthesized and released, helix P4 grows and invades into helix P1 in the aptamer structure. From step 101 when the stacks in $P4_b$ can be closed, the populations flow from state C6 to the more stable state C7. The structure of this state is similar to the structural model proposed from the fluorescence spectroscopic approach [28]. With gradual elongation of helix P4, it will invade into P2 helix from the bottom. When state C8 becomes more stable than C7 at step 115, the RNAs begin to transit

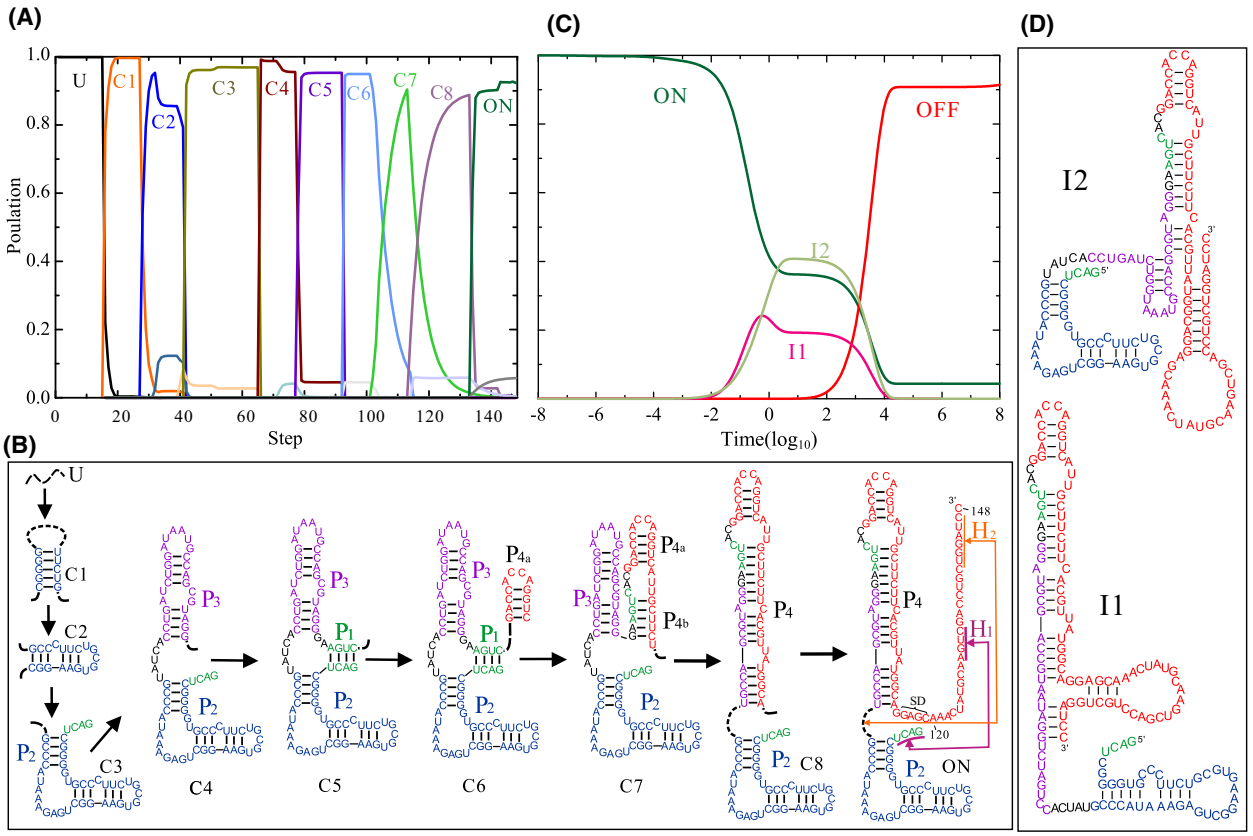

Fig. 3. The co-transcriptional folding behaviors of the *ThiM* riboswitch at an elongation rate of 50 nt·s^{-1} in the absence of TPP. The populational kinetics of intermediate states (a cutoff population of 5%) and the secondary structure of main states (a cutoff population of 15%) are shown in (A–D) describes the equilibration from ON state to OFF state and the structure of intermediates. The black dotted lines in (B) denote tails of a helix or a linker between two helices.

Table 1. Free energies of the intermediate states formed during the transcription (in unit of kcal·mol^{-1}).

State	C1	C2	C3	C4	C5	C6	C7	C8	ON	OFF
ΔG	−2.54	−3.49	−10.37	−19.93	−21.83	−27.0	−33.7	−36.66	−39.0	−44.5

from states C7 to C8. At step 130, state C8 has a population close to 90%, which is the genetic ON model reported by Rentmeister *et al.* [23].

With more nucleotides in the expression platform are released, some short helices can be formed. Among them, there are two helices (H1 and H2) compatible with all helices in state C8. They are formed by the 3′ end pairing with the 5′ end and the junction between helices P2 and P4, respectively (Fig. 3B). During the course of transcription, state C8 quickly transits into other two states by forming helices H1 or H2. Like state C8, a previously proposed ON model [23,54], these two states have the unstructured SD sequence. Besides, both of them share high similarities in stability and structural organization. Hence, they are treated as one state, called ON state here.

Our results suggested that, at the end of transcription, almost all the riboswitches folded into ON state. Even without the ligand, it would still equilibrate into OFF state due to the lower energy. The equilibrium time scale therefore is very important to understand switch function of the *ThiM* riboswitch. Supposing the initial state is ON state, the results in Fig. 3C,D showed that it requires at least 50 min for most of the RNAs to fold into OFF state. Considering the half-life of the *ThiM* mRNA (about 2 min) [24,71], this time interval is long enough for translation. It implies that ON state formed during the transcription is able to trap the riboswitch until the downstream gene is expressed. The intermediate, long-trapped state becomes the genetic ON state of the *ThiM* riboswitch.

The structural model of the unbound functional state (ON) here is different from the structural model proposed by Winkler *et al*. [27] Except the pseudoknot that frees the SD sequence, the structure of their model is almost the same as that of OFF state. The high structural similarity but difference in accessibility of the SD sequence can explain the ligand-dependent regulation and the comparable binding abilities of the isolate aptamer and the full-length RNA. However, based on our calculation for pseudoknots [40,72], this proposed structure cannot be formed as the pseudoknot is very unstable. As the full-length RNA favors OFF state, the comparable binding abilities of the two different constructs are from the same folding pattern in the aptamer domain. The main predicted folding behaviors agree with the results from the fluorescence experiments on transcriptional intermediate mimics [28], and the structural mode of ON state is the same as that proposed from experiments [23,54], except the small helices H1 and H2.

The efficiency of *ThiM* gene expression depends on the transcription process

The above results suggest that the *ThiM* riboswitch is trapped into ON state for a long time during the transcription without TPP. As the binding pocket is largely broken in this state, TPP binding will be significantly hampered if it is formed, which can be known from the fluorescence experiment [28]. Thus, although the riboswitch can sense TPP both co- and post-transcriptionally [24], its response to the intracellular TPP concentrations would be more efficient in a co-transcriptionally TPP sensing manner. In fact, most riboswitches are often transcribed in the presence of its metabolites.

Under a concentration of 100 μM TPP (Fig. 4A), most of the RNAs can bind to the ligand within ten transcriptional steps because of the great effective binding rate (8.7 s^{-1}). Formation of the bound aptamer $C5^b$ could lock the conformation within the aptamer domain and direct folding helix R within the expression platform. At the end of transcription, about 98% of the riboswitches fold into OFF^b state, which consists of a bound aptamer and stem R covering the SD sequence. Consistent with the experimental observation, the presence of sufficient TPP can trigger the switch from an activator to an inhibitor of gene expression [26,73]. Compared to that without the ligand (Fig. 3A), TPP binding prevents the invasion of helix P4, thereby avoiding formation of the long-trapped ON state. However, when the ligand concentration decreases to 10 μM, only about 37% of the riboswitches bind to TPP and then fold to OFF^b state through $C6^b$ (Fig. 4B). The rest riboswitches are unable to bind the ligand before helix P4 becomes stable under this ligand concentration. To effectively turn off gene expression, TPP is required to present at a sufficient level *in vivo*.

A low ligand level returns a small effective binding rate, which yields a long time to reach the binding equilibrium. Due to the sequential folding, the time window allowed for ligand binding is limited, which can be estimated from step 78 when the aptamer structure C5 is able to be formed to around step 108 when helix P1 is broken by helix P4. If this time interval ($30/v$ s) is much shorter than the time required for binding, most of the RNAs will not bind to the ligand, like that in Fig. 4B. Obviously, a small transcription rate leads to a long binding time window ($30/v$ s). When the transcription rate decreased to 15 nt·s^{-1}, the population of the bound state increased to about 75% near step 105 (Fig. 4C). But with elongation of the nascent RNA chain, about 5% of the bound RNAs fell apart. This is because the lower elongation rate also gives more time for TPP to dissociate from the bound aptamer, which is not as stable as that formed at high TPP levels.

Proper transcriptional pausing can provide extra time for ligand binding, which is found in several riboswitches [36,68,74,75]. Single-round pausing assays on the *ThiM* riboswitch revealed a major pause site located around the 95-th nucleotide [54]. Considering about ~15nt covered in RNAP and 1~2 nt RNAP backtracking [31,55,56], transcriptional pausing is presumed to occur at the 78th step for 10 s. Within the binding time window of ($30/v$ + 10) s, there are around 95% of the RNAs folding into the bound state even at a low concentration of 10 μM TPP (Fig. 4D). When the transcription is ended, the population of OFF^b state increases from about 37% (Fig. 4B) to 94% (Fig. 4D). This suggests that the pause site located directly after the aptamer domain can facilitate ligand binding and prevent formation of the unbound functional structure.

The biological function of the *ThiM* riboswitch is linked to its wild-type sequence and the intracellular environment

To induce downstream gene expression, the *ThiM* riboswitch should transcriptionally form the long-trapped ON state before equilibrating into the more stable OFF state, where the SD sequence is sequestered. Whether the riboswitch can form ON state or not decides whether it has the ability to act as a gene-

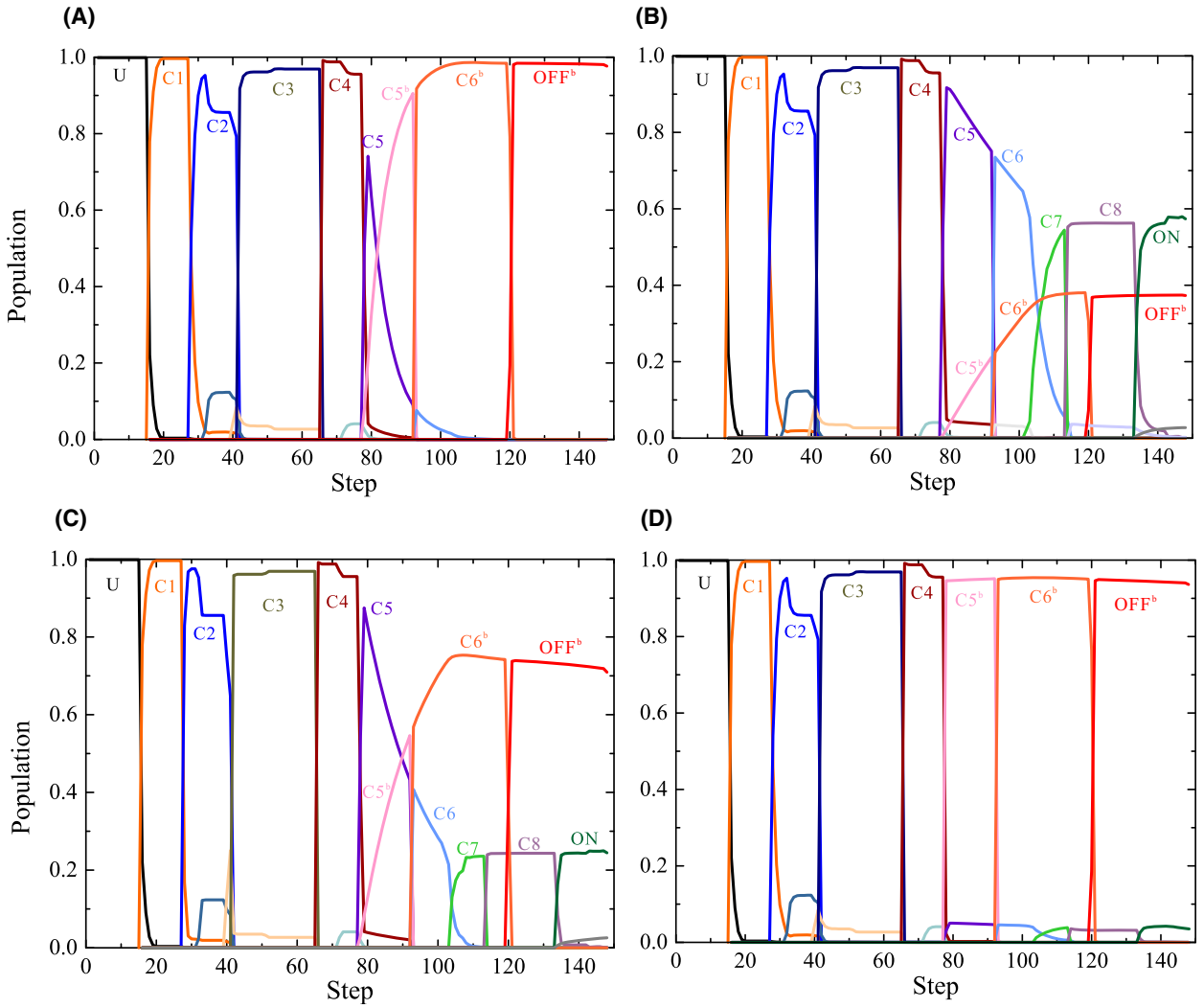

Fig. 4. The populational kinetics of the *ThiM* riboswitch during the transcription under different conditions: (A) 100 μM TPP, 50 nt·s^{-1}; (B) 10 μM TPP, 50 nt·s^{-1}; (C) 10 μM TPP, 15 nt·s^{-1}; (D) 10 μM TPP, 50 nt·s^{-1} with pausing at step 78 (10 s). The superscript 'b' denotes the corresponding state with TPP bound.

regulatory device. Different from that of co-transcriptional folding, ON state is unable to trap most of the full-length riboswitches during the refolding process (Fig. 5). It means that co-transcriptional folding significantly impacts how effectively the *ThiM* riboswitch functions.

The switch function of the *ThiM* riboswitch is not only greatly influenced by the intracellular environment, but also highly depends on its wild-type sequence. Reporter-gene assays suggested the GG to CC mutation in the expression platform (Fig. 1) can largely repress gene expression even without TPP [26]. For this lack of genetic control, structural probing provides a possible explanation, that the mutated construct adopts a structure similar to the secondary structure resembling the TPP-bound form of the wild-type riboswitch (OFF state) [23]. However, as co-transcriptional folding differs from *in vitro* RNA folding, whether this structure is the functional structure that represses translation needs more evidences from its co-transcriptional folding.

After the mutated nucleotides are free to form structures during the transcription process, the mutated RNA begins to exhibit different folding behaviors from that of the wild-type sequence (Fig. 6). From step 92 where the mutated site locates, the aptamer structure C5 successively folds into states S1, S2, and S3. With elongation of the RNA chain, a small part of RNAs transit from C5 to states S4 and S5 as soon as they can be formed near step 115. When the

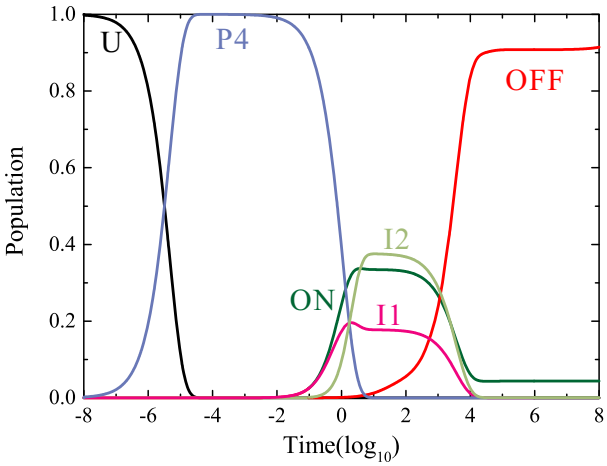

Fig. 5. The refolding populational kinetics of the *ThiM* riboswitch, where state P4 is the structure consisted by helix P4 shown in Fig. 3.

transcription is ended, there are about 15% of the RNAs folding through S5 (ΔG_{S5} = −25.14 kcal·mol^{-1}) to the more stable state S7 (ΔG_{S7} = −33.01 kcal·mol^{-1}). By adding a short helix, most of the rest RNAs transit from state S6 into S8, which is the most stable state at step 135 (ΔG_{S8} = −34.11 kcal·mol^{-1}). At the end of transcription, S7 and S8 occupy about 15% and 70% of the populations, respectively. Just like OFF[b] state, formation of state S8 sequesters the SD sequence via forming a helix. Thus, there are no TPP-induced structural alterations in the SD sequence for most of the mutated RNAs, in accordance with the experimental observation [26].

Although the mutated full-length RNA construct may finally adopt a structure similar to OFF state due to its high stability, formation of state S8 is the reason that causes the complete loss of regulation in cells. As the mutation breaks helix P4a in ON state, the

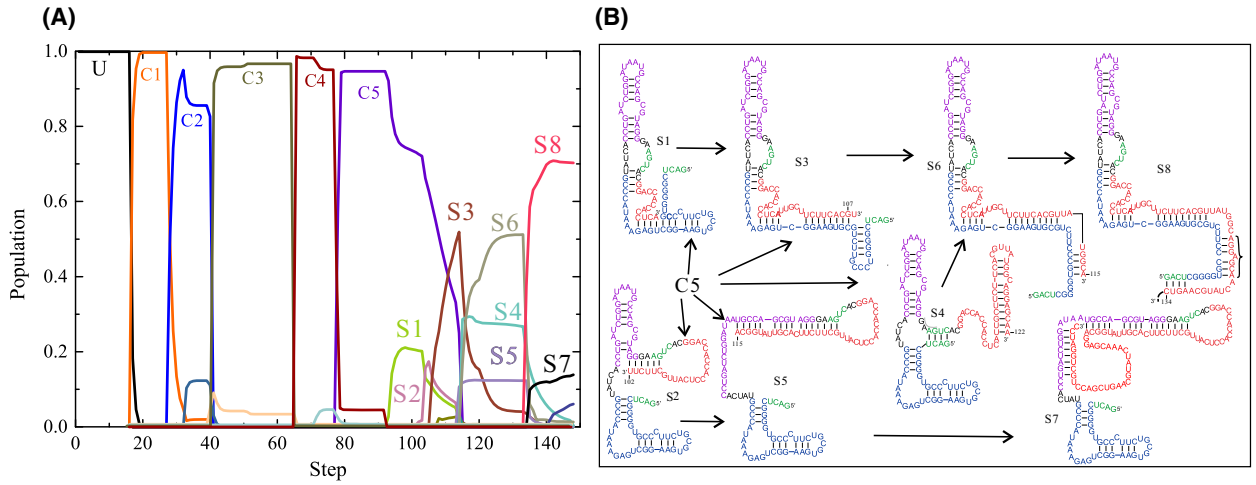

Fig. 6. The populational kinetics (A) and main intermediate structures formed during the transcription of the mutated *ThiM* riboswitch variant at an elongation rate of 50 nt·s^{-1}.

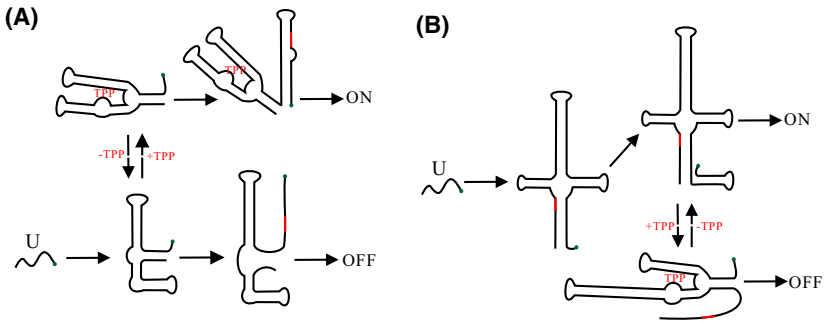

Fig. 7. Schematic illustration of conformational decision points during the synthesis of two TPP riboswitches: *ThiM* (A) and *NMT1* (B). The switchpoint of the *ThiM* riboswitch determines whether the aptamer structure binds to its ligand or is broken by formation of helix P4. The only decision point of the *NMT1* riboswitch is near the end of transcription, which involves the competition between the two functional states. The SD sequence, the splice site (for *NMT1* riboswitch), and 3′ end of RNA are colored red and green, respectively.

invasion of P4 helix into the aptamer structure is prevented during the course of transcription. Instead of ON state, the mutated riboswitches are co-transcriptionally trapped into state S8. In addition, the populations at most steps for the mutated riboswitch are not as concentrated in one or two main intermediate structures as that for the wild-type sequence. This implies that efficient co-transcriptional folding is one of evolutionarily selected properties in the sequence determination of the *ThiM* riboswitch.

Conclusion

As a kind of RNA-based gene devices, riboswitches regulate downstream gene expression by forming one of the two alternative functional states (unbound/bound) in response to specific signals. Like the *NMT1* TPP and *metF* riboswitches [35,76], the *ThiM* riboswitch has a well-known bound functional state but an unclear ligand-free functional state, which is difficult to refine by crystallography or NMR spectroscopy. According to the populational kinetics of states formed during the transcription, the helix-based RNA folding theory successfully reveals the TPP-induced switch mechanism and its functional structures. When TPP is present at a low level, full formation of the long helix P4 invades into the aptamer structure and unfolds helices P1 and P3. These structural rearrangements lead to formation of ON state with elongation of the RNA chain, consistent with the experimental observation that efficient translation at low TPP concentrations [24]. Conversely, the presence of sufficient TPP can lock the conformation within the aptamer domain by formation of the bound aptamer to prevent the invasion of helix P4. This competition between ligand binding and formation of helix P4 can be represented by the switchpoint in Fig. 7A, where one conformational decision takes place. As TPP binding is significantly hampered when the aptamer structure is seriously broken, the time window allowed for effective ligand binding is limited. Thus, both pausing at a strategic location, a general strategy to provide extra time for ligand binding, and the transcription rate, could affect the conformational choice made by the riboswitch at this decision point. If the switchpoint locates near the end of transcription, such as that in the *NMT1* TPP (Fig. 7B) and SAM-III riboswitches [35,76], the above factors will not have a big impact on their performances. For these riboswitches, ligand binding could perturb the equilibrium toward the bound functional state at the decision point. Accordingly, this kind of riboswitches, may exhibit some thermodynamic characters.

Formation of ON state is a valuable strategy and vital step to induce gene expression for the kinetically controlled *ThiM* riboswitch. It is not the most stable state, but has the ability to trap the RNA for a long time to ensure translation initiation. Any actions or conditions that break the potential to form ON state will result in a loss of gene regulation. For example, by preventing formation of ON state, the GG/CC mutated riboswitch variant behaves as a constitutive-OFF switch. Compared to *in vitro* folding where only a small proportion of the RNAs are trapped into ON state, the *ThiM* riboswitch functions better under an intracellular environment. All these suggested that the wild-type *ThiM* sequence has highly evolved to fulfill its role of genetic control *in vivo*.

Contrary to the complex intracellular environment, the energy parameters in the helix-based folding model are of RNAs at 1 M NaCl solution condition. The effects of cofactors, such as Mg^{2+}, on these parameters and tertiary interactions are neglected in the model. Even so, the folding behaviors of the *ThiM* riboswitch and other functional RNAs under this simple condition can provide a basis for the elucidation of their associated mechanisms *in vivo*. Hence, to reveal RNA folding details by using this theory or other similar methods is of fundamental importance to understand their intracellular role in a comprehensive way.

Acknowledgements

We acknowledge financial support from Huanggang science and technology project [grant no. XQYF2021000056], National Natural Science Foundation of China (grant no. 31600592, 12005321), and Hubei Key Laboratory of Economic Forest Germplasm Improvement and Resources Comprehensive Utilization, Hubei Collaborative Innovation Center for the Characteristic Resources Exploitation of Dabie Mountains (grant no. 201932003).

Author contributions

CD and SG conceived and supervised the project. CD and YW performed the calculations and wrote the paper. All authors have contributed in the improving of the manuscript, and read and approved its final version.

Data accessibility

The data that support the findings of this study are available in the figures/tables of this article.

References

1 Marraffini LA and Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181–190.
2 Seetin MG and Mathews DH (2012) RNA structure prediction : an overview of methods. *Methods Mol Biol* **905**, 99–122.
3 Strulson CA, Molden RC, Keating CD and Bevilacqua PC (2012) RNA catalysis through compartmentalization. *Nat Chem* **4**, 941–946.
4 Zhuang X (2000) A single-molecule study of RNA catalysis and folding. *Science* **288**, 2048–2051.
5 Yamada-Okabe T, Mio T, Kashima Y, Matsui M, Arisawa M and Yamada-Okabe H (1999) The *Candida albicans* gene for mRNA 5'-cap methyltransferase: identification of additional residues essential for catalysis. *Microbiology* **145**, 3023–3033.
6 Zhong G, Wang H, He W, Li Y, Mou H, Tickner ZJ, Tran MH, Ou T, Yin Y, Diao Y *et al.* (2020) A reversible RNA on-switch that controls gene expression of AAV-delivered therapeutics *in vivo*. *Nat Biotechnol* **38**, 169–175.
7 Strobel B, Spöring M, Klein H, Blazevic D, Rust W, Sayols S, Hartig JS and Kreuz S (2020) High-throughput identification of synthetic riboswitches by barcode-free amplicon-sequencing in human cells. *Nat Commun* **11**, 714.
8 Kierzek E and Kierzek R (2020) Conscious uncoupling of riboswitch functions. *J Biol Chem* **295**, 2568–2569.
9 Serganov A and Nudler E (2013) A decade of riboswitches. *Cell* **152**, 17–24.
10 DebRoy S, Gebbie M, Ramesh A, Goodson JR, van Cruz MR, Hoof A, Winkler WC and Garsin DA (2014) A riboswitch-containing sRNA controls gene expression by sequestration of a response regulator. *Science* **345**, 937–940.
11 Reining A, Nozinovic S, Schlepckow K, Buhr F, Fürtig B and Schwalbe H (2013) Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature* **499**, 355–359.
12 Chauvier A, Picard-Jean F, Berger-Dancause JC, Bastet L, Naghdi MR, Dubé A, Turcotte P, Perreault J and Lafontaine DA (2017) Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. *Nat Commun* **8**, 13892.
13 Miranda-Ríos J (2007) The THI-box riboswitch, or how RNA binds thiamin pyrophosphate. *Structure* **15**, 259–265.
14 Chen B, Zuo X, Wang Y-X and Dayie TK (2012) Multiple conformations of SAM-II riboswitch detected with SAXS and NMR spectroscopy. *Nucleic Acids Res* **40**, 3117–3130.
15 Mellin JR, Tiensuu T, Bécavin C, Gouin E, Johansson J and Cossart P (2013) A riboswitch-regulated antisense RNA in Listeria monocytogenes. *Proc Natl Acad Sci USA* **110**, 13132–13137.
16 Chen J and Gottesman S (2014) Riboswitch regulates RNA. *Science* **345**, 876–877.
17 Henkin TM (2008) Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* **22**, 3383–3390.
18 Breaker RR (2012) Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol* **4**, 1–15.
19 Deigan KE and FerrÉ-D'AmarÉ AR (2011) Riboswitches: discovery of drugs that target bacterial gene-regulatory RNAs. *Acc Chem Res* **44**, 1329–1338.
20 Tucker BJ and Breaker RR (2005) Riboswitches as versatile gene control elements. *Curr Opin Struct Biol* **15**, 342–348.
21 Delfosse V, Bouchard P, Bonneau E, Dagenais P, Lemay J-F, Lafontaine DA and Legault P (2010) Riboswitch structure: an internal residue mimicking the purine ligand. *Nucleic Acids Res* **38**, 2057–2068.
22 Cheah MT, Wachter A, Sudarsan N and Breaker RR (2007) Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* **447**, 497–500.
23 Rentmeister A, Mayer G, Kuhn N and Famulok M (2007) Conformational changes in the expression domain of the *Escherichia coli* thiM riboswitch. *Nucleic Acids Res* **35**, 3713–3722.
24 Bastet L, Chauvier A, Singh N, Lussier A, Lamontagne A-M, Prévost K, Massé E, Wade JT and Lafontaine DA (2017) Translational control and Rho-dependent transcription termination are intimately linked in riboswitch regulation. *Nucleic Acids Res* **45**, 7474–7486.
25 Serganov A, Polonskaia A, Phan AT, Breaker RR and Patel DJ (2006) Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441**, 1167–1171.
26 Mayer G, Raddatz M-SL, Grunwald JD and Famulok M (2007) RNA ligands that distinguish metabolite-induced conformations in the TPP riboswitch. *Angew Chem* **119**, 563–566.
27 Winkler W, Nahvi A and Breaker RR (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952–956.
28 Lang K, Rieder R and Micura R (2007) Ligand-induced folding of the thiM TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach. *Nucleic Acids Res* **35**, 5370–5378.
29 Lutz B, Faber M, Verma A, Klumpp S and Schug A (2014) Differences between cotranscriptional and free riboswitch folding. *Nucleic Acids Res* **42**, 2687–2696.
30 Zemora G and Waldsich C (2010) RNA folding in living cells. *RNA Biol* **7**, 634–641.
31 Lubkowska L, Maharjan AS and Komissarova N (2011) RNA folding in transcription elongation complex. *J Biol Chem* **286**, 31576–31585.

32 Herschlag D (1995) RNA chaperones and the RNA folding problem. *J Biol Chem* **270**, 20871–20874.
33 Schroeder R, Grossberger R, Pichler A and Waldsich C (2002) RNA folding *in vivo*. *Curr Opin Struct Biol* **12**, 296–300.
34 Frieda KL and Block SM (2012) Direct observation of cotranscriptional folding in an adenine riboswitch. *Science* **338**, 397–400.
35 Gong S, Du C and Wang Y (2020) Regulation of the thiamine pyrophosphate (TPP)-sensing riboswitch in NMT1 mRNA from *Neurospora crassa*. *FEBS Lett* **594**, 625–635.
36 Gong S, Wang Y and Zhang W (2015) Kinetic regulation mechanism of pbuE riboswitch. *J Chem Phys* **142**, 015103.
37 Gong S, Wang Y and Zhang W (2015) The regulation mechanism of yitJ and metF riboswitches. *J Chem Phys* **143**, 045103.
38 Wang Y, Wang Z, Liu T, Gong S and Zhang W (2018) Effects of flanking regions on HDV cotranscriptional folding kinetics. *RNA* **24**, 1229–1240.
39 Chen J and Zhang W (2012) Kinetic analysis of the effects of target structure on siRNA efficiency. *J Chem Phys* **137**, 225102.
40 Chen J, Gong S, Wang Y and Zhang W (2014) Kinetic partitioning mechanism of HDV ribozyme folding. *J Chem Phys* **140**, 025102.
41 Zhao P, Zhang W and Chen S-J (2011) Cotranscriptional folding kinetics of ribonucleic acid secondary structures. *J Chem Phys* **135**, 245101.
42 Wolfinger MT and Svrcek-seiler WA (2004) Exact folding dynamics of RNA secondary structures. *J Phys A Math Gen* **37**, 4731–4741.
43 Chadalavada DM, Senchak SE and Bevilacqua PC (2002) The folding pathway of the genomic hepatitis delta virus ribozyme is dominated by slow folding of the pseudoknots 1 1Edited by J. Doudna. *J Mol Biol* **317**, 559–575.
44 Stoddard CD, Gilbert SD and Batey RT (2008) Ligand-dependent folding of the three-way junction in the purine riboswitch. *RNA* **14**, 675–684.
45 Lemay J-F, Penedo JC, Tremblay R, Lilley DMJ and Lafontaine DA (2006) Folding of the adenine riboswitch. *Chem Biol* **13**, 857–868.
46 Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS and Tacker M (1994) Fast folding and comparison of RNA secondary structures. *Monatsh Chem* **188**, 167–188.
47 Geis M, Flamm C, Wolfinger MT, Tanzer A, Hofacker IL, Middendorf M, Mandl C, Stadler PF and Thurner C (2008) Folding kinetics of large RNAs. *J Mol Biol* **379**, 160–173.
48 Poot RA, Tsareva NV, Boni IV and van Duin J (1997) RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proc Natl Acad Sci USA* **94**, 10110–10115.
49 Lin JC, Hyeon C and Thirumalai D (2012) RNA under tension: Folding landscapes, kinetic partitioning mechanism, and molecular tensegrity. *J Phys Chem Lett* **3**, 3616–3625.
50 Pan T, Artsimovitch I, Fang X-W, Landick R and Sosnick TR (1999) Folding of a large ribozyme during transcription and the effect of the elongation factor NusA. *Proc Natl Acad Sci USA* **96**, 9545–9550.
51 Zhang L, Bao P, Leibowitz MJ and Zhang Y (2009) Slow formation of a pseudoknot structure is rate limiting in the productive co-transcriptional folding of the self-splicing *Candida intron*. *RNA* **15**, 1986–1992.
52 Wong TN, Sosnick TR and Pan T (2007) Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc Natl Acad Sci USA* **104**, 17995–18000.
53 Boyle J, Robillard GT and Kim S-H (1980) Sequential folding of transfer RNA. *J Mol Biol* **139**, 601–625.
54 Wong TN and Pan T (2009) RNA folding during transcription: protocols and studies. *Methods Enzymol* **468**, 167–193.
55 Komissarova N, Becker J, Solter S, Kireeva M and Kashlev M (2002) Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. *Mol Cell* **10**, 1151–1162.
56 Komissarova N and Kashlev M (1998) Functional topography of nascent RNA in elongation intermediates of RNA polymerase. *Proc Natl Acad Sci USA* **95**, 14699–14704.
57 Zhang W and Chen S-J (2006) Exploring the complex folding kinetics of RNA hairpins: I. General folding kinetics analysis. *Biophys J* **90**, 765–777.
58 Zhang W and Chen S-J (2006) Exploring the complex folding kinetics of RNA hairpins: II. Effect of sequence, length, and misfolded states. *Biophys J* **90**, 778–787.
59 Zhao P, Zhang W-B and Chen S-J (2010) Predicting secondary structural folding kinetics for nucleic acids. *Biophys J* **98**, 1617–1625.
60 Zhao Y, Huang Y, Gong Z, Wang Y, Man J and Xiao Y (2012) Automated and fast building of three-dimensional RNA structures. *Sci Rep* **2**, 734.
61 Theimer CA, Wang Y, Hoffman DW, Krisch HM and Giedroc DP (1998) Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot. *J Mol Biol* **279**, 545–564.
62 Mathews DH, Sabina J, Zuker M and Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911–940.
63 Xu X, Yu T and Chen SJ (2016) Understanding the kinetic mechanism of RNA single base pair formation. *Proc Natl Acad Sci USA* **113**, 116–121.

64 Wang Y-J, Wang Z, Wang Y-L and Zhang W-B (2017) Molecular dynamic simulation of the thermodynamic and kinetic properties of nucleotide base pair. *Chinese Phys B* **26**, 128705.
65 Wang Y, Gong S, Wang Z and Zhang W (2016) The thermodynamics and kinetics of a nucleotide base pair. *J Chem Phys* **144**, 115101.
66 Wang Y, Liu T, Yu T, Tan Z-J and Zhang W (2020) Salt effect on thermodynamics and kinetics of a single RNA base pair. *RNA* **26**, 470–480.
67 Wang Y, Wang Z, Wang Y, Liu T and Zhang W (2018) The nearest neighbor and next nearest neighbor effects on the thermodynamic and kinetic properties of RNA base pair. *J Chem Phys* **148**, 045101.
68 Guedich S, Puffer-Enders B, Baltzinger M, Hoffmann G, Da Veiga C, Jossinet F, Thore S, Bec G, Ennifar E, Burnouf D *et al*. (2016) Quantitative and predictive model of kinetic regulation by *E. coli* TPP riboswitches. *RNA Biol* **13**, 373–390.
69 Vogel U and Jensen KF (1997) NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA. *J Biol Chem* **272**, 12265–12271.
70 Pan T and Sosnick T (2006) RNA folding during transcription. *Annu Rev Biophys Biomol Struct* **35**, 161–175.
71 Caron M-P, Bastet L, Lussier A, Simoneau-Roy M, Masse E and Lafontaine DA (2012) Dual-acting riboswitch control of translation initiation and mRNA decay. *Proc Natl Acad Sci USA* **109**, E3444–E3453.
72 Gong S, Wang YY, Wang Z, Wang Y and Zhang W (2018) Genetic regulation mechanism of the yjdF riboswitch. *J Theor Biol* **439**, 152–159.
73 Lünse CE, Scott FJ, Suckling CJ and Mayer G (2014) Novel TPP-riboswitch activators bypass metabolic enzyme dependency. *Front Chem* **2**, 1–8.
74 Perdrizet GA, Artsimovitch I, Furman R, Sosnick TR and Pan T (2012) Transcriptional pausing coordinates folding of the aptamer domain and the expression platform of a riboswitch. *Proc Natl Acad Sci USA* **109**, 3323–3328.
75 Wickiser JK, Winkler WC, Breaker RR and Crothers DM (2005) The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell* **18**, 49–60.
76 Gong S, Wang Y, Wang Z, Wang Y and Zhang W (2016) Reversible-switch mechanism of the SAM-III riboswitch. *J Phys Chem B* **120**, 12305–12311.