doi:10.1016/j.jmb.2009.08.062 J. Mol. Biol. (2009) 393, 993–1003

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Analysis of Riboswitch Structure and Function by an Energy Landscape Framework

Giulio Quarta[1]†, Namhee Kim[1]†, Joseph A. Izzo[1] and Tamar Schlick[1,2]*

[1]*Department of Chemistry, New York University, 251 Mercer Street, New York, NY 10012, USA*

[2]*Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA*

Received 29 June 2009; received in revised form 15 August 2009; accepted 27 August 2009
Available online 3 September 2009

Edited by D. E. Draper

The thiamine pyrophosphate (TPP) riboswitch employs modular domains for binding TPP to form a platform for gene expression regulation. Specifically, TPP binding triggers a conformational switch in the RNA from a transcriptionally active "on" state to an inactive "off" state that concomitantly causes the formation of a terminator hairpin and halting of transcription. Here, clustering analysis of energy landscapes at different nucleotide lengths suggests a novel computational tool for analysis of the mechanics of transcription elongation in the presence or absence of the ligand. Namely, we suggest that the riboswitch's kinetics are tightly governed by a length-dependent switch, whereby the energy landscape has two clusters available during transcription elongation and where TPP's binding shifts the preference to one form. Significantly, the biologically active and inactive structures determined experimentally matched well the structures predominant in each computational set. These clustering/structural analyses combined with modular computational design suggest design principles that exploit the above features to analyze as well as create new functions and structures of RNA systems.

Keywords: thiamine pyrophosphate (TPP) riboswitch; aptamer; energy landscape; cluster analysis; gene regulation

Introduction

Riboswitches regulate gene expression by exploiting the effects of selective binding of small molecules on secondary and tertiary structural changes.[1,2] In prokaryotes, these modular RNA elements are normally found in the 5′ untranslated regions (UTRs) of genes and affect expression levels through various mechanisms: formation or destruction of transcription terminator hairpins,[3–5] sequestration of ribosome binding sites,[6] or emergence of alternative cleavage sites.[7–9]

The thiamine pyrophosphate (TPP) riboswitch is a prominent example. This riboswitch can adapt two structures, depending on whether it is bound or unbound to TPP.[10,11] TPP can bind only when the secondary structure that links the aptamer domain to the expression platform, termed *thi-box*, forms, as shown in Fig. 1a. However, it is currently not known how this riboswitch achieves such specific structural changes without the aid of proteins. Ligand binding triggers a conformational switch in the entire RNA from a transcriptionally active "on" state to an inactive "off" state that causes the formation of a terminator hairpin. Due to a high activation barrier between these states, we suggest that this "conformational switch" is actually regulated by a thermodynamic/kinetic mechanism whereby the presence or absence of the ligand favors one particular folding pathway and shifts the folded state from one structural conformational to the other. That is, if the ligand is absent, the anti-terminator conformation is favored and transcription of the downstream gene is turned on (Fig. 1b); when TPP is present, that cascade of events is hampered due to the terminator hairpin configuration. Previous studies on the function of FMN and TPP terminator hairpin riboswitches in *Bacillus subtilis*[11] suggested a model for small-molecule binding to nascent RNA riboswitches, whereby a decision between two alternative structures is likely to be made during transcription. Fluorescent experiments of the full-length *thiM* TPP riboswitch led to the conclusion that little to no structural changes occur upon TPP binding,[12] while

*Corresponding author. Department of Chemistry, New York University, 251 Mercer Street, New York, NY 10012, USA. E-mail address: schlick@nyu.edu.

† G.Q. and N.K. contributed equally to this work.

Present address: G. Quarta, NYU School of Medicine, 550 First Avenue, New York, NY 10016, USA.

Abbreviations used: TPP, thiamine pyrophosphate; UTR, untranslated region; mfe, minimum free energy.

0022-2836/$ - see front matter

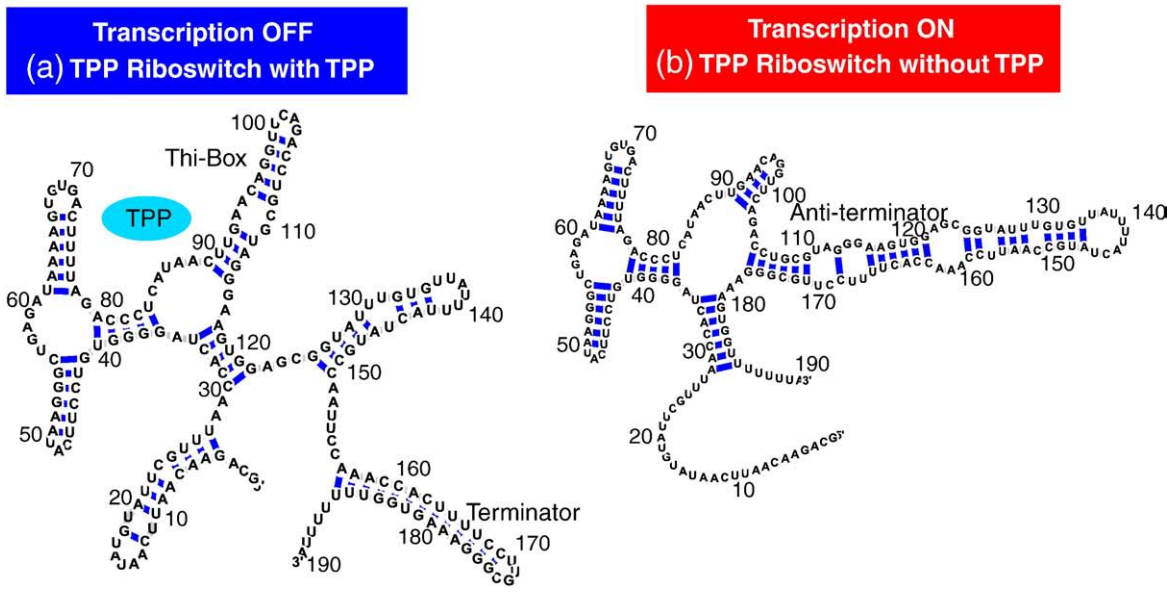

Fig. 1. Two alternative structures of TPP riboswitch in *B. subtilis*. (a) Full-length riboswitch when TPP present. *Thi-box* and terminator are formed (lowest-energy structure predicted by Mfold). (b) TPP riboswitch when TPP is absent. Anti-terminator is paired with terminator [lowest-energy structure with the constraint that prohibits base pairing in terminator stem–loop in (a)].

shorter riboswitch intermediates displayed competing alternative folds.

In this work, we examine this structural-equilibrium regulation in the TPP riboswitch as well as related systems using various computational tools. Our investigations show that the predominant structure in the nascent RNA is conformationally sensitive to the binding of TPP, which is present in the RNA up to a specific length (170 nt), after which the more stable structure is the anti-terminator structure. This mechanism points to a "switching threshold" for the riboswitch folding pathway, where one functional state is favored over the other at a critical length of the RNA. Our structural and thermodynamic analyses during transcription elongation help interpret how the presence or absence of the ligand regulates this conformation-dependent gene expression process and thus suggest an application of the tools offered here to RNA design.

Results

In the first subsection, we describe simulation results of the transcription folding process by structural and clustering analysis. These analyses help interpret key structural changes occurring when TPP binds and transcription terminates. For structural analysis, intermediate sequences are folded into their native secondary structures by free-energy minimization. For thermodynamic analysis, we sample alternative secondary structures and plot the distance between the minimum free energy (mfe) structure and all other structures against the free energy of folding to illustrate the folding pathways, which we call the energy landscape plot. Using a clustering algorithm, we partition the entire set of structures into their respective clusters and analyze the most dominant secondary structural features in the clusters (see details in Materials and Methods). These results are summarized in Table 1.

Table 1. Riboswitches and energy landscape clustering

Riboswitch name (reference)	Ligand	Number of clusters in structure set	Proportion of structures in termination set (%)
tenA, *B. subtilis*[11]	TPP (Vit. B_1)	2	85
Mutant 118	TPP	1	~100
Mutant 30	TPP	2	39.8
Mutant 80	TPP	2	87
Mutant 97	TPP	2	8.1
ribD, *B. subtilis*[11]	Flavin mononucleotide (FMN)	2	85
ypaA, *B.subtilis*[13,14]	FMN	2	77
gcvT, *B. subtilis*[15]	Glycine	1	~100
VCI-II, *Vibrio cholerae*[15]	Glycine	1	51
btuB, *E. coli*[16]	Coenzyme B_{12} (AdoCbl)	2	92.3
thiM, *E. coli*[4]	TPP	2	35.1

Clusters in each energy landscape are determined using the *k*-means algorithm (see Materials and Methods). Proportions of structures are based on the set containing the transcriptionally inactive, ligand-binding structures.

Switches of TPP riboswitch at 170 nt (State 1) and at 175 nt (State 2)

Structural predictions for the TPP riboswitch from 50 nt to the full-length 190-nt sequence were performed in 5-nt increments using secondary-structure prediction tools such as Mfold,[17] as described in Materials and Methods. The *thi-box* domain 5′ start site begins at 77 nt relative to the transcription start site and ends at 122 nt.[18] At each length, we "folded" the subsequence into its mfe structure for comparison. As Fig. 2 shows, from 125

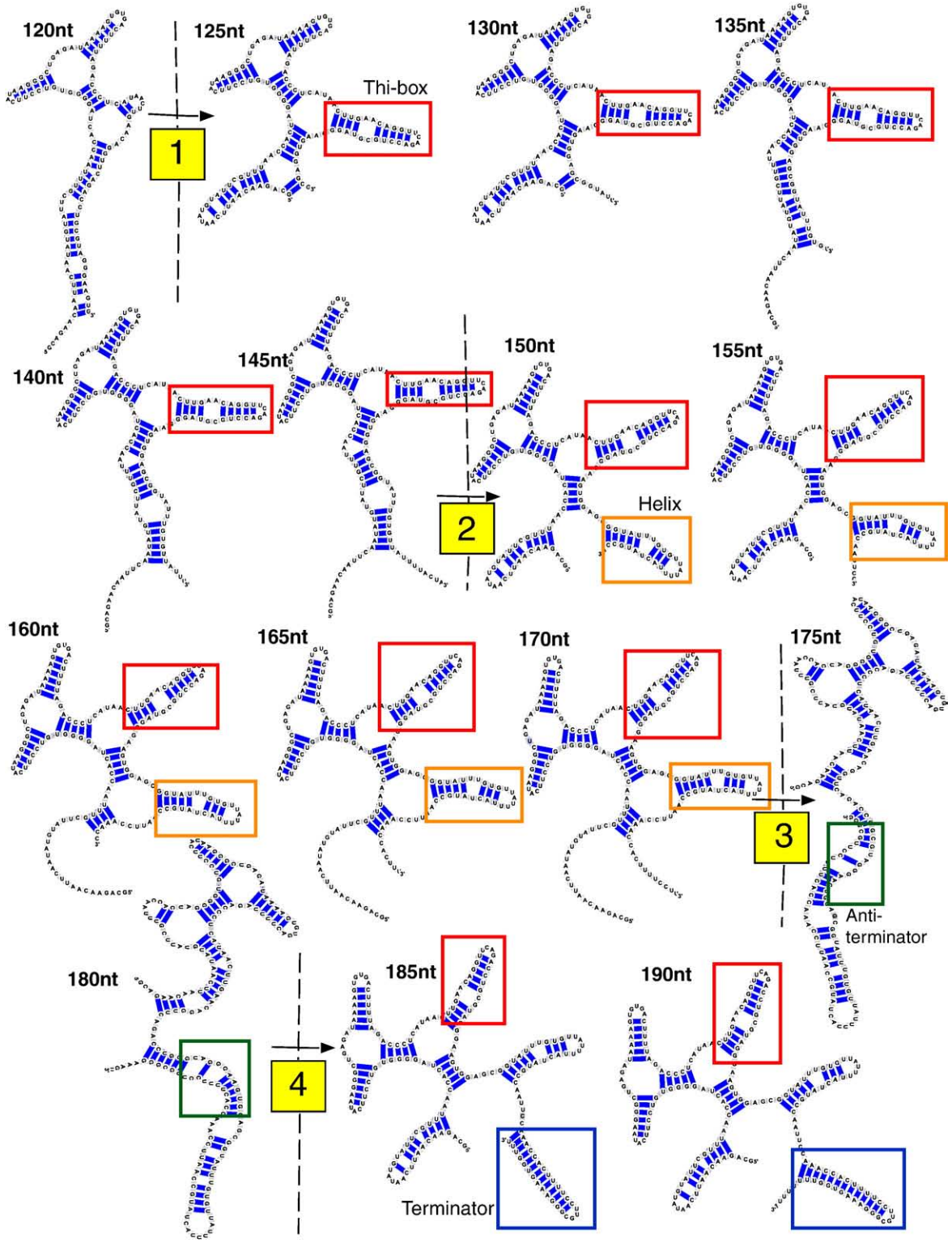

Fig. 2. TPP riboswitch folding pathway snapshots from 120 to 190 nt (full length) with the 5-nt incremental step. Note four large-scale structural changes (120 to 125 nt, 145 to 150 nt, 170 to 175 nt, and 180 to 185 nt).

until 170 nt, the RNA riboswitch forms a structure that would permit ligand binding since the *thi-box* is fully formed. Based on the availability and energetic favorability of this ligand-binding domain, we propose a model where complete transcription of the TPP binding region (121) remains intact up until 170 nt. During this length interval, the riboswitch can sense or detect the presence of intracellular ligand. Preorganization of the ligand-binding domain has been confirmed through force-based spectroscopy[19] and crystallographic evidence.[2]

If the ligand is present at sufficiently high concentrations for binding, the ligand-bound configuration of the RNA (Fig. 1a) is stabilized, and thus, formation of alternative structures is prevented. Support for this theory comes from the high specificity and strength of binding between the *thi-box* domain and ligand[4] as well as a similar model for an FMN-sensitive riboswitch.[11]

At a particular transcriptional length, the non-ligand-bound state can become energetically favored, and a concomitant misfolding of the *thi-box* results. Past 175 nt, the favored RNA structure has different topology (Fig. 2) and corresponds to that which forms the anti-terminator hairpin. This configuration may arise in TPP-poor environments due to the lack of stability provided by ligand binding.

Clustering of energy landscape of TPP riboswitches: Two groups or one group

We evaluate the feasibility of the above model by performing a series of energy landscape plots at varying transcriptional lengths. We generate the energy landscape plots for the TPP riboswitch from 50 to 190 nt in 5- or 1-nt increments. Using clustering techniques (see Materials and Methods), we found that the entire space of the sequence displays two distinct clusters in the energy plot (Fig. 3). As shown in Fig. 3a, the RNA forms two dominant clusters at the two end sections of the elongation process (120–124 nt and 181–190 nt), and these represent the two riboswitch conformations. The energy plot of 145 nt shows one cluster (Fig. 3b), the majority of which folds into a helical structure with an open *thi-box* (see Fig. 3b representative structure). On the other hand, the energy landscape at 190 nt shows two distinct clusters with two different structures (Fig. 3c), corresponding to the transcriptionally active and inactive structures. The structures in the smaller-distance cluster closely resemble the transcriptionally inactive structure, where the *thi-box* domain has formed along with the downstream transcription terminator hairpin. Conversely, the structures in the larger-distance cluster resemble the riboswitch in the anti-terminator form, which permits transcription to proceed. Thus, the presence or absence of TPP shifts the balance from one configuration over another.

The average structures in each cluster agree well with their respective RNA secondary structures found *in vitro* in the TPP-bound and TPP-free state.[10] Each cluster is structurally homogeneous (88% and 84%, respectively), but the difference between the two clusters is large (base-pair distance between cluster centers: 48.3). A significant energetic barrier is evident by a low-density area of structures between the two peaks (Fig. 3c, energy landscape plot). This suggests that only two global topologies are permissible for the folded riboswitch, each of which is represented by a cluster on the landscape.

Cluster analysis of mutants

To test this hypothesis further, we analyzed a set of mutants constructed by the Nudler group[11] with noted varying effects on the efficiency of transcription termination *in vitro*. A series of consecutive mutations inserted into the putative anti-terminator region of mutant sequence 118 (G118C, T119A, G120C, and G121C) increases termination efficiency up to ~100% (Table 1).

Thermodynamic analysis of mutant 118 showed a continuous energy funnel with no energy barrier but high structural similarity (81.1%) when treated as a single cluster. This contrasts the wild-type sequence that displays two clusters (Fig. 3c). As predicted, the terminator hairpin was present in all structures sampled in the set (Fig. 4a).

Mutant sequence 30 (C30G, C31G, A32T, and C33G) was designed to allow extensive base pairing between the *thi-box* (77–122) and mutated upstream bases, thus disrupting the TPP-binding domain. The energy landscape of mutant 30 displays two clusters with little separation, both representing the terminator hairpin state; as predicted, none of the structures displayed the open *thi-box* domain (Fig. 4b).

Mutants 80 (C80A, C81A, and C82A) and 97 (G97C, G98C, and T99A) have point mutations of the *thi-box* domain that disable the riboswitch's ability to bind to a ligand. Clustering confirms the presence of two structural clusters, extensively overlapping (Fig. 4c and d), yet with high intra-cluster structural similarity (76% and 82%, respectively). Similarly, the effect of TPP on transcription termination was completely abolished in mutant 97. Mutant 97 displays a two-cluster landscape predominantly composed (91.9%) of the low-energy family of anti-terminator structures (Fig. 4d).

Crystallographic studies show that binding between the *thi-box* and TPP is dependent on an induced-fit mechanism between highly conserved residues, divalent cations, and the moieties of TPP.[20–22] These interactions were disrupted in mutations 80 and 97, whose substitutions had deleterious effects on the ability of TPP to favor one of the configurations (*in vitro*, both mutants showed a +2% termination efficiency increase in the presence of TPP; wild type showed +71%). However, 2D folding analyses for these mutants display two-cluster energy landscapes, which suggest preservation of function to some degree. Thus, this mutation has a deleterious effect on binding but not on folding.

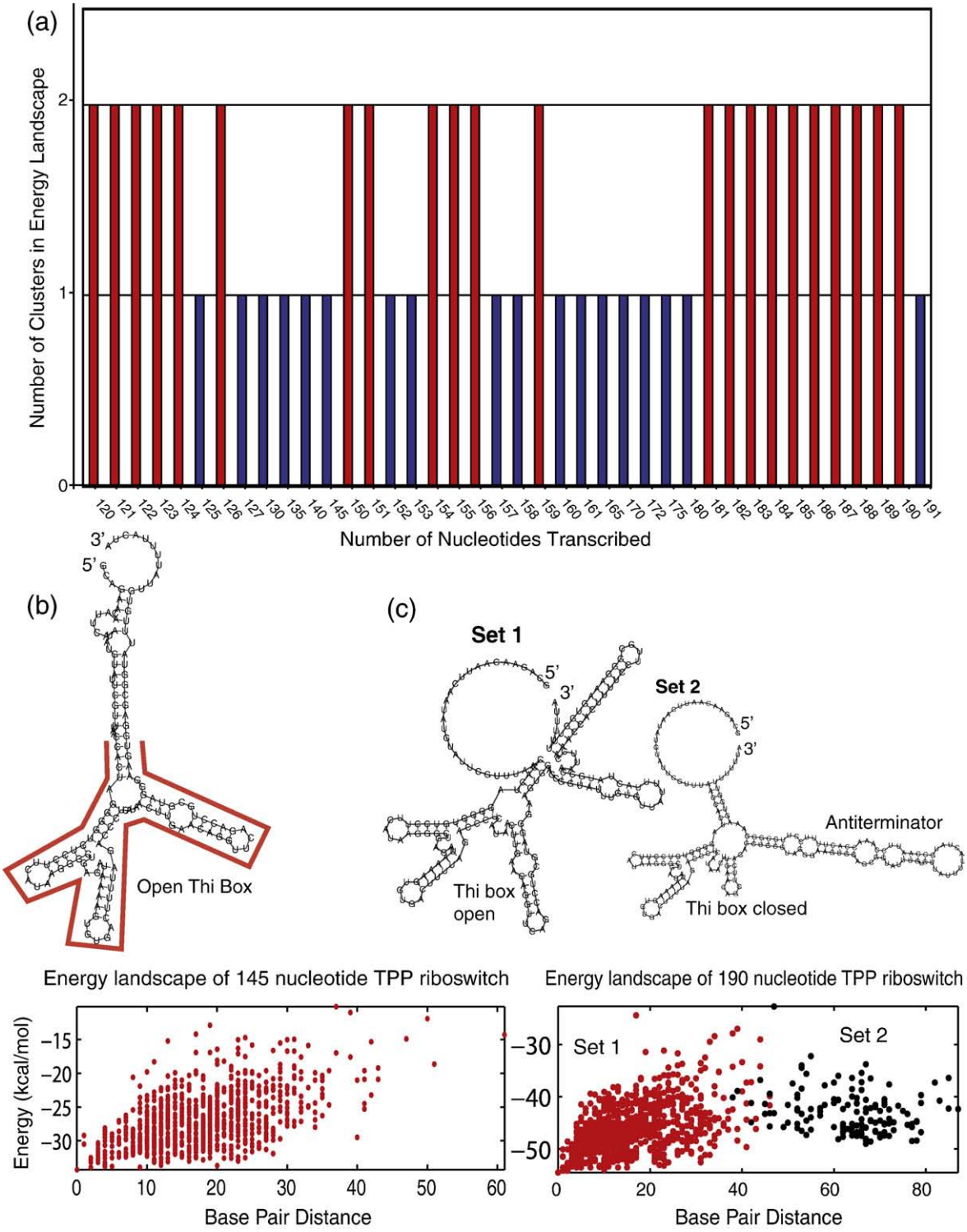

Fig. 3. Energy landscape analysis of TPP riboswitch from 120 to 190 nt (full length) with 1- to 5-nt incremental step. (a) Clustering results of TPP riboswitch energy landscape along the length of the sequence. Note that two clusters are present after 180 nt. Two representatives of energy landscapes and structures display (b) one cluster at 145 nt or (c) two distinct clusters at 190 nt. The Set 1 (left) and Set 2 (right) structures approximate the average structures represented by set 1 (red) and set 2 (black) on the energy landscape, respectively.

Clustered landscapes are present in diverse riboswitch systems

We further applied our method to a number of riboswitch systems that utilize other ligands and alternative mechanisms of genetic control (see Table 1). The *thiM* (TPP)[4] and *ypaA* (FMN)[13] mRNA 5′ UTR sequences are translation terminator riboswitches that function by base pairing of the Shine–Dalgarno sequence. Both sequences were

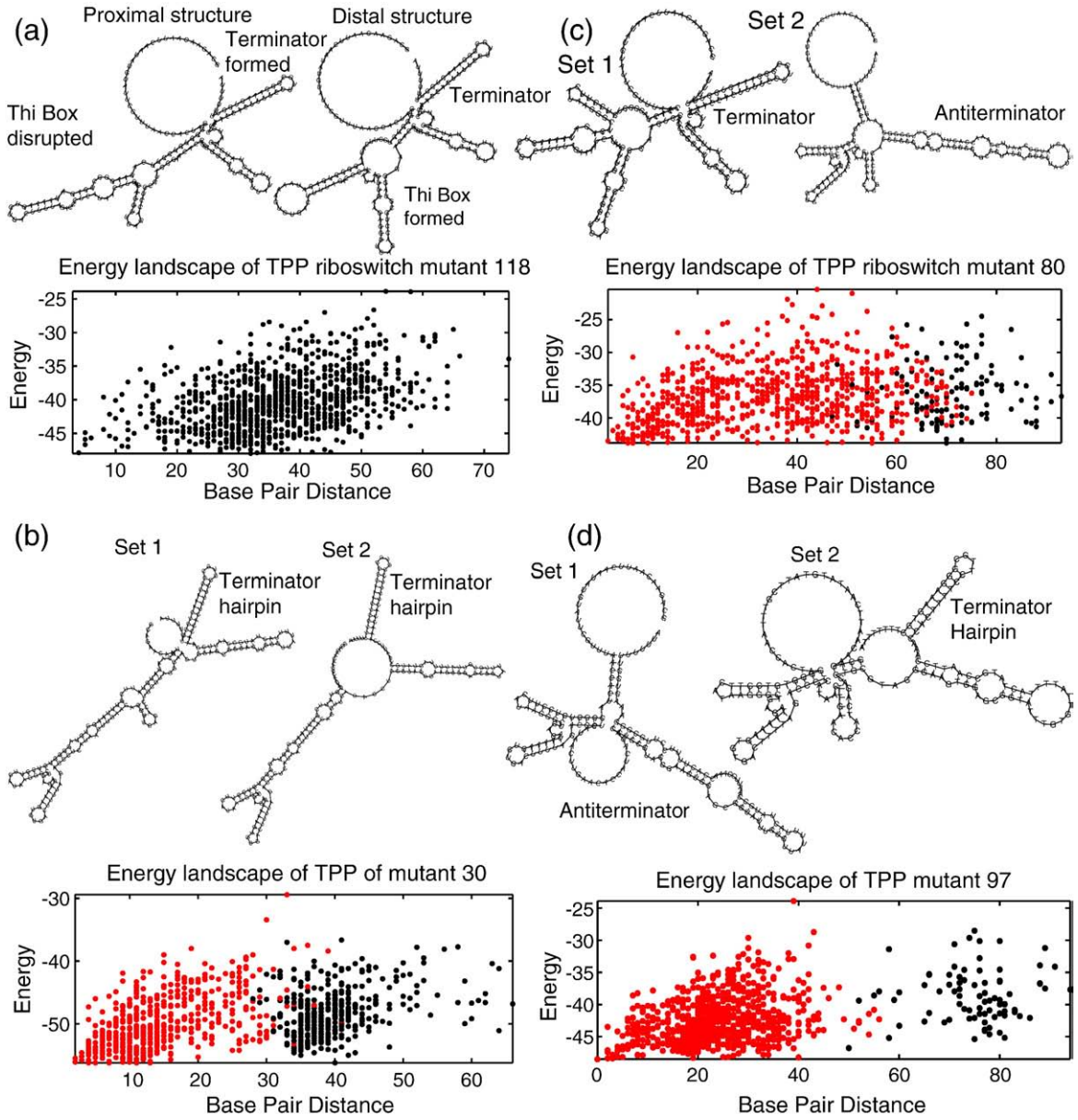

Fig. 4. Energy landscape and representative structures of riboswitch mutant 118 (a), mutant 30 (b), mutant 80 (c), and mutant 97 (d). Note that the two structures shown for mutant 118 do not represent different sets of structures but are only two sample points along the energy landscape. For mutants 30, 80, and 97, Set 1 (left) and Set 2 (right) structures approximate the average structures represented by set 1 (red) and set 2 (black) on the energy landscape, respectively.

found to display clustered energy landscape properties similar to those of the *tenA* riboswitch (see Fig. 5a and b). However, the *ypaA* riboswitch is unique in that the higher-energy cluster (Set 2) corresponds to the FMN-bound structure, opposite of what was found for the wild-type *tenA* TPP riboswitch. This suggests that the lowest-energy state is the non-bound form. Microarray analysis indicates that the *ypaA* sequence shows no change in the quantitative level of transcripts when *B. subtilis* strains are grown in the presence or absence of riboflavin, alluding to the function of *ypaA* as a translation inhibitor.[14] In addition, the coenzyme B_{12}-sensing *btuB* leader sequence in *Escherichia coli* shows a two-state energy landscape (see Fig. 5c). This riboswitch utilizes a variety of tertiary interactions and pseudoknots to perform silencing of downstream genes in the presence of a ligand.[16] These studies thus suggest that the energy landscape properties developed here are generalizable to different classes of riboswitches.

The *gcvT* and VCI-II elements are dual, tandem-aptamer RNA riboswitches that display cooperative binding of glycine and genetic control through transcription termination.[15,23] Each aptamer domain can bind one separate molecule of glycine. Noteworthy is the fact that the presence of glycine decreases the level of *gcvT* termination, represent-

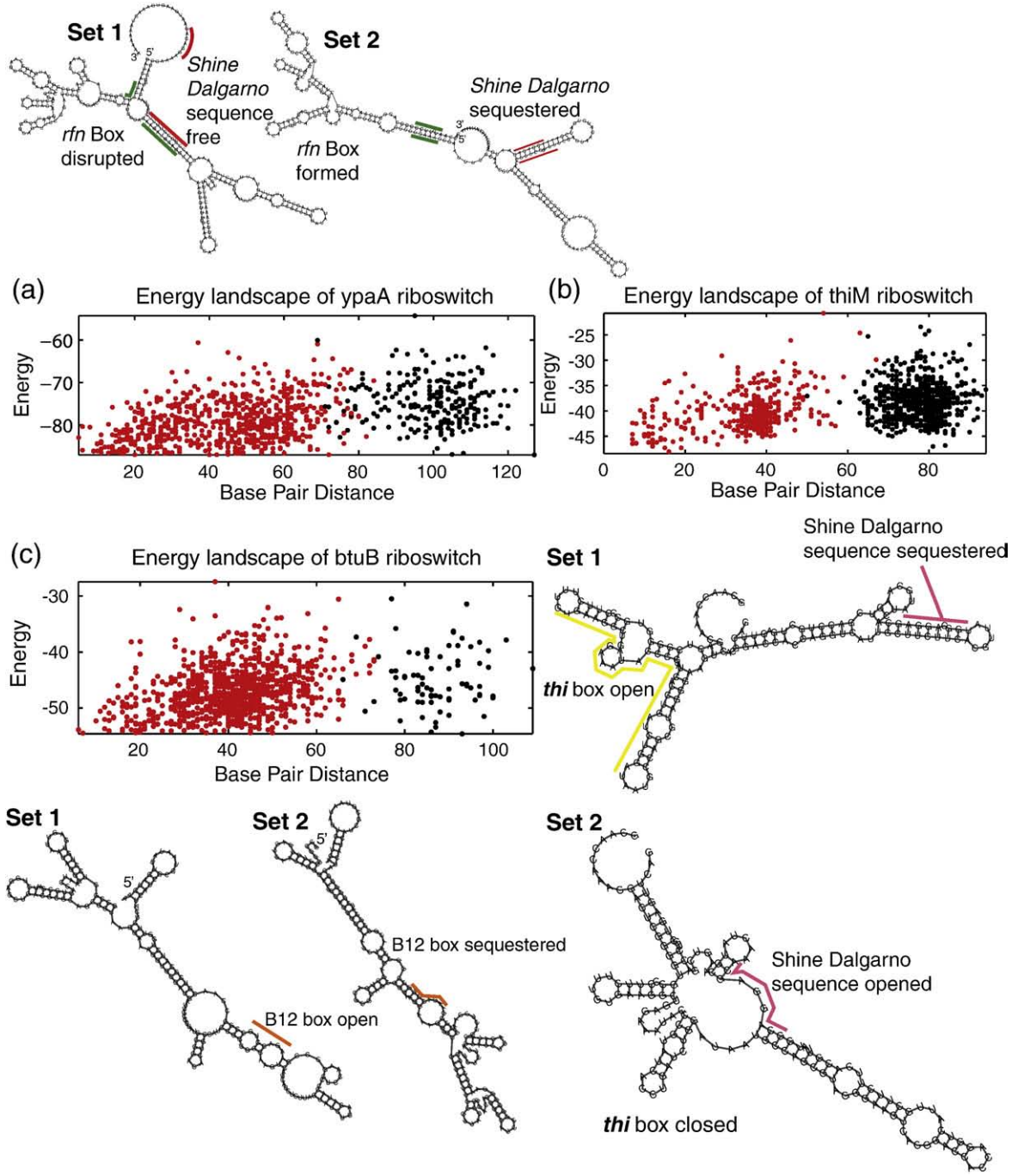

Fig. 5. Energy landscape plot and structures of the *ypaA* (a), *ribD* (b), and *btuB* (c) riboswitches. The Set 1 (left) and Set 2 (right) structures approximate the average structures represented by set 1 (red) and set 2 (black) on the energy landscape, respectively.

ing a case where the ligand is required to activate gene transcription. Computational analysis of both glycine riboswitches, however, reveals no clustering in the energy landscape (data not shown). This may be attributable to the unique function of the riboswitch as a subtle sensor that significantly modifies its structure with only moderate changes in ligand concentration.[24] Thus, more global changes in secondary structure may not be required to affect riboswitch function. It is this cooperative nature of ligand binding that may account for this difference in energy landscape properties.

Discussion

The total number of conformations that a linear macromolecule such as RNA can adopt is astronomical. RNA's ability to navigate this huge folding space quickly has been continuously subject to selective pressure. Many factors in the folding

process, including thermally driven chain fluctuations, ion-mediated electrostatics, base pairing and stacking, and other noncanonical interactions guide the folding process.[25] The advantage of the energy landscape approach presented here for analyzing the structural properties of functional RNAs is a simplified representation that approximates the feasible secondary structures at the low-energy portion of the energy funnel. Funnel energy landscapes have proven invaluable in protein folding studies.[26–28] The mfe approach, while theoretically most stable, represents a single point on the energy landscape and is subject to small changes in energy parameters and kinetic barriers to folding.[29] A growing RNA molecule may not fold immediately into the most stable structure but explore multiple folding routes or fluctuate between nearby "suboptimal" structures. Additionally, significant kinetic traps have been found to hold long RNAs in suboptimal folds.[30]

Thus, it remains unknown whether RNA folding is a direct pathway to the native fold, singularly controlled by the more stabilizing interactions of the native fold, or a process complicated by a series of frustrated intermediate states, in which the RNA inhabits suboptimal structures. The former notion suggests that the folding landscape is a smoothly sloping energy funnel with a unique mfe structure. The latter theory supports an exhaustive search between all possible sub-interactions in which the energy landscape is decorated with local minima that may trap the macromolecule. Our investigations suggest that the folding landscape of certain classes of riboswitches is length dependent: during the process of transcription, certain lengths utilize a smooth transition energy funnel to the mfe, while at other lengths, the energy landscape forms a frustrated equilibrium between two major topologies. This suggests that a complex interplay between kinetics and thermodynamics is required for the structure to attain the proper conformation. For all the riboswitches tested, we suggest that ligand binding occurs on top of a preformed macromolecular backbone and that this step determines the future outcome of gene regulation. This has been confirmed recently in thermodynamic simulations of the *S*-adenosylmethionine riboswitch aptamer domain, in which a helical platform for SAM binding is preformed.[31] Ligand binding then stabilizes the rate-limiting step to folding.

Similarly, protein folding is hypothesized to be co-translational, which means that smaller, compactly folded intermediates and completed native-like structures are attached to the ribosome during the process of translation.[32] This hypothesis is strongly supported by experimental results, such as enzymatic activities immediately after the protein's release from the ribosome.[32–34] The long history of research into protein folding has shown that the free-energy surface is also frustrated,[35] decorated with many minima aside from the native state, separated by varying energy barriers. Like *in vivo* protein folding, naturally occurring RNAs have been found to fold co-transcriptionally. Experimental studies done by the Pan and Sosnick groups have shown that RNA co-transcriptional folding in the cell is facilitated by pausing-induced nonnative structures.[36,37] Our computational approach identified such possible switches along the transcription elongation through a structural and thermodynamic analysis of the *B. subtilis tenA* TPP binding riboswitch. Though various folding algorithms analyzing RNAs during transcription have been developed—using genetic algorithms,[30,38] Monte Carlo simulations,[39–41] and stochastic methods[42]—this is the first attempt to study mfe folding as a function of nucleotide length, to the best of our knowledge.

Two major classes have emerged from crystal structure studies of riboswitches, functioning through differing allosteric mechanisms.[2] Type 1 riboswitches, characterized by the TPP riboswitch, display global conformational changes upon ligand binding. Type 2 riboswitches, exemplified by the purine and *S*-adenosylmethionine riboswitches, function through subtle tertiary interactions stemming from changes in the binding pocket.[2] In agreement with our results of the TPP riboswitch, those riboswitches that display global conformational clustering may indeed fall into the Type 1 classification. In contrast, it may be possible that the *gcvT* and VCI-II glycine riboswitches, characterized by their cooperative nature of ligand binding, may act as Type 2 riboswitches and, as a result, do not display the same energy landscape properties in our analysis.

Interestingly, we also found that the experimentally reported efficiency of functional gene regulation is somewhat correlated to the density of the ligand-bound structure set in our corresponding computed energy landscapes. Because the experimental values reflect nonuniform conditions in each experiment (see Supplementary Table 1), exact comparisons to our energy-landscape-derived values are not prudent, but overall, we note a general correspondence for most cases examined. This general correspondence suggests a structure–function connection of our energy landscape views to transcription activity.

Riboswitches can be exploited to design new functional RNAs for biotechnology or biomedical applications, for example, using rational modular design to engineer assemblies of riboswitches and other aptamer modules. Such design efforts can be made more productive by selecting RNAs with desired structural and thermodynamic properties, as examined in this work. Recently, Wieland *et al.* designed a TPP riboswitch–hammerhead ribozyme fusion that controls ribosome binding through the presence or absence of TPP.[43] This process required exhaustive search of approximately 4000 sequences for activity. We anticipate that a computational framework, such as the one we presented, could streamline the search for functional transcripts. Applications of such computational tools to novel RNA design are currently underway.

Materials and Methods

Simulation of transcription elongation

Various programs[17,44] have been used to predict the "optimal" secondary structure of short RNA sequences by minimizing the free energy of folding[45] from a standard set of energetic parameters.[46,47] These methods are based on the principles of hierarchal folding of RNA so that secondary structural elements form a scaffold upon which tertiary interactions are then achieved.[48–50] It is also assumed that, at equilibrium, the molecule will thermodynamically favor its lowest-energy state. Such RNA prediction of secondary structures has been shown to be around 70% accurate on average for short RNAs of less than 200 nt,[46] despite the approximations of thermodynamic parameters. Additionally, the kinetics of folding, tertiary interactions, and pseudoknot formation are not fully taken into account. Thus, the structures predicted by free-energy minimization provide a valuable but incomplete view of structures. For recent approaches to RNA folding, see Ref. 51.

In both Mfold and Vienna RNAfold, the mfe and suboptimal structures of a single RNA sequence are predicted by the algorithm of Zuker and Stiegler.[52] Essentially, the overall free energy is approximated by the sum of the loop and base-pair energies.[53] These energy parameters are estimated based on melting temperature studies of synthetically constructed oligoribonucleotides at arbitrary temperatures.[46]

In our applications, we simulate the elongation of the TPP-binding riboswitch from the 5′ UTR of the *tenA* gene from *B. subtilis* at 1- to 5-nt incremental steps up to the transcription start site (191 nt) to investigate the structural and thermodynamic switches along the transcription elongation pathway. For structural analysis, subsequences were folded into their native ("optimal") secondary structures using Mfold[17] by free-energy minimization.

Clustering analysis of energy landscape

For thermodynamic analysis, we sample 1000 possible ("suboptimal") secondary structures using the RNAsubopt module of the Vienna RNA package.[54] Sampling was performed at 55 °C from the Boltzmann-weighted distribution of secondary structures because we found that the sampling at elevated temperatures produces greater diversity and was more representative of *in vitro* experimental results. Our simulation shows that temperatures below 55 °C generate structures that are mostly similar to the optimal secondary structures, while higher temperatures generate unfolded structures. Obtaining the partition function for folding at elevated temperatures requires extrapolation of free-energy parameters from their reference at 37 °C. However, this has been shown to accurately predict folding landscape properties and, more importantly, effectively enhance the range of conformational states sampled.[55] Free energies of sampled structures were recalculated at 37 °C using RNAeval to be physiologically relevant.

We compute the base-pair distance matrix between all sampled structures using RNAdistance.[56] The base-pair distance measures the number of base pairings that require breaking or forming in order to convert one structure into another. A plot of the distance between the mfe structure and all other structures (i.e., the first column of the distance matrix) against the free energy of folding produces an illustrative representation of the folding pathways, which we call the energy landscape plot. Such a plot indicates the range of possible RNA secondary structures for a given sequence with relationship to the mfe represented as the lowest point on the energy axis.

To assign individual secondary structures of the energy landscape to clusters of structures with similar topological characteristics, we use the k-means algorithm (with $k=2$) for partition clustering in the R statistical software package.[57] The algorithm aims to partition the points into k groups such that the sum of squares between the assigned cluster centers and each point is minimized.[58] For cluster validation, the average silhouette width was used, quantifying a measure of the clustering (for full description, see Ref. 58). We define a threshold value of ≥ 0.4 as a well-clustered result (note that the highest silhouette coefficient with $k=3$, 4, and 5 for the *B. subtilis* TPP riboswitch is 0.35, while the value with $k=2$ is 0.62).

We partition the entire set of structures into their respective clusters. We then developed an automated procedure to analyze the most dominant secondary structural features in the clusters. In abstract form, a given RNA secondary structure can be represented as a set of balanced parentheses and points associated with each nucleotide to indicate base-pairing patterns. For example, the string ((((.....)))) represents the secondary structure of a simple 5-bp helix. In each cluster, we compute the frequency of finding one of the string characters at each position in the string. The most dominant characters (i.e., signature corresponding to the greatest number of structures) were assembled together into a string of characters. If no single element was present in the majority of the structures, an underscore character was assigned. The complete string of characters approximates the dominant structural elements in the cluster, which we term the *average structure* of the cluster. As an example, a set of structures was generated for the helix presented above (see Table 2).

To determine the accuracy of our method, we calculate the number of structural elements that matched well our average structure to structures derived experimentally or through sequence alignment analysis (see Table 3). The aptamer domains of many riboswitches are known to be highly conserved across species,[3,6] and thus, we limited this comparison to the aptamer region only. In each riboswitch data set analyzed, we found high structural

Table 2. Base-pair distance between an mfe structure and suboptimal structure

Structure set	Base-pair distance
(((((.....)))))	mfe structure
((((.....)))).	9
((((.....))))	1
(((.(.....).)))	1
(.(((.....))).)	1
(((.....)))	2
((((_.....))))	Average structure

The average structure is composed of the most common structural elements at each position in the set of structures. The second structure requires unzipping of a short helix (four moves), followed by matching the distal 5′ and 3′ ends (five moves) to reproduce the mfe structure. In the last entry, at position 5, the average structure is marked with an underscore since no character is present in >50% of the set.

Table 3. Percentage of matching base pairs between the ligand-binding domains of the computationally predicted average structure (see Materials and Methods) and experimentally derived structure

Riboswitch name (sequence)	Percentage of matching structural elements
tenA, B. subtilis thi-box[11,a]	99
ypaA, B. subtilis RFN element[13]	65
ribD, B. subtilis RFN element[13]	62
gcvT, B. subtilis[15]	64
VCI-II, *V. cholerae*[15]	81
btuB, E. coli B12 box[59]	62
thiM, E. coli thi-box[4]	83

[a] Experimental structures were published using Mfold.[17]

similarity, indicating the validity of our computational approach.

Acknowledgements

We are deeply grateful to Dr. Hin Hark Gan for insightful comments and recommendations regarding various aspects of this work. Funding support from the National Science Foundation, the National Institutes of Health, and the Human Frontier Science Project is gratefully acknowledged.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.08.062

References

1. Serganov, A., Yuan, Y. R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T. *et al.* (2004). Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.* **11**, 1729–1741.
2. Montange, R. K. & Batey, R. T. (2008). Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **37**, 117–133.
3. Nudler, E. & Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29**, 11–17.
4. Winkler, W., Nahvi, A. & Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.
5. Winkler, W. C. & Breaker, R. R. (2003). Genetic control by metabolite-binding riboswitches. *ChemBioChem*, **4**, 1024–1032.
6. Winkler, W. C. & Breaker, R. R. (2005). Regulation of bacterial gene expression by riboswitches. *Annu. Rev. Microbiol.* **59**, 487–517.
7. Roth, A., Nahvi, A., Lee, M., Jona, I. & Breaker, R. R. (2006). Characteristics of the glmS ribozyme suggest only structural roles for divalent metal ions. *RNA*, **12**, 607–619.
8. Cheah, M. T., Wachter, A., Sudarsan, N. & Breaker, R. R. (2007). Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, **447**, 497–500.
9. Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S. *et al.* (2003). Thiamine-regulated gene expression of *Aspergillus oryzae* thiA requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. *FEBS Lett.* **555**, 516–520.
10. Rentmeister, A., Mayer, G., Kuhn, N. & Famulok, M. (2007). Conformational changes in the expression domain of the *Escherichia coli* thiM riboswitch. *Nucleic Acids Res.* **35**, 3713–3722.
11. Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A. *et al.* (2002). Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, **111**, 747–756.
12. Lang, K., Rieder, R. & 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.
13. Winkler, W. C., Cohen-Chalamish, S. & Breaker, R. R. (2002). An mRNA structure that controls gene expression by binding FMN. *Proc. Natl Acad. Sci. USA*, **99**, 15908–15913.
14. Lee, J. M., Zhang, S., Saha, S., Santa, A. S., Jiang, C. & Perkins, J. (2001). RNA expression analysis using an antisense *Bacillus subtilis* genome array. *J. Bacteriol.* **183**, 7371–7380.
15. Mandal, M., Lee, M., Barrick, J. E., Weinberg, Z., Emilsson, G. M., Ruzzo, W. L. & Breaker, R. R. (2004). A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science*, **306**, 275–279.
16. Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L. & Breaker, R. R. (2002). Genetic control by a metabolite binding mRNA. *Chem. Biol.* **9**, 1043.
17. Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415.
18. Miranda-Rios, J., Navarro, M. & Soberon, M. (2001). A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl Acad. Sci. USA*, **98**, 9736–9741.
19. Greenleaf, W. J., Frieda, K. L., Foster, D. A., Woodside, M. T. & Block, S. M. (2008). Direct observation of hierarchical folding in single riboswitch aptamers. *Science*, **319**, 630–633.
20. Thore, S., Leibundgut, M. & Ban, N. (2006). Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science*, **312**, 1208–1211.
21. Edwards, T. E. & Ferre-D'Amare, A. R. (2006). Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure*, **14**, 1459–1468.
22. Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R. & Patel, D. J. (2006). Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, **441**, 1167–1171.
23. Lipfert, J., Das, R., Chu, V. B., Kudaravalli, M., Boyd, N., Herschlag, D. & Doniach, S. (2007). Structural transitions and thermodynamics of a glycine-dependent riboswitch from *Vibrio cholerae*. *J. Mol. Biol.* **365**, 1393–1406.
24. Welz, R. & Breaker, R. R. (2007). Ligand binding and gene control characteristics of tandem riboswitches in *Bacillus anthracis*. *RNA*, **13**, 573–582.
25. Chen, S. J. (2008). RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu. Rev. Biophys.* **37**, 197–214.

26. Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19.
27. Chan, H. S. & Dill, K. A. (1998). Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins*, **30**, 2–33.
28. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. (2008). The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316.
29. Ding, Y. (2006). Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA*, **12**, 323–331.
30. Gultyaev, A. P., van Batenburg, F. H. & Pleij, C. W. (1995). The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250**, 37–51.
31. Whitford, P. C., Schug, A., Saunders, J., Hennelly, S. P., Onuchic, J. N. & Sanbonmatsu, K. Y. (2009). Nonlocal helix formation is key to understanding *S*-adenosylmethionine-1 riboswitch function. *Biophys. J.* **96**, L7–L9.
32. Komar, A. A. (2009). A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* **34**, 16–24.
33. Hamlin, J. & Zabin, I. (1972). β-Galactosidase: immunological activity of ribosome-bound, growing polypeptide chains. *Proc. Natl Acad. Sci. USA*, **69**, 412–416.
34. Kudlicki, W., Kitaoka, Y., Odom, O. W., Kramer, G. & Hardesty, B. (1995). Elongation and folding of nascent ricin chains as peptidyl-tRNA on ribosomes: the effect of amino acid deletions on these processes. *J. Mol. Biol.* **252**, 203–212.
35. Guo, Z. & Thirumalai, D. (1996). Kinetics and thermodynamics of folding of a de novo designed four-helix bundle protein. *J. Mol. Biol.* **263**, 323–343.
36. Wong, T. N., Sosnick, T. R. & Pan, T. (2007). Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc. Natl Acad. Sci. USA*, **104**, 17995–18000.
37. Pan, T. & Sosnick, T. (2006). RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 161–175.
38. Shapiro, B. A., Bengali, D., Kasprzak, W. & Wu, J. C. (2001). RNA folding pathway functional intermediates: their prediction and analysis. *J. Mol. Biol.* **312**, 27–44.
39. Tang, X., Thomas, S., Tapia, L., Giedroc, D. P. & Amato, N. M. (2008). Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.* **381**, 1055–1067.
40. Liu, F. & Ou-Yang, Z. C. (2005). Monte Carlo simulation for single RNA unfolding by force. *Biophys. J.* **88**, 76–84.
41. Schmitz, M. & Steger, G. (1996). Description of RNA folding by "simulated annealing". *J. Mol. Biol.* **255**, 254–266.
42. Xayaphoummine, A., Bucher, T., Thalmann, F. & Isambert, H. (2003). Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl Acad. Sci. USA*, **100**, 15310–15315.
43. Wieland, M., Benz, A., Klauser, B. & Hartig, J. S. (2009). Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angew. Chem., Int. Ed. Engl.* **48**, 2715–2718.
44. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188.
45. Zuker, M. (1989). Computer prediction of RNA structure. *Methods Enzymol.* **180**, 262–288.
46. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940.
47. Xia, T., Santa Lucia, J., Jr., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X. *et al.* (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*, **37**, 14719–14735.
48. Brion, P. & Westhof, E. (1997). Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 113–137.
49. Tinoco, I., Jr & Bustamante, C. (1999). How RNA folds. *J. Mol. Biol.* **293**, 271–281.
50. Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F. & Zehl, M. (2001). Design of multistable RNA molecules. *RNA*, **7**, 254–265.
51. Shapiro, B. A., Yingling, Y. G., Kasprzak, W. & Bindewald, E. (2007). Bridging the gap in RNA structure prediction. *Curr. Opin. Struct. Biol.* **17**, 157–165.
52. Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9**, 133–148.
53. Eddy, S. R. (2004). How do RNA folding algorithms work? *Nat. Biotechnol.* **22**, 1457–1458.
54. Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431.
55. Bonhoeffer, S., McCaskill, J. S., Stadler, P. F. & Schuster, P. (1993). RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur. Biophys. J.* **22**, 13–24.
56. Fontana, W., Konings, D. A., Stadler, P. F. & Schuster, P. (1993). Statistics of RNA secondary structures. *Biopolymers*, **33**, 1389–1404.
57. R Development Core Team (2007). R: A Language and Environment for Statistical Computing, Vienna, Austria.
58. Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York, NY.
59. Nahvi, A., Barrick, J. E. & Breaker, R. R. (2004). Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res.* **32**, 143–150.