Three-dimensional structures of RNA obtained by means of knowledge-based interaction potentials

Oscar Taxilaga-Zetina, Patricia Pliego-Pastrana, and Mauricio D. Carbajal-Tinoco
*Departamento de Física, Centro de Investigación y de Estudios Avanzados del IPN,
Apartado Postal 14-740, 07000 México, Distrito Federal, Mexico*

(Received 21 July 2009; revised manuscript received 24 March 2010; published 19 April 2010)

We derive a set of effective potentials describing the interaction between pairs of nucleotides that belong to an RNA molecule. Such interaction potentials are then used as the main constituents of a simplified simulation model, which is tested in the description of small secondary structure motifs. Our simulated RNA hairpins are consistent with the experimental structures obtained by NMR.

DOI: 10.1103/PhysRevE.81.041914 PACS number(s): 87.14.gn, 87.15.ak, 87.15.bd

I. INTRODUCTION

RNA is one of the most important molecules for life. For instance, different types of RNA molecules perform multiple tasks that include the translation of genetic information, biological catalysis, and structural scaffold [1]. A specific example is the case of an RNA called XIST that has the power to turn off an entire chromosome [2]. Like proteins, RNAs adopt complex three-dimensional (3D) structures to achieve their functions. Therefore, a classic problem of biophysics is the determination of the 3D native state of an RNA molecule, based on its linear sequence of nucleotides or primary structure. The solution to this problem will be of outstanding value for medical and biochemical applications.

The problem of RNA folding has been studied with varied approaches. One direct strategy is to start with a secondary (base paired) structure of a given molecule and then use it as a template to construct a compatible 3D structure [3]. The entire procedure eventually demands an interactive manipulation or a further refinement by molecular-dynamics (MD) simulations. Otherwise, a number of coarse-graining models can be found in the literature. Among them, we can mention the discrete molecular dynamics (DMD) that has been used to fold RNA molecules with relative precision [4]. The noncontinuous energy function of such model includes base pairing, base stacking, and hydrophobic interactions. Another example of a coarse-graining model is the nucleic acid simulation tool (NAST) [5]. Different types of statistical information contribute to the NAST protocol such as short-range interactions (distances, angles, dihedrals, and excluded volume) and long-range contact interactions. In this model, the secondary structure has to be provided as well. On the other hand, it is possible to study the whole folding process in full detail. A method of this kind consists of using an optimized MD algorithm within an all-atom model in the explicit solvent [6]. This approach, however, requires a vast computational effort even in the case of a relatively small RNA molecule. It is thus desirable to combine the advantages of these methods, i.e., simplicity and accuracy.

In this paper, we propose a simulation model of RNA that is based on a series of continuous pairwise potential functions that are extracted from experimental data. In our scheme, each nucleotide of an RNA sequence is replaced by its corresponding center of mass. The interaction between centroids is mediated by effective pair potentials (EPPs), which are either angular or radial, and each EPPs is independent of the other ones. Our model utilizes a total amount of 21 EPPs (11 angular and 10 radial), but the number of degrees of freedom is significantly reduced, in comparison with an all-atom model. Otherwise, the resulting structures of our Monte Carlo (MC) simulations reproduce some of the most relevant features found in experimental molecules.

II. EXPERIMENTAL INTERACTION POTENTIALS

We define an EPP as the potential energy between two "particles" that recreates their corresponding pair-correlation function [7]. Accordingly, each EPP has the integrated contribution of both direct interactions (e.g., Coulomb, hydrogen bonds, or excluded volume), as well as indirect ones (solvent, surrounding ions, among others). In the case of biomolecules, we have already tested a model that makes use of EPPs as principal constituents. As a result, our model of polypeptides reproduced three secondary structures of proteins [8,9]. At this point, it is important to emphasize that models based on EPPs are expected to generate structural features that are only consistent with the conditions of the original experiments. For example, the experimental RNAs analyzed in the present study are characterized by a pH of 7 ± 1 and a temperature of 281 ± 7 K.

Our knowledge-based potentials were obtained from the analysis of a series of 20 crystallographic structures of RNAs of high molecular weight from the Protein Data Bank (PDB), mainly ribosomes such as 1ffk, 1njp, and 2b66. These large structures are assumed to be in thermodynamic equilibrium [9,10]. In the present model of RNA, we are interested in three types of pairwise interactions, namely, bending, torsion, and distance-dependent energies. All of them are derived from their corresponding pair-correlation functions $g_{\mu\gamma}(\xi)$, where ξ is the independent variable and the subscripts μ and γ stand for A, C, G, or U. In order to take into account finite density effects, the radial-dependent effective potentials $\beta u_{\mu\gamma}(r)$ ($\beta^{-1}=k_BT$ is the thermal energy) can be related to the functions $g_{\mu\gamma}(r)$ by means of the multicomponent Ornstein-Zernike equations and a suitable closure approximation. In all cases under study [8–10], as well as in the case of RNA molecules, we found that the many-body effects are negligible. Thus, our EPPs (both, radial and an-

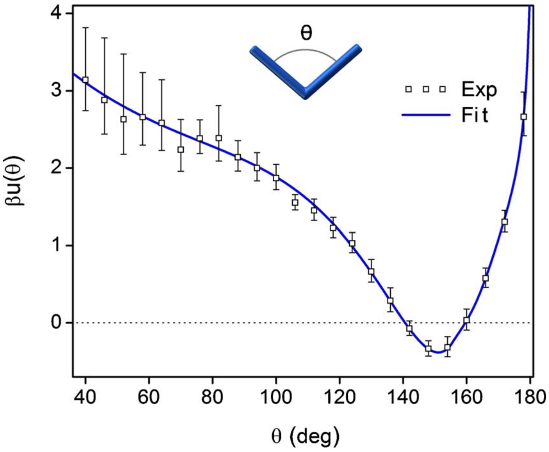

FIG. 1. (Color online) Bending effective pair potential $\beta u(\theta)$ extracted from the experimental correlation function $g(\theta)$ (open squares), as explained in the text. The continuous line is a Padé-like fit of this potential.

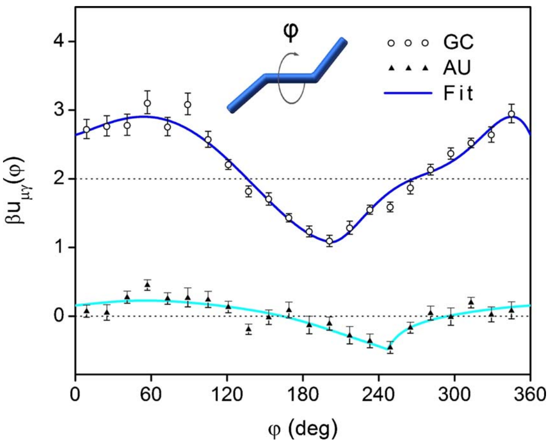

FIG. 2. (Color online) Torsional EPPs $\beta u_{\mu\gamma}(\varphi)$ between the pairs GC (open circles) and AU (full triangles). To enhance clarity, $\beta u_{GC}(\varphi)$ is presented with an offset of 2 k_BT. The continuous lines are Padé-like fits of the potentials GC (upper blue line) and AU (lower cyan line).

gular) are determined through the low-density limit expression [11],

$$\beta u_{\mu\gamma}(\xi) = -\ln[g_{\mu\gamma}(\xi)]. \tag{1}$$

Let us first start by computing the bending correlation functions from the averaged properties of the experimental data. The vector associated with the pairs of nucleotides $(i,i+1)$ is $\mathbf{a}_{i+1}=\mathbf{r}_{i+1}-\mathbf{r}_i$, with $\mathbf{r}_i$ being the position of the base i. The bending angle θ is obtained from the relation $\theta = \arccos[-\mathbf{a}_{i+2}\cdot\mathbf{a}_{i+1}/(|\mathbf{a}_{i+2}||\mathbf{a}_{i+1}|)]$. Hence, such correlation functions are determined on the understanding that $\pi\rho_\theta g_{\mu\gamma}(\theta)d\theta$ is the number of nucleotides of species μ and γ, which are located between two cones of apertures θ and $\theta+d\theta$, and centered in a nucleotide of arbitrary species. Here, the number density is $\rho_\theta=1/90°$. Within experimental errors, all correlation functions $g_{\mu\gamma}(\theta)$ are very similar. Therefore, we averaged them in a single correlation function $g(\theta)$, which is then used in Eq. (1) to obtain $\beta u(\theta)$. In Fig. 1, we plot the bending effective potential together with a Padé-type fit of the experimental data. The position of the minimum of the potential well is $(151\pm4)°$.

The torsional correlation functions are calculated in a similar way to the previous case. Our torsion angle is defined as $\varphi=\pi+\arctan 2[|\mathbf{a}_{i+2}|\mathbf{a}_{i+1}\cdot(\mathbf{a}_{i+2}\times\mathbf{a}_{i+3}),(\mathbf{a}_{i+1}\times\mathbf{a}_{i+2})\cdot(\mathbf{a}_{i+2}\times\mathbf{a}_{i+3})]$. Correspondingly, $\rho_\varphi g_{\mu\gamma}(\varphi)d\varphi$ is the number of nucleotides found between the angles φ and $\varphi+d\varphi$. The number density in this case is $\rho_\varphi=1/180°$. The interaction potentials are then obtained from the correlation functions $g_{\mu\gamma}(\varphi)$ by means of Eq. (1). For example, the torsional EPPs corresponding to the pairs of bases GC and AU are plotted in Fig. 2, as well as their fitting curves. As it can be observed, the two EPPs are quantitatively different in regard to the position and depth of their respective potential wells. The positions of the potential wells of curves GC and AU are (202 ± 8) and $(247\pm8)°$, respectively.

The distance-dependent correlation functions are derived from the remaining pairs of nucleotides (i,j), with $j\geq i+4$. Although we have selected some of the largest RNA crystallographic structures, it is necessary to take into account the effect of finite size of such molecules. Let us first define a test sphere of radius $r_{\max}$ and volume $V=4\pi r_{\max}^3/3$ that contains one or two types of bases of species μ and γ, with at least 50 nucleotides of each type and homogeneously distributed inside the test sphere. For a given distance r, the radial distributions $g_{\mu\gamma}(r)$ are determined through the equation [9],

$$g_{\mu\gamma}(r) = \frac{NV}{N_\mu N_\gamma}\left(\frac{h'(r)}{N4\pi r^2 dr - N'V_c(r)}\right), \tag{2}$$

where N_μ and N_γ are the numbers of particles of species μ and γ inside the volume V, $N=N_\mu+N_\gamma$, and $h'(r)$ is the total number of nucleotides of the same species between two concentric spheres of radii r and $r+dr$, about a central one, as shown in Fig. 3. If the central nucleotide is found inside the sphere of radius $r_{\max}-r$ then $V_c(r)=0$ (region I). For a central nucleotide located outside region I and still inside the

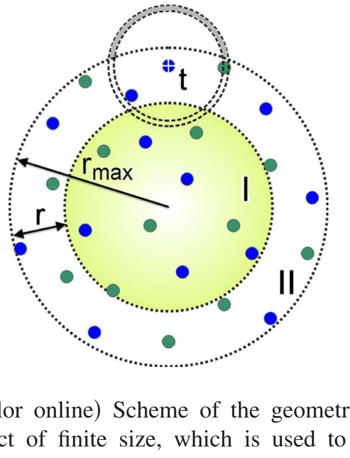

FIG. 3. (Color online) Scheme of the geometry employed to correct the effect of finite size, which is used to derive $g_{\mu\gamma}(r)$. Given a test sphere of radius $r_{\max}$ and a certain distance r, we define regions I (inside a sphere of radius $r_{\max}-r$) and II (between the test sphere and region I). For nucleotides located inside region II (t, for example), there is an excess volume (gray filled) that has to be subtracted for an appropriate normalization (see text).

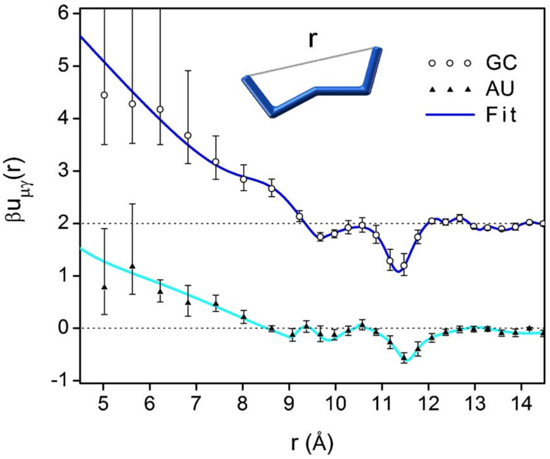

FIG. 4. (Color online) Distance-dependent EPPs $\beta u_{\mu\gamma}(r)$ between the Watson-Crick base pairs GC (open circles) and AU (full triangles). To enhance clarity, $\beta u_{GC}(r)$ is plotted with an offset of 2 k_BT. The continuous lines are Padé-like fits of the potentials GC (upper blue line) and AU (lower cyan line).

test sphere (region II), $V_c(r)=\pi[(r_{\max}^2-r^2)\ln(1-r/r_{\max})+3r^2/2+rr_{\max}]dr$, with N' [in Eq. (2)] being the number of particles found in region II. The function $V_c(r)$ is thus a correction to the effect of finite size. The 10 different potentials $\beta u_{\mu\gamma}(r)$ were calculated using Eqs. (1) and (2). Not surprisingly, the potential curves with the deepest potential wells are those identified as Watson-Crick base pairs, namely, GC and AU, which are plotted in Fig. 4. The potential wells of curves GC and AU are located at 11.4 ± 0.3 and 11.5 ± 0.3 Å, respectively, while their corresponding depths are −0.92 and −0.61 k_BT.

III. SIMULATION MODEL OF RNA

The continuous versions of the EPPs described in the previous section are the basis for a simulation model of RNA (the Padé fits are included in Ref. [12]). Our dressed polymer model (DPM) consists of a freely jointed chain made out of N rigid segments of distinct lengths $a_{\mu\gamma}$. The ends of each segment represent the positions of the centroids of two nucleotides of species μ and γ, and the actual lengths $a_{\mu\gamma}$ are the average distances extracted from the experimental data. The interactions between the second and third nearest neighbors in the chain are mediated by the potentials $\beta u(\theta)$ and $\beta u_{\mu\gamma}(\varphi)$, respectively, and the remaining couples of bases interact through $\beta u_{\mu\gamma}(r)$. The purpose of the three dressing EPPs is to capture, at least qualitatively, some of the most important mechanical and structural properties of RNA chains that include stiffness, chirality, and long-range order.

In our basic simulation algorithm, the RNA chains are grown in a progressive way. Starting with a single segment, more segments are added one by one, in successive steps. Each segment is represented by a vector, thus the chain (partially formed or complete) is just the sum of these individual vectors. A trial move consists of randomly rotating one or more arbitrary segments and then reforming the chain. The trial move is accepted with a probability given by [13],

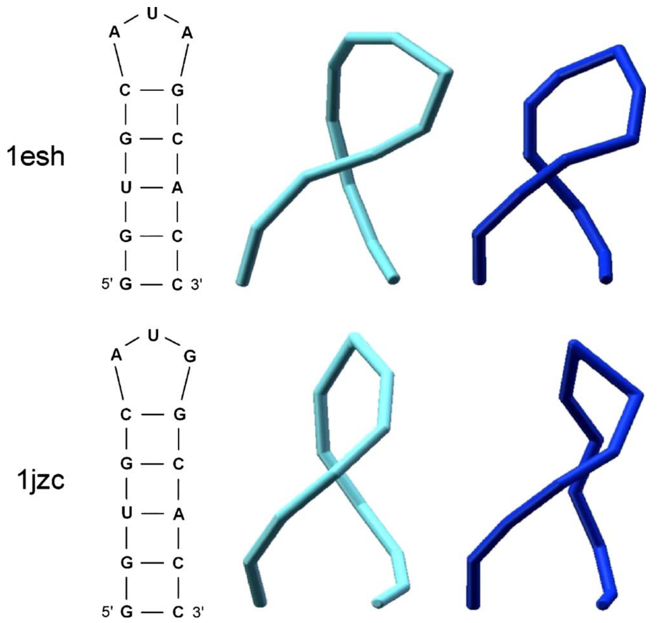

FIG. 5. (Color online) RNA hairpin structures 1esh (top) and 1jzc (bottom). The experimental molecules (blue lines, right) and simulated RNAs (cyan lines, center) are drawn together with the secondary structures obtained with the program MFOLD (left), as explained in the text.

$$P_{acc}=\min[1,\exp(-\Delta\beta E_{RT})],\tag{3}$$

with $\Delta\beta E_{RT}=\beta E_T-\beta E_R$ is the change in potential energy between the conformations R (reference) and T (test). This algorithm has the following features. It does not require any further information and it allows to obtain the most stable

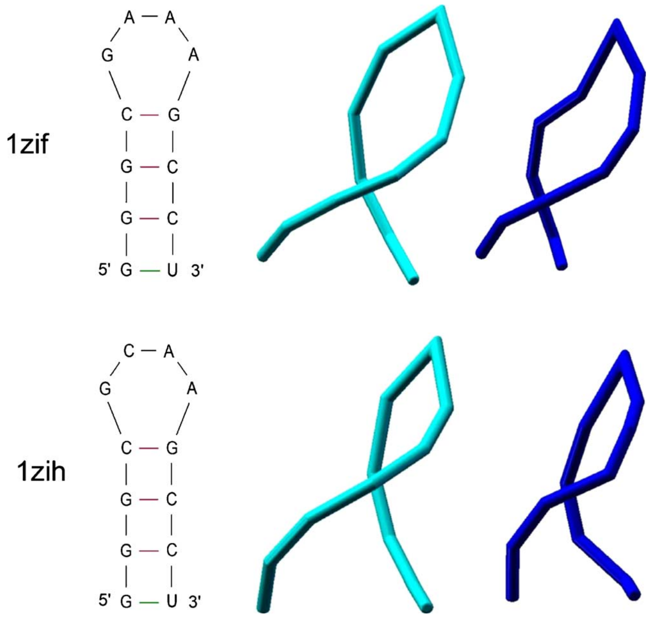

FIG. 6. (Color online) RNA hairpin molecules 1zif (top) and 1zih (bottom). The experimental structures (blue lines, right) and simulated RNAs (cyan lines, center) are plotted together with the secondary structures determined with the program MFOLD (left).

TABLE I. A comparison of the characteristic distances $\bar{d}_{WC}$ (mean distance between the centroids of the Watson-Crick base pairs), R_g (radius of gyration), L_c (contour length), and the ratio L_c/R_g that were obtained from the experimental (1esh and 1jzc) and simulated RNA molecules (s1esh and s1jzc). The average energy $\beta\bar{E}$ of the simulated hairpins is also included.

	$\beta\bar{E}$	$\bar{d}_{WC}$ (Å)	R_g (Å)	L_c (Å)	L_c/R_g
1esh		11.0 ± 0.9	8.1	56.7	7.0
s1esh	-15.4	11.6 ± 1.0	9.5	65.8	6.9
1jzc		10.5 ± 1.2	8.7	61.2	7.0
s1jzc	-16.2	11.6 ± 0.5	9.1	64.8	7.1

chain configuration. In regard to the computing time, we have not done any cutoff assumption and for that reason the code makes on the order of N^2 calculations of potential energy per MC step. In its present form, however, our code is not optimized to perform an efficient exploration of the energy landscape that becomes increasingly complicated for larger chains. We thus restricted our computations to solve the structure of small molecules. Otherwise, both thermodynamic and structural properties were obtained from an ensemble of about 100 configurations of $\sim10^9$ MC steps each one, requiring ~22 hrs of single processor time per configuration.

We carried out MC simulations of small RNA hairpins, which are among the most common secondary structure motifs. Moreover, understanding the folding of small molecules is a crucial step to describe the folding of larger RNAs. In order to validate the results of our model, we compared them with two distinct pairs of NMR structures that have the following features. The sequences of each pair of molecules are almost identical between them, except for a single nucleotide. In the first case, the PDB code of the experimental structures is 1esh ($^{5'}$GGUGCAUAGCACC$^{3'}$) [14] and 1jzc ($^{5'}$GGUGCAUGGCACC$^{3'}$) [15]. Both structures consist of a Watson-Crick base paired stem capped with a loop of three unpaired nucleotides ($^{5'}$AUA$^{3'}$ and $^{5'}$AUG$^{3'}$). On the other hand, we have also examined the sequences 1zif ($^{5'}$GGGCGAAAGCCU$^{3'}$) and 1zih ($^{5'}$GGGCGCAAGCCU$^{3'}$) [16], which are two hairpins characterized by the tetraloops ($^{5'}$GAAA$^{3'}$) and ($^{5'}$GCAA$^{3'}$), respectively. It should be mentioned that tetraloops are of great importance in the ribosome [17]. Our results are shown in Figs. 5 and 6. As a reference, we plot (left in both figures) the 2D secondary structures obtained with the program MFOLD of Zuker and Turner [18], which provides an estimate of the main interactions. In Figs. 5 and 6 we present our simulated molecules (center), together with the experimental RNAs (right). As it can be observed, the simulated structures are clearly similar to the experimental ones. The similitude is not only qualitative but also quantitative.

TABLE II. A comparison of the same characteristic distances that are studied in Table I for the experimental (1zif and 1zih) and simulated structures (s1zif and s1zih). The average energy $\beta\bar{E}$ of the simulated RNA molecules is also presented.

	$\beta\bar{E}$	$\bar{d}_{WC}$ (Å)	R_g (Å)	L_c (Å)	L_c/R_g
1zif		11.1 ± 0.4	8.4	52.8	6.3
s1zif	-14.5	11.4 ± 0.2	8.8	59.7	6.8
1zih		11.2 ± 0.2	7.7	50.6	6.6
s1zih	-15.7	11.3 ± 0.3	8.7	59.2	6.8

In Tables I and II, we compare some characteristic distances of the simulated molecules with the corresponding measurements done in the experimental RNAs. Such distances include the mean distance between the centers of mass of the Watson-Crick base pairs $\bar{d}_{WC}$, the radius of gyration $R_g=\sqrt{(\sum_{i=1}^{n}|\mathbf{r}_i-\mathbf{r}_c|^2)/n}$, the contour length $L_c=(\sum_{i=2}^{n}|\mathbf{r}_i-\mathbf{r}_{i-1}|)/n$, and the ratio L_c/R_g. Here, $\mathbf{r}_c$ is the position of the centroid of the RNA molecule with $n=13$ (Table I) and $n=12$ (Table II). First, it can be noticed that, within the error bars, the simulated values of $\bar{d}_{WC}$ are consistent with the experimental data. On the other hand, both the radius of gyration and the contour length are slightly overestimated in the simulated molecules (due to the fixed lengths $a_{\mu\gamma}$). In both cases, however, the ratios L_c/R_g are basically identical to the experimental ones.

IV. CONCLUSIONS

In conclusion, we have introduced a self-consistent model of RNA folding, which reduces, by construction, the number of accessible configurations in comparison with an all-atom model. Such reduction is possible because the interaction between the divers pairs of nucleotides is modeled with a series of effective potential functions that are characterized by a low number of potential minima. Of course, the folding problem is still nontrivial. We have tested our DPM in the case of numerical simulations of small RNA molecules. The model can reproduce, with a certain degree of accuracy, some structural features that are found in the experiments. Otherwise, our coarse-graining model could be used to accelerate the convergence of an all-atom model, provided that there are algorithms capable to do the reconstruction of the full atomic structure [19].

ACKNOWLEDGMENTS

Enlightening discussions with I. Tinoco Jr. are acknowledged. This work was supported by CONACyT (Mexico) under Grants No. 49486 and No. 60595.

[1] R. T. Batey, R. P. Rambo, and J. A. Doudna, Angew. Chem., Int. Ed. **38**, 2326 (1999).

[2] J. Chow, Z. Zen, S. Ziesche, and C. Brown, Annu. Rev. Genomics Hum. Genet. **6**, 69 (2005).

[3] Y. G. Yingling and B. A. Shapiro, J. Mol. Graphics Modell. **25**, 261 (2006).

[4] F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan, RNA **14**, 1164 (2008).

[5] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman, RNA **15**, 189 (2009).

[6] G. R. Bowman, X. Huang, Y. Yao, J. Sun, G. Carlsson, L. J. Guibas, and V. S. Pande, J. Am. Chem. Soc. **130**, 9676 (2008).

[7] P. González-Mozuelos and M. D. Carbajal-Tinoco, J. Chem. Phys. **109**, 11074 (1998); N. Bagatella-Flores and P. González-Mozuelos, *ibid.* **117**, 6133 (2002).

[8] P. Pliego-Pastrana and M. D. Carbajal-Tinoco, J. Chem. Phys. **122**, 244908 (2005).

[9] P. Pliego-Pastrana and M. D. Carbajal-Tinoco, J. Phys. Chem. B **110**, 24728 (2006).

[10] P. Pliego-Pastrana and M. D. Carbajal-Tinoco, Phys. Rev. E **68**, 011903 (2003).

[11] D. A. McQuarrie, *Statistical Mechanics* (Harper, New York, 1976).

[12] See supplementary material at http://link.aps.org/supplemental/10.1103/PhysRevE.81.041914 for the set of EPPs fits.

[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[14] C.-H. Kim, C. C. Kao, and I. Tinoco, Jr., Nat. Struct. Biol. **7**, 415 (2000).

[15] C.-H. Kim and C. C. Kao, RNA **7**, 1476 (2001).

[16] F. M. Jucker, H. A. Heus, P. F. Yip, E. H. Moors, and A. Pardi, J. Mol. Biol. **264**, 968 (1996).

[17] C. R. Woese, S. Winker, and R. R. Gutell, Proc. Natl. Acad. Sci. U.S.A. **87**, 8467 (1990).

[18] M. Zuker, Nucleic Acids Res. **31**, 3406 (2003); D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, J. Mol. Biol. **288**, 911 (1999).

[19] M. A. Jonikas, R. J. Radmer, and R. B. Altman, Bioinformatics **25**, 3259 (2009).