Supporting Information to

Computer folding of RNA tetraloops? Are we there yet?

Petra Kührová,[§] Pavel Banáš,^{§†} Robert B. Best,[‡] Jiří Šponer,^{†,&} and Michal Otyepka^{*§†}*

[§] Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacky University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic e-mail: michal.otyepka@upol.cz, pavel.banas@upol.cz

[‡] Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

[†] Institute of Biophysics, Academy of Sciences of the Czech Republic, Kralovopolska 135, 612 65 Brno, Czech Republic

[&] CEITEC – Central European Institute of Technology, Masaryk University, Campus Bohunice, Kamenice 5, 625 00 Brno, Czech Republic

*corresponding authors

The replica exchange setup and analysis of exchange ratio

The temperature ranges for both tetraloops were slightly different. The temperature ladder for each tetraloop was dictated by targeting a given exchange rate between consecutive replicas. We set the acceptance ratio to be approximately 20-30%. The real acceptance ratio during REMD simulation was indeed between 21% and 30% (see Tables S1 and S2). For verification of a traveling of structures between temperatures, we analyzed the time evolution of temperatures of replicas for UUCG TL starting from folded state (Figures S1-S6). Our results show that the most replicas sampled the whole temperature space and the majority of replicas traveled repeatedly between the lowest and the highest temperatures and vice versa.

Table S1. The acceptance rate for the replica exchanges between replicas of UUCG TL starting from folded structure over 400 ns.

Exchange between replicas	Acceptance rate	Exchange between replicas	Acceptance rate
0↔1	0.26	24↔25	0.27
1↔2	0.26	25↔26	0.27
2↔3	0.21	26↔27	0.28
3↔4	0.21	27↔28	0.29
4↔5	0.21	28↔29	0.26
5↔6	0.22	29↔30	0.26
6↔7	0.24	30↔31	0.27
7↔8	0.24	31↔32	0.28
8↔9	0.25	32↔33	0.29
9↔10	0.26	33↔34	0.26
10↔11	0.26	34↔35	0.26
11↔12	0.28	35↔36	0.27
12↔13	0.28	36↔37	0.28
13↔14	0.29	37↔38	0.29
14↔15	0.30	38↔39	0.26
15↔16	0.30	39↔40	0.24
16↔17	0.26	40↔41	0.30
17↔18	0.26	41↔42	0.29
18↔19	0.27	42↔43	0.30
19↔20	0.23	43↔44	0.30
20↔21	0.23	44↔45	0.30
21↔22	0.24	45↔46	0.29
22↔23	0.25	46↔47	0.30
23↔24	0.26

Figure S1-S6: The time evolution of exchanges of replicas in the temperature space for UUCG TL starting from folded state.

Table S2. The acceptance rate for the replica exchanges between replicas of GAGA TL starting from folded structure over 400 ns.

Exchange between replicas	Acceptance rate	Exchange between replicas	Acceptance rate
0↔1	0.25	24↔25	0.23
1↔2	0.27	25↔26	0.26
2↔3	0.27	26↔27	0.27
3↔4	0.28	27↔28	0.27
4↔5	0.29	28↔29	0.28
5↔6	0.30	29↔30	0.25
6↔7	0.25	30↔31	0.27
7↔8	0.25	31↔32	0.27
8↔9	0.26	32↔33	0.23
9↔10	0.23	33↔34	0.27
10↔11	0.23	34↔35	0.28
11↔12	0.25	35↔36	0.26
12↔13	0.25	36↔37	0.27
13↔14	0.26	37↔38	0.28
14↔15	0.27	38↔39	0.28
15↔16	0.28	39↔40	0.25
16↔17	0.28	40↔41	0.26
17↔18	0.25	41↔42	0.28
18↔19	0.26	42↔43	0.26
19↔20	0.27	43↔44	0.27
20↔21	0.29	44↔45	0.27
21↔22	0.24	45↔46	0.28
22↔23	0.25	46↔47	0.29
23↔24	0.27

Figure S7. Conformational transition maps between folded, misfolded and unfolded states of the investigated UUCG (a) and GAGA (b) tetraloops (TLs). The states are depicted as secondary structure schemes with base-pair and base-phosphate interactions annotated according to the standard classification.^{1,2} Arrows between clusters indicate conformational transitions observed during the REMD simulations. The color of arrows indicate frequency of transitions: red corresponds to transitions frequently observed in our REMD simulations and blue stands for rare but still observed transitions. Number of transition is given above arrow. Each misfolded state is accompanied by the average RMSD from the native state (shown in red). The loop residues are labeled as L₁–L₄. G_{L4} of the UUCG tetraloop having *syn* orientation is shown in red. The nucleotides in circle are flexible and their positions are not well-defined compared to the other nucleotides in the TL.

Figure S8. The density plot of relative population of conformational states of the (a) UUCG and (b) GAGA TLs with x-axis and y-axis corresponding to RMSD of the stem bases and four

TL nucleotides, respectively. The RMSD is measured from the native reference tetraloop structure during all REMD simulations starting from the folded state. The red lines correspond with RMSD of experimental structures of each tetraloop (labeled according to PDB ID).

Figure S9. Relative populations of folded, misfolded and unfolded states as a function of temperature calculated from REMD started from the folded structures. However, note that these populations cannot be considered as converged (see the main text for more details)

UUCG

GAGA

0-100 ns

0-200 ns

0-300 ns

0-400 ns

Figure S10. The convergence of relative populations of folded (blue), misfolded (gray) and unfolded (black) states as a function of temperature and cumulative simulation time in REMD simulation started from the folded state. While rapid decrease of the population of folded state in expense of unfolded state (especially at high temperatures) occurs on first tens of ns, subsequent extension of the simulation time up to 400ns only slightly modulates the population profiles. Still, decrease of the population of the native state with time is visible even between 200-300 and 300-400 ns intervals.

Rarely observed non-native structures of studied TLs

As discussed in the main text, both GNRA and UUCG TLs have precisely defined structures that can be repeatedly found in the X-ray and NMR structures of different types of RNA and within different structural context. However, some rare exceptions can be also found.

In case of UUCG TL we found two different X-ray structures (PDB ID 3AM1 and 1P6V) having non-native conformation of TL that corresponds rather to misfolded-anti state that was identified in our REMD simulations, i.e. the G_{L4} guanine is in *anti*- rather than native *syn*-conformation (Figure S11). On the other hand, it should be noted that both these structures have rather poor resolution of the UUCG TL (having G_{L4} in *anti*) with high b-factors reported for this region. Thus it cannot be ruled out that these conformations of the UUCG TL in these particular X-ray structures might be compromised by limited resolution.

Figure S11. The three dimensional structures of cUUCGg with G_{L4} in *anti*-conformation resolved by X-ray crystallography: A) Structure from complex of O-Phosphoseryl-tRNA kinase with anticodon-stem/loop truncated tRNA(Sec) (PDB ID:3AM1, resolution 2.4 Å),³ B) Structure

from complex of the tRNA domain of transfer-messenger RNA with small protein B (PDB ID:1P6V, resolution 3.2 Å)⁴ and C) its structural environment. The UUCG TL in 3AM1 structure does not have any tertiary contacts.

Figure S12. The three dimensional structures of the misfolded compact structure of cUUUCGg populated in REMD simulations.

In contrast, in case of GNRA TL, there is clear experimental evidence of structural adaptation of TL upon binding of ribotoxin restrictocin. The X-ray structure (PDB ID 1JBR) of GNRA bound to restrictocin (Figure S13) shows the topology of GNRA TL similar to conformation of misfolded-bulged state as observed in our REMD simulation.

Figure S13. A) Structure of cGAGAg TL from complex of the ribotoxin restrictocin and a 31-mer sarcin-ricin domain (PDB ID:1JBR, resolution 2.15 Å)⁵ and B) its structural context.

Calculation of free energy difference between native (G_{L4-syn}) and misfolded-anti ($G_{L4-anti}$)

In our REMD simulations we observed preferential formation of UUCG misfolded-anti state over the native state. On the other hand, it is obvious that such preferential formation of misfolded state must be artifact of the simulations as there are clear experimental evidences that the native structure is the global free energy minimum (e.g. high resolution NMR structure, ref. 69 in the main text). There are two possible explanations for preferential formation of the misfolded-anti state: (i) It is a consequence of limited sampling (despite relatively long time scale, 400 ns, in combination with REMD technique effectively increasing sampling, the folding events are still very rare and simulations are far from convergency in this aspect), or (ii) it might be also caused by artificial overstabilization of misfolded anti state in force field, i.e. it might be a force field artifact. In order to test the latter possibility, we tried to estimate the free energy difference between native and misfolded anti state (under force field description).

During the transition between native and misfolded anti state, the G_{L4} base must lose its contact to U_{L1} , bulge out, flip from *syn* to *anti* (or vice versa), and reestablish the U_{L1} - G_{L4} base pair contact. Thus, the transition pathway corresponding to the G_{L4} *syn/anti* flip is rather complicated and any method calculating free energies using explicit definition of the reaction coordinate (such as umbrella sampling, potential of mean force integration or adaptive biasing force) necessarily

suffers by insufficient sampling of bulged state and challenging description of the reaction coordinate. To our best knowledge, the only method avoiding sampling of bulged state is thermodynamic integration, in which G_{L4} in *syn*-conformation is disappearing, while G_{L4} in *anti*- is appearing with lambda parameter coupling hamiltonians of UUCG TL with G_{L4} in *anti* and *syn*, respectively. Unfortunately, the G_{L4} *syn/anti* flip significantly changes the structure and shape of TL. As a consequence, a relatively big disappearing/appearing region must be taken into account (Figure S14). This, however, pushes the studied system toward (if not beyond) limits of applicability of thermodynamic integration method and compromises its accuracy. Namely, at the low and high lambda values (near lambda of 0 or 1), where either *syn*- or *anti*-states are about to disappear, the estimated mean derivative of total energy with respect to lambda is most likely highly inaccurate (see mean derivative of total energy with respect to lambda as a function of lambda in right panel of Figure S15).

Figure S14. Disappearing/appearing region used for thermodynamic integration between native G_{L4} -*syn* and misfolded-*anti* (G_{L4} -*anti*) states of UUCG TL is shown in stick. Blue lines represent structure of misfolded-*anti* state.

In order to estimate at least some portion of the inaccuracy, we performed a bootstrap analyzes, where we estimated a confidence interval of each $dV/d\lambda$ mean value. We used fifty 100-ps-long frames. Data of each frame were averaged in order to decorrelate the data set. Subsequently, the obtained frame averages were used for calculation of the half width of total average confidence interval (Fig. S15, left panel). Finally, the total accuracy was estimated as a square root of weighted sum of half width squares for each lambda, where the same weighting coefficient as used in Gaussian quadrature were used. Thus final estimate of free energy

difference between G_{L4} *syn* and *anti* state equal to $\sim 15 \pm 8$ kcal/mol. However, we would like to stress that the estimate of accuracy does not take into account effects coming from too large appearing/disappearing region (i.e. inherent limitation of used method), which is supposed to be rather systematic and does not contribute to the fluctuations of derivative of total energy with respect to lambda. On the other hand, the contributions of middle lambda values to the total free energy difference are still positive, which allow us to suggest that despite all inaccuracies, it is safe to predict that total free energy difference is positive and thus the native system is more stable (as described by the force field) than the misfolded-anti state.

Taken together, it is challenging to estimate the free energy difference between native and misfolded-anti state. To our best knowledge, the thermodynamic integration method represents the best but still very limited choice among all available methods for calculation of free energy differences. The inaccuracy of estimated result ($\sim 15 \pm 8$ kcal/mol) contains only effects of limited sampling while not including the effects connected with large appearing/disappearing region. Thus it is likely that the total inaccuracy might be even bigger than the estimated 8 kcal/mol. On the other hand, when looking to those lambdas, where size of the appearing/disappearing region does not matter, their contribution is still positive and thus we can suggest that the total free energy difference most likely remains positive, i.e., the native system is the global minimum even under force field description and the preferential formation of misfolded-anti state is most likely caused by insufficient sampling.

Figure S15. Mean derivative of total energy with respect to lambda parameter coupling hamiltonians of UUCG TL with G_{L4} in *anti* and *syn*, respectively, obtained by thermodynamic integration at 298 K (right panel). The total free energy difference between UUCG misfolded-anti conformation having G_{L4} in *anti* and native conformation with G_{L4} in *syn* can be obtained by integration of this curve over lambda using Gaussian quadrature method, which results in a sum of weighted mean derivative of total energy with respect to lambda (left panel) and equals $\sim 15 \pm 8$ kcal/mol (thus native state being favored even in force field).

Calculation of free energy of folding and melting temperatures from REMD

In the main text we argued that after careful analyzes of our REMD simulations, we have to state that these simulations are not sufficiently converged to allow accurate estimation of melting temperatures and free energy of folding. Nevertheless, despite such inaccuracy, we decided to apply similar approach for calculation of melting temperature and ΔG_{unfold} between the native and the unfolded state as used in previous study,⁶ i.e., to calculate the ΔG_{unfold} from relative populations of the native state using following equation.

$$\Delta G_{\text{unf.}} = -RT \ln \left(\frac{1 - f_{\text{folded}}}{f_{\text{folded}}} \right),$$

In case of UUCG TL we used both our definition of native folded state as well as the definition of folded state used in the recent study, i.e., the native state corresponding to all structures with RMSD from NMR structure below 4 Å.⁷ In case of GAGA TL we used only our definition (see Methods in main text). From the dependence of ΔG_{unfold} on temperature, the melting temperature can be easily estimated as the temperature, where ΔG_{unfold} equals to zero (Figure S16). The experimental melting temperatures for the UUCG and GAGA TLs with identical sequence as the systems used in this study is 342 K⁸ and 324 K⁹, respectively, while our estimates equal to 328 K and 344 K, respectively (assuming only our definition of native state). However, we would again like to stress that our REMD simulations are still not fully converged and the value of ΔG_{unfold} at the given temperature should be rather assumed as a semi-quantitative estimate. As a consequence of above mentioned inaccuracy and as the ΔG_{unfold} decreases relatively slowly with temperature, the estimate of melting temperature from REMD simulation should be even more inaccurate.

Figure S16. The free energy of unfolding as a function of temperature. Red and blue solid lines correspond with calculations with the natives states according our strict definition (see Methods in the main text), while the red circles correspond to the data obtained when using definition of native state according to Garcia and Paschek.

Despite the above mentioned inaccuracies, the ΔG_{unfold} of both TLs estimated from our REMD simulation reach positive values at lowest temperatures (Figure S16). Together with the estimation of the melting temperatures above 310 K, it suggests that the TLs under approximate force field description (using most recent ff99bsc0 χ_{OL3} force field) are stable at room temperature in agreement with experimental observations (e.g. for UUCG the unfolding free energy of 3.7 kcal/mol was reported⁸). This finding is in a sharp contrast with previous REMD studies^{6,7} carried out using older ff99 force field, which reported negative ΔG_{unfold} values indicating underestimated stability of native fold of TL in the simulations. The negative unfolding free energies were explained by two possible reasons.¹⁰ First coming from the difference of the salt concentrations used in simulations and experiments and the other one corresponding to the absolute convergence of recent REMD simulations. We may simply add that another reason for this observation might stem from insufficient quality of force field parameters. An inherent instability of GNRA and UUCG TLs caused by force field imbalances was indeed reported in our recent study.^{11,12}

Note that here we used a revised version of ff99 force field, namely ff99bsc0 χ_{OL3} . This reparameterized AMBER force field involves corrected parameters of glycosidic χ torsion and prevents artificial transition of canonical A-RNA duplex into meaningless ladder-like structure.¹³ Recently, we have shown that UNCG and GNRA TLs are prone to the structural degradation in ff99 force field.¹¹ Therefore we suggest that REMD simulations using ff99 force field were biased by serious force field artifacts. On the other hand, when using the refined ff99bsc0 χ_{OL3} force field, the REMD simulations seem to be able to correctly describe both studied TLs as thermodynamically stable at room temperature albeit still some potential artificial influence of force field are apparent (e.g. similar propensity of UUCG and GAGA TLs for making misfolded-bulged state, which is according the experimental structural data rather unlikely).

References

- (1) Leontis, N. B.; Westhof, E. *RNA* **2001**, 7, 499.
- (2) Zirbel, C. L.; Sponer, J. E.; Sponer, J.; Stombaugh, J.; Leontis, N. B. *Nucleic Acids Res* **2009**, 37, 4898.
- (3) Sherrer, R. L.; Arais, Y.; Aldag, C.; Ishitani, R.; Ho, J. M.; Soll, D.; Nureki, O. *Nucleic acids research* **2011**, 39, 1034.
- (4) Gutmann, S.; Haebel, P. W.; Metzinger, L.; Sutter, M.; Felden, B.; Ban, N. *Nature* **2003**, 424, 699.
- (5) Yang, X. J.; Gerczei, T.; Glover, L.; Correll, C. C. *Nat. Struct. Biol.* **2001**, 8, 968.
- (6) Zuo, G.; Li, W.; Zhang, J.; Wang, J.; Wang, W. *J. Phys. Chem. B* **2010**, 114, 5835.
- (7) Garcia, A. E.; Paschek, D. *J. Am. Chem. Soc.* **2008**, 130, 815.
- (8) Williams, D. J.; Hall, K. B. *J. Mol. Biol.* **2000**, 297, 1045.
- (9) Hernandez, B.; Baumruk, V.; Leulliot, N.; Gouyette, C.; Huynh-Dinh, T.; Ghomi, M. *J. Mol. Struct.* **2003**, 651, 67.
- (10) Deng, N. J.; Cieplak, P. *Biophys. J.* **2010**, 98, 627.

- (11) Banas, P.; Hollas, D.; Zgarbova, M.; Jurecka, P.; Orozco, M.; Cheatham, T. E.; Sponer, J.; Otyepka, M. *J. Chem. Theory Comput.* **2010**, 6, 3836.
- (12) Zgarbova, M.; Otyepka, M.; Sponer, J.; Mladek, A.; Banas, P.; Cheatham, T. E., 3rd; Jurecka, P. *J. Chem. Theory Comput.* **2011**, 7, 2886.
- (13) Ditzler, M. A.; Otyepka, M.; Sponer, J.; Walter, N. G. *Acc. Chem. Res.* **2010**, 43, 40.