Cite this: DOI: 00.0000/xxxxxxxxxx

Salt dependent mesoscopic model for RNA with multiple strand concentrations†

Izabela Ferreira,[a,b] Tauanne D. Amarante,[c] and Gerald Weber‡[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Mesoscopic models can be used for the description of the thermodynamic properties of RNA duplexes. With the use of experimental melting temperatures, its parametrization can provide important insights into its hydrogen bonds and stacking interactions as has been done for high sodium concentrations. However, the RNA parametrization for lower salt concentrations is still missing due to the limited amount of published melting temperature data. While the Peyrard-Bishop (PB) parametrization was found to be largely independent of strand concentrations, it requires that all temperatures are provided at the same strand concentrations. Here we adapted the PB model to handle multiple strand concentrations and in this way we were able to make use of an experimental set of temperatures to model the hydrogen bond and stacking interactions at low and intermediate sodium concentrations. For the parametrizations we make a distinction between terminal and internal base pairs, and the resulting potentials were qualitatively similar as we obtained previously for DNA. The main difference from DNA parameters, was the Morse potentials at low sodium concentrations for terminal r(AU) which is stronger than d(AT), suggesting higher hydrogen bond strength.

1 Introduction

RNA plays an essential role in many cellular processes such as transcription, translation, and conservation of genetic information. Double stranded (ds) RNAs are present in cells and perform a variety of biological functions.[1,2] For instance, small non-coding dsRNA that mediate neuronal differentiation,[3] dsRNA segments of special lengths, known as siRNA, can inhibit the translation of mRNA molecules into proteins through attaching to mRNAs,[4,5] and RNAs of more than 30 base pairs of length can be key activators of the innate immune response against viral infections.[6]

Similarly to dsDNA, the interchain interactions stabilizing the structure of dsRNA are very sensitive to environmental conditions such as temperature and salt concentration.[7–10] For example, a reduction in salt concentration increases the binding affinity between the protein kinase R (PKR) and the dsRNA, improving the recognition pathway.[11]

dsRNA form helices in an A-form which has a much deeper/narrower major groove and a wider minor groove than the B-form of dsDNA, which concedes a very different surface electrostatic potential for dsDNA and dsRNA.[12] These different ion binding modes for dsDNA and dsRNA have been suggested to be responsible for the different multivalent ion-dependent condensation behaviours[13] and flexibilities for dsDNA and dsRNA.[14,15] Therefore, the presence of mono and divalent cations plays a fundamental role in the stabilization of RNA secondary and tertiary structures by neutralizing the negative charge and reducing the repulsion of the phosphates.[16,17] Although magnesium ions are much more stabilizing,[18–20] monovalent ions like sodium are important and the general conclusion is that sodium ions are essential as they mediate the long-range interactions that are crucial for folding and assembly of RNA tertiary structures.[21,22]

There is some NMR evidence that group I monovalent ions, Na$^+$ and K$^+$ in particular, remain well hydrated in the presence of RNA[23] interacting with it on a diffuse way[24,25] or may even be chelated by irregular RNA structures.[26,27] Those factors may relate the sensibility of RNA tertiary structure to the size of the monovalent cations that are present, in contrast to the weak discrimination shown by DNA helices in their interaction with different group I ions.[28] Another aspect that affects the thermodynamic stability of the double stranded duplex is a process known as "base fraying", which is the breaking of base-pairing interactions at the termini of a RNA or DNA. Frayed states are intermediaries in zipping and unzipping processes and have been suggested to be important for the interactions of RNA with proteins,[29,30] are required for secondary structure rearrangements for riboswitch function,[31] and may be relevant for strand migration.[32]

The effect of monovalent ions in RNA has been investigated with several theoretical methods, such as molecular dynamics (MD),[14,33–36], coarse-grained models,[37] Debye-Hückel models,[38] and tightly-bound ion theory.[16,39] For instance, MD simulations such as made by Beššeová *et al.*[33 33,34] concluded that the force field and salt effects are sequence-dependent and the helix compactness is sensitive to the salt and water conditions. Salt effects and stability on the tridimensional structure of RNA

[a] *Departamento de Física, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brazil; E-mail: gweberbh@gmail.com*

[b] *Programa Interunidades de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brazil*

[c] *MRC Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre, Cambridge Biomedical Campus, Cambridge, UK*

† Electronic Supplementary Information (ESI) available: Supplementary Tabs. S1— S5 show the number of internal and terminal parameters. Supplementary Tabs. S6— S8 show the summary of logarithmic grouping for coarseness f. Supplementary Tabs. S9— S12 show the number of elements per group and its respective C_t. The validation sequences are shown in Tab. S13. Figs. S1— S14 contain the average regression parameters for the rounds EU for each L_f. Figs. S15— S28 show the Morse potentials for each L_f over the range of salt concentration. Figs. S29— S42 show the stacking potentials for each L_f over the range of salt concentration.

were also explored by Monte-Carlo simulations and an increase in Na$^+$ concentration tend to improve the folding of RNA hairpins, suggesting that the base-pair adjacent to the terminal is not stable due to the reduction of stacking.[40] Debye-Hückel models concluded that a decrease in salt concentration generally destabilize the folding of RNA and lowers its denaturation temperatures.[37,38] Mesoscopic modelling, based on the Peyrard-Bishop (PB) description, using experimental melting temperatures as input data, have been restricted to high sodium concentrations.[41] Existing RNA melting temperature data at lower sodium concentrations exists at varying strand concentrations, however the mesoscopic approach requires all temperatures to be at a single strand concentration.[42] Here, we extend this mesoscopic model to handle multiple strand concentration, thus overcoming the current limitations of this approach.

Base fraying is an important, yet still poorly understood aspect of RNA stability, in particular it is unclear how fraying dependents on salt concentration. Melting temperature measurements indicate that the 5′ ends are substantially more stable when the purine is positioned at the 3′ end, which determine the stability of sequential mismatches as well.[43] NMR measurements concluded that the opening and closing rates of r(AU) base-pairs are much larger than those observed for d(AT), despite comparable stability.[44] MD simulations have had difficulties to deal with base fraying as existing force fields were inadequate for terminal AU bases.[45] However, more recently this limitation seems to have been resolved and Pinamonti *et al.*[46] concluded that 5′ ends containing UApCG or AUpGC have a slower fraying due to a larger stability assigned to stacking interactions. This suggests that terminal adenine base pairs have stronger stacking interaction when compared with uracils.[46] In contrast to MD, for mesoscopic PB models[47] and coarse-grained models,[48] end-fraying is well represented and they have in principle no difficulty in dealing with AU terminal pairs.[41,49] Nearest-neighbour (NN) models are typically limited to temperature prediction, and terminal effects are included as an energy penalty.[50] The salt dependence of these terminal factors were studied by us recently,[51] and we observed a marked quadratic dependence in the enthalpies and entropies with salt concentration which are compensated to form almost linear Gibbs free energies.[51]

Here, we adapt the mesoscopic PB model to RNA with varying salt dependence, multiple strand concentrations and including terminal effects. In part we applied a similar approach as from our previous work on DNA salt-dependent terminal effects,[52] which enables us to compare RNA and DNA terminal effects and discuss their differences. However, for RNA the available melting temperatures are scattered into a non-uniform range of strand concentrations, see Fig. 1.[51] This represents a challenge for the mesoscopic model which usually requires that all temperatures are at the same concentration.[53] The reason for this is that the PB model is a single molecule calculation, and the melting temperatures are correlated to experimental values at a single strand concentration.[54] To work with the existing set of temperatures we adapted the model to handle multiple strand concentrations simultaneously. To achieve this we grouped the strand concentrations into logarithmic groups and then worked out the corresponding model parameters. We tested various levels of grouping and, surprisingly, the model parameters had very little dependence on the grouping factors. Once we established the best level of grouping we were able to compare the new salt-dependent parameters to our previous DNA parameters. In general, we found that the Morse potential representing the hydrogen bonds of RNA followed very closely that of DNA, except for low salt concentrations where d(AT) had an important reduction which we did not observe for r(AU).

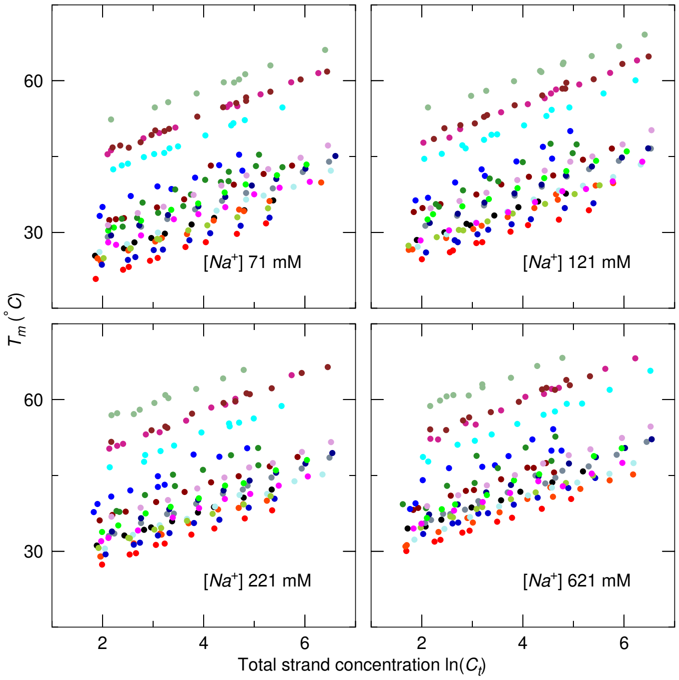

Fig. 1 Experimental melting temperatures from Ref. 51 as logarithmic function of total strand concentration $\ln(C_t)$. Each colour represents a specific sequence.

2 Methods

2.1 Model

The configurational part of the PB Hamiltonian is written as[47,55]

$$U_{i,i+1} = \frac{k_{\alpha,\beta}}{2}\left(y_i - y_{i+1}\right)^2 + D_\alpha\left(e^{-y_i/\lambda_\alpha} - 1\right)^2, \tag{1}$$

which describes the interaction of a base pair of type α, at sequence position i, with its nearest-neighbour of type β at position $i+1$. The Morse potential, which describe the hydrogen bond between the base pairs, uses two more parameters to characterize its depth and width of the ith base pair of type α, D_α, λ_α, respectively. The stacking interaction between adjacent base-pairs or the nearest-neighbours is represented by an elastic constant $k_{\alpha,\beta}$, and the coordinate y represents the relative displacements between the bases.

Therefore the sum for the Eq. (1) over all N base-pairs is carried out using its partition function:

$$Z_y = \int_{y_{min}}^{y_{max}} dy_1 \int_{y_{min}}^{y_{max}} dy_2 \cdots \int_{y_{min}}^{y_{max}} dy_N \int \prod_{n=1}^{N} e^{-\beta U(y_i,y_{i+1})} \tag{2}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant and T the absolute temperature. Subsequently, the integral over all possible configurations of base pair displacements, y_i is performed. Thus, all possible Morse potentials and stacking interactions are considered simultaneously during the evaluation. From the partition function, Eq. (2) an adimensional index τ is calculated and it is directly correlated to the experimental melting temperatures as we will see in the next sections.[42]

Furthermore, the average base pair displacement, $\langle y_m \rangle$, at the mth position in the sequence can be obtained from

$$\langle y_m \rangle = \frac{1}{Z_y} \int_{y_{min}}^{y_{max}} dy_1 \int_{y_{min}}^{y_{max}} dy_2 \cdots \int_{y_{min}}^{y_{max}} dy_N y_m \int \prod_{n=1}^{N} e^{-\beta U(y_i, y_{i+1})} \qquad (3)$$

2.2 Notation

To reliably distinguish terminal base pairs from internal base pairs we need to establish an unambiguous notation. Consider the following example sequence

$$\begin{matrix} 5'\text{-AGAUAUCU-}3' \\ 3'\text{-UCUAUAGA-}5' \end{matrix}$$

where we separate the terminal and internal base pairs

$$\underbrace{\begin{matrix}\text{A}\\ \text{U}\end{matrix}}_{\text{terminal}} \cdot \underbrace{\begin{matrix}\text{GCAAGU}\\ \text{CGUUCA}\end{matrix}}_{\text{internal}} \cdot \underbrace{\begin{matrix}\text{U}\\ \text{A}\end{matrix}}_{\text{terminal}}$$

For the terminal base pairs we will use a superscript *, in our example this would be AU* at the 5′-side and UA* at the 3′-side. For Morse potentials, AU* is equivalent to UA* , as well as CG* is to GC*, and will share the same parameters D and λ, see Eq. (1).

For the nearest-neighbour (NN) stacking parameter k there will be a mixed notation of terminal and internal base pairs. The first NN pair of our example sequence would be AU*pGC, that is a terminal AU* followed by an internal GC. The AU*pGC pair is symmetric to CGpUA*:

$$\begin{matrix} 5'\text{-}\textbf{AG}\text{AUAUCU-}3' \\ 3'\text{-}\textbf{UC}\text{UAUAGA -}5' \end{matrix} \leftrightarrow \begin{matrix} 5'\text{-AGAUAU}\textbf{CU}\text{-}3' \\ 3'\text{-UCUAUA}\textbf{GA}\text{-}5' \end{matrix}$$

As both can be described by the same stacking parameter k, we keep just the one that precedes alphabetically, in this case AU*pGC. Therefore, as stated above, the stacking parameter for AUpGC NN pairs will be divided into three separate parameters, namely AU*pGC for terminal AU*, AUpGC* for terminal CG* and the internals which we maintain the original notation AUpGC.

Some base pairs will have only one additional terminal-related parameter as a result of the NN pair symmetry. Such as CGpGC which has only one terminal related NN CGpGC* since it is symmetric to CG*pGC.

2.3 Melting temperature sets

The melting temperatures used here fall into two very different categories: four are at lower sodium concentrations where all sequences are self-complementary and of the same length; and a single one at high sodium concentrations with a mix of self-complementary and non-self-complementary sequences and variable lengths. This requires different theoretical approaches depending on the type of temperature set, and therefore we will distinguish them by low salt (LS) and high salt (HS).

2.3.1 (LS) 71 to 621 mM [Na$^+$]

We used the set of RNA melting temperatures from Ferreira *et al.*[51], consisting of 18 RNA duplexes at four different [Na$^+$] concentrations (71, 121, 221, 621 mM). For each sequence and salt concentrations there are at least 9 measurements at different strand concentrations C_t in the range of 5 μM to 700 μM, see Fig. 1. All sequences are self-complementary and either are 6 or 8 base-pairs (bp) in length.

2.3.2 (HS) 1021 mM [Na$^+$]

For the higher salt concentration we used the melting temperature set from Xia *et al.*[50] which was complemented by two sequences from Chen and Znosko[56]. Unlike the LS data, they have varying lengths, include non-self-complementary sequences, are at the same strand concentration and from various sources.

2.4 Temperature correlation with melting index

The PB model describes the thermodynamics via a coefficient τ_i, obtained from the partition function Eq. (2), for the ith duplex in the data set which provides temperature prediction T_i'

$$T_i'(P) = a_0 + a_1 \tau_i(P) \qquad (4)$$

where the coefficients $a_{0,1}$ are calculated via a linear regression of the experimental melting temperatures T_i at a single strand concentration C_t, and P is a set of tentative model parameters. The regression of the coefficients used in Eq. (4) is typically carried out at a single strand concentration C_t. However, for the LS dataset, there are multiple strand concentrations which require a different approach as we will discuss next.

2.4.1 LS temperature regression

Previous studies have confirmed that the resulting model parameters are independent on the strand concentration.[41,53] However, when the melting temperature set involves multiple strand concentrations the regression of Eq. (4) needs to be carried separately for each concentration C_t, that is

$$T_i'(P) = a_0(C_t) + a_1(C_t)\tau_i(P) \qquad (5)$$

where the coefficients $a_{0,1}$ are now functions of C_t, which requires a minimum amount of melting temperatures for each value of C_t as otherwise the regression calculation cannot be carried out. In other words, there needs to be subsets of melting temperatures grouped to the same C_t. However, here the dataset has measurements scattered over a wide range of C_t and there is no single subset was measured at the same C_t, see Fig. 1. This does not represent a problem for the nearest-neighbour model,[51] but for the PB model it becomes necessary to group the melting temperatures together to the closest value of C_t.

Table 1 Summary of logarithmic grouping for coarseness f. Shown are the number of groups n_f, total number of grouped elements N_f, and total number of ungrouped elements U_f for $[Na^+]$ 71mM.

f	n_f	N_f	U_f
5.0	22	179	6
4.0	18	180	5
3.0	14	182	3
2.0	10	184	1
1.5	8	184	1
1.4	7	184	1
1.3	7	183	2
1.2	7	184	1
1.1	6	184	1
1.0	5	182	3
0.9	5	184	1
0.8	5	182	3
0.7	4	184	1
0.6	4	184	1
0.5	3	184	1
0.4	3	185	0

2.4.2 Strand concentration grouping

Since the melting temperatures scale with $\ln(C_t)$,[57] it makes sense to introduce a logarithmic group index

$$L_f = \frac{\text{round}[f \ln(C_t/C_0)]}{f} \tag{6}$$

where f is a factor that controls the coarseness of groups, and C_0 is a fixed reference concentration taken as 1 μM to ensure that L_f is adimensional. As we will perform a linear regression for each group, we only consider groups with at least 3 elements. For each available melting temperature we work out to which group L_f it belongs depending on its C_t and the coarseness factor f which results in n_f groups with a total of N_f members. A small f will create a small number of groups n_f with many elements, while a large f results in many groups with few elements. The upper limit of f is when there are too few melting temperatures per group to perform a meaningful linear regression (at least 3 elements), and the lower limit of f is when there is only a single group that contains all temperatures. Table 1 shows the summary of the logarithmic grouping L_f that is considered in this work for $[Na^+]$ 71 mM. See supplementary tables S6, S7 and S8 for a summary of the remaining LS salt concentrations, and a detailed breakdown in supplementary tables S9–S12.

Using the logarithmic grouping, we now replace Eq. (5) with

$$T_i'(P) = a_0(L_f) + a_1(L_f)\tau_i(P). \tag{7}$$

and the regression coefficients are obtained independently for each group L_f.

2.5 HS temperature regression

For the HS data, which are all given at the same strand concentration and are available at varying sequence lengths N, the linear regression is performed separately for each group of base pair length N[54]

$$T_i'(P) = a_0(N) + a_1(N)\tau_i(P), \tag{8}$$

similarly as used in our previous work,[41] and gives better results than the single regression Eq. (4).

2.6 Optimization

The parameter sets P needed for the calculation of the melting index $\tau_i(P)$, Eqs. (7) and (8), contains the model parameters used in the Hamiltonian Eq. (1) for each type of base pair and nearest-neighbour present in the sequence set. Therefore, we will need to find the optimal set of L parameters, $P_j = \{p_1, p_2, \ldots, p_L\}$ that will provide the temperature predictions $T_i'(P)$ that are closest to the experimental temperatures T_i. The P parameters are varied until we minimize the squared difference

$$\chi_j^2 = \sum_{i=1}^{M} \left[T_i'(P_j) - T_i\right]^2. \tag{9}$$

where P_j is the jth tentative set of parameters and M is the number of experimental melting temperatures. Each parameter within the P_j is sampled between $0.1p_u$ and $1.1p_u$, where p_u is the uniform parameter calculated previously for high salt concentration[53]. For LS we use $M = N_f$ which is the total number of grouped temperatures for a given coarseness factor f. The numerical parameter optimization is performed by a downhill simplex multidimensional minimization algorithm.[58] We will also refer to another quality parameter which is average melting temperature deviation

$$\langle \Delta T \rangle = \frac{1}{M} \sum_{i=1}^{M} \left|T_i'(P_j) - T_i\right|. \tag{10}$$

As a result of the terminal/internal (T/I) notation, we will be dealing with 4 Morse potentials (2 internal, 2 terminal) and 26 NN stacking potentials (10 internal and 16 terminal), representing $L = 30$ parameters. For comparison, we will also perform all calculations without the distinction between terminal/internal which we will call uniform (UN) parameters and represents $L = 12$ variables. Next, we will detail the optimization steps used here.

2.6.1 MR1 (LS and HS)

The first minimization round (MR1) of the parameter optimization was performed by varying the initial Morse and stacking potentials[41,53] randomly over an interval which averages to the initial values. For the T/I scheme, initial parameters are assumed to have same values, although designated by different variables (AU and AU*), so they can vary separately. The minimization procedure was repeated 100 times for each f. The same procedure was carried out for HS, the only difference being the use of Eq. (8), applied during the minimization.

2.6.2 MR2 (LS and HS)

For the next round, we calculate the average of those parameters with lowest χ^2 from MR1 to be used as a new fixed initial set of parameters for a second round of minimizations (MR2), following the same procedure described for MR1. Here, that is a way to refine the parameters and reduce the difference between each minimization and consequently reduce the parameter standard deviation. Once more, this was repeated 100 times for HS and for each f (LS).

2.6.3 EU-HS

The last step is to evaluate the impact of the experimental uncertainty (EU) by changing the temperatures of the dataset by small amounts, such that the standard deviation between the original set and the optimized set approaches the declared experimental uncertainty. We then run again the minimization procedure, however, unlike for MR1/2 we keep the initial parameters fixed and only disturb the melting temperatures. A standard deviation of 0.5°C was considered and the minimization carried out for 100 rounds for each f.

2.6.4 EU-LS

For the LS-type datasets, in addition to the impact of the melting temperature uncertainty, we also need to evaluate the impact of the strand concentration grouping procedure described in section 2.4.2. For this, we proceeded in a very similar way as for the temperature perturbation described in the previous section: we disturb the C_t by small amounts and rerun the minimization again. The estimated uncertainty for C_t was reported as 5%, using absorbance reading at 260 nm at 80 °C.[56] Again, this was repeated 100 times for each f and gives us an estimate of the uncertainty over the calculated parameters. Therefore, the final results shown here are the averages over these minimizations. All those steps were carried out independently for each LS salt concentration.

2.7 Validation

For a validation set we collected 25 sequences and their melting temperatures at low and medium salt concentrations and various species concentrations from Refs. 43,59–65, which are shown in supplementary Tab. S13.

3 Results

3.1 Logarithmic groups

The available LS melting temperatures are scattered over a wide range of strand concentrations C_t. Here, we will attempt to group these temperatures according to a logarithmic grouping scheme described in section 2.4.2. The first question we need to address is how this logarithmic grouping impacts the parameter optimization and what is best the coarseness factor f. If f is too small, the melting temperatures are separated into very few large groups, if it is too large they end up scattered into many sparsely populated groups. To answer this question, we performed all minimization independently for f ranging between 0.4 and 5.0, see Tab. 1 and supplementary tables S9–S12.

In Fig. 2 we show the final merit function χ^2 that was minimized during rounds MR1, MR2 and LS-EU for the UN (blue circles) and T/I minimization (red boxes). Both show the same behaviour as function of f. χ^2 levels off after $f = 2$ and there is little difference between 3 and 5. The regression coefficients, $a_{0,1}$, Eq. (7), for $f = 1$ and 5 are shown in Fig. 3. See Figs. S1— S14. At $f = 1$, both a_0 and a_1 show a relatively uniform behaviour for all salt concentrations, with a_1 increasing slightly with L_f. However, for the larger $f = 5$ this uniformity is lost due to the low number of melting temperatures in some L_f groups. This is especially evident for the lowest salt concentration 71 mM, see also the first column in supplementary table S9.

In Fig. 4 we show the Morse potentials for three f factors,

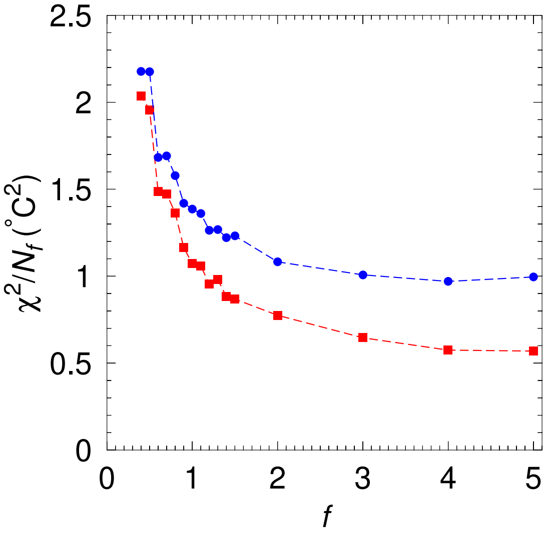

Fig. 2 Final merit function χ^2/N_f as a function of the grouping coarseness factor f for [Na$^+$] 121 mM. Red boxes and blue circles represent T/I and UN optimizations, respectively.

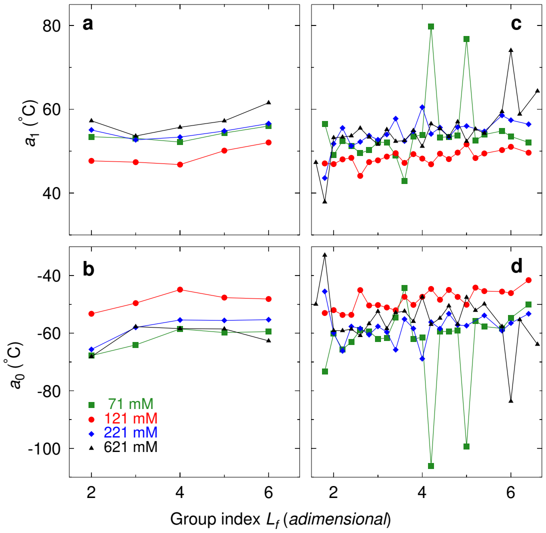

Fig. 3 Regression parameters for (a,b) $f = 1$ and (c,d) $f = 5$ for salt concentrations 71 mM (green boxes), 121 mM (red bullets), 221 mM (blue diamonds) and 621 mM (black triangles).

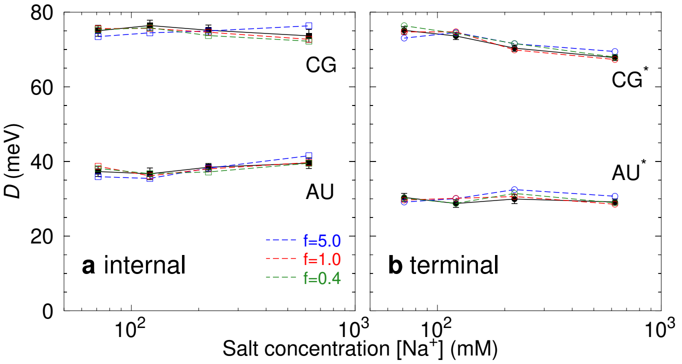

Fig. 4 Morse potentials averaged over all f (black squares) for (a) internal and (b) terminal base pairs. The error bars represent the standard deviation within the f sets. Specific results for $f = 0.4, 1, 5$ are shown in green, red and blue, respectively.

namely 0.4 (3 groups), 1.0 (5 groups) and 5.0 (21 to 24 groups, depending on salt concentration). Comparatively we also show the Morse average over the results for all calculated f factors. The standard deviation within the sets is marginally small with $f = 5$ showing the most pronounced deviation from the average (5 meV). Moreover, for the stacking parameters the higher deviation occurs for the nearest-neighbour UApAU and the rest remain nearly equal within the average. We also compute the average for the stacking potentials which is shown in the Fig. 5. Even displaying more unstable regression parameters, higher values of f still derive parameters consistent within the set and with previous works[51] and on the average producing similar results.

Finally, to answer the question of the most adequate coarseness factor f, it would seem that balancing a low merit factor χ^2 with uniform regression coefficients $a_{0,1}$ points toward an f around 1. It is desirable to deal with monotonic regression coefficients as they allow us to interpolate new coefficient for missing salt concentrations which is not possible for large f. On the other hand, for the optimized parameters shown in Figs. 4 and 5, the actual value of the coarseness factor f appears to be of little importance. Therefore, for the remainder of this article we will discuss the results for $f = 1$, unless noted otherwise.

Since the T/I minimization has a substantially lower merit factor χ^2 than the UN parameters, Fig. 2, there is a possibility of overfitting for the T/I minimization due to the larger number of parameters. To verify if overfitting may have occurred we apply these parameters to the prediction of melting temperatures of an independent validation set of sequences that was not used for the optimization, see supplementary table S13. Using UN parameters, for $f = 1$, we obtain $\langle \Delta T \rangle = 2.04$ °C and a $\chi^2 = 217.76$ °C^2. However, using the parameters derived from T/I minimization, also for $f = 1$, we obtain an important reduction, $\langle \Delta T \rangle = 1.68$ °C and a $\chi^2 = 143.68$ °C^2, which gives us confidence that no overfitting occurred for T/I minimization.

3.2 Parameters at $f = 1$

In Fig. 6 we show the final average Morse potentials for $f = 1$. For comparison, we also show previous salt-dependent results for DNA from Ref. 52. The internal base pair Morse potentials are always larger than for the internal base pairs, which is consistent with our previous calculations for DNA which were calculated at a very low strand concentration (2 μM). The major difference to the DNA results is that we do not observe a reduced Morse potential for terminal r(AU*) at very low salt concentrations. Therefore, it would seem that the hydrogen bonding of r(AU*) is less susceptible to the sodium concentration. However, there is still a considerable difference between internal r(AU) and terminal r(AU*) Morse potentials which makes the terminal base pairs even more vulnerable to end-fraying. The Morse potentials of internal r(AU) base pairs are consistently higher than their d(AT) counterparts, confirming our previous findings for high sodium concentrations[41].

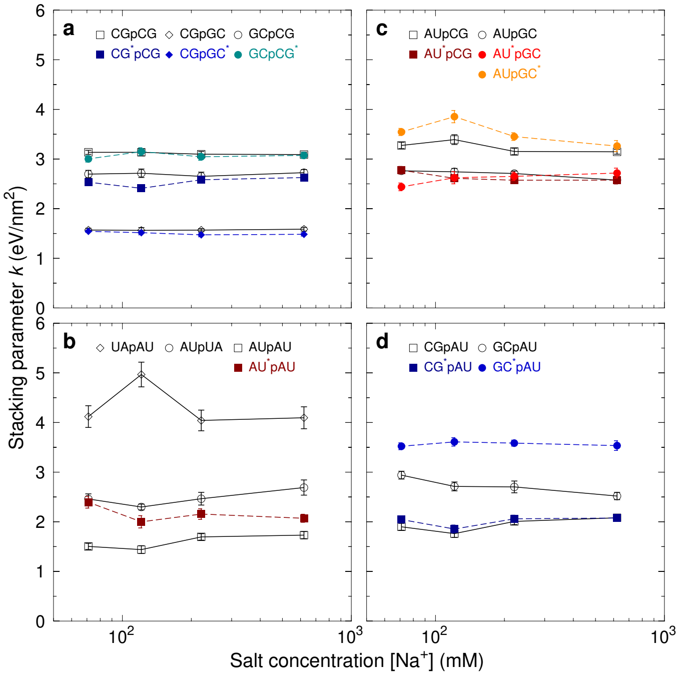

Fig. 5 Stacking potentials averaged over all f. Panels (a,b) show the symmetric NN and panels (c,d) the non-symmetric NN. The error bars represent the standard deviation within the f sets. Dashed lines are for NNs with terminal base pairs.

For r(CG*) Morse potentials we found similar values to r(CG) which at very low and low salt concentrations which is not observed for their DNA analogs. In other words, RNA appears to be less susceptible to end-fraying than DNA at low salt concentrations. We attribute the shift towards higher Morse potentials for HS in Fig. 6 to the substantial difference between the LS and HS datasets, as described in the methods sections.

The calculated stacking parameters are shown in Fig. 7, grouped into symmetric and asymmetric NNs. Note that not all combinations of NNs with terminal base pairs were present in the dataset, therefore not all terminal analogues of internal NNs could be calculated. Except for UApAU NNs, most stacking interactions show little change with salt concentrations. Similarly, our previous results for DNA have shown little dependence of stacking with sodium except for AT*pAT, TApAT* and ATpGC*.[52]

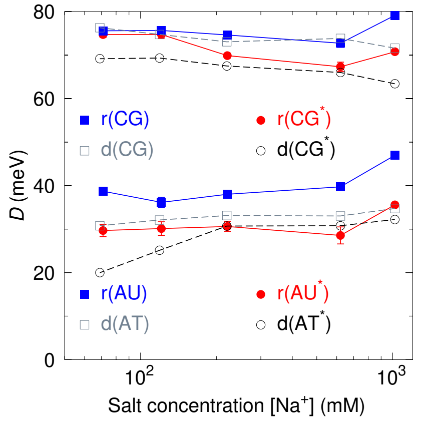

Fig. 6 Average Morse potentials as function of salt concentration. LS results are for $f = 1$. Error bars were estimated in the EU-LS/HS minimization round. For comparison, we show the analogous DNA parameters as grey boxes (internal) and black circles (terminal).[52] The lines connecting the data points are only intended as guide to the eyes.

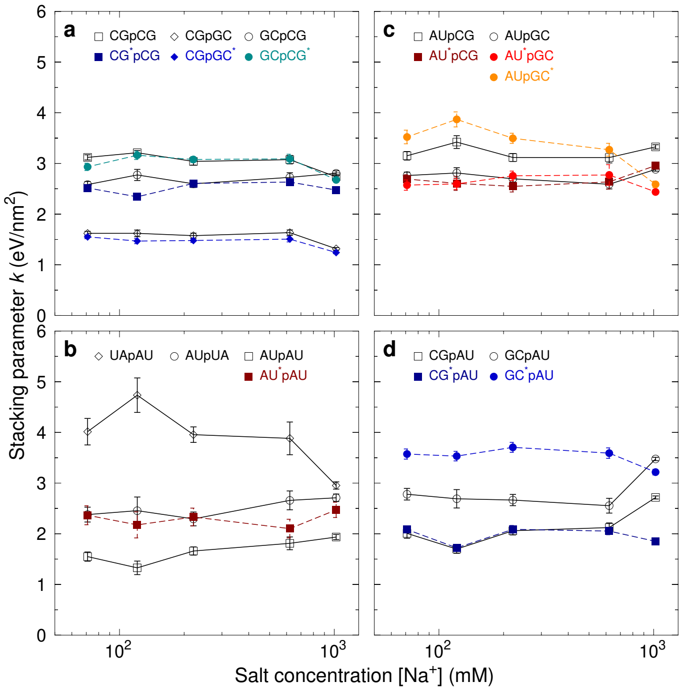

Fig. 7 Average stacking potentials as function of sodium concentration. LS results are for $f = 1$. Panels (a,b) show the symmetric NN and panels (c,d) the asymmetric NN. Error bars were estimated in the EU-LS/HS minimization round. Solid lines are for internal NNs and dashed lines for NNs with terminal base pairs. The lines connecting the data points are only intended as guide to the eyes.

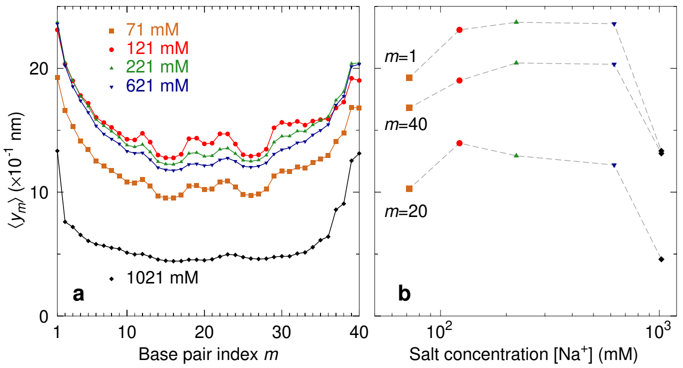

Fig. 8 (a) Average opening profile for the dsRNA sequence from Ref. 67. Squares, diamonds, triangle up, triangle down and bullets represent the average opening for RNA in 71, 121, 221, 621 and 1021 mM [Na$^+$] for the T/I calculation, respectively. Calculation was carried out at 300 K which has no relation to the melting temperature. (b) Comparative opening for the $5'$ terminal ($m = 1$), central ($m = 20$) and $3'$ terminal ($m = 40$) base-pair in function of salt concentration.

4 Discussion

Due to the non-linear Hamiltonian, Eq. (1), it is not straightforward to visualize how the Morse and stacking potentials will affect the opening of the base pairs. For this, we use the average displacement profiles using Eq. (3), that is the $\langle y_m \rangle$ where m is the base pair index. The average displacements indicates which base pairs are likely open first at a given temperature and can be qualitatively related to the to the root-mean-squared distance (RMSD) or root-mean-squared fluctuation (RMSF) used in MD and coarse-grained simulations.[66]

In Fig. 8a we show the average opening for the sequence used in coarse-grained calculations reported in Refs. 37,67. Similarly to the coarse-grained calculations[37] we observe larger fluctuations at the terminal base pairs which increase up to intermediate salt concentrations. However, for higher salt concentrations we observe a saturation and even a substantial drop in $\langle y_m \rangle$ at 1021 mM [Na$^+$]. This saturation between 121 and 621 mM is better seen in Fig. 8b where we show $\langle y_m \rangle$ as a function of sodium concentration for three locations in the sequence. It is unclear if the reduction of $\langle y_m \rangle$ at HS is due to the large difference between the melting temperature datasets, but nevertheless it does not support the continuous increase in RMSF with salt concentration calculated by Jin *et al.*[37]. The terminal $5'$ shows a considerable wider opening than the $3'$ end, see Fig. 8b. This is dissimilar to the calculations by Jin *et al.*[37], yet consistent with results from O'Toole *et al.*[68].

In Fig. 9 we show an example for sequence II from Ref. 44, comparing RNA to the equivalent DNA sequence at 121 mM [Na$^+$]. The calculation temperature in this case was 180 K, which has no relation to the melting temperature correlation of Eqs. (7) and (8). Fig. 9 shows that for internal base pairs, $\langle y_m \rangle$ is somewhat larger at the r(AU) tract than the equivalent d(AT) tract, despite the larger r(AU) Morse potential. The reason for this is that the internal $\langle y_m \rangle$ is pushed up by the terminal r(CG*), which illustrates the cooperativity of the base pairs at the termini affects the internal base pairs as well. In the specific case of sequence II

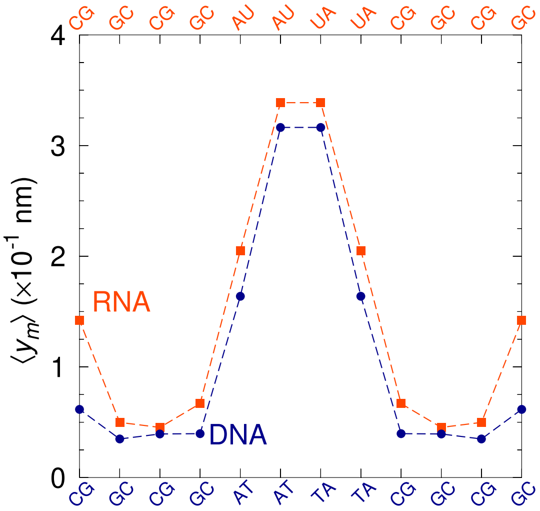

Fig. 9 Average opening profile for sequence IV from Ref. 44. Open and closed squares (bullets), represent the average opening for RNA (DNA) in 121 mM [Na^+], for the UN and T/I calculation, respectively.

from Snoussi and Leroy[44], their NMR measurements indicated a shorter r(AU) lifetime than for d(AT), which would be consistent with a larger displacement for r(AU) seen in Fig. 9. On the other hand, contrary to their results, we observe larger base-fraying for r(CG*) which can be understood from the larger difference between internal and terminal Morse potentials for CG at this salt concentration.

5 Conclusions

We introduce a new technique to parametrise the PB model at multiple strand concentrations by the use of a logarithmic groups. The resulting parameters show little dependence on the coarseness of the grouping which evidences that this technique is robust, and enabled us to make use of a large salt dependent RNA melting temperature dataset. We calculated new salt dependent PB parameters, including specific parameters for the sequence terminals. Unlike d(AT), the Morse potentials for r(AU), which are related to hydrogen bonding, showed no important reduction at low sodium concentrations. Most stacking interactions show little change with salt concentration, however for some terminal contexts stronger stacking interactions were found, similar to our previous study for DNA.[52]

Conflicts of interest

"There are no conflicts to declare".

Acknowledgements

G.W. and I.F. acknowledge the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) for the financial support.

Notes and references

1 F. Michelini, A. P. Jalihal, S. Francia, C. Meers, Z. T. Neeb, F. Rossiello, U. Gioia, J. Aguado, C. Jones-Weinert, B. Luke *et al.*, *Chem. Rev.*, 2018, **118**, 4365–4403.

2 S. Nellimarla and K. L. Mossman, *Journal of Interferon & Cytokine Research*, 2014, **34**, 419–426.

3 T. Kuwabara, J. Hsieh, K. Nakashima, K. Taira and F. H. Gage, *Cell*, 2004, **116**, 779–793.

4 G. J. Hannon, *Nature*, 2002, **418**, 244–251.

5 G. Meister and T. Tuschl, *Nature*, 2004, **431**, 343–349.

6 S. Akira and K. Takeda, *Nature Reviews Immunology*, 2004, **4**, 499–511.

7 J. Lipfert, S. Doniach, R. Das and D. Herschlag, *Annual Review of Biochemistry*, 2014, **83**, 813–841.

8 S. A. Woodson, *Curr. Opin. Chem. Biol.*, 2005, **9**, 104–109.

9 D. E. Draper, *Biopoly.*, 2013, **99**, 1105–1113.

10 E. Koculi, C. Hyeon, D. Thirumalai and S. A. Woodson, *J. Am. Chem. Soc.*, 2007, **129**, 2676–2682.

11 P. A. Lemaire, E. Anderson, J. Lary and J. L. Cole, *J. Mol. Biol.*, 2008, **381**, 351–360.

12 K. Xi, F.-H. Wang, G. Xiong, Z.-L. Zhang and Z.-J. Tan, *Biophys. J.*, 2018, **114**, 1776–1790.

13 Y.-Y. Wu, Z.-L. Zhang, J.-S. Zhang, X.-L. Zhu and Z.-J. Tan, *Nucleic Acids Res.*, 2015, **43**, 6156–6165.

14 S. Kirmizialtin and R. Elber, *J. Phys. Chem. B*, 2010, **114**, 8207–8220.

15 A. V. Drozdetski, I. S. Tolokh, L. Pollack, N. Baker and A. V. Onufriev, *Phys. Rev. Lett.*, 2016, **117**, 028101.

16 Z.-J. Tan and S.-J. Chen, *Met. Ions Life Sci.*, 2011, **9**, 101.

17 E. D. Holmstrom, J. L. Fiore and D. J. Nesbitt, *Biochem.*, 2012, **51**, 3732–3743.

18 J. L. Chen, A. L. Dishler, S. D. Kennedy, I. Yildirim, B. Liu, D. H. Turner and M. J. Serra, *Biochem.*, 2012, **51**, 3508–3522.

19 L. G. Laing, T. C. Gluick and D. E. Draper, *J. Mol. Biol.*, 1994, **237**, 577–587.

20 A. Pyle, *JBIC Journal of Biological Inorganic Chemistry*, 2002, **7**, 679–690.

21 D. E. Draper, *RNA*, 2004, **10**, 335–343.

22 D. Lambert, D. Leipply, R. Shiman and D. E. Draper, *J. Mol. Biol.*, 2009, **390**, 791–804.

23 M. Egli, *Angew. Chem., Int. Ed. Engl.*, 1996, **35**, 1894–1909.

24 V. B. Chu, Y. Bai, J. Lipfert, D. Herschlag and S. Doniach, *Curr. Opin. Struct. Biol.*, 2008, **12**, 619 – 625.

25 G. L. Conn, A. G. Gittis, E. E. Lattman, V. K. Misra and D. E. Draper, *J. Mol. Biol.*, 2002, **318**, 963–973.

26 Y. V. Bukhman and D. E. Draper, *J. Mol. Biol.*, 1997, **273**, 1020–1031.

27 K. Takamoto, Q. He, S. Morris, M. R. Chance and M. Brenowitz, *Nat. Struc. Biol.*, 2002, **9**, 928–933.

28 R. Shiman and D. E. Draper, *J. Mol. Biol.*, 2000, **302**, 79–91.

29 L.-T. Da, F. Pardo-Avila, L. Xu, D.-A. Silva, L. Zhang, X. Gao, D. Wang and X. Huang, *Nature Comm.*, 2016, **7**, 11244.

30 J. F. Sydow, F. Brueckner, A. C. Cheung, G. E. Damsma, S. Dengl, E. Lehmann, D. Vassylyev and P. Cramer, *Mol. Cell*, 2009, **34**, 710–721.

31 A. Serganov and E. Nudler, *Cell*, 2013, **152**, 17–24.

32 W. Huang, J. Kim, S. Jha and F. Aboul-Ela, *PLoS Comput. Biol.*, 2013, **9**, e1003069.

33 I. Bešševá, M. Otyepka, K. Réblová and J. Šponer, *Phys. Chem. Chem. Phys.*, 2009, **11**, 10701–10711.

34 I. Bešševá, P. Banáš, P. Kührová, P. Koşinová, M. Otyepka and J. Šponer, *J. Phys. Chem. B*, 2012, **116**, 9899–9916.

35 J. Virtanen, T. Sosnick and K. Freed, *J. Chem. Phys.*, 2014, **141**, 12B604_1.

36 L. Bao, J. Wang and Y. Xiao, *Phys. Rev. E*, 2019, **99**, 012420.

37 L. Jin, Y.-Z. Shi, C.-J. Feng, Y.-L. Tan and Z.-J. Tan, *Biophys. J.*, 2018, **115**, 1403–1416.

38 T. R. Einert and R. R. Netz, *Biophys. J.*, 2011, **100**, 2745–2753.

39 Z.-J. Tan and S.-J. Chen, *Biophys. J.*, 2007, **92**, 3615–3632.

40 Y.-Z. Shi, F.-H. Wang, Y.-Y. Wu and Z.-J. Tan, *J. Chem. Phys.*, 2014, **141**, 09B606_1.

41 G. Weber, *Nucleic Acids Res.*, 2013, **41**, e30.

42 G. Weber, N. Haslam, J. W. Essex and C. Neylon, *J. Phys.: Condens. Matter*, 2009, **21**, 034106.

43 K. Clanton-Arrowood, J. McGurk and S. J. Schroeder, *Biochemistry*, 2008, **47**, 13418–13427.

44 K. Snoussi and J. L. Leroy, *Biochemistry*, 2001, **40**, 8898–8904.

45 M. Zgarbová, M. Otyepka, J. Sponer, F. Lankas and P. Jurečka, *J. Chem. Theory Comput.*, 2014, **10**, 3177–3189.

46 G. Pinamonti, F. Paul, F. Noé, A. Rodriguez and G. Bussi, *J. Chem. Phys.*, 2019, **150**, 154123.

47 Y.-L. Zhang, W.-M. Zheng, J.-X. Liu and Y. Z. Chen, *Phys. Rev. E*, 1997, **56**, 7100–7115.

48 P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. Doye and A. A. Louis, *J. Chem. Phys.*, 2012, **137**, 135101.

49 P. Šulc, F. Romano, T. E. Ouldridge, J. P. Doye and A. A. Louis, *J. Chem. Phys.*, 2014, **140**, 235102.

50 T. Xia, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox and D. H. Turner, *Biochem.*, 1998, **37**, 14719–14735.

51 I. Ferreira, E. A. Jolley, B. M. Znosko and G. Weber, *Chem. Phys.*, 2019, **521**, 69–76.

52 I. Ferreira, T. D. Amarante and G. Weber, *J. Chem. Phys.*, 2015, **143**, 175101.

53 G. Weber, J. W. Essex and C. Neylon, *Nat. Phys.*, 2009, **5**, 769–773.

54 G. Weber, N. Haslam, N. Whiteford, A. Prügel-Bennett, J. W. Essex and C. Neylon, *Nat. Phys.*, 2006, **2**, 55–59.

55 M. Peyrard and A. R. Bishop, *Phys. Rev. Lett.*, 1989, **62**, 2755–2757.

56 Z. Chen and B. M. Znosko, *Biochem.*, 2013, **52**, 7477–7485.

57 S. Schreiber-Gosche and R. A. Edwards, *J. Chem. Educ.*, 2009, **86**, 644.

58 W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, 1988.

59 G. T. Walker, *Nucleic Acids Res.*, 1988, **16**, 3091–3099.

60 P. J. Mikulecky and A. L. Feig, *Biochem.*, 2006, **45**, 604–616.

61 S. Nakano, M. Fujimoto, H. Hara and N. Sugimoto, *Nucleic Acids Res.*, 1999, **27**, 2957.

62 A. Pasternak and J. Wengel, *Nucleic Acids Res.*, 2010, **38**, 6697–6706.

63 J. I. Gyi, G. L. Conn, A. N. Lane and T. Brown, *Biochem.*, 1996, **35**, 12538–12548.

64 S. Wang and E. T. Kool, *Biochem.*, 1995, **34**, 4125–4132.

65 R. I. Hara, M. Kageyama, K. Arai, N. Uchiyama and T. Wada, *RSC Advances*, 2017, **7**, 41297–41303.

66 L. M. Oliveira, A. S. Long, T. Brown, K. R. Fox and G. Weber, *Chem. Sci.*, 2020, **11**, 8273–8287.

67 L. Bao, X. Zhang, Y.-Z. Shi, Y.-Y. Wu and Z.-J. Tan, *Biophys. J.*, 2017, **112**, 1094–1104.

68 A. S. O'Toole, S. Miller and M. J. Serra, *RNA*, 2005, **11**, 512–516.