Evaluating Hydrogen Bonds and Base Stacking of Single, Tandem and Terminal GU Mismatches in RNA with a Mesoscopic Model

Tauanne D Amarante* and Gerald Weber*

Departamento de Física, Universidade Federal de Minas Gerais, 31270-901 Belo Horizonte-MG, Brazil

E-mail: tauamarante@gmail.com; gweberbh@gmail.com

Phone: +55 31 3409 5633. Fax: +55 31 3409 5600

Abstract

GU mismatches are crucial to the stability of the RNA double helix and need to be considered in RNA folding algorithms for numerous biotechnological applications. Yet despite its importance many aspects of GU base pairs are still poorly understood. There is also a lack of parametrization which prevents it to be considered in mesoscopic models. Here we adapted the mesoscopic Peyrard-Bishop model to deal with context-dependent hydrogen bonds of GU mismatches and calculated the model parameters related to hydrogen bonding and base stacking from available experimental

melting temperatures. The context-dependence causes a proliferation of parameters which made the problem computationally very demanding. We were able to overcome this problem by systematically regrouping the parameters during the minimization procedure. Our results not only provide the much needed parametrization but also answer several questions about the general properties of GU base pairs, as they can be associated straightforwardly to hydrogen bonding and base stacking. In particular we found a very small Morse potential for tandem $5'-$GU$-3'$ which confirms a single hydrogen bond for this configuration, answering a long standing question over conflicting experimental findings. Terminal GU base pairs are known to increase the duplex stability but it is not clear why. Our results suggests that the increased terminal stability is mostly due to stronger hydrogen bonding.

Introduction

The mismatch GU is the most commonly found base pair besides AU and CG in RNA. The unique chemical and structural properties of GU wobble pairs make them special sites for recognition of some biomolecules.[1] The conservation of this motif in specific sites along evolution is another evidence of its functional importance. For example, most living organisms presents one GU mismatch at the third position of tRNA^{Ala} that allows the recognition by the enzyme that attaches the amino acid alanine to its tRNA.[2] The GU mismatch is also linked to the RNA catalysis function, for example, in nearly all organisms, in group I self-splicing introns, there is a GU pair at the site of cleavage.[1]

In the wobble hypothesis, Crick proposed that GU could form two hydrogen bonds similar to the canonical AU pair.[3] This was confirmed by early spectroscopic measurements which obtained GU free energies similar to AU.[4,5] However, later experimental studies showed that the stability of internal GU mismatches depends very much on the neighbor context.[6] In particular, thermal stability of GU symmetric tandem base pairs depends on the direction in which they are arranged, with $5'-$UG$-3'$ being generally more stable than $5'-$GU$-3'$. The

stability is also influenced by flanking Watson Crick base pairs.[7] Early NMR studies[6] did not attribute this change in stability to a difference in hydrogen-bonding pattern. However, later NMR experiments concluded that the symmetric tandem GU base pair may have either one or two hydrogen bonds depending on mismatches sequence and flanking pairs.[8]

In terms of theoretical studies, a comprehensive analysis of nearest-neighbor (NN) parameters for GU was carried out recently by Chen et al.[9]. However, the NN model reveals very little about the intra-molecular interactions due to its fundamental limitation of not being able to separate the hydrogen bonds from the stacking interactions.[10] At the other extreme of the theoretical complexity are molecular dynamics simulations which due to finite computational resources are typically limited to the analysis of just few sequences.[11,12] Even more restricted are density functional theroy (DFT) calculations which study the interactions of isolated GU wobble pairs but do not include the RNA backbone and no stacking interaction.[13–18] Mesoscopic approaches such as the Peyrard-Bishop model[19] overcome the limitations of nearest-neighbor models by treating separately the hydrogen bonds and stacking parameters. The combination of this fundamental property of mesoscopic models with experimental melting temperatures provides a way to calculate the strength of hydrogen bonds[20] which is a difficult property to measure[21] or calculate.[22,23] This has enabled us to obtain fundamental insights into DNA,[20] RNA,[24] and more recently on deoxyinosine.[25] These recent advances are a major motivation to extend this type of analysis to GU mismatches.

The simplified model proposed by Peyrard and Bishop[19] consists in a 2D Hamiltonian that takes into account the stacking interaction and hydrogen bond as separate potentials. The model Hamiltonian is easily modified to include different aspects of nucleotide interactions and can be used either in the framework of equilibrium physical statistics or dynamics. Some examples of recent applications are: the analysis of cynobacterial promoters,[26] protein

induced DNA bubbles[27] and fast prediction of bubble openings.[28] Modified Hamiltonians were proposed to add additional barriers for base pair to model A-DNA[29] or the unzipping induced by force.[30] Further modifications of the model Hamiltonian include the description of structural parameters such as the rise of the helical steps[31] and salt-dependent Morse potentials.[32]

Here we use the Peyrard-Bishop model to obtain the hydrogen bond and stacking parameters for GU base pairs in RNA from published melting temperatures.[9] In comparison to our previous work on RNA CG and AU base pairs,[24] this represents a considerable challenge since we can not assume *a priori* an uniform hydrogen bond strength for GU mismatches. In other words, we are not dealing with a single value for hydrogen bonds as previously for CG or AU.[24] Instead, we need to consider the possibility of multiple values for hydrogen bond strengths depending on context, increasing the parameter searching space dramatically to 114 parameters (40 Morse potentials and 74 stacking parameters). Not only does this represent a huge computational challenge for a non-linear minimization, it also exceeds the number of sequences in the dataset.

To overcome the challenges represented by the large number of parameters we approached the problem in several steps. First we optimized only the 40 Morse potentials independently and kept all stacking parameters constant. Then we gradually reduced the number of Morse potentials into groups of similar strengths until we ended with just 5 different potentials. By reducing the number of Morse potentials we were able to reduce also the number of stacking parameters from 74 to 40, that is, reducing the problem to optimizing just 45 parameters in total.

This process was successful into optimizing the GU Morse potentials into the known groups of one or two hydrogen bonds. We also obtained larger hydrogen bonds for terminal GUs which are known for their increased stability. In contrast to the variable hydrogen bonds, the GU stacking parameters are very similar in order of magnitude to RNA Watson-Crick base pairs. We found an overall agreement with independent experimental measurements

such as NMR and we believe we were able to settle a specific question about the possibility of a single hydrogen bond for tandem GU.

Materials and Methods

The mesoscopic model

We used the model proposed by Peyrard and Bishop[19] with harmonic stacking interaction which is the model with the smallest number of parameters. This simple model has provided good results in a variety of situations.[20,24,25]

The main components of this model are the hydrogen bond represented by a Morse potential,

$$V(y_i) = D_i \left(e^{-y_i/\lambda_i} - 1\right)^2 \tag{1}$$

and the stacking interaction

$$w(y_i, y_{i+1}) = \frac{k_{i,i+1}}{2} \left(y_i^2 - 2y_i y_{i+1} \cos\theta + y_{i+1}^2\right), \tag{2}$$

where the parameters D and λ depend on the base pair i, and the elastic constant k is related to the interaction between subsequent pairs i and $i+1$. The small angle (0.01 rad) θ was introduced to avoid numerical problems in the partition function integral.[33]

Equations (1) and (2) form the model Hamiltonian which is used in the classical partition function.[19] The partition function is calculated numerically using the transfer integral technique.[33] For the integration of Eq. (14) of Ref. 34 we used 400 points over the interval $y_{\min} = -0.1$ nm to $y_{\max} = 20.0$ nm, and a cut-off of P =10 of Eq. (22) of Ref. 34. The calculation of the thermal index τ is carried out at 370 K. Please note that this temperature is unrelated to the temperatures obtained from the regression method. For further details on the model implementation please see Refs. 20,34,35.

Experimental data used

Experimental melting temperature data were taken from the comprehensive review by Chen et al.[9]. They reported an expanded database of 80 sequences that provides all possible combinations of base triplets containing GU pair flanked by canonical pairs, that is AU and CG, in different orientations. This broad set of melting measurements was achieved by adding some oligonucleotides designed to extend a previous database reported by Mathews et al.[36]. These melting temperatures were retrieved from different groups which makes it difficult to estimate a consistent experimental uncertainty for this set, especially as this was not explicitly provided. We recalculated the melting temperatures of self-complementary sequences to 200 μM, following the same approach as used by Xia et al.[37] and to be consistent with our previous calculations for RNA,[24] see Supporting Table S1.

Sequence decomposition and notation

In this work the GU base pairs are uniquely described in terms of context trimers NGN/NUN or NUN/NGN, where N stands for any base. For instance, a GU flanked by UA on the 5′ side and a AU on the 3′ side is identified by the trimer context U<u>G</u>A/A<u>U</u>U. For clarity, the central GU base pairs are underlined. If the GU mismatch is located at one of the terminals of the helix this will be indicated as a terminal-dimer N<u>G</u>*/N<u>U</u>*, which we will treat simply as yet another trimer with a pseudo-base pair **. Please note that the GU base pair may be flanked by another GU for the case of tandem mismatches. The following example illustrates the procedure adopted in this work. Consider a sequence containing a GU at the 5′ terminal and an internal tandem GU,

5′-UCAUGUGG-3′
3′-GGUGUACC-5′

its initial decomposition into trimers would be,

<u>U</u>C/<u>G</u>G, UCA/GGU CAU/GUG, A<u>U</u>G/U<u>G</u>U, U<u>G</u>U/G<u>U</u>A, GUG/UAC UGG/ACC GG*/CC*

where the asterisk * represents the terminal of the helix. All trimers are subsequently symmetry-reduced. For instance out of the two symmetry equivalent trimers

$$\begin{array}{ll} 5'\text{-A}\underline{\text{UG}}\text{-}3' & 5'\text{-}\underline{\text{UG}}\text{U-}3' \\ 3'\text{-U}\underline{\text{GU}}\text{-}5' & 3'\text{-}\underline{\text{GU}}\text{A-}5' \end{array}$$

we retain only the one with lowest lexical order

$$\begin{array}{l} 5'\text{-A}\underline{\text{UG}}\text{-}3' \\ 3'\text{-U}\underline{\text{GU}}\text{-}5' \end{array}$$

that is, we keep A$\underline{\text{UG}}$/U$\underline{\text{GU}}$ because it alphabetically precedes $\underline{\text{UG}}$U/$\underline{\text{GU}}$A. Therefore, the $\underline{\text{UG}}$U/$\underline{\text{GU}}$A trimer of our example sequence will be replaced by its equivalent A$\underline{\text{UG}}$/U$\underline{\text{GU}}$.

Similarly, for terminal-dimers the symmetry lexical reduction results in

$$\begin{array}{l} 5'\text{-*}\underline{\text{UC}}\text{-}3' \\ 3'\text{-*}\underline{\text{G}}\text{G-}5' \end{array} \Rightarrow \begin{array}{l} 5'\text{-G}\underline{\text{G}}\text{*-}3' \\ 3'\text{-C}\underline{\text{U}}\text{*-}5' \end{array}$$

therefore, all terminal-dimers considered here will be shown with */* at the 3′-side, since * has a higher lexical order than A, C, G or U.

After applying all symmetry reductions the final trimer sequence of our example would be

G$\underline{\text{G}}$*/C$\underline{\text{U}}$*, UCA/GGU, CAU/GUG, A$\underline{\text{U}}$G/U$\underline{\text{G}}$U, A$\underline{\text{U}}$G/U$\underline{\text{G}}$U, CAU/GUG, GG*/CC*

Independent Morse potentials will be attributed for each context trimer or groups of context trimers. For example, we may consider an independent Morse parameter $D^{\text{UGA/AUU}}$ uniquely for UGA/AUU contexts. Alternatively, we may use a parameter D^{X} for a group X containing a collection of context trimers which will be independent of all other context groups. On occasion, we will refer to a generic Morse potential D^{α} where α stands for specific context trimers or to D^{β} where β stands for context groups.

Base-pair notation

Context specific base-pairs will be shown with an added superscript α, that is, GU$^{\alpha}$. From the point of view of base-pair parameters such as the Morse potential D^{α}, GU$^{\alpha}$ is equivalent to UG$^{\alpha}$.

Nearest-neighbor notation

Adapting the typical intra-strand notation 5′-GU-3′/3′-UG-5′ to the notation above would lead to something as unwieldy as 5′-G^{A}U^B-3′/3′-U^{A}G^B-5′. Therefore we preferred to keep a base-pair oriented notation GU^{A}pUG^B, that is, a base pair GU^A followed by another UG^B, and drop the redundant 5′ and 3′ notation altogether.

Melting temperature prediction

Given a set of tentative model parameters $P = \{p_1, p_2 \dots p_L\}$ consisting of Morse potentials and stacking parameters, we calculate an adimensional melting index $\tau_i(P)$ for each sequence i from the partition function of the Peyrard-Bishop Hamiltonian.[38] The melting temperature $T_i'(P)$ resulting from the tentative set of parameters P is then obtained from the following linear equation,

$$T_i'(P) = a_0(N) + a_1(N)\tau_i(P), \tag{3}$$

where the regression coefficients are dependent on the sequence length N

$$a_k(N) = b_{0,k} + b_{1,k}N^{1/2}, \quad k = 0, 1 \tag{4}$$

since we have found that the coefficients $a_{0,1}$ are essentially linear with $N^{1/2}$.[38]

Minimization procedure

Optimization method

Here, we briefly outline the optimization method used to obtain the model parameters, which is described in detail in Refs. 20,38. A similar method was also used successfully to calculate the parameters for the Gibbs free energy nearest-neighbor model.[39]

For each tentative set of model parameters P_j we calculate the predicted melting temperatures $T_i'(P_j)$, Eq. (3), and compare them to the experimental temperatures T_i. The model parameters (P_j) are then varied until we minimize the squared differences

$$\chi_j^2 = \sum_{i=1}^{N} \left[T_i'(P_j) - T_i\right]^2 . \tag{5}$$

The minimization is implemented numerically by the Nelder-Mead or downhill simplex method.[20] To refine the optimized parameters, the minimization is repeated two more times using as new starting points the parameters from the previous round.

Occasionally, we also refer in this work to an average melting temperature deviation

$$\langle \Delta T \rangle = \frac{1}{N} \sum_{i=1}^{N} \left| T_i' - T_i \right| . \tag{6}$$

Due to the large number of possible GU mismatch contexts the minimization procedure of Eq. (5) was carried out in five separate minimization rounds.

Initial parameters

For the first four rounds (MR1–MR4) we vary the initial parameters randomly around a given value p in the interval of $[0.5p, 1.5p]$ such that for every minimization step we try to approach the global minima from a different direction.

Minimization round 1 (MR1)

In this step we considered that the hydrogen bonding pattern for a GU mismatch is unique for each trimer context. In other words, an independent Morse parameter D^α was associated to each of the 40 different trimer contexts α present in the dataset. Stacking parameters associated to GU were fixed at 2.5 eV/nm^2. The λ parameters were kept constant at 0.03 nm for all minimization rounds. For the remaining AU and CG base pairs we used the RNA parameters recently calculated for the PB model in Ref. 24. In order to avoid local minima

during the minimization of Eq. (5) the procedure was carried out 300 times, each time with different set initial parameters,[20] as described in the previous section. These initial parameters D were randomly chosen between 15 meV and 45 meV, that is ±50% of Morse potential calculated for AU.[24] The final total squared difference for MR1 was $\chi^2 = 1453\ ^\circ\mathrm{C}^2$ and required of the order of 6000 h on 2 GHz processors.

Minimization round 2 (MR2)

From the results of MR1, the trimer contexts were grouped together into 7 context groups, W, M1, M2, S1, S2, S3, S4, chosen by the similarity of their calculated Morse potentials. This time we used as initial values for the Morse potentials the averaged values for each context group. The minimization was repeated again as for MR1, but only 40 times as this step was only to test the initial arrangement. The final total squared difference for MR2 was $\chi^2 = 1426\ ^\circ\mathrm{C}^2$ and required 280 h of computation time.

Minimization round 3 (MR3)

We evaluated again the resulting Morse potentials from MR2 and identified the possibility of reducing further the number of context groups by joining groups S1 and S2 into SA, and S3 and S4 into SB. We used the averaged Morse potentials as initial value. This minimization was carried out 300 times. This round resulted in $\chi^2 = 1431\ ^\circ\mathrm{C}^2$ and took 1500 h of computation time.

Minimization round 4 (MR4)

After verifying the results from MR3, we allowed the stacking constants to vary as well, adding further 40 parameters to the minimization. The final value for the total squared difference was $\chi^2 = 1023\ ^\circ\mathrm{C}^2$ and required 3600 h of computation time.

Minimization round 5 (MR5)

To obtain an error estimate of the influence of the experimental uncertainty,[20] we carried out one final minimization. This uses the averaged results of MR4, but instead of varying the initial parameters we now varied the experimental melting temperatures. For each round a small random amount δT_i (positive or negative) was added to the reported melting temperature T_i. The random δT_i follows a Gaussian distribution such that the resulting standard deviation of the modified set matches the experimental uncertainty of 1.3 °C. This procedure was repeated 300 times and provides an estimate of the error for each parameter involved in the minimization. The final parameters, obtained by averaging these runs, correspond to $\chi^2 = 920$ °C^2 with 3900 h of calculation time. For further assessment of the influence of the experimental uncertainty, the complete calculation of MR5 was carried out again with 0.5 °C deviation which we refer to as MR5′.

Results

The experimental dataset used here[9] contains 80 sequences with 32 unique context trimers and 8 terminal dimers containing a GU mismatch as shown in Table 1. In Methods we describe the details on how the sequences are divided into context trimers and terminal dimers. For the remaining of the discussion we will consider the terminal dimers as yet another trimer with a */* representing helix terminal as described in Methods.

Unlike our previous parametrization for RNA,[24] we can not attribute a uniform Morse potential D for GU mismatches. One possibility would be to collect information of GU stability trends from the literature to form groups of contexts which would reduce the number of parameters to minimize. Unfortunately, information about GU stability trends, hydrogen bonding and stacking interaction is available only for few sequence contexts and the experimental techniques and conditions differ significantly. Furthermore, we would be introducing an undesirable bias into our calculations. Therefore, we decided to consider an independent

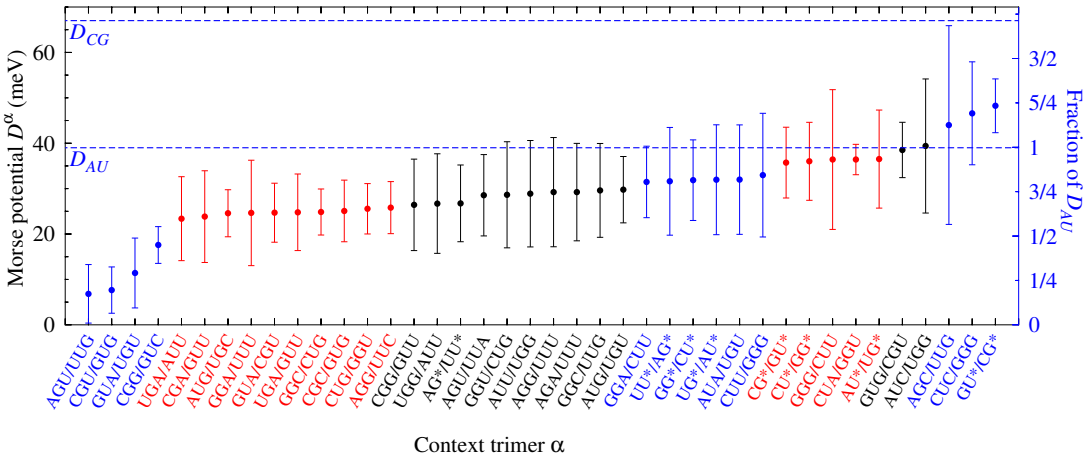

Figure 1: Average Morse potentials D^{α} obtained for each trimer context α from MR1. Trimers are shown in order of increasing D^{α}. AU and CG Morse parameters are from Ref. 24 and are shown as blue dashed lines for reference. Each shaded stripe correspond to the ranges of 1/4 of the hydrogen bond strength D_{AU} and are used as visual aid for the regrouping of the GU context. Left blue scale shows the Morse potentials as fractions of the AU Morse parameter D_{AU}. The colors of the trimers on the horizontal axis refer to context arrangement I used for MR2.

Morse potential for each of the 40 possible trimer contexts. In other words, each GU Morse potential is considered independent from all others. While this is interesting as it prevents grouping biases for the initial minimization, it is challenging to minimize over such a large number of parameters. This is the reason why had to keep all stacking parameters at a single constant value for the initial minimization (MR1) as otherwise we would be adding further 74 parameters to the searching space. For this same reason, it would not be feasible to use more complex potentials which require more parameters, such as including the rise distance proposed in Ref. 31. Also the Morse potential width λ was kept constant for all GU contexts and for all minimization rounds. This was based on our observation[20,24] that λ has no significant influence on the final χ^2 value and consequently does not influence the values of D and k.

We performed the first round of minimizations (MR1) by letting all 40 Morse potentials vary freely. In Figure 1 we show the averaged Morse potentials resulting from the first minimization round MR1. Most Morse potentials resulted in the range of 25 meV to 50 meV,

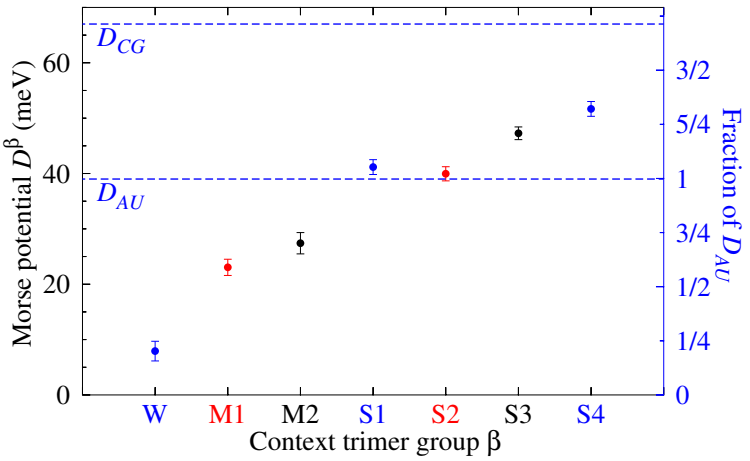

Figure 2: Average Morse potentials D^β obtained for each trimer context group β from MR2. Context groups are from arrangement I shown in Table 1. Remaining figure elements are as in Figure 1.

which supports the early notions of two hydrogen bonds for most GU mismatches.[3] For comparison, the Morse potential for AU was estimated as 38 meV for the same type of mesoscopic model.[24] However, there is a group of GU context with considerable smaller Morse potentials, in the range of 8–20 meV which suggest much weaker hydrogen bonds.

The error bars shown in Figure 1 are not indicative of the statistical uncertainty of the minimization but represent the numerical difficulty in performing a finite number of minimization rounds over a 40-dimensional parameter space of a non-linear model. In other words, if we were given an unlimited amount of time, computer resources or perhaps a more efficient minimization algorithm these error bars should tend to zero, that is, they should all tend to the same global minimum. Nevertheless, they are helpful in providing a guidance for our first attempt in grouping together trimer contexts with similar Morse potential. Another source of numerical difficulty is that by considering 40 different Morse parameters there is only a reduced amount of occurrences of GU mismatch for each trimer context as shown in Table 1.

From the analysis of Figure 1 we selected a tentative arrangement of 7 trimer context groups (see color coding), called arrangement I. This increases considerably the number of

GU occurrences per trimer group as shown in Table 1 and reduces the number of Morse potentials to 7. The minimization MR2 is the first test of this arrangement where the stacking parameters are still kept constant. The optimized Morse parameters for MR2 are shown in Figure 2. The context groups W (W=weak), M1 and M2 (M=medium) are close to the average values of the individual contexts of Figure 1. However, for context groups S1–S4 (S=strong) we observe a somewhat larger Morse potentials than for the individual contexts. The error bars of Figure 2, as for Figure 1, are representative of the numerical non-convergence of the multidimensional minimization. However, they are now considerably smaller since we started out the minimization with a much better initial knowledge of the Morse potentials.

The analysis of Figure 2 suggest that further grouping should be possible. Therefore, we decided to join groups S1 and S2 into group SA, and S2 and S4 into group SB for the similarity of their Morse potentials. This is now arrangement II, see Table 1, which reduces the Morse potentials to just 5. A more extensive minimization MR3 was carried out which confirmed the stability of this new arrangement. After this round, we now included the stacking parameters into the minimization which increases the searching space to 45 parameters and carried out round MR4. The resulting Morse potentials of MR3 and MR4 are shown in Supporting Figure S1. We also note that the merit function χ^2 gradually reduces from 1453 °C^2 for MR1 to 1431 °C^2 for MR3. The stability of χ^2 indicates that at no point the system was under-determined, that is, that the number of data points was sufficient such that no parameter has become a function of any other parameter.

Recalling that the stacking interaction parameter k was kept fixed during MR1–MR4, there is now the question if the arrangements could have had a different outcome if a different fixed value for k had been selected. Fortunately, the form of the Hamiltonian, composed of the sums of Equations (1) and (2), assures that as long as k is fixed for all Morse potentials the arrangement will not change. A different value of k would simply shift uniformly all the Morse potentials resulting in the same arrangement.

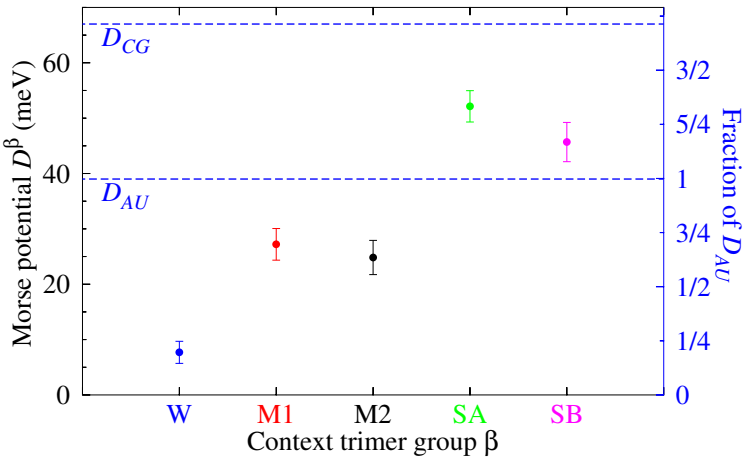

Figure 3: Average Morse potentials D^β obtained for each trimer context group β from MR5. Context groups are from arrangement II shown in Tab. 1. Remaining figure elements are as in Fig. 1.

With the results of MR4, we performed the final minimization MR5 which differs from the previous ones by varying the experimental dataset, see Methods for details. Figure 3 shows the final Morse potential for the remaining 5 context groups and Tables 2 and 3 show the 40 stacking parameters for arrangement II and MR5. MR5 differs from MR4 in that we now randomly change the melting temperatures by small amounts which allows us to evaluate the effect of the estimated 1.3 °C experimental uncertainty on the resulting parameters. Therefore the error bars shown in Figure 3, as well as the displayed uncertainties of Tables 2 and 3 are now of statistical significance. One should note that since the dataset comes from various sources no explicit experimental uncertainty was given.[9] Therefore, we kept the value of 1.3 °C from our previous work on canonical RNA[24] for consistency. However, a smaller uncertainly value basically just reduces the size of the error bars as shown in supplementary Figs. S2 and S3, comparing minimization MR5 and MR5′. The average prediction deviation ΔT for the final parameters of MR5 is 2.7 °C which is moderately smaller than the prediction deviation of 3.0 °C evaluated for NN model (calculated from Table 2 of Ref. 9).

Discussion

Tandem GU

Tandem mismatches are possibly the most extensively studied GU group,[6–9,11,12,40,41,43,45–49] which gave rise to a confusing variety of notations. To aid the following discussion we compiled the correspondence of our flat notation and some common forms found in the literature in Table 4.

Early NMR measurements by He et al.[6] hinted at the possibility of weaker hydrogen bonds for some GUpUG tandem mismatches, perhaps even with just a single bond. Other measurements however[40,41] essentially confirmed the long held view that all GU mismatches, even in tandem configuration, form two hydrogen bonds. A few years later, new NMR data by the same group[8] revived the idea of a single hydrogen bond for a specific sequence r(GGCGUGCC)$_2$ with a GUpUG tandem mismatch. X-ray crystal structure analysis by Jang et al.[43] however was unable to confirm this and attributed their difficulties on different experimental conditions as well as on limitations of NMR technique. On the other hand, molecular dynamics and quantum mechanical calculations by Pan et al.[11] appeared supportive of the one hydrogen bond hypothesis, but also pointed out that the lower stability of GUpUG could be also due to stacking interactions.

Our results largely support the single hydrogen bond hypothesis for this tandem GU motif. Only two tandem GUpUG configurations appear in the dataset for this motif, one is GU^{W}pUG^W where both base pairs are in the weakest W group with a mere 8 meV, in stark contrast to the 39 meV for two-hydrogen bonded AU.[24] The r(GGCGUGCC)$_2$ sequence studied by Chen et al.[8] is exactly of this type as are three of the sequences by He et al.[6]. The other is GU^{M2}pUG^W with one W-type base pair and a medium-strength M2 Morse potential of 25 meV for which we are not aware of any independent experimental NMR or X-ray measurements. Concerning the possibility raised by Pan et al.[11] that stacking interactions could be responsible for a reduced stability of GUpUG, we found no particular difference

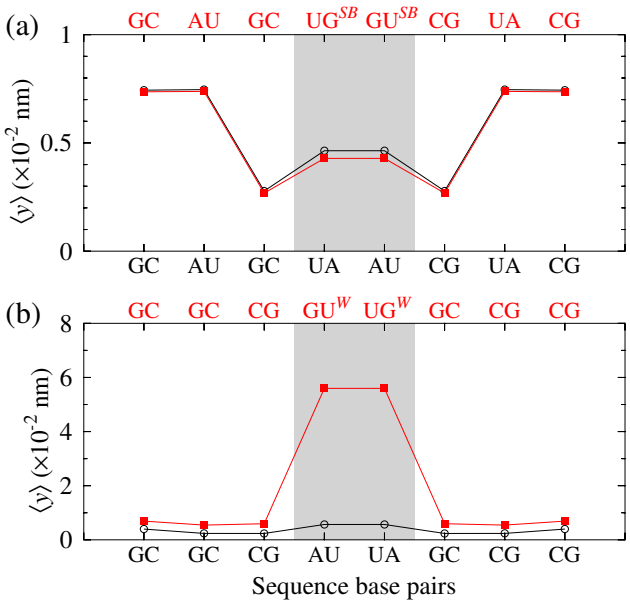

Figure 4: Average opening profile calculated for sequences containing symmetric tandem GUpUG mismatches. Shown are the opening profile (red boxes, upper horizontal axes) for sequence (a) r(GGCGUGCC)$_2$ and (b) r(GAGUGCUC)$_2$ analyzed by Chen et al.[8]. The canonical sequence analogues (black circles, lower horizontal axes) were obtained by replacing G with A. The shaded area highlights the sequence positions where mismatch and canonical sequences differ. Opening profiles were calculated at T= 150 K, note that this temperature is unrelated to the melting temperatures.

in regard to other tandem motifs as shown in Table 3. In fact the stacking parameters do not show any particular difference to canonical CG or AU base pairs either.[24] Therefore, stacking interactions do not appear to be the primary cause for GUpUG instability which leaves only a reduced hydrogen bond as plausible explanation.

The reduced Morse potential alone has a dramatic influence on opening profiles as exemplified in Figure 4, calculated with the new GU parameters using our free-software TfReg.[35] Comparing the strong SB group, Figure 4a, to the weakest W Morse potentials in Figure 4b we notice a 6-fold difference in the average opening profile $\langle y \rangle$. Note that the stacking parameters for both tandem mismatches are virtually the same, see Table 3.

The other two tandem motifs, the symmetric UGpGU and the non-symmetric GUpGU, have Morse potentials in the medium (M2, M1) to strong (SB) groups as shown in Table 3. This is consistent with the experimental NMR data for UGpGU[7,8,41] and X-ray diffraction

for GUpGU,[44,45] which unanimously attribute two hydrogen bond for these tandem motifs. In particular, Kondo et al.[50] reported the role played by water molecules in the stabilization of the wobble pairs UGpGU in tandem that could explain why we observed D values larger than those for AU. In addition, the stability of UGpGU tandems can be further refined in terms of flanking base pairs.[7,42] For instance, the stablest UGpGU is the one flanked symmetrically by GC base pairs

$$\begin{array}{l} 5'\text{-}\mathbf{G}\text{UGC-}3' \\ 3'\text{-CGU}\mathbf{G}\text{-}5' \end{array}$$

In Table 5 we show all trimers associated to a $5'$ flanking base which follows closely the predicted trend $5'\text{G} > 5'\text{C} > 5'\text{U} \geq 5'\text{A}$,[7,42] remembering that the M2 Morse potential is somewhat smaller than for M1. A similar trend is observed for GUpUG tandems where $5'$G has higher Morse potentials than the other contexts. However, we should point out that the $5'$-G flanking base is inferred only from Morse potentials as there is no actual sequence with such configuration in the dataset.

Terminal GU

For terminal GU pairs there is a general consensus that they stabilize the helix.[47–49,51] What is less clear is where this stabilization comes from, if due to hydrogen bonding or due to stacking interactions. Our results place the 8 terminal GU pairs into the context groups with highest Morse potential, SA (52 meV) and SB (46 meV), with the only exception of AG*/UU* with which lies in the intermediate group M2 (25 meV). This suggest that, for the SA and SB groups, this stabilization is due to an increased hydrogen bonding.

It was suggested that the stacking interaction plays a role in the stability of GU terminal mismatches as well, in particular, those with G at the $5'$ position are reportedly more stable due to larger stacking overlap.[47,51] In Table 6 we listed the terminal GU mismatches according the position of the G base being either at $5'$ or $3'$. The table is further organized such that each row shows the same type of flanking base pair (BP), respecting the strand direction.

For our results this is the case for GC, CG and UA flanking pairs as shown in Table 6, largely reflecting the current consensus.[47,51] For the the AU flanking pair the stacking parameter is smaller for G at 5′, but this appears to be compensated by a much larger Morse potential in the SA group.

Single mismatches

In contrast to tandem and terminal GU mismatches, single GU has seemingly not attracted that much attention. Possibly, because early measurements[6] gave no indication of anything but two hydrogen bonds for single GU mismatches, which was confirmed by X-ray measurements for some specific contexts.[52] Indeed this would appear to be mostly the case for our results as well, as shown in Figure 3. However, a closer inspection of group W, Table 1, shows that a non-tandem trimer C<u>G</u>G/G<u>U</u>C also appears in this group with very low average Morse potential of 8 meV. This would suggest a single hydrogen bond for this particular GU context. In contrast, the crystal structure analyzed by Kondo et al.[50] contains a single mismatch in this particular context and predicts two hydrogen bonds. One possible reason for this disagreement could be the crystallization of the RNA sample for performing x-ray diffraction experiments, while the experimental data used here are for RNA in solution.

For stacking parameters, AUpGU^W also stands out with a much smaller than usual value as shown in Table 2. However, reversing the GU pair as in AUpUG^W shows a stronger than average stacking parameter. While these stacking parameters are nowhere as extreme as recently calculated for deoxyinosine mismatches,[25] they still could influence the melting cooperativity in some important ways. Unfortunately, we are not aware of any independent measurements that could be used for comparison in this case as most structural measurements do not provide an estimate of stacking interaction strengths.

Single mismatches would be the only situation where we could possibly try a comparison to DFT calculations[13–18] as they do not consider the RNA backbone and do not include the sequence context. In other words, they are single mismatches without the flanking base

pairs. While they are able to consider several type of base pair geometries such as cis Watson-Crick/sugar edge[15] or sugar edge/sugar edge[14] they cannot specify which type of conformation will be assumed for a given sequence. On the other hand, even though our model can predict the stability of the GU mismatch depending on context, our results for stacking interactions are not detailed enough do infer this base-pair geometry which means that a direct comparison of our results to the DFT calculations is actually not possible. On a broader basis, some DFT calculations suggest a possibility of up to three hydrogen bonds,[17] however our SA/SB groups do not confirm this.

Conclusion

Here we applied successfully a mesoscopic model to GU mismatches. The method involved considering multiple values for Morse potentials depending on the flanking base pair of the GU pair. This provides a way to obtain estimates of hydrogen bond and stacking interaction strengths which are independent from the traditional NMR and crystallographic measurements. This is of importance since there are cases where these measurements provide conflicting results, and a third experimentally-derived method could be helpful to resolve those questions. For instance, we confirmed a single hydrogen bond for GUpUG tandem configurations in agreement with NMR measurements[8] while X-ray measurements were suggesting two hydrogen bonds.[43] In some cases we were able to provide predictions for which, to our knowledge, there are currently no NMR or X-ray measurements available. These encouraging results pave the way to apply the method to other GU configurations such as GU flanked by mismatches[53] or multiple terminal GU.[48] However, at present our analysis does not cover sequence position dependence of the GU parameters which would require a considerable number of additional experimental data. Combining the new RNA parameters for GU with the previously calculated AU and CG[24] allows a more comprehensive application of the Peyrard-Bishop model to this important molecule.

Acknowledgement

This work was supported by Fundação de Amparo a Pesquisa do Estado de Minas Gerais (Fapemig); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Nível Superior (Capes).

Supporting Information Available

Supporting Table S1 and Figures S1, S2 and S3. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Abbreviations

PB, Peyrard-Bishop; NMR, nuclear magnetic resonance; MR, minimization round; NN, nearest-neighbor; BP, base pair; MD, molecular dynamics; W, weak; M, medium; S, strong;

References

(1) Varani, G.; McClain, W. H. The G·U Wobble Base Pair. *EMBO Rep.* **2000**, *1*, 18–23.

(2) Naganuma, M.; Sekine, S.-i.; Chong, Y. E.; Guo, M.; Yang, X.-L.; Gamper, H.; Hou, Y.-M.; Schimmel, P.; Yokoyama, S. The Selective tRNA Aminoacylation Mechanism Based on a Single G·U Pair. *Nature* **2014**, *510*, 507–511.

(3) Crick, F. H. C. Codon–Anticodon Pairing: the Wobble Hypothesis. *J. Mol. Biol.* **1966**, *19*, 548–555.

(4) Freier, S. M.; Kierzek, R.; Caruthers, M. H.; Neilson, T.; Turner, D. H. Free Energy Contributions of G·U and other Terminal Mismatches to Helix Stability. *Biochem.* **1986**, *25*, 3209–3213.

(5) Sugimoto, N.; Kierzek, R.; Freier, S. M.; Turner, D. H. Energetics of Internal GU Mismatches in Ribooligonucleotide Helixes. *Biochem.* **1986**, *25*, 5755–5759.

(6) He, L.; Kierzek, R.; SantaLucia, J., Jr; Walter, A. E.; Turner, D. H. Nearest-Neighbor Parameters for G·U Mismatches: 5′GU3′/3′UG5′is Destabilizing in the Contexts CGUG/GUGC, UGUA/AUGU, and AGUU/UUGA but Stabilizing in GGUC/CUGG. *Biochem.* **1991**, *30*, 11124–11132.

(7) Wu, M.; McDowell, J. A.; Turner, D. H. A Periodic Table of Tandem Mismatches in RNA. *Biochem.* **1995**, *34*, 3204–3211.

(8) Chen, X.; McDowell, J. A.; Kierzek, R.; Krugh, T. R.; Turner, D. H. Nuclear Magnetic Resonance Spectroscopy and Molecular Modeling Reveal that Different Hydrogen Bonding Patterns are Possible for G·U Pairs: One hydrogen Bond for Each G·U Pair in r(GGCGUGCC)$_2$ and Two for Each G·U Pair in r(GAGUGCUC)$_2$. *Biochem.* **2000**, *39*, 8970–8982.

(9) Chen, J. L.; Dishler, A. L.; Kennedy, S. D.; Yildirim, I.; Liu, B.; Turner, D. H.; Serra, M. J. Testing the Nearest Neighbor Model for Canonical RNA Base Pairs: Revision of GU Parameters. *Biochem.* **2012**, *51*, 3508–3522.

(10) Yildirim, I.; Turner, D. H. RNA Challenges for Computational Chemists. *Biochem.* **2005**, *44*, 13225–13234.

(11) Pan, Y.; Priyakumar, U. D.; MacKerell, A. D. Conformational Determinants of Tandem GU Mismatches in RNA: Insights from Molecular Dynamics Simulations and Quantum Mechanical Calculations. *Biochem.* **2005**, *44*, 1433–1443.

(12) Ananth, P.; Goldsmith, G.; Yathindra, N. An Innate Twist Between Crick's Wobble and Watson-Crick Base Pairs. *RNA* **2013**, *19*, 1038–1053.

(13) Brandl, M.; Meyer, M.; Sühnel, J. Water-mediated Base Pairs in RNA: A Quantum-Chemical Study. *J. Phys. Chem. A* **2000**, *104*, 11177–11187.

(14) Šponer, J. E.; Leszczynski, J.; Sychrovskỳ, V.; Šponer, J. Sugar Edge/Sugar Edge Base Pairs in RNA: Stabilities and Structures from Quantum Chemical Calculations. *J. Phys. Chem. B* **2005**, *109*, 18680–18689.

(15) Šponer, J. E.; Špačková, N.; Kulhánek, P.; Leszczynski, J.; Šponer, J. Non-Watson-Crick Base Pairing in RNA. Quantum chemical Analysis of the Cis Watson-Crick/Sugar-Edge Base Pair Family. *J. Phys. Chem. A* **2005**, *109*, 2292–2301.

(16) Šponer, J. E.; Špačková, N.; Leszczynski, J.; Šponer, J. Principles of RNA Base Pairing: Structures and Energies of the Trans Watson-Crick/Sugar-Edge Base Pairs. *J. Phys. Chem. B* **2005**, *109*, 11399–11410.

(17) Kelly, R. E.; Kantorovich, L. N. Planar Heteropairing Possibilities of the DNA and RNA bases: An ab Initio Density Functional Theory study. *J. Phys. Chem. C* **2007**, *111*, 3883–3892.

(18) Bhattacharyya, D.; Koripella, S. C.; Mitra, A.; Rajendran, V. B.; Sinha, B. Theoretical Analysis of Noncanonical Base Pairing Interactions in RNA Molecules. *J. Biosci.* **2007**, *32*, 809–825.

(19) Peyrard, M.; Bishop, A. R. Statistical Mechanics of a Nonlinear Model for DNA denaturation. *Phys. Rev. Lett.* **1989**, *62*, 2755–2757.

(20) Weber, G.; Essex, J. W.; Neylon, C. Probing the Microscopic Flexibility of DNA from Melting Temperatures. *Nat. Phys.* **2009**, *5*, 769–773.

(21) Pervushin, K.; Ono, A.; Fernández, C.; Szyperski, T.; Kainosho, M.; Wüthrich, K. NMR Scalar Couplings Across Watson–Crick Base Pair Hydrogen Bonds in DNA Observed

by Transverse Relaxation-Optimized Spectroscopy. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14147–14151.

(22) Kenny, P. W. Hydrogen Bonding, Electrostatic potential, and Molecular Design. *J. Chem. Inf. Model.* **2009**, *49*, 1234–1244.

(23) Szatyłowicz, H.; Sadlej-Sosnowska, N. Characterizing the Strength of Individual Hydrogen Bonds in DNA Base Pairs. *J. Chem. Inf. Model.* **2010**, *50*, 2151–2161.

(24) Weber, G. Mesoscopic Model Parametrization of Hydrogen Bonds and Stacking Interactions of RNA from Melting Temperatures. *Nucleic Acids Res.* **2013**, *41*, e30.

(25) Maximiano, R. V.; Weber, G. Deoxyinosine Mismatch Parameters Calculated with a Mesoscopic Model Result in Uniform Hydrogen Bonding and Strongly Variable Stacking Interactions. *Chem. Phys. Lett.* **2015**, *631–632*, 87–91.

(26) Tapia-Rojo, R.; Mazo, J. J.; Hernández, J. Á.; Peleato, M. L.; Fillat, M. F.; Falo, F. Mesoscopic Model and Free Energy Landscape for Protein-DNA Binding Sites: Analysis of Cyanobacterial Promoters. *PLoS Comput. Biol.* **2014**, *10*, e1003835.

(27) Traverso, J. J.; Manoranjan, V. S.; Bishop, A.; Rasmussen, K. Ø.; Voulgarakis, N. K. Allostery through Protein-Induced DNA Bubbles. *Sci. Rep.* **2015**, *5*, 9037.

(28) Zrimec, J.; Lapanje, A. Fast Prediction of DNA Melting Bubbles using DNA Thermodynamic Stability. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2015**, avaliable at DOI:10.1109/TCBB.2015.2396057.

(29) Valle-Orero, J.; Wildes, A.; Theodorakopoulos, N.; Cuesta-López, S.; Garden, J.; Danilkin, S.; Peyrard, M. Thermal Denaturation of A-DNA. *New J. Phys.* **2014**, *16*, 113017.

(30) Bergues-Pupo, A. E.; Falo, F.; Fiasconaro, A. Resonant Optimization in the Mechanical Unzipping of DNA. *EPL* **2014**, *105*, 68005.

(31) Amarante, T. D.; Weber, G. Analysing DNA Structural Parameters using a Mesoscopic Model. *J. Phys.: Conf. Ser.* **2014**, *490*, 012203.

(32) Singh, A.; Singh, N. Effect of Salt Concentration on the Stability of Heterogeneous DNA. *Phys. A (Amsterdam, Neth.)* **2015**, *419*, 328–334.

(33) Zhang, Y.-L.; Zheng, W.-M.; Liu, J.-X.; Chen, Y. Z. Theory of DNA melting based on the Peyrard-Bishop model. *Phys. Rev. E* **1997**, *56*, 7100–7115.

(34) Weber, G.; Haslam, N.; Essex, J. W.; Neylon, C. Thermal Equivalence of DNA Duplexes for Probe Design. *J. Phys.: Condens. Matter* **2009**, *21*, 034106.

(35) Weber, G. TfReg: Calculating DNA and RNA Melting Temperatures and Opening Profiles with Mesoscopic Models. *Bioinformatics* **2013**, *29*, 1345–1347.

(36) Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol.* **1999**, *288*, 911–940.

(37) Xia, T.; SantaLucia, J., Jr.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs. *Biochem.* **1998**, *37*, 14719–14735.

(38) Weber, G.; Haslam, N.; Whiteford, N.; Prügel-Bennett, A.; Essex, J. W.; Neylon, C. Thermal Equivalence of DNA Duplexes Without Melting Temperature Calculation. *Nat. Phys.* **2006**, *2*, 55–59.

(39) Weber, G. Optimization Method for Obtaining Nearest-Neighbour DNA Entropies and Enthalpies Directly from Melting Temperatures. *Bioinformatics* **2015**, *31*, 871–877.

(40) McDowell, J. A.; Turner, D. H. Investigation of the Structural Basis for Thermodynamic

Stabilities of Tandem GU Mismatches: Solution Structure of (rGAGGUCUC)$_2$ by Two-Dimensional NMR and Simulated Annealing. *Biochem.* **1996**, *35*, 14077–14089.

(41) McDowell, J. A.; He, L.; Chen, X.; Turner, D. H. Investigation of the Structural Basis for Thermodynamic Stabilities of Tandem GU Wobble Pairs: NMR Structures of (rGGAGUUCC)$_2$ and (rGGAUGUCC)$_2$. *Biochem.* **1997**, *36*, 8030–8038.

(42) Biswas, R.; Wahl, M. C.; Ban, C.; Sundaralingam, M. Crystal Structure of an Alternating Octamer r(GUAUGUA)dC with Adjacent G·U Wobble Pairs. *J. Mol. Biol.* **1997**, *267*, 1149–1156.

(43) Jang, S. B.; Hung, L.-W.; Jeong, M. S.; Holbrook, E. L.; Chen, X.; Turner, D. H.; Holbrook, S. R. The Crystal Structure at 1.5 Å Resolution of an RNA Octamer Duplex Containing Tandem G·U Basepairs. *Biophys. J.* **2006**, *90*, 4530–4537.

(44) Trikha, J.; Filman, D. J.; Hogle, J. M. Crystal Structure of a 14 bp RNA Duplex with Non-Symmetrical Tandem G·U Wobble Base Pairs. *Nucleic Acids Res.* **1999**, *27*, 1728–1739.

(45) Deng, J.; Sundaralingam, M. Synthesis and Crystal Structure of an Octamer RNA r(guguuuac)/r(guaggcac) with G·G/U·U Tandem Wobble Base Pairs: Comparison with other Tandem G·U Pairs. *Nucleic Acids Res.* **2000**, *28*, 4376–4381.

(46) Xu, D.; Landon, T.; Greenbaum, N. L.; Fenley, M. O. The Electrostatic Characteristics of G·U Wobble Base Pairs. *Nucleic Acids Res.* **2007**, *35*, 3836–3847.

(47) Masquida, B.; Westhof, E. On the Wobble GoU and Related Pairs. *RNA* **2000**, *6*, 9–15.

(48) Nguyen, M.-T.; Schroeder, S. J. Consecutive Terminal GU Pairs Stabilize RNA Helices. *Biochem.* **2010**, *49*, 10574–10581.

(49) Utsunomiya, R.; Suto, K.; Balasundaresan, D.; Fukamizu, A.; Kumar, P. K.; Mizuno, H.

Structure of an RNA duplex r(GGCG$_{Br}$UGCGCU)$_2$ with Terminal and Internal Tandem G·U Base Pairs. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62*, 331–338.

(50) Kondo, J.; Dock-Bregeon, A.-C.; Willkomm, D. K.; Hartmann, R. K.; Westhof, E. Structure of an A-form RNA Duplex Obtained by Degradation of 6S RNA in a Crystallization Droplet. *Acta Crystallogr., Sect. F: Struct. Biol. Cryst. Commun.* **2013**, *69*, 634–639.

(51) Mizuno, H.; Sundaralingam, M. Stacking of Crick Wobble Pair and Watson-Crick Pair: Stability Rules of GU Pairs at Ends of Helical Stems in tRNAs and the Relation to Codon-Anticodon Wobble Interaction. *Nucleic Acids Res.* **1978**, *5*, 4451–4462.

(52) Mueller, U.; Schuebel, H.; Sprinzl, M.; Heinemann, U. Crystal Structure of Acceptor Stem of tRNA^{Ala} From *Escherichia coli* Shows Unique G·U Wobble Base Pair at 1.16 Å Resolution. *RNA* **1999**, *5*, 670–677.

(53) Davis, A. R.; Znosko, B. M. Thermodynamic Characterization of Naturally Occurring RNA Single Mismatches with GU Nearest Neighbors. *Biochem.* **2008**, *47*, 10178–10187.

Graphical TOC Entry

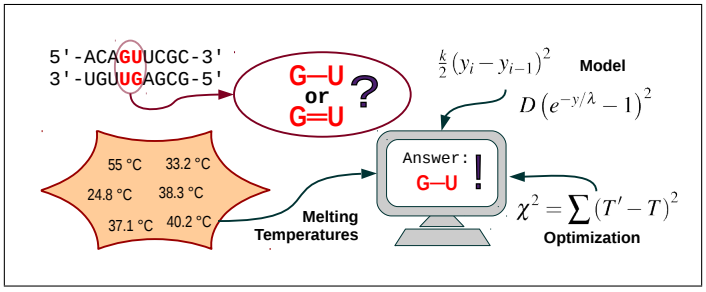

Table 1: Number of occurrences n of GU mismatches per context trimers or context group. The trimers contexts listed in ascending Morse potential order from MR1, which is also the same order shown in Fig. 1. Also shown are the tentative context arrangements and respective number of context trimers contained within each group.

Trimer context	n	Arrangement I	Arrangement II
AGU/UUG	6	23 (W)	23 (W)
CGU/GUG	8		
GUA/UGU	5		
CGG/GUC	4		
UGA/AUU	4	39 (M1)	39 (M1)
CGA/GUU	2		
AUG/UGC	5		
GGA/UUU	5		
GUA/CGU	4		
UGA/GUU	4		
GGC/CUG	2		
CGC/GUG	4		
CUG/GGU	5		
AGG/UUC	4		
CGG/GUU	5	35 (M2)	35 (M2)
UGG/AUU	2		
AG*/UU*	4		
AGU/UUA	4		
GGU/CUG	1		
AUU/UGG	1		
AGG/UUU	2		
AGA/UUU	2		
GGC/UUG	6		
AUG/UGU	8		
GGA/CUU	8	23 (S1)	45 (SA)
UU*/AG*	4		
GG*/CU*	4		
UG*/AU*	4		
AUA/UGU	2		
CUU/GGG	1		
CG*/GU*	4	22 (S2)	
CU*/GG*	4		
GGG/CUU	4		
CUA/GGU	4		
AU*/UG*	6		
GUG/CGU	5	8 (S3)	18 (SB)
AUC/UGG	3		
AGC/UUG	3	10 (S4)	
CUC/GGG	3		
GU*/CG*	4		

Table 2: Final stacking parameters k in eV/nm^2, for non-tandem GU mismatches, from MR5. Calculated uncertainties are shown in compact notation.

NN	k	NN	k	NN	k
AUpGU^{M2}	2.9(4)	AUpGU^{SB}	3.6(9)	AUpGU^{M1}	3.0(5)
AUpGU^{W}	0.9(2)	AUpUG^{M2}	4.2(5)	AUpUG^{SB}	2.8(6)
AUpUG^{SA}	2.5(3)	AUpUG^{M1}	3.7(7)	CGpGU^{M2}	2.4(4)
CGpGU^{SA}	1.9(2)	CGpGU^{M1}	2.1(3)	CGpGU^{W}	2.8(3)
CGpUG^{SB}	2.4(4)	CGpUG^{SA}	1.8(2)	CGpUG^{M1}	2.6(3)
CGpUG^{W}	4.7(5)	GCpGU^{M2}	2.7(6)	GCpGU^{SB}	3.2(6)
GCpGU^{SA}	1.7(2)	GCpGU^{M1}	2.4(3)	GCpUG^{M2}	3.0(5)
GCpUG^{SB}	2.9(4)	GCpUG^{M1}	3.1(4)	GU^{M2}pAU	2.1(5)
GU^{SA}pAU	2.2(3)	GU^{M1}pAU	1.5(2)	UApGU^{M2}	2.0(6)
UApGU^{SA}	2.1(3)	UApGU^{M1}	2.4(3)	UApGU^{W}	1.6(4)

Table 3: Final tandem stacking parameters k in eV/nm^2, from MR5. Calculated uncertainties are shown in compact notation. Also shown are references which independently determined the number of hydrogen bonds for each stacking NN configuration.

motif	NN	k	Independent measurements of hydrogen bonds
UGpGU	UG^{SB}pGU^{SB}	1.8(5)	2 hydrogen bonds NMR,[8,40] MD[11]
	UG^{SB}pGU^{M1}	2.6(5)	
	UG^{M2}pGU^{M2}	3.1(7)	2 hydrogen bonds NMR,[6,41] X-ray,[42] MD[11]
	UG^{M1}pGU^{M1}	3.3(9)	2 hydrogen bonds NMR,[40] MD[11]
GUpUG	GU^{W}pUG^{W}	1.9(5)	1 hydrogen bond NMR,[6,8] MD[11] 2 hydrogen bonds NMR,[41] Xray[43]
	GU^{M2}pUG^{W}	2.6(5)	
GUpGU	GU^{M2}pGU^{M2}	1.9(5)	2 hydrogen bonds X-ray[44,45]
	GU^{M2}pGU^{M1}	1.9(4)	
	GU^{SA}pGU^{M2}	3.2(8)	
	GU^{M2}pGU^{SA}	3.5(1)	

Table 4: Correspondence between the flat nearest-neighbor (NN) notation used in this work and elsewhere in the literature.

NN	structure	groups	equivalent notation and reference
UGpGU	5′-UG-3′ 3′-GU-5′	M1, M2, SB	Motif I[45–47]
			5′−UG−3′[11] 5′<u>UG</u>3′[8]
			5′UG/3′GU[9,48]
			5′UG/GU3′[49]
			5′U·G/G·U3′[12]
			U·G/G·U[45]
			5′-UG-3′/3′-GU-5′[45,46]
GUpUG	5’-GU-3’ 3’-UG-5’	W, M2	Motif II[45–47]
			5′−GU−3′[11] 5′<u>GU</u>3′[8]
			5′GU/3′UG[9,43,48]
			5′G·U/U·G3′[12]
			G·U/U·G[45]
			5′-GU-3′/3′-UG-5′[45,46]
GUpGU	5’-GG-3’ 3’-UU-5’	M1, M2, SA	Motif III[45–47]
			5′GG/3′UU[9,48]
			G·G/U·U[45]
	5′-UU-3′ 3′-GG-5′		5′-UU-3′/3′-GG-5′[45,46]
			5′UU/3′GG[9,48]

Table 5: Identification of the trimers associated to GU mismatches in symmetric tandem according to the direction and the flanking base pairs.

	flanking base-pairs/trimer context (group)			
NN	5′G	5′C	5′U	5′A
UGpGU	5′-**G**UG CGU (SB)	5′-**C**UG GGU (M1)	UGA GU**U**-5′ (M1)	5′-**A**UG UGU (M2)
GUpUG	*5′-**G**GU CUG (M2)	5′-**C**GU GUG (W)	GUA UG**U**-5′ (W)	5′-**A**GU UUG (W)

*No sequence contains this trimer associated to a symmetric tandem GU in the dataset of Ref. 9.

Table 6: Association of terminal GU trimers and stacking groups. Each row is for a flanking base pair (BP). Stacking parameters k (eV/nm^2) are repeated from Table 2.

BP	3′-end	NN	k	5′-end	NN	k
AU	AG*/UU*	AUpGU^{M2}	2.9	UU*/AG*	GU^{SA}pAU	2.2
GC	GG*/CU*	GCpGU^{SA}	1.7	CU*/GG*	CGpUG^{SA}	1.8
CG	CG*/GU*	CGpGU^{SA}	1.9	GU*/CG*	GCpUG^{SB}	2.9
UA	UG*/AU*	UApGU^{SA}	2.1	AU*/UG*	AUpUG^{SA}	2.5