Estimating uncertainty in predicted folding free energy changes of RNA secondary structures

JEFFREY ZUBER[1] and DAVID H. MATHEWS[1,2]

[1]Department of Biochemistry and Biophysics and Center for RNA Biology, University of Rochester Medical Center, Rochester, New York, 14642, USA

[2]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York, 14642, USA

ABSTRACT

Nearest neighbor parameters for estimating the folding stability of RNA are commonly used in secondary structure prediction, for generating folding ensembles of structures, and for analyzing RNA function. Previously, we demonstrated that we could quantify the uncertainties in each nearest neighbor parameter by perturbing the underlying optical melting data within experimental error and rederiving the parameters, which accounts for the substantial correlations that exist between the parameters. In this contribution, we describe a method to estimate uncertainty in the estimated folding stabilities of RNA structures, accounting for correlations in the nearest neighbor parameters. This method is incorporated in the RNA structure software package.

Keywords: Gibbs free energy change; nearest neighbor rules

INTRODUCTION

Noncoding RNAs (ncRNA), functional RNAs that are not transcribed into protein, have functions that range from enzymatic catalysis (ribozymes) (Doudna and Cech 2002) to regulation of gene expression (siRNA, miRNA, and riboswitches) (Wu and Belasco 2008; Serganov and Nudler 2013) and to target identification (guide RNAs) (Yu and Meier 2014). Many of these RNA sequences function by either having a structure or by forming a structure, e.g., a helix, with a second RNA sequence.

RNA structure is hierarchical (Tinoco and Bustamante 1999). The primary structure is the linear sequence of covalently linked nucleotides. The secondary structure is the set of canonical base pairs in the RNA, which are A-form helices. These helices flank regions that are termed loops, including hairpin loops (with one exiting helix), bulge loops (with two exiting helices and all unpaired nucleotides on one strand of the loop), internal loops (with two exiting helices and unpaired nucleotides on both strands of the loop), and multibranch loops (with three or more exiting helices). The tertiary structure includes additional contacts and is defined by the positions of all atoms in the RNA. The secondary structure generally forms faster (Woodson 2000) and is generally more thermostable (Crothers et al. 1974; Onoa and Tinoco 2004) than tertiary structure, allowing RNA secondary structure to be predicted independently of tertiary structure (Tinoco and Bustamante 1999).

A set of nearest neighbor parameters can be used to estimate the free energy change of folding to a secondary structure from random coil (Mathews et al. 2004). In the nearest neighbor rules, the stability of a given motif, such as a stack of two base pairs or a set of unpaired nucleotides called a loop, is assumed to be determined by the sequence of the motif and the adjacent base pairs. These parameters approximate the folding free energy change of a secondary structure as the sum of the energies of neighboring structural motifs, and they were derived using linear regression on a database of folding stabilities determined by optical melting data of small model RNA structures (Andronescu et al. 2014). The most recent complete set of Turner rules, assembled in 2004 (Mathews et al. 2004; Turner and Mathews 2010), includes the Watson–Crick terms determined in 1998 (Xia et al. 1998) and are enumerated in the nearest neighbor database (NNDB), along with examples of their use (Turner and Mathews 2010).

The nearest neighbor parameters are used widely in software for RNA secondary structure prediction (Andronescu et al. 2003; Zuker 2003; Ding et al. 2004; Reuter and

Corresponding author: David_Mathews@urmc.rochester.edu

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.069203.118.

Mathews 2010; Lorenz et al. 2011). It is popular to use dynamic programming algorithms to predict lowest free energy structures (Seetin and Mathews 2012; Hofacker 2014), base-pairing probabilities across an ensemble of structures for a given sequence (McCaskill 1990; Mathews 2004), or a set of structures to represent the Boltzmann ensemble (Ding and Lawrence 2003). Additionally, methods that infer folding parameters from the set of sequences with known structure generally use the same functional forms (Do et al. 2006; Andronescu et al. 2010; Rivas et al. 2012).

Despite their wide application, methods to determine the uncertainties in the folding free energy predictions calculated using the nearest neighbor parameters have been unavailable. There are a number of barriers that have impeded the development of these methods. First, uncertainties for many of the nearest neighbor parameters were unreported. Second, due to the complex relationships between the parameters during the derivation of the nearest neighbor parameters, an accurate estimate of the parameter covariation was difficult (Zuber et al. 2018). Finally, many of the reported uncertainties determined from the standard error of the linear regressions were inaccurate due to the practice of using correlated values in the regressions (Zuber et al. 2018). We recently reported the uncertainty in parameters and the correlation between parameters, addressing these barriers.

Uncertainties in folding free energy estimates of secondary structure formation have a number of important applications. Uncertainties can be used to determine whether the probabilities of two distinct secondary structures are significantly different from each other. The uncertainty in the folding free energy of the lowest free energy structure could also be used to determine the optimal difference from the lowest free energy structure from which to sample suboptimal structures (Wuchty et al. 1999; Mathews 2006).

Here we report a new method, implemented in the RNAstructure software package, for estimating the experimental uncertainty in a folding free energy change for an RNA secondary structure, accounting for the uncertainty in each parameter and the correlation between parameters. Four things are required for this estimate: a count of the number of times each data table entry (corresponding to one or more of the nearest neighbor parameters) was used in the software, a method to map data table usage to the underlying nearest neighbor parameters, an estimate of the uncertainty of each nearest neighbor parameter, and an accurate estimate of the parameter covariation.

RESULTS

Calculating uncertainty in predicted folding free energy changes

The predicted folding free energy change determined using the nearest neighbor thermodynamic model is a linear sum of parameter values:

$$\Delta G = \sum_{i}^{NNs} P_i \times NN_i, \quad (1)$$

where P_i is the parameter coefficient (the parameter usage count) and NN_i is the value of the i^{th} nearest neighbor parameter. The variance in the free energy change ($\sigma^2_{\Delta G}$) can be determined by

$$\sigma^2_{\Delta G} = \sum_{i}^{NNs}\sum_{j}^{NNs} \sigma_i \times \sigma_j \times P_i \times P_j \times \rho_{ij}, \quad (2)$$

where, ρ_{ij} is the linear correlation coefficient between the i^{th} and j^{th} nearest neighbor parameter values (where $\rho_{ii} = 1$), P_i and P_j are the parameter usage counts for the i^{th} and j^{th} nearest neighbor parameters, and σ_i and σ_j are the estimated uncertainties for the i^{th} and j^{th} parameters. $\sigma^2_{\Delta G}$ is the variance in the free energy change, and therefore $\sigma_{\Delta G}$ is the estimated experimental uncertainty for the folding free energy change. It is convenient to instead use the variance-covariance form of the equation:

$$\sigma^2_{\Delta G} = \sum_{i}^{NNs}\sum_{j}^{NNs} P_i \times P_j \times cov_{ij}. \quad (3)$$

In this form, cov_{ij} is the covariance for parameters i and j for i≠j and the variance for parameter i otherwise ($cov_{ij} = \sigma_i \times \sigma_j \times \rho_{ij}$). Equation 3 can be represented using matrix notation:

$$\sigma^2_{\Delta G} = \mathbf{P} \times \mathbf{\Sigma_P} \times \mathbf{P}^\top, \quad (4)$$

where $\boldsymbol{P}$ is the parameter usage count vector, $\boldsymbol{\Sigma_P}$ is the variance-covariance matrix (not to be confused with the summation operator), and $\boldsymbol{P}^\top$ is the transpose of the parameter usage count vector. The variance-covariance matrix for the free energy parameters is available from our prior work (Zuber et al. 2018). The variance-covariance matrix was determined with complete sets of thermodynamic parameters derived from optical melting data that had been randomly perturbed within the experimental limits. This method accounts for all the explicit and implicit relationships between all the parameters in the nearest neighbor thermodynamic model that are derived from experimental data.

To validate our method, we estimated the experimental uncertainty in the folding free energy change for 1450 RNA structures (listed in Materials and Methods) ranging in length from 54 to 2927 nt using two methods. In the first, the experimental uncertainty was measured as the standard deviation of the distribution of free energy changes determined using 100 nearest neighbor parameter sets that are equivalent, but fit to experimental data perturbed within the experimental error. These parameter sets were available from our previous work (Zuber et al. 2018). In the second, the square root of Equation 4 was used to calculate the uncertainty using variance and covariance

values we determined previously (Zuber et al. 2018). A high linear correlation ($r^2 = 0.9952$) was observed between the two estimates (Fig. 1). This demonstrates our method that uses uncertanties in individual parameters and the covariances of these parameters matches the explicit determination of error.

Calculating uncertainty in predicted folding free energy changes at temperatures other than 37°C

In order to calculate the uncertainty in the predicted folding free energy change at multiple temperatures, we needed to determine the variance-covariance matrix for the enthalpy change nearest neighbor parameters. This requires that the enthalpy and free energy parameters be determined simultaneously from the experimental optical melting data for sets of data perturbed within the experimental error. We randomly perturbed the enthalpy and entropy change values for the optical melting experiments within experimental uncertainty, accounting for the known correlation between enthalpy change and entropy change (Xia et al. 1998). The experimental values were then used to derive the folding free energy change and enthalpy change values for each nearest neighbor parameter (Lu et al. 2006). 100,000 perturbed parameter sets were generated, from which the variance-covariance matrices were calculated for the enthalpy change parameters.

Parameter covariation contributes to reduced uncertainty magnitudes

We calculated the uncertainty in the predicted folding free energy for known RNA structures from a set of 1450 sequences (listed in Materials and Methods). This calculation was done while including the effects of parameter covariation and while ignoring covariation (Fig. 2). The covariation

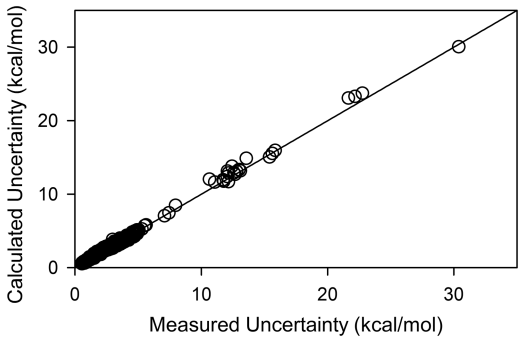

FIGURE 1. Calculated uncertainties agree with measured uncertainties. Experimental uncertainty values calculated using Equation 4 are compared to uncertainties measured from the distribution of free energy predictions using 100 randomly perturbed parameter sets. Each point represents the uncertainty in the predicted free energy for a single secondary structure. The diagonal line represents the ideal case where the two methods agree.

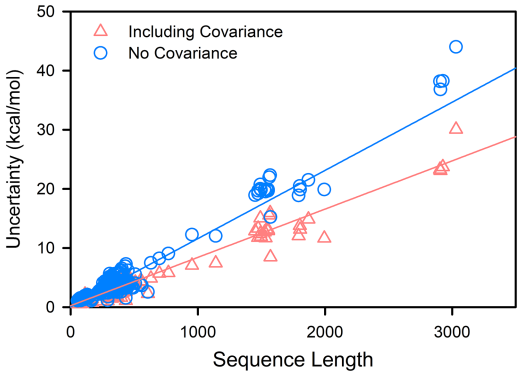

FIGURE 2. Correlation between uncertainty and sequence length. The uncertainty in the folding free energy for the accepted structures of the sequences in the structure archive is plotted against sequence length, both considering correlation and neglecting correlation. Best fit lines are shown to represent the trend.

of the parameters is an important consideration when estimating the uncertainty in the folding free energy changes (Zuber et al. 2018). Neglecting the parameter covariation has the effect of increasing the estimated uncertainty by ~40% compared to the uncertainties estimated when parameter covariations are included (Fig. 2).

Uncertainty in folding free energy is correlated with sequence length

We found that the estimated uncertainty is highly correlated with sequence length ($r^2 = 0.939$; Fig. 2). Plotting the distribution of the fractional uncertainty relative to the folding free energy value reveals a compact distribution with a long tail, where the tail is mainly due to accepted secondary structures with low magnitude predicted folding free energies (Fig. 3).

DISCUSSION

To our knowledge, this is the first technique to estimate the uncertainty of RNA folding free energy calculations, providing information that is required to assess the significance of many predictions. This capability allows users of RNAstructure to differentiate between true suboptimal secondary structures and structures within experimental uncertainty of the lowest free energy structure. The incorporation of uncertainty estimates opens up the field to more rigorous statistical analyses of the predictions made.

Additionally, this analysis provides guidance for future implementations of RNA structure prediction. For example, in RNAstructure, the program *Fold* can generate suboptimal structures, up to a specified maximum percentage difference between the folding free energies of the optimal and suboptimal structures. By default, the maximum percentage difference scales with sequence length. However, the fractional uncertainties, the calculated

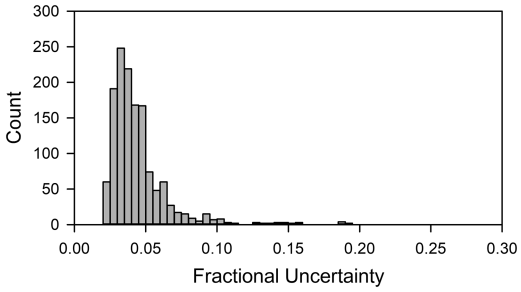

FIGURE 3. Fractional uncertainty distribution. The fractional uncertainty (uncertainty/predicted energy) was calculated for the accepted secondary structures for 1450 sequences. The distribution of the calculated fractional uncertainty is shown.

uncertainties divided by the predicted folding free energies, are largely independent of the sequence length. In fact, the width of the fractional uncertainty distribution becomes narrower as length increases (Fig. 4). This reflects the fact that the fractional uncertainty converges as the number of distinct parameters that are used increases. We hypothesize that this is because the parameter usage for a secondary structure approaches the parameter usage distribution of the general ensemble of RNA, i.e., the "average" parameter usage, as sequence length increases. Therefore, combinations of parameter usage that disproportionately increase or decrease uncertainty through parameter variation and covariation are more likely to influence the fractional uncertainty in short sequences than in large sequences.

Previously, we estimated the parameter usage frequencies among the ensemble of RNA secondary structures by counting parameter usage in a large stochastic sample of structures (Zuber et al. 2017). By sampling from this parameter usage distribution and calculating the resulting free energies and uncertainties for each sample, we reproduce the relationship between sample size and fractional uncertainty distribution (Fig. 5). This evidence supports our hypothesis that longer sequences use a wider variety and number of nearest neighbor parameters which more closely approximate the "average" parameter usage than short sequences.

This work estimates sources of uncertainty in nearest neighbor calculations that originate in the experimental uncertainty from random errors. An additional source of uncertainty is the systemic errors that are present in both the analysis of optical melting data and in nearest neighbor parameter derivation. The systematic error is harder to estimate. One source of systemic error is that optical melting experiments are analyzed assuming two-state behavior. Our previous work shows that this leads to errors in the Watson–Crick parameters (Spasic et al. 2018). Another source of systematic error is that the nearest neighbor model is incomplete. For example, recent optical melting studies demonstrate sequence-dependence to RNA folding that is not included in the existing rules (O'Toole et al. 2005; Blose et al. 2007; Clanton-Arrowood et al. 2008; Thulasi et al. 2010; Chen et al. 2012). Part of the effect of this systematic error is indirectly included in this work because it manifests as error in the parameter values fit using an incomplete model. A final systematic error is one that arises from using theoretical or empirical models to extrapolate free energy changes. We do not include uncertainty in these parameters, which include three sets of parameters: Jacobson–Stockmayer parameters, pseudoknot parameters, and the maximum asymmetry penalty. The Jacobson–Stockmayer polymer theory is used to estimate the loop closure entropy for loops larger than those that have been measured (Jacobson and Stockmayer 1950; Mathews et al. 2004). The energy parameters used to calculate the energy of a pseudoknot are similarly derived from polymer theory (Aalberts and Nandagopal 2010; Hajdin et al. 2013). Additionally, there is an empirically determined nearest neighbor parameter for the maximum asymmetry penalty for internal loops (Jaeger et al. 1989).

This new uncertainty calculation is incorporated in the folding free energy calculator (program *efn2*) in the RNAstructure package. This is available under the GNU GPL open source license and downloadable at https ://rna.urmc.rochester.edu/RNAstructure.html.

MATERIALS AND METHODS

Data table usage counts

In order to measure the number of times each data table entry is used in a folding free energy calculation, the data tables in

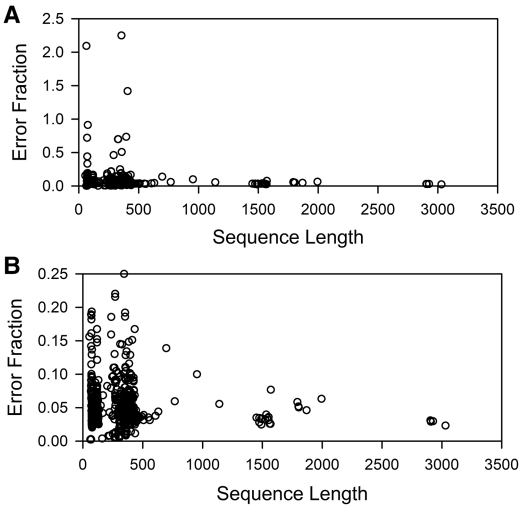

FIGURE 4. Fractional uncertainty converges with increased sequence length. (*A*) The fractional uncertainty (uncertainty in folding free energy/predicted folding free energy) is plotted against sequence length for each of the sequences in the structure archive. (*B*) Same as *A* except the *y*-axis is truncated to show more detail.

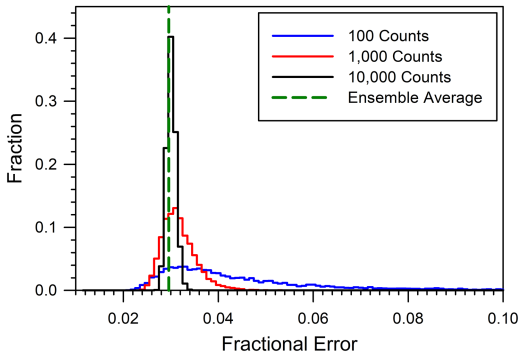

FIGURE 5. Random parameter usage sampling reproduces fractional uncertainty distributions. Parameter usage counts were randomly sampled from the ensemble parameter usage distribution and fractional uncertainties were calculated for each sample. Sample sizes were 100, 1000, and 10,000 parameter counts. As more parameter counts are included in each sample, the fractional uncertainty converges toward the fractional uncertainty for the total ensemble.

RNAstructure were implemented as a C++ class that automatically tracks the number of times each variable is accessed, as described in Zuber et al. (2018). These data table usage counts were then aggregated into a single vector.

Mapping to nearest neighbor parameters

Each data table entry used in the RNAstructure software is a linear combination of one or more nearest neighbor parameters. The data tables have a total of 13,172 entries. To map data table usages back to the 294 independent nearest neighbor parameters, the contributions of each nearest neighbor parameter to the value of each data table entry needed to be determined (Zuber et al. 2017).

One example involves terminal mismatches. Because not every terminal mismatch was experimentally measured, some are estimated from the sum of two dangling end measurements. For example, $^{5'\text{-AC-}3'}_{3'\text{-UU-}5'}$ was not measured, and its value was approximated by the sum of the constituent dangling end parameters: $^{5'\text{-AC-}3'}_{3'\text{-UU-}5'} = {}^{5'\text{-AC-}3'}_{3'\text{-U -}5'} + {}^{5'\text{-A -}3'}_{3'\text{-UU-}5'}$. However, the dangling end term $^{5'\text{-A -}3'}_{3'\text{-UU-}5'}$ was not measured and was approximated from the values of two dangling end terms: $^{5'\text{-A -}3'}_{3'\text{-UU-}5'} = \frac{1}{2}\left({}^{5'\text{-A -}3'}_{3'\text{-UA-}5'} + {}^{5'\text{-A -}3'}_{3'\text{-UC-}5'}\right)$. Therefore, a usage of $^{5'\text{-AC-}3'}_{3'\text{-UU-}5'}$ from the terminal mismatch data table must map to a single count for the $^{5'\text{-AC-}3'}_{3'\text{-U -}5'}$ parameter and a half count each for the $^{5'\text{-A -}3'}_{3'\text{-UA-}5'}$ and $^{5'\text{-A -}3'}_{3'\text{-UC-}5'}$ parameters in the dangling end data tables.

Other examples are bulge loop initiations for loops with four or more unpaired nucleotides. Because there were no experimental measurements for those loops, the initiation energies were linearly extrapolated from the initiation terms for loops of size 2 and 3 unpaired nucleotides. Therefore, for a bulge loop with four unpaired nucleotides:

$$\Delta G^\circ_{Bulge\,4} = \Delta G^\circ_{Bulge\,3} + (\Delta G^\circ_{Bulge\,3} - \Delta G^\circ_{Bulge\,2}), \quad (5)$$

and for a bulge loop of five unpaired nucleotides:

$$\Delta G^\circ_{Bulge\,5} = \Delta G^\circ_{Bulge\,3} + 2 \times (\Delta G^\circ_{Bulge\,3} - \Delta G^\circ_{Bulge\,2}). \quad (6)$$

Therefore, a count for $\Delta G^\circ_{Bulge\,5}$ maps to +3 counts for $\Delta G^\circ_{Bulge\,3}$ and −2 counts for $\Delta G^\circ_{Bulge\,2}$.

A third example is the $^{5'\text{-GG-}3'}_{3'\text{-CC-}5'}$ base pair stack and $^{5'\text{-CC-}3'}_{3'\text{-GG-}5'}$ base pair stack. These each have entries in the data tables for computational efficiency, but they are the same nearest neighbor parameter. Additionally, this parameter is also used in the flush coaxial stacking data tables (Walter et al. 1994). All of these data table usage counts need to be mapped to the same parameter usage count.

To determine the contributions of each nearest neighbor parameter to each data table entry, a new data table format was devised, where every nonindependent data table entry is described in terms of other entries. This format was originally devised so that changes in parameter values could be propagated to all the dependent data table values (Zuber et al. 2017). However, for this analysis, the same data table format was used to recursively determine the linear combination of nearest neighbor parameters that determine the value of every dependent data table entry. The coefficients of this linear combination can be used to map data table usage counts to the independent nearest neighbor parameters. The parameter mapping matrix is generated by concatenating all the linear combination vectors, forming a 2-D matrix where each column represents the linear combination vector for an individual data table entry. The nearest neighbor parameter usage counts can then be calculated using the parameter mapping matrix:

$$\mathbf{P} = \mathbf{M} \times \mathbf{D}, \quad (7)$$

where $\boldsymbol{P}$ is the parameter usage vector of independent parameters (294 × 1 using the 2004 Turner rules), $\boldsymbol{M}$ is the parameter map (294 × 13,172 elements), and $\boldsymbol{D}$ is the data counts vector (13,172 × 1). The vector $\boldsymbol{P}$ represents the linear combination of nearest neighbor parameters required to calculate the folding free energy of a given secondary structure.

Accounting for parameter covariance

To accurately assess the experimental uncertainty from the linear combination of multiple parameters requires the estimation of parameter covariance. We previously determined the free energy parameter covariance at 37°C (Zuber et al. 2018). To extrapolate to other temperatures, the parameter covariance was estimated by perturbing the experimental $\Delta H°$ and $\Delta S°$ values for the optical melts that were used to derive the nearest neighbor parameters within experimental uncertainty and rederiving the nearest neighbor parameters. In order to generate an input data set that accurately reflects the high Pearson correlation (0.9996) between the enthalpy and entropy measurements (Xia et al. 1998), the matrix **c** needed to be determined for each experiment, such that

$\boldsymbol{c} \times \boldsymbol{c}^{\top} = \boldsymbol{\Sigma_E}$, where Σ_E is the desired covariance matrix for the experiment, defined by

$$\boldsymbol{\Sigma_E} = \begin{bmatrix} \boldsymbol{\sigma^2_{\Delta H}} & \boldsymbol{\rho} \times \boldsymbol{\sigma_{\Delta H}} \times \boldsymbol{\sigma_{\Delta S}} \\ \boldsymbol{\rho} \times \boldsymbol{\sigma_{\Delta H}} \times \boldsymbol{\sigma_{\Delta S}} & \boldsymbol{\sigma^2_{\Delta S}} \end{bmatrix}, \quad (8)$$

where $\boldsymbol{\rho}$ is the correlation coefficient between enthalpy and entropy measurements (Xia et al. 1998) and $\sigma_{\Delta H}$ and $\sigma_{\Delta S}$ are the experimental uncertainty in enthalpy and entropy, respectively. The matrix $\boldsymbol{c}$ can be computed using Cholesky decomposition (Watkins 2002).

During the calculation of the multibranch loop folding free energy change and enthalpy change, there is an optimization to determine the lowest free energy configuration of dangling ends, terminal mismatches, and coaxial stacks in the multibranch loop (Mathews and Turner 2002; Tyagi and Mathews 2007; Turner and Mathews 2010). This optimization was performed using free energy changes; the folding enthalpy calculation needs to use the optimal configuration. To do this, the folding free energies and the enthalpies were determined simultaneously.

Generating parameter sets by randomly perturbing experimental enthalpy and entropy values for the 802 optical melting experiments results in a data set that can be used to calculate parameter variances and covariances for both folding free energies and enthalpies (Zuber et al. 2018). The resulting variance-covariance matrix can be used to estimate the experimental uncertainty using two matrix multiplications using Equation 4, repeated here:

$$\boldsymbol{\sigma}^2 = \mathbf{P} \times \Sigma_{\mathbf{P}} \times \mathbf{P}^{\top}, \quad (4)$$

where $\boldsymbol{\Sigma_P}$ is the variance-covariance matrix for the nearest neighbor parameters (not to be confused with $\boldsymbol{\Sigma_E}$ in Equation 8) and $\boldsymbol{P}$ is parameter usage determined using Equation 7. The effects of covariation can be ignored by setting $\boldsymbol{\Sigma}_{ij} = 0 : i \neq j$. In that case, only the parameter variances contribute to the estimated uncertainty in the predicted folding free energy change.

Calculating experimental uncertainty in free energy calculations

The program *efn2*, in the RNAstructure package, was instrumented to track the data table usage when calculating the folding free energy of an RNA secondary structure. The software reads the parameter mapping matrix and the variance-covariance matrix from disk. The software then performs the matrix operations required to convert data table usage to uncertainty in folding free energy change for a given RNA secondary structure. A fractional uncertainty can be calculated by dividing the calculated uncertainty in folding free energy by the predicted folding free energy of a given secondary structure.

To calculate uncertainties at temperatures other than 37°C, $\sigma_{\Delta G37}$ and $\sigma_{\Delta H}$ are first calculated using the parameter usage counts and the variance-covariance matrices for free energies and enthalpies. From the equation to determine the folding free energy: $\Delta G^{\circ} = \Delta H^{\circ} - T \times \Delta S^{\circ}$, the equation for uncertainty in the folding free energy is

$$\sigma^2_{\Delta G} = \sigma^2_{\Delta H} + \mathrm{T}^2\sigma^2_{\Delta S} - 2\rho\mathrm{T}\sigma_{\Delta H}\sigma_{\Delta S}, \quad (9)$$

where ρ is the correlation coefficient between ΔH and ΔS. Given a $\sigma^2_{\Delta G37}$ at 37 °C and $\sigma^2_{\Delta H}$, $\sigma_{\Delta S}$ can be determined by solving for the root of the quadratic equation $T^2_{37}\sigma^2_{\Delta S} - 2\rho T_{37}\sigma_{\Delta H}\sigma_{\Delta S} + \sigma^2_{\Delta H} - \sigma^2_{\Delta G37} = 0$, which yields

$$\sigma_{\Delta S} = \frac{\rho\sigma_{\Delta H} - \sqrt{\rho^2\sigma^2_{\Delta H} - \sigma^2_{\Delta H} + \sigma^2_{\Delta G37}}}{\mathrm{T}}. \quad (10)$$

This $\sigma_{\Delta S}$ is then used in Equation 9 to calculate the uncertainty in the folding free energy at arbitrary temperature.

Randomly sampling parameter usage counts

To verify that the length dependence of the fractional uncertainty can be attributed to variations in sample size, sets of synthetic parameter usage counts were generated. A parameter usage count distribution was generated by summing the parameter usage counts for 10,000 stochastically sampled secondary structures for 1650 RNA sequences (16,500,000 total predicted secondary structures) (Zuber et al. 2017). Sets of parameter usages were randomly sampled from that distribution with sample sizes of 100, 1000, and 10000 parameters. The uncertainty in the free energy prediction for each sample was calculated using Equation 4. The free energy for each sample was calculated by taking the dot product of the parameter usage count vector and the vector of parameter values.

Structure archive

The uncertainty estimation method was performed on a set of 1450 sequences. The RNA families in this collection include 5S rRNA (309 sequences; 119.5 nt mean length), 16S rRNA (21 sequences, 1512.7 nt mean length), 23S rRNA (four sequences, 2577.5 nt mean length), tRNA (484 sequences, 77.5 nt mean length), tmRNA (462 sequences, 366.0 nt mean length), Group I Introns (25 sequences, 343.0 nt mean length), Group II Introns (3 sequences, 668.7 nt mean length), RNase P RNA (15 sequences, 378.7 nt mean length), SRP RNA (91 sequences, 267.9 nt mean length), and telomerase RNA (37 sequences, 444.5 nt mean length). These structured RNA sequences were previously assembled for structure prediction accuracy benchmarks (Bellaousov and Mathews 2010).

ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences (R01GM076485).

Received October 8, 2018; accepted April 2, 2019.

REFERENCES

Aalberts DP, Nandagopal N. 2010. A two-length-scale polymer theory for RNA loop free energies and helix stacking. *RNA* **16:** 1350–1355. doi:10.1261/rna.1831710

Andronescu M, Aguirre-Hernandez R, Condon A, Hoos HH. 2003. RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res* **31:** 3416–3422. doi:10.1093/nar/gkg612

Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. 2010. Computational approaches for RNA energy parameter estimation. *RNA* **16:** 2304–2318. doi:10.1261/rna.1950510

Andronescu M, Condon A, Turner DH, Mathews DH. 2014. The determination of RNA folding nearest neighbor parameters. *Methods Mol Biol* **1097:** 45–70. doi:10.1007/978-1-62703-709-9_3

Bellaousov S, Mathews DH. 2010. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA* **16:** 1870–1880. doi:10.1261/rna.2125310

Blose JM, Manni ML, Klapec KA, Stranger-Jones Y, Zyra AC, Sim V, Griffith CA, Long JD, Serra MJ. 2007. Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops. *Biochemistry* **46:** 15123–15135. doi:10.1021/bi700736f

Chen JL, Dishler AL, Kennedy SD, Yildirim I, Liu B, Turner DH, Serra MJ. 2012. Testing the nearest neighbor model for canonical RNA base pairs: revision of GU parameters. *Biochemistry* **51:** 3508–3522. doi:10.1021/bi3002709

Clanton-Arrowood K, McGurk J, Schroeder SJ. 2008. 3′ terminal nucleotides determine thermodynamic stabilities of mismatches at the ends of RNA helices. *Biochemistry* **47:** 13418–13427. doi:10.1021/bi801594k

Crothers DM, Cole PE, Hilbers CW, Shulman RG. 1974. The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J Mol Biol* **87:** 63–88. doi:10.1016/0022-2836(74)90560-9

Ding Y, Lawrence CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* **31:** 7280–7301. doi:10.1093/nar/gkg938

Ding Y, Chan CY, Lawrence CE. 2004. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res* **32:** W135–W141. doi:10.1093/nar/gkh449

Do CB, Woods DA, Batzoglou S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22:** e90–e98. doi:10.1093/bioinformatics/btl246

Doudna JA, Cech TR. 2002. The chemical repertoire of natural ribozymes. *Nature* **418:** 222–228. doi:10.1038/418222a

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci* **110:** 5498–5503. doi:10.1073/pnas.1219988110

Hofacker IL. 2014. Energy-directed RNA structure prediction. *Methods Mol Biol* **1097:** 71–84. doi:10.1007/978-1-62703-709-9_4

Jacobson H, Stockmayer WH. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J Chem Phys* **18:** 1600–1606. doi:10.1063/1.1747547

Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci* **86:** 7706–7710. doi:10.1073/pnas.86.20.7706

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 26. doi:10.1186/1748-7188-6-26

Lu ZJ, Turner DH, Mathews DH. 2006. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res* **34:** 4912–4924. doi:10.1093/nar/gkl472

Mathews DH. 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **10:** 1178–1190. doi:10.1261/rna.7650904

Mathews DH. 2006. Revolutions in RNA secondary structure prediction. *J Mol Biol* **359:** 526–532. doi:10.1016/j.jmb.2006.01.067

Mathews DH, Turner DH. 2002. Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry* **41:** 869–880. doi:10.1021/bi011441d

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci* **101:** 7287–7292. doi:10.1073/pnas.0401799101

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119. doi:10.1002/bip.360290621

Onoa B, Tinoco I. 2004. RNA folding and unfolding. *Curr Opin Struct Biol* **14:** 374–379. doi:10.1016/j.sbi.2004.04.001

O'Toole AS, Miller S, Serra MJ. 2005. Stability of 3′ double nucleotide overhangs that model the 3′ ends of siRNA. *RNA* **11:** 512–516. doi:10.1261/rna.7254905

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11:** 129. doi:10.1186/1471-2105-11-129

Rivas E, Lang R, Eddy SR. 2012. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **18:** 193–212. doi:10.1261/rna.030049.111

Seetin MG, Mathews DH. 2012. RNA structure prediction: an overview of methods. *Methods Mol Biol* **905:** 99–122. doi:10.1007/978-1-61779-949-5_8

Serganov A, Nudler E. 2013. A decade of riboswitches. *Cell* **152:** 17–24. doi:10.1016/j.cell.2012.12.024

Spasic A, Berger KD, Chen JL, Seetin MG, Turner DH, Mathews DH. 2018. Improving RNA nearest neighbor parameters for helices by going beyond the two-state model. *Nucleic Acids Res* **46:** 4883–4892. doi:10.1093/nar/gky270

Thulasi P, Pandya LK, Znosko BM. 2010. Thermodynamic characterization of RNA triloops. *Biochemistry* **49:** 9058–9062. doi:10.1021/bi101164s

Tinoco I Jr, Bustamante C. 1999. How RNA folds. *J Mol Biol* **293:** 271–281. doi:10.1006/jmbi.1999.3001

Turner DH, Mathews DH. 2010. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* **38:** D280–D282. doi:10.1093/nar/gkp892

Tyagi R, Mathews DH. 2007. Predicting helical coaxial stacking in RNA multibranch loops. *RNA* **13:** 939–951. doi:10.1261/rna.305307

Walter AE, Turner DH, Kim J, Lyttle MH, Muller P, Mathews DH, Zuker M. 1994. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci* **91:** 9218–9222. doi:10.1073/pnas.91.20.9218

Watkins D. 2002. *Fundamentals of matrix computations*. John Wiley & Sons, Inc., New York, NY.

Woodson SA. 2000. Recent insights on RNA folding mechanisms from catalytic RNA. *Cell Mol Life Sci* **57:** 796–808. doi:10.1007/s000180050042

Wu L, Belasco JG. 2008. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol Cell* **29:** 1–7. doi:10.1016/j.molcel.2007.12.010

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165. doi:10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G

Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37:** 14719–14735. doi:10.1021/bi9809425

Yu YT, Meier UT. 2014. RNA-guided isomerization of uridine to pseudouridine–pseudouridylation. *RNA Biol* **11:** 1483–1494. doi:10.4161/15476286.2014.972855

Zuber J, Sun H, Zhang X, McFadyen I, Mathews DH. 2017. A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction. *Nucleic Acids Res* **45:** 6168–6176. doi:10.1093/nar/gkx170

Zuber J, Cabral B, McFadyen I, Mauger DM, Mathews DH. 2018. Analysis of RNA nearest neighbor parameters reveals interdependencies and quantifies the uncertainty in RNA secondary structure prediction. *RNA* **24:** 1568–1582. doi:10.1261/rna.065102.117

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415. doi:10.1093/nar/gkg595