6168–6176 *Nucleic Acids Research, 2017, Vol. 45, No. 10* *Published online 15 March 2017*
doi: 10.1093/nar/gkx170

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

Jeffrey Zuber[1,†], Hongying Sun[1,†], Xiaoju Zhang[1], Iain McFadyen[2] and David H. Mathews[1,3,*]

[1]Department of Biochemistry & Biophysics and Center for RNA Biology, University of Rochester Medical Center, Rochester, NY 14642, USA, [2]Computational Sciences, Moderna Therapeutics, Cambridge, MA 02141, USA and [3]Department of Biostatistics & Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA

Received October 16, 2016; Revised March 01, 2017; Editorial Decision March 03, 2017; Accepted March 10, 2017

ABSTRACT

Nearest neighbor parameters for estimating the folding energy changes of RNA secondary structures are used in structure prediction and analysis. Despite their widespread application, a comprehensive analysis of the impact of each parameter on the precision of calculations had not been conducted. To identify the parameters with greatest impact, a sensitivity analysis was performed on the 291 parameters that compose the 2004 version of the free energy nearest neighbor rules. Perturbed parameter sets were generated by perturbing each parameter independently. Then the effect of each individual parameter change on predicted base-pair probabilities and secondary structures as compared to the standard parameter set was observed for a set of sequences including structured ncRNA, mRNA and randomized sequences. The results identify for the first time the parameters with the greatest impact on secondary structure prediction, and the subset which should be prioritized for further study in order to improve the precision of structure prediction. In particular, bulge loop initiation, multibranch loop initiation, AU/GU internal loop closure and AU/GU helix end parameters were particularly important. An analysis of parameter usage during folding free energy calculations of stochastic samples of secondary structures revealed a correlation between parameter usage and impact on structure prediction precision.

INTRODUCTION

It is increasingly clear that RNA sequences serve many essential roles aside from their functions in the expression of proteins. Non-coding RNAs (ncRNA), functional RNAs that are not transcribed into protein, perform diverse functions, including regulation of gene expression as siRNA or miRNA (1), reaction catalysis as ribozymes (2), metabolite detection as riboswitches (3) and target identification as guide RNAs (4).

The functions of many RNAs are determined by their structure. RNA structure is hierarchical (5). The primary structure is the linear sequence of nucleotides, connected by covalent bonds. The secondary structure is the canonical base pairing between nucleotides in the RNA, and these base pairs are organized as A-form helices. The tertiary structure is the positions of all atoms in the RNA in three dimensions, which is organized by hydrogen bonds and base stacking. The secondary structure generally forms faster (6) and is generally more thermostable (7,8) than tertiary structure, therefore secondary structure can be predicted independently of tertiary structure.

To estimate the free energy change of folding to a secondary structure from random coil, a set of parameters called the nearest neighbor parameters can be used (9). 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 (10). These parameters are used widely in software programs for RNA secondary structure prediction (11–13). Additionally, methods that infer folding parameters from the set of sequences with known structure also generally use the same functional forms (14–16). The nearest neighbor

*To whom correspondence should be addressed. Tel: +1 585 275 1734; Fax: +1 585 275 6007; Email: David_Mathews@urmc.rochester.edu
†These authors contributed equally to the paper as first authors.

database (NNDB) provides the set of current RNA folding parameters and also provides examples for their use (17).

Most prior work benchmarking nearest neighbor parameters focused on the accuracy of secondary structure prediction (9,11,18–20). Another aspect that has received less attention is how uncertainty in the values of parameters results in implicit uncertainty in structure prediction, i.e. the precision of structure prediction. In one study, parameters were adjusted within experimental uncertainty to generate alternative secondary structures with the goal of providing alternative hypotheses for the structure to improve the structure prediction of a given sequence (21). Another study showed that randomly perturbing all the thermodynamic parameters simultaneously results in different predicted structures, and that highly probably base pairs, as determined by partition function calculations, are more robust to changes in thermodynamic parameters (22). A recent parametric analysis of the multibranch loop initiation parameters demonstrated that overall RNA branching topology is not sensitive to changes in the three multibranch loop parameters (23).

In this work, a sensitivity analysis was performed to determine the extent to which errors in the estimates of the nearest neighbor parameters result in uncertainty in RNA structure prediction, focusing on the estimates of ensemble base pairing probability from partition function calculations. The sensitivity analysis was performed by varying each parameter, one at a time, up or down in value. The magnitude of the change for the parameter was either related to the experimental uncertainty of the parameter or to a flat fixed value across all parameters, which facilitated the comparison of sensitivity across parameters. The uncertainty was then quantified as a root mean squared deviation (RMSD) of the base pairing probability estimates as compared to those calculated using the current, reference parameters or as changes in structure prediction. In order to identify factors that determine the impact of a given parameter, the relative frequencies of use for the different nearest neighbor parameters in probable RNA secondary structures were determined. This comprehensive analysis of the contribution to the uncertainty by each parameter on the variability of the pair probability estimates and secondary structure predictions identified parameters and functional forms that should be refined by future experimental studies. The analysis also identified the most significant parameters that need to be determined precisely for the precise modeling of RNA secondary structures with modified alphabets, e.g. synthetic nucleotides or modified nucleotides. This analysis is the first performed on the nearest neighbor parameters that has systematically quantified the impact of each individual parameter on structure predictions. It is also the first that has been done using experimental uncertainties for all parameters because not all the uncertainties in the loop parameters have been reported previously.

MATERIALS AND METHODS

Software

Calculations were performed using the RNAstructure package (13). Specifically, partition function (program *partition*) (24), stochastic sampling (program *stochastic*), *ProbKnot* (25), secondary structure comparison (program *scorer*) and the folding free energy calculator (program *efn2*) were used.

Tabulating RNA thermodynamic parameter standard errors

This work used the 2004 set of folding free energy parameters. For the loop parameters, these were previously reported to tenths of a kcal/mol precision (9,17). For this work, these parameters were recalculated to a higher precision, i.e. to hundredths of a kcal/mol. Additionally, error estimates for each parameter were determined through the propagation of errors, calculation of the standard error of means or the standard error from a regression analysis, as appropriate. Experimental errors were determined by approximating the uncertainty in change in enthalpy as 12% of the measured ΔH° and the uncertainty in the change in entropy as 13.5% of the measured ΔS°, following (26). The uncertainty in the measured folding free energy is then determined by propagating those uncertainties through the free energy calculation, taking into consideration the correlation between enthalpy and entropy (26). When a parameter is a mean of up to five experiments, errors are propagated using the error propagation method:

$$\sigma^2 = \sum_i \sigma_i^2 \left(\frac{\partial \Delta G^\circ}{\partial \Delta G_i^\circ} \right)^2$$

where σ is the error estimate in the nearest neighbor parameter ΔG°, σ_i is the error estimate for experiment i, and ΔG°_i is the free energy from experiment i. For means, the error propagation reduces to:

$$\sigma^2 = \sum_{i=1}^{N} \left[\sigma_i^2 \left(\frac{1}{N} \right)^2 \right]$$

for N experiments. For five parameters, the parameter is the mean of six or more experimentally determined values and the standard error of the mean is the estimated error. For the parameters determined by linear regression, the standard error of the regression is the estimated error.

Parameters used by RNAstructure are stored in plain text files, organized by parameter classes. Files exist for the following classes: helical stacking for canonical base pairs, dangling ends, terminal mismatches, coaxial stacking, loop initiation, hairpin loops with stability not well modeled by generic terms and with length 3, 4 or 6 unpaired nucleotides (triloops, tetraloops and hexaloops), coaxial stacking for stacks without an intervening mismatch, mismatch-mediated coaxial stacking with an intervening mismatch, multibranch loop terminal mismatches, hairpin loop terminal mismatches, internal loop terminal mismatches, 1 × n internal loop terminal mismatches, 2 × 3 internal loop terminal mismatches, 1 × 1 internal loops, 1 × 2 internal loops and 2 × 2 internal loops. In addition, there are a number of implicit parameters that do not appear in the final tables themselves but are used to generate other parameters that are included. For example the table used to lookup energies for internal loop first mismatch terms has a total of 96 parameters. However each parameter is simply a combination of AU/GU closure, GA or AG first mismatch, GG first mismatch or UU first

mismatch terms. In total, there are 13 254 parameters either explicitly or implicitly included in the data tables. The NNDB (http://rna.urmc.rochester.edu/NNDB) defines the structure classes, provides the tables, and also provides instruction for using the parameters (17).

For this project, the set of independent parameters, i.e. the set of adjustable parameters, was identified. This is a smaller set of 291 parameters. The total parameters (13 254) include duplicate parameters due to symmetry, precalculated approximations (using the implicit parameters), and redundant parameters used in functional forms that are not implemented. For symmetry, the tables have redundant entries, where the same entry appears in two strand orientations. For example, in the base pair stack table, the stability for a stack of a GC base pair followed by a GC base pair is the same as CG base pair followed by a CG base pair. In the former case, the consecutive Gs are oriented in the top strand, and, in the latter, the two Cs are oriented in the top strand. The unimplemented functional forms are those that are implemented in software, but not used by the 2004 nearest neighbor parameters. For example, RNAstructure supports different parameter values for terminal mismatches in multibranch and exterior loops, but these are identical using the current nearest neighbor rules.

A compilation of the nearest neighbor parameters, grouped by parameter class and their error estimates are provided in an Excel file in the Supplementary Data. Included in the file are all the calculations that were required to derive the parameters as well as the list of references from which the optical melting data were sourced.

New data table formats

For this project, the 2004 nearest neighbor parameters were implemented in an improved data table format for RNAstructure. The new data table format removed unnecessary entries and made the tables more human and machine readable.

In addition, for this project, another data table format was implemented that allowed for the propagation of parameter values. The second data table format allowed parameters to be defined based on the values of other parameters, making explicit the relationships between parameters. This allows parameters to remain consistent (such that symmetric parameters are always equal to each other) and for changes in parameters to propagate through the dependent parameters. This ensured that changing the value of a stacking term will always also change the value of the symmetric stack. This also ensured that the values of the pre-calculated approximations are updated when the value of an implicit parameter is changed.

Sequence archive

There were 1663 sequences used in this analysis. The sequence families in this archive include 5S rRNA (309 sequences), 16S rRNA (21 sequences), 23S rRNA (4 sequences), tRNA (484 sequences), tmRNA (462 sequences), Group I Introns (25 sequences), Group II Introns (3 sequences), RNase P RNA (15 sequences), SRP RNA (91 sequences), mRNAs (100 sequences), telomerase RNA (37 sequences) and randomly shuffled sequences (100 sequences). The structural RNA sequences were previously assembled for structure prediction accuracy benchmarks (25). The mRNAs were from the RefSeq database and included 5' and 3' UTRs (27). The mRNAs were randomly selected from ∼90 000 human mRNA sequences, limited to those that were <1.5 kb in length. The shuffled RNA sequences were randomly selected from the archive and shuffled such that the dinucleotide frequency was maintained. The shuffled sequences were generated using the Python module uShuffle, which implements the Euler algorithm to randomly permute a sequence while maintaining *k*-let frequencies for an arbitrary *k* (28).

Sensitivity analysis

The sensitivity analysis was performed by perturbing each independent parameter with perturbations ranging from −3 σ to +3 σ, in increments of one σ, where σ is either the standard error for the parameter or a flat value of 0.5 kcal/mol. Using standard error reveals those parameters that have a large impact on structure prediction relative to how well defined that parameter is, suggesting parameter classes that can be the focus of future experiments. Using a flat value allows a comparison of the impact of different parameters, identifying those parameters whose precise values are the most important to determine for non-standard nucleotides.

The standard error for a parameter is the estimate of the magnitude of the error for the mean of the parameter, and the standard error scales with the reciprocal of the square root of the number of measurements (29). The standard error is the proper estimate of the error for a parameter because the major source of error is random experimental errors; therefore taking multiple measurements reduces the error in the parameter estimate. Standard deviation, in contrast, is an estimate of the width of the distribution of a parameter and is a reflection of the magnitude of the random errors. As such, standard error is used throughout the sensitivity analysis.

Using the perturbed parameter sets, new data tables for RNAstructure were generated following the rules outlined in the NNDB (17). This ensured that symmetric parameters for base pairs and internal loops always had equivalent values. Additionally, the precalculated approximations, such as those for unmeasured 1 × 1, 2 × 1 and 2 × 2 internal loop parameters are updated to reflect the perturbed parameter values. The perturbed data tables were then used to calculate the pair probability of each possible base pair of each sequence in the archive using the programs *partition* and *ProbabilityPlot*. The program *ProbabilityPlot* outputs the probability of all possible base pairs, which are those base pairs that can form an allowed pair (A-U, G-C, G-U) and can form a run of two or more base pairs.

RMSDs of the pair probabilities were calculated for each sequence, comparing pair probabilities calculated from each of the perturbed data tables to the probabilities calculated with unperturbed data tables (the reference parameter set):

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{\mathrm{All\ BP}} (P_N - P_R)^2}{N_{\mathrm{BP}}}}$$

where N_{BP} is the number of possible base pairs, P_N is the base pair probability calculated with the perturbed data tables and P_R is the base pair probability calculated with the reference data tables. N_{BP} is the sum, for each sequence, of the total number of possible canonical (AU, CG and GU) pairs for that sequence, where pairs are also required to be able to form a helix with at least two stacked base pairs.

Structures were predicted from the pair probabilities (both perturbed and reference parameter sets) using *ProbKnot* (25). *ProbKnot* is a method to predict maximum expected accuracy structures (14). It assembles structures with base pairs of nucleotides that are mutually maximal base pairing partners. Thus, i is paired with j if and only if the nucleotide with highest pairing probability with i is j and the nucleotide with the highest pairing probability for j is i.

To quantify the difference in predicted structures between a perturbed data set and the reference data set, a sensitivity defect and a positive predictive value (PPV) defect were calculated for the secondary structures predicted using perturbed parameter tables as compared to secondary structures predicted using the reference-parameter tables. Sensitivity defect and PPV defect were defined as a measure of the difference in the two predicted structures:

$$\text{Sensitivity Defect} = 100 \times \left(1 - \frac{N_{\text{BP with both tables}}}{N_{\text{BP with reference tables}}}\right)$$

$$\text{PPV Defect} = 100 \times \left(1 - \frac{N_{\text{BP with both tables}}}{N_{\text{BP with perturbed tables}}}\right)$$

where $N_{\text{BP with both tables}}$ is the number of pairs that appear in both predicted structures, $N_{\text{BP with perturbed tables}}$ is the number of pairs in the structure predicted with the perturbed tables and $N_{\text{BP with reference tables}}$ is the number of base pairs predicted with the standard nearest neighbor rules. A sensitivity defect of 0 indicates that all pairs predicted by the reference parameters are also predicted by the perturbed parameters. A PPV defect of 0 indicates that all the pairs predicted by perturbed parameters are also predicted by the reference parameters. Base pairs were considered identical even if one of the nucleotides in the pair was shifted by up to one nucleotide in either direction. Therefore, pair i-j for one set of parameters would be considered the same pair as i–j, (i + 1) − j, (i − 1) − j, i − (j + 1) or i − (j − 1). This is because thermal energies are sufficient for pairs to fluctuate in this manner (30,31).

Parameter usage counting by stochastic sampling

To calculate how frequently each parameter is used for estimating folding free energies for probable structures, 10 000 secondary structures were sampled from the Boltzmann ensemble for each sequence in the archive using the program *stochastic*, based on calculations using unperturbed data tables (32). Then, parameter usage was counted while the free energy change of each of the secondary structures in the stochastic sample was calculated using a free energy change calculator, *efn2*.

efn2 was modified with the addition of a custom data type that returns a parameter value while also counting how often that parameter value was called. Both multibranch and exterior loops can adopt multiple potential configurations of coaxial stacks, terminal mismatches, and dangling ends. The functions calculating the folding free energies of multibranch and exterior loops use recursive algorithms to determine the energy of the optimal configuration and had to be modified so that parameter usage counts were not incremented during recursive calculations and only counted during the traceback steps of those functions. Additionally, *efn2* was modified to increment the counts of those parameters that are used in a multiplicative fashion by the multiplier. For example, the multibranch loop per helix penalty needed to be counted once per branching helix.

RESULTS

One-at-a-time sensitivity analysis with experimental parameter errors

To determine the impact of experimental uncertainty in independent parameter values on the precision of pair probability estimation, single independent parameters were adjusted from their reference values by ±3, ±2 or ±1 σ, where σ is the experimentally-derived standard error for each parameter, resulting in perturbed parameter sets. Partition function calculations were performed to estimate base pairing probabilities for each of 1663 sequences for each parameter set. Mean base pair probability RMSD was calculated for each of these single parameter changes as compared to the reference parameters. The estimated base pairing probabilities were then used to predict a secondary structure for each sequence using *ProbKnot*, which predicts a maximum expected accuracy secondary structure, including those with pseudoknots (25). To quantify the change in predicted secondary structure as compared to the reference parameters, two structural defect metrics (Sensitivity Defect and PPV Defect) were calculated for each sequence.

This analysis illustrates the impact of each parameter on the precision of base pairing probabilities relative to how well defined that parameter is. The average base pair probability RMSDs for each independent parameter are shown in Figure 1A for ±3 standard errors. The same trends were observed for parameter sets with a single parameter adjusted by ±2 or ±1 standard errors, with smaller magnitudes of RMSDs, sensitivity defects and PPV defects. These data are available in an Excel file provided in the Supplementary Data.

A high linear correlation was observed between RMSD and sensitivity defect (R^2 = 0.989, Supplementary Figure S1) and also between sensitivity defect and PPV defect (R^2 = 0.998, Supplementary Figure S2). The correlations depend on the RNA family being studied, and the correlations for each family are available in the Excel file in the Supplementary Data.

Parameters whose errors had the greatest impact on estimated base pair probabilities include canonical pair stacking in helices (stacking parameters in Figure 1), multibranch loop terms (miscellaneous loop parameters in Figure 1), hairpin and bulge loop initiations (loop initiation in Figure 1) and coaxial stacking parameters. Parameters with minimal impact on the estimated base pair probabilities include hairpin loop folding free energies for specific sequences and specific internal loop parameters.

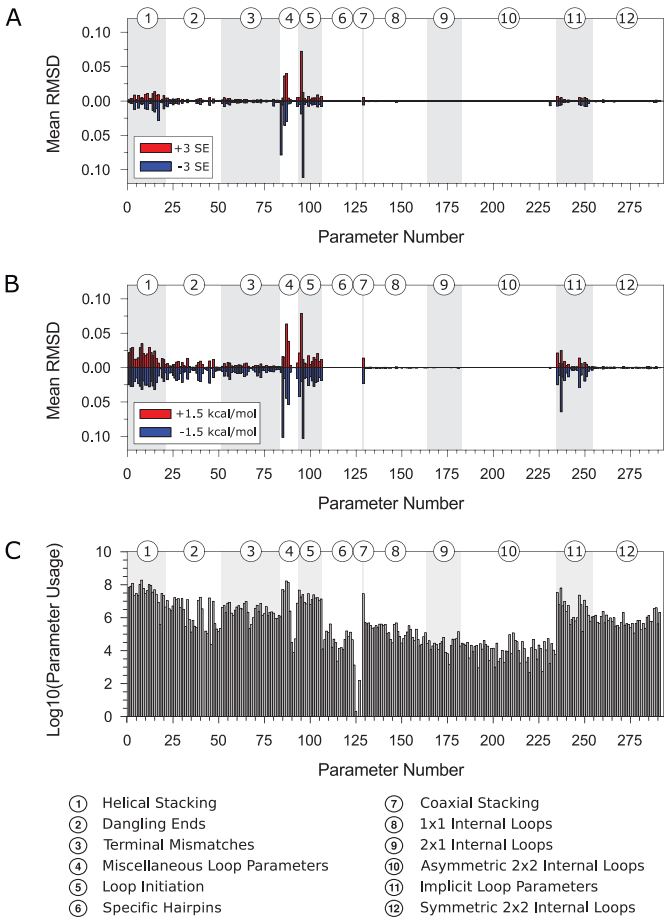

Figure 1. Sensitivity analysis. In each panel, independent parameters are along the x-axis, organized by motif type and with a key below the plot. (**A**) Mean base pair probability RMSD for the entire sequence archive except randomized sequences for ±3 standard errors. The RMSDs for +3 standard errors are shown above the x-axis, while the RMSDs for −3 standard error are shown below the x-axis. (**B**) The sensitivity analysis using flat errors across all parameters. The analysis was performed as in Figure 1A, except a σ value of 0.5 kcal/mol was used for each parameter instead of using the experimentally determined errors. (**C**) The counts of parameter use. Use counts for each parameter were tabulated for folding free energy calculations for secondary structures sampled from the Boltzmann ensamble. This measurement was performed for all sequences. The counts for the dependent parameters were attributed to the independent parameters on which the dependent parameters depend.

One-at-a-time sensitivity analysis with flat parameter errors

The comparison of the magnitude of effects of perturbing individual parameters was complicated by the varying magnitudes of experimental errors across the parameters. For example, the mean standard error for stacks of Watson–Crick pairs is 0.07 kcal/mol, but the mean standard error for all independent parameters is 0.38 kcal/mol, with variation from 0.03 to 1.47 kcal/mol. Therefore, to compare the influence of each parameter relative to other parameters, sensitivity analysis was repeated using a flat σ value of 0.5 kcal/mol for each parameter. The average RMSDs from this analysis are shown in Figure 1B.

Compared to the results using experimental errors for each parameter, the flat errors resulted in several differences. Notably, the stacking parameters had a greater effect on base pair probability estimate precision with the flat errors, which is not surprising considering the relatively low estimated errors for those parameters. In addition, loop initiation parameters and the implicit internal loop parameters had larger impacts on base pair probabilities with the flat error value compared to the results using the experimental errors, reflecting the relatively low experimental uncertainty (0.05–0.31 kcal/mol) for these parameters.

Increasing the stability of the internal loop asymmetry parameter by subtracting 1.5 kcal/mol resulted in a number of sequences for which there is no predicted secondary structure. This is because the asymmetry term became favorable, making increasingly large asymmetric loops dominantly favorable. As a result, *ProbKnot* does not predict any helices as long as the default minimum allowed helix length (3 bp) and all the base pairs are thus removed. For the affected sequences, the PPV defect was set to 100%. Approximately 2.5% of the sequences exhibited this behavior for this particular parameter set. No other parameter sets were affected.

Parameter usage counting

One method to track parameter usage is to track the number of times that a parameter is called by the partition function. However, due to the recursive nature of the dynamic programming algorithm used by the partition program, this approach would only return the explicit usage counts, ignoring the implicit parameter usage caused by recursion to prior calculated values as part of the dynamic programming algorithm. Instead, the energy calculator program *efn2* was instrumented to track the total number of times each nearest neighbor parameter was used in the calculation of free energy changes for secondary structures stochastically sampled from the Boltzmann distribution (32). This approach returns the parameter usage for a set of secondary structures representative of the ensemble. The cumulative parameter usage counts were tracked for the entire sequence archive (Figure 1C). The most-used parameters were the helical stacking parameters, AU/GU helical end terms, multibranch loop parameters, internal loop asymmetry, single nucleotide bulge loop initiation and the mismatch-mediated coaxial stacking parameter.

Figure 2 shows a plot of logarithm of RMSD from the analysis using flat errors, as a function of logarithm of parameter usage count for all parameters. This plot demonstrates that the effect on pairing probability estimate precision for a thermodynamic parameter varies as a function of the number of uses of that parameter. Parameters used more often to evaluate the free energies in the folding ensemble are associated with higher RMSD.

The average change in pair probability with parameter perturbation is a function of pair probability

To test whether the magnitude in change in pairing probability depended on the pairing probability, the mean absolute value of pairing probability change as function of pairing probability was plotted (Figure 3A). In this analysis, all possible base pairs across all sequences were binned (in intervals of 1%) according to their probability estimated using the reference thermodynamic parameters. The set of all

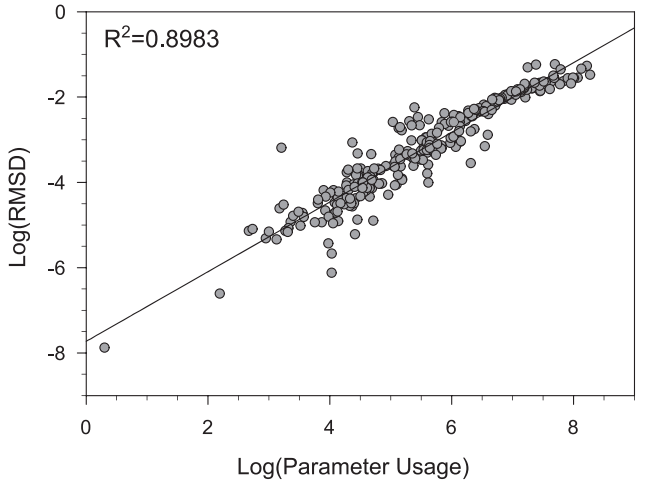

Figure 2. Parameter usage counts correlate with RMSD. The $\log_{10}$ of RMSD as a function of the $\log_{10}$ of the thermodynamic parameter usage count for calculating folding energies of a stochastic sample across all sequences. RMSD was calculated using a flat error estimate of +3σ (1.5 kcal/mol). A best fit line is shown and the linear correlation coefficient, R^2, is 0.8983.

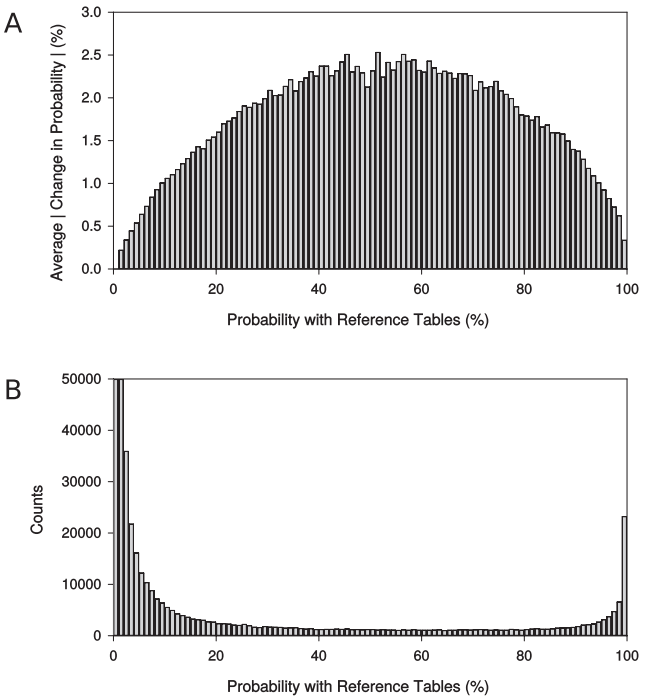

Figure 3. The sensitivity of base pairing probability to parameter change is a function of the probability of the pair. (**A**) The mean absolute value of change in pairing probability plotted as a function of pairing probability. The change in each base pair probability in the entire sequence archive was averaged over every independent parameter change of −3 standard errors. The changes were then averaged for every pair probability bin. (**B**) A plot of the pair probability distribution. Shown is a histogram of the reference base pair probabilities. Note that ∼98% of the pair probabilities have a value <1%; the y-axis was limited to 50 000 counts per bin (the number of counts for the 0–1% bin is 17.44 million and the number of counts in the 1–2% bin is 69 865).

possible base pairs are all base pairs that can form a canonical base pair (A-U, G-C, G-U) and can also form a helix of two or more stacked base pairs. Then the mean of the absolute value of the change in pairing probability was averaged for all perturbed parameter sets for pairs in each bin. This analysis found that the base pairs with intermediate pairing probability tend to fluctuate the most in probability as the parameter values are changed.

Additionally, the number of pairs in each pairing probability bin was plotted (Figure 3B). As expected, a large number of pairs had low pairing probability (close to zero). Starting at a pairing probability of about 20%, the bin population increased as the pair probabilities decreased. This is expected because the total number of possible pairs grows with the square of the sequence length, but the number of pairs with high probability can only grow linearly with sequence length. Thus, most of the possible base pairs must have low formation probability. At the same time, the bin population increases as the pairing probability increases above 90%.

Because the ratio of probable base pairs to low probability pairs (close to zero) is proportional to the reciprocal of the sequence length, it is difficult to compare base pair probability RMSDs between RNA families, which can vary in length. In fact, the slope of a linear fit of sensitivity defect as a function of pair probability RMSD for the different RNA families is dependent on mean family sequence length (Supplementary Figure S3). A corrected RMSD value can be calculated by limiting the mean to base pairs with relevant probability:

$$\mathrm{cRMSD} = \sqrt{\frac{\sum_{\text{Relevant BP}} (P_N - P_R)^2}{N_{\text{Relevant BP}}}}$$

Assuming that base pairs with relevant probability will dominate the total base pair probability and that the number of significant base pairs is the square root of the number of total potential base pairs, then the corrected RMSD reduces to:

$$\mathrm{cRMSD} = \sqrt{\frac{\sum_{\text{All BP}} (P_N - P_R)^2}{\sqrt{N_{\text{BP}}}}}$$

On average, the top $\sqrt{N_{\text{BP}}}$ potential base pairs account for ∼85% of the total base pair probability (Supplementary Figure S4). By correcting for the number of probable base pairs, the length dependence for the ratio of mean sensitivity defect and mean cRMSD is removed (Supplementary Figure S5).

DISCUSSION

This work provides several new insights into the prediction of RNA secondary structure. First, there are parameters that are crucial for high-quality base pair estimates, and these parameters should be the focus of additional experiments to improve the accuracy and precision of secondary structure prediction. Second, there are nearest neighbor parameters for which errors in the estimates have little impact on the precision of base pairing probability estimates (Figure 1). This means that these parameters do not need to be

precisely determined for robust structure prediction. For example, in a set of folding nearest neighbor parameters developed for modified chemistries, these parameters could be estimated based on fewer experiments than were used with RNA, without compromising the precision of base pairing probability estimates. Third, the imprecision of base pairing probability estimates varies across pair probabilities. High and low probability pairs are less prone to imprecision in the parameters (Figure 3).

From the sensitivity analysis using a flat σ value, a number of parameters were highlighted as being particularly important for predicting RNA secondary structure with high precision. As expected, the helical stacking parameters are important for precise RNA structure prediction. However, there are a number of other parameters that are observed to be just as important, such as multibranch loop parameters (indices 85–86 in Figure 1), the terminal AU/GU penalty (index 87 in Figure 1), bulge loop initiations (indices 94–95 in Figure 1) and the AU/GU closure of 1 × n internal loops (index 236 in Figure 1). Also important are the hairpin and internal loop initiation energies.

Other parameters appear to have little impact on the estimates of pair probabilities when perturbed. These include the parameters for specific triloops, tetraloops and hexaloops sequences, as well as many of the internal loop parameters. The parameters with least impact are those parameters that apply to specific sequences. For example, the tetraloop parameter tables contains the folding free energy change of 16 tetraloops that are known by experiment to be poorly predicted using the standard hairpin loop parameters. These tetraloops are 6 nt long (including the sequence of the closing base pair) and therefore a specific tetraloop stability would not apply in calculations for sequences in which the 6-mer motif is not found.

Figure 1C shows the tally for the parameter usages when calculating folding free energy changes for a stochastic sample. This is an estimate for the importance-weighted use of each parameter when calculating the partition function. The most frequently called parameters are those for coaxial stacking, helical stacking, AU/GU end penalties, multibranch loop initiation parameters and bulge loop initiations. Figure 2 plots the logarithm of the mean RMSD from the sensitivity analysis using flat errors as a function of the logarithm of the parameter usage for each parameter, clearly showing the correlation between the two. However, there are parameters whose effects on structure prediction are poorly predicted by parameter usage (Supplementary Figure S6). Examples include the parameters for bulge loop initiations, AU/GU end penalty, AU/GU closure of internal loops, internal loop asymmetry and the multibranch loop per helix penalty, which all have greater effects on structure prediction than other parameters with similar parameter usage counts. The bulge loop initiation parameters with the greatest impact are the loop initiation terms for bulge loops of two and 3 nt. The initiation parameters for bulge loops of 4–6 nt are linear extrapolations of those two parameters, while the initiation terms for bulge loops >6 are extrapolated from the initiation of 6 nt bulge loops using polymer theory (9). This means perturbations in terms for bulge loops of 2 and 3 nt are propagated for estimates of larger bulge loops. One effect is that

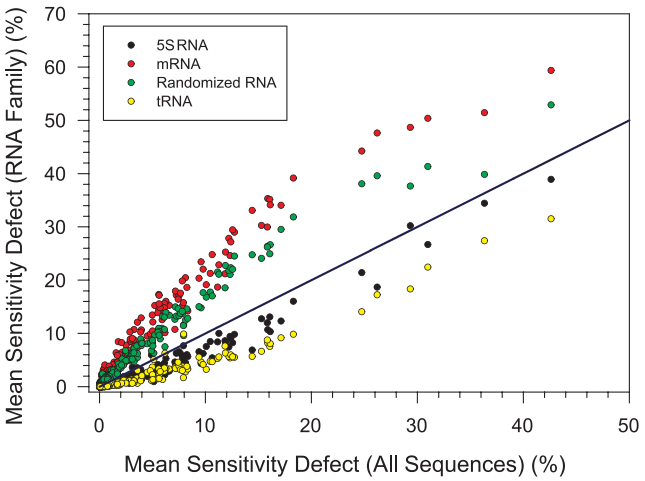

Figure 4. The sensitivity to parameter changes is family dependent. The scatter plots show the sensitivity defect from changing a parameter by +3 standard errors for specific RNA families as a function of the average for all sequences (where the average is the mean of the per family RMSDs). Therefore, this plot has one point per family for each of the independent parameters. If the sensitivity defect for a parameter for an individual RNA family is identical to the average across all families, it would fall on the diagonal line (shown in black). The mRNA and shuffled RNA sequences experience a greater sensitivity defect than the average (their points are generally above the diagonal line), while 5S rRNAs and tRNAs have a lower sensitivity defect than the average (the points generally fall below the line).

for some perturbed parameter sets, the slope of the extrapolation changes, making larger bulge loops more favorable than small bulge loops and this artifact explains why perturbation of bulge loop initiation parameters stand out. The AU/GU helix end penalty is a case where the thermodynamic model changed since the 2004 nearest neighbor rules. In the most recent parameter derivation (33), the GU helix end term is set to 0 in light of new data, indicating that the parameter is not being correctly applied in the 2004 parameter set used here. This might also hold for the parameter for AU/GU closure of internal loops. Similarly, it is known that the functional form that is used to calculate multibranch loop energies in the dynamic programming algorithm poorly models the measured experimental data (34,35). Therefore, parameters for which the mean RMSD is larger than expected for the number of uses of the parameters appear to identify parameters for loop nearest neighbor models that do not model folding stability as well as other loop models.

The impact of perturbing parameters depended on identity of the RNA family being analyzed. RNA families such as 5S rRNA and tRNA, were more resistant to changes in the parameters than the average for all sequences, while other RNA families such as mRNA and randomly shuffled sequences were more sensitive to parameter changes than the average for all sequences (Figure 4). However, it should be noted that other structured ncRNA, like 23S and 16S rRNAs, behaved similarly to mRNA and randomized RNAs (Interactive plot in the Excel file included in Supplementary Data), indicating that RNA structure was not the only factor that determined this response.

When the average change in base pair probability was plotted against the initial base pair probability calculated from unperturbed data tables, highly probable base pairs were found to be resistant to changes in a single thermodynamic parameter (Figure 3A). This suggests that the pairs predicted with greatest confidence in RNA secondary structure prediction are also robust to errors in estimates of parameters (24). Additionally, the low probability base pairs were also resistant to changes in pair probabilities with changes in a single parameter. One reason for this is that, as shown in Figure 3B, there is a large set of pairs that have little to no probability of forming. As parameters are perturbed, it is simply unlikely that change in a single parameter would dramatically increase the pairing probability for these unlikely pairs.

Another observation is that there is a general asymmetry of the effects of parameter deviations, with changes that make a parameter more stable tending to have a greater impact than changes that make the parameter less stable (Supplementary Figure S7). For example, the helical stacking parameters have 40% greater impact on RMSD when perturbed by −3 standard errors than by +3 standard errors. Additionally a plot of the difference between the average probability changes between −3 and +3 standard error changes as a function of initial base pair probability shows that the −3 standard error parameter sets disproportionally affect the base pair probabilities of more likely base pairs (Supplementary Figure S8).

An important use of this sensitivity analysis is to provide focus on the parameters that need the most immediate attention of additional experiments to refine the parameters or improve the underlying models. The parameters that stand out as most in need of additional attention are bulge loop initiations, multibranch loop parameters and the helical stacking parameters with GU base pairs. Interestingly, some of these categories have already been addressed by additional experiments performed after the parameters were last assembled in 2004. For example, bulge loops and GU pairs have been studied with additional optical melting experiments (33,36–39). The results here show the importance of integrating the new data into the complete set of nearest neighbor parameters in current use.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dave Mauger, Joe Cabral, John Reynders and Vlad Presnyak for helpful discussions. The authors also thank Christine Heitsch for suggesting the plot in Figure 3A and for suggesting the form it would take.

FUNDING

Moderna Therapeutics and the National Institutes of Health Grant [R01GM076485 to D.H.M.]. Funding for the open access charge: National Institutes of Health [R01GM076485 to D.H.M.].
Conflict of interest statement. None declared.

REFERENCES

1. Wu,L. and Belasco,J.G. (2008) Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell*, **29**, 1–7.
2. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
3. Serganov,A. and Nudler,E. (2013) A decade of riboswitches. *Cell*, **152**, 17–24.
4. Yu,Y.T. and Meier,U.T. (2014) RNA-guided isomerization of uridine to pseudouridine–pseudouridylation. *RNA Biol.*, **11**, 1483–1494.
5. Tinoco,I. Jr and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
6. Woodson,S.A. (2000) Recent insights on RNA folding mechanisms from catalytic RNA. *Cell Mol. Life Sci.*, **57**, 796–808.
7. Crothers,D.M., Cole,P.E., Hilbers,C.W. and Schulman,R.G. (1974) The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J. Mol. Biol.*, **87**, 63–88.
8. Onoa,B. and Tinoco,I. Jr (2004) RNA folding and unfolding. *Curr. Opin. Struct. Biol.*, **14**, 374–379.
9. Mathews,D.H., Disney,M.D., Childs,J.L., Schroeder,S.J., Zuker,M. and Turner,D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 7287–7292.
10. Andronescu,M., Condon,A., Turner,D.H. and Mathews,D.H. (2014) The determination of RNA folding nearest neighbor parameters. *Methods Mol. Biol.*, **1097**, 45–70.
11. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.
12. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
13. Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
14. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
15. Rivas,E., Lang,R. and Eddy,S.R. (2012) A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, **18**, 193–212.
16. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.
17. Turner,D.H. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.
18. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
19. Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 7706–7710.
20. Doshi,K.J., Cannone,J.J., Cobaugh,C.W. and Gutell,R.R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.
21. Le,S.Y., Chen,J.H. and Maizel,J.V. Jr (1993) Prediction of alternative RNA secondary structures based on fluctuating thermodynamic parameters. *Nucleic Acids Res.*, **21**, 2173–2178.
22. Layton,D.M. and Bundschuh,R. (2005) A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res.*, **33**, 519–524.
23. Hower,V. and Heitsch,C.E. (2011) Parametric analysis of RNA branching configurations. *Bull. Math. Biol.*, **73**, 754–776.
24. Mathews,D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178–1190.
25. Bellaousov,S. and Mathews,D.H. (2010) ProbKnot: fast prediction of rna secondary structure including pseudoknots. *RNA*, **16**, 1870–1880.
26. Xia,T., SantaLucia,J. Jr, Burkard,M.E., Kierzek,R., Schroeder,S.J., Jiao,X., Cox,C. and Turner,D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA

duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719–14735.
27. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
28. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192.
29. Bevington,P.R. and Robinson,D.K. (2003) *Data Reduction and Error Analysis for the Physical Sciences*. 3rd edn. McGraw-Hill, Boston.
30. Woodson,S.A. and Crothers,D.M. (1987) Proton nuclear magnetic resonance studies on bulge-containing DNA oligonucleotides from a mutational hot-spot sequence. *Biochemistry*, **26**, 904–912.
31. Znosko,B.M., Silvestri,S.B., Volkman,H., Boswell,B. and Serra,M.J. (2002) Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry*, **41**, 10406–10417.
32. Ding,Y. and Lawrence,C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, **31**, 7280–7301.
33. Chen,J.L., Dishler,A.L., Kennedy,S.D., Yildirim,I., Liu,B., Turner,D.H. and Serra,M.J. (2012) Testing the nearest neighbor model for canonical RNA base pairs: revision of GU parameters. *Biochemistry*, **51**, 3508–3522.
34. Mathews,D.H. and Turner,D.H. (2002) Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, **41**, 869–880.
35. Liu,B., Diamond,J.M., Mathews,D.H. and Turner,D.H. (2011) Fluorescence competition and optical melting measurements of RNA three-way multibranch loops provide a revised model for thermodynamic parameters. *Biochemistry*, **50**, 640–653.
36. Tomcho,J.C., Tillman,M.R. and Znosko,B.M. (2015) Improved model for predicting the free energy contribution of dinucleotide bulges to RNA duplex stability. *Biochemistry*, **54**, 5290–5296.
37. Murray,M.H., Hard,J.A. and Znosko,B.M. (2014) Improved model to predict the free energy contribution of trinucleotide bulges to RNA duplex stability. *Biochemistry*, **53**, 3502–3508.
38. Nguyen,M.T. and Schroeder,S.J. (2010) Consecutive terminal GU pairs stabilize RNA helices. *Biochemistry*, **49**, 10574–10581.
39. Strom,S., Shiskova,E., Hahm,Y. and Grover,N. (2015) Thermodynamic examination of 1- to 5-nt purine bulge loops in RNA and DNA constructs. *RNA*, **21**, 1313–1322.