Secondary structure prediction for RNA sequences including N6-methyladenosine

Elzbieta Kierzek, Xiaoju Zhang, Richard M. Watson, Scott D. Kennedy, Marta Szabat, Ryszard Kierzek & David H. Mathews

There is increasing interest in the roles of covalently modified nucleotides in RNA. There has been, however, an inability to account for modifications in secondary structure prediction because of a lack of software and thermodynamic parameters. We report the solution for these issues for N6-methyladenosine (m6A), allowing secondary structure prediction for an alphabet of A, C, G, U, and m6A. The RNAstructure software now works with user-defined nucleotide alphabets of any size. We also report a set of nearest neighbor parameters for helices and loops containing m6A, using experiments. Interestingly, N6-methylation decreases folding stability for adenosines in the middle of a helix, has little effect on folding stability for adenosines at the ends of helices, and increases folding stability for unpaired adenosines stacked on a helix. We demonstrate predictions for an N6-methylation-activated protein recognition site from MALAT1 and human transcriptome-wide effects of N6-methylation on the probability of adenosine being buried in a helix.

https://doi.org/10.1038/s41467-022-28817-4

1 Institute of Bioorganic Chemistry Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland. 2 Department of Biochemistry and Biophysics and Center for RNA Biology, School of Medicine and Dentistry, University of Rochester, 601 Elmwood Avenue, Box 712, Rochester, NY 14642, USA.

✉email: Elzbieta.Kierzek@ibch.poznan.pl; David_Mathews@urmc.rochester.edu
It has long been appreciated that covalent modification of RNA is used by nature to expand the chemical repertoire of the four common nucleotides. tRNAs, in particular, are known to have prevalent modifications, and the roles of some of these have been elucidated. For mRNAs and long non-coding RNAs (lncRNAs), it had been harder to identify sites of modification until recently when methods were developed using next generation sequencing technologies to identify modifications. Modifications including deamination to inosine, pseudouridylation, 5-methylation of cytosine, and N6-methylation of adenosine can now be localized transcriptome-wide.

N6-methyladenosine (m6A) is considered the most prevalent modification in mRNA, and m6A is also widespread in lncRNAs. It is known to have writers that apply the modifications to specific positions (methyltransferases including METTL3 and METTL14), readers that identify sequences with N6-methylation (RNA-binding proteins including YTHDF2 and the YTH family), and erasers that can remove the modification (demethylases including FTO and ALKBH5), restoring the base to adenine. Furthermore, there are hundreds of sites for which the m6A modification is known to cause structural switches and to expose protein binding sites that are otherwise not available for binding. Additionally, m6A can regulate splicing. RNA secondary structure prediction is in widespread use to help determine function-structure relationships, but has not been generally available for understanding the roles of covalent modifications. For unmodified sequences, secondary structure prediction has been used to identify microRNA binding sites, design siRNAs, identify protein binding sites, and discover functional RNA structures. These types of calculations have not been able to account for modifications without extensive user intervention because a set of nearest neighbor parameters are needed for estimating the folding stability of structures that include modifications. A number of studies demonstrated an impact on folding stability by modifications, but no complete set of parameters has been available for RNA folding, as there are for RNA folding with the four prevalent bases. The same time, no software has been available for handling a larger alphabet of sequences containing modifications. This led to a chicken-and-egg problem; without software, there was no impetus to assemble parameters and without parameters there was no reason to write the software.

In this work, we developed a full set of nearest neighbor parameters for a folding alphabet of m6A, A, C, G, and U nucleotides. These parameters account for helix and loop formation, and they are based on optical melting experiments for 32 helices with m6A-U base pairs and 13 oligonucleotides with m6A in loop motifs. We also modified the RNAstructure software package to accept user-defined folding alphabets and to read and utilize thermodynamic parameters for these extended alphabets. Together, these advances allow the prediction of RNA secondary structures for sequences with m6A. We demonstrate, for calculations with human mRNA sequences known to contain m6A, that N6-methylation alters the folding landscape so that m6A is less likely to be buried in a helix, i.e., stacked between two base pairs. We also provide a model for the change in protein binding affinity caused by N6-methylation of the lncRNA metastasis-associated lung adenocarcinoma transcript (MALAT1).

Results

Overview of methods. Secondary structure prediction for RNA sequences including m6A requires both a set of nearest neighbor folding parameters and software capable of using the set of parameters. An overview of the methods is illustrated in Fig. 1. RNA secondary structure prediction requires both parameters for evaluating folding stability and a search algorithm to identify the optimal structure given the parameters. In our RNAstructure software, we use nearest neighbor parameters to estimate folding free energy change and a set of dynamic programming algorithms that predict optimal structures.

We built a database of optical melting experiments of oligonucleotides including m6A and then used linear regression to fit nearest neighbor parameters. We also extended the functionality of RNAstructure to recognize modified nucleotides in sequences and to use parameters for sequence alphabets beyond the four common nucleotides. The m6A modification parameters take advantage of this additional feature.
Fig. 2 A comparison of base pairing stability for m\(^6\)A to A. A The nearest neighbor parameters for helix stacks. The position of the m\(^6\)A is indicated by 6. The stacking parameters, determined by linear regression, are compared for methylated (blue; i.e., m\(^6\)A-U base pairs) and unmethylated (red; i.e., A-U base pairs) sequences for analogous nearest neighbors. The unmethylated stacks (i.e., A-U base pairs) are those of Xia et al.\(^{44}\) for adjacent Watson–Crick pairs and those of Chen et al.\(^{79}\) for adjacent G-U pairs. Stacks with m\(^6\)A-U pairs are generally less stabilizing than analogous stacks with A-U pairs. Uncertainty estimates are the standard errors of the regression. B Terminal m\(^6\)A-U pairs are not destabilizing. Plotted are the duplex stabilities as folding free energy change from the linear fit to the \(T_M^{-1}\) vs. ln(Cr/a) plots of the optical melting data. The top two sequences (Watson–Crick paired with a complementary strand) have the same nearest neighbor stacks, but the second helix has two terminal A-U pairs\(^{44}\). This costs 0.7 kcal/mol of stability. The bottom two sequences also have the same nearest neighbor stacks, but the second has two terminal m\(^6\)A-U pairs. Here the stability cost is 0.18 kcal/mol and not outside of the uncertainty estimate, which is approximated as 4% of the total free energy change\(^{44}\). On average, terminal A-U pairs cost 0.45 kcal/mol of stability\(^{46}\), but terminal m\(^6\)A-U pairs are not destabilizing.

Linear regression was used to fit the nearest neighbor parameters for folding free energy change. Figure 2A shows the increments in comparison to the same stack with A-U pairs and Supplementary Table S2 provides the values. The free energy changes range from \(-1.79\pm0.25\) kcal/mol to \(+1.45\pm0.57\) kcal/mol. As expected based on prior optical melting experiments for duplexes with m\(^6\)A-U pairs\(^{34,36}\), nearest neighbor stacks for methylated A-U pairs are less stable than stacks for unmethylated A-U pairs. On average, the stacks with m\(^6\)A-U pairs are 0.4 kcal/mol less stable per methylation. There are exceptions, however; an m\(^6\)A-U pair followed by a U-A pair is as stable as an A-U pair followed by a U-A pair (\(-1.10\) kcal/mol). The most unstable stack has two m\(^6\)A-U pairs. Like A-U pairs, when the m\(^6\)A-U pair is adjacent a G-C it is more stable than when adjacent to A-U. Also like A-U pairs, m\(^6\)A-U pairs adjacent to G-U are less stable than those adjacent to A-U pairs.

An unexpected feature of terminal m\(^6\)A-U pairs is that they require no terminal penalty, although terminal A-U pairs receive a +0.45 ± 0.04 kcal/mol penalty per A-U pair at the end of a helix\(^{44}\). Two findings support this. First, when a terminal parameter is included in the linear regression fit, the value is +0.13 ± 0.17 kcal/mol, which is not significantly different from 0 kcal/mol. Second, our dataset includes two helices with the same nearest neighbor stacks, but with different helix ends (Fig. 2B). Previously, it was noted that this pair of helices, when unmethylated, had markedly different stability (0.70 ± 0.47 kcal/mol), with the helix with A-U ends less stable\(^{44}\). For the methylated helices, the difference is small (0.18 ± 0.45 kcal/mol). This demonstrates that a terminal m\(^6\)A-U base pair has overall similar stability to a terminal A-U base pair because a terminal A-U pair has a more favorable stack but requires the terminal A-U penalty.

Fig. 3 M\(^6\)A stacking on a helix end stabilizes secondary structure as compared to A stacking. The \(\Delta G°_{37}\) (kcal/mol) for dangling ends and terminal mismatches as a result of N\(^6\)-methylation (Supplementary Table S4) is shown, where negative values mean greater folding stability for m\(^6\)A than A. The motifs shown here have a terminal base pair (left side of motif), and either a dangling end or terminal mismatch right (right side of motif). On average, the methylated motifs are more stabilizing than the unmethylated motifs, although the extent of the stabilization is sequence dependent. Uncertainty estimates are propagated from the uncertainty from the individual optical melting experiments (see “Error propagation” in the “Methods”).

Loop nearest neighbor parameters for m\(^6\)A. For secondary structure prediction, parameters need to also be extrapolated for loop formation. The stability of a 3′ dangling end or as a component in a terminal mismatch stabilizes secondary structure formation to a greater extent than an analogous A. On average, the m\(^6\)A dangling end is \(-0.43\pm0.15\) kcal/mol more stable than the analogous A dangling end for the 3′ and 5′ dangling ends studied here. Terminal mismatches for m\(^6\)A-m\(^6\)A, G-m\(^6\)A, m\(^6\)A-G, and m\(^6\)A-C on Watson–Crick or G-U terminal pairs are on average \(-0.28\pm0.26\) kcal/mol more stabilizing than the analogous A-A, G-A, A-G, or A-C terminal mismatches. This stabilizing effect is sequence dependent; the \(\Delta G°_{37}\) ranges from \(-0.74\) kcal/mol (G-m\(^6\)A mismatch on a U-G pair) to +0.02 kcal/mol (G-m\(^6\)A mismatch on a U-G pair) to +0.43 kcal/mol (G-m\(^6\)A mismatch on a U-G pair).
mismatch on an A-U pair). An m6A-m6A mismatch on an m6A-U pair is more stable than the m6A-m6A mismatch on an A-U pair by −0.42 ± 0.40 kcal/mol.

The hairpin loop structure with m6A is marginally less stable than the analogous hairpin loop with A (ΔΔG°37 = 0.23 ± 0.24 kcal/mol; Supplementary Table S4). The 2 × 2 internal loop with tandem G-m6A pairs is also marginally less stable than the analogous loop with tandem G-A pairs (ΔΔG°37 = 0.33 ± 0.53 kcal/mol; Supplementary Table S4). Both stability changes are within the uncertainty estimates, suggesting that they are not significant differences.

Additional experiments to test the parameters. To test our parameters, we performed additional melts of duplexes (Supplementary Table S5A). The first is a duplex with all base pairs, incorporating a consensus N6-methylation site, GGCACU, where we determined the helix stability with and without methylation. The second is an additional Y dangling m6A to test our assumption that dangling m6A are stabilized by −0.3 kcal/mol compared to dangling A. Then we studied A bulges containing bulge loops, with and without methylation, four of which are loops that were previously studied45 and one of which models the m6A site in the hepatitis C viral genome46. In these bulge loops, m6A is the bulged nucleotide, in the base pair closing the single bulge, or in a base pair adjacent to the pair closing the bulge loop. Supplementary Table S5A provides the stability determined by optical melting and Supplementary Table S6A shows how well the stabilities are estimated with our nearest neighbor parameters.

We conclude from these tests that the nearest neighbor parameters are accurate enough to be used for RNA secondary structure prediction42,43. With the exception of one duplex with m6A bulged, the estimates for the duplex stabilities are within the uncertainties propagated for the experiment and the nearest neighbor parameters (ΔΔG°37, column of Supplementary Table S6A). The unmethylated consensus duplex (GGACUA-GUCC2) is estimated by nearest neighbor parameters to be more stable (by −0.48 ± 0.73 kcal/mol) than it is by experiment. The methylated consensus duplex is estimated by nearest neighbors to be less stable than it is (by 0.85 ± 0.97 kcal/mol). These deviations are 2.9 and 5.5% of the experimentally determined values. The estimated stability of the duplex with the dangling m6A closely matches the experimental value (ΔΔG°37 of 0.01 ± 0.84 kcal/mol).

The bulge loops closed by m6A-U pairs are closely modeled by the nearest neighbor parameters with ΔΔG°37 of 0.05 ± 0.84, 0.07 ± 0.82, and 0.39 ± 0.87 kcal/mol. For the two duplexes with bulged m6A, the bulges are systematically less stable than the analogous bulged A duplex. The nearest neighbor rules do not consider the sequence identity of single bulged nucleotides, and therefore modeling this is outside the current functional form33. In one case, the stability of the bulged A is more accurately modeled by the sequence-independent stability and in one case, the stability of the bulged m6A is more accurately modeled. This points to a limitation in the nearest neighbor parameters, which has previously been documented for single nucleotide bulge loops47,48. In the worst case, the nearest neighbor rules are incorrect by 1.12 kcal/mol, which is only 9.4% of the total folding free energy change of the duplex. For the duplex with an m6A-U base pair adjacent to a G-C pair that closes a bulge loop, the agreement is also excellent, with ΔΔG°37 of 0.01 ± 0.94 kcal/mol.

Our optical melting experiments to derive the nearest neighbor parameters were performed in 1 M Na+ to be consistent with the experiments used to derive the Turner nearest neighbor parameters33. To test whether m6A is stabilized by Mg2+, we also performed experiments in 150 mM K+ and 5 mM Mg2+, chosen to mimic physiological salt conditions for monovalent and divalent cations49. Supplementary Table SSB shows the optical melting data for the Mg2+-containing buffer for the consensus duplex and for the eight bulge loop sequences.

We find similar folding stabilities between 1 M Na+ and the Mg2+-containing buffer (Supplementary Table S6B). The largest difference in stability was observed for the consensus methylation site duplex at −1.79 ± 0.84 kcal/mol more stability in 1 M Na+, a difference of 12.3% between the two buffers. The bulge loop-containing duplexes were found to have no systematic bias in stability between methylated and unmethylated sequences. Prior tests of optical melting experiments in similar Mg2+-containing buffers generally demonstrated similar folding stabilities as 1 M Na+50-55, with the Loop E motif in 0.1 M Na+ and 50 mM Mg2+ as a notable exception56. A study of hairpin stem loops with m6A showed a magnesium dependence on relative folding stability with and without methylation, using a relatively low monovalent salt of 25 mM NaCl and 15 mM sodium phosphate45. In the higher monovalent K+ we used, which is similar to intracellular conditions, this methylation-dependent change in stability is not observed.

RNAstructure software modifications. To predict RNA secondary structures for sequences with A, C, G, U, and m6A, we modified the command line programs in the RNAstructure software package to accept extended alphabets of nucleotides40. By default, the software interprets sequences as standard RNA, but a command line switch can specify an alternative alphabet. For example, the nearest neighbor parameters for a DNA alphabet composed of A, C, G, and T has long been available. Now, because of this work, the nearest neighbor parameters for an m6A alphabet are available.

The key to an extended alphabet is the specification of the nucleotides and base pairs (Supplementary Fig. S2). A common architecture across the RNAstructure programs means that the command line programs are capable of using the extended alphabets, which can include any number of characters. This includes the prediction of minimum free energy structures, base pair probabilities, maximum expected accuracy structures, and folding stability for structures. Each nucleotide must be encoded by a single character, and we chose “6” or “M” as the characters to encode m6A in sequences and in the m6A nearest neighbor parameter tables. The Methods section details our estimates for the m6A nearest neighbor parameters. The parameter tables are provided as Supplementary Materials.

Nearest neighbor parameter tables are read from disk as programs start. Each parameter table requires additional rows and columns to provide the nearest neighbor parameters values for those nucleotides, although the dimensionality of the tables stays the same. For example, a base pair stack table is four-dimensional because the sequence of four positions is required to estimate the stacking stability of two base pairs. When m6A is included with RNA, the size of each dimension is increased from four. The largest table is the 2 × 2 internal loop lookup table33, which is eight dimensional because it includes the sequence of the two closing base pairs.

Modeling the accessibility changes in MALAT1 as a result of methylation. It has been established that N6-methylation can alter protein binding accessibility. To test our m6A nearest neighbor parameters and software, we made a quantitative prediction for the accessibility of protein binding to a structure in the lncRNA MALAT1 that is the binding target of heterogeneous nuclear ribonucleoprotein C (HNRNPC). This has been characterized by Tao Pan and co-workers in an in vitro system with a single stem-loop structure57. Filter binding experiments demonstrated that the methylated RNA is more accessible to protein
binding than the unmethylated RNA. Additionally, enzymatic cleavage by RNase S1, which has specificity for loop regions of RNA, demonstrated increased cleavage 5' and 3' to the methylated A, supporting increased accessibility.

We used RNAstructure to predict the lowest free energy structure of the 32 nucleotide RNA. We predicted the structure to be a hairpin stem-loop as previously modeled for the unmethylated sequence (Fig. 4A, closed structure)\(^\text{57}\). However, our structure prediction was unchanged for the methylated sequence. For HNRNPC to bind, the site composed of the five Us starting at position 10 must be unpaired\(^\text{58}\). We can therefore estimate the accessibility as the \(\Delta G_{37}\) for the breaking of the three base pairs adjacent to the hairpin loop (Fig. 4A, open structure)\(^\text{59}\). The \(\Delta G_{37}\) is the difference in the accessibility with and without the N\(^6\)-methylation at position 22 is 0.6 ± 0.4 kcal/mol (where the uncertainty is propagated from the uncertainties of two nearest neighbor stacks that change from the introduction of the methylation). This is consistent with the \(\Delta G_{37}\) of 1.26 ± 0.22 kcal/mol determined by filter binding assay\(^\text{57}\).

Our model therefore is that the N\(^6\)-methylation of the MALAT1 hairpin does not open the binding site, but instead facilitates the HNRNPC protein-mediated breaking of base pairs at the binding site. To test this, we probed the 32 nucleotide RNA (extended with a 3′ structural cassette) by chemical mapping with CMCT (1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide methyl-p-toluenesulfonate), DMS (dimethyl sulfate), and kethoxal (Supplementary Fig. S3). The chemical mapping data and the prior enzymatic mapping\(^\text{57}\) data are consistent with the proposed structure, and only minor differences are observed with and without the methylation. We also collected NMR spectra with and without methylation and assigned the imino resonances, following previous NMR studies on the same sequence\(^\text{60}\). The NOESY walk (Supplementary Fig. S4) confirms the hairpin conformation, and Supplementary Fig. S5 shows that the number of NMR peaks are observed at high RNA concentration because of duplex formation. Supplementary Fig. S6 also shows that the imino resonances in the chemical shifts are largely unchanged in a higher Na\(^+\) buffer (95 mM NaCl and 38 mM sodium phosphate) as a result of N\(^6\). methylation, although chemical shift changes close to the methylation site are observed because of the proximity of the methyl group (Supplementary Table S7). Additionally, we tested whether Mg\(^{2+}\) alters the conformation of the N\(^6\)-methylated sequence differently than the unmethylated sequence. The Mg\(^{2+}\) titration in Supplementary Fig. S7 shows that the imino resonances are similarly affected for the two sequences. The U10 imino resonance shifts by ~0.2 ppm in both sequences, and this was resolved by NOESY spectra (Supplementary Fig. S8). Taken together, the modeling and experimental data show that the MALAT1 RNA conformation is not directly switched by N\(^6\). methylation. Instead, the methylation poises the RNA to be more accessible to HNRNPC binding.

Transcriptome-wide predictions with m\(^6\)A. To further test our m\(^6\)A nearest neighbor parameters and software, we predicted structures for 18,026 mRNAs that were identified as having N\(^6\)A methylation by whole transcriptome sequencing\(^\text{61}\) and for which PARS structure mapping data are available\(^\text{62}\). We used the nearest neighbor parameters and RNAstructure package to estimate the probability that the methylation site is buried in a helix, i.e., in a base pair stacked between two other base pairs, for both the unmethylated and methylated sequence (Fig. 4B). We used 800 nucleotide fragments of local sequence to estimate the pairing probability because we previously found that pairing probability estimates for 800 nucleotide fragments closely match those for global secondary structure prediction\(^\text{28}\). This is a reasonable balance between accuracy and total calculation time.

We find that the unmethylated A at the methylation site is less likely to be buried in a helix than adjacent nucleotides (Fig. 4B). This is intuitive because adjacent nucleotides at the consensus site are often G or C, and A is more predominant in loops in RNAs with known structure. There is a substantial shift in the
probability of m6A being buried in a helix relative to A (21% for A and 13% for m6A). This suggests there could be widespread structural switching being affected by N6-methylation. We can also compare our results to PARS data for the same sequences (Fig. 4C)36,62. A PARS score quantifies the enzymatic cleavage estimate of local pairing and the experiment is performed transcriptome wide. A lower PARS score indicates greater nuclease S1 cleavage relative to nuclease V1 and thus a greater extent of unpairing because nuclease S1 has specificity for loops and nuclease V1 has specificity for helices63. The PARS scores at the methylation site also demonstrate a propensity to be unpaired at the methylation site, but the minimum average PARS score is at the nucleotide 5′ to the m6A site. A possible explanation for the discrepancy is that PARS attributes S1 cleavages to the base 5′ to the cleavage site, assuming that the base 5′ to the cleavage is unpaired. Cleavage can also occur when the base 3′ to the cleavage site is unpaired and therefore the PARS scores 5′ to the methylation site might be overestimating the propensity of being unpaired, in that some of the propensity of being unpaired should be attributed to the methylation site. For example, the prior S1 mapping of 5S rRNA structure is consistent with cleavages both 5′ and 3′ to unpaired nucleotides (Supplementary Fig. S9)64. Prior analysis of PARS scores for methylation sites also concluded that the data indicate the m6A is positioned in structures at the transition between base paired regions and loop regions, consistent with our structure prediction estimates36.

Discussion

Here we provide the complete nearest neighbor model for a folding alphabet including modified nucleotides. Because m6A is considered the most abundant modification in mRNA and is known to affect folding stability, we chose m6A as the modification to study. The full nearest neighbor model for secondary structure prediction requires both helical stack parameters and also loop parameters. We know from a sensitivity analysis of secondary structure prediction that, for loops, accurate parameters are most important for dangling ends and terminal mismatches42,43 accordingly we focused our experimental effort on these motifs. We also observed marginal differences in stability for hairpin and internal loops containing m6A as compared to the same sequences without the N6-methylation. Subsequent studies could be focused on understanding and modeling folding stability differences for loops with m6A.

The other component of this study was advancing RNA-structure to work with sequences with nucleotides beyond A, C, G, and U. We provide command line tools that are ready to make quantitative predictions of structure and folding stability for sequences with m6A. The software is available for free download under the GNU GPL 2 at https://rna.urmc.rochester.edu/RNAsstructure.html. Given the software, we plan expand our work in the future to include alphabets with inosine and pseudouridine. Both have helical nearest neighbor parameters available for stacks on Watson–Crick pairs37,38, and both could be extended to full nearest neighbor parameters sets with additional optical melting experiments. The updated software also makes it possible to take advantage of engineered65,66 or natural67 modifications in designs of sequences for nanostructures or catalysts and a wide range of applications in biomedicine and biotechnology68.

The two loops studied here with N6-methylations both had marginally less folding stability than the analogous unmethylated loops (Supplementary Table S4). Solution structures are available for each of the A-containing loops, and these structures provide clues as to why the stabilities would be only marginally changed by methylation. The hairpin loop, GGCGUAUAUGCC, has the first A in the loop (A6; the site of our m6A-methylation) stacked at the apex of the loop on the adjacent A (A7)69. Because A6 is not hydrogen bonding in the structure, a methylation at N6 can be accommodated in the preferred syn orientation by the structure without change70. For the internal loop with tandem G-A pairs, the pairs are trans-sugar-Hoogsteen pairs, i.e., the N6 position of the A is hydrogen bonded with the G at the N3 position71. For each methylated A, one hydrogen of N6 is available to form this hydrogen bond, placing the methyl in the preferred syn orientation72. However, the second hydrogen of A N6 is close to O′ of the G (ranging from 2.34 to 3.36 Å in the 15 deposited NMR models). A crystal structure with this motif shows that N6 methylation disrupts pairing, with the A N6 hydrogen bonding to the O′ of the opposite G72. Interestingly, this loss of pairing cost only 0.33 kcal/mol of folding stability, suggesting that conformational rearrangements do not necessarily result in a large loss of folding stability.

Recent studies demonstrated the ability of computational methods to estimate folding free energy changes73–78. In this work, we performed optical melting experiments to determine the folding stabilities of small model systems with m6A and fit nearest neighbor parameters to these data. Future work, however, could rely on computation or a mixture of computation and experimentation. Hopfinger et al., for example, estimated helical stacking nearest neighbor parameters for the eight stacks with m6A-U pairs adjacent to Watson–Crick pairs53. Overall the agreement of their estimates against our experimental values is excellent, with a root mean squared deviation of 0.30 kcal/mol. The largest single deviation is for a U-m6A pair followed by a G-C pair, where their estimate overstabilized the stack by 0.6 kcal/mol (Supplementary Fig. S10). Loop folding stabilities continue to be more of a challenge to estimate using computational methods because the conformational flexibility requires extensive sampling76.

With this work, we demonstrate the position of m6A in a structure determines whether folding stability is increased, decreased, or unchanged relative to the same structure with A. It was previously known that N6-methylation of an A-U pair in the middle of a helix would decrease the helix folding stability34,36. Our stacking parameters now quantify this sequence-dependent change (Fig. 2A). It was also previously known that m6A stacking on the end of a helix would stabilize the helix more than an analogous A. In this work, we also discovered that an m6A-U base pair at the terminal position of a helix provides roughly the same folding stability as an analogous A-U base pair. This is because terminal A-U pairs destabilize helices with a penalty of +0.45 kcal/mol44 that is not needed for terminal m6A-U base pairs (Fig. 2B). Recently, it was also discovered that terminal G-U base pairs in helices do not need an end penalty79. These results, taken together, show why N6-methylation is a potent switch of secondary structure.

Our transcriptome-wide calculations also suggest that structure switches from N6-methylation might be widespread (Fig. 4B). It will be interesting to perform similar calculations with other widespread covalent modifications, such as inosine. There is potential to identify structural mechanisms by which covalent modifications exert changes in protein binding, transcript stability, or gene expression.

Methods

Synthesis of oligonucleotides with m6A. Oligonucleotides were synthesized on a BioAutomation MerMade12 DNA/RNA synthesizer using β-cyanoethyl phosphoramidite chemistry and commercially available RNA phosphoramidites (ChemGenes, GenePharma) and protected N8-methyladenosine phosphoramidite.
which was synthesized according to a standard protocol. Synthesis of N6-methyladenosine was conducted via Dimroth rearrangement followed by protection of the 5'-hydroxyl with dimethoxytrityl and 2'-hydroxyyl with tert-butyldimethylsilyl. Next, 5'- and 2'-protected N6-methyladenosine was treated with 2-cyanoethyl N,N,N',N'-tetraisopropylphosphorodiamidite. Oligoribonucleotides were deprotected with aqueous ammonia/ethanol (3/1 v/v) for 16 h at 55 °C. Silylex-protecting groups were cleaved by treatment triethylamine trithylfluoride. Deprotected oligoribonucleotides were purified by silica gel thin layer chromatography in 1-propanol/aqueous ammonia/water (55/35/10 v/v/v). Optical melting data. The thermodynamic measurements were performed for nine various concentrations of RNA duplex in the range of 0.1 mM to 1 μM in buffer containing 1 M sodium chloride, 20 mM sodium cacodylate, and 0.5 mM Na2EDTA, pH 7. A subset of additional optical melting experiments (Supplementary Table S5b) was performed in a buffer containing 150 mM KCl, 20 mM sodium cacodylate acid, and 5 mM MgCl2. The temperature was increased from 0 to 90 °C on JASCO V-650 spectrophotometer with a thermostatically controlled melting medium. The melting curves were analyzed and the thermodynamic parameters were calculated from a two-state model of the program MeltWin 3.552. For most model RNAs, the ΔHΔG derived from Tm−1 vs. ln(ΔC/i)(a) plots, where a is 4 for non-self-complementary and a is 1 for self-complementary duplexes, is within 15% of that derived from averaging the fits to individual melting curves, as expected if both the two-state model is reasonable. Figure S11 shows the UV absorbance as a function of temperature and the Tm−1 vs. ln(C0/a) plots for 5′CCGGU/Gm6AUCG2 in a buffer with 1 M NaCl, 5′GGGACGm6AUCG3′ CCGGUGAG in a buffer with 1 M NaCl, and 5′GGGACGm6AUCG3′ CCGCUGAG in a buffer with 150 mM KCl and 5 mM MgCl2.

Linear regression. Linear least-squares fitting to determine RNA stacking stabilities was performed with a custom Python program using the statsmodels ordinary least-squares class. For each duplex, to determine the stabilities to be fit, the fixed terms were subtracted, including the stability of base pair stacks with Watson–Crick and G–U pairs only, the duplex initiation term, the terminal A–U penalty term (when needed), and the symmetry term (when needed). The free energy changes from the Tm−1 vs. ln(C0/a) fits of the optical melting data were used. The fit was excellent, with coefficient of determination, R2, of 0.984. Uncertainty estimates (Fig. 2A and Supplementary Table S2) are the standard errors of the regression. Supplementary Table S8 shows the stability to be fit and the estimate of the fit. Supplementary Table S9 shows the number of occurrences of each stacking parameter in the set of helices. The Python code for fitting nearest neighbor parameters is provided as Supplementary Software 1.

Loop motif stability calculations. Loop motif stabilities (Supplementary Table S4) are calculated by subtracting the helical component of stability, the free energy changes from the Tm−1 vs. ln(C0/a) fits of the optical melting data were used. For the dangling ends and terminal mismatches, twice the stability increment of the motif is determined by subtracting a reference helix stability from the stability of the duplex with the motif:

\[ 2 \times \Delta G_{\text{loop, motif}} = \Delta G_{\text{duplex}} - \Delta G_{\text{reference duplexes without the motif}} \]

The factor of two is present because the self-complementary duplexes have two instances of the motif.

For the hairpin loop, the stability of the loop motif is determined by subtracting the stability of the helical stacks (estimated with nearest neighbor parameters) from the total stability:

\[ \Delta G_{37 \text{hairpin loop}} = \Delta G_{37 \text{stem}} - \Delta G_{37 \text{reference helical stacks}} \]

The total helical stack stability is reported as the Reference \( \Delta G_{37} \) in Supplementary Table S4.

For the internal loop, the stability is the total stability of the duplex minus the helical stacks and helix initiation (estimated with nearest neighbor parameters) and minus the stability cost of symmetry (because the duplex is self-complementary):

\[ \Delta G_{37 \text{internal loop}} = \Delta G_{37 \text{duplex}} - \Delta G_{37 \text{helical stacks}} - \Delta G_{37 \text{initiation}} - \Delta G_{37 \text{symmetry}} \]

The Reference \( \Delta G_{37} \) reported in Supplementary Table S4 is the sum of the helical stacks and symmetry free energy increments.

Error propagation. To estimate uncertainties in free energies (δ), we propagate uncertainty estimates for experiments and nearest neighbor parameters using the standard method for uncorrelated parameters:

\[ \sigma^2 = \sum_i \left( \frac{\partial \Delta G}{\partial \Delta G_i} \right)^2 \]

where δG_i is the ith term (or experimental value) and σ_i is the uncertainty in the ith term. For the sum of terms used here, this simplifies to

\[ \sigma^2 = \sum_i \left( \sigma_i \right)^2 \]

where \( n \) is the number of occurrences of the ith parameter or the ith experiment.

For uncertainty estimates for optical melting experiments, we use 4% of the magnitude of \( \Delta G_{37} \). This was chosen as a conservative estimate of the precision of optical melting by Xia et al. Optical melting data (Average of Curve Fits and Analysis of Tm Dependence) for a database of optical melting experiments.

Nearest neighbor parameter determination. Nearest neighbor parameters were developed to estimate the folding stability (\( \Delta G_{37} \)) of sequences with A, C, G, U, and m6A. Nearest neighbor parameters are inherited from the 2004 Turner RNA93, for which a summary of their derivation can be found in Zubarev et al. and examples for their use are available on the Nearest Neighbor Database website. Helical stacking tables are from Xia et al. and Watson–Crick and G–U stacks from Chen et al. for stacks that contain G–U pairs, supplemented with the stacks determined for m6A–U pairs in this work. Following Chen et al., terminal G–U base pairs in a helix are not penalized.

Dangling end m6As are stabilized as compared to the analogous A dangling end by the mean additional stability found here (−0.4 kcal/mol). Dangling ends on m2A–U pairs are assumed to be the same stability as dangling ends on A–U base pairs when the stability is measured by an experiment, the measured value is used (Supplementary Table S4).

Terminal mismatches involving m2A are estimated to be more stable than the analogous A terminal mismatch by the mean value found for the terminal mismatches in this study (−0.3 kcal/mol). m2A–m2A terminal mismatches receive only −0.6 kcal/mol additional stabilization. A terminal mismatch involving m2A is also stabilized by −0.3 kcal/mol compared to the analogous mismatch on an A–U pair. These effects are additive; an m2A-containing terminal mismatch on an m2A–U pair receives an additional −0.6 kcal/mol stability than the analogous terminal mismatch with all A parameters. When the stability is measured by an experiment, the measured value is used (Supplementary Table S4).

Hairpin, internal, and bulge loop initiation costs are length-dependent. The same length-dependent costs are used here, which assumes that m2A does not alter the initiation costs.

In total, 1 × 1, 2 × 1, and 2 × 2 internal loop loops are stored in lookup tables. The stabilities for loops with unpaired m2A are taken from the analogous loops with A. And m2A–U-closed loops are taken from analogous A–U-closed loops with one change. A–U-closed loops have a 0.7 kcal/mol stability penalty per closure95; for m2A–U-closed loops, this cost has been removed compared to the analogous A–U-closed loop. Larger internal loops use a terminal stacking table to assign a stability increment for the sequence of the closing pair and first mismatch. Separate tables are used for loops of size 1 n, 2 × 3, and (2 × 2)3. These terminal stack tables use the analogous A parameter for stacks with m2A. The one exception is that the +0.7 kcal/mol internal loop A–U pair closure penalty is removed for m2A–U closures.

Hairpin loop tables for triloop, tetraloops, and hexaloops are unchanged. These tables include stabilities for specific hairpin sequences known by experiment to not be well predicted using nearest neighbor rules. Other hairpin loops are estimated with the sum of a terminal mismatch and a length-dependent initiation. The terminal mismatches for m2A use the analogous A parameter.

Multibranch loop initiation parameters are from an experimental fit using a simple linear model. Coaxial stacking is included in multibranch and exterior loops. Coaxial stacking between two adjacent helices is assumed to be as stable as a helical stack. For coaxial stacks with an intervening mismatch, there are two coaxial stacking increment for the stack where the backbone is not continuous was previously found to be independent of sequence, and the sequence-independent value is used here for stacks involving one or more m2As. The other stack is identical to the terminal mismatch stack table.

Extended alphabet implementation in RNAstructure. RNAstructure is a software package written in C++, with a C++ class library that is also wrapped using SWIG to be available to JAVA or Python programs. It is open source and is provided for free under the GNU GPL license version 2 at [https://rna.urmc.rochester.edu](https://rna.urmc.rochester.edu). A number of the command line programs have been updated to be capable of using extended alphabets, including Fold (secondary structure prediction by free energy minimization), en2f (estimation of folding free energy changes for secondary structures), and partition (partition function calculation for estimating terminal parameters) has been added. A number of command line programs that rely on the partition function calculations are therefore also able to consider extended alphabets, including design (design of a sequence to fold to a specific secondary structure), EnsembleEnergy (calculation of the ensemble folding free energy change), MaxExpect (prediction of maximum expected accuracy structures), FoldPlot (prediction of the output structure that can include pseudoknots), ProbScan (estimation of motif probabilities), and stochastic (stochastic sampling from the Boltzmann ensemble).
The underlying secondary structure prediction algorithms remain unchanged, although substantial changes were made to expand the software from the four nucleotide nucleic acid alphabet to a fully user-customizable folding alphabet. To do this, the command line tools read the thermodynamic parameters at startup. The switch “—alphabet” is used to specify the set of parameters to be used. The default is “rna”, the current (2004) Turner rules for estimating RNA folding free energy changes55,56. Included with the latest RNAstructure release (version 6.3) is also “rna54”, the parameters discussed herein, and “dna”, a set of nearest neighbor rules for DNA secondary structure prediction. The files are a plain text format that was updated (in version 6.0) for extended alphabets. The specification file (Supplementary Fig. S2) is read first, and it defines the alphabet and base pairs. Dynamax memory is used to provide the memory needed to store the tables. The parameters themselves are then read from the files.

The 2004 Turner rules gave a terminal base pair penalty for any base pair (A-U or G-U) at the end of a helix that contained a U334. In this work, we found that terminal m²A-U pairs did not require this terminal base pair penalty. Additionally, the revised G-U stack parameters57, used with the m²A parameters we derived, do not require a terminal base pair penalty. Therefore, we changed the implementation of the energy function to account for this change. The default “rna” rules are unchanged and still penalize terminal G-U base pairs.

Chemical mapping of RNA and data analysis. DMS (to modify adenosine and cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridine and guanosine) and kethoxal (to modify cytidine), CMCT (to modify uridan...
References

1. Phizicky, E. M. & Hopper, A. K. RNA biology charges to the front. Genes Dev. 24, 1832–1860 (2010).
2. Li, X., Xiong, X. & Yu, C. Epitranscriptome sequencing technologies: decoding RNA modifications. Nat. Methods 14, 23–31 (2016).
3. Gilbert, W. V., Bell, T. A. & Schaening, C. Messenger RNA modifications: form, distribution, and function. Science 352, 1408–1412 (2016).
4. Sakurai, M. et al. A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. Genome Res. 24, 522–534 (2014).
5. Carlile, T. M. et al. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515, 143–146 (2014).
6. Schwartz, S. et al. Transcriptome-wide imaging reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell 159, 148–162 (2014).
7. Lovejoy, A. F., Riordan, D. P. & Brown, P. O. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases specify unique modifications in S. cerevisiae. PLoS ONE 9, e107994 (2014).
8. Squires, J. E. et al. Widespread occurrence of 5-methylcytosine in human DNA. Nucleic Acids Res. 40, 5023–5033 (2012).
9. Schwartz, S. et al. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell 155, 1493–1504 (2013).
10. Linder, B. et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat. Methods 12, 767–772 (2015).
11. Ke, S. et al. A majority of m6A residues are in the last exons, allowing the potential for 3′ UTR regulation. Genes 29, 2057–2053 (2015).
12. Dominissini, D. et al. Topology of the human and mouse m6A RNA methylation revealed by m6A-seq. Nature 485, 201–206 (2012).
13. Meyer, K. D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell 149, 1635–1646 (2012).
14. Chen, K. et al. High-resolution N(6)-methyladenosine (m6A) A map using pho-cros-oligomer-aided mapping A sequencing. Angew. Chem. Int. Ed. Eng. 54, 1587–1590 (2015).
15. Esteller, M. & Pandolfi, P. P. The epitranscriptome of noncoding RNAs in cancer. Cancer Disco. 7, 359–368 (2017).
16. Pan, T. N6-methyl-adenosine modification in messenger and long non-coding RNA. Trends Biochem. Sci. 38, 204–209 (2013).
17. Zaccara, S., Ries, R. J. & Jaffrey, S. R. Reading, writing and erasing mRNA regulation reveals regulated pseudouridylation of ncRNA and mRNA. Nat. Methods 15, 148–149 (2018).
18. Jiang, T., Kennedy, S. D., Moss, W. N., Kierzek, E. & Turner, D. H. Secondary structure prediction and analysis. BMC Bioinform. 11, 129 (2010).
19. Rivas, E. The four ingredients of single-sequence RNA secondary structure prediction. RNA 10, 1185–1196 (2003).
20. Zuber, J., Cabral, B. J., McFadyen, I., Mauger, D. M. & Mathews, D. H. Analysis of RNA nearest neighbor parameters reveals interdependencies and quantifies the uncertainty in RNA secondary structure prediction. RNA 24, 1568–1582 (2018).
21. Zuber, J., Sun, H., Zhang, X., McFadyen, I. & Mathews, D. H. 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 (2017).
22. Xue, T. et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. Biochemistry 37, 14719–14735 (1998).
23. Liu, B. et al. A potentially abundant junctional RNA motif stabilized by m(6)A and Mg(2+). Nat. Commun. 9, 2761 (2018).
24. Gygi, S. V. et al. N6-methyladenosine in flaviviridae viral RNA genomes regulates infection. Cell Host Microbe 20, 654–665 (2016).
25. Blowe, J. M. et al. Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of I single-nucleotide bulge loops. Biochemistry 46, 15123–15135 (2007).
26. Kent, J. L. et al. Non-nearest-neighbor dependence of stability for group III intron RNA single nucleotide bulge loops. RNA 20, 825–834 (2014).
27. Leamy, K. A., Assmann, S. M., Mathews, D. H. & Bevilacqua, P. C. Bridging the gap between in vitro and in vivo RNA folding. Q. Rev. Biophys. 49, e10 (2016).
28. Diamond, J. M., Turner, D. H. & Mathews, D. H. Thermodynamics of three-way multibranch loops in RNA. Biochemistry 40, 6971–6981 (2001).
29. Jäger, I. A., Zuker, M. & Turner, D. H. Melting and chemical modification of a cyclized self-splicing group I intron: similarity of structures in 1 M Na2SO4 in 10 mM Mg2+ and in the presence of substrate. Biochemistry 29, 10147–10158 (1990).
30. McDowell, J. A. & Turner, D. H. Investigation of the structural basis for thermodynamic stabilities of tandem GU mismatches: Solution structure of (GAGGCUCUC), by two-dimensional NMR and simulated annealing. Biochemistry 35, 14077–14089 (1996).
31. Xia, T., McDowell, J. A. & Turner, D. H. Thermodynamics of nonsymmetric tandem mismatches adjacent to GC base pairs in RNA. Biochemistry 36, 12486–12487 (1997).
32. Schroeder, S. J. & Turner, D. H. Factors affecting the thermodynamic stability of small asymmetric internal loops in RNA. Biochemistry 39, 9257–9274 (2000).
33. Jiang, T., Kennedy, S. D., Moss, W. N., Kierzek, E. & Turner, D. H. Secondary structure of a conserved domain in an intron of influenza A M1 mRNA. Biochemistry 35, 5236–5248 (2014).
34. Serra, M. J. et al. Effects of magnesium ions on the stabilization of RNA oligomers of defined structures. RNA 8, 307–323 (2002).
35. Liu, N. et al. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. Nature 518, 560–564 (2015).
36. Cienikova, Z., Damberger, F. F., Hall, J., Allain, F. H. & Maris, C. Structural and mechanistic insights into poly(uridylate) tract recognition by the hnRNP C RNA recognition motif. J. Am. Chem. Soc. 136, 14536–14544 (2014).
