Original Paper

Exploring the Accuracy of Ab Initio Prediction Methods for Viral Pseudoknotted RNA Structures: Retrospective Cohort Study

Vasco Medeiros[1], BSc, MSc, AS; Jennifer Pearl[2], BSc, AS; Mia Carboni[3], BSc; Stamatia Zafeiri[1], BSc, MSc

[1]Stevenage Bioscience Catalyst, Stevenage, United Kingdom

[2]Stanford University, Santa Clara, CA, United States

[3]University of Turin, Torino, Italy

Corresponding Author:
Vasco Medeiros, BSc, MSc, AS
Stevenage Bioscience Catalyst
Gunnels Wood Rd Stevenage
Stevenage, SG1 2FX
United Kingdom
Phone: 44 07534150352
Email: vasco.miguel.medeiros@gmail.com

Related Articles:
Preprint (bioRxiv): https://www.biorxiv.org/content/10.1101/2024.03.21.586060v1
Peer-Review Report by Daniela Saderi, Vaishnavi Nagesh, Randa Salah Gomaa Mahmoud, Toba Olatoye, and Femi Qudus Arogundade: https://bio.jmirx.org/2024/1/e65154
Authors' Response to Peer-Review Reports: https://bio.jmirx.org/2024/1/e67586

Abstract

Background: The prediction of tertiary RNA structures is significant to the field of medicine (eg, messenger RNA [mRNA] vaccines, genome editing) and the exploration of viral transcripts. Though many RNA folding software programs exist, few studies have condensed their locus of attention solely to viral pseudoknotted RNA. These regulatory pseudoknots play a role in genome replication, gene expression, and protein synthesis.

Objective: The objective of this study was to explore 5 RNA folding engines that compute either the minimum free energy (MFE) or the maximum expected accuracy (MEA), when applied to a specified suite of viral pseudoknotted RNAs that have been previously confirmed using mutagenesis, sequence comparison, structure probing, or nuclear magnetic resonance (NMR).

Methods: The folding engines used in this study were tested against 26 experimentally derived short pseudoknotted sequences (20-150 nt) using metrics that are commonplace while testing software prediction accuracy: percentage error, mean squared error (MSE), sensitivity, positive predictive value (PPV), Youden's index (J), and F_1-score. The data set used in this study was accrued from the Pseudobase++ database containing 398 RNAs, which was assessed using a set of inclusion and exclusion criteria following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Base pairings within a given RNA sequence were deemed correct or incorrect following Mathews' parameters.

Results: This paper reported RNA prediction engines with greater accuracy, such as pKiss, when compared to previous iterations of the software and when compared to older folding engines. This paper also reported that when assessed using metrics such as the F_1-score and the PPV, MEA folding software does not always outperform MFE folding software in prediction accuracy when applied to viral pseudoknotted RNA. Moreover, the results suggested that thermodynamic model parameters will not ensure accuracy if auxiliary parameters, such as Mg^{2+} binding, dangling end options, and hairpin-type penalties, are not applied.

Conclusions: This is the first attempt at applying a suite of RNA folding engines to a dataset solely comprised of viral pseudoknotted RNAs. The observations reported in this paper highlight the quality between difThis is the first attempt at applying a suite of RNA folding engines to a data set solely comprising viral pseudoknotted RNAs. The observations reported in this paper highlight the quality between different ab initio prediction methods, while enforcing the idea that a better understanding of intracellular thermodynamics is necessary for a more efficacious screening of RNAs.ferent ab initio prediction methods while enforcing the idea that a better understanding of intracellular thermodynamics is necessary for a more efficacious screening of RNAs.

(JMIRx Bio 2024;2:e58899) doi: 10.2196/58899

KEYWORDS

pseudoknot; viral RNA; MFE; minimum free energy; MFE prediction; MEA; maximum expected accuracy; MEA prediction; virus; virology; computational biology

Introduction

Computational biology is 1 of the key tools we possess to understand RNA folding and is used in pharmacokinetics, drug discovery, and pharmacology. In silico predictions of catalytic RNAs help narrow down and consolidate a surfeit of data, while expediting the search for potential drug targets. As of now, we know that catalytic RNA controls for ribozymes, riboswitches, messenger RNA (mRNA) vaccines, thermosensors, and essential elements of genome editing [1-4]. This is due to RNA's ability to fold itself into tertiary structures (pseudoknots), forming binding pockets and active site clefts that can act as targets for active pharmaceutical ingredients (APIs) [5]. As artificial intelligence (AI), computer processing, and data throughput continue to advance, we are witnessing these methodologies more frequently implemented in the field of virology [6]. This paper explores the different ways in which stochastic folding engines predict viral pseudoknotted RNAs and the accuracy of these approaches.

Pseudoknots are structural motifs found in almost all classes of RNA. Although most RNA forms planer secondary structures, these 3D structures embody up to 30% of tertiary nonplaner motifs in G+C-rich RNA sequences [7]. In the context of viruses, these pseudoknots control gene expression and protein synthesis. Many catalytic RNAs regulate the mechanisms of action associated with viral replication, viral translation, or both. An example of this can be seen in satellite viruses (eg, hepatitis delta virus, satellite tobacco necrosis virus 1) that encode ribozymes that are folded by pseudoknotted structures [8,9].

For decades, scientists have explored and created different prediction software to better elucidate the complicated nature of RNA folding. Though RNA is a biopolymer that folds in a specific manner, it can be difficult to discern which prediction algorithms, alignment sequences, or applied mathematics would result in the most accurate model. This difficulty becomes more apparent when noting how much the field has changed over the years and how small changes in the underlying formalisms and constraints can result in drastic differences in the final predicted structure.

RNA folding occurs through populated intermediates and is accomplished in a hierarchical manner, where secondary planer forms come prior to tertiary contacts [10]. This allows software engines to model both canonical and noncanonical base pairs, making it so the inputs within V(i,j) base pairs have a range of integer values (rather than binary values) dependent on the base pairs they form [11,12]. V(i,j) in this case is the real symmetric contact matrix of N×N, where N represents the number of nucleotides on a given polymer chain.

Using V(i,j), and other mathematical formalisms, derived from prior experimental data to model the effects of salinity, pH, temperature, loop entropies, and stacking formations, we can generate a "pseudo-energy model." This grants us a measure of the relative probability of different RNA secondary structures, expounding on the ensemble free energy, and the equilibrium concentrations of all possible structures, all of which correspond to the topological character of the RNA strand [13].

Within the RNA template, the first base at the 3' terminus is regarded as 1, and the final base found at the 5' terminus is regarded as N. In the total secondary structure, made up of V(i,j) base pairs, the index $1 \leq i < j \leq N$ should be set. Each integer within a given matrix will represent the i-th nucleotide being paired with the j-th nucleotide. The base pairs, (G-C), (A-U), and, at times, (G-U), dependent on the algorithm/software used, are the integer values that contribute to the matrix field. The entire structure of length N is regularly represented in Feynman diagrams, also known as arc and chord diagrams (Figure 1A,B [14-18]), where each nucleotide is represented as a point on the chain, while each arc represents a base pair forming between any nucleotides i and j.

Gilbert and coworkers [6] defined a pseudoknot as follows: "Pseudoknots are formed upon base pairing of a single-stranded region of RNA in the loop of a hairpin to a stretch of complementary nucleotides elsewhere in the RNA chain."

We proposed more specified definitions for the sake of clarity posing 1 definition of a pseudoknot as "a template of RNA in which nucleotides within a loop pair with regions that do not pertain to the helices that close said loop." Another definition could be "an RNA secondary structure that forms base pair regions upstream or downstream, resulting in stem-loop structures." Their topology is more varied than most other assemblies of RNA, presenting a challenge for in silico prediction software.

Although comparative approaches exist in the solving of optimal RNA structures [19,20], including web servers, such as KNetFold [21] and pAliKiss [22], this review focused on the ab initio topological predictions based primarily on the RNA secondary structure. The accuracy of these ab initio stochastic RNA folding software programs will be assessed in relation to a catalog of 26 distinct viral pseudoknotted RNAs taken from PseudoBase++ [23], whose wild-type structures have been previously determined via sequence comparison, structure probing, mutagenesis, and nuclear magnetic resonance (NMR). The RNA predictions generated in this paper, imparted by the underlying formalisms of the software, will result in structures that represent either the minimum free energy (MFE) or the maximum expected accuracy (MEA) as both models are compared.

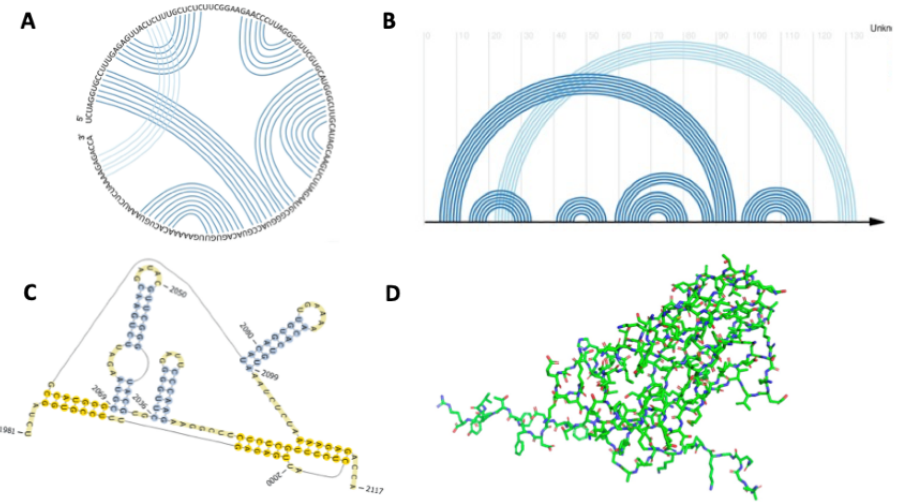
It can generally be posited that base pairs, when formed, lower the Gibbs free energy of a ribonucleic strand, making use of the attractive interactions between the complementary strands. MFE prediction algorithms assess these by solving for the maximum number of nucleotide pairings, via their thermodynamic

properties, which generally results in the lowest energy form. It is important to note, however, that in nature, kinetic barriers, environmental conditions, and other factors may influence RNA folding intermediates, resulting in a physiologically favored RNA that does not coincide with the MFE structure [24].

Conversely, MEA models compute the final RNA structure via a partition function (a function that is used to calculate the thermodynamic properties of a system) that implements hard and soft constraints based on electrostatic interactions, stacking interactions, adjacent complementary base pairs, and other variables, depending on the software in question [24]. This results in a final structure that may not necessarily encompass the lowest-possible free energy of a system.

The current literature compares the accuracy of RNA folding software when applied to viral RNAs and cellular RNAs, suggesting there exists no difference in accuracy between the 2 [25]. These investigations have explored viral RNAs of various lengths, accounting for the positive predictive value (PPV), sensitivity, and F_1-scores. However, to the best of our knowledge, few papers exist that address the accuracy of RNA folding software when applied to viral pseudoknotted RNA transcripts alone. Moreover, the literature does not expound on the differences in accuracy regarding MEA and MFE modalities when applied to viral pseudoknotted RNA. This paper aimed to address this knowledge gap within the literature by forming a highly specific investigation of a data set of 26 short pseudoknotted sequences (20-150 nt), with updated versions of existing stochastic RNA prediction algorithms. This investigation addressed whether MEA prediction modalities are, in fact, more accurate than MFE modalities and which of the 5 folding software programs are more accurate when applied to the data set of 26 pseudoknotted RNAs.

Figure 1. Ways in which to model pseudoknotted RNA. (A) Circular arc and chord diagram of the viral tRNA-like brome mosaic virus [14,15]. (B) Planer arc and chord drawing of the viral tRNA-like brome mosaic virus made using the R-chie package [16]. (C) Planer representation of viral tRNA-like brome mosaic virus using nitrogenous bases [17]. (D) Three-dimensional model of tRNA-like brome mosaic virus in stick format made using PyMOL Molecular Graphics System version 2.0 [18]. tRNA: transfer RNA.



Methods

RNA Folding Engines, RNA Classes, and Genuses Assayed

The 5 RNA secondary structure prediction servers used in this paper are listed in Table 1.

See Sections S1-S5 in Multimedia Appendix 1 to view further details regarding the auxiliary parameters enforced by each software program. Access/links to the data set, as well as the 5 stochastic RNA folding web servers, are provided in Table S1 in Multimedia Appendix 1, in accordance with the FAIR (findability, accessibility, interoperability, and reusability) principles of data sharing.

The knowledge we possess pertaining to pseudoknots and their metabolic functions holds much of its origins in the study of viral biology, mounted on well-studied strains, such as flaviviruses, influenza viruses, and mosaic viruses [32-34]. These structures (Figure 2 [14,35-38]) encompass the regulatory elements of some viruses, controlling various phases of gene expression and function. Though many forms of pseudoknot classification have been conjectured [24,39,40], in this paper, all pseudoknots fall under 1 of the following 6 categories, building upon the grouping proposed by Legendre et al [24]:

- Hairpin-type (H-type) pseudoknot.
- Kissing hairpin–type (HHH-type) pseudoknot.
- Hairpin loop outer (HLout) pseudoknot: pertains to pseudoknots that form base pairs residing outside of a hairpin loop. This structure typically involves a larger loop enclosing the pseudoknot, which results in a more extended configuration.
- Hairpin loop inner (Hlin) pseudoknot: pertains to pseudoknots that form base pairs residing inside a hairpin loop. The pseudoknot is nested within the loop structure, creating overlapping interactions within the loop itself.
- HLout,HLin pseudoknot.
- Loop-loop (LL) pseudoknot: forms between 2 distinct loops within the RNA structure. This base pair crossing results in steric interactions and stacking interactions across separate looped regions.

Table 1. MEA[a] and MFE[b] folding engines applied to viral RNA structures.

Name of the folding engine	Method of prediction	Thermodynamic model parameters	Pseudoknots enforced	Auxiliary parameters enforced	Study
Chiba Institute of Technology's Vsfold 5: RNA Secondary Structure Prediction Server	MEA	Jacobson-Stockmayer (standard parameters)	Yes	Kuhn length, Mg^{2+} binding, contiguous stems, minimum stem length, length of leading stem	Dawson et al [26]
Universitat Bielefeld's BiBiServ's (pKiss)	MFE	Turner model	Yes	H-type penalty, K-type penalty, maximal pseudoknot size, minimal hairpin length, lonely base pairs	Janssen and Giegerich [21]
Institut Curie's Kinefold	MFE	Turner model	No	Cotranscriptional fold, simulated molecular time, tracing and forcing helices	Xayaphoummine et al [27]
NUPACK 3.0	MFE	Turner model	No	Mg^{2+} binding, dangling end options, input of multiple interactions	Zadeh et al [28], Fornaceet al [29]
Vienna RNAfold[c]	MFE	Turner model	—[d]	Avoiding isolated base pairs, incorporation of G-quadruplex formation into the structure prediction algorithm, dangling end options, addition of modified base pairs	Hofacker et al [30]

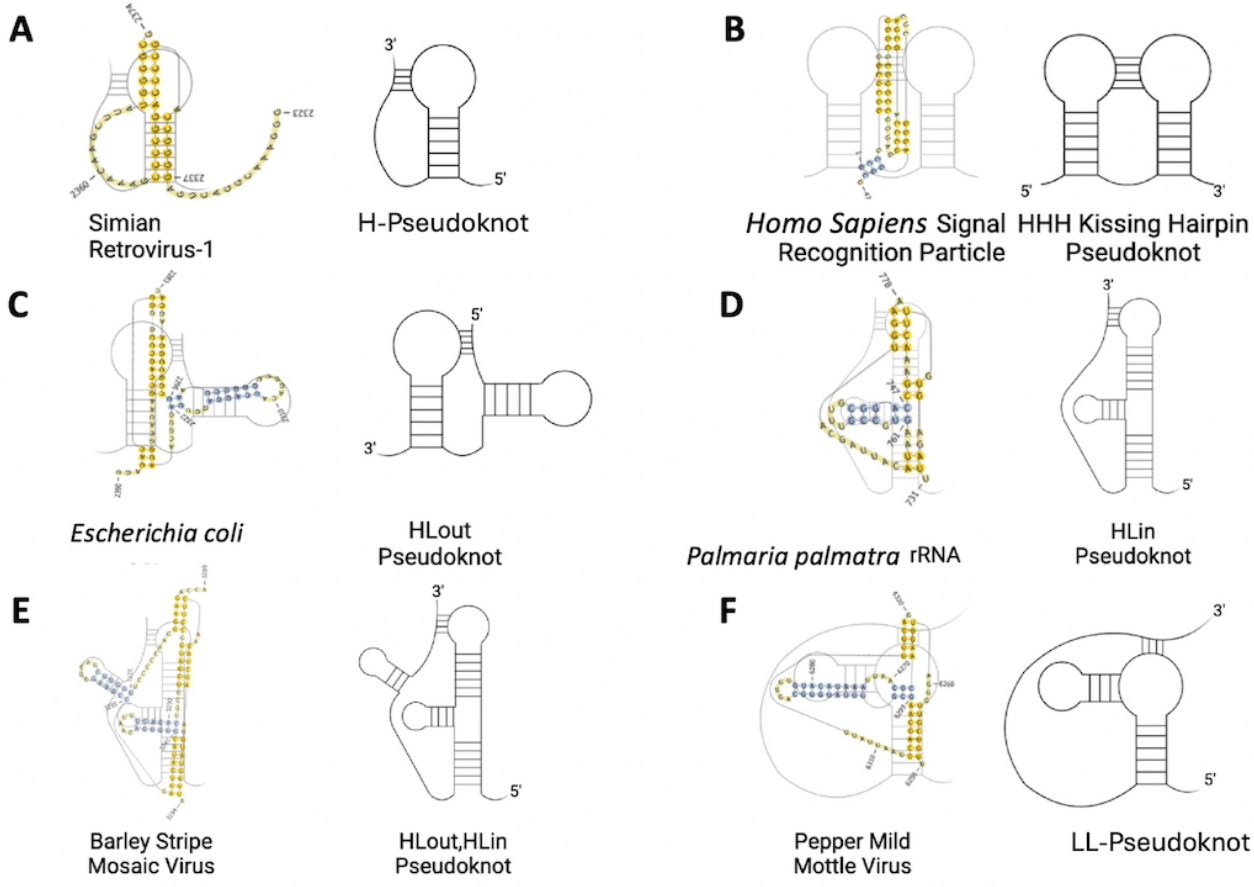
[a]MEA: maximum expected accuracy.

[b]MFE: minimum free energy.

[c]The Vienna RNAfold engine does not compute for pseudoknots and is implemented as a negative control [31].

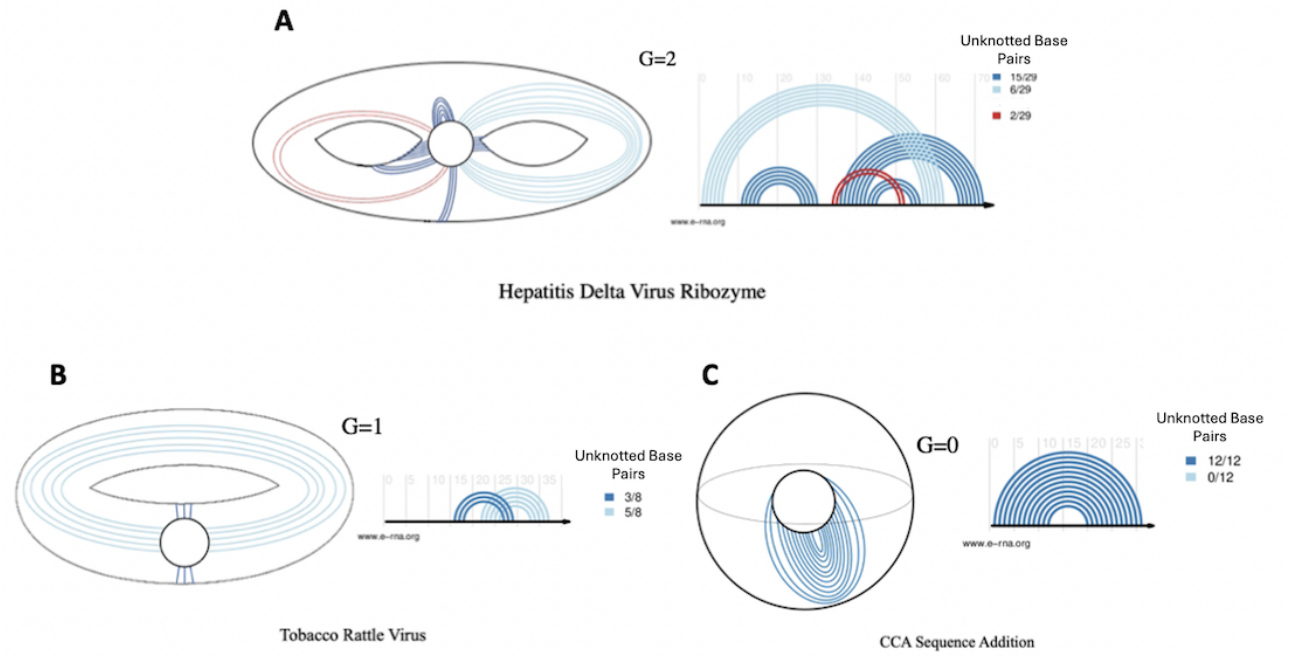
[d]Not applicable.

Figure 2. Six classes of pseudoknot. The figure depicts the class of pseudoknot (skeletal structure) on the right, with an example of an accepted structure on the left. (A) Simian retrovirus-1 [35]. (B) *Homo sapiens* signal recognition particle [36]. (C) *Escherichia coli* [37]. (D) *Palmaria palmatra* rRNA [38]. (E) Barley stripe mosaic virus [14]. (F) Pepper mild mottle virus [14]. H: hairpin; HHH: kissing hairpin; Hlin: hairpin loop inner; HLout: hairpin loop outer; LL: loop-loop; rRNA: ribosomal RNA.



Given that most classified pseudoknots in data banks (both proteins and nucleic acids) have a genus of 1 [41], all pseudoknots expounded on in this investigation had a genus of 1. The "genus" of a pseudoknot refers to a mathematical concept that correlates to the topology of a surface. The genus is a positive integer value, which corresponds to the minimum number of handles of the embedding surface of a structure, or, more simply, the number of times the RNA molecule intersects with itself in 3D space. A genus of 0 means that the graph can be drawn without any crossing on a sphere. A genus of 1 means that the graph can be drawn on a torus (doughnut shape), without any base pairs crossing (Figure 3 [14,41-43]).

Figure 3. Depiction of genus 2, 1, and 0 RNAs. (A) Hepatitis delta virus ribozyme in 3D space and a planer arc and chord diagram [41,42]. (B) Tobacco rattle virus in 3D space and a planer arc and chord diagram [14]. (C) CCA sequence adding polymerase in 3D space and a planer arc and chord diagram [43].



Generating a Viral Pseudoknotted RNA PseudoBase++ Data Set

For the 398 RNAs found in the Pseudobase++ database, 205 (51.5%) peer-reviewed papers referenced in these sites were vetted thoroughly following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [44]. The search identified 87 (42.4%) eligible studies. In addition to solely using pseudoknots of a genus 1 class, the computed MFE of all pseudoknots was also considered (see Table S2 in Multimedia Appendix 1). Any pseudoknot resulting in too high an MFE (P<.05, df=1) was excluded from the data set and considered an outlier. What remained were 26 (31.7%) RNAs of varying sizes from 20 to 150 nt, conforming to 1 of the 6 structures depicted in Figure 2 (see Table S3 in Multimedia Appendix 1 for further information about individual structures). Structures were derived from both plant and animal species, with each structure corresponding to a unique viral genome.

As mentioned previously, all pseudoknots expounded in this report fall under at least 1 of the 6 categories described in Figure 2. Of the 26 RNAs assessed, 17 (65.4%) consist of H-type pseudoknots, while the other 9 (34.6%) fall under 1 of the other configurations listed. This skewness is intentional, and true to nature, given that H-type pseudoknots are more common by far [6]. In addition, of the pseudoknotted RNAs assessed, 16 (61.5%) harbor viral transfer RNA (tRNA)–like motifs, 5 (19.2%) harbor viral 3' untranslated region (UTR)–like motifs, and 4 (15.4%) harbor viral frameshifts. Each motif plays an essential role in the replication of viruses and was thus included in the data set (see Table S3 in Multimedia Appendix 1 for further information). For example, the pseudoknotted UTRs of positive-strand RNA viruses regulate and initiate protein synthesis and replicase enzymes, L-shaped pseudoknots that resemble tRNAs propagate viral proteins, and viral frameshifts result in the production of unique proteins (especially in retroviruses) [6].

Assessing Prediction Software Efficacy Through Percentage Error and Mean Squared Error Metrics

Base pairings were deemed correct following Mathews' parameters [45]. Through this system, base pairing within a reference sequence of length N, between base pairs i and j (where $1 \leq i < j \leq N$), was considered correct if i was paired with either j, j – 1, or j + 1 or if j was paired with i, i – 1, or i + 1. If a pairing did not fall under these conditions, it was considered a false positive (FP). If a pairing that fell under these conditions was missed by a given prediction software program, it was classified as a false negative (FN). This model allows for some elasticity and leniency in prediction, deeming base pairing correct even if base pairs are displaced by 1 nucleotide either up- or downstream. This is the standard for in silico RNA computation. Further benchmarks used by Mathews [45] include the following:

- A large set of well-established reference/accepted structures to compare against experimentally derived data
- Tests for statistical significance
- Different RNA families/RNA types (see Table S3 in Multimedia Appendix 1) that should be used

First, the percentage error (represented in the results as the mean absolute error [MAE]) was used to assess the prediction accuracy of total base pairs and knotted base pairs. The percentage error is expressed as the absolute value of the difference between base pairings derived from 1 of the 5 RNA folding engines (Table 1) and base pairs from the PseudoBase++ database derived from sequence comparison, structure probing, and NMR, in percentage format:

$$\%\ \text{Error} = \frac{|\text{Number of base pairs predicted} - \text{number of base pairs of accepted structures}|}{\text{Number of base pairs of accepted structures}} \times 100 \quad (1)$$

This manner of assessment is robust against outliers and better suited to delineating between software that can or cannot predict pseudoknots.

The mean squared error (MSE) was applied to all experimental and control conditions. The MSE for a given folding engine is simply

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i)^2 \quad (2)$$

where n is the number of data points, Y_i is observed values, and X_i is predicted values.

A better model will possess a better fit to the data should the resultant MSE be closer to 0. Often used in machine learning and regression analysis, this metric is appropriate when applied to evaluating the functioning of predictive models. However, these 2 metrics do not encompass a software program's ability to holistically assess both knotted and total base pairs in tandem. For this reason, many papers are instead opting for sensitivity (recall) and the PPV (precision) [45,46].

Assessing Prediction Software Efficacy Through Sensitivity, PPV, and Youden Index Metrics

Sensitivity and the PPV are common and important systems of measurement when predicting software accuracy. The former assesses a folding engine's ability to identify correct base pairs, while the latter assesses a software program's propensity to incorrectly identify base pairs (resulting in a value of 1-0).

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positves} + \text{false negatives}} \quad (3)$$

$$\text{PPV} = \frac{\text{True positives}}{\text{True positives} + \text{false positives}} \quad (4)$$

Here, sensitivity reduces from 1 when an RNA folding engine misses pairings, while the PPV reduces from 1 the more an RNA folding engine predicts bases that are not of the original secondary structure. Though these metrics are commonplace in almost all branches of bioinformatics, such as genomic variant calling and drug targeting prediction [47,48], this report imposed them onto folding engines.

Youden index values (J) were also used to compare the accuracy of each model, with higher values being indicative of models with higher discriminative ability. Simply defined by the following equation:

$$J = \text{Sensitivity} + \text{specificity} - 1, \quad (5)$$

this metric can range from –1 (denoting completely incorrect detection for a given group) to 1 (denoting completely correct detection for a given group). It is important to clarify that the value of J is often represented on a receiver operating characteristic (ROC) curve, with recall (true positive [TP] rate, T_{pr}, on the y axis and the FP rate [F_{pr}] on the x axis), when applied to 1 reference structure using 1 predictive model. However, this paper compared all relative J values imposed by all models across all 26 viral pseudoknotted RNAs in column format.

Assessing Prediction Software Efficacy Through F_1-Score Metrics

Once the PPV and sensitivity have been calculated, F_1-scores can then be derived for each structure. F_1-scores are a standard method used to evaluate prediction software [24,45,46,49] and assess the harmonic mean between sensitivity and the PPV by considering 3 of the 4 confusion matrix categories (TP, FP, and FN). It should be noted once more that for a base pairing prediction to be considered incorrect, it must fall outside the parameters established by Mathew [45], where base i is paired with j + (≥2) or j – (≥2) or base j is paired with i + (≥2) or i – (≥2)—or, in layperson terms, when the base pair is displaced by 2 nucleotides, either upstream or downstream, relative to the native structure.

F_1-scoring remains an archetype for binary classification problems [49] and integrates the boons of sensitivity and the PPV to produce a higher-caliber performance metric:

$$F_{1-\text{score}} = \frac{2 \times \text{Sensitivity} \times \text{PPV}}{\text{Sensitivity} + \text{PPV}} \quad (6)$$

Normality and lognormality testing included Kolmogorov-Smirnov tests and Shapiro-Wilk tests to confirm normal (gaussian) distribution of data (α=.05). Robust outlier (ROUT) tests were used to identify outliers (Q=1%), and statistical analysis included 1-sample *t* tests, Wilcoxon tests, and 1- and 2-way ANOVA testing. Testing was performed using GraphPad Prism v.9.5 software (Graph-Pad Software), and accuracy metrics were reported as decimals.

Ethical Considerations

Ethical considerations were taken into regard for this study, though the authors have nothing to declare, given the in silico nature of this paper.

Results

Assessment of Percentage Error and Mean Squared Error Metrics

In this study, we tested the 5 folding engines used against 26 experimentally derived short pseudoknotted RNA sequences

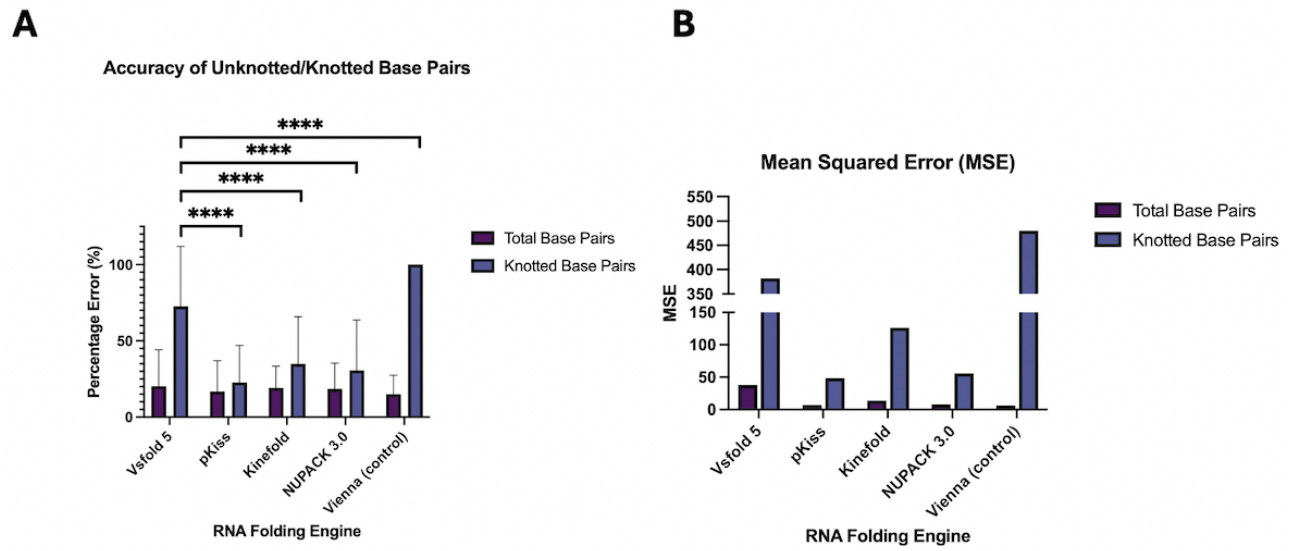
selected following PRISMA guidelines (Figure S1 in Multimedia Appendix 1). No significant difference in percentage error was observed between the total number of base pairs computed by all RNA folding algorithms (Figure 4), with the largest discernible difference between Vsfold 5 and Vienna (adjusted *P*=.99). This agrees with current studies, as the challenge biotechnicians face instead lies within predictions in O(N3) and O(N4) time and space, where N is the sequence length (using big O notation) [50]. Although certain algorithms can reduce higher-ordered structures to O(N3), thereby reducing computational complexity, a growing minimal N value correlates with more possible pseudoknots, making the algorithm less accurate [21].

The mean percentage error of total base pairs generated by each software program was 17.95%, with Vsfold 5 exhibiting the highest (mean 20.23%, SD 23.94%) and the negative control exhibiting the lowest (mean 15.07%, SD 12.35%) values, which was expected. Though the current literature now advocates for software that computes the MEA, rather than the MFE [24], the Vienna package is specifically designed to predict planer secondary structures.

A much higher percentage error was exhibited in knotted viral RNA structures produced by the MEA engine Vsfold 5, relative to its MFE-computing counterparts (apart from the negative control). Vsfold 5 was expected to give the lowest percentage error. However, the lowest percentage error for knotted base pairs was, instead, exhibited by MFE structures generated by pKiss (mean 22.37%, SD 24.2%), while Vsfold 5 retained a mean percentage error of 69.91% (SD 39.3%).

The values of the MSE drew parallels to those of the percentage error, with Vsfold 5 exhibiting the highest values of the experimental controls (382.29 for knotted bases and 37.85 for unknotted bases) and pKiss exhibiting the lowest values (48.4 for knotted bases and 6.88 for unknotted bases). Kolmogorov-Smirnov tests indicated that the data were normally distributed.

Figure 4. Percentage error and MSE. (A) Percentage error of total base pairs and knotted base pairs. Data correspond to the mean (SD) percentage error. The mean (SD) percentage error, across all 4 MFE RNA folding engines, was compared to that of the MEA computations of Vsfold 5, with statistical analysis performed using 2-way ANOVA, followed by a Tukey multiple-comparison test. ****P<.001 (df=4). The ROUT test was performed to identify outliers (Q=1%). (B) MSE (mean squared deviation). Kolmogorov-Smirnov tests performed on both "accuracy of unknotted/knotted base pairs" and "MSE" confirmed normal (gaussian) distribution of data (α=.05). MEA: maximum expected accuracy; MSE: mean squared error; ROUT: robust outlier.



Sensitivity, PPV, and Youden Index of Folding Engines

The second metric used to assess RNA folding engines was sensitivity, coupled with the PPV. The highest mean sensitivity and PPV were derived from pKiss, with mean 0.88 (SD 0.14) and mean 0.82 (SD 0.16), respectively. Conversely, the lowest mean sensitivity and PPV were derived from Kinefold, with mean 0.14 (SD 0.23) and mean 0.171 (SD 0.31), respectively (Figure 5). Ultimately, pKiss outperformed the mean of the Vsfold 5 MEA prediction software by mean 0.296 with respect to sensitivity and by mean 0.176 with respect to the PPV.

We would like to note that PPV values derived from free-energy minimization (ie, MFE folding engines) have been shown to be lower than sensitivity values [51,52]. This is likely because structures accepted in the literature can be missing base pairs that may occur experimentally and because the thermodynamics imposed by MFE algorithms often overshoot the number of canonical base pairings (because it is the formation of base pairs that innately lowers the Gibbs free energy of a structure) [53]. However, this trend does not present itself in all 4 experimental conditions but only in pKiss and NUPACK 3.0. This is because updated software such as this implements a more accurate assessment of the thermodynamic properties of the structure, removing unwanted pairs and improving overall performance [54].

In addition to the PPV and sensitivity, the Youden index (sometimes denoted as J) provides an additional framework to assess detection accuracy. As shown in Figure 6, pKiss exhibited the highest J value in raw and normalized data sets (mean 0.713, SD 0.267, and mean 100.0, SD 19.1, respectively), while

Kinefold exhibited the lowest J values in raw and normalized data sets (mean –0.68, SD 0.57, and mean 0.0, SD 41.0, respectively), reflecting the results in Figure 5.

Figure 5. Sensitivity and PPV of RNA folding engines on pseudoknotted viruses. (A) Data corresponding to the mean (SD) sensitivity and PPV across all 5 experimental conditions were compared to the sensitivity and PPV generated by Vsfold 5. Statistical analysis was performed using 2-way ANOVA, followed by a Tukey multiple-comparison test. ****P<.001 (df=1), **P≤.002 (df=1). The ROUT test was performed to identify outliers (Q=1%). (B) Sensitivity and PPV values plotted in a 1×1 matrix. The Shapiro-Wilk test confirmed normal (gaussian) distribution of data (α=.05). PPV: positive predictive value; ROUT: robust outlier.

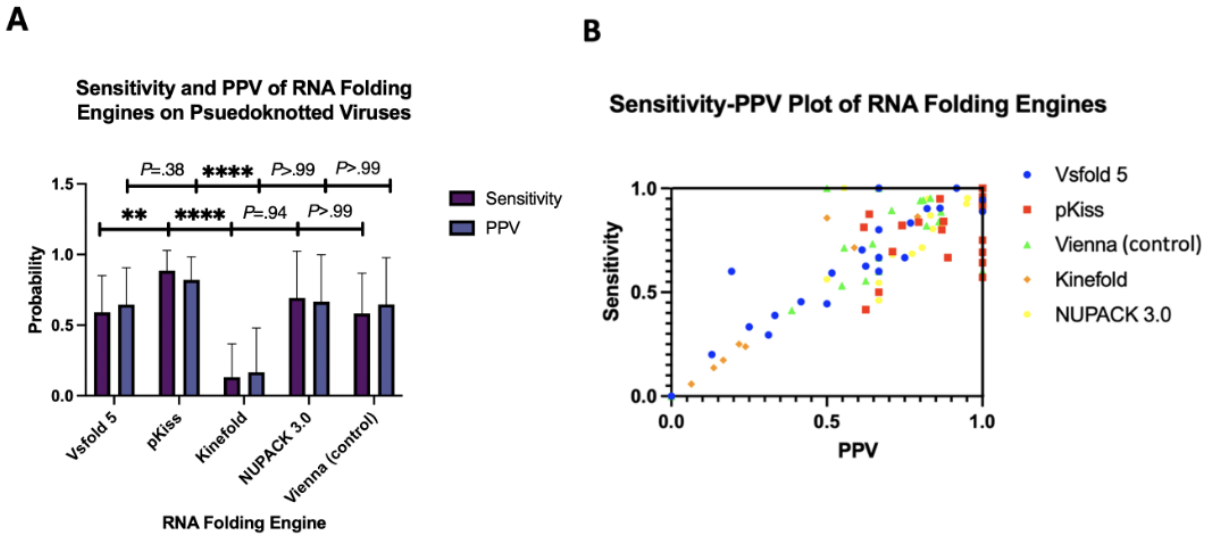
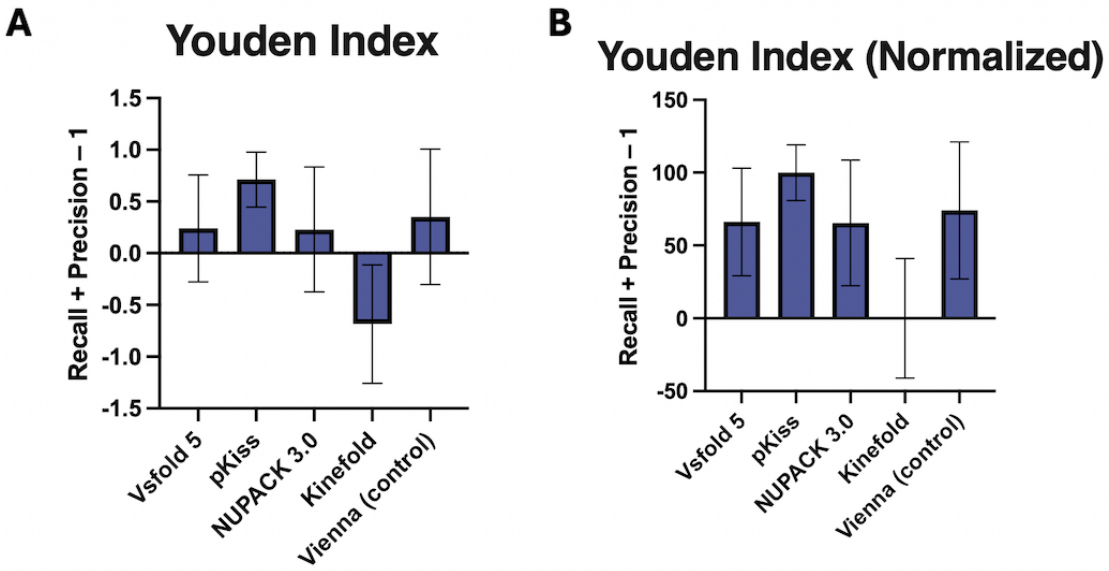


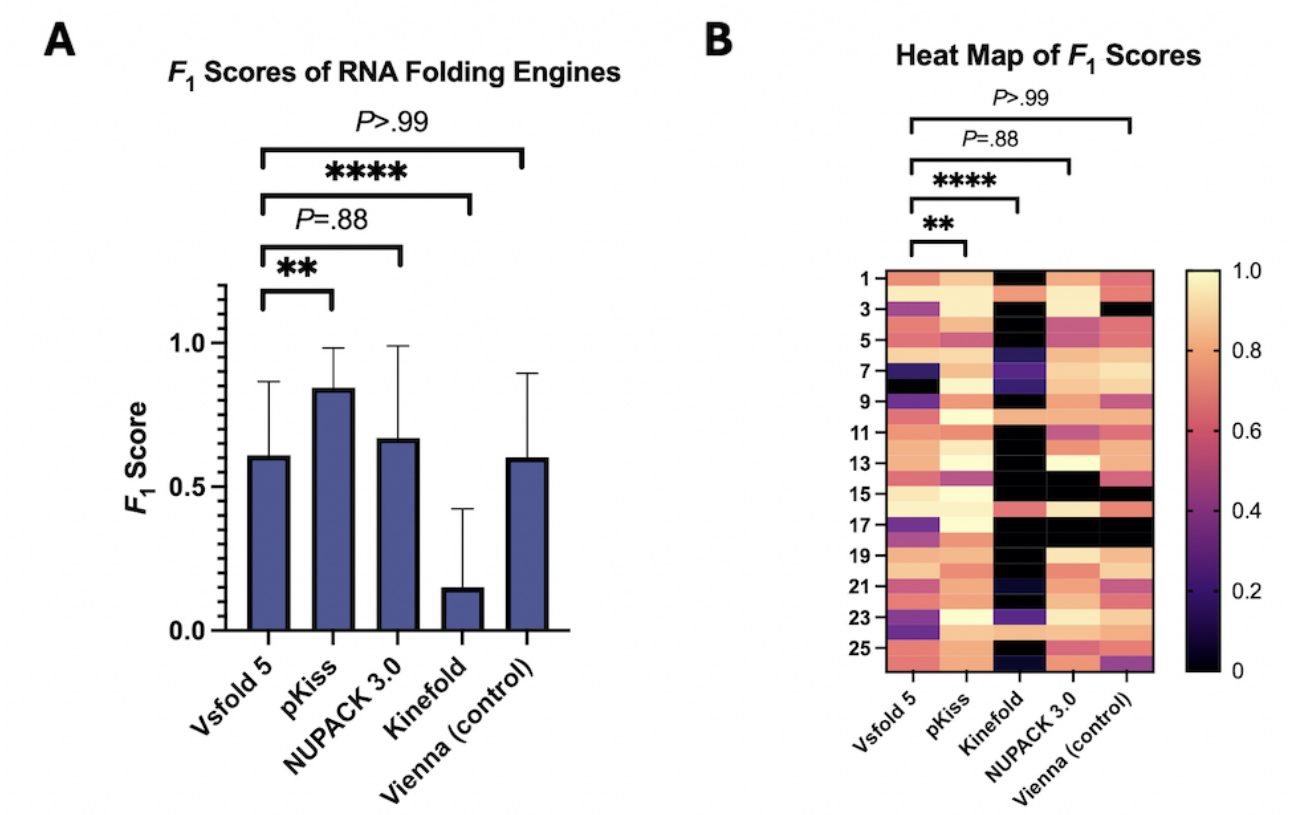
Figure 6. Youden index of RNA folding engines on pseudoknotted viruses. (A) Youden index of raw values. Data corresponding to the mean (SD) across all 5 experimental conditions. Row statistics were performed, alongside a 1-sample t test and Wilcoxon test (t4=0.7354). Data passed normality (gaussian) and logarithmic tests (which included the Shapiro-Wilk test and the Kolmogorov-Smirnov test). *P<.332, ****P<.001. (B) Graph displaying normalized data. This entailed the averaging of subcolumns and normalization of the means, where 0% is defined as the smallest mean in each data set and 100% is defined as the largest mean in each data set.



Quality of Prediction Software Assessed via F_1-Scoring

F_1-scores were derived from the sensitivity and PPV values (Figure 7). This modality is often adapted to assess prediction accuracy so long as the reference structure is given. Of all average F_1-scores generated from all 5 stochastic folding algorithms, the pKiss engine computed the largest values (mean 0.844, SD 0.138), while Kinefold computed the lowest values (mean 0.150, SD 0.273). Of all MFE folding algorithms used, pKiss was the only one to significantly outperform the mean F_1-score of the Vsfold 5 MEA engine by a value of 0.235. Outliers were found in Vsfold 5, Kinefold, and NUPACK 3.0, corresponding to large data sets.

Figure 7. F_1-scores of the Pseudobase++ database. (A) Mean (SD) F_1-scores depicted as a bar graph. (B) Individual F_1-scores generated by all software programs depicted as a heatmap, with white equating to the best performance (F_1-score=1) and black equating to the worst performance (F_1-score=0). Data corresponding to the mean (SD) F_1-scores. Mean (SD) F_1-scores across all 5 experimental conditions were compared to those of Vsfold 5, with statistical analysis performed using 2-way ANOVA, followed by a Tukey multiple-comparison test. ****P<.001, **P≤.002. The ROUT test was performed to identify outliers (Q=1%). The Kolmogorov-Smirnov test was performed, confirming that pKiss was the only experimental condition to conform to normal (gaussian) distribution (α=.05).



Discussion

Principal Findings

The primary aims of this study were to evaluate how MEA folding modalities compare to MFE folding modalities in the context of pseudoknotted viral RNAs and to see which of the 4 experimental conditions (folding software) provides the most accurate models. Although differences in the percentage error (MAE) of "total base pair" predictions were nonsignificant across all software programs (which was expected), the differences in the percentage error of "knotted base pair" predictions were vastly different across the software programs.

The pKiss MFE folding engine exhibited the highest prediction accuracy across all MFE and MEA folding engines and across all performance metrics used (percentage error of total base pairs and knotted base pairs, MSE, sensitivity, PPV, Youden index, F_1-scores), outperforming the MEA folding software Vsfold 5 on all accounts. In contrast, Kinefold exhibited the lowest values for sensitivity, PPV, Youden index, and F_1-scores, even when compared to the control. We have provided evidence suggesting that MEA software is not always the optimal method of topological prediction when applied to short viral pseudoknotted RNAs.

Origins of Stochastic RNA Folding Engines

The underlying functions and computational modalities for RNA prediction algorithms have greatly evolved since Nussinov's dynamic programming algorithm [55], a formalism derived in 1978 and arguably the genesis of predictive RNA folding, whereby:

$$V_{(i,j)} = \max\begin{cases} V(i,j-1) & S_j \text{ unpaired} \\ \max_{(i\le k<(j-1)} V(i,k-1)+V(k+1,j-1)+1 & \text{if } V_kV_j \text{ complement base pair} \end{cases} \quad (7)$$

Here,

$$V(i,j) = \begin{cases} 1 \text{ if i and j can pair} \\ 0 \text{ otherwise} \end{cases} \quad (8)$$

Using these systems, it is possible to generate MFE predictions in kcal/mol of a structure via tracebacking, resulting in a probabilistic model of the nature of the RNA template in vivo. This original scheme has been amended many times over, with Zuker's algorithm [56] being the most notable change—an amendment that is still used to this day.

The realization of these formalisms coincides with the discovery of the first pseudoknotted plant viruses discovered in the early 1980s. Today, they make up many of the pseudoknots found in various online databases/databanks and are recognized as common motifs that allow for viral mRNA function, ribosome function, and replication [6]. The inherent vastness of viral RNA in nature and, consequently, within pseudoknot databases (like

those taken from PseudoBase++ [23] and the RNA Secondary Structure and Statistical Analysis Database (RNA STRAND [57]) was the primary rationale for this investigation.

Evaluation of Percentage Error and MSE Performance Metrics

MEA software outperformed MFE software. The first metric used to evaluate the accuracy of the 5 RNA folding engines was the percentage error of total base pairs and knotted base pairs. These values were computed first by considering all total base pairs in the given model (purple) and then again by considering only pseudoknotted base pairs (blue), as shown in Figure 4A. To reiterate, the difference in percentage error between the total number of base pairs computed by all RNA folding algorithms was nonsignificant (as expected), while pKiss (an MFE model) resulted in a lower percentage error for knotted bases (mean 22.37%, SD 24.2%) when compared to its MEA counterpart Vsfold 5 (mean 69.91%, SD 39.3%). Although some might assume that the lower percentage error exhibited by pKiss could be the result of the pseudoknot "enforce" constraint embedded in the software, it is more likely that this outcome was multivariable, equating to the Turner energy model used and the sensitive auxiliary parameters enforced by the program (refer to Table 1 and Section S2 in Multimedia Appendix 1). This was later tested and proven, showing that small changes within the auxiliary parameters resulted in drastic changes in pseudoknot prediction accuracy (refer to Figure S2 in Multimedia Appendix 1).

The MSE was additionally used, as it is a commonplace tool for assessing predictive models (especially those models that incorporate continuous variables), as shown in Figure 4B. A prerequisite for applying the MSE is that observations are normally distributed [58], which was confirmed using the Kolmogorov-Smirnov test. It was found that for knotted bases, Vsfold 5 resulted in the highest MSE value (382.29), while pKiss resulted in the lowest MSE value (48.4). This greatly buttresses the claim that MFE software may, at times, result in a more favorable model, especially when considering the MSE's sensitivity to outliers.

Evaluation of Sensitivity, PPV, and Youden Index Performance Metrics

With all 26 viral RNAs considered, the highest mean sensitivity and PPV were derived from pKiss, with values of 0.88 (SD 0.14) and 0.82 (SD 0.16), respectively. These values trump those of prior versions, reporting lesser sensitivity (mean 0.80, SD 0.24) and PPV (mean 0.75, SD 0.27) [24].

It is important to emphasize the increase in both the PPV and sensitivity of the newer folding engines listed in Table 1 when compared to older folding engines previously reported in the literature. Such examples include ProbKnot, with a mean sensitivity of 0.693 and a mean PPV of 0.613 [54], and PKNOTS (the older version of pKiss), with older papers reporting a mean sensitivity of 0.828 and a mean PPV of 0.789 and newer papers reporting a mean sensitivity of 0.855 and a mean PPV of 0.808 [59]. Papers promoting RNA software that shares MEA and MFE properties, such as BiokoP dating back 5 years prior, have also reported lesser sensitivity (mean 0.81, SD 0.22) and PPV (mean 0.75, SD 0.26) values [24]. Note that confounding variables between this paper and the referenced literature are minimal, as all reports screened for a diverse set of RNA pseudoknots.

Concerning J, we saw that pKiss continued to show the highest mean value of 0.713 (SD 0.267) for raw values (Figure 6A). Though this metric is similar to the F_1-score, in the sense that it incorporates sensitivity and the PPV, it optimally predicts the probability cutoff by increasing the difference between TP and FP rates. This system of measurement remains inherently more sensitive than F_1-scoring, with minimums and maximums ranging from 1 to –1 rather than from 1 to 0 [60].

Regarding the raw data set, we know that a value of 0 indicates that the experimental variable tested (eg, folding software) has no diagnostic value, while a value of 1 indicates a perfect test, yielding no FPs or FNs. Therefore, pKiss, the MFE model, outperformed Vsfold 5, the MEA model, by a significant amount (mean 0.713, SD 0.267, vs mean 0.241, SD 0.516), enforcing the idea that MEA models are not always the optimal method for topological prediction. It should also be noted that J values were normalized and put into graphical format (Figure 6B) for visual clarification via elimination of negative values (as negative values for J are not defined).

Quality of Prediction Software Assessed via F_1-Scoring

Among the F_1-scores (Figure 7), the pKiss MFE folding engine computed the most promising values (mean 0.844, SD 0.138), significantly outperforming the mean F_1-score computed by the Vsfold 5 MEA engine by a value of 0.235. These values exceed those of previously reported folding engines, such as CCJ and ProbKnot with mean F_1-scores of 0.644 and 0.738, respectively, while, at the same time, underperforming when compared to deep learning algorithms, such as ATTfold with a mean F_1-score of 0.966 [46,61]. Moreover, in all the 5 RNA folding engines assessed, the F_1-score derived from pKiss was the most consistent by far (most symmetrical, with the least amount of skewness), being the only one to have passed the Kolmogorov-Smirnov test for gaussian distribution (α=.05).

It is important to keep in mind that minor improvements have been made to previous MFE reports, although they still underperform existing deep learning and machine learning algorithms.

Of the metrics that were used under the remits of this investigation, none demonstrated that the MEA algorithm used, Vsfold 5, was inherently superlative to its MFE counterparts. This suggests that some short viral pseudoknotted RNAs (20-150 nt) may often result in their lowest free-energy model (granted that salinity, Mg^{2+} concentration, and other environmental variables remain constant). This conclusion is shared among viral pseudoknotted RNAs that are of different structures (eg, H-type, LL-type) and of different motifs (eg, viral tRNA-like, viral 3 UTR).

It should be noted that the thermodynamics within the cell, as well as the many auxiliary folding pathways of RNA, become muddled when the extensive cellular environment is explored

in vivo. This is where ab initio and comparative approaches come into play. However, one should consider that in silico studies will often lack direct cellular relevance, as researchers remain aloof to the broader physiological consequences of change [5]. Therefore, it should be emphasized that both in vitro and in silico approaches are necessary to explore the nature of viral RNAs.

Limitations

Error is inherently fixed to in silico predictions such as these. In nature, kinetic barriers, environmental conditions, and other factors may influence RNA folding intermediates, resulting in a physiologically favored RNA that does not coincide with the predicted results, even if these natural factors are accounted for. Though the methodologies provide novel data to better help us understand genus 1, short (20-150 nt) viral pseudoknotted RNAs, this is a niche subsection of plant and animal viromes. This confers a limited use, should one endeavor to use the RNA folding prediction software on larger, more heterogeneous data sets.

Finally, those who have intellectual rights to these RNA folding web servers [21,26-30] can amend the software and change the parameters of whatever in silico modus they are using for the sake of improvement/refinement/issuing better predictions. This limits the reproducibility of this work, should someone wish to input the same values/predict the same structures used in this investigation.

Future Directions

The conclusions derived from this report further the understanding of how to predict the tertiary, 3D conformation of viral pseudoknotted RNAs (under the context of MEA and MFE prediction software). However, further work could certainly be conducted in this field, which could be brought about by either expanding the current data set or providing further analysis of the already established data set.

Addressing the former point, by softening the exclusion criteria (see Figure S1 in Multimedia Appendix 1), either by allowing more lengthy pseudoknots (nucleotides>150), allowing for a broader scope of MFE structures (MFE<–60 kcal/mol), or assaying a wider variety of RNA classes (rather than just viral tRNA-like, viral 3 UTR, and viral frameshifts), we can increase our understanding of how these moieties behave. Additionally, other RNA data sets, such as RNA STRAND [57], could be adopted in this investigation.

Addressing the latter point, further "prediction accuracy metrics" exist within computational biology that could bolster the claims this investigation made. One metric that the authors advocate for is the *M*-score (also known as the macroaveraged F_1-score), which is a form of weighted F_1-score. It is calculated by summing all F_1-scores for a data set with n classes and then dividing the total by n [62].

$$F_1\text{-score} = \frac{\sum_{i=1}^{n} F1\text{-score}_i}{n} \qquad (9)$$

This macroaveraged F_1-score is most applicable when the data set in question has equal amounts of data points, for each class n (which, in this investigation, happened to be the case). Yet, data sets found in the real world often comprise skewed data, are class-imbalanced, and can encompass nonnormalized data.

In these cases, where there are limited samples in a small class n or where the data are imbalanced, when performing binary classification evaluation, a more appropriate tool to use would be the Mathews correlation coefficient (MCC) [63,64]. The following equation (considering 3 of the 4 confusion matrices, analogous to F_1-scoring):

$$\text{MCC} = \frac{\text{TN} \times \text{TP} - \text{FN} \times \text{FP}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \in [-1,1], \qquad (10)$$

serves to measure the relationship between predicted values and real values (TN refers to true negative). Following the Pearson correlation coefficient directly [64], this metric is commonplace in bioinformatics and has been benchmarked to a wide array of open source data sets within the literature. When applied to either imbalanced support vector machine (SVM) or the MCC-Bayes data sets, this tool offers a good balance between training time and computational efficacy.

Conclusion

To the best of our knowledge, this paper is the first attempt at applying a suite of RNA folding engines to a data set solely comprising viral pseudoknotted RNAs. The data computed in this paper were founded upon different MEA and MFE software programs that have received updates in recent years, and the accuracy of these RNA folding engines was benchmarked following Mathews' parameters. The evidence provided suggests that viral pseudoknotted RNAs may conform to the MFE structure in some cases, rather than the MEA structure. Under the scope of these quality folding engines, pKiss provided the most accurate structures when compared to data experimentally derived from mutagenesis, sequence comparison, structure probing, and NMR, while Kinefold resulted in the least accurate structures. This indicates that the veracity of the underlying thermodynamic model parameters (eg, Turner model, Jacobson-Stockmayer model) is compromised if the auxiliary parameters are not enforced (eg, Mg^{2+} binding, dangling end options, H-type penalties).

To expedite the screening of RNAs, whether they are knotted or planer, we must achieve a better understanding of the thermodynamics associated with cellular processes and how they govern the shaping of RNA. The explored ab initio methodologies provide more accurate results than previously reported, though they do not outperform deep learning algorithms. The exploration of RNA outside the wet lab might seem counterintuitive; however, the computing power we now possess lends to efficacious predictions. Limitations are present in both in vitro and in silico methodologies, leading to the conclusion that both are necessary to further the exploration of drug targets, mRNA vaccines, thermosensors, and RNA-based genome editing.

Acknowledgments

We would like to thank Ajay Singh, Jamie M Robertson, and the Effective Writing for Healthcare program (Harvard Medical School) for reviewing and editing this work. We thank the Das and Barna laboratories (Stanford University), as well as the entirety of the EteRNA community for their contributions to the project. We would also like to thank Henri Orland for providing supplemental aid to the content of the paper. Finally, we would like to acknowledge the Eterna OpenKnot labs project for inspiring this work and the National Instate of Health for generating its Pseudokbase++ database. This research received no external funding.

Data Availability

The data analyzed in this study are present within the main paper, as well as Multimedia Appendix 1.

Authors' Contributions

Conceptualization, methodology, and writing—original draft preparation were handled by VM; formal analysis and data curation by VM and MC; and writing—review and editing by VM, SZ, and JP. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary findings related to the manuscript.
[DOC File , 9458 KB-Multimedia Appendix 1]

References

1. Lewin AS, Hauswirth WW. Ribozyme gene therapy: applications for molecular medicine. Trends Mol Med. May 2001;7(5):221-228. [doi: 10.1016/s1471-4914(01)01965-7] [Medline: 11325634]
2. Walter NG, Engelke DR. Ribozymes: catalytic RNAs that cut things, make things, and do odd and useful jobs. Biologist (London). Oct 2002;49(5):199-203. [FREE Full text] [Medline: 12391409]
3. Zuber J, Schroeder SJ, Sun H, Turner DH, Mathews DH. Nearest neighbor rules for RNA helix folding thermodynamics: improved end effects. Nucleic Acids Res. May 20, 2022;50(9):5251-5262. [FREE Full text] [doi: 10.1093/nar/gkac261] [Medline: 35524574]
4. Wu L, Belasco JG. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. Mol Cell. Jan 18, 2008;29(1):1-7. [FREE Full text] [doi: 10.1016/j.molcel.2007.12.010] [Medline: 18206964]
5. Leamy KA, Assmann SM, Mathews DH, Bevilacqua PC. Bridging the gap between in vitro and in vivo RNA folding. Q Rev Biophys. Jan 2016;49:e10. [FREE Full text] [doi: 10.1017/S003358351600007X] [Medline: 27658939]
6. Brierley I, Pennell S, Gilbert RJC. Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nat Rev Microbiol. Aug 2007;5(8):598-610. [FREE Full text] [doi: 10.1038/nrmicro1704] [Medline: 17632571]
7. Xayaphoummine A, Bucher T, Thalmann F, Isambert H. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. Proc Natl Acad Sci U S A. Dec 23, 2003;100(26):15310-15315. [FREE Full text] [doi: 10.1073/pnas.2536430100] [Medline: 14676318]
8. Pu H, Li J, Li D, Han C, Yu J. Identification of an internal RNA element essential for replication and translational enhancement of tobacco necrosis virus A(C). PLoS One. Feb 27, 2013;8(2):e57938. [FREE Full text] [doi: 10.1371/journal.pone.0057938] [Medline: 23460916]
9. Taylor JM. Structure and replication of hepatitis delta virus RNA. In: Casey JL, editor. Hepatitis Delta Virus. Berlin, Heidelberg. Springer; 2006:1-23.
10. Tinoco I, Bustamante C. How RNA folds. J Mol Biol. Oct 22, 1999;293(2):271-281. [doi: 10.1006/jmbi.1999.3001] [Medline: 10550208]
11. Nikolova EN, Zhou H, Gottardo FL, Alvey HS, Kimsey IJ, Al-Hashimi HM. A historical account of Hoogsteen base-pairs in duplex DNA. Biopolymers. Dec 2013;99(12):955-968. [FREE Full text] [doi: 10.1002/bip.22334] [Medline: 23818176]
12. Gernot A, Jinho B, Philippe DF, Henri O, Graziano V. The Oxford Handbook of Random Matrix Theory. Oxford, UK. Oxford University Press; 2011:872-897.
13. Lim CS, Brown CM. Know your enemy: successful bioinformatic approaches to predict functional RNA structures in viral RNAs. Front Microbiol. Jan 4, 2017;8:2582. [FREE Full text] [doi: 10.3389/fmicb.2017.02582] [Medline: 29354101]
14. Mans RM, Pleij CW, Bosch L. tRNA-like structures. Structure, function and evolutionary significance. Eur J Biochem. Oct 15, 1991;201(2):303-324. [FREE Full text] [doi: 10.1111/j.1432-1033.1991.tb16288.x] [Medline: 1935928]

15. Felden B, Florentz C, Giegé R, Westhof E. Solution structure of the 3'-end of brome mosaic virus genomic RNAs. Conformational mimicry with canonical tRNAs. J Mol Biol. Jan 1994;235(2):508-531. [doi: 10.1006/jmbi.1994.1010] [Medline: 8289279]
16. Lai D, Proctor J, Zhu J, Meyer I. R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic Acids Res. Jul 2012;40(12):e95. [FREE Full text] [doi: 10.1093/nar/gks241] [Medline: 22434875]
17. Byun Y, Han K. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. Bioinformatics. Jun 01, 2009;25(11):1435-1437. [doi: 10.1093/bioinformatics/btp252] [Medline: 19369500]
18. Larson SB, Lucas RW, McPherson A. Crystallographic structure of the T=1 particle of brome mosaic virus. J Mol Biol. Feb 25, 2005;346(3):815-831. [FREE Full text] [doi: 10.1016/j.jmb.2004.12.015] [Medline: 15713465]
19. Choudhary K, Deng F, Aviran S. Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. Quant Biol. Mar 2017;5(1):3-24. [FREE Full text] [doi: 10.1007/s40484-017-0093-6] [Medline: 28717530]
20. Lavender CA, Gorelick RJ, Weeks KM. Structure-based alignment and consensus secondary structures for three HIV-related RNA genomes. PLoS Comput Biol. May 2015;11(5):e1004230. [FREE Full text] [doi: 10.1371/journal.pcbi.1004230] [Medline: 25992893]
21. Janssen S, Giegerich R. The RNA shapes studio. Bioinformatics. Feb 01, 2015;31(3):423-425. [FREE Full text] [doi: 10.1093/bioinformatics/btu649] [Medline: 25273103]
22. Bindewald E, Shapiro BA. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA. Mar 2006;12(3):342-352. [FREE Full text] [doi: 10.1261/rna.2164906] [Medline: 16495232]
23. Taufer M, Licon A, Araiza R, Mireles D, van Batenburg FHD, Gultyaev AP, et al. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. Nucleic Acids Res. Jan 2009;37(Database issue):D127-D135. [FREE Full text] [doi: 10.1093/nar/gkn806] [Medline: 18988624]
24. Legendre A, Angel E, Tahi F. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. BMC Bioinform. Jan 15, 2018;19(1):13. [FREE Full text] [doi: 10.1186/s12859-018-2007-7] [Medline: 29334887]
25. Sato K, Kato Y. Prediction of RNA secondary structure including pseudoknots for long sequences. Brief Bioinform. Jan 17, 2022;23(1):bbab395. [FREE Full text] [doi: 10.1093/bib/bbab395] [Medline: 34601552]
26. Dawson WK, Fujiwara K, Kawai G. Prediction of RNA pseudoknots using heuristic modeling with mapping and sequential folding. PLoS One. Sep 19, 2007;2(9):e905. [FREE Full text] [doi: 10.1371/journal.pone.0000905] [Medline: 17878940]
27. Xayaphoummine A, Bucher T, Isambert H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. Nucleic Acids Res. Jul 01, 2005;33(Web Server issue):W605-W610. [FREE Full text] [doi: 10.1093/nar/gki447] [Medline: 15980546]
28. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, et al. NUPACK: analysis and design of nucleic acid systems. J Comput Chem. Jan 15, 2011;32(1):170-173. [doi: 10.1002/jcc.21596] [Medline: 20645303]
29. Fornace ME, Huang J, Newman CT, Porubsky NJ, Pierce MB, Pierce NA. NUPACK: analysis and design of nucleic acid structures, devices, and systems. ChemRxiv preprint. Preprint posted online November 10, 2022. [doi: 10.26434/chemrxiv-2022-xv98l]
30. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem. Feb 1994;125(2):167-188. [doi: 10.1007/bf00818163]
31. Wayment-Steele HK, Kladwang W, Strom AI, Lee J, Treuille A, Becka A, Eterna Participants, et al. RNA secondary structure packages evaluated and improved by high-throughput experiments. Nat Methods. Oct 03, 2022;19(10):1234-1242. [FREE Full text] [doi: 10.1038/s41592-022-01605-0] [Medline: 36192461]
32. Zammit A, Helwerda L, Olsthoorn RCL, Verbeek FJ, Gultyaev AP. A database of flavivirus RNA structures with a search algorithm for pseudoknots and triple base interactions. Bioinformatics. May 17, 2021;37(7):956-962. [FREE Full text] [doi: 10.1093/bioinformatics/btaa759] [Medline: 32866223]
33. Mlera L, Melik W, Bloom ME. The role of viral persistence in flavivirus biology. Pathog Dis. Jul 2014;71(2):137-163. [FREE Full text] [doi: 10.1111/2049-632X.12178] [Medline: 24737600]
34. Creager ANH. Tobacco mosaic virus and the history of molecular biology. Annu Rev Virol. Sep 29, 2022;9(1):39-55. [FREE Full text] [doi: 10.1146/annurev-virology-100520-014520] [Medline: 35704746]
35. Du Z, Holland JA, Hansen MR, Giedroc DP, Hoffman DW. Base-pairings within the RNA pseudoknot associated with the simian retrovirus-1 gag-pro frameshift site. J Mol Biol. Jul 18, 1997;270(3):464-470. [doi: 10.1006/jmbi.1997.1127] [Medline: 9237911]
36. Zwieb C, Samuelsson T. SRPDB (Signal Recognition Particle Database). Nucleic Acids Res. Jan 01, 2000;28(1):171-172. [FREE Full text] [doi: 10.1093/nar/28.1.171] [Medline: 10592215]
37. De Rijk P, Robbrecht E, de Hoog S, Caers A, Van de Peer Y, De Wachter R. Database on the structure of large subunit ribosomal RNA. Nucleic Acids Res. Jan 01, 1999;27(1):174-178. [FREE Full text] [doi: 10.1093/nar/27.1.174] [Medline: 9847172]
38. Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T, De Wachter R. The European Small Subunit Ribosomal RNA database. Nucleic Acids Res. Jan 01, 2000;28(1):175-176. [FREE Full text] [doi: 10.1093/nar/28.1.175] [Medline: 10592217]

39. Peselis A, Serganov A. Structure and function of pseudoknots involved in gene expression control. Wiley Interdiscip Rev RNA. 2014;5(6):803-822. [FREE Full text] [doi: 10.1002/wrna.1247] [Medline: 25044223]
40. Lucas A, Dill KA. Statistical mechanics of pseudoknot polymers. J Chem Phys. Jul 22, 2003;119(4):2414-2421. [doi: 10.1063/1.1587129]
41. Chiu JKH, Chen YP. Conformational features of topologically classified RNA secondary structures. PLoS One. Jul 5, 2012;7(7):e39907. [FREE Full text] [doi: 10.1371/journal.pone.0039907] [Medline: 22792195]
42. Ferré-D'Amaré AR, Zhou K, Doudna JA. Crystal structure of a hepatitis delta virus ribozyme. Nature. Oct 08, 1998;395(6702):567-574. [doi: 10.1038/26912] [Medline: 9783582]
43. Tomita K, Ishitani R, Fukai S, Nureki O. Complete crystallographic analysis of the dynamics of CCA sequence addition. Nature. Oct 26, 2006;443(7114):956-960. [doi: 10.1038/nature05204] [Medline: 17051158]
44. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. Oct 2009;62(10):e1-e34. [FREE Full text] [doi: 10.1016/j.jclinepi.2009.06.006] [Medline: 19631507]
45. Mathews DH. How to benchmark RNA secondary structure prediction accuracy. Methods. Jun 01, 2019;162-163:60-67. [FREE Full text] [doi: 10.1016/j.ymeth.2019.04.003] [Medline: 30951834]
46. Wang Y, Liu Y, Wang S, Liu Z, Gao Y, Zhang H, et al. ATTfold: RNA secondary structure prediction with pseudoknots based on attention mechanism. Front Genet. 2020;11:612086. [FREE Full text] [doi: 10.3389/fgene.2020.612086] [Medline: 33384721]
47. An J, Meng F, Yan Z. An efficient computational method for predicting drug-target interactions using weighted extreme learning machine and speed up robot features. BioData Min. Jan 20, 2021;14(1):3. [FREE Full text] [doi: 10.1186/s13040-021-00242-1] [Medline: 33472664]
48. Sun Y, Liu F, Fan C, Wang Y, Song L, Fang Z, et al. Characterizing sensitivity and coverage of clinical WGS as a diagnostic test for genetic disorders. BMC Med Genomics. Apr 13, 2021;14(1):102. [FREE Full text] [doi: 10.1186/s12920-021-00948-5] [Medline: 33849535]
49. Fu L, Cao Y, Wu J, Peng Q, Nie Q, Xie X. UFold: fast and accurate RNA secondary structure prediction with deep learning. Nucleic Acids Res. Feb 22, 2022;50(3):e14. [FREE Full text] [doi: 10.1093/nar/gkab1074] [Medline: 34792173]
50. Sikka J, Satya K, Kumar Y, Uppal S, Shah R, Zimmermann R. Learning based methods for code runtime complexity prediction. 2020. Presented at: Advances in Information Retrieval: 42nd European Conference on IR Research (ECIR 2020); April 14–17, 2020:313-325; Lisbon, Portugal. [doi: 10.1007/978-3-030-45439-5_21]
51. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. Aug 2004;10(8):1178-1190. [FREE Full text] [doi: 10.1261/rna.7650904] [Medline: 15272118]
52. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. Jul 15, 2006;22(14):e90-e98. [FREE Full text] [doi: 10.1093/bioinformatics/btl246] [Medline: 16873527]
53. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. Oct 20, 1998;37(42):14719-14735. [doi: 10.1021/bi9809425] [Medline: 9778347]
54. Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA. Oct 2010;16(10):1870-1880. [FREE Full text] [doi: 10.1261/rna.2125310] [Medline: 20699301]
55. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. Algorithms for loop matchings. SIAM J Appl Math. Jul 1978;35(1):68-82. [doi: 10.1137/0135006]
56. Zuker M, Sankoff D. RNA secondary structures and their prediction. Bull Math Biol. 1984;46(4):591-621. [doi: 10.1016/s0092-8240(84)80062-2]
57. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: the RNA Secondary Structure and Statistical Analysis Database. BMC Bioinform. Aug 13, 2008;9(1):340. [FREE Full text] [doi: 10.1186/1471-2105-9-340] [Medline: 18700982]
58. Holst E, Thyregod P. A statistical test for the mean squared error. J Stat Comput Simul. Jul 1999;63(4):321-347. [doi: 10.1080/00949659908811960]
59. Li H, Zhu D, Zhang C, Han H, Crandall KA. Characteristics and prediction of RNA structure. Biomed Res Int. 2014;2014:690340. [FREE Full text] [doi: 10.1155/2014/690340] [Medline: 25110687]
60. Andres K. Formulas. In: Laboratory Statistics (Second Edition). Amsterdam, the Netherlands. Elsevier; 2018:1-140.
61. Jabbari H, Wark I, Montemagno C. RNA secondary structure prediction with pseudoknots: contribution of algorithm versus energy model. PLoS One. 2018;13(4):e0194583. [FREE Full text] [doi: 10.1371/journal.pone.0194583] [Medline: 29621250]
62. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. Mar 13, 2024;14(1):6086. [FREE Full text] [doi: 10.1038/s41598-024-56706-x] [Medline: 38480847]
63. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. Jan 02, 2020;21(1):6. [FREE Full text] [doi: 10.1186/s12864-019-6413-7] [Medline: 31898477]
64. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One. Jun 2, 2017;12(6):e0177678. [FREE Full text] [doi: 10.1371/journal.pone.0177678] [Medline: 28574989]

Abbreviations

FN: false negative
FP: false positive
MAE: mean absolute error
MCC: Mathews' correlation coefficient
MEA: maximum expected accuracy
MFE: minimum free energy
mRNA: messenger RNA
MSE: mean squared error
PPV: positive predictive value
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNA STRAND: RNA Secondary Structure and Statistical Analysis Database
ROUT: robust outlier
tRNA: transfer RNA
UTR: untranslated region
TP: true positive

Edited by T Leung, A Schwartz; submitted 27.03.24; peer-reviewed by D Saderi, V Nagesh, RSG Mahmoud, T Olatoye, F Qudus Arogundade; comments to author 09.07.24; revised version received 31.08.24; accepted 04.10.24; published 05.11.24

Please cite as:
Medeiros V, Pearl J, Carboni M, Zafeiri S
Exploring the Accuracy of Ab Initio Prediction Methods for Viral Pseudoknotted RNA Structures: Retrospective Cohort Study
JMIRx Bio 2024;2:e58899
URL: https://bio.jmirx.org/2024/1/e58899
doi: 10.2196/58899
PMID: