BMC Bioinformatics

Open Access

ENTRNA: a framework to predict RNA foldability

Congzhe Su[1], Jeffery D. Weir[2], Fei Zhang[3], Hao Yan[3] and Teresa Wu[1*]

Abstract

Background: RNA molecules play many crucial roles in living systems. The spatial complexity that exists in RNA structures determines their cellular functions. Therefore, understanding RNA folding conformations, in particular, RNA secondary structures, is critical for elucidating biological functions. Existing literature has focused on RNA design as either an RNA structure prediction problem or an RNA inverse folding problem where free energy has played a key role.

Results: In this research, we propose a Positive-Unlabeled data- driven framework termed ENTRNA. Other than free energy and commonly studied sequence and structural features, we propose a new feature, Sequence Segment Entropy (SSE), to measure the diversity of RNA sequences. ENTRNA is trained and cross-validated using 1024 pseudoknot-free RNAs and 1060 pseudoknotted RNAs from the RNASTRAND database respectively. To test the robustness of the ENTRNA, the models are further blind tested on 206 pseudoknot-free and 93 pseudoknotted RNAs from the PDB database. For pseudoknot-free RNAs, ENTRNA has 86.5% sensitivity on the training dataset and 80.6% sensitivity on the testing dataset. For pseudoknotted RNAs, ENTRNA shows 81.5% sensitivity on the training dataset and 71.0% on the testing dataset. To test the applicability of ENTRNA to long structural-complex RNA, we collect 5 laboratory synthetic RNAs ranging from 1618 to 1790 nucleotides. ENTRNA is able to predict the foldability of 4 RNAs.

Conclusion: In this article, we reformulate the RNA design problem as a foldability prediction problem which is to predict the likelihood of the co-existence of a sequence-structure pair. This new construct has the potential for both RNA structure prediction and the inverse folding problem. In addition, this new construct enables us to explore data-driven approaches in RNA research.

Keywords: Data-driven, Foldability, Sequence segment entropy

Background

Ribonucleic acid (RNA), as an emerging nanoscale building block, is regarded as one of the most promising candidates to create nano-architectures and nano-devices for therapeutic and diagnostic purposes. Due to its unique biochemical properties and functionalities [1], such as catalysis of metabolic reactions [2], regulation of gene expression [3], and organization of proteins into large machineries [4], RNA has attracted great attention from both academia and industry resulting in broad applications. For example, the success in clinical trials has proved that RNA-based therapeutics hold great potential

to overcome the limitation of existing medicine that can only target a limited number of proteins [5]. To fully explore and utilize RNA functions, the cornerstone is to study the multi-levels of complicated RNA structures to include the linear ribonucleotide sequence (primary structure), the 2D fold based on canonical Watson-Crick and wobble base-pairings (secondary structure), the 3D fold (tertiary structure), and the complex spatial arrangement of multiple folded molecules (quaternary structure) [6]. The folding of RNA molecules is broadly considered as a hierarchical process in which the secondary structure will be folded first representing the most relevant characteristic of an RNA molecule [7]. Therefore, studying the RNA secondary structure is one of the fundamental steps towards understanding function-related RNA structures.

* Correspondence: Teresa.Wu@asu.edu
[1]School of Computing, Informatics, Decision Systems Engineering, Arizona State University, Tempe, AZ 85281, USA
Full list of author information is available at the end of the article

Su *et al. BMC Bioinformatics*   (2019) 20:373

Page 2 of 11

In general, RNA secondary structure research falls into two categories: The RNA structure prediction problem, which is to predict the folding result of base pairs given the RNA sequence; and the RNA inverse folding problem, which is to identify the appropriate assignment of nucleotides so that a targeted RNA secondary structure can be folded with certainty. For the RNA structure prediction problem, researchers have developed a variety of computational approaches to increase the prediction accuracy. One early effort is to use the comparative approach to infer a consensus secondary structure by aligning the given sequence with other existing RNA sequences. This requires large collections of RNA sequences for the analysis. A major challenge of this approach is the limited availability of RNA [8]. An alternative is using a thermodynamic model to predict the secondary structure, which is based on the assumption that a structure with smaller free energy tends to be more stable. Therefore, an optimization problem with the objective being to minimize the free energy is constructed to identify the structures with minimum free energy (MFE). A number of research tools have been developed to serve this purpose. One tool is Mfold [9]. It employs a dynamic programming algorithm to predict the RNA secondary structure with MFE. While promising, the prediction accuracy of Mfold is less than satisfactory leading to some research efforts to improve its performance. For example, RNAstructure [8] incorporates the constraints from experimental data to improve the prediction accuracy. Realizing the uncertainties in the folding process, RNAfold [10] provides the estimated probabilities of base pairs. For the RNA inverse folding problem, the objective is to identify the appropriate sequences minimizing the distance metric (e.g., the number of common base pairs) between the structure folded from the designed sequence to the target secondary structure. One of the first tools is RNAinverse [11]. In RNAinverse, a random sequence is generated, changes of the nucleotide assignment are made locally to minimize the dissimilarities between the structures. Apparently, such a local search strategy may be trapped in a local optimum and the designed sequences are highly depended on the initial seed solution. To address this issue, RNA-SSD [12] is proposed to assign initial bases probabilistically attempting to avoid local trapping. incaRNAtion [13] uses global sampling and weighted sampling techniques to avoid the seed bias in local search. In antaRNA [14], ant colony optimization, an efficient bio-inspired optimization algorithm is implemented to expedite the searching process with high accuracy. All of the algorithms reviewed assume the designed sequence will fold into the MFE structure, which will be used to calculate the distance to a target secondary structure.

As noted, previous research in both structure prediction and inverse folding has heavily relied on free energy as the metric to evaluate the stability of RNA structures [9–16]. The hypothesis here is, given an RNA sequence, the secondary structure with the MFE will be the stable structure which it would fold into with highest likelihood and thus is considered "optimum"; and given a structure, the sequence shall be assigned with nucleotides in the way that MFE is achieved. To test the hypothesis, we started by collecting 167 existing pseudoknot-free RNA sequences from the Protein Data Bank (PDB), it is observed that only 53 RNAs (32%) are in MFE secondary structures. This finding indicates MFE alone may not be a sufficient condition in guiding RNA design. In other words, not all existing RNA structures are folded with the energy being MFE. Often, RNA can still be folded at an energy level close to MFE, we call them suboptimal RNAs. As indicated in Laing [6], RNA may have a large number of alternative suboptimum folding which is known as the multi-conformation RNA issue.

Recognizing the limitations from MFE algorithms, some research has proposed to generate a set of possible structures with near-optimal free energy instead of the MFE secondary structure alone. For example, RNAsubopt provides all the secondary structures within δ difference from the MFE [28]. However, the number of possible structures grows exponentially with the increment of different δ. Others have developed alternative metrics calculated from partition functions to evaluate the accessibility of the possible secondary structures. These include IPknot, Sfold [29], RNAshapes [30] and RNA profiling [31]. However, although efforts in the field have focused on exploring different metrics, researchers have not reached the consensus on which metrics should be broadly adopted.

In this research, we introduce a new concept: RNA foldability. Let the RNA structure prediction problem be considered as sequence → structure*, and the RNA inverse folding problem be considered as the structure → sequence*. Our foldability is defined as $l(structure, sequence)$, which measures the likelihood of the co-existence of the structure – sequence pair. One motivation of developing this new construct is it can be potentially applied to both the structure prediction and inverse folding problems. For example, given a sequence, a number of possible structures could be folded, foldability $l(structure, sequence)$ can be used to identify the structure with high likelihood. For an inverse folding problem, a number of sequencing candidates can be first identified for a targeted structure, again, foldability $l(structure, sequence)$ here can be used to identify the sequence most likely to fold into the structure. A second motivation of this foldability concept is it enables us to

Su *et al. BMC Bioinformatics* (2019) 20:373

Page 3 of 11

explore data-driven approaches to RNA research. By extracting features from both sequence and structure, multi-parametric machine learning models can be developed to obtain the foldability measures. To achieve this, in conjunction with free energy and other commonly used RNA structural design features (e.g., GC content and base pair percentage), we introduce a new metric to evaluate the diversity of RNA sequence segments termed Sequence-Segment entropy (SSE). A Positive-Unlabeled (PU) learning based data driven framework, ENTRNA, is developed using the features to predict RNA foldability. After training on both pseudoknot-free and pseudoknotted RNAs, ENTRNA shows promising accuracy in predicting RNA foldability. Specifically, it successfully identifies 80% pseudoknot-free RNAs and pseudoknotted RNAs can be folded into the desired structures.

There are two main contributions from our proposed ENTRNA. First, RNA design is reformulated as a foldability prediction problem ($l(structure, sequence)$) which can evaluate the successful rate of a given pair of sequence and structure. This new formulation can fundamentally tackle the challenging issues in RNA design, that is, one RNA sequence may fold into multiple structures, and one RNA structure may have multiple sequence assignments. The second contribution lies in the new metric on assessing the RNA sequence segment diversity. In the remainder of the paper, the ENTRNA is presented in Section 2 followed by validation experiments in Section 3. The conclusion and discussion are drawn in Section 4.

Methods

RNA foldability prediction problem

Most existing computational algorithms formulate RNA secondary structure prediction as a deterministic optimization problem which aims to find the global optimal secondary structure for the given sequence. It provides a single best guess for the secondary structure with the assumption that the RNA sequence will only fold into the optimal secondary structure (i.e. MFE secondary structure). Unfortunately, such an assumption has notable limitations as some RNAs (i.e. highly structured ribosomal RNAs) often exist in multiple conformations [17]. Deterministic optimization approaches fail to discover multiple RNA secondary structures.

To address the multi-conformation RNA challenge, we look at RNA design from a different perspective. Specifically, we propose to develop a predictive model to estimate the likelihood $l(structure, sequence)$ of a given RNA sequence folding into a given secondary structure. We call this approach RNA foldability prediction. RNA foldability prediction fundamentally differs from RNA secondary structure prediction and the RNA inverse folding problem, as the later ones only require RNA

sequences or secondary structure as a single input. RNA foldability prediction will require both sequence and secondary structure to be provided. As such, it enables foldability evaluation on one sequence vs. its several potential secondary structures. Similarly, it can be used to evaluate one secondary structure vs. its multiple sequence candidates which is the RNA inverse folding problem.

ENTRNA for RNA foldability prediction

RNA foldability prediction could be regarded as a classification problem. To train a classification model, both successful and failed examples are needed. In the RNA foldability prediction problem, any reported successful synthetic RNA or natural existing RNA can be regarded as a positive example. However, failed RNAs have rarely been reported in the literature. To address this issue, we propose the application of the Positive-Unlabeled Learning technique (PU) to fill in the failed examples. Two different sets of RNA features are defined and extracted for pseudoknot-free and pseudoknotted RNAs respectively. By mapping RNAs into a length-free feature space, it enables us to fully learn and explore all the existing RNAs together. In addition, a new metric is proposed to evaluate the diversity of RNA sequences (see Section 2.2.2). Together with free energy (see Section 2.2.3), base pairing probability (see Section 2.2.4) and other RNA domain knowledge driven features (Section 2.2.5), ENTRNA is developed as a data-driven framework to predict RNA foldability.

Generate training dataset for PU learning

PU Learning is originally used to solve the text classification problem, which is to assign predefined labels to a new document [18]. Two datasets are needed for training: a positive labeled training set P and an unlabeled mixed set U. The positive set P has the positive examples, the mixed set U is assumed to have both positive and negative examples, but no explicit class label. Generally, PU learning is a two-step approach. First, it identifies a set of reliable negative examples from the mixed set U based on the knowledge of positive set P. Next, it builds predictive models on those positive and "negative" examples iteratively and then selects the best model among them.

In the RNA foldability prediction problem, a pair of existing RNA sequence and its corresponding secondary structure is considered a successful example in the positive training set P. The challenge lies in the unlabeled dataset U as it is not publically available. We decide to generate synthetic RNAs computationally as the examples composing U. The rationale here is the synthetic sequences generated by the computational algorithms are believed to be folded into targeted secondary

structures, yet not empirically validated through lab testing, thus could be treated as part of the unlabeled dataset **U**.

In this research, we use the secondary structures existing in **P** as seeds to generate possible sequences. For a given secondary structure in **P**, instead of randomly assign sequences, we generate a number of possible sequences satisfying three constraints. The first two constraints are the same as in Williams et al. [19]: base pairing and repetition. Base-pairing constraint states only Watson-Crick and G-U base pairs are valid. The repetition constraint sets the longest sequence of bases that can all be the same. For example, if the repetition limit is 4, then AAAA may not appear in the structure, though AAAC can. Given the unique property of RNA folding, the third constraint on GC percentage is added, that is, the minimum and maximum percent of bases in the structure that must be either guanine (G) or cytosine (C). The set of sequences for the given structures consists of our unlabeled dataset **U**.

Next, we apply PU Learning to identify "reliable" negatives from **U**. Note we use "reliable" instead of "true" negatives as there is no ground truth to validate the negatives. We make the assumption "reliable" negatives are the ones furthest from the true positives in **P** which is known as a prior. For simplicity, we propose to use the Euclidean distance of feature values (see sections 2.2.2–2.2.5 for details on the features) to identify these negatives. Normalization has been done to eliminate the scaling issue of different features. Let $f_{u_i, j}$ and $f'_{p_k, j}$ denote the values of feature j for example ui from **U** and example pk from **P** respectively. d_{u_i} is calculated as follows to measure the maximum distance between example u_i to the positive set **P**:

$$d_{u_i} = \max d_{u_i, p_k} \forall p_k \, \epsilon \, P \tag{1}$$

where

$$d_{u_i, p_k} = \sqrt{\sum_{k=1}^{m} \left(f_{u_i, j} - f'_{p_k, j} \right)^2} \tag{2}$$

and *m* is the number of features.

With true positives from **P** and "reliable" negatives from **U**, we are able to develop a classification model (see section 2.2.5) to predict foldability, *l(structure, sequence)* for any pair of structure - sequence.

ENTRNA feature: sequence segment entropy

Due to the incomplete and inaccurate thermodynamic parameters, a great number of RNAs are trapped in the suboptimal structures that are near the predicted global free energy minimum [6]. Meanwhile, the sequence is more likely to be trapped into its suboptimal secondary structures if it has diverse secondary structures.

Therefore, a new metric measuring the secondary structure diversity, is needed in addition to free energy.

Entropy, derived from thermodynamics and information theory [20], is used to measure the amount of uncertainty and disorder within a system. Since its inception, entropy has been applied to a diverse set of research fields including structural RNA research. For example, conformational entropy is considered an important factor in protein-ligand discrimination [21]. Positional entropy is introduced to measure the certainty of being unpaired considering all nucleotides [22]. However, the base pairing probability is required for all the existing entropy-based metrics, which is calculated based on the free energy value. Hence, it is still dependent on thermodynamic parameters and it is not capable for pseudoknotted RNAs. Therefore, a pseudoknotted-RNA capable and thermodynamic parameter free metric is needed to evaluate the structural diversity.

The k-mer concept has been widely used in bioinformatics research. For example, in genome, k-mer has been applied to de novo assembly of large genomes from short read sequences [32] and detecting mis-assemblies [33]. In RNA, Sailfish, a k-mer based algorithm, is developed to quantify the abundance of RNA isoforms [34]. In this research, we introduce sequence segment entropy (SSE) to measure the diversity of RNA sequence segments, which is motivated by the k-mer concept. For generalization, assume an RNA sequence of length *n* nucleotides (nt_1, nt_2, ...,nt_n), let *w* be the segment size referring to the number of consecutive nucleotides in order. To derive the SSE, we need to evaluate the entire RNA sequence. Thus, we use the moving window concept to list the segments. In that case, the segments of the RNA sequence can be written as:

$$\textbf{\textit{Seg}}_w = \left[Seg_w^1, Seg_w^2, ..., Seg_w^{n+1-w} \right],$$

where

$$Seg_w^1 = (nt_1, nt_2, ..., nt_w), Seg_w^2$$
$$= (nt_2, nt_3, ..., nt_{w+1}), Seg_w^{n+1-w}$$
$$= (nt_{n+1-w}, nt_{n+2-w}, ...nt_n).$$

Let $\textbf{\textit{SegU}}_w$ be the set representing the collection of distinct segments, we have

$$\textbf{\textit{SegU}}_w = \left[SegU_w^1, SegU_w^2, ..., SegU_w^s \right], where \, s$$
$$= |\textbf{\textit{SegU}}_w|.$$

Following the entropy calculation, we define $V_{ent,w}$ as:

$$V_{ent,w} = -\sum_{i=1}^{s} p\left(SegU_w^i \right) \log_2 p\left(SegU_w^i \right) \tag{3}$$

where

$$p\left(SegU_w^i\right) = \frac{\#of\ SegU_w^i\ occurence\ in\ \mathbf{Seg_w}}{n+1-w}\ for\ i = 1, ..., s \tag{4}$$

Since the value range of SSE is highly dependent on the length of an RNA sequence, we normalize SSE as $RV_{ent,\ w}$:

$$RV_{ent,w} = \frac{V_{ent,w}}{V_{ent,w}^*} \tag{5}$$

where $V_{ent,w}^*$ is the maximum SSE for segment size w, which is proven to be:

$$V_{ent,w}^* = \begin{cases} -\log_2\left(\frac{1}{n+1-w}\right) & if\ n+1-w \le 4^w \\ -b*\frac{a+1}{n+1-w}*\log_2\left(\frac{a+1}{n+1-w}\right)-(4^w-b)*\frac{a}{n+1-w}*\log_2\left(\frac{a}{n+1-w}\right), & o/w \end{cases} \tag{6}$$

where

$$a = \left\lfloor \frac{n+1-w}{4^w} \right\rfloor,\ \ b = (n+1-w)\ \ mod\ 4^w.$$

[**Proposition 1**] Suppose we have two sequences of the same size with probability density set $\{p_1, p_2, p_3..., p_{n+1-w}\}$ and $\{p_1 + \epsilon, p_2 - \epsilon, p_3, ..., p_{n+1-w}\}$ and $p_1 = p_2 = ... = p_{n+1-w} = p > 0$, $\epsilon > 0$. The first SSE minus the second SSE equals $-plog_2p - plog_2p + (p + \epsilon)\log_2(p + \epsilon) + (p - \epsilon)\log_2(p - \epsilon)$

Since f(x) = $-xlog(x)$ is a concave function, according to Jensen's inequality,

$$\begin{aligned} &\frac{1}{2}\left(\ (p+\epsilon)\ \log_2(p+\epsilon) + (p-\epsilon)\ \log_2(p-\epsilon)\ \right) \\ &= \frac{1}{2}*f(p+\epsilon) + \frac{1}{2}*f(p-\epsilon) \\ &< f\left(\frac{1}{2}*(p+\epsilon) + \frac{1}{2}*(p-\epsilon)\right) \\ &= f(p) = -plog_2p \end{aligned}$$

Hence, the SSE of the first sequence is greater than the second one. Therefore, the sequence segment should be as uniform as possible to achieve the maximum SSE.

[**Proof on maximum SSE**]. The total number of distinct sequence segments with size w is 4^w, since 4 different nucleotides could be assigned to each position arbitrarily. Therefore we have two cases depending on the cardinality of Seg_w.

- In the cases where $n + 1 - w \le 4^w$, the most uniform probability density set will occur when all elements of Seg_w are unique and then each element of $SegU_w$ would have probability $\frac{1}{n+1-w}$.
- In the cases where $n + 1 - w > 4^w$ there must exist elements Seg_w that are not unique. The most

uniform probability density set will occur when Seg_w is partitioned into two groups of segments. The first group of segments will contain in $b = (n + 1 - w)\ mod\ 4^w$ out of 4^w and occur more frequently than the remaining group of $4^w - b$, which occur in equal amounts. For the group occurring in equal amounts, they must occur exactly $a = \lfloor\frac{n+1-w}{4^w}\rfloor$ times giving them a probability of $\frac{a}{n+1-w}$. Therefore, the probability for the b remaining elements must be $\frac{a+1}{n+1-w}$.

Substituting the optimal probability density sets into Eq. (3), we get Eq. (6).

[**Illustration Example on SSE**] Suppose we have two RNA sequences:

seq$_1$ = 'GAAAAAAAAAAAAAAAAAAC'
seq$_2$ = 'GACCGUCGUGAGACAGGUUA'

First, we calculate the scaled sequence segment entropy value of **seq**$_1$, take segment size 3 as an example:

Seg$_3$ = ['GAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAA', 'AAC'];

SegU$_3$ = ['GAA', 'AAA', 'AAC'];

P('GAA') = $\frac{1}{18}$ = 0.056; P('AAA') = $\frac{16}{18}$ = 0.889; P('AAC') = $\frac{1}{18}$ = 0.056;

V$_{ent,3}$ = $-\left(\frac{1}{18}*\log_2\frac{1}{18} + \frac{16}{18}*\log_2\frac{16}{18} + \frac{1}{18}*\log_2\frac{1}{18}\right)$ = 0.614;

a = $\lfloor\frac{(20 + 1 - 3)}{4^3}\rfloor$ = 0;

b = $(20 + 1 - 3)\ mod\ 4^3$ = 18;

V$_{ent,3}^*$ = $-\log_2\frac{1}{18}$ = 4.170;

RV$_{ent,3}$ = $\frac{0.614}{4.170}$ = 0.147;

Following the same steps above, we get RV$_{ent,\ 3}$ of **seq**$_2$ is 0.947. The second sequence (**seq**$_2$) may fold into more possible structures than the first one. This is reflected by scaled segment entropy value. The $RV_{ent,\ 3}$ of first sequence is 0.147, while the value of second sequence is 0.947. The higher scaled segment entropy value means the lower certainty of base pairings between RNA segments.

As the segment size increases, SSE converges to 1. To determine the appropriate segment size, we extract 342 RNA sequences from the PDB database and calculate their normalized SSE with different segment sizes starting with 3 and increment by 1. For each SSE calculated, we also calculate a condition index to check the linear dependency. Following Grewal [23], if the condition index is greater than 30, we conclude there exist high linear dependencies among the SSEs (from varied segmentation size). This is the indicator that at least one SSE with a specific segment size can be derived from a linear combination of SSEs from other segment sizes. In that case,

Su *et al. BMC Bioinformatics*   (2019) 20:373

Page 6 of 11

adding more SSE would not contribute to distinguishing the RNA sequence. As seen in Table 1, the maximum condition indices reach > 30 when the segment size 9 is added. Therefore, we determine that the segment size should be 3 to 8. As a result, six SSE features are to be derived for the ENTRNA classification model.

ENTRNA feature: free energy

Free energy is used to measure stability of an RNA structure quantitatively. For pseudoknot-free RNAs, both the free energy value (V_{fe}) of a given pair of sequence and structure and the minimum free energy value (V_{mfe}) that the sequence could achieved would be calculated. The program RNAeval [10] of the ViennaRNA – package calculates the free energy value (V_{fe}) of any pair of sequence and secondary structure. We use RNAfold [10] of the ViennaRNA-package to calculate the minimum free energy value so that we could measure the distance between the current structure to the MFE structure in terms of free energy value.

Unlike the easily computed free energy of pseudoknot-free RNAs, the free energy of pseudoknotted RNA is hard to compute directly due to the inaccurate and incomplete parameters. Inspired by Sato's idea to decompose pseudoknotted structures into several pseudoknot-free substructures [24], we propose to decompose pseudoknotted structures into a base substructure and knotted substructure(s) (See Fig. 1).

A pseudoknot is typically formed from the base pairings between the unpaired bases in a hairpin loop and those outside the hairpin. Hence, we treat the pseudoknotted structures as the result of two-step folding: First, a pseudoknot-free base substructure is formed as the skeleton structure. Second, the unpaired bases in the hairpin formed by the base substructure form new base pairs with bases outside the hairpin. Specifically, the base substructure is the pseudoknot-free structure that keeps the maximum number of base pairs [25]. It shares the same sequence of the pseudoknotted structure but keeps bases in the knotted substructures unpaired. As a result of further improving structural stability, knotted substructures are formed by keeping the portion of the original sequence that contains additional base pairs that are not knotted. From this viewpoint, it enables the decomposition on arbitrary pseudoknots.

Since both the base substructure and knotted substructures are pseudoknot-free, free energy can be easily calculated. The following free energy based features are

Table 1 Maximum Condition Index

Segment Size	3–5	3–6	3–7	3–8	3–9
Maximum Condition Index	3.1	6.4	9.8	21.2	35.4

extracted for each pseudoknotted RNA by RNAeval [10] and RNAfold [10]:

- *Base substructure free energy* (V_{bfe}): The free energy value given to the sequence and base substructure. It is used to quantitatively measure stability of the base structure;
- *Base substructure minimum free energy* (V_{bmfe}): The minimum free energy value that the sequence could achieve without forming pseudoknots;
- *Knotted substructure free energy* (V_{kfe}): The free energy reduction brought on by the pseudoknots. In addition, we remove the energy increase caused by the "hairpin" since the hairpin is artificially created during the decomposition process.

ENTRNA features from base pair probabilities

MFE-based prediction algorithms are generally far from perfect. In general, less than 40% of base pairs could be predicted correctly if a RNA is more than 500 nucleotides [35]. Base pairing uncertainty is considered one of the top reasons. To quantitatively evaluate the base pairing uncertainty, it is assumed that the probability of a secondary structure s in equilibrium follows Boltzmann distribution:

$$p(s) \propto e^{-E(s)/RT} \tag{6}$$

where $E(s)$ is the free energy of the structure, R is the gas constant and T the thermodynamic temperature of the system. After normalization, the probability of being in secondary structure s is:

$$p(s) = \frac{e^{-E(s)/RT}}{Z} \tag{7}$$

where Z is partition function by summing over all the possible structure:

$$z = \sum_{s} e^{-E(s)/RT} \tag{8}$$

Base pairing probability p_{ij} is derived by summing up the secondary structure probability with i and j paired, q_i is the probability of base i being unpaired. The following two metrics, calculated by using base pairing probability, have been widely used to evaluate the pseudoknot-free RNA secondary structure uncertainty, which can serve as features in ENTRNA for pseudoknot-free modeling:

- *Ensemble Diversity* (V_{ed}): It measures the expected distance between the target secondary structure and all the other secondary structure. The lower ensemble diversity means the sequence has less ensemble diversity, which further implies the sequence would fold into the target secondary structure with high certainty.
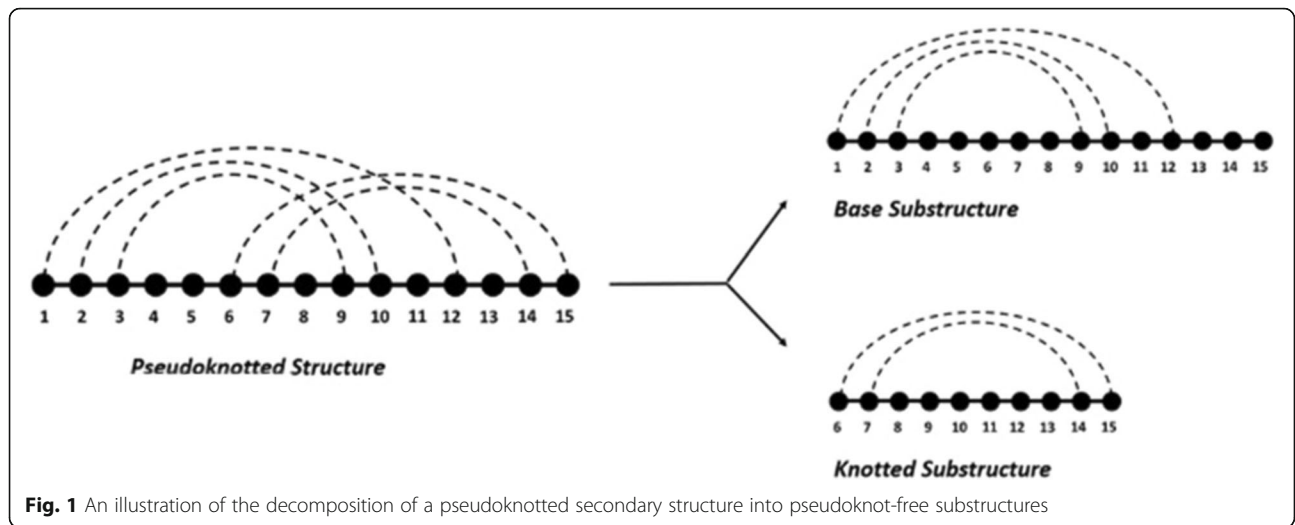
Fig. 1 An illustration of the decomposition of a pseudoknotted secondary structure into pseudoknot-free substructures

- *Expected Accuracy* (V_{ea}): It measures the expected number of bases that are in correct base pairing status. The higher expected accuracy means more bases are expected to appear in the target secondary structure, which further implies the sequence would fold into the target secondary structure with high certainty.

ENTRNA features from RNA domain knowledge

In addition to SSE, free energy and base pairing features, two more features are extracted from domain knowledge:

- *GC Content* (Per_{GC}): The percentage of guanine or cytosine nucleotides in the sequence. This is a sequence-based feature. GC content is believed to have an impact on RNA stability [26];
- *Base pair percentage* (Per_{bp}): The percentage of base pairs for a given structure. This is a structure-based feature. Base pairs bring free energy reduction in most cases, which influences the structure stability.

In Tables 2, 3 and 4, we summarize all the features including our proposed SSE, free energy, sequence and structural features used for the classification model's development.

Classification model

Based on the training dataset generated, ENTRNA applies logistic regression as a classifier to predict the foldability using 11 features (Tables 2 and 3) for pseudoknot-free and 11 features (Tables 2 and 4) for pseudoknotted RNAs separately. Compared to other classifiers, one advantage of logistic regression is that the result is a continuous value instead of a binary class, which could be explained as the probability of being in

the positive class. In this research, the prediction result could be regarded as the foldability for the given pair of sequence and secondary structure. Specifically, we set the foldability threshold as 0.5, which means the given pair of sequence and secondary structure would be classified as a successful case if its foldability value is greater than 0.5. It is our intention to conduct sensitivity analysis on this threshold as one of the future tasks.

Results

To evaluate the performance of ENTRNA, we measure the model accuracy as the mean of sensitivity and specificity:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

where TP is the number of positive examples that are

Table 2 ENTRNA: Pseudoknot-free and Pseudoknotted RNAs Common Features

Features	Calculation	Description
Per_{GC}	$\frac{\#of\ G\ nts + \#of\ C\ nts}{n}$	GC percentage
Per_{bp}	$\frac{\#of\ nts\ that\ are\ base\ paired}{n}$	Base pair percentage
$RV_{ent,\ 3}$	$\frac{V_{ent,3}}{V^*_{ent,3}}$	Normalized SSE with segment size 3
$RV_{ent,\ 4}$	$\frac{V_{ent,4}}{V^*_{ent,4}}$	Normalized SSE with segment size 4
$RV_{ent,\ 5}$	$\frac{V_{ent,5}}{V^*_{ent,5}}$	Normalized SSE with segment size 5
$RV_{ent,\ 6}$	$\frac{V_{ent,6}}{V^*_{ent,6}}$	Normalized SSE with segment size 6
$RV_{ent,\ 7}$	$\frac{V_{ent,7}}{V^*_{ent,7}}$	Normalized SSE with segment size 7
$RV_{ent,\ 8}$	$\frac{V_{ent,8}}{V^*_{ent,8}}$	Normalized SSE with segment size 8

Table 3 ENTRNA: Pseudoknot-free RNA Only Features

Features	Calculation	Description
RV_{fe}	$\frac{\lvert V_{fe}-V_{mfe}\rvert}{\lvert V_{mfe}\rvert}$	Pseudoknot-free RNA normalized free energy
V_{ed}	$\frac{\sum_{(i,j)\in s}(1-p_{ij})+\sum_{(i,j)\notin s}p_{ij}}{n}$	Ensemble Diversity
V_{ea}	$\frac{\sum_{(i,j)\in s}2p_{ij}+\sum_{i\in up}q_i}{n}$	Expected Accuracy

correctly predicted as positive, TN is the number of negative examples correctly predicted as negative, FP is the number of negative examples that are incorrectly predicted as positive and FN is the number of positive examples that are incorrectly predicted as negative.

In order to identify the best feature combinations and parameter settings, we investigate ENTRNA performance exhaustively and record the best parameter settings and feature combinations in terms of Leave-One-Out cross validation accuracy. In addition, a blind test is conducted to evaluate the robustness and generalization of the proposed ENTRNA.

Dataset

In this research, we prepare 3 separate datasets to train, cross-validate and blind test ENTRNA. The details are as follows:

- Dataset I: 2084 (1024 pseudoknot-free + 1060 pseudoknotted) RNAs from the RNASTRAND database [36]. The length ranges from 4 to 1192 nucleotides. This serves as the training dataset
- Dataset II: 299 (206 pseudoknot-free + 93 pseudoknotted) RNAs extracted by CompaRNA [27] from the PDB database. The length ranges from 20 to 1495 nucleotides. This is used as the test dataset
- Dataset III: 5 laboratory-tested pseudoknotted RNAs with synthetic sequences. All 5 RNA strands were obtained through in vitro transcription and further purified by gel electrophoresis. The RNA strands folded themselves in a buffer solution with a slow cooling process. Among the 5 sequences, 4 of them were not able to produce the designed well-formed rectangle nanostructures. The length of RNA

Table 4 ENTRNA: Pseudoknot RNA Only Features

Features	Calculation	Description
RV_{bfe}	$\frac{\lvert V_{bfe}-V_{bmfe}\rvert}{\lvert V_{bmfe}\rvert}$	Pseudoknotted RNA base substructure normalized free energy
RV_{kfe}	$\frac{\lvert V_{bfe}-V_{kfe}\rvert}{\lvert V_{bfe}\rvert}$	Pseudoknotted RNA knotted substructure normalized free energy
Per_{kbp}	$\frac{\#of\ knot\ base\ pairs}{\#of\ total\ base\ pairs}$	Percentage of knot base pairs

sequences ranges from 1618 to 1790 nucleotides. This is used to test ENTRNA on long structural-complex pseudoknotted RNAs

During the training process, all the RNAs in Dataset I are treated as the positive dataset *P*. To create the unlabeled dataset *U*, we generate 100 sequences for each secondary structure by using existing computational algorithms. Specifically, we use secondary structures in the positive dataset as seed structures, generate the sequence solutions by three different RNA inverse folding algorithms(RNAinverse [11], incaRNAtion [13] and antaRNA [14]). The reason multiple inversion folding algorithms are used is to improve the diversity of the sequence-secondary structure pairs. A pair of seed secondary structure and corresponding sequence defines an example in unlabeled dataset.

Experiment I: pseudoknot-free RNA

The first experiment is to evaluate ENTRNA on pseudoknot-free RNA. We train and cross-validate the model using 1024 pseudoknot-free RNAs from RNAS-TRAND to identify the best parameter settings and feature combinations. The model is then blindly tested using 206 RNAs from PDB database. To balance the positive and negative examples, we identify the same number of examples from the unlabeled dataset as "reliable" negative examples. After exhaustively evaluating all the feature combinations, the best performing model, leave-one-out cross validated, is built with the following 5 features:

- Normalized SSE with segment size 3 ($RV_{ent,\,3}$)
- GC percentage (Per_{gc})
- Ensemble Diversity (V_{ed})
- Expected Accuracy (V_{ea})
- Pseudoknot-free RNA normalized free energy (RV_{fe})

Since extensive research uses minimum free energy as the single metric to guide RNA design, we provide the MFE result as a reference. Specifically, we implement RNAfold [10] to estimate the MFE structure from the sequence and assess the consistency between the real RNA secondary structure and the MFE predicted RNA secondary structure. If the two structures are identical, the pair of RNA secondary structure and sequence is considered as a positive example under MFE criteria. Table 5 summarizes the comparison between ENTRNA and MFE model on the training and testing datasets.

As observed, in the training and testing, only 76 out of 1024 and 52 out of 206 RNAs are in their MFE secondary structure, which yields the MFE sensitivity to 7.4 and 25.7% separately. In the training procedure, ENTRNA is able to correctly predict 886 pairs of RNA sequence and

Su *et al. BMC Bioinformatics*   (2019) 20:373

Page 9 of 11

Table 5 Prediction result of ENTRNA on pseudoknot-free RNA

Dataset	ENTRNA Sensitivity	MFE Sensitivity
Training (1024 RNAs from RNASTRAND database)	86.5%	7.4%
Test (206 RNAs from PDB database)	80.6%	25.7%

Table 6 Prediction result of ENTRNA on pseudoknotted RNA

Dataset	ENTRNA Sensitivity
Training (1060 RNAs from RNASTRAND database)	81.5%
Test (93 RNAs from PDB database)	71.0%

secondary structure (leave-one-out sensitivity: 86.5%). By directly applying the trained model on the 206 RNAs (blind testing), 165 RNAs are correctly predicted. We conclude ENTRNA model is robust in predicting the foldability of pseudoknot-free RNAs.

Experiment II: ENTRNA on Pseduoknotted RNA

Following the same procedure as Experiment I, this experiment is to evaluate the performance of ENTRNA on pseudoknotted RNAs. Here we train and leave-one-out cross-validate the model using 1060 pseudoknotted RNAs from RNASTRAND and blindly tested using 93 RNAs from PDB database. The following 3 features are identified in the best performing model:

- Normalized SSE with segment size 3 ($RV_{ent,\ 3}$)
- Normalized SSE with segment size 8 ($RV_{ent,\ 8}$)
- Pseudoknotted RNA base substructure normalized free energy (**RV_{kfe}**)

The free energy calculation for pseudoknotted RNA is still unavailable. Therefore, we only provide the training and test accuracy of ENTRNA, which are summarized in Table 6.

From Table 6, we observe in the leave-one-out cross validated training procedure, ENTRNA is able to correctly predict 864 out of 1060 RNAs (sensitivity: 80.6%). Blind test on the PDB data gives 71.0% sensitivity, that is, 66 out of 93 pseduoknotted RNAs are correctly predicted with foldability. While it is expected blind test will have inferior performance than the training, it is our intention to further explore potential features that could be gathered to improve the predictions.

Next, we validate the model generated from the second experiment blindly on the 5 laboratory long RNA strands. Please note the first two experiments have shown that ENTRNA is able to predict positive examples with high accuracy, while the ability of predicting negative examples could not be validated due to the lack of failed RNAs. Dataset III consists of four failed RNA and one successful RNA which enables us to test the performance of ENTRNA on both sensitivity and specificity. We use the best model trained from Experiment II to predict the foldability of the give RNAs. The model is able to correctly predict the foldability of the one positive example and three out of four negative examples, which yields 100% sensitivity and 75% specificity.

Discussion

In this paper, we propose a new concept: foldability. It transforms the RNA design problem to a foldability prediction problem - predicting the folding success rate for a given pair of sequence and structure. RNA sequence and secondary structure is a many-to-many mapping, known as multi-conformation. Specifically, each RNA secondary structure could be folded from several RNA sequences and vice versa. In addition, RNA folding is a stochastic process. For each RNA sequence, it will fold into several different secondary structure with certain probabilities. This research proposes a data-driven approach taking the RNA sequence and secondary structure jointly to predict its foldability. The result shows the approach is able to predict RNA foldability with high sensitivity and specificity. This implies the potential promise of the new formulation and its uses in both RNA structure prediction and inverse folding problems.

While successfully, there is room for improvement. First, it is our intention to explore extracting more features to enrich the description of RNA for improved prediction power. Second, we plan to explore the robustness of ENTRNA. One potential issue for all data-driven approaches is the performance is highly dependent on training dataset. In ENTRNA framework, the real world RNAs are not only used in training model, but also identifying reliable negative RNA examples. A larger RNA dataset with both successful and failed (instead of negative) RNA examples will certainly help improve the robustness of the model.

Conclusion

Introducing thermodynamics (free energy) into RNA folding has been a revolutionary milestone since more than three decades ago. It provides the foundation to computational algorithms for RNA design based on three assumptions: (1) One RNA sequence has a single unique target conformation. (2) The thermodynamic parameters are accurate to derive the free energy characterizing a specific structure. (3) An RNA structure at minimum free energy (MFE) is the most stable structure. The "stable" here refers to the thermodynamic stability calculated in silico. However, recent research has proven that the same RNA sequence may fold into several structures, known as multi-conformation. The thermodynamic parameters used in calculating free energy are only estimates using nearest neighborhood methods. And, many natural RNAs

discovered in cells are in an alternative structure with higher-than-the-minimum free energy.

The issues with the three assumptions motivate us to reformulate the RNA structure prediction problem to an RNA foldability prediction problem. As a result, one sequence with its respected multiple potential structures, and one structure with its respected multiple sequences can all be assessed with a unified foldability prediction model. We propose ENTRNA as a data-driven framework for the RNA foldability prediction. In addition, we propose a new metric sequence segment entropy (SSE) as an additional feature for ENTRNA in conjunction with free energy and other RNA domain commonly used features (e.g., GC percentage). Since the unique challenge in designing data-driven approaches for RNA designs is the lack of failure examples, we propose the application of PU (Positive-Unlabeled) learning to make up the failed RNA sequence-structure pairs for the training dataset.

The performance of ENTRNA is validated using both pseudoknot-free and pseudoknotted datasets. In addition, 5 laboratory tested long structural-complex pseudoknotted RNAs with synthetic sequences are used to blindly test the model performance. The superior experiment results show that our method is able to learn from existing RNAs and apply its learning in predicting foldability of unknown RNAs. Unlike previous computational based methods, our method stands at the machine learning perspective to understand and exploit reported RNAs.

Abbreviations
FN: False Negative; FP: False Positive; MFE: Minimum Free Energy; PU: Positive-Unlabeled; SSE: Sequence Segment Entropy; SVM: Support Vector Machine; TN: True Negative; TP: True Positive

Availability of data and materials
The ENTRNA source code and other necessary resources can be obtained from https://github.com/sucongzhe/ENTRNA.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Computing, Informatics, Decision Systems Engineering, Arizona State University, Tempe, AZ 85281, USA. [2]Department of Operational Sciences, Graduate School of Engineering and Management, Air Force Institute of Technology, Wright-Patterson AFB, Dayton, OH 45433, USA. [3]Biodesign Center for Molecular Design and Biomimetics, The Biodesign Institute & School of Molecular Sciences, Arizona State University, Tempe, AZ 85281, USA.

References
1. Afonin KA, Lindsay B, Shapiro BA. Engineered RNA nanodesigns for applications in RNA nanotechnology. DNA RNA Nanotechnol. 2013;1(1).
2. Doherty EA, Doudna JA. Ribozyme structures and mechanisms. Annu Rev Biophys Biomol Struct. 2001;30(1):457–75.
3. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature. 2001;411(6836):494–8.
4. Shajani Z, Sykes MT, Williamson JR. Assembly of bacterial ribosomes. Annu Rev Biochem. 2011;80:501–26.
5. Bramsen JB, Kjems J. Development of therapeutic-grade small interfering RNAs by chemical engineering. Front Genet. 2012;3:154.
6. Laing C, Schlick T. Computational approaches to 3D modeling of RNA. J Phys Condens Matter. 2010;22(28):283101.
7. Thirumalai D, Lee N, Woodson SA, Klimov DK. Early events in RNA folding. Annu Rev Phys Chem. 2001;52(1):751–62.
8. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinf. 2010;11(1):1.
9. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981; 9(1):133–48.
10. Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26.
11. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh Chem/Chem Mon. 1994;125(2):167–88.
12. Andronescu M, Fejes AP, Hutter F, Hoos HH, Condon A. A new algorithm for RNA secondary structure design. J Mol Biol. 2004;336(3):607–24.
13. Reinharz V, Ponty Y, Waldispühl J. A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. Bioinformatics. 2013;29(13):i308–15.
14. Kleinkauf R, Mann M, Backofen R. antaRNA: ant colony-based RNA sequence design. Bioinformatics. 2015;31(19):3114–21.
15. Parisien M, Major F. The MC-fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008;452(7183):51–5.
16. Hofacker IL, Stadler PF. Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics. 2006;22(10):1172–6.
17. Woods CT, Lackey L, Williams B, Dokholyan NV, Gotz D, Laederach A. Comparative visualization of the RNA suboptimal conformational ensemble in vivo. Biophys J. 2017;113(2):290–301.
18. Liu B, Dai Y, Li X, Lee WS, Yu PS. Building text classifiers using positive and unlabeled examples. In: Data mining, 2003. ICDM 2003. Third IEEE international conference on: IEEE; 2003;3:179–188.
19. Williams S, Lund K, Lin C, Wonka P, Lindsay S, Yan H. Tiamat: a three-dimensional editing tool for complex DNA structures. In: International workshop on DNA-based computers. Berlin: Springer; 2008. p. 90–101.
20. Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3–55.
21. Garcia-Martin JA, Clote P. RNA thermodynamic structural entropy. PLoS One. 2015;10(11):e0137859.
22. Huynen M, Gutell R, Konings D. Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol. 1997;267(5):1104–12.
23. Grewal R, Cote JA, Baumgartner H. Multicollinearity and measurement error in structural equation models: implications for theory testing. Mark Sci. 2004;23(4):519–29.

24. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics. 2011;27(13):i85–93.
25. Smit S, Rother K, Heringa J, Knight R. From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. RNA. 2008;14(3):410–6.
26. Isaacs FJ, Dwyer DJ, Ding C, Pervouchine DD, Cantor CR, Collins JJ. Engineered riboregulators enable post-transcriptional control of gene expression. Nat Biotechnol. 2004;22(7):841–7.
27. Puton T, Kozlowski LP, Rother KM, Bujnicki JM. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. Nucleic Acids Res. 2013;41(7):4307–23.
28. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999;49(2):145–65.
29. Ding Y, Chan CY, Lawrence CE. S fold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res. 2004;32(suppl_2):W135–41.
30. Steffen P, Voß B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 2005;22(4):500–3.
31. Rogers E, Heitsch CE. Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble. Nucleic Acids Res. 2014;42(22):e171.
32. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.
33. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 2008;9(3):R55.
34. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462.
35. Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinf. 2004;5(1):105.
36. Andronescu M, Bereg V, Hoos HH, Condon A. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics. 2008;9(1):340.

Publisher's Note