Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 25):684
https://doi.org/10.1186/s12859-019-3258-7

BMC Bioinformatics

RESEARCH

Open Access

Predicting RNA secondary structure via adaptive deep recurrent neural networks with energy-based filter

Weizhong Lu[1], Ye Tang[1], Hongjie Wu[1,2*], Hongmei Huang[1], Qiming Fu[1], Jing Qiu[1] and Haiou Li[1]

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference
Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: RNA secondary structure prediction is an important issue in structural bioinformatics, and RNA pseudoknotted secondary structure prediction represents an NP-hard problem. Recently, many different machine-learning methods, Markov models, and neural networks have been employed for this problem, with encouraging results regarding their predictive accuracy; however, their performances are usually limited by the requirements of the learning model and over-fitting, which requires use of a fixed number of training features. Because most natural biological sequences have variable lengths, the sequences have to be truncated before the features are employed by the learning model, which not only leads to the loss of information but also destroys biological-sequence integrity.

Results: To address this problem, we propose an adaptive sequence length based on deep-learning model and integrate an energy-based filter to remove the over-fitting base pairs.

Conclusions: Comparative experiments conducted on an authoritative dataset RNA STRAND (RNA secondary STRucture and statistical Analysis Database) revealed a 12% higher accuracy relative to three currently used methods.

Keywords: RNA, Secondary structure prediction, Recurrent neural network, LSTM, Pseudoknots

Introduction

RNA is a carrier of genetic information, and its structure plays a crucial role in gene maturation, regulation, and function [1–3]. Studying the relationship between RNA function and structure and determining the form and frequency of RNA folding are important to reveal the role of RNA molecules in the life process [4–6]. The most common way to manipulate RNA structures algorithmically is to reduce them to base pairs (i.e., secondary structures) abstracted from the actual spatial arrangement of nucleotides. For a valid secondary structure, each base, *i*, can only interact with at most one other base, *j*, and form one base pair (*i, j*) [7, 8].

The secondary structure of an RNA molecule represents base-pair interactions that fundamentally determine overall structure [9–11]. Current studies of RNA molecular structure emphasize the difficulty of RNA secondary structure analysis [12, 13].

Pseudoknots are substructures of RNA secondary structure that describe crossed base pairs [(*i, j*) and (*k, l*)] in a sequence, where $i < k < j < l$. RNA secondary structure prediction in the absence of pseudoknots has been studied using dynamic programming algorithms described by Zuker [14] and Mathews [15, 16] and employing m-fold [17] and GT-fold [18]. From an algorithmic standpoint, RNA pseudoknotted secondary structure prediction represents an NP-hard optimization problem [19]; therefore, in

* Correspondence: Hongjie.wu@qq.com
[1] School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiang 215000, China
[2] Anhui Key Laboratory of Intelligent Building Energy Efficiency, Anhui Jianzhu University, Hefei, Anhui 230601, China

order to reduce computational complexity, most algorithms ignore pseudoknots [7].

Common RNA secondary structure prediction models mainly include thermodynamic models, homology comparison models, and statistical-learning models [20]. The thermodynamic model assumes that RNA molecules are subject to the laws of thermodynamics, and that RNA structures with a lower free energy are more stable. Therefore, from all possible secondary structures, that with smallest free energy represents the optimal predicted result. A homologous comparison model searches for commonly mutated base pairs in sequences from the same source, although homologous sequences need to be additionally provided as input [21]. The statistical-learning model can predict the regularity of known RNA structures through machine learning and other methods, with accuracies that can potentially exceed those associated with the method targeting the minimum free energy.

After translating the RNA secondary structure-prediction problem into a classification problem of base pairings in the sequence by using machine-learning algorithms, computational complexity can be reduced, but only to a certain extent. However, there remain two difficulties. First, the existence of pseudoknots makes the folding of RNA sequences more complicated; therefore, bases cannot be distinguished using only three categories of “unpaired bases”, “paired bases near the head”. and “paired bases near the end”. The E-NSSE method, which divides bases into five categories, can only predict plane pseudoknots, but cannot predict complex structures involving nonplanar pseudoknots. Second, a machine-learning model using a fixed-sized vector as the input feature is unsuitable for processing sequence-type data and cannot process RNA sequences of variable length.

As a research hotspot in the field of machine learning, deep learning can mine deeper hidden features from data [22–24]. A recurrent neural network is a sequence-oriented neural-network model for deep learning that displays excellent performance in natural-language processing, image recognition, and speech recognition [25, 26].

However, common deep neural-network models are restricted to features with a fixed shape and, therefore, cannot model RNA primary structures with variable sequence lengths. Here, we applied a long short-term memory (LSTM) network to establish a secondary structure-prediction method that is adaptable to RNA sequences of variable length. A previous study by Mathews [27] showed that a higher base-pairing probability calculated by the partition function resulted in a greater the probability of its appearance in the real structure. Therefore, the type of base and the output of its partition function was selected as the feature of the base. Additionally, we introduced a mask vector to eliminate the effect of the extended sequence on the model, which allowed the model to process variable length RNA primary sequences. A weight vector was used to dynamically regulate the proportion of the loss function associated with each base in the total loss function of the sequence, thereby alleviating the unbalanced distribution of the samples. However, there may be some conflicting predicted bases in the predicted result of LSTM, for example, the i-th base is predicted to be paired with the j-th base, but the j-th base is predicted to be an unpaired base or paired with another base. To solve the problem, a energy-based filter is also proposed to filter the conflicting predicted result of LSTM.

Methods

RNA secondary structure prediction

A, G, C, U are four different bases in RNA molecules, several bases are arranged in order to form the primary structure of RNA [28, 29]. The primary structure of an RNA sequence S consisting of n bases can be expressed as $S = s_1, s_2, \ldots s_{\mathrm{n}}$, where s_1 is the base near the 5′ side, s_n is the base near the 3′ side, s_i is the i-th base in sequence S and $s_i \in \{\mathrm{A, G, C, U}\}$.

RNA secondary structure prediction problem, with the purpose of calculating the pairing results y_i of each base s_i in sequence S when the primary structure of S is known, is a classification problem. According to different categories of classification, RNA secondary structure prediction problems can be divided into the following categories:

(1) Two categories classification: pairing results consist of the category of paired base ($y_i = 1$) and the category of not paired base ($y_i = 0$).
(2) Three categories classification: pairing results include the category of paired base near 5′ side ($y_i = 1$), the category of paired base near 3′ side ($y_i = 2$) and the category of not paired base ($y_i = 0$).
(3) Multi-category classification: for a sequence with pseudoknots, pairing results $y_i = j (j = 0,1 \ldots \mathrm{n})$ means the i-th base is paired with the j-th base if $j > 0$, otherwise the i-th base is not paired.

This paper focus on (1, 3).

Adaptive LSTM with energy-based model

The scheme of a RNA secondary structure-prediction model based on Adaptive LSTM and energy-based filter is shown in Fig. 1.

The input feature matrix of a sequence is a two-dimensional tensor X. $X\ [i, j]$ represents the j-th feature of the i-th base, the $N + 5$ dimensional vector $X\ [i]$ is the

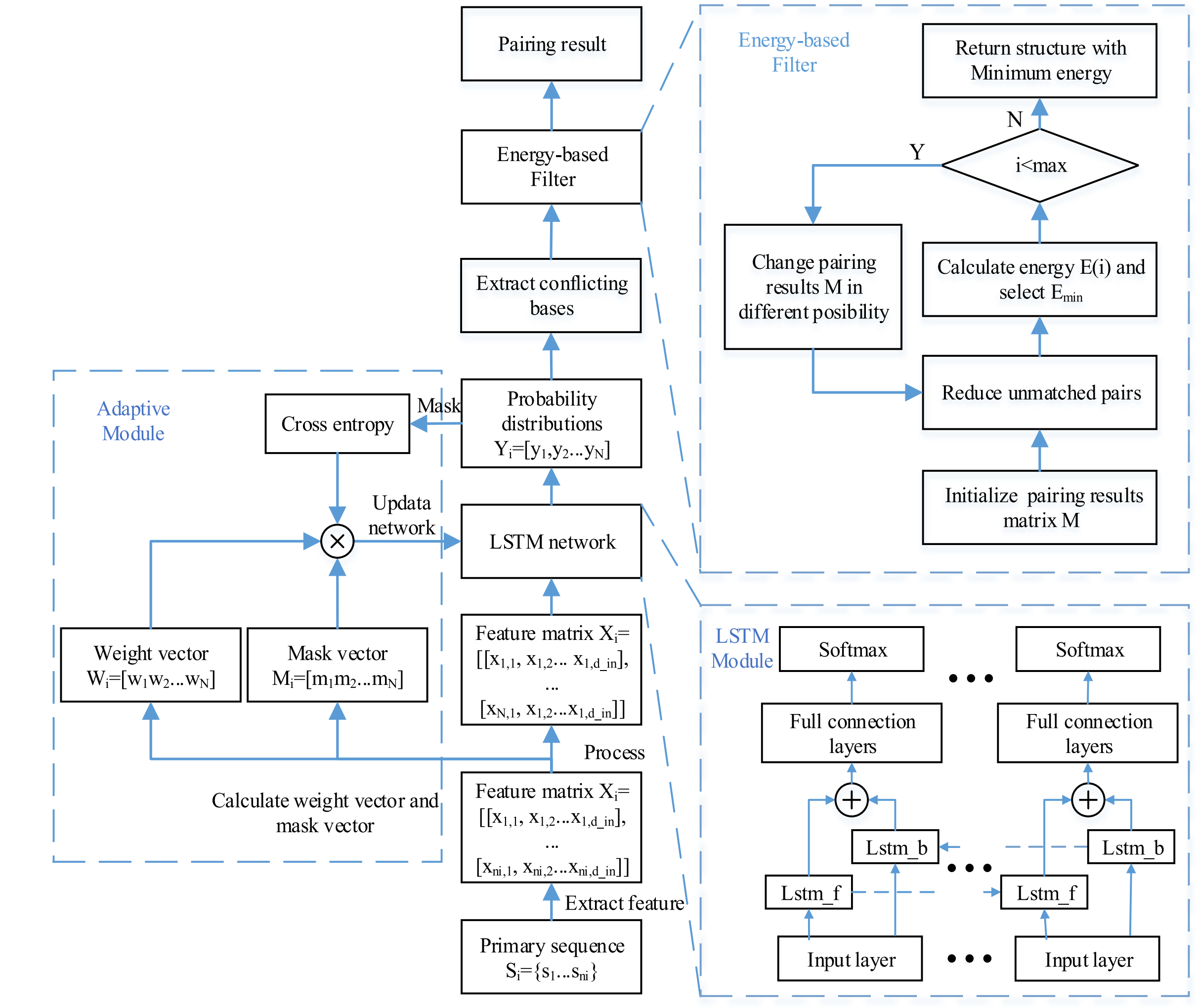

Pairing result
Energy-based Filter
Extract conflicting bases
Probability distributions $Y_i=[y_1,y_2...y_N]$
LSTM network
Feature matrix $X_i=$ $[[x_{1,1}, x_{1,2}... x_{1,d\_in}],$ ... $[x_{N,1}, x_{1,2}...x_{1,d\_in}]]$
Process
Feature matrix $X_i=$ $[[x_{1,1}, x_{1,2}...x_{1,d\_in}],$ ... $[x_{ni,1}, x_{ni,2}...x_{ni,d\_in}]]$
Extract feature
Primary sequence $S_i=\{s_1...s_{ni}\}$
Adaptive Module
Cross entropy
Mask
Updata network
Weight vector $W_i=[w_1w_2...w_N]$
Mask vector $M_i=[m_1m_2...m_N]$
Calculate weight vector and mask vector
Energy-based Filter
Return structure with Minimum energy
N
Y
i<max
Change pairing results M in different posibility
Calculate energy E(i) and select E_{min}
Reduce unmatched pairs
Initialize pairing results matrix M
LSTM Module
Softmax
Softmax
Full connection layers
Full connection layers
Lstm_b
Lstm_b
Lstm_f
Lstm_f
Input layer
Input layer

Fig. 1 Framework $X \in R^{I_1^{author} \times I_2^{paper} \times I_3^{term}}$ of the Adaptive-LSTM with filter

input feature vector of a base, the first element of x [i] is the frequency of the base in the sequence, the second to 5th elements represent the type of base, replacing A,U,C, G and unknown type by [0,0,0,1], [0,0,1,0], [0,1,0,0], [1,0, 0,0] and [0,0,0,0]. The last N elements are the outputs of the partition function. The output label is set as the multi-category classification in section 2.1.

This model capable of predicting RNA pseudoknots comprises adaptive module, LSTM module and energy-based filter module.

LSTM module consists of an input layer, an LSTM layer, fully connected layers, and a softmax layer. The input layer maps the input features into higher dimensional feature vectors and inputs them into the forward LSTM unit lstm_f and the backward LSTM unit lstm_b in the LSTM layer. After splitting the outputs of the forward LSTM and the backward LSTM, they are input into the back-propagation neural network comprising the full connection layer and the output layer for classification. The fully connected layers use tanh as an activation function, and a dropout layer was added to improve fitting to the test data. The output layer uses softmax as an activation function and converts the output of the fully connected layer into a probability distribution vector that represents the probability that each base in the sequence belongs to each output category.

The adaptive module is used to make the model handle sequences with variable length and reduce the impact from imbalanced samples.

The energy-based filter module is used to filter the over-fitting pairs by selecting the structure with lower free energy. In the output of LSTM, each base s_i, of the sequence has an independently predicted category y_i, representing the most likely base index to be paired with. However, it is possible that some predicted results could have been duplicated or might have shown inconsistencies. For example, the *i*-th base was predicted as being paired with the *j*-th base ($y_i = j$), but the base with which the *j*-th base was predicted to be paired with was not the *i*-th base ($y_i \neq j$). The energy-based filter correct some of these conflicting bases to make the predicted structure conform to the principle of base pairing and have lower free energy.

Adaptive LSTM

A mask vector was introduced to enable the model to effectively process sequences of different length and to ensure that the extended base will not affect the normal training of the model.

Let the maximum length of the sequence the model can accept be *N*. For the *k*-th sequence, its length is n_k, and the original sequence features comprising its feature of n_k bases are $x(1)$, x (2) ... x (n_k). Expand the sequence feature by setting the features of the $n_k + 1$-th to *N*-th base, x ($n_k + 1$), x ($n_k + 2$)...$x(N)$, with an arbitrary value, so that all lengths of all sequences can be unified into a fixed value, *N*. An *N*-dimensional mask vector, $\mathrm{M_k}$, is generated ($M_k = [M_k(i) \mid i = 1, 2 \ldots N]$, $M_k(i) = 1$) when $i \leq n_k$, and $M_k(i) = 0$ when $n_k < i \leq N$. The mask vector is used to distinguish between the original and extended parts of a sequence, and the new sequence represents the input to the LSTM network. Therefore, the cross-entropy loss function of this sequence is calculated as follows:

$$L_k\left(y, y'\right) = \frac{1}{n}\sum_{i=1}^{n} M_k(i) \sum_{j=0}^{m} y'[i, j] \log(y[i, j]) \tag{1}$$

where *n* is the length of the sequence, m is the number of categories, the two-dimensional arrays *y* and *y’* represent the prediction result and the real label, respectively, and *y* [*i*, *j*] indicates the probability that the *i*-th base belongs to the *j*-th category.

When calculating the cross-entropy of the *i*-th base in the *k*-th sequence, $M_k(i)$ will appears as a product in the calculation of the gradient formula. When the *i*-th base is from the original sequence, ($i \leq n_k$) [$M_k(\mathrm{i}) = 1$], the gradient value will not be changed by $M_k(i)$, and the network weight will be updated similar to that in a conventional method. When the base is not found in the original sequence ($n_k < i \leq N$) [$M_k(i) = 0$], the gradient value will be 0, and the network weight will not be updated. Therefore, the invalid prediction by the model of extended bases will not affect the update of the model.

Dynamic weighting method

Among the two categories used for classification of RNA secondary structure prediction, the ratio of the number of bases belonging to the paired and unpaired categories ~ 6:4 in the RNA STRAND dataset; however, in the multi-category classification problem, the ratio is ~ $6/n$: $6/n$: ...: $6/n$:4. As n increases, there will be an uneven distribution of samples, which might lead a model to predict all bases as in the unpaired category, because the number of bases belonging to that category in the real structure is much larger than that of any other categories.

To address this problem, a dynamic weighting method was added to the model. For the *k*-th sequence, a weight vector, W_k, was generated. If the *i*-th base is a paired base in the real structure, the value of $W_k(i)$ equals the number of unpaired bases in the sequence; otherwise, the value of $W_k(i)$ is 1. After adding the dynamic weighting method, the loss function of the *k*-th sequence is as follows:

$$L_k\left(y, y'\right) = \frac{1}{w}\sum_{i=1}^{n} W_k(i) M_k(i) \sum_{j=0}^{m} y'[i, j] \log(\mathrm{y}[i, j])$$

$$\mathrm{m} = \sum_{k=1}^{N} M_k, \mathrm{w} = \sum_{k=1}^{N} W_k \cdot M_k,$$

$$W_k = \begin{cases} 1 & y'[i, 0] \neq 1 \\ \sum_{i=1}^{n} y'[i, 0] & y'[i, 0] = 1 \end{cases} \tag{2}$$

Energy-based filter

As the result of translating the RNA secondary structure-prediction problem into a classification problem of base pairings, there exist some conflicting pairing result in the output of LSTM. The energy-based filter is used to deal with this problem.

In laws of thermodynamics, RNA structures with a lower free energy are more stable [16], so the energy-based is used to randomly change the label of conflicting base pairings according to the free energy of the structure to make the structure more likely to its real structure.

According to the Watson-Crick base complementary pairing principle [23], each base $s(i)$, can only interact with at most one other base, $s(j)$, to form one base pair ($s(i)$, $s(j)$) and $\{s(i), s(j)\} \in \{\{\mathrm{A,U}\},\{\mathrm{C,G}\},\{\mathrm{G,U}\}\}$. As a result, the predicted result of *i*-th base $y(i)$ can be reserved if the two following conditions are met:

(1) $y(y(i))=i$
(2) ($s(i)$, $s(y(i))$) is in {(A, U), (U, A), (C, G), (G, C), (G, U), (U, G)}

If these conditions are not met, $y(i)$ should be set as 0 to classify the *i*-th base as unpaired base.

By setting all the conflicting bases to unpaired bases, it may incorrectly turn false positive samples into false negative samples, energy-based filter is improved from reducing unmatched pairs to filter pairing results by the free energy of secondary structure. The flow chart of energy-based filter is shown in the right upper of Fig. 1 and the energy-based filter algorithm is as follows:

In Extract_bases(y), bases with conflicting pairing result will be extracted from y (the predicted result of Adaptive LSTM) into conflicting base list $S(i)$ and its index $Id(i)$, $i\in\{0,1 \ldots m^{-1}\}$, $S(i)$, $Id(\mathrm{i})\in\{1,2 \ldots k\}$, where m is the number of conflicting bases, k is the length of the primary sequence, i is the index of each base in the conflicting bases list, $Id(\mathrm{i})$ is the index of the i-th base in its primary sequence, $S(i)$ represents the predicted result of Adaptive LSTM of i-th conflicting base (the $Id(i)$-th base in its primary sequence).

The pairing result matrix M_{mxn}, denoting n pairing results, is randomly initialized with 0 or 1, each column of the matrix represents the pairing result of the m conflicting bases. In the j-th pairing result, if $M(i,j) = 1$, it means the predicted result of Adaptive LSTM of i-th conflicting base is retained, if $M(i,j) = 0$, it means the predicted result of Adaptive LSTM of i-th conflicting base is not accepted and $S(i)$ should be set as 0 to classify this base to unpaired base.

Correct($M[:,j]$) is to correct the j-th pairing result according to the j-th column of M: y' is a copy of y, in the j-th pairing result, for each conflicting base, if $M(i,j) =$ 1and the y $(Id(i))$-th base in primary sequence is conflicting base or unpaired base, keep $y'(Id(\mathrm{i}))$ and set $y'(y\ (Id(i))) = Id(i)$, else set $y'(Id(i)) = 0$.

Reducing unmatched pairs is to correct the predicted result according to the two conditions.

In Calculate_energy(y'), $E(j)$ is the free energy of the j-th pairing result. And M_{best} is the structure with lowest free energy in M till now.

Change each pairing result: for each base of each pairing result, if $M(i,j) = M_{best}(j)$, set $M(i,j) = 1\text{-}M(i,j)$ in a low possibility $p1$, else set $M(i,j) = 1\text{-}M(i,j)$ in a high possibility $p2$.

Evaluation metrics

Accuracy (ACC) is a commonly used evaluation metrics in classification. Sensitivity (SEN) and specificity (PPV) are commonly used in RNA secondary structure prediction [30]. Matthews correlation coefficient (MCC) is an evaluation metrics that combines sensitivity and specificity. This paper uses SEN, PPV, MCC and ACC to evaluate the model, they are calculated as follows:

$$sen = \frac{TP}{TP + FN} \tag{3}$$

$$ppv = \frac{TP}{TP + FP} \tag{4}$$

$$mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Where TP (true positives) means the number of correctly predicted bases; FN (false negatives) means the number of bases that are not correctly predicted; FP (false negatives) means the number of unpaired bases that predicted to be paired; TN (true negatives) means the number of correctly predicted unpaired bases [31]. The range of SEN, PPV and ACC is between 0 and 1, while the range MCC is between − 1 and 1, and the higher these evaluation metrics are, the better the model is.

Results and discussion

Dataset. The dataset of this paper comes from authoritative dataset RNA STRAND [32], including five subsets: TMR (The tmRNA website [33]),SPR (Sprinzl tRNA Database [34]),SRP (Signal recognition particle database [35]),RFA (The RNA family database [36])and ASE (RNase P Database [37]).There are 2493 sequences in the 5 datasets, the maximum and average length is 553 and 267.37 respectively. The number of sequences, the average sequence length, the minimum length and the maximum length are shown in Table 1. 90% of these sequences are randomly selected as training data and the rest 10% are testing data.

Comparison between adaptive-LSTM with and without energy-based filter

To prove the validity of the energy-based filter, a comparative experiment was carried out on the five datasets. Figure 2 shows the accuracy comparison of adaptive LSTM with energy-based filter(y-axis) and adaptive LSTM without energy-based filter(x-axis) on 249 test RNAs and the size of each points indicates the length of sequence. The number of points above the dotted line y = x are much more than the number of points under the dotted line. There are 172 RNAs (above the dotted line y = x) predicted by adaptive-LSTM with filter with higher accuracy than the accuracy predicted by adaptive-

Table 1 Datasets

Dataset	Number	Average	Max	Min
TMR	721	361.1	463	102
ASE	454	332.6	486	189
SPR	622	77.3	93	54
SRP	383	224.7	533	66
RFA	313	118.9	553	40

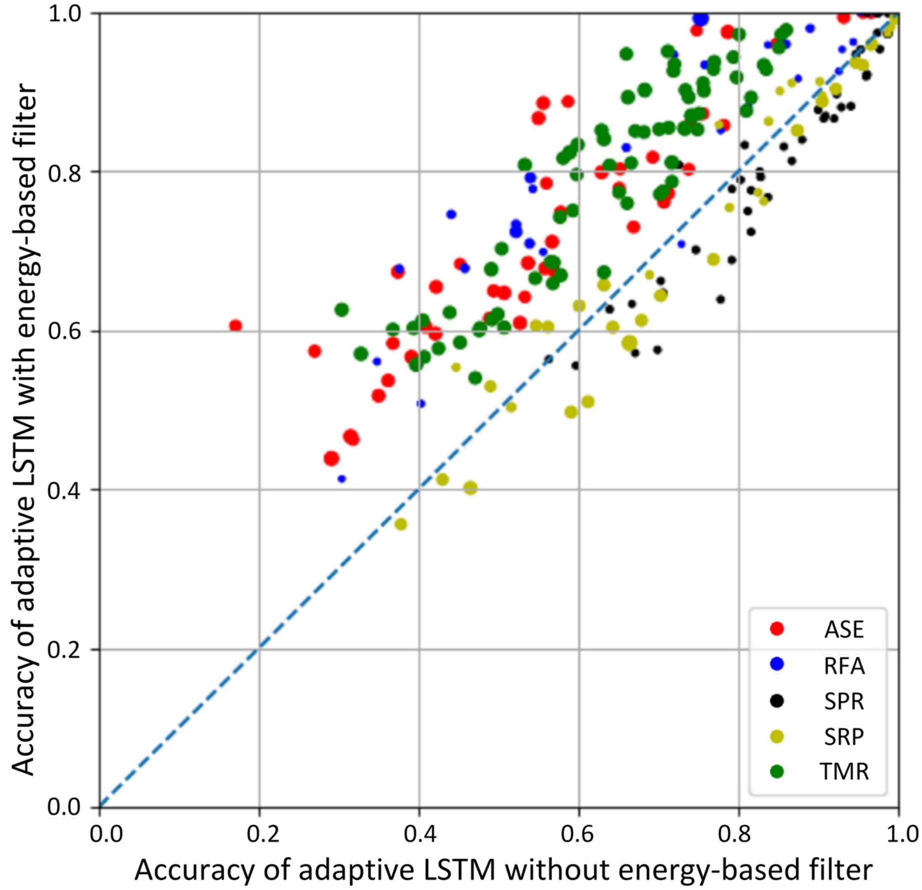


Fig. 2 Scatter of accuracy comparison between adaptive-LSTM with and without filter

LSTM. On the dataset ASE and TMR adaptive-LSTM with filter predicted 100 and 100% of test RNAs are better than the prediction by adaptive-LSTM.

In the Fig. 2, most of the RNAs in dataset SPR (in black) with energy-based filter have lower accuracy than without the filter. There are two possible reasons. The first one is the average length of the sequences in SPR dataset is much shorter than other datasets and their folding structures are relatively simple. Less free energy gap of the simple structures leads to energy-based filter failure. The second possible reason is the percentage of bases with unknown base type is 11.44% on SPR dataset, while the percentage on TMR, ASE, SRP, RFA are 0.04, 0.02, 5.93%, which will cause inaccuracy of free-energy. Since the energy-base filter heavily relies on the free energy, the two reasons may make the energy-based filter failed over adaptive-LSTM.

Comparison between adaptive LSTM and other three classical methods

ProbKnot [38] assembles maximum expected accuracy structures from computed base-pairing probabilities in $O(N^2)$ time. Cylofold [39] is an RNA secondary structure prediction method with no algorithmic restriction in terms of pseudoknot complexity. Centroidfold [40] use novel estimators to maximize an objective function which is the weighted sum of the expected number of the true positives and that of the true negatives of the base pairs.

Comparison experiments between adaptive LSTM with energy-based filter and these three classical method was operated, the *SEN*, *PPV*, *ACC* and *MCC* of ProbKnot are 0.757, 0.587, 0.646 and 0.319, the *SEN*, ppv, *ACC* and *MCC* of Cylofold are 0.414, 0.319, 0.406 and 0.014, the *SEN*, *PPV*, *ACC* and *MCC* of Centroidfold are 0.673, 0.605, 0.654 and 0.307, the *SEN*, *PPV*, *ACC* and *MCC* of Adaptive LSTM are 0.927, 0.613, 0.689 and 0.483, the *SEN*, *PPV*, *ACC* and *MCC* of adaptive-LSTM with energy-based filter are 0.685, 0.883, 0.780 and 0.592.

Table 2 shows the *MCC* and *ACC* of the adaptive-LSTM (Adaptive), adaptive-LSTM with energy-based filter (Filter) and other classic RNA secondary structure-prediction methods. Adaptive LSTM is better than other three methods in all metrics and energy-based filter can further improve the *ACC* and *MCC*. Because the Cylofold algorithm does not allow for missing bases, it

Table 2 MCC and ACC of adaptive LSTM and other three methods

Dataset	Metrics	ProbKnot	Cylofold	CentroidFold	Adaptive	Filter
TMR	MCC	0.105	−0.043	0.106	0.434	**0.581**
	ACC	0.531	0.485	0.561	0.630	**0.786**
SPR	MCC	0.591	*	0.668	**0.786**	0.751
	ACC	0.796	*	0.834	**0.891**	0.870
SRP	MCC	0.262	−0.184	0.177	0.421	**0.475**
	ACC	0.613	0.396	0.584	**0.708**	0.690
RFA	MCC	0.398	0.256	0.299	**0.451**	0.699
	ACC	0.677	0.624	0.650	0.661	**0.834**
ASE	MCC	0.238	0.043	0.286	0.323	**0.484**
	ACC	0.611	0.523	0.642	0.556	**0.720**
Average	MCC	0.319	0.014	0.307	0.483	**0.592**
	ACC	0.646	0.406	0.654	0.689	**0.780**

Boldface represents the highest MCC or ACC in comparison with the other three methods

*indicates Cylofold does not generate results on SPR dataset, since Cylofold can not accept the sequence with missing bases in SPR dataset

generated no results for SPR datasets in which the sequence information was incomplete.

Compared with the three classical methods, Adaptive LSTM have higher *ACC* in four datasets and higher MCC in all datasets, after adding the energy-based filter, the *MCC* is further improved to 0.78 on average. Because the Cylofold algorithm does not allow for missing bases, it generated no results for SPR datasets in which the sequence information was incomplete.

Case study on RNA with pseudoknots

RFA_00633 (hepatitis delta virus ribozyme) is an RNA sequence with pseudoknots and represents a noncoding RNA molecule found in the hepatitis delta virus that is necessary for viral replication and reportedly the only catalytic RNA required by a human pathogen for viability [41]. The length of its primary sequence is 91 bases, and its native secondary structure is shown in Fig. 3.

Figure 4 shows the predictive results from the ProbKnot, adaptive LSTM with energy-based filter, Centroidfold and Cylofold. There were no pseudoknots in the predicted structures, and the ACC values were 51.6 and 40.7% for ProbKnot and centroidfold, respectively Fig. 4a and c. Figure 4d shows the predicted structure with pseudoknots by cylofold, with an ACC of 63.7%.

Adaptive LSTM without energy-based filter cannot predict a valid result to construct a secondary structure because there are conflicting pairing bases. The structure predicted by the adaptive LSTM showed an *ACC* of 93.4%, which exceed that of the other three methods.

Conclusions

This paper proposed a RNA secondary structure predicting method based on Adaptive LSTM with energy-based filter. This method addressed problems associated with truncating sequences in order to address problems associated with the variability in RNA-sequence length, with truncation often resulting in the loss of sequence information and incompleteness. Additionally, we added a dynamic weighting algorithm to alleviate problems related to the unbalanced distribution of samples and use energy-based filter to remove the conflicting pairing result. Experimental results showed that this method effectively improved the accuracy of RNA secondary structure prediction, as the *MCC* metrics were 16% higher than other 3 classical algorithms on average, and energy-based filter can further improve the *MCC* to 59.2% which is 28% higher than other methods. For tests using the TMR dataset harboring sequences with large span lengths, *MCC* and *ACC* values were 43.4 and 63%, respectively, which were 33 and 9.9% higher than that of the ProbKnot algorithm, respectively.

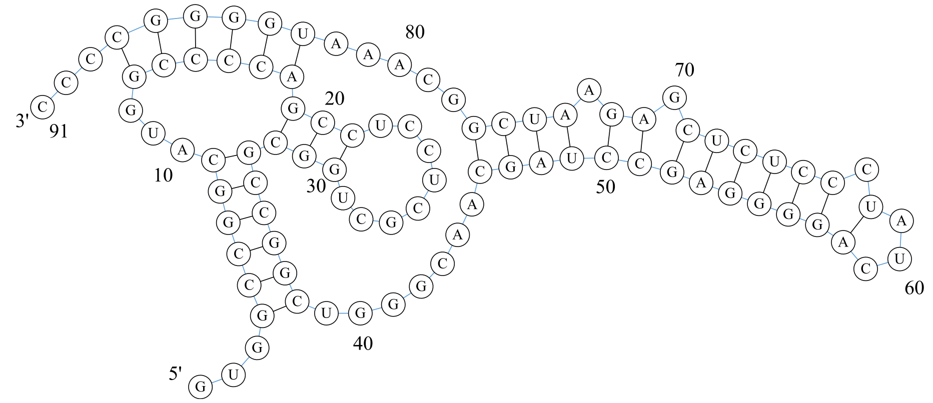
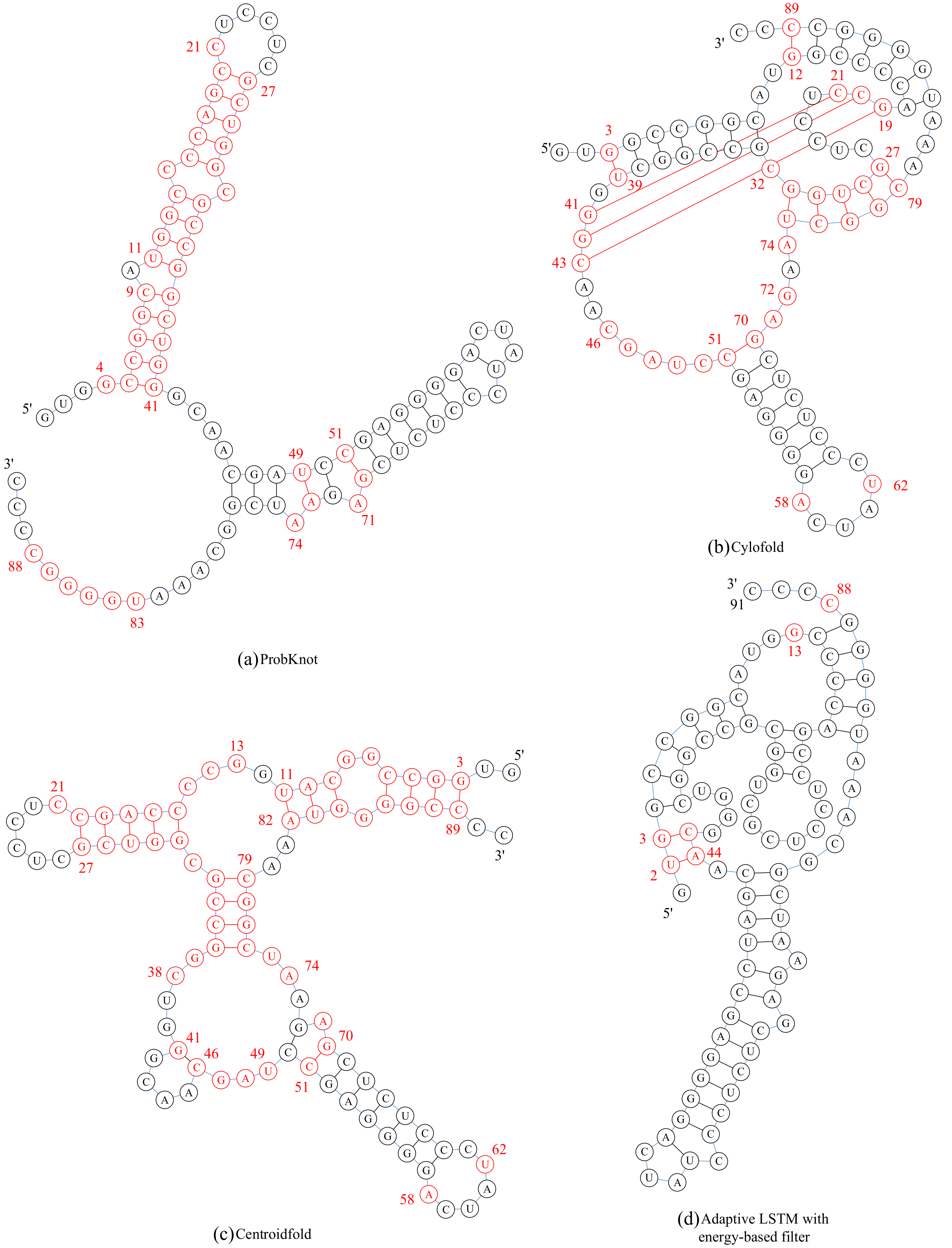


Fig. 3 Native secondary structure of RFA_00633



(a) ProbKnot

(b) Cylofold

(c) Centroidfold

(d) Adaptive LSTM with energy-based filter

Fig. 4 Predicted secondary structure of RFA_00633. **a** ProbKnot. **b** Cylofold. **c** Centroidfold. **d** Adaptive LSTM with energy-based filter

Our future research will focus on improving the network structure by adding convolutional layer or attention layer and use cross validation, which could further enhance the predictive accuracy of the model.

Abbreviations

A: Adenine; ACC: Accuracy; ASE: RNase P Database; C: Cytosine; FN: False negatives; FP: False positives; G: Ganciclovir; LSTM: Long short-term memory; MCC: Matthews correlation coefficient; NP: Non-deterministic polynomial; PPV: Specificity; RFA: The RNA family database; RNA STRAND: RNA secondary STRucture and statistical Analysis Database; RNA: Ribonucleic Acid; SEN: Sensitivity; SRP: Signal recognition particle database; TMR: The tmRNA website; TN: True negatives; TP: True positives; U: Uracil

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 25, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-25.

Authors' contributions

WL and YT wrote the program and executed the experiments. HW and QF designed the protocol. HW and JQ organized the manuscript. HH and HL collected and analyzed the data. All authors read and approved of the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China (No. 61772357, 61902272, 61672371, 61876217, 61902271, 61750110519), and Suzhou Science and Technology Project (SYG201704, SNG201610, SZS201609).

Availability of data and materials

Datasets, source codes and results can be access from http://eie.usts.edu.cn/prj/AdaptiveLSTMRNA/index.html.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Published: 24 December 2019

References

1. Anderson-Lee J, Fisker E, Kosaraju V, et al. Principles for predicting RNA secondary structure design difficulty. J Mol Biol. 2016;428(5):748.
2. Liu B, Weng F, et al. iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics. 2018;34(22):3835–42.
3. Zhu L, Deng SP, et al. Identifying spurious interactions in the protein-protein interaction networks using local similarity preserving embedding. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(2):345–52.
4. Wu JS, Zhou ZH. Sequence-based prediction of microRNA-binding residues in proteins using cost-sensitive Laplacian support vector machines. IEEE/ACM Trans Comput Biol Bioinforma. 2013;10(3):752–9.
5. Liu B, Weng F, et al. iRO-3wPseKNC: identifying DNA replication origins by three-window-based PseKNC. Bioinformatics. 2018;34(18):3086–93.
6. Guo WL, Huang DS. An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency. Mol BioSyst. 2017;13(9):1827–37.
7. Lorenz R, Wolfinger MT, Tanzer A, et al. Predicting RNA Secondary Structures from Sequence and Probing Data. Methods. 2016;103:86–96.
8. Chuai GH, Ma H, Yan JF, et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol. 2018;19(1):80.
9. Jelena T, Anders EW, Michal S, et al. RNA packaging motor: from structure to quantum mechanical modelling and sequential-stochastic mechanism. Comput Math Methods Med. 2008;9(3–4):351–69.
10. Guo X, Gao L, Wang Y, et al. Large-scale investigation of long noncoding RNA secondary structures in human and mouse. Curr Bioinforma. 2018;13:450–60.
11. Huang DS, Yu HJ. Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(2):457–67.
12. Huang DS, Lei Z, et al. Prediction of protein-protein interactions based on protein-protein correlation using least squares regression. Curr Protein Pept Sci. 2014;15(6):553–60.
13. Deng SP, Huang DS. SFAPS: an R package for structure/function analysis of protein sequences based on informational spectrum method. Methods. 2014;69(3):207–12.
14. Zuker M, Sankoff D. RNA secondary structures and their prediction. Bull Math Biol. 1984;46(4):591–621.
15. Mathews DH, Turner DH. Prediction of RNA secondary structure by free energy minimization[J]. Current Opinion in Structural Biology. 2006;16(3): 270–278.
16. Mathews DH, Using the RNAstructure Software Package to Predict Conserved RNA Structures.[J]. Current Protocols in Nucleic Acid Chemistry, 2000;2(1):9–14.
17. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.
18. Mathuriya A, Bader DA, Heitsch CE, et al. GTfold:a scalable multicore code for RNA secondary structure prediction. ACM Symp Appl Comput. 2009;1(1):981–8.
19. Lyngsø RB, Pedersen CN. RNA pseudoknot prediction in energy-based models. J Comput Biol. 2000;7(3):409.
20. Zhang HW, Yang YC, Lu Z. Bioinformatics Methods for Noncoding RNAs: RNA Structure Prediction and Its Applications. Chin Bull Life Sci. 2014;3:219–27.
21. Legendre A, Angel E, Tahi F. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. BMC Bioinformatics. 2018;19(1):13.
22. Wu HJ, Wang K, Lu LY, et al. A deep conditional random field approach to Transmembrane topology prediction and application to GPCR three-dimensional structure modeling. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(5):1106–14.
23. Wu HJ, Cao CY, Xia XY, et al. Unified deep learning architecture for modeling biology sequence. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(1):1445–52.
24. Li HS, Yu H, Gong XJ. A deep learning model for predicting RNA-binding proteins only from primary sequences. J Comput Res Dev. 2018;55(1):93–101.
25. Shen Z, Bao WZ, et al. Recurrent neural network for predicting transcription factor binding sites. Sci Rep. 2018;8(1):15270.
26. Zhu L, Zhang HB, et al. Direct AUC optimization of regulatory motifs. Bioinformatics. 2017;33(14):i243–51.
27. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. 2004;10(8):1178.
28. Wu HJ, Lv Q, Quan LJ, et al. Structural topology modeling of GPCR Transmembrane Helix and its prediction. Chin J Comput. 2013;36(10): 2168–78.
29. Wu HJ, Lv Q, Wu JZ, et al. A parallel ant Colony method to predict protein skeleton and its application in CASP8/9. Sci Sin Inf. 2012;42(8):1034–48.
30. Chakraborty D, Wales DJ. Energy landscape and pathways for transitions between Watson-crick and Hoogsteen Base pairing in DNA. J Phys Chem Lett. 2018;9(1):229.
31. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics. 2004;5(1):140.
32. Mirela A, Vera B, Hoos HH, et al. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics. 2008;9(1):1–10.
33. Hudson CM, Williams KP. The tmRNA website. Nucleic Acids Res. 2015;43:138–40.
34. Jühling F, Mörl M, Hartmann RK, et al. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009;37:159–62.
35. Zwieb C, Samuelsson T. SRPDB (signal recognition particle database). Nucleic Acids Res. 2000;28(1):171–2.

36. Griffithsjones S, Moxon S, Marshall M, et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 2005;33:121.
37. Brown JW. The Ribonuclease P database. Nucleic Acids Res. 1994;22(17): 3660.
38. Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA. 2010;16(10):1870–80.
39. Eckart B, Tanner K, Shapiro BA. CyloFold: secondary structure prediction including pseudoknots. Nucleic Acids Res. 2010;38:368–72.
40. Hamada M, Kiryu H, Sato K, et al. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics. 2009; 25(4):465–73.
41. Wu HJ, Tang Y, Lu WZ, et al. RNA Secondary Structure Prediction Based on Long Short-Term Memory Model, International Conference on Intelligent Computing. Cham: Springer; 2018. p. 595–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.