Modified Free Energy Model to improve RNA secondary structure prediction with pseudoknots

Kwok-Kit Tong, Kwan-Yau Cheung, Kin-Hong Lee, Kwong-Sak Leung
Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin N.T. Hong Kong
Email: {kktong, kycheung, khlee, ksleung}@cse.cuhk.edu.hk

***Abstract*— The free energy (evaluation) models used in RNA secondary structure prediction are one of the most important reasons that makes the prediction a challenging computational problem in Bioinformatics. These models are the key factor determining the accuracy of the prediction algorithms. Previously we have developed a method called GAknot that has obtained good performance on predicting RNA secondary structures with pseudoknots. In this paper, we propose a new free energy model. We first select a number of RNA sequences from a database which contains known RNA secondary structures as a training dataset for learning this new model. From the training dataset, we then extract base pairs patterns in subsequences of pairs of k-mers from the stems of each sequence in the training data and use the patterns to formulate penalty factors. We modify the energy model by adding these penalty factors. Combined with the new modified energy model, the prediction performance of GAknot has been improved significantly. GAknot with the new modified energy model is shown to be the best method in comparison with two state-of-the-art algorithms using a commonly used testing dataset. The penalty factors of the new energy model and dataset can be downloaded at `http://appsrv.cse.cuhk.edu.hk/~kktong/NewModel`**

***Index Terms*— rna secondary structure prediction; energy model; pseudoknot**

I. Introduction

RNA secondary structure prediction including pseudoknots is an important problem in Bioinformatics. It is because predicting RNA secondary structure can provide estimation on the 3D structure and the functions of RNA [1], [2]. In addition, psueodknots are found in many RNAs, like ribosomal RNAs, telomerase RNAs and viral RNAs [3], [4], like HIV-1 [5], and they are involved in many biological functions such as splicing, ribosomal frameshifting, rival genome replication and regulation of translation [6]–[9]. Figure 1 shows the simplest type of pseudoknots, which is called H-type pseudoknots.

There are many computational methods for predicting RNA secondary structure and they can be roughly classified to two streams, which are comparative sequences approaches and single sequence approaches. Comparative sequences approaches exploit the conservation of evolutionary information in multiple homologous sequences alignment. Basically this type of approaches can get more accurate results compare to single sequence approaches, but required sequence alignments are not sufficient so that this type of approaches may not be always feasible. Single sequence approaches use

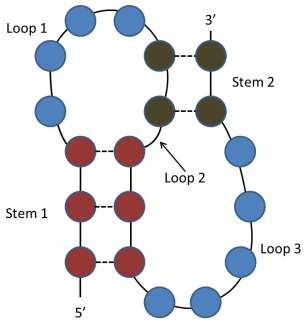

Fig. 1. A simple H-type pseudoknot. H-type pseudoknot is composed by 2 stems and 3 loops. Usually, loop 2 has zero or one base.

an energy model as a scoring function to find a structure that have a minimum free energy (MFE). Since predicting pseudoknots by finding MFE has been proven a NP-complete problem [10], there are two types of methods that can solve this problem in a reasonable time. The first type is to predict restricted classes of pseudoknots by dynamic programming. The second type uses heuristic algorithms, which can predict more classes of pseudoknots and usually more efficient than dynamic programming approaches, but not guaranteed to find the MFE structure.

The energy model used in secondary structure prediction is composed by a list of structural features (like hairpin loops, bulge loops and stacked pairs, etc.), free energy change parameters and a function that assigns total energy change to a secondary structure [11]. The Turner energy model [12] is the most widely used model for pseudoknot-free secondary structure prediction. For pseudoknots prediction, the Dirks-Pierce (DP) model is widely used and it can achieve better performance in predicting pseudoknots compared to the Turner energy model [13].

Since the accuracy of prediction methods highly depends on the energy model, the performance of prediction will be improved if the energy model can model the reality more accurately. In this paper, we propose a modified version of an existing energy model called DP09, which is an improved version of DP model proposed by Andronescu et al. [11], to increase the accuracy of RNA secondary structure prediction with pseudoknots by a set of penalty factors learnt from a dataset of validated RNA secondary structures as the training

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

data.

The paper layout is as follows: The data collection, analysis and modified energy model will be described in Section II. The results are shown in Section III. Discussion and conclusion are given in Section IV.

II. Materials And Methods

We propose a modified energy model with new parameter from a training dataset. In Section II A, we discuss the process of data collection and analysis. Then we talk about the formulation of the new energy model in Section II B.

A. Training Data and statistics analysis

To train new parameters for the energy model, we first get the validated RNA secondary structures from a database called RNA STRAND [14]. We select 1057 RNA structures with pseudoknots as the training dataset. These set of RNAs includes different lengths, from 20nt to 4400nt, and different types of pseudoknots, like simple H-type and kissing hairpins.

From the training dataset, we want to learn which parts of the RNA sequences are likely to form base pairs. Therefore, we extract the stems of each sequence from the training data and scan the stems to obtain lists of 3-mers, 4-mers and 5-mers respectively. A k-mer means a subsequence with k base-pairs. Figure 2 shows an example of the lists of 3-mers, 4-mers and 5-mers. The reason why we use 3-mers, 4-mers and 5-mers is that from the statistics of RNA STRAND, the average stem length of RNA secondary structures is concentrated between 3nt and 5nt . After that, we count the respective numbers of each 3-mer, 4-mer and 5-mer patterns, and normalize the occurrences of them by the total number of occurrences of the corresponding k-mer respectively. Please note that the normalization of patterns of 3-mers, 4-mers and 5-mers are independent. Figure 3 shows an example of the normalization. Now for each 3-mers, 4-mers and 5-mers, there is a normalized occurrence value attached to it, and we treat each value as a penalty factor for the respective 3-mers, 4-mers and 5-mers. Given a k-mer, we can obtain the corresponding penalty factor by the following function:

$$Pen^k(mer_i^k) = \frac{\#mer_i^k}{\#mer^k} \tag{1}$$

for $k = \{3, 4, 5\}$, where mer_i^k is a particular k-mer, $\#mer_i^k$ is the number of occurrence of this particular k-mer, and $\#mer^k$ is the total number of occurrences of respective k-mer. For example, $\#mer^3$ is 25 in Figure 3.

B. The New energy model

After we have extracted the stem patterns, namely 3-mers, 4-mers and 5-mers, in Section II A, we have found 6666 patterns in total, 216 for 3-mers, 1261 for 4-mers, and 5189 for 5-mers. To make a new energy model, we use DP09 (Dirks-Pierce with new parameters) proposed by Andronescu et al. [11] as the base energy model and combine it with our penalty factors defined in Section II A. To calculate the new energy, we first find out the stems of a predicted

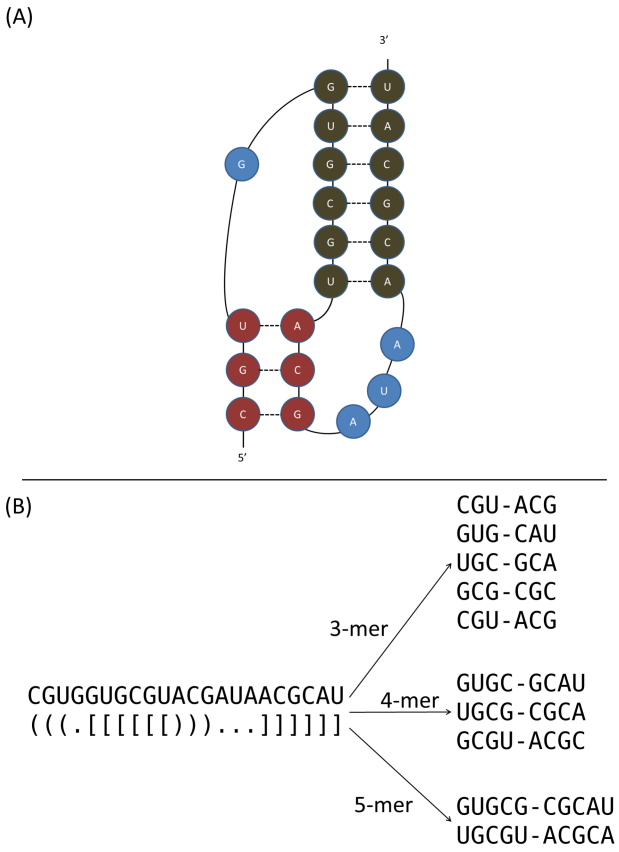

Fig. 2. An example of how to extract the 3-mers, 4-mers and 5-mers from training data. In this example, we have two stems in the secondary structure, which are 3nt and 6nt in length respectively. Then there is one 3-mer, but no 4-mer and 5-mer can be extracted from the 3nt long stem. Similarly, there are four 3-mers, three 4-mers and two 5-mers can be extracted from the 6nt long stem. Suppose we want to extract the stem patterns of the secondary structure in figure (A). The input is an equivalent representation shown in the left of figure (B). Then by this representation, we get the lists of 3-mers, 4-mers and 5-mers.

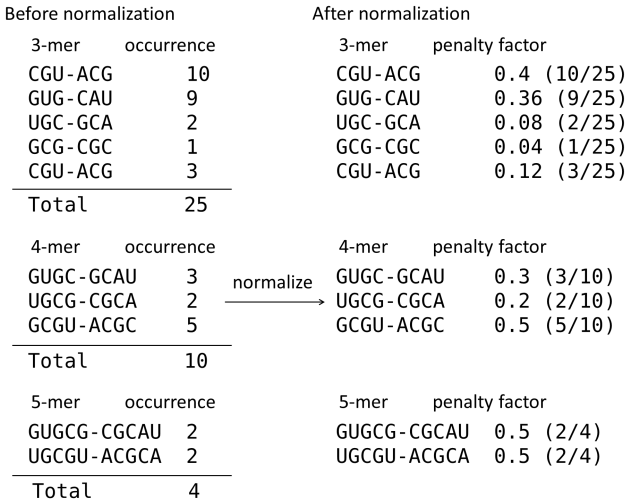

Fig. 3. An example of normalization. We normalize 3-mer, 4-mer and 5-mer independently, based on its total occurrences. Please note that the occurrence count for each pattern is summed by all training data in practice.

secondary structure and then for each stem, we use three sliding windows of sizes 3, 4 and 5, to extract the 3-mers, 4-mers and 5-mers. Then we look up the penalty factors obtained from the training dataset described in Section II A and figure 3 to calculate the sum of penalty factors, and subtract it from the energy of the predicted secondary structure given by DP09. We sum the penalty factors if the extracted 3-mers, 4-mers and 5-mers exist in the training dataset. Otherwise, the value of penalty factor is set to zero. Therefore, the new energy model is given by the following formula:

$$E = E_{DP09} - \sum_{k=3}^{5} \sum_{i} (Pen^k(mer_i^k)) \qquad (2)$$

where E_{DP09} is the energy given by DP09, $Pen^3()$ returns the corresponding penalty factor of 3-mers, $Pen^4()$ returns the penalty factor of 4-mers, $Pen^5()$ returns the penalty factor of 5-mers, mer_i^3, mer_i^4 and mer_i^5 are the respective numbers of the 3-mers, 4-mers and 5-mers which appear in the predicted secondary structure respectively. Figure 4 shows an example of the calculation of the new energy model. The reason why we formulate the new energy model as (2) is based on the intuitive belief that the longer the stem, the more stable the structure will be. Therefore, in our model, if a predicted stem is longer, it will have a higher penalty factor to lower the energy, which implies a more stable predicted structure.

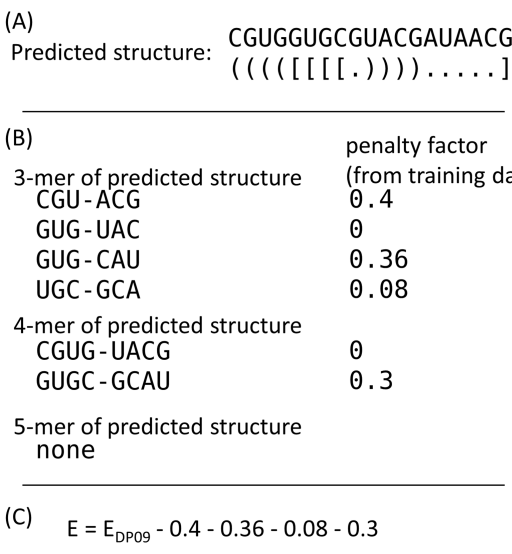

Fig. 4. An example of calculating energy of a predicted RNA secondary structure using the new energy model. Suppose we have the predicted secondary structure in figure (A). Then we extract the 3-mers, 4-mers and 5-mers of this predicted structure and find out the corresponding penalty factor, which have been get in the normalization step, as shown in figure (B). Please note that if the pattern cannot be found, which means the pattern does not appear in the training data, we will give the penalty factor as zero. Finally, we can calculate the energy of the predicted structure by (2), as shown in figure (C).

III. Results

To evaluate our new energy model, we select a well-known dataset as the testing dataset and it will be described in Section III A. In Section III B, we choose an RNA secondary structures prediction algorithm, called GAknot [15], to show the differences of performance between our new energy model and DP09, and compare them with two state-of-the-art algorithms using different energy models. The two state-of-the-art algorithms are HotKnots [11] and IPknot [16] respectively.

The training dataset, testing dataset and the penalty factors can be found at `http://appsrv.cse.cuhk.edu.hk/~kktong/NewModel`

A. Testing Data

To validate our new energy model, we use a dataset of well known, widely used and validated RNA secondary structures with pseudoknots. This dataset contains 41 sequences, which is a subset of sequences used in HotKnots [17]. We exclude 2 sequences because they do not have full sequence information. We call this dataset as HK41. The sequences in this dataset have length 27nt to 230nt.

B. Evaluation of the new energy model

We choose our previously published RNA secondary structures prediction algorithm, GAknot, as the base testing algorithm [15]. The reason is that it has been proven the searching power of GAknot is better than other existing algorithms for pseudoknots prediction. Another reason is GAknot can output more than one predicted structure. Originally, GAknot use DP09 as the scoring function. We call GAknot using the new energy model as GAknot 2.0 and we can easily evaluate the accuracy of the proposed new energy model by comparing to the original GAknot, which also means we can easily compare our new energy model and DP09 through GAknot. We also compare the results of GAknot 2.0 with two state-of-the-art algorithms, which are HotKnots and IPknot respectively. We choose these two algorithms because they are the state-of-the-art and get quite good accuracies on predicting pseudoknots. Moreover, they use different energy models as the scoring functions. HotKnots uses DP09 and IPknot uses DP.

To evaluate the performance of prediction methods, sensitivity (Sen) and positive predictive value (PPV) are used. The definition of Sensitivity and PPV are given as follows:

$$PPV = \frac{TP}{TP + FP}, Sen = \frac{TP}{TP + FN}$$

where TP (true positive) is the number of correctly predicted base pairs, FP (false positive) is the number of incorrectly predicted base pairs, and FN (false negative) is the number of base pairs in the known structure that were not predicted. In this evaluation, we run both GAknot and GAknot 2.0 10 times to get average values of PPV and sensitivity because they both base on genetic algorithm which is a stochastic algorithm and 10 times are a reasonable number of runs to show the stability of a stochastic algorithm.

TABLE I

10 RUNS OF GAKNOT 2.0 ON HK41. BOLD VALUES INDICATE THE BEST RESULT AMONG 10 RUNS.

Run	1	2	3	4	5	6	7	8	9	10	ave
Sen (%)	83.4	82.3	81.0	83.7	84.0	82.9	84.4	**85.2**	82.2	82.9	83.2
PPV (%)	75.8	74.9	73.5	75.9	76.8	75.2	77.1	**77.1**	75.5	75.5	75.7

TABLE II

COMPARISON OF DIFFERENT ALGORITHMS ON HK41. BOLD VALUES INDICATE THE BEST RESULT AMONG THE ALGORITHMS.

Algorithms	Sen (%)	PPV (%)
GAKnot 2.0 (average of 10 runs)	**83.2**	**75.7**
GAKnot (average of 10 runs)	80.1	74.9
HotKnots	64.6	68.4
IPknot with DP model	50.8	57.5

We test the algorithms on dataset HK41 defined in Section III A. Table I shows the results of GAknot 2.0 (GAknot using our proposed new energy model) in 10 runs on HK41 and the average result. From this table, the best result (run 8) has 85.2% sensitivity and 77.1% PPV. The average of the 10 runs also has 83.2% sensitivity and 75.7% PPV. In addition, the performances of GAknot 2.0 are very stable among 10 runs. Table II shows the comparison of different algorithms on this dataset, in terms of PPV and sensitivity. Note that we test IPknot using DP model except the second sequence (PDB ID: 1Y0Q), which is predicted using McCaskill model [18]. The reason is due to the length of 1Y0Q, which has 299 nt and it is too long for IPknot to use DP model. From Table II, GAknot 2.0 is the best algorithm and it is shown that the new energy model can improve the prediction sensitivity by 3.1% (80.1% to 83.2%) and PPV by 0.8% (74.9% to 75.7%), compared to original GAknot (which uses DP09). We can see that even the worst result of GAknot 2.0 (run 3 in Table I) is better than the average result of GAknot using DP09. Therefore, from this dataset and using the same algorithm, the accuracy of the new energy model is better than DP09. In addition, GAknot 2.0 is better than the two state-of-the-art algorithms, which are using erengy models DP09 and DP respectively.

IV. DISCUSSION AND CONCLUSION

From the results in Section III, we can conclude that the performance of the new model is better than that of DP09 and the original DP model when dealing with pseudoknots prediction. The model is still not perfect, since it cannot get 100% accuracy, but it can give some hints on designing a more accurate energy model. More specifically, we can see that by recognizing the stems patterns, which are k-mers in this work, we can capture which patterns can easily form stems and we can exploit it to improve the energy model.

In this article, we have proposed a new energy model for RNA secondary structure prediction with pseudoknots. We have shown that how we use the existing, validated RNA secondary structure to formulate penalty factors to improve the prediction accuracy by the concept of k-mers' normalized occurrences. For future improvements, we may collect more validated data to improve the factors and we may try other ways to get more information from the data and not just limited to k-mers.

ACKNOWLEDGMENT

This research is partially supported by the Direct Grant of CUHK and the General Research Fund (Project Number: LU310111) of Hong Kong SAR, China.

REFERENCES

[1] Jr, I. T. and Bustamante, C. (1999) How RNA folds. *Journal of Molecular Biology,* **293**(2), 271 – 281.

[2] Shapiro, B. A., Yingling, Y. G., Kasprzak, W., and Bindewald, E. (2007) Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology,* **17**(2), 157 – 165 Theory and simulation / Macromolecular assemblages.

[3] van Batenburg, F. H. D., Gultyaev, A. P., and Pleij, C. W. A. (2001) PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Research,* **29**(1), 194–195.

[4] Chen, J.-L. and Greider, C. W. (2005) Functional analysis of the pseudoknot structure in human telomerase RNA. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(23), 8080–8085.

[5] Sakuragi, J., Ode, H., Sakuragi, S., Shioda, T., and Sato, H. (2012) A proposal for a new HIV-1 DLS structural model. *Nucleic Acids Research,*.

[6] Brierley, I., Gilbert, R., and Pennell, S. (2008) RNA pseudoknots and the regulation of protein synthesis. *Biochemical Society Transactions,* **36**(4), 684–689.

[7] Staple, D. W. and Butcher, S. E. (06, 2005) Pseudoknots: RNA Structures with Diverse Functions. *PLoS Biol,* **3**(6), e213.

[8] Giedroc, D. P. and Cornish, P. V. (2009) Frameshifting RNA pseudoknots: Structure and mechanism. *Virus Research,* **139**(2), 193 – 208 Structural motifs controlling the replication cycle of positive strand RNA viruses.

[9] Giedroc, D. P., Theimer, C. A., and Nixon, P. L. (2000) Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *Journal of Molecular Biology,* **298**(2), 167 – 185.

[10] Lyngsø, R. and Pedersen, C. (2000) RNA pseudoknot prediction in energy-based models. *Journal of computational biology,* **7**(3-4), 409–427.

[11] Andronescu, M., Pop, C., and Condon, A. (2010) Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA,* **16**(1), 26–42.

[12] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology,* **288**(5), 911 – 940.

[13] Dirks, R. M. and Pierce, N. A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry,* **24**(13), 1664–1677.

[14] Andronescu, M., Bereg, V., Hoos, H., and Condon, A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC bioinformatics,* **9**(1), 340.

[15] Tong, K. K., Cheung, K. Y., Lee, K. H., and Leung, K. S. (2013) GAknot: RNA secondary structures prediction with pseudoknots using Genetic Algorithm. *Proceedings of Computational Intelligence in Bioinformatics and Computational Biology 2013,*.

[16] Sato, K., Kato, Y., Hamada, M., Akutsu, T., and Asai, K. (2011) IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics,* **27**(13), i85–i93.

[17] Ren, J., Rastegari, B., Condon, A., and Hoos, H. (2005) HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *Rna,* **11**(10), 1494–1504.

[18] McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* **29**(6-7), 1105–1119.