Chapter 3. Using Log Linear Models to Study Evolutionary Influence on RNA Secondary Structure and Alignment

Abstract

Comparative analysis is widely used to predict RNA secondary structures. It is based on the assumption that RNA sequences from a common ancestor that serve the same function have conserved structures. A machine learning method called a log linear model was used to quantify the structural conservation among RNA sequences of common ancestry. For two RNA sequences that descended from a common ancestor, the log linear model was used to quantify 1) the structural constraint imposed by evolution on first sequence of the pair when its alignment with and the structure of the second sequence are known; 2) the constraint imposed by evolution on the alignment when the structures of the two sequences are known. In order to model the structural constraint, an extra set of free energy parameters were added to the free energy model of the first sequence. The parameters represent the influence of alignment between the two sequences and the second sequence's structure. In order to model the constraint imposed by evolution on a pairwise RNA sequence alignment with the structures of the two RNA sequences known, extra parameters were added to the hidden Markov models (HMM) for the alignment of the two sequences. Maximum likelihood estimation was used to estimate both sets of extra parameters. The estimation showed the extra parameters favored the structures or alignments that conform to the RNA structural conservation.

Introduction

The log linear model is a machine learning model that can model the probability of a set of labels on sequence inputs (*e.g.*, RNA secondary structure on sequences, grammatical structure on sentences). It has its origin in natural language processing (1) and can be applied in document classification (2) and part-of-speech tagging (3). Many other machine learning methods also have their origins or the most important application in natural language processing, *e.g.*, stochastic context free grammar (SCFG) (4) and HMM (5). In this project, the log linear model is used to model RNA secondary structure and sequence alignment under evolutionary constraint.

The model has the general form:

$$P(y'|x) = \frac{e^{-\sum_i n_{y'}(\theta_i) f(\theta_i)}}{\sum_y e^{-\sum_i n_y(\theta_i) f(\theta_i)}} \quad \mathbf{1}$$

where x is the sequence input which can be nucleotide sequences and y is the label on the sequence input which can be RNA secondary structure of the sequence input or alignment between them. θ_i is a certain feature dependent on the sequence input and the label. $n_y(\theta_i)$ is the number of times θ_i appears given the label and the input. $f(\theta_i)$ is the parameter associated with θ_i. $P(y'|x)$ is the conditional probability of label y' given the sequence input x. Because the log linear model is flexible in terms of how the features can be extracted, a lot of models can be seen as log linear models. Both the free energy model of RNA secondary structure and HMM sequence alignment can be viewed as log linear models of which the derivation will be shown later. A maximum likelihood

estimation method can be used to estimate the parameters $\{f(\theta_i)\}$ given a set of sequence inputs and their known label (6).

In automated comparative analysis methods, there are parameters that represent structural evolution, *e.g.*, gap penalties in Dynalign (7) and Dynalign II (8) and probabilities of SCFG (9,10) production rules. The values of these parameters quantify the influence of evolution on RNA structure and sequence alignment. Because both the free energy model of RNA secondary structure and HMM for sequence alignment can be viewed as log linear models, the log linear models were used to expand the structure and alignment models to include the influence of evolution.

For studying the influence of evolution on RNA secondary structure, the probability of a structure given a sequence can be modeled using a set of nearest neighbors which correspond to $\{\theta_i\}$ in log linear models. The free energies of the nearest neighbors $\{\Delta G(\theta_i)\}$ correspond to $\{f(\theta_i)\}$. When the sequence to be modeled with the log linear model appears in a sequence pair as the first sequence, and the structure of the second sequence and the alignment between the two sequences are known, the set of features $\{\theta_i\}$ in the log linear model were expanded to include features representing the influence of evolution.

For studying the influence of evolution on RNA sequence alignment, the HMM probability of an alignment between two sequences can be modeled using a log linear model as well. The set of features $\{\theta_i\}$ correspond to nucleotide alignments/insertion and $\{f(\theta_i)\}$ correspond to the HMM probability parameters. If the structures of the two

sequences are known, then to model the influence of evolution, the set of features and the parameters were expanded to model the evolutionary influence on the alignment.

A maximum likelihood method (6) was used to estimate the added parameters in the expanded log linear models. The results showed evolution favors RNA secondary structures to be conserved.

Methods

Quantify evolutionary constraints on RNA structures

For a certain structure π' of an RNA sequence s, the structure's probability given the sequence, $P(\pi'|s)$, can be modeled using:

$$P(\pi'|s) = \frac{e^{\frac{-\Delta G(\pi')}{RT}}}{\sum_{\pi} e^{\frac{-\Delta G(\pi)}{RT}}} \quad \mathbf{2}$$

where $\Delta G(\pi)$ is the free energy change of structure π. The denominator sums the Boltzmann factors of all the possible structures of sequence s and is called the partition function. Using nearest neighbor models (11,12), $P(\pi'|s)$ can be further modeled as:

$$P(\pi'|s) = \frac{e^{\frac{-\sum_i n_{\pi'}(\theta_i)\Delta G(\theta_i)}{RT}}}{\sum_{\pi} e^{\frac{-\sum_i n_{\pi}(\theta_i)\Delta G(\theta_i)}{RT}}} \quad \mathbf{3}$$

where θ_i represents a certain nearest neighbor, $\Delta G(\theta_i)$ is the free energy change parameter associated with nearest neighbor θ_i and $n_\pi(\theta_i)$ is the number of times θ_i appears in structure π. This probability only includes terms describing the intrinsic

properties of a sequence *s*. When a pair of sequences is considered, in order to describe the constraint brought by the structural conservation, when the second sequence's structure and the pairwise alignment of the two RNA sequences are known, a modified equation which is conceived by the author of the thesis can be used to describe $P(\pi'|s)$:

$$P(\pi'|s) = \frac{e^{\frac{-\sum_i n_{\pi'}(\theta_i)\Delta G(\theta_i) - \sum_k n_{\pi'}(\varphi_k)\Delta G(\varphi_k)}{RT}}}{\sum_\pi e^{\frac{-\sum_i n_\pi(\theta_i)\Delta G(\theta_i) - \sum_k n_\pi(\varphi_k)\Delta G(\varphi_k)}{RT}}} \quad \mathbf{4}$$

where *s* is the first sequence, φ_k represents the structural evolutionary constraint when the structure of the second RNA sequence and its alignment to the first sequence are known. $n_\pi(\varphi_k)$ represents the number of times φ_k appears in structure π and $\Delta G(\varphi_k)$ is the free energy parameter associated with the constraint φ_k. Using this model, from a database of multiple RNA homologs whose structures and alignments are known, we drew a sample of sequence pairs and used a maximum likelihood method to estimate $\{\Delta G(\varphi_k)\}$. The more negative $\Delta G(\varphi_k)$ is, the more favorable φ_k is.

For the sequence pair given in Figure 3.1 where pairwise RNA structural alignment and the structure of the second sequence (the bottom one in Figure 3.1) are known, when its partition function is calculated, the Boltzmann factor of any structure containing the red base pair will be multiplied by a parameter ($exp(-\Delta G(\varphi_1)/RT)$), representing the influence of base pair conservation, which is denoted as φ_1; the Boltzmann factor of any structure containing the blue base pair will be multiplied by a parameter ($exp(-\Delta G(\varphi_2)/RT)$) representing the influence of base pair insertion. The list of parameters $\{\Delta G(\varphi_k)\}$ is in Table 3.1.

The probability of a certain structure π' of a certain sequence s, $P(\pi'|s)$, therefore is modeled using equation 4 instead of equation 3 under the evolutionary constraint of the second sequence. In order to estimate the parameters $\{\Delta G(\varphi_k)\}$, a maximum likelihood method was used (6). With a training set of sequence pairs whose structures and alignments are known and one sequence for each pair is denoted as the first sequence, the objective that is maximized is:

$$L = \sum_i logP(\pi_i' | s_i) - \frac{1}{2}\alpha \sum_k \left[\frac{\Delta G(\varphi_k)}{RT}\right]^2 \quad \textbf{5}$$

where i is the index of a sequence pair, s_i is the first sequence of the i^{th} sequence pair and π_i' is the known structure of the first sequence of the i^{th} sequence pair. The second term in the right half of the equation is the L_2 regularization term with α as the hyperparameter. The objective function was optimized *w.r.t.* $\{\Delta G(\varphi_k)/\mathrm{RT}\}$. The optimization method for log-linear models is well established (6). The gradient of the objective function is:

$$\frac{\partial L}{\partial \frac{\Delta G(\varphi_k)}{RT}} = \sum_i [n_{\pi_i'}(\varphi_k) - E\left(n_{\pi_i}(\varphi_k)|s_i\right)] - \alpha \frac{\Delta G(\varphi_k)}{RT} \quad \textbf{6}$$

where $n_{\pi i'}(\varphi_k)$ is the number of times φ_k appears in π_i', and $E(n_{\pi i}(\varphi_k)|s_i)$ is the expected number of times φ_k appreas in s_i. $E(n_{\pi i}(\varphi_k)|s_i)$ can be calculated using dynamic programming algorithm. For a φ_k involving the forming of a base pair,

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = E\left(\sum_{p,q} I_{\pi_i}(p,q) \,|p,q \text{ base pair conforms to } \varphi_k, s_i\right) \quad \textbf{7}$$

where $I_{\pi i}(p,q)$ is the indicator variable that is 1 when the p^{th} and q^{th} nucleotides form a base pair in π_i and 0 when the p^{th} and q^{th} nucleotide do not. With some rearrangements, we can know that

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = \sum_{p,q \text{ base pair conforms to } \varphi_k} P_{bp}(p,q|s_i) \quad \textbf{8}$$

where $P_{bp}(p, q| s_i)$ is probability that the p^{th} and q^{th} nucleotides form a base pair given s_i which can be calculated using dynamic programming algorithm(13). For a φ_k involving the forming a single stranded nucleotide,

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = E\left(\sum_p I_{\pi_i}(p) \,|p \text{ conforms to } \varphi_k, s_i\right) \quad \textbf{9}$$

where $I_{\pi i}(p)$ is the indicator variable that is 1 when the p^{th} nucleotide forms a single stranded nucleotide in π_i and 0 when the p^{th} nucleotide does not. With some rearrangements, we can know that

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = \sum_{p \text{ nucleotide conforms to } \varphi_k} P_s(p|s_i) \quad \textbf{10}$$

where $P_s(p| s_i)$ is the probability that the p^{th} nucleotide is single stranded given s_i which can be calculated using dynamic programming algorithm.

Quantify evolutionary constraints on RNA sequence alignment

An analogy can be made to model the alignment constraint of two RNA sequences given their structures are known. A pair-HMM (14) used to model a pairwise sequence alignment can be written as:

$$P(A, s_1, s_2) = \prod_k P(\mu_k)^{n_A(\mu_k)} \quad \textbf{11}$$

where $P(A,s_1,s_2)$ is the joint probability of the two homologous sequences and the alignment A. μ_k is a certain action of the HMM model, i.e., transition between states and emissions. $P(\mu_k)$ is the probability associated with the action. $n_A(\mu_k)$ is the number of times μ_k appears in the alignment A. Because

$$P(A, s_1, s_2) = e^{\sum_k n_A(\mu_k) logP(\mu_k)} \quad \textbf{12}$$

and

$$P(A'|s_1, s_2) = \frac{P(A',s_1,s_2)}{\sum_A P(A,s_1,s_2)}, \quad \textbf{13}$$

we can derive that:

$$P(A'|s_1, s_2) = \frac{e^{\sum_k n_{A'}(\mu_k) logP(\mu_k)}}{\sum_A e^{\sum_k n_A(\mu_k) logP(\mu_k)}} \quad \textbf{14}$$

As shown, the equation of the conditional probability of a certain alignment A' given the sequences, s_1 and s_2 is very similar to the equation of the conditional probability of an RNA structure given the sequence. They both conform to a log-linear model. Therefore, we can use the same method to model alignment constraint given the structures of s_1 and s_2 by adding extra parameters. The resulting modified equation is:

$$P(A'|s_1, s_2) = \frac{e^{\sum_k n_{A'}(\mu_k) logP(\mu_k) + \sum_l n_{A'}(\tau_l) logP(\tau_l)}}{\sum_A e^{\sum_k n_A(\mu_k) logP(\mu_k) + \sum_l n_A(\tau_l) logP(\tau_l)}} \quad \textbf{15}$$

where τ_l represents a certain evolutionary constraint on alignments. For the pair of RNA structures given in Figure 3.2, when running the forward-backward algorithm of the pair HMM algorithm for alignment of the two RNA sequences, extra terms can be

multiplied to emission probabilities of aligned nucleotides or inserted nucleotides according to the structures of the two sequences. For example, the emission probability of the red nucleotide pair will be multiplied by $P(\tau_1)$, representing the influence of conserved structures on alignment; The emission probability of the blue nucleotide pair will be multiplied by $P(\tau_2)$, representing the influence of a base paired nucleotide mutating into an unpaired one on alignment. The whole list of parameters $\{P(\tau_l)\}$ is in Table 3.2.

The probability of a alignment *A'* given sequences, s_1 and s_2, $P(A'|s_1, s_2)$ therefore is modeled using equation 15 instead of equation 14. In order to estimate the parameters $\{ logP(\tau_l)\}$, a maximum likelihood method was used. With a training set of sequence pairs whose structures and alignments are known, the objective that is maximized is:

$$L = \sum_i logP(A_i{}'|\, s_{i1}, s_{i2}) - \frac{1}{2}\alpha \sum_l [logP(\tau_l)]^2 \qquad \textbf{16}$$

where i is the index of a sequence pair, s_{i1} and s_{i2} are the sequences in the pair and A_i' is the known alignment of the i^{th} sequence pair. The second term in the right half of the equation is the L_2 regularization term with α as the hyperparameter. The objective function was optimized *w.r.t.* $\{logP(\tau_l)\}$. The gradient of the objective function is:

$$\frac{\partial L}{\partial logP(\tau_l)} = \sum_i \left[E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) - n_{A_i{}'}(\tau_l)\right] - \alpha logP(\tau_l) \qquad \textbf{17}$$

where $n_{Ai'}(\tau_l)$ is the number of times τ_l appears in A_i', and $E(n_{Ai}(\tau_l)|s_{i1}, s_{i2})$ is the expected number of times τ_l appears in A_i. For a τ_l involving alignment of two nucleotides,

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) = E\left(\sum_{p,q \text{ nucleotide alignment conforms to } \tau_l} I_{A_i}(p,q) \mid s_{i1}, s_{i2}\right) \quad \textbf{18}$$

where $I_{Ai}(p,q)$ is the indicator variable that is 1 when the p^{th} nucleotide in the 1^{st} sequence and q^{th} nucleotide in the 2^{nd} sequence are aligned in A_i and 0 when the two nucleotides are not. With some rearrangements, we can know that

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) = \sum_{p,q \text{ nucleotide alignment conforms to } \tau_l} P_a(p,q|s_{i1}, s_{i2}) \quad \textbf{19}$$

where $P_a(p, q| s_{i1}, s_{i2})$ is the probability that the p^{th} nucleotide in the 1^{st} sequence and q^{th} nucleotide in the 2^{nd} sequence are aligned given s_{i1} and s_{i2} which can be calculated using the forward backward algorithm. For a τ_l involving the insertion of a nucleotide,

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) =$$
$$E\left(\sum_{p \text{ nucleotide insertion conforms to } \tau_l} I1_{A_i}(p) + \right.$$
$$\left.\sum_{q \text{ nucleotide insertion conforms to } \tau_l} I2_{A_i}(q) \,|s_{i1}, s_{i2}\right) \quad \textbf{20}$$

where $I1_{Ai}(p)$ is the indicator variable that is 1 when the p^{th} nucleotide in the 1^{st} sequence is inserted in the alignment A_i and 0 when the nucleotide is not inserted. $I2_{Ai}(q)$ has a similar meaning for the 2^{nd} sequence. With some rearrangements, we can know that

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) =$$
$$\sum_{p \text{ nucleotide insertion conforms to } \tau_l} P1_{ins}(p) + \sum_{q \text{ nucleotide insertion conforms to } \tau_l} P2_{ins}(q)$$

21

where $P1_{ins}(p)$ is the probability that the p^{th} nucleotide in the 1st sequence is which can be calculated as:

$$P1_{ins}(p) = 1 - \sum_q P_a(p, q|s_{i1}, s_{i2}) \quad \textbf{22}$$

Optimization

Because it has been shown log liner models are concave (6), an implementation of a quasi-Newton method called limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) in the dlib c++ library (15) was used along with the implementations of both the objective functions and their gradients to optimize the parameters $\{\Delta G(\varphi_k)/\mathrm{RT}\}$ and $\{logP(\tau_l)\}$.

Results

50 tRNA sequence pairs and 50 5S rRNA sequence pairs with their known structures and alignment were drawn from the databases (16,17) as the training set for both models. A grid search was performed for different hyperparameters in both models within $\{0, 1 \times 10^{-5}, 1 \times 10^{-4}, \ldots, 1 \times 10^{10}\}$. The performances for different hyperparameters were evaluated on a separate data set of 50 tRNA and 50 5S rRNA sequence pairs using the objective functions without the regularization of the separate data set for both models. The optimal hyperparameters (α in both equations) for both models are 0. The optimal parameters are shown in Tables 3.1 and 3.2.

As expected, the optimal parameters for both models favor conserved alignments and structures. For the structure model, the most favored structure by the optimal parameters is an unpaired nucleotide aligned with another one ($exp(-\Delta G(\varphi_5)/RT)$=3.23), and the most disfavored structures by the optimal parameters are a base pair aligned with two unpaired nucleotides($exp(-\Delta G(\varphi_3)/RT)$=0.55) and an unpaired nucleotide aligned with a paired one ($exp(-\Delta G(\varphi_7)/RT)$=0.59). For the alignment model, the most favored alignment by the optimal parameters is the alignment of two nucleotides paired with other nucleotides in the same direction ($P(\tau_1)$=4.60), and the most disfavored alignment is a paired nucleotide aligned with an unpaired nucleotide ($P(\tau_2)$=0.21). From the optimal parameters, we can see that evolution favors conserved structural alignment, and disfavored structural mutation (unpaired mutated to paired/paired mutated to unpaired). Surprisingly, insertions/deletions tend to be neither favored nor disfavored by the optimal parameters.

Discussion

The parameters are optimized using maximum likelihood estimation, meaning 1)for studying structural evolutionary constraint, given two sequences with the structure of the second sequence and the alignment known, the optimized parameters on average give the biggest conditional probability of the known structure of the first sequence; 2)for studying sequence alignment evolutionary constraint, given two sequences with known structures, the optimized parameters on average give the biggest conditional probability of the known pairwise alignment. This means the log linear model with

these optimized parameters explains the training data best and the parameters quantify the evolutionary constraint on RNA structure and alignment best. And because the parameters are extra terms added to the free energy model of RNA secondary structures and the pair HMM sequence alignment model, it is easy to compare the parameters to the free energy parameters and the HMM parameters and understand how important they are in terms of determining RNA secondary structure and sequence alignment.

References

1. Berger, A.L., Pietra, V.J.D. and Pietra, S.A.D. (1996) A maximum entropy approach to natural language processing. *Comput Linguist*, **22**, 39-71.
2. Pang, B., Lee, L. and Vaithyanathan, S. (2002), *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, pp. 79-86.
3. Ratnaparkhi, A. (1996), *Proceedings of the conference on empirical methods in natural language processing*. Philadelphia, USA, Vol. 1, pp. 133-142.
4. Baker, J.K. (1979) Trainable grammars for speech recognition. *J Acoust Soc Am* **65**, S132-S132.
5. Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, **77**, 257-286.
6. Smith, N.A. (2004) Log-linear models.
7. Mathews, D.H. and Turner, D.H. (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, **317**, 191-203.

8. Fu, Y., Sharma, G. and Mathews, D.H. (2014) Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res*, **42**, 13939-13948.

9. Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, **31**, 3423-3428.

10. Holmes, I. (2004) A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics*, **5**, 166.

11. Mathews, D.H. and Turner, D.H. (2002) Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, **41**, 869-880.

12. Xia, T., SantaLucia, J., Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C. and Turner, D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, **37**, 14719-14735.

13. Mathews, D. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, **10**, 1178-1190.

14. Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

15. King, D. (2009) Dlib-ml: A machine learning toolkit. *J Mach Learn Res*, **10**, 1755-1758.

16. Szymanski, M., Barciszewska, M.Z., Erdmann, V.A. and Barciszewski, J. (2002) 5S Ribosomal RNA Database. *Nucleic Acids Res*, **30**, 176-178.

17. Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res*, **37**, D159-162.

Figures

```
AGCGGUGAAAACACCGCU
A-CAG-GGA-CC-CAGAU
(-(-(-((--))-)-)-)
```

Figure 3.1. A pairwise alignment of 2 RNA sequences and the structure of the bottom one.

```
AGCGGUGAAAACACCGCU
(((((((|||)))))))
ACAGGGACCCAGAU
((|(((|)))|)|)
```

Figure 3.2. Two RNA structures.

Tables

Table 3.1. Parameters for modeling the influence of RNA homolog's structure and the structural alignment on RNA folding. "-" represents an alignment gap. "|" represents a single stranded nucleotide.

Symbol	Illustration	Explanation	Boltzman Factor Optimal Value	Free Energy Optimal Value
φ_1	A...U C...G (...)	A base pair aligned with a base pair	$exp(-\Delta G(\varphi_1)/RT)$ 1.30	$\Delta G(\varphi_1)$ -0.16kcal/mol
φ_2	A...U -...-	A base pair inserted	$exp(-\Delta G(\varphi_2)/RT)$ 1.00	$\Delta G(\varphi_2)$ 0.00kcal/mol
φ_3	A...U G...C |...|	A base pair aligned with two unpaired bases	$exp(-\Delta G(\varphi_3)/RT)$ 0.55	$\Delta G(\varphi_3)$ 0.37kcal/mol
φ_4	A...U G...- |...-	A base pair aligned with one base and one gap	$exp(-\Delta G(\varphi_4)/RT)$ 0.98	$\Delta G(\varphi_4)$ 0.01kcal/mol
φ_5	A G |	An unpaired base aligned with another	$exp(-\Delta G(\varphi_5)/RT)$ 3.23	$\Delta G(\varphi_5)$ -0.72kcal/mol
φ_6	A -	An unpaired base inserted	$exp(-\Delta G(\varphi_6)/RT)$ 1.02	$\Delta G(\varphi_6)$ -0.01kcal/mol
φ_7	A G (	An unpaired base aligned with a paired base	$exp(-\Delta G(\varphi_7)/RT)$ 0.59	$\Delta G(\varphi_6)$ 0.59kcal/mol

Table 3.2. Parameters for modeling the influence of RNA structures on their alignment. "**-**" represents an alignment gap. "l" represents a single stranded nucleotide.

Symbol	Illustration	Explanation	Multiplier Optimal Value
τ_1	**A** **(** **C** **(**	Alignment of bases with conserved structures	$P(\tau_1)$ 4.60
τ_2	**A** **(** **C** **|**	Paired base aligned with unpaired base	$P(\tau_2)$ 0.21
τ_3	**A** **|** **C** **|**	Alignment of unpaired bases	$P(\tau_3)$ 1.70
τ_4	**A** **(** -	A paired base inserted	$P(\tau_4)$ 0.23
τ_5	**A** **|** **–**	An unpaired base inserted	$P(\tau_4)$ 0.98