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 that descended from one common ancestor have conserved structures. A machine learning method called 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 on the target sequence when its alignment with and the structure of the other sequence of a common ancestry are known; 2) the constraint on the alignment when the structures of the two sequences of a common ancestry are known. 1) In order to model the structural constrain, an extra set of free energy parameters were added to the target. The parameters represent the influence of alignment between the target and the other sequence and the other sequence's structure. 2) In order to model the constraint on a pairwise RNA sequence alignment caused by the structures of two RNA sequences with a common ancestry whose structures are known, extra parameters were added to the hidden Markov models (HMM) for the alignment of the two sequences. These extra parameters represent the alignment constraint brought by the structures. 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 most accurate method to predict RNA secondary structures is comparative analysis where structures of RNA sequences are inferred using multiple homologs. There are also automated methods that take multiple RNA sequences as input, assuming they are homologs and predict their structures[1, 2]. These methods are based on the assumption that RNA sequences that descended from a common ancestor have conserved structures. In this project, a machine learning method, log-linear model[3], was used to quantify this conservation.

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

$$P(\pi'|s) = \frac{e^{\frac{-\Delta G(\pi')}{RT}}}{\sum_{\pi} e^{\frac{-\Delta G(\pi)}{RT}}} \quad (1)$$

where $\Delta G(\pi)$ is the free energy of structure π. The denominator sums up the Boltzmann factor of all the possible structures of sequence s. Using nearest neighbor models[4, 5], $P(\pi'|s)$ can be further modeled as:

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

where θ_i represents a certain nearest neighbor, $\Delta G(\theta_i)$ is the free energy 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*. In order to describe the constraint brought by the structural conservation, when another RNA homolog's structure and the pairwise alignment of the two RNA sequences are known, a modified equation can be used to describe $P(\pi'|s)$:

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

where φ_k represents the a certain structural constraint brought by another RNA homolog when its structure and their alignment are known. 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 free energy parameters associated with the structural constraints. The smaller $\Delta G(\varphi_k)$ is, the more favorable φ_k is.

An analogy can be made to model the alignment constraint of two RNA sequences given their structures are known. A pair-HMM[6] which is used to model a pairwise sequence alignment can be written as:

$$P(A, s_1, s_2) = \prod_k p(\mu_k)^{n_A(\mu_k)} \quad (4)$$

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*. With some mathematical manipulations, we can derive that:

$$P(A'|s_1, s_2) = \frac{e^{\sum_k n_{A'}(\mu_k) log p(\mu_k)}}{\sum_A e^{\sum_k n_A(\mu_k) log p(\mu_k)}} \quad (5)$$

as we can see, 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 a RNA structure given the sequence. They both conform to 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:

$$P(A'|s_1, s_2) = \frac{e^{\sum_k n_{A'}(\mu_k) log p(\mu_k) + \sum_l n_{A'}(\tau_l) log p(\tau_l)}}{\sum_A e^{\sum_k n_A(\mu_k) log p(\mu_k) + \sum_l n_A(\tau_l) log p(\tau_l)}} \quad (6)$$

where τ_l represents a certain constraint on alignments. Likewise, the same optimization method was used to optimize the model of alignment constraints.

Methods

Quantify evolutionary constraints on RNA structures

Given a pairwise RNA structural alignment and the structure of one of the sequence (the bottom one in Fig. 1),

AGCGGUGAAAACACCGCU
A–CAG–GGA–CC–CAGAU
(– (– (– ((– –)) –) –) –)

Fig 1. A pairwise alignment of 2 RNA sequences and the structure of the bottom one.

a log-linear model can be used to model the folding of the other sequence (the top one). The partition function[7] of the folding of the top sequence can be calculated with extra free energy parameters added according to the pairwise alignment and the structure of the bottom sequence. For example, when calculating the partition function, 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 etc. The whole list of parameters are in Table 1.

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 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.

The probability of a certain structure π' of a certain sequence s $P(\pi'|s)$, therefore changes from (2) to (3). In order to estimate the parameters $\{\Delta G(\varphi_k)\}$, a maximum likelihood method was used[3]. With a training set of sequence pairs whose structures and alignments are known and one sequence for each pair is denoted as the target sequence, the objective that is maximized is:

$$L = \sum_i logP(\pi_i' | s_i) - \frac{1}{2}\alpha \sum_k \left[\frac{\Delta G(\varphi_k)}{RT}\right]^2 \quad (7)$$

where i is the index of a sequence pair, s_i is the target sequence of the i^{th} sequence pair and π_i' is the known structure of the target 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[3]. The gradient of the objective function is:

$$\frac{\partial L}{\partial \frac{\Delta G(\varphi_k)}{RT}} = \sum_i [n_{\pi_i'}(\varphi_k) - E\left(n_{\pi_i}(\varphi_k)|s_i\right)] - \alpha \frac{\Delta G(\varphi_k)}{RT} \quad (8)$$

where $n_{\pi i'}(\varphi_k)$ is the number of times φ_k appreas 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 a base pair,

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = E(\sum_{p,q} I_{\pi_i}(p,q)\,|p,q \text{ base pair conforms to } \varphi_k, s_i) \quad (9)$$

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

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = \sum_{p,q \text{ base pair conforms to } \varphi_k} p_{bp}(p,q|s_i) \quad (10)$$

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[7]. For a φ_k involving the forming a single stranded nucleotide,

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = E(\sum_p I_{\pi_i}(p)\,|p \text{ conforms to } \varphi_k, s_i) \quad (11)$$

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

$$E\left(n_{\pi_i}(\varphi_k)|s_i\right) = \sum_{p \text{ nucleotide conforms to } \varphi_k} p_s(p|s_i) \quad (12)$$

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

Given two RNA structures,

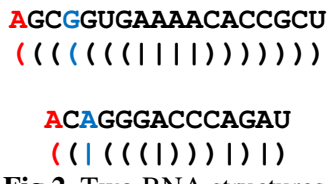

Fig 2. Two RNA structures.

a log-linear model can be used to model the sequence alignment between them. When running the forward-backward algorithm of the pair HMM algorithm for sequence alignment, extra terms can be multiplied to emission probabilities of aligned nucleotides or inserted nucleotides according to the structures of the two sequences. For example, when running the forward-backward algorithm, 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 is in Table 2.

The probability of a certain alignment *A'* given certain sequences, s_1 and s_2, $P(A'|s_1, s_2)$ therefore changes from (5) to (6). 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

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

Table 2. Parameters for modeling the influence of RNA structures on their alignment. "**-**" represents an alignment gap. "|" represents a single stranded nucleotide.

that is maximized is:

$$L = \sum_i logP(A_i' \mid s_{i1}, s_{i2}) - \frac{1}{2}\alpha \sum_l [logp(\tau_l)]^2 \quad (13)$$

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:

$$\frac{\partial L}{\partial logp(\tau_l)} = \sum_i [E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) - n_{A_i'}(\tau_l)] - \alpha logp(\tau_l) \quad (14)$$

where $n_{Ai'}(\tau_l)$ is the number of times τ_l appreas in A_i', and $E(n_{Ai}(\tau_l)|s_{i1}, s_{i2})$ is the expected number of times τ_l appreas in A_i. $E(n_{Ai}(\tau_l)|s_{i1}, s_{i2})$ can be calculated using dynamic programming algorithm (forward-backward algorithm)[6]. For a τ_l involving alignment of two nucleotides,

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) = E(\sum_{p,q \text{ nucleotide alignment conforms to } \tau_l} I_{A_i}(p, q) \mid s_{i1}, s_{i2}) \quad (15)$$

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

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) = \sum_{p,q \text{ nucleotide alignment conforms to } \tau_l} p_a(p, q|s_{i1}, s_{i2}) \quad (16)$$

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

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) = E\left(\sum_{p \text{ nucleotide insertion conforms to } \tau_l} I1_{A_i}(p) + \sum_{q \text{ nucleotide insertion conforms to } \tau_l} I2_{A_i}(q) \,|s_{i1}, s_{i2}\right) \quad (17)$$

where $I1_{Ai}(p)$ is the indicator variable that is 1 when the p^{th} nucleotide in the 1st 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 2nd sequence. With some rearrangements, we can know that

$$E\left(n_{A_i}(\tau_l)|s_{i1}, s_{i2}\right) = \sum_{p \text{ nucleotide insertion conforms to } \tau_l} p1_{ins}(p) + \sum_{q \text{ nucleotide insertion conforms to } \tau_l} p2_{ins}(q) \quad (18)$$

where $p1_{ins}(p)$ is the probability that the p^{th} nucleotide in the 1st sequence is which can be calculated using dynamic programming algorithm.

Optimization

Because it has been shown log liner models are convex[3]. Therefore an implementation of a quasi-Newton method limited-memory BFGS in dlib c++ library[8] was used along with the implementations of the both 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[9, 10] as the training set for both models. A grid search was performed for different hyperparameters in both models within $\{0, 1e\text{-}5, 0.0001, \ldots, 1e10\}$. 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 Table 1 and Table 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 paramters, 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

Apart from studying evolutionary influence on RNA structures and alignments, the optimized parameters can also be used for RNA alignment/conserved structure prediction. The optimal parameters for the alignment model can be directly used for pair HMM alignment when the structures of the two RNA sequences are involved. The optimal parameters for the structure model, with some adaption, can be used for conserved RNA structure prediction. For example, in Dynalign[11], the sum of the two input RNA sequences' structures and the penalty for alignment gaps is minimized:

$$\sum_i n_{\pi_1}(\theta_i)\Delta G(\theta_i) + \sum_i n_{\pi_2}(\theta_i)\Delta G(\theta_i) + n_A(gap)\Delta G_{gap} \quad (19)$$

where π_1 is the structure is 1st sequence and π_2 is the structure is 2nd sequence. ΔG_{gap} is the penalty for a single gap and $n_A(gap)$ is the number of gaps in the alignment. This can be turned into:

$$\sum_i n_{\pi_1}(\theta_i)\Delta G(\theta_i) + \alpha \sum_k n_{\pi_1}(\varphi_k)\Delta G(\varphi_k) + \sum_i n_{\pi_2}(\theta_i)\Delta G(\theta_i) + \alpha \sum_k n_{\pi_2}(\varphi_k)\Delta G(\varphi_k) \quad (20)$$

where α can act as a scaling factor for the alignment penalty.

1. Havgaard JH, Gorodkin J. RNA structural alignments, part I: Sankoff-based approaches for structural alignments. Methods Mol Biol. 2014;1097:275-90.
2. Asai K, Hamada M. RNA structural alignments, part II: non-Sankoff approaches for structural alignments. Methods Mol Biol. 2014;1097:291-301.
3. Smith NA. Log-linear models. Cited by. 2004;3.
4. Mathews DH, Turner DH. Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. Biochemistry. 2002;41(3):869-80. PubMed PMID: 11790109.
5. Xia T, SantaLucia J, Jr., Burkard ME, Kierzek R, Schroeder SJ, Jiao X, et al. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998;37(42):14719-35. doi: 10.1021/bi9809425. PubMed PMID: 9778347.
6. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids: Cambridge university press; 1998.
7. Mathews D. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. 2004;10(8):1178-90.
8. King D. Dlib-ml: A machine learning toolkit. J Mach Learn Res. 2009;10:1755-8.
9. Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. 5S Ribosomal RNA Database. Nucleic Acids Res. 2002;30(1):176-8. PubMed PMID: 11752286; PubMed Central PMCID: PMC99124.
10. Juhling F, Morl M, Hartmann RK, Sprinzl M, Stadler PF, Putz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009;37(Database issue):D159-62. doi: 10.1093/nar/gkn772. PubMed PMID: 18957446; PubMed Central PMCID: PMC2686557.
11. Mathews DH, Turner DH. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol. 2002;317(2):191-203. doi: 10.1006/jmbi.2001.5351. PubMed PMID: 11902836.