Using a Log Linear Model to Study the Influence of Evolutionary Constraints on RNA Secondary Structures

Yinghan Fu

Department of Biochemistry & Biophysics and Center for RNA Biology, University of Rochester Medical Center, Rochester, New York, United States of America

Introduction

Homologous RNA sequences have conserved structures. Numerous RNA secondary structure prediction methods have been developed to exploit this feature [1, 2]. In this project, a log linear model was used to study the influence of evolutionary constraint on RNA secondary structures. A set of parameters that represent the influence of evolutionary constraints on RNA secondary structures were optimized using maximum likelihood method. The method (conserved_training) was implemented as part of the author's version of RNAstructure available at https://github.com/franfyh.

Methods and Results

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 [3] of the folding of the top sequence can be calculated with extra 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 representing the influence of base pair conservation; the Boltzmann factor of any structure containing the blue base pair will be multiplied by a parameter representing the influence of base pair insertion etc.

There are seven parameters in the model: 1) the parameter for inserted single stranded nucleotides (for the gray nucleotide to be single stranded, denoted as $para_1$); 2) the parameter for conserved single stranded nucleotides (for the purple nucleotide to be single stranded, denoted as $para_2$); 3) the parameter for inserted base pairs (for the blue base pair to form, denoted as $para_3$); 4) the parameter for conserved base pairs (for the red base pair to form, denoted as $para_4$); 5) the parameter for single stranded nucleotides aligned with structured nucleotides (for the brown nucleotide to be single stranded, denoted as $para_5$); 6) the parameter for a base pair aligned with two unpaired nucleotides (for the yellow base pair to form, denoted as $para_6$); 7) the parameter for a base pair aligned with one singled stranded nucleotide and a gap (for the green base pair to form, denoted as $para_7$). The probability of any structure π' when its alignment with the other sequence and the structure of the other sequence is known can be modelled as:

$$P(\pi' | \boldsymbol{\theta}, \boldsymbol{para}, \Pi, \mathrm{A}) = \frac{e^{\sum_l n_l(\pi')\theta_l + \sum_k m_k(\pi') ln(para_k)}}{\sum_{\pi} e^{\sum_l n_l(\pi)\theta_l + \sum_k m_k(\pi) ln(para_k)}} \quad (1)$$

where $\boldsymbol{\theta}$ is the vector of all the thermodynamic parameters for different secondary structure units, ***para*** is the vector of all the parameters influencing base pairs and single stranded nucleotides under different evolutionary situations as mentioned above, Π is the structure of the other sequence, A is the alignment between the two sequences. $n_l(\pi)$ is the number of structure units in π whose thermodynamic parameter is θ_l. $m_k(\pi)$ is the number of structure units in π under such structural alignment situation that they will bring an extra term into the Boltzmann factor of π modelled as a pseudo thermodynamic parameter $ln(para_k)$.

This is the form of a log linear model. Given the known structure of the sequence to be modelled, the model is a convex function of $\{ln(para_i)\}_{1 \le i \le 7}$. A maximum likelihood method can be used to optimize $\{ln(para_i)\}_{1 \le i \le 7}$ [4], therefore ***para***. In this project, 50 tRNA sequence pairs and 50 5s rRNA sequence pairs with their known structures and alignment were drawn from the databases as the training set [5, 6]. The objective that was maximized is:

$$\frac{\sum_i lnP(\pi'_i | \boldsymbol{\theta}, \boldsymbol{para}, \Pi_i, \mathrm{A}_i)}{N} - \frac{1}{2}\alpha \sum_h [ln(para_h)]^2 \quad (2)$$

where i represents different sequence pairs in the training set, π'_i is the known structure of the sequence whose probability is to be optimized in the i^{th} sequence pair. α is the parameter for L_2 regularization. The optimization was carried out using the limited-memory BFGS method in the dlib c++ library [7]. A grid search was performed for different αs within {0, 3e-5, 0.0003, … , 3e10}. The performances for different αs were evaluated on a separate data set of 50 tRNA and 50 5s rRNA sequence pairs using the sum of the logarithm of the conditional probabilities of the structures in the data set. The optimal ***para*** with α=0 is:

Table 1. The optimal parameters representing evolutionary constraints.

$para_1$	$para_2$	$para_3$	$para_4$	$para_5$	$para_6$	$para_7$
1.02	3.23	1.00	1.30	0.59	0.55	0.98

Conclusion

According to the optimal parameters, the most favorable structural units are those with conserved structures in the homologous sequences. The most unfavorable structural units are those aligned with nucleotides with different structures in the homologous sequences. The implemented method can be used to study evolutionary constraints for different RNA families or build more complex models.

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. 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.
4. Smith N. Log-linear models. 2004.
5. 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.
6. 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.
7. King D. Dlib-ml: A machine learning toolkit. J Mach Learn Res. 2009;10:1755-8.