HAIRPIN MODELING IMPROVEMENT FOR NON-CODING RNA GENE SEARCH WITH COVARIANCE MODELS

Jennifer A. Smith*

Electrical and Computer Engineering Department, Boise State University,
1910 University Drive, Boise, Idaho 83725-2075, USA
**Email: jasmith@boisestate.edu*

The effect of more detailed modeling of the interface between stem and loop in non-coding RNA hairpin structures on efficacy of covariance-model-based non-coding RNA gene search is examined. Currently, the prior probabilities of the two stem nucleotides and two loop-end nucleotides at the interface are treated the same as any other stem and loop nucleotides respectively. Laboratory thermodynamic studies show that hairpin stability is dependent on the identities of these four nucleotides, but this is not taken into account in current covariance models. It is shown that separate estimation of emission priors for these nucleotides and joint treatment of substitution probabilities for the two loop-end nucleotides leads to improved non-coding RNA gene search.

1. INTRODUCTION

Laboratory studies indicate that there is a significant effect on RNA hairpin stability of the specific nucleotides at the interface between stem and loop[3]. Covariance models as currently used for database non-coding RNA gene search can not capture the thermodynamic regularities know from these laboratory studies. Ideally, modification of the covariance-model-based search algorithms such as Infernal[2] to jointly model the probabilities of the four nucleotides at the interface would solve this problem, but at the expense of significant programming effort. However, some of the benefits of joint modeling can be had by tricking the existing algorithms by using a P-type node for the loop ends and using a new set of priors for these nodes that depend on the consensus closing pair.

Limited testing on the fourteen shortest Rfam[1] families with a hairpin and without a pseudoknot show that specificity does seem to improve given fixed sensitivity when this trick is employed.

2. CHANGES TO MODEL STRUCTURE AND ESTIMATION

In the work of Vecenie and Serra[3] a number of regularities are noted regarding the thermodynamic stability of hairpin structures when different nucleotides are present in the stem-loop interface. For example, they note that if the closing pair is CG or GC and loop ends are GA or UU (but not AG), then the hairpin is much more stable.

It is hypothesized here that some RNA families may not be able to function as well with less stability in one or more of their hairpins. If this is so, then it would be desirable to penalize database search scores when the database sequence implies a mutation away from one of the very stable consensus configurations noted above. If the two loop-end L nodes are replaced by a single P node modeling these loop ends, expression of the joint probabilities of emission is possible. This can be accomplished simply by marking the two loop ends as if they were consensus base pairs in the input multiple alignment file to the *cmbuild* program of the Infernal program suite.

3. EXPERIMENTAL RESULTS

First, the entire Rfam 8.1 database was processed and all 26,644 hairpin structures in all the seed sequences extracted. Table 1 shows the log-likelihood ratios generated from counts of the number of observed loop-end pairs for each observed closing pair. Since wobble closing pairs are infrequent, they were not compiled separately, but are including the "All" column (such that the AU, UA, CG and GC columns do not add up to the All column).

The log-likelihood ratios of Table 1 were used as priors for loop-end P nodes on the fourteen shortest RNA families in the Rfam database which contained a hairpin without a pseudoknot.

[*] Corresponding author.

Table 1. Base-2 log-likelihood ratios of stem closing pairs given loop end nucleotides. Corrected for background nucleotide frequencies.

Loop	Stem Closing Pair				
End	AU	UA	CG	GC	All
AA	0.16	0.03	0.98	0.48	0.65
AC	-0.93	-2.89	-1.24	-1.75	-1.36
AG	-0.88	-2.76	-0.22	-2.33	-0.89
AU	-1.15	-1.94	-1.06	-1.70	-1.36
CA	1.91	2.77	0.32	-1.60	0.90
CC	1.43	-0.69	-1.76	-1.22	-0.55
CG	-1.64	0.11	1.12	-2.07	0.25
CU	-0.41	-0.61	0.19	-1.27	-0.29
GA	-0.25	-0.25	0.78	1.98	1.16
GC	-1.07	-1.66	-1.57	-1.97	-1.64
GG	-0.69	0.57	-1.04	-1.39	-0.70
GU	-1.90	-2.45	-2.49	-1.69	-1.98
UA	0.55	-0.96	-1.07	-1.02	-0.75
UC	0.18	0.69	-1.32	-0.38	-0.33
UG	-1.46	-3.01	0.75	-1.17	-0.11
UU	-0.02	-0.64	0.57	2.09	1.11

Table 2. Ratios of E-values using stem closing pair specific priors to E-values using standard priors on the full set (seed plus those found by search) of sequences in 14 Rfam families

RF	Family Properties		E-value ratios		
Acc.	Length	Number	Mean	Max	Min
00032	26	1046	1.64	2.20	1.02
00037	28	318	1.91	2.25	1.58
00453	33	30	2.67	3.60	1.81
00196	35	8	1.21	1.83	0.75
00180	36	30	1.82	3.01	1.08
00469	36	344	0.24	0.34	0.16
00385	41	41	1.66	2.42	1.09
00496	42	13	0.86	0.97	0.75
00164	42	302	1.32	1.91	0.87
00207	44	6	1.41	2.20	0.86
00617	45	426	1.47	2.43	1.16
00197	45	25	0.99	1.13	0.87
00500	45	5	1.58	2.63	0.66
00522	46	63	1.63	2.91	0.94
Mean			1.46		

Table 2 shows the results of the computational experiment. The E-value ratios shown are the ratio of the E-value using the standard covariance model divided by the E-value with the loop-end P node. Ratios greater than one mean that using the loop-end P node has more power than the standard model. A E-value ratio of two means that we expected twice as many false alarms from the standard model. On average, in only two cases (Rfam accession numbers RF00469 and RF00496) did modeling the loop ends jointly do significantly worse and in most cases it did quite a bit better.

4. CONCLUSIONS

Additional testing is needed to be more conclusive. In order to make this feasible, a more automated way to generate parameter files for Infernal needs to be developed (currently, it involves manual cut and paste and running a side program). Also, access to a computer cluster is needed to calculate E-values for many more and much longer sequences. These tasks are currently being undertaken by the author.

Acknowledgments

This material is based upon work supported by the National Institute of General Medical Sciences under grant NIH R15GM087646. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

1. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* 2005; **33**: D121-D124.
2. Eddy SR. *Infernal User's Guide*, version 1.0.2. http://infernal.rfam.org.
3. Vecenie C, Serra M. Stability of RNA hairpin loops closed by AU base pairs. *Biochemistry* 2004; **43**: 11813-11817.