Joint Loop End Modeling Improves Covariance Model Based Non-coding RNA Gene Search

Jennifer Smith

Electrical and Computer Engineering Department, Boise State University,
1910 University Drive, Boise, Idaho 83725-2075, USA
JASmith@BoiseState.edu

Abstract. 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.

Keywords: Sequence analysis, RNA gene search, covariance models.

1 Introduction

Covariance models are an effective method of capturing the joint probability information inherent in the intramolecularly base-paired positions of a non-coding RNA molecule [1, 2]. Unlike profile hidden Markov models [3, 4], which have a set of four emission probabilities over the possible nucleotides at each consensus sequence position, covariance models allow consensus base pairs to be assigned sixteen joint probabilities over the possible ordered nucleotide pairs. Covariance models also allow the probability of insertion or deletion of a base pair to be different than the sum of the marginal probabilities of insertion or deletion of the individual nucleotides. The profile hidden Markov model can be viewed as a special form of a covariance model with no base pairs specified.

Covariance models are finite state machines which require the estimation of state emission and state transition probabilities as well as model structure. This is normally done using a family of known sequences in a multiple alignment with secondary structure annotation. Counts of nucleotide frequencies in unpaired consensus columns or nucleotide pair frequencies in couples of base-paired consensus columns form the basis for emission probabilities. Counts of missing nucleotides in consensus columns and of nucleotide presence in non-consensus columns can be used to generate transition probabilities in and out of deletion and insertion states respectively.

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 27–37, 2010.

Conceptually, estimation of emission and transition probabilities is as simple as calculating the observed frequency of occurrence in the multiple alignment. The reality is much more complex. The very small number of family sequences that most RNA family models are estimated from is a major problem. In the *Rfam* 9.1 (December 2008) database of RNA alignments and covariance models, more than half of the 1371 family models are estimated from ten or fewer sequences [5, 6]. Most of the possible mutations, insertions, or deletions are never observed even though we have no particular reason to believe that they should be excluded from consideration. At very least pseudocounts need to be added to all possibilities such that the probability estimates do not outright exclude them. Pseudocounts are a form of prior information used in the estimation.

Far more informative priors than simple pseudocounts are needed for effective estimation of family models formed from so few sequences. Generic mutation, insertion, and deletion probabilities are obtained via observed frequency from the entire database of all RNA families. The generic emission and transition probabilities are found separately for base-paired and non-base-paired positions and with dependence on whether adjacent positions are paired or not. It will be demonstrated that these classifications are not quite fine enough later in this paper. In order to automatically uncover groups of mutation, deletion, or insertion patterns that tend to be observed together, these generic priors are estimated as a Dirichlet mixture [7] in recent versions of the Infernal [8] suite of programs for covariance-model-based RNA family analysis and search.

When combining the observed-frequency information from the multiple alignment of a specific family with the generic prior information, it is necessary to obtain a weighting based on our confidence in the family specific data versus our generic information. Having more sequences in the specific family increases our confidence in that data. However, simple counts of number of sequences are not very effective because our set of known sequences is rarely a random sample of actual sequences from the true complete family. We may have many sequences that are nearly identical and only a few with lots more diversity. This causes a simple count of number of sequences to overestimate the true information content. The usual solution to this problem is to employ entropy weighting based on the variability of the known family sequences [9].

There is a large literature on RNA secondary structure estimation based on primary sequence [10, 11]. Much of this literature uses the results of laboratory thermodynamic studies of RNA as its basis. These thermodynamic measurements are not used in covariance-model-based RNA family modeling. Instead, observed mutations, insertions, and deletions within the family or over the entire database (the priors) are used. However, it may be useful to study the regularities in RNA free energy measurements in the laboratory to guide choices in how covariance models are constructed. From the laboratory, we know that the identities of the nucleotides at the interface between the stem and the loop of a hairpin structure greatly affect thermodynamic stability of the hairpin structure. We also know that the length of the loop is a factor in stability. The mechanisms to capture these regularities are weak and nonexistent, respectively, in current covariance modeling practice. This paper will examine the stem/loop interface, but not loop length.

Some initial evidence that interface nucleotides and loop length might be important was found by Smith and Wiese [12]. This paper presents much more evidence for the stem/loop interface. It also looks at implementing a new type of node in the covariance model that can get around some of the problems encountered in tricking the existing Infernal program suite into handling the loop end nucleotides jointly.

The next section will review covariance models and estimation of model parameters in more detail. Section 3 looks at the regularities in free energy change when forming RNA hairpins observed in the laboratory. Changes to covariance model structure and parameter estimation procedure that can capture the observed thermodynamic regularities is presented in Section 4. Results of computational experiments on data from the Rfam database are presented in Section 5, followed by conclusions.

2 Covariance Model Structure and Parameter Estimation

Covariance models are finite state machines composed of emitting and silent states and directed edges connecting some of the states to some of the others. There is a unique starting state (called the root start state) and one or more terminal states (called end states). Given any nucleotide sequence it is possible to find the most probable mapping of the sequence onto model state visits and the associated overall probability of this mapping. Given a family of sequences, it is possible to find a set of state emission and state transition probabilities such that the overall probability when mapping a family member to the model is high and of mapping a dissimilar sequence to the model is low.

2.1 Model Structure

The states of a covariance model and the connectivity of these states can be determined from a consensus secondary structure of the RNA family. RNA secondary structure is a listing of pairs of sequence positions that intramolecularly base pair. The state structure can be described at a high level through the use of node trees, where nodes of a given class have identical internal state structure.

Figure 1 shows an example of a consensus secondary structure for an RNA family (right). The letters refer to the consensus nucleotides and the subscripts to the consensus sequence positions. The figure also shows the covariance model node tree for the same secondary structure. S, B, and E-type nodes contain no consensus emitting states. L and R-type nodes contain a single-emission consensus state and P-type nodes contain a pair-emission consensus state. The model is entered at the root start state located in the S0 node and has two exit points at the end states contained in nodes E12 and E22.

The node tree is simply a guide for constructing the underlying state model. The state model is the final model of interest. Figure 2 shows internal state structure of some of the nodes from the node tree in Figure 1. Nodes of the same type have the same internal structure, so constructing the state machine from the node tree is straightforward. There is a standard rule for how to connect edges from states in one

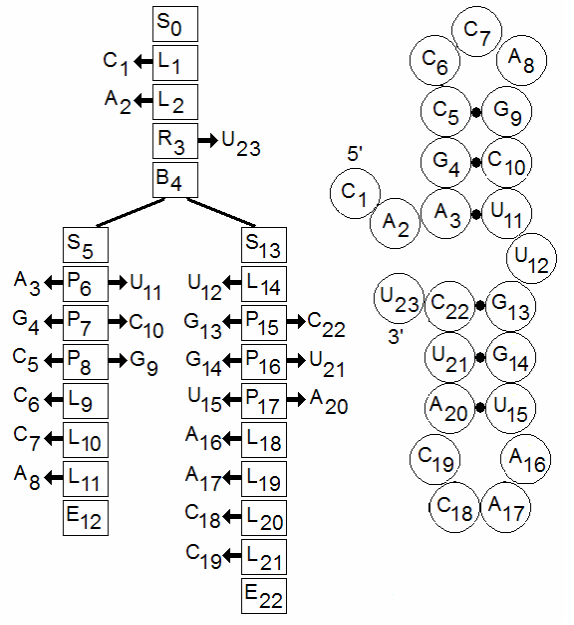

Fig. 1. An example consensus RNA secondary structure (right) and associated covariance model node tree (left)

node to states in an adjacent node. Each node contains one consensus state and varying numbers of non-consensus states. P, L, R, IL, and IR states types are emitting and all others are silent. D states allow for deletions relative to the consensus and IL or IR states allow for insertions.

2.2 Model Parameters

Once we have state structure, it is necessary to estimate emission probabilities for emitting states and transition probabilities for each edge connecting states. These probabilities are converted to log-likelihood ratios so that the total (log) probability of a particular path can be computed as the sum of transition and emission probabilities along the path. Dynamic programming can then be used to find the most probable path for a given sequence.

The parameters are estimated through a weighted combination of observed frequency of events in the family multiple alignment and the prior for the parameter. The priors in turn depend on the type of node holding the state and on adjacent node types. As an example, transition probabilities into and out of the D state in the R3 node at the top of Figure 2 would depend in part on the count of the number of gap characters in the twenty-third consensus column of the family multiple alignment. The R state in the R3 node is the consensus state which emits a consensus U and the D state in the R3 node is used to bypass this emission when a sequence has a deletion at this position relative to the consensus. Even though U is the consensus nucleotide for position 23, there are actually four emission probabilities associated with the R state in node R3. The probability for U is simply the highest of the four.

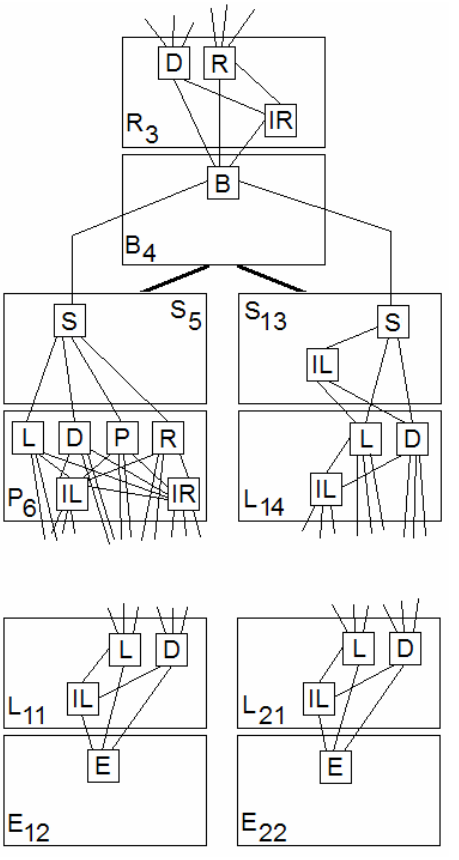

Fig. 2. Internal state structure of portions of the example covariance node tree from Figure 1

3 Thermodynamic Regularities

The thermodynamic stability of RNA hairpins is a fairly well studied topic [13-18]. Using calorimetry observations of the folding of short synthetic strands of RNA, models of the free energy of larger hairpin structures can be inferred. These models are used extensively in algorithms to predict secondary structure of RNA from sequence. These algorithms are based on the idea that the final conformation of an RNA molecule will be close to that of the minimum free-energy conformation.

Two of the major observations from the laboratory data is that hairpin stability depends on the number of nucleotides in the loop and on the identities of the four nucleotides at the stem-loop interface. The loop-length observation is relevant to covariance models and should be addressed, but the focus in this paper is on the stem-loop interface observation.

In Figure 3, the stem-loop interface is composed of the closing pair U15 and A20 as well as the loop ends A16 and C19. Although the structure appears symmetric in the figure, the free energy of the structure shown for GGUAACCAUC is different than its mirror CUACCAAUGG. In other words, it maters which side of the

stem-loop interface is 5' and which is 3'. Covariance model P nodes can emit any of the sixteen possible ordered pairs of nucleotides. In the middle of a stem it makes sense to allow all sixteen possibilities since a mutation from a Watson-Crick or wobble base pair (a canonical base pair) to a non-canonical pair can be held together by adjacent base pairs in the stem without necessarily destroying the stem. If the closing pair becomes non-canonical, then effectively the loop length increases by two and the next pair up the stem becomes the closing pair. So, there are really only six consensus ordered pairs to consider for the closing pair: AU, UA, CG, and GC (Watson-Crick) as well as the wobble pairs GU and UG. In the Rfam database, consensus wobble pairs are very infrequent at the closing pair position (observed only about 4.1% of the time in version 8.1).

In the work of Vecenie and Serra [13] a number of regularities are noted regarding the thermodynamic stability of hairpin structures when different nucleotides are present in the stem-loop interface. 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. They also note that if the closing pair has a purine (A or G) on the 5' side, the GG loop ends are particularly 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. Unfortunately, covariance model structure and parameter priors do not allow for these thermodynamic regularities to be expressed either directly or indirectly.

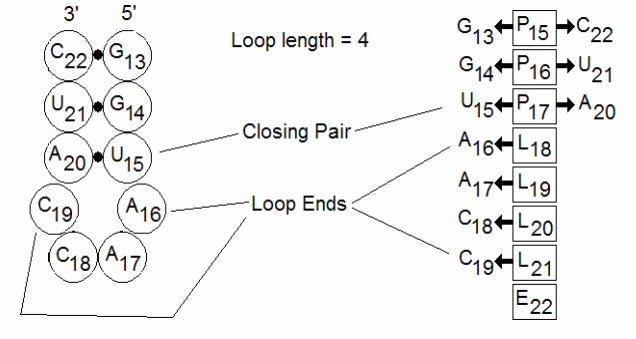

Fig. 3. A portion of the RNA secondary structure and covariance node tree from Figure 1 showing a single hairpin with the locations of the stem's closing pair and the loop ends labeled

4 Changes to Model Structure and Estimation

A major problem making expression of the thermodynamic regularities described in the previous section not possible is that the four nucleotides in the stem-loop interface are contained in three covariance model nodes with independent emission probabilities. Another problem is that the priors used for these emission probabilities are estimated as a mixture of database locations corresponding to stem-loop interfaces and to other structures.

To allow for expression of a regularity such as stable GG loop ends when the 5' side of a closing pair is A or G requires a new type of covariance model node. Such a node replaces a P node and two L nodes of a hairpin structure. In Figure 3, these are the P17, L18, and L21 nodes. Two hundred fifty six joint emission probabilities are needed for the consensus state of this node type. Since 160 of these combinations are not seen in practice (the combinations with non-canonical closing pairs), they can simply be assigned a very low probability, leaving only 104 emission probabilities that need to be estimated. Since wobble pairs are relatively rare, it may also be desirable to treat them as a class with a single emission probability (but a different value than for non-canonical pairs). This would leave 64 emission probabilities to be estimated for the Watson-Crick closing pairs. Clearly, heavy reliance on priors for these probabilities is needed since so few families have known sequences numbering in the hundreds and even fewer have enough variation in the observed stem-loop interface nucleotide combinations.

Implementation of a new node type requires significant programming effort to rewrite program suites such as Infernal. A partial solution is to at least express the joint probability of the two loop end nucleotides by tricking the existing algorithms. 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. In Figure 3, the L18 and L21 nodes would be removed and replaced by a single P18 node directly below the existing P17 closing-pair node. In practice 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.

Using the P-node substitution trick does cause a couple of problems with priors. Firstly, The closing-pair P node will now use priors associated with a P node with P node child rather than the correct P node with L node child priors. This first problem can be solved by running the *cmbuild* program twice, once with and once without the loop ends marked as base paired. Then parameter estimates for the closing-pair P node in the second run are used in place of the estimates in the first run. The second problem is that the priors for the fake loop-end P node are completely wrong. The standard P node priors are generated from stem locations in the overall Rfam database with high probabilities for Watson-Crick base pairs, somewhat lower probabilities for wobble pairs and very low probabilities for non-canonical pairs. Instead, sets of priors for these loop-end P nodes are estimated on the side, one set for each possible consensus closing pair.

The loop-end P-node trick allows for a one-way dependence of loop-end emission probabilities on consensus closing pairs. It would be possible to also regenerate sixteen sets of priors for closing-pair P nodes and use the one associated with a given family's consensus loop ends. This two-way dependence would still not be quite as good as full use of joint probabilities.

5 Experimental Results

This section looks at results of using a P-node to model loop ends with non-standard priors on the loop-end P node only (and not for the closing pair P node).

First, the entire Rfam 8.1 database was processed and all 26,644 hairpin structures in all the seed sequences extracted. Since some RNA families have no hairpins and others have multiple hairpins, this number is different than the total number of seed sequences in the database. Table 1 shows the raw counts of 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). These raw counts are not that useful because the background frequencies of A, C, G and U are not each one quarter. To remedy this, Table 2 shows the same data as base-2 log-likelihood ratios. The log form is what is used by Infernal in order that the algorithm calculate additions instead of multiplications and it is visually useful since positive values are more likely than chance and negative less likely.

Some of the regularities noted in section 3 are apparent in Table 2. GA and UU loop ends are overrepresented by a factor of four when the closing pair is GC and by a factor of two when the closing pair is CG (but not for AU or UA closing pairs). Some other combinations have deviations of up to a factor of eight (for example UG loop ends on a UA closing pair).

The log-likelihood ratios of Table 2 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. Pseudoknots are a situation where at least one pair of base pairs is such that neither base pair is completely between the other in sequence [19]. Covariance models use stochastic context-free grammars [20], which are incapable of describing a pseudoknot. Covariance models handle pseudoknots by treating some of the actually base-paired positions as if they were unpaired. Since what appears to be a hairpin in the node tree of pseudoknotted RNA families is actually something somewhat more complex, they will not be considered. The amount of computation time require to calculate E-values for covariance models is extremely high and goes up by more than the square of sequence length and short sequences are the most difficult to find in database search, so short sequences were chosen for this experiment.

Table 3 shows the results of the computational experiment. The first two columns show the length of the consensus sequence and the number of known family sequences. Both the seed sequences used to construct the family models and those found through database search by the curators of Rfam are included in this number. E-values are calculated by the Infernal program suite by reshuffling the known sequence many times (5000 times chosen for this study), scoring each reshuffled sequence against the family covariance model and then and fitting the resulting scores to a Gumble extreme value distribution [21]. The score of the unshuffled sequence is then used to find the probability of matching or exceeding the unshuffled score by pure chance. Lower E-values imply better specificity given that the threshold is set such that the sequence is just barely accepted as a true positive. 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.

Table 1. Counts of loop-end nucleotides in the full Rfam database (in 26,644 hairpins from all seed sequences from Rfam 8.1)

Loop End	Stem Closing Pair				
	AU	UA	CG	GC	All
AA	318	302	2173	1098	4054
AC	94	25	293	147	628
AG	113	32	694	114	1013
AU	110	66	454	208	859
CA	671	1269	865	163	3007
CC	301	72	128	133	692
CG	42	146	1099	86	1405
CU	115	104	678	175	1133
GA	175	182	1387	2270	4202
GC	62	43	170	92	378
GG	94	235	285	160	844
GU	48	34	123	153	410
UA	359	131	450	332	1318
UC	174	257	238	324	1104
UG	65	23	1158	219	1495
UU	207	140	1204	2459	4102
All	2948	3061	11399	8133	26644

Table 2. Base-2 log-likelihood ratios using raw data from Table 1 (corrected for background frequencies of A, C, G, and U)

Loop End	Stem Closing Pair				
	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 3. 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		

6 Conclusions

Laboratory studies indicate that there is a significant effect on RNA hairpin stability of the specific nucleotides at the interface between stem and loop. 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 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 than depend on the consensus closing pair.

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

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. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
2. Eddy, S.R., Durbin, R.: RNA Sequence Analysis Using Covariance Models. Nucleic Acids Research 22, 2079–2088 (1995)
3. Karplus, K., Barrett, C., Hughey, R.: Hidden Markov Models for Detecting Remote Protein Homologies. Bioinformatics 14, 846–856 (1998)
4. Eddy, S.R.: Hidden Markov Models. Curr. Opp. Structural Biology 6, 361–365 (1996)
5. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., Bateman, A.: Rfam: Annotating Non-coding RNAs in Complete Genomes. Nucleic Acids Research 33, D121–D124 (2005)
6. Rfam, R.N.A.: Families Database of Alignments and Covariance Models, version 9.1 (2008), `http://rfam.janelia.org`
7. Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., et al.: Dirichlet Mixtures: a Method for Improving Detection of Weak but Significant Protein Structure Momology. Comp. Appl. Biosci. 12, 327–345 (1996)
8. Eddy, S.R.: Infernal User's Guide, version 1.0.2 (2010), `http://infernal.rfam.org`
9. Nawrocki, E., Eddy, S.R.: Query-dependent Banding (QDB) for Faster RNA Similarity Searches. PLoS Comp. Bio. 3, 540–554 (2007)
10. Zucher, M.: Computer Prediction of RNA Structure. Methods Enzymology 180, 262–288 (1989)
11. Wiese, K.C., Hendricks, A.: A Hybrid Clustering/Evolutionary Algorithm for RNA Folding. In: Symp. Comp. Intelligence Bioinformatics Comp. Biol., pp. 15–21. IEEE Press, New York (2008)
12. Smith, J.A., Wiese, K.C.: Integrating Thermodynamic and Observed-Frequency Data for Non-coding RNA Gene Search. In: Priami, C., Dressler, F., Akan, O., Ngom, A. (eds.) Trans. Computational Systems Biology X, pp. 124–142. Springer, Berlin (2008)
13. Vecenie, C., Serra, M.: Stability of RNA Hairpin Loops Closed by AU Base Pairs. Biochemistry 43, 11813–11817 (2004)
14. Dale, T., Smith, R., Serra, M.: A Test of the Model to Predict Unusually Stable RNA Hairpin Loop Stability. RNA 6, 608–615 (2000)
15. Serra, M., Little, M., Axenson, T., Schadt, C., Turner, D.: RNA Hairpin Loop Stability Depends on Closing Base Pair. Nucleic Acids Research 21, 3845–3849 (1993)
16. Serra, M., Axenson, T., Turner, D.: A Model for the Stabilities of RNA Hairpins Based on a Study of the Sequence Dependence of Stability for Hairpins with Six Nucleotides. Biochemistry 33, 14289–14296 (1994)
17. Giese, R., Beschart, K., Dale, T., Riley, C., Rowan, C., Sprouse, K., Serra, M.: Stability of RNA Hairpins Closed by Wobble Base Pairs. Biochemistry 37, 1094–1100 (1998)
18. Freier, S., Kierzek, R., Jaeger, J., Sugimoto, N., Caruthers, M., Neilson, T., Turner, D.: Improved Free-Energy Parameters for Predictions of RNA Duplex Stability. Proc. Natl. Acad. Sci. USA 83, 9373–9377 (1986)
19. Staple, D., Butcher, S.: Pseudoknots: RNA Structures with Diverse Functions. PLoS Bio. 3, 956–959 (2005)
20. Chomsky, N.: Three Models for the Description of Language. IRE Trans. Information Theory 2, 113–124 (1956)
21. Gumbel, J.: Statistics of Extremes. Columbia University Press, New York (1958)